

Predicting repayment success in IT-mediated lending: the value of soft information

Dennis M. Steininger, T. Wagenfuehrer, Daniel Veit

Angaben zur Veröffentlichung / Publication details:

Steininger, Dennis M., T. Wagenfuehrer, and Daniel Veit. 2012. "Predicting repayment success in IT-mediated lending: the value of soft information." In Proceedings of SIG-IQ Pre-ICIS Workshop, Orlando, Florida, USA. New York, NY: AISel.



Predicting Repayment Success in IT-Mediated Peer-to-Peer Lending: The Value of Soft Information

Abstract

The growth of IT-mediated online peer-to-peer (P2P) lending may have repercussions on the development of the financial industry. Due to the low costs and high benefits of deception, we hypothesize that borrowers with private information about their true high risk have incentives to misrepresent to improve credit conditions and funding success. To test our proposition, we derive linguistic artifacts of deceptive language. Using Content Analysis, we examine the relation of these artifacts to repayment performance through hard information and descriptions of 1099 loan projects on the P2P lending platform LendCo. While we observe that results are only robust for texts beyond 48 words in length, we find that (i) lenders make inefficient lending decisions, (ii) available hard information, such as the credit grade, is not sufficient for an accurate risk indication and (iii) borrowers who compose more expressive, affective and less complex loan descriptions are significantly more likely to default.

Introduction

Peer-to-peer (also P2P, person-to-person, or social) lending allows private lenders to issue small loans to borrowers via an online platform without traditional intermediaries. Having been declared a breakthrough business idea in the Harvard Business Review in 2009 (Benyus et al. 2009), the online P2P lending market has seen both its purpose diversify as well as its growth rapidly increase across fields. *Gartner* assigns P2P lending the potential to capture up to 10 percent of the small loan market within a few years, facilitating a loan volume of about € 3.8 bn until 2013 (Gartner IT Research 2010). This growth could precede fundamental changes in the financial sector (Wang et al. 2009). Literature delineates the market-creating impact of the Internet and its dependence on adjacent mechanisms such as information asymmetry and signaling in P2P lending (Wang et al. 2009). This paper is motivated by a problem inherent to P2P lending: Lenders need to decide whether or not to lend solely on the borrower's information such as financials and project descriptions on a website. Many researchers suggest that the incapability of private lenders to decide on trust- and creditworthiness ex-ante puts the business model at risk. Two reasons amplify this issue. First, borrowers might be adversely selected if they use P2P lending since they do not qualify for a traditional loan (Akerlof 1970). Second, in the absence of professional screening by a third party, borrowers are tempted to deceive or present facts (signaling) in an overly optimistic way in order to increase their *funding success*. Literature suggests that the ignorance or misinterpretation of relevant publicly available information is a frequent reason for misvaluation (Hirshleifer 2001). Previous P2P literature compliments this through assessing the predictive capacity of soft information on lending profitability by trying to understand if borrower-provided texts comprise hints on creditworthiness that lenders overlook (Greiner and Wang 2010; Herzenstein et al. 2008; Larrimore et al. 2011; Moulton 2007). If lenders were incapable of processing available information, then the imperfection of online disintermediation would diminish the competitiveness.

Despite its importance, the assessment of the relationship between ex-ante borrower-provided data and repayment performance has been impossible to date (Berger and Gleisner 2009). Due to the newness of P2P lending, research focused on determinants of *funding success* since ex-post repayment data was not available (Herzenstein et al. 2011). However, the first 36-month loans have recently expired and become available for research. To the best of our knowledge, to date there exist only two papers that take advantage of this fact by analyzing P2P lending. The first notes that many borrowers offer similar explanations for why they need a loan. These explanations might be classified into different categories such as “*denial of hard data*”, which significantly predict *funding success* (Sonenshein et al. 2011). The second article investigates the impact of identities (e.g. “*hard-working*”) that borrowers create in self-

description texts (Herzenstein et al. 2011). They find a significant effect of identities on *funding success*, and in a follow-up analysis, on *repayment success*. However, both focus on the prediction of *funding success* and hence rather take the perspective of a borrower, as opposed to a lender. By analyzing language particularities, both papers indirectly follow the idea that psychology plays an important role on P2P lending websites. However, they do not systematically question the motivation or incentive that creates the observed effects. To fill this gap, we analyze the incentives of asymmetrically informed higher-risk borrowers (lemons) to falsely signal low-risk, and the capability of lenders to draw correct inferences from verifiable hard information in spite of such signals. Assuming that false signals are a form of deception, we approach our analysis using the *Interpersonal Deception Theory*, which is also used in criminalistics. The theory suggests that deception is “*imperfect strategic behavior*” (Buller and Burgoon 1996). The imperfection leaves language artifacts or cues that can be detected when analyzing language style. Therefore, we aim to answer the research question:

Can the occurrence of deceptive cues in soft information of IT-mediated P2P lending project descriptions explain repayment success?

Based on the structuring foundations of agency theory (Akerlof 1970), we use ‘soft’ textual descriptions and hard information (as controls) such as the credit grade from 1099 loan projects of the P2P lending platform *LendCo*. After modifying a dictionary to entirely cover the *Interpersonal Deception Theory*, computer-supported Content Analysis and multiple regressions are applied analogous to previous deception research for evaluating on the research question (Zhou et al. 2004a). To our best knowledge, this is the first study to blend research on automated deception detection in a computer-mediated context with the evaluation of economic transaction outcomes. Our findings can add to the ongoing discussions in IS on the analysis of online reviews and profiles (Pavlou and Dimoka 2006). Finally, for practitioners, we can support or reject the potential of automated deception detection for a future real-life application and provide the blueprint for deception recognition software. The remainder is structured as follows: we introduce agency and interpersonal deception theory section two and develop our hypotheses. We familiarize the reader with our used methodology in a next step and present the results, which we discuss and interpret. We conclude and provide directions for future research in the last section.

Theoretical Foundations and Hypotheses

We draw on agency theory to structure our research, understand information asymmetries in P2P lending and the incentives for borrowers for over-optimistic signaling (e.g. lying). Agency theory conceptualizes the goal conflicts between two partners in economic transactions where bounded rationality, fears of opportunism and information asymmetries exist (Milgrom and Roberts 1992). It has been applied in IS for numerous studies (e.g. Dibbern et al. 2004; Pavlou et al. 2007). To investigate on the explanatory

power of deception as a predictor of *repayment success*, we first need to understand the decision-making process of investors, the value of borrower signals and the accuracy by which ex-ante hard information predicts *repayment success*. Some pieces of the information consist of exogenously verified hard facts; others (i.e. soft information such as the textual description why a credit is needed) can be influenced by the borrower and thus represent purposeful signals which might be susceptible to deception. P2P lending captures a niche in terms of additional risk evaluation through soft information which banks fail to take into account. Hard information is found to be less accurate for low *credit grades*, where soft information explains up to 39 percent of risk, which indicates that soft facts are more important when hard information convey a negative image of creditworthiness (Moulton 2007, 2007). Borrowers supposedly try to mitigate the negative effects of exogenous hard information by changing endogenous hard information (e.g., adapting the *interest rate*) or conveying a positive image by providing reassuring soft information. This may be particularly true for high-risk borrowers, whose incentive to persuade seems exponential as they could gain relatively higher benefits from it (Iyer et al. 2009).

Soft Information, Lying and Interpersonal Deception Theory

To further specify our deception hypothesis, we need a definition of Lying. A lie can be defined as “*intentional misrepresentation of information [...] to achieve some preconceived end*” (Ford et al. 1988, p. 554). Deception is defined as “*intentional control of information [...] to create a false belief in the receiver*” (Hancock 2007, p. 290; Zhou et al. 2008, p. 119). Hence, both lie and deception are defined as an intent to create a false impression. Hence, we use both notions synonymously with the meaning of *intentional control or misrepresentation of information to create the image of higher-than-true creditworthiness with the ultimate goal of funding success*. The incentive to deceive depends on individually perceived costs and benefits (Hurkens and Kartik 2009). Potential benefits from deception depend on the accuracy of verifiable hard facts: the more risk can still be explained by soft information, the more a borrower can differentiate by providing overly positive soft information. Relatively riskier borrowers within a given risk category have, *ceteris paribus*, a higher incentive to misrepresent themselves than lower-risk ones since they can expect higher potential benefits. The anonymous and unmediated online setting generally makes deception easier, not only in lending (Caspi and Gorsky 2006; Herzenstein et al. 2011; Horne et al. 2007; Larrimore et al. 2011; Utz 2005). The Internet has increased “*physical, psychological, cultural [and] social distance*” (Jones 1991, p. 372) between transaction parties as well as between decision and effects. Accordingly, research suggests that the reduction in personal communication has decreased mutual empathy (Logsdon and Patterson 2010) and increased ambiguity and uncertainty in message decoding due to the lack of nonverbal cues (Daft and Lengel 1986). These

tendencies increase the ease of and eagerness to engage in deception (Joinson and Dietz-Uhler 2002; Logsdon and Patterson 2010, 2010). In other words, it seems that the Internet has reduced lying cost. Researchers agree that the main factors determining these cost are the consequences, the proximity to the victim, and the societal view if it was to be considered a fair or unfair deception (Jones 1991; McMahon and Harvey 2006).

We assume that, for lemons, lying benefits exceed costs in P2P lending and lying costs are generally low: it is difficult to establish a valid legal argument for discrepancies between repayment failure and the promises made in descriptions. Moreover, social costs are low since the community at *LendCo* is not yet evolved. Also, for most one-off borrowers, a degradation of the future credit score is unproblematic. Finally, we have elaborated that psychological costs are generally lower in computer-mediated communication. Assuming low lying costs, the incentives to deceive might be turned into actions particularly by those with potentially high benefits – lemons. The lower accuracy of hard information for lower *credit grades* reported in existing studies illustrates that the higher default rate could be connected to successful deceptive behavior of lemons. Intuitively, loans can only default if they were successfully funded before. Hence, soft information of high-risk projects on *LendCo* seems, on average, persuasive (otherwise lenders would not have funded them) but also, on average, deceptive, since otherwise they would not have defaulted on an above-average level.

How can we relate deception to default? The style of language represents an informational meta-function of texts which Freud already in his *Psychopathology of Everyday Life* (1901) assumes to unveil hidden intentions. The *Interpersonal Deception Theory* institutionalizes the detection of liars through their language style based on a set of linguistic constructs also used in criminalistics (Buller and Burgoon 1996). Buller and Burgoon (1996) propose that liars subconsciously stand out since their motives leave artifacts in their language. Tests and meta-analyses of the *Interpersonal Deception Theory* have confirmed that liars manipulate clarity, relevance, association, truthfulness and completeness for these reasons (Buller and Burgoon 1996; Burgoon et al. 1996; DePaulo et al. 2003). We find numerous studies operationalizing such cues to deception (Anderson et al. 1999; DePaulo et al. 2003) but only include empirical studies based on linguistic cues since non-verbal cues are not available in online profiles (Toma and Hancock 2010). We summarize the results of our review in Table 1 as a concept matrix (Webster and Watson 2002). Hence, researchers expect to find artifacts of deceptive communication, e.g. liars would communicate in a less *complex* and *diverse* manner since they lack real knowledge about the issue (Zhou et al. 2004b): To systematize the heterogeneity, literature spots two moderating factors for effect and direction: the medium (email vs. instant chat, etc.) and mode (verbal vs. written).

Table 1. Concept Matrix – Linguistic Deception Detection (Only Empirical Studies)

Author(s)	Journal (J),		Setting	Quantity	Expressivity	Pos. Affect	Neg. Affect	Informality	Uncertainty	Immediacy	Complexity	Diversity	Specificity
	Conference (C)	Medium											
(Burgoon et al. 2003)	ISI '03 (C)	Audio, chat	Sync.	+ [#]	+ [#]	+ [#]	+ [#]	+		-	- [*]	-	- [*]
(Zhou et al. 2003) ³	ISI (C)	Email	Async.	+ [*]	+	+ [*]	+ [*]	+ [*]		-	+	- [*]	-
(Newman et al. 2003)	PSPB (J)	FtF ¹ , essay	both				+ [*]			- [*]	- [*]		
(Zhou et al. 2004a)	GDN (J)	Email	Async.	+	+		(+)	+	+ [*]	- [*]	- [*]	- [#]	- [#]
(Hancock et al. 2004) ⁴	SIGCHI '04 (C)	FtF	Sync.	- [#]	+	+	+			- [*]	-		
(Zhou et al. 2004b)	JMIS (J)	Email	Async.	+ [*]	+ [*]		(+)	+ [*]	+ [*]	- [*]	- [*]	- [*]	-
(Qin et al. 2005)	HICSS '05 (C)	Text, audio, FtT	both	-		-	-		+ [#]	- [*]	- [*]	- [*]	- [#]
(Zhou and Zhang 2006)	SGR (J)	Chat (IM)	Sync.	+	+						+ [*]		+
(Zhou and Zhang 2008)	CACM (J)	Email (SM ²)	Async.	+ [*]	+ [*]	+	+ [*]	+ [*]	+ [*]	- [*]	- [*]	- [*]	
(Hancock et al. 2007)	DP (J)	Email (SM)	Async.	+			+			- [*]			+ [*]
(Toma and Hancock 2010)	ACM '10 (C)	Dating profile	Async.	- [*]			+ [#]			- [*]			

Legend: “+”=hypothesis that liars would use more of the respective construct; “-”=hypothesis that liars use less of the respective construct; “()”=tentative finding; blank=not analyzed; “*”= significance hypothesis supported; “#”= opposed direction found significant; ¹FtF”=Face to Face; ²SM”=Short messages; ³Predecessor of Zhou et al. 2004; ⁴Predecessor of Hancock et al. 2008

Carlson et al. (2004) are first to integrate the *Computer-Mediated Communication Theory* and the *Interpersonal Deception Theory*. They argue that linguistic features are well suited to detect deception in computer-mediated communication since it is well documented through IS. Characteristics of a medium facilitate or prevent the occurrence of deception (e.g. capacity to store and edit text) (Carlson et al. 2004, p. 13). This indicates that P2P lending platforms might facilitate deception. Drawing from findings on computer-mediated communication, Zhou et al. suggest that on the Internet, the distance between sender and receiver decreases negative emotions experienced when lying. Moreover, the ease of communication control and editability of messages provoke that the persuasive intent crowds out other goals. As a consequence of computer-mediation deceivers would write more to be more persuasive (higher *quantity*), also using more (instead of less) *affect* and *expressivity* on purpose. However, liars would still exhibit cues such as *informal* language and others outlined above (Zhou et al. 2004b). The only other paper on deception detection which is also based on online profiles is a study on deception in online dating profiles, which finds weak support for the deception hypothesis (Toma and Hancock 2010). Summing up, we infer that the significant occurrence of deception constructs in a borrower’s loan description in the direction predicted by deception theory would suggest deceptive intent. Moreover, we propose that deceptive intent is an indicator for high true risk since the benefit of lying is highest for lemons while we assume that costs are uniformly low for all borrowers. Since high true risk would result in low *repayment success*, we can analyze the relation between the frequency at which deceptive constructs are used in loan descriptions and the loan’s respective *repayment success* to prove our proposition. Therefore, we posit:

Hypothesis 3a-i: Borrowers that exhibit (a) higher quantity, (b) expressivity, (c) affect, (d) informality and (e) uncertainty, but lower (f) immediacy, (g) complexity, (h) diversity and (i) specificity in their loan descriptions have a lower repayment success.

Summary of Hypotheses

Figure 1 gives an overview of all hypotheses and controls in our conceptual model. **H1a-i** test the effect of the constructs on *repayment success*, controlling for signals, promised return and ex-ante risk indicators. **C1** controls for *gender, occupation, age* and the *quantity* of signals on promised return. **C2a-c** control the effect of borrower signals and ex-ante risk indicators on the lender’s profitability assessment. **C3a-e** control the effect of individual signals, promised return and indicators on *repayment success*, while **C4** is controlling for *gender, occupation* and *age* on *repayment success*.

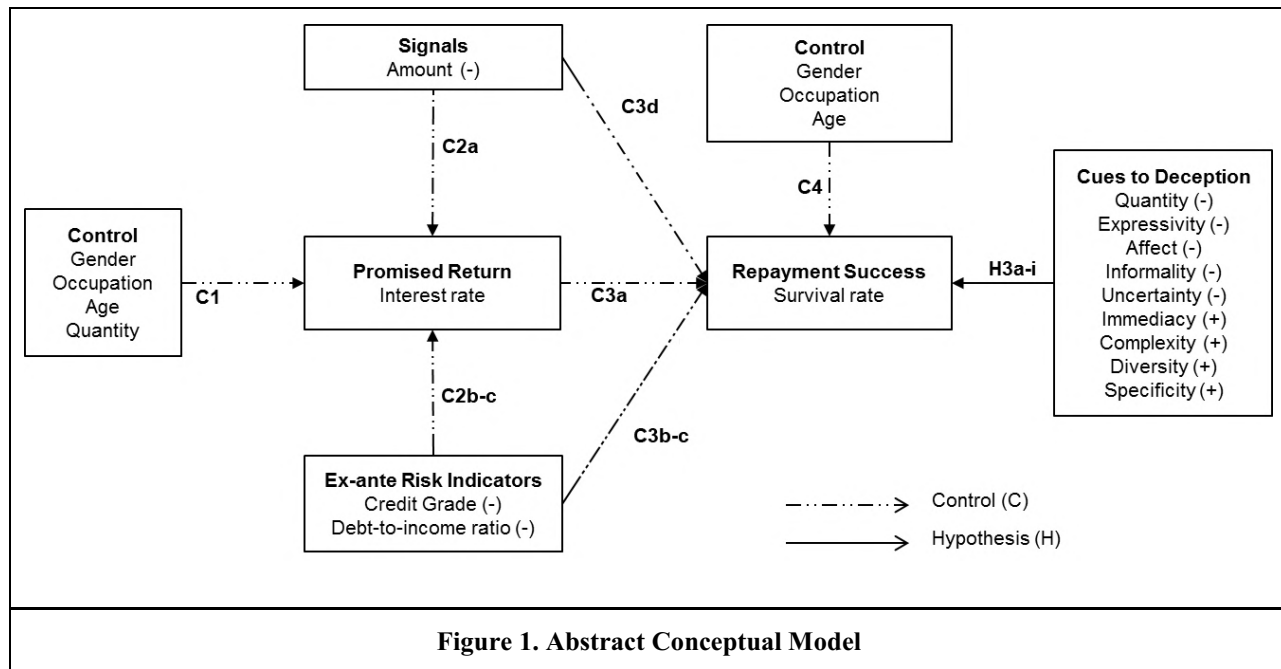


Figure 1. Abstract Conceptual Model

Methodology

Research Design

In our analysis, we assume that deception is an inverse indicator for the true ability to repay since we suggest that the deceptive intent is caused by a low ability to repay. The true ability to repay causes *repayment success* (or failure). To support causation assumptions, we require to show ‘*association*’ (correlation), ‘*isolation*’ (the rejection of alternative hypotheses) and ‘*temporal precedence*’ for the relationship between the deception constructs and our dependent variable *survival rate* (Cook and Campbell 1979; Gefen et al. 2000, p. 40). Since borrowers compose their loan descriptions before they

can default, temporal precedence is naturally established. We demonstrate association and isolation with the help of Content Analysis and statistical tools for a confirmatory data analysis (Shi and Tao 2008). In confirmatory data analysis, we infer association from a correlation of our independent variables with the degree of *repayment success*.

Data Collection and Sampling

We chose *LendCo* as our data source due to very transparent lending processes and the availability of textual project descriptions from the borrowers for each funded project. To make inferences from a sample on the population of *LendCo* users and P2P borrowers in general, the sample needs to be appropriately sized to minimize both type I errors – that we do reject our null hypothesis although it is true – and type II errors – that we do not reject our null hypothesis though it is truly wrong (Petter et al. 2007). Maxwell elaborates that he considers a ratio of 70:1 to be appropriate (Maxwell 2000, p. 454). For our 13 predictors, this would require a sample of at least 910 observations. We consider loans with a 36 months duration that were issued in 2007 and 2008 to only include contracts with a completed period. This results in 1,099 observations.

Data Pre-Processing and Content Analysis

For our correlation analysis, qualitative constructs based on text, such as *quantity* and quantitative variables, such as the *credit grade*, need to be tested in a single empirical analysis. We use content analysis to transform qualitative textual constructs into measurable factors (Berelson 1952). The use of Content Analysis has several advantages. It enables the examination of rich communication data previously untapped by merely quantitative studies. Moreover, its observing perspective on communication (Barley et al. 1988) avoids the risk of influencing the behavior of borrowers, which other methods, such as surveys, would be susceptible to, especially when asking lyers. We follow the procedures proposed by Insch et al. (1997) and deductively derive cues to deception from literature for our coding scheme. Similar to Insch et al. (1997), Homburg and Giering subdivide validity into content (also called construct), convergent, discriminant and nomological validity (Homburg and Giering 1998). We ensure nomological validity by embedding the constructs in the *Interpersonal Deception Theory*. We test for discriminant validity across constructs in multicollinearity pre-tests where we exclude irregularly cross-correlated categories (Morris 1994; Weber 1990). Finally, we aim to maximize content- and convergent validity through selecting coding categories from previous deception literature. To reduce the correlation between the *word count* and the other variables, we standardize all variables dividing them by *word count* analogous to previous deception research, e.g. (Zhou et al. 2003, 2004b).

Statistical Data Analysis

With our correlation analysis, we would like to make a *prediction* about the standalone and joint effect of linguistic variables on one metric dependent variable. Multiple linear regression tools would fit these requirements which estimate a function that predicts the *survival rate* based on a linear combination of linguistic variables. We first test for multicollinearity, also to verify the empirical discriminant validity of our independent variables. The usual measures for multicollinearity include the *variance inflation factor* (VIF) and the *condition number* for the independent variables (Baum 2006; O'Brien 2007). For our sample, STATA reports VIF values below ten.

Second, we observe the distribution of the dependent variable in our final sample. Naturally, the fractional variable *survival rate* is bounded by zero and one. Moreover, the data are strongly skewed towards the upper boundary since most borrowers do not default. These characteristics lead to inaccuracies when using an OLS regression (Gefen and Rigdon 2011) or SEM (Gefen et al. 2000) since both base on the assumption of normally distributed error terms. When analyzing our data, we find that heteroscedasticity and non-linearity in the regression residuals violate two of the usual Gauss-Markov assumptions for classical linear models. Both the Breusch-Pagan/Cook-Weisberg test and the White test for heteroscedasticity in our sample strongly reject the null hypothesis that the variance of the residuals is homogeneous ($\chi^2=230.98$, $p<.00001$). Using Cameron and Trivedi's decomposition measure, we also need to reject the null hypothesis that the distribution is evenly distributed ($\chi^2=171.11$, $p<.00001$) or mesokurtic ($\chi^2=59.93$, $p<.00001$). We find a trend to the residuals versus predicted values which should not occur in case of homoscedasticity (Baum 2006). We chose the linear regression to analyze our data. For multiple regressions, standard solutions to the violations discussed above would suggest to either take the natural logarithm of the dependent variable (Gefen and Rigdon 2011; Neter 1996), analogous to Herzenstein et al.'s approach (2011), or to square all independent variables (Baum 2006). However, since fractional data is naturally bounded by [0,1], the OLS model predicts values outside the possible range when applying these remedies (Papke and Wooldridge 1996). Papke and Wooldridge (1996) found a solution for the issue. They develop a generalized least square model (GLS) specifically for skewed fractional dependent variables. To do so, they specify the GLS model as part of the *binomial family* with a *logit link function* to account for its fractional nature and *robust standard errors* to balance the heteroscedasticity. Analogously to Papke et. al (1996), we model:

$$E \left[Survivalrate \left| \begin{matrix} \rightarrow & \rightarrow & \rightarrow \\ S_i & H_j & C_k \end{matrix} \right. \right] = G(\alpha + \beta_i \times S_i + \beta_j \times H_j + \beta_k \times C_k)$$

where $E[\textit{survival rate}|S_j, H_j, C_k]$ represents the expected value of the *survival rate* under the conditions specified through the vector of signals S_i , hard information H_j and linguistic categories C_k . G represents the logit link function (Papke and Wooldridge 1996, p. 621), α the estimated intercept and β_i , β_j and β_k the estimated coefficients of variables S_i , H_j and C_k , respectively. By standard assumption, the sum of error terms in this fitted equation is zero. We test the regression coefficients for the null hypothesis that they are not significantly different from zero (Barclay et al. 1995; Gefen et al. 2000). We assume a normally distributed population for the error terms of each independent linguistic variable and accordingly t-distributed sample error terms. In our two-tailed test, we assume a standard confidence level of 95 percent, i.e. we consider our results significant at $p < .05$, highly significant at $p < .01$ and weakly significant at $p < .1$ which is a usual convention (Neter 1996). For the regressions concerning the accuracy of hard information, we need a measure to indicate how large the maximum benefit from lying could be for deceivers, measured by the variance that is *not* explained by hard information. The coefficient of determination (R^2) represents such a measure. However, for our GLS based regression model, we cannot use the standard coefficient of determination because it only applies to OLS models (Zheng and Agresti 2000). A widely applied alternative goodness-of-fit test measures the correlation between the realization of a dependent variable with its prediction (Zheng and Agresti 2000). The measure is sensitive to outliers; however, the bounded nature of our values mitigates this disadvantage. Therefore, we use this approach to measure the goodness-of-fit for our model.

We finally introduce several robustness tests which ensure that our results are not provoked by omitted-variable bias, fixed effects, model misspecification, or specificities of subsamples. Regarding the establishment of causality, the tests are supposed to *isolate* deception as indicator of high risk which causes the default (Cook and Campbell 1979).

Presentation of Results on Soft Information and Deception Detection

The second and third column in Table 2 show the coefficients and t-statistics from the GLS regression of all linguistic constructs on *survival rate* using *interest rate*, *amount*, *credit grade* and *debt-to-income ratio* as control variables. We observe that the model can explain 18.5 of the variation in *repayment success*. Relative to previous results, the model can explain 6.4 percent more risk than the model based only on hard information. This number does not change upon the exclusion of the (collinear) *credit grade*. The absolute improvement in fit seems low; however, it is substantial in relative terms, explaining 35 percent more variation than hard information only. We find that the coefficient for *self-references*, measuring the construct *immediacy*, is highly significant ($\beta = -3.11$, $t = -3.66$, $p < .001$) but, contrary to the prediction from theory, negatively related to the *survival rate*. The coefficients of *first-person plural pronouns* measuring

immediacy ($\beta=1.99$, $t=2.03$, $p<.05$), *spatial specifications gauging specificity* ($\beta=1.59$, $t=2.07$, $p<.05$), and *pausality*, our measure for *complexity* ($\beta=3.25$, $t=2.48$, $p<.05$), are all significant and are, as proposed, positively related to the *survival rate*. The coefficients of all other measures are not significant, including the *credit grade*. We assume that this is the result of the collinearity with the *interest rate* which explains the same variation in *repayment success* as the *credit grade*. Due to the low amount of additionally explained risk and the seemingly theory-contradicting findings, we consider factors that might have influenced our results. We find that the validity of categories can only be guaranteed as far as the dictionary allows when the analyzed texts exceed a *word count* of 50 (Pennebaker et al. 2006). Examining this, we compare a subsample including only observations with a word count below 50 with its complement. Interestingly, while the means are roughly identical for both samples, the categories' standard deviations at a lower *word count* are between 20 and 700 percent higher than at a higher *word count*. Hence, we find that the minimum length of the texts should not undercut a word count of 48. We decide to repeat our analyses with this subsample (cf. Table 2).

Dependent Variable:	Survival Rate				
	Sample:	Entire		Word Count $\geq 48^B$	
		Coefficient ^A (T-Statistic)		Coefficient ^A (T-Statistic)	
Interest Rate	-3.15***	(-4.25)	-3.70***	(-3.59)	
Credit Grade	-.08	(-.15)	.03	(.04)	
Debt-to-Income Ratio	-1.82***	(-7.10)	-2.12***	(-6.61)	
Requested Amount	-.98*	(-1.87)	-.04	(-.06)	
Quantity: Word Count	-.74	(-.82)	-1.00	(-.81)	
Expressivity: Modifiers	-.89	(-1.11)	-3.12**	(-2.40)	
Expressivity: Perceptual Verbs	6.89	(.68)	7.52	(.54)	
Affect: Positive Affect	-.10	(-.16)	-2.11*	(-1.88)	
Informality: Typos	-.32	(-.35)	-3.36	(-1.10)	
Uncertainty: Modal Verbs	.32	(.49)	.58	(.51)	
Immediacy: Self-References	-3.11***	(-3.66)	-1.60	(-.99)	
Immediacy: Plural Pronouns	1.99**	(2.03)	2.44*	(1.93)	
Complexity: Pausality	3.25**	(2.48)	6.27**	(2.48)	
Diversity: Unique Words	-.15	(-.16)	.54	(.37)	
Specificity: Temporal Specifications	.89	(1.36)	-.65	(-.43)	
Specificity: Spatial Specifications	1.59**	(2.07)	.93	(.64)	
Intercept (Constant)	5.45***	(5.94)	5.71***	(3.51)	
Number of Observations	973		532		
R-squared (Modified^C)	.185		.260		

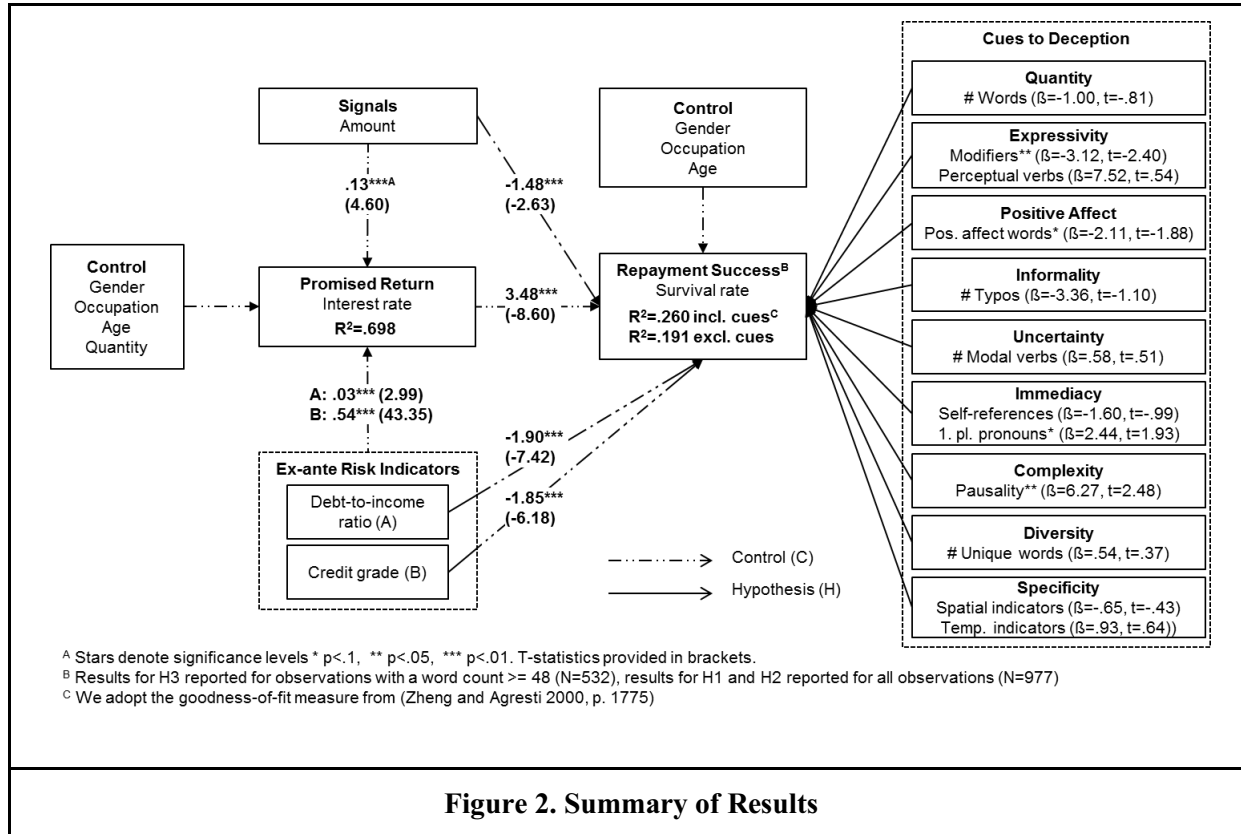
^A Stars denote significance levels * $p<.1$, ** $p<.05$, *** $p<.01$. T-statistics provided in brackets.

^B Subsample including only observations with a word count > 50 to balance LIWC's inaccuracy at a word count < 48 .

^C For the GLS model, we use a modification of the goodness-of-fit measure adopted from (Zheng and Agresti 2000, p. 1775), indicating the correlation between realization and prediction.

First, we observe that the coefficient of determination increases to 26 percent, which demonstrates twice as much explanatory power than our model that only includes hard information (however, running the subsample on hard information also raises R^2 to .19).

Second, we find that the coefficient of *pausality* ($\beta=6.27$, $t=2.48$, $p<.05$), gauging *complexity*, is now significantly positively related to *repayment success* as predicted. Equally, the coefficient of *modifiers* ($\beta=-3.12$, $t=-2.40$, $p<.05$), which measures *expressivity*, is now, as predicted by theory, in a significantly negative relation to it. The coefficients of *positive affect* ($\beta=-2.11$, $t=-1.88$, $p<.1$) and *first-person plural pronouns* ($\beta=2.44$, $t=1.93$, $p<.05$) are only slightly significant, but also follow predicted directions. To prove the stability of these results, we remove all non-significant independent variables and run the regression again. We observe that the model fits the data (almost) equally well, lowering R-squared by only .6 percent. However, the t-statistics of the coefficients do not show a substantial improvement. For the subsample with a word count above 48, the coefficient of *first-person plural pronouns* do not prove significant in a stepwise regression. Moreover, we find that the *all affect* measure captures the essence of *positive affect*, its coefficient being slightly more significant. Overall, for the subsample, our results (cf. Figure 2) provide stable support for **H1b** (*expressivity*) and **H1g** (*complexity*), weak support for **H1c** (*affect*) and **H1f** (*immediacy*). They provide no support for **H1a**, **H1d**, **H1e**, **H1h** and **H1i**.



Discussion of Results

Our results support **H1b** stating that defaulting borrowers (lemons) are more *expressive* by coloring their loan descriptions through the use of adjectives and adverbs, **H1c** stating that they are using significantly more *positive affect words* and **H1g** stating that they use a less *complex* writing. The findings are robust towards the inclusion of control variables, fixed effects and the use of different statistical models. The evidence supports the hypothesized motive of persuasion and the restriction of language complexity caused by the necessity to fabricate information, which is in line with Zhou et al.'s (2004a, 2008) and Newman (2003). However, we receive no evidence for **H1a** (*quantity*), **H1d** (*informality*), **H1e** (*uncertainty*), **H1h** (*diversity*) and **H1i** (*specificity*) and mixed evidence for the assumed *disassociation* in **H1f**. We attempt to disentangle the causes and implications of these results in the following.

First, the covariance matrix indicates a potential distortion of the nomological validity of *diversity*, *informality* and *quantity* (Homburg and Giering 1998). Their interrelation suggests that they might unwantedly gauge ability, ease of writing and communicativeness in addition to capturing the realization of imperfect language use upon deception. These opposed effects might have cancelled out the effect of deceptive intent. The fact that these constructs are among the five non-significant ones gives some credit to this apprehension. Second, the pretests also specify a strongly positive correlation between *temporal specifications* and *modifiers*. However, while the coefficient of the latter is highly significant and negatively related to *repayment success*, we do not make an equivalent observation for *temporal specifications*. This indicates that the general tendency that borrowers are also more *specific* if they are more *expressive* does not apply for borrowers who default. Hence, although none of the measures for *specificity* is significant, the fact that they do not show the same negative prediction as *expressivity* indirectly supports the assumption that lemons are less *specific*. Most recent deception studies find that liars *disassociate* themselves from their lies (Hancock et al. 2004, 2007; Qin et al. 2005; Toma and Hancock 2010; Zhou and Zhang 2008; Zhou et al. 2004a, 2004b). Therefore, we expected that lemons – given that they exhibit deceptive intent – would use fewer *self-references*, i.e. be less *immediate*. However, our findings provide support for the opposite in the entire sample, and deliver none in the reduced sample. We come up with a theory-based- and a setup-based explanation for this deviation. First, we observe a sample correlation between *self-references* and *positive affect*. When a borrower explains himself, he might appear most trustworthy if he refers to himself and his potentially miserable situation. Herzenstein et al. (2011) show that borrowers can improve credit conditions on P2P lending websites if they shape their identity as being in an “*economic hardship*”. Hence, packaging of explanations can

greatly influence perceptions of trustworthiness (Elsbach and Eloffson 2000). Therefore, if a deceptive lemon's intention to persuade and create familiarity overcompensates his intention to disassociate from the deception, his increased use of *positive affect* might go along with an increased use of *self-references*. Our finding that *positive affect* is significantly and negatively related to *repayment success* indicates that this is the case. Second, the finding may also be an artifact of an obtrusive setup of previous experiments in which constructs were developed. In most of the experiments, selected participants were given the task to consciously lie about certain facts before measuring their language use (Burgoon et al. 2003; Hancock et al. 2007; Marett and George 2004; Newman et al. 2003; Qin et al. 2005; Toma and Hancock 2010; Zhou et al. 2003, 2004a, 2004b, 2004b). Potentially, by increasing psychological lying costs, the requirement to lie is what triggers the intent to disassociate oneself from it in the first place. Hence, if the decision to deceive is made by the person the motive to disassociate might be less relevant than the incentive to be persuasive. The missing support for several constructs could also be grounded in the violation of our implicit assumptions beyond the validity of employed constructs and measures. Finally, we can expect mixed results. For instance, significant results are obstructed if the sought group of borrowers overlaps with the group who uses the constructs but does not default, and with the overconfident group who defaults but does not use the constructs. Mixed support seems to support the latter explanation. However, what we demonstrate is that lemons in online P2P lending write loan descriptions which are significantly different in style than those of good risks. We can assume that they consciously or subconsciously know about their high risks and express this knowledge in a particular language style. As a preliminary conclusion, the applicability of the usual deception detection categories depends on the base rate of words and the number of observations that underlie such an analysis. We can give support to our proposition that borrowers have a significantly higher propensity of defaulting when they create more *expressive*, *affective* and less *complex* loan descriptions.

Conclusion and Implications

We evaluate on 'Can the occurrence of deceptive cues in soft information of IT-mediated P2P lending project descriptions explain repayment success?' basing on agency and Interpersonal Deception Theory. We apply content analysis and multiple regressions to 'soft' textual descriptions and hard information from 1099 loan projects of the P2P lending platform *LendCo* for investigation. To our best knowledge, this is the first study to blend research on automated deception detection in a computer-mediated context with the evaluation of economic transaction outcomes. We add to the discussion on intermediaries (Bailey and Bakos 1997; Datta and Chatterjee 2008) by observing that provided hard facts can only explain 12 percent of the variation in true risk and that lenders make inefficient decisions. In addition, with an

explained variance of 26 percent, we can support that lemons attempt to deceive by finding four significant cues to deception impacting repayment outcomes. These findings have theoretical and practical consequences for IS research since signaling and screening processes evidently are not yet far enough evolved to minimize transaction costs. Our evidence demonstrates that the inefficiency of a self-regulated P2P lending market still requires intermediaries to reduce these transaction costs (Bhattacharya and Thakor 1993; Stigler 1961; Williamson 1981). Hence, our findings reject the electronic marketplace hypothesis. Just as Bailey and Bakos predict (Bailey and Bakos 1997), the roles of these intermediaries must be different. They do not need to execute the screening process themselves, like banks, but they need to prepare and verify information since borrowers seem to be unable to correctly decide how information should be evaluated and weighed.

For IS practitioners, we reason that the role of an intermediary could be executed through the P2P lending platform itself, independent third parties or through borrower groups (Ashta and Assadi 2010; Berger and Gleisner 2009; Chircu and Kauffman 2000). The creation of borrower groups is probably the cheapest, but also least acknowledged way of implementation. Moreover, P2P lending platforms need to provide more verifiable information to increase the value and credibility of borrower signals. Additionally, the information must be presented in a more intuitive way to be understood even by inexperienced lenders. Apparently, a written explanation of a risk measure's value is not sufficient. A more graphical visualization might be helpful. Alternatively, the *interest rate* could be set professionally by the platform, as successfully executed by the currently most successful American P2P lending platform *LendingClub* (Lending Club 2011). Our study has also practical implications for the automation of deception detection in settings that involve computer-mediated communication. Since we find the measures of *expressivity*, *affect* and *complexity* to be significant predictors for risk, we believe that deception detection research has potential and should further be pursued.

References

- Akerlof, G. A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics* (84:3), August, pp. 488–500.
- Anderson, D. E., DePaulo, B. M., Ansfield, M. E., Tickle, J. J., and Green, E. 1999. "Beliefs about cues to deception: Mindless stereotypes or untapped wisdom?," *Journal of Nonverbal Behavior* (23:1), pp. 67–89.
- Anderson, J. C., and Gerbing, D. W. 1988. "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin* (103:3), May, pp. 411–423.
- Anderson, J. C., Gerbing, D. W., and Hunter, J. E. 1987. "On the assessment of unidimensional measurement: Internal and external consistency, and overall consistency criteria," *Journal of Marketing Research* (24:4), November, pp. 432–437.
- Ashta, A., and Assadi, D. 2010. "The Use of Web 2.0 Technologies in Online Lending and Impact on Different Components of Interest Rates," in *Advanced Technologies for Microfinance: Solutions and Challenges*, France: Groupe ESC Dijon Bourgogne, pp. 206–224.

- Bailey, J. P., and Bakos, Y. 1997. "An exploratory study of the emerging role of electronic intermediaries," *International Journal of Electronic Commerce* (1:3), April, pp. 7–20.
- Barclay, D., Higgins, C., and Thompson, R. 1995. "The partial least squares (PLS) approach to causal modeling: Personal computer adoption and use as an illustration," *Technology Studies* (2:2), pp. 285–309.
- Barley, S. R., Meyer, G. W., and Gash, D. C. 1988. "Cultures of culture: Academics, practitioners and the pragmatics of normative control," *Administrative Science Quarterly*, pp. 24–60.
- Baum, C. F. 2006. *An introduction to modern econometrics using Stata*, 1. ed., College Station, USA: Stata Press.
- Benyus, J., Bremmer, I., Pujadas, J., Christakis, N., Collier, P., Warnholz, J., et al. 2009. "Breakthrough ideas for 2009," *Harvard Business Review* (87:2), pp. 19–40.
- Berelson, B. 1952. *Content analysis in communication research*, 1. ed., Michigan, USA: University of Michigan.
- Berger, S. C., and Gleisner, F. 2009. "Emergence of financial intermediaries in electronic markets: The case of online P2P lending," *Business Research* (2:1), pp. 39–65.
- Bhattacharya, S., and Thakor, A. V. 1993. "Contemporary banking theory," *Journal of Financial Intermediation* (3:1), pp. 2–50.
- Buller, D. B., and Burgoon, J. K. 1996. "Interpersonal deception theory," *Communication Theory* (6:3), August, pp. 203–242.
- Burgoon, J., Blair, J., Qin, T., and Nunamaker, J. 2003. "Detecting deception through linguistic analysis," Berlin, Heidelberg: Springer Verlag, pp. 91–101.
- Burgoon, J. K., Buller, D. B., Guerrero, L. K., Afifi, W. A., and Feldman, C. M. 1996. "Interpersonal deception: Information management dimensions underlying deceptive and truthful messages," *Communications Monographs* (63:1), pp. 50–69.
- Carlson, J. R., George, J. F., Burgoon, J. K., Adkins, M., and White, C. H. 2004. "Deception in computer-mediated communication," *Group Decision and Negotiation* (13:1), pp. 5–28.
- Caspi, A., and Gorsky, P. 2006. "Online deception: Prevalence, motivation, and emotion," *CyberPsychology & Behavior* (9:1), pp. 54–59.
- Chircu, A. M., and Kauffman, R. J. 2000. "Reintermediation strategies in business-to-business electronic commerce," *International Journal of Electronic Commerce* (4:4), pp. 7–42.
- Cook, T. D., and Campbell, D. T. 1979. *Quasi-experimentation: Design & analysis issues for field settings*, Boston, USA: Rand McNally College.
- Daft, R. L., and Lengel, R. H. 1986. "Organizational information requirements, media richness and structural design," *Management Science* (32:5), pp. 554–571.
- Datta, P., and Chatterjee, S. 2008. "The economics and psychology of consumer trust in intermediaries in electronic markets: The EM trust framework," *European Journal of Information Systems* (17:1), February, pp. 12–28.
- DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. 2003. "Cues to deception," *Psychological Bulletin* (129:1), p. 74.
- Dibbern, J., Goles, T., Hirschheim, R., and Jayatilaka, B. 2004. "Information systems outsourcing: a survey and analysis of the literature," *ACM SIGMIS Database* (35), November, pp. 6–102.
- Elsbach, K. D., and Eloffson, G. 2000. "How the packaging of decision explanations affects perceptions of trustworthiness," *Academy of Management Journal* (43:1), pp. 80–89.
- Ford, C. V., King, B. H., and Hollender, M. H. 1988. "Lies and liars: Psychiatric aspects of prevarication," *American Journal of Psychiatry* (145:5), p. 554.
- Gartner IT Research 2010. *Press statements on development of P2P lending*. Retrieved 28. November, 2011, from <http://www.gartner.com/technology/research.jsp>.
- Gefen, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in online shopping: An integrated model," *MIS Quarterly* (27:1), pp. 51–90.
- Gefen, D., and Rigdon, E. E. 2011. "An update and extension to SEM guidelines for administrative and social science research," *MIS Quarterly* (35:2), p. iii–A7.
- Gefen, D., Straub, D. W., and Boudreau, M. C. 2000. "Structural equation modeling and regression: Guidelines for research practice," *Communications of the Association for Information Systems* (4:7), pp. 1–70.
- Greiner, M. E., and Wang, H. 2010. "Building consumer-to-consumer trust in e-finance marketplaces: An empirical analysis," *International Journal of Electronic Commerce* (15:2), pp. 105–136.
- Hair Jr, J. F., Anderson, R. E., Tatham, R. L., and William, C. 1998. *Multivariate data analysis*, Upper Saddle River, NJ, USA: Prentice Hall.
- Hancock, J. T. 2007. "Digital deception," *Oxford handbook of internet psychology*, pp. 289–301.
- Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. 2007. "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication," *Discourse Processes* (45:1), pp. 1–23.

- Hancock, J. T., Thom-Santelli, J., and Ritchie, T. 2004. "Deception and design: The impact of communication technology on lying behavior," pp. 129–134.
- Herzenstein, M., Andrews, R. L., Dholakia, U., and Lyandres, E. 2008. "The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities," *Boston University School of Management Research Paper*.
- Herzenstein, M., Sonenshein, S., and 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:Special Issue), pp. 138–149.
- Hirshleifer, D. 2001. "Investor psychology and asset pricing," *The Journal of Finance* (56:4), August, pp. 1533–1597.
- Homburg, C., and Giering, A. 1998. "Konzeptualisierung und Operationalisierung komplexer Konstrukte: Ein Leitfaden für die Marketingforschung," in *Die Kausalanalyse*, Stuttgart, Germany: Schäffer-Poeschel, pp. 111–146.
- Horne, D. R., Norberg, P. A., and Ekin, A. C. 2007. "Exploring consumer lying in information-based exchanges," *Journal of Consumer Marketing* (24:2), pp. 90–99.
- Hurkens, S., and Kartik, N. 2009. "Would I lie to you? On social preferences and lying aversion," *Experimental Economics* (12:2), June, pp. 180–192.
- Iyer, R., Khwaja, A. I., Luttmer, E. F. P., and Shue, K. 2009. *Screening in new credit markets: can individual lenders infer borrower creditworthiness in peer-to-peer lending?*
- Joinson, A. N., and Dietz-Uhler, B. 2002. "Explanations for the perpetration of and reactions to deception in a virtual community," *Social Science Computer Review* (20:3), p. 275.
- Jones, T. M. 1991. "Ethical decision making by individuals in organizations: An issue-contingent model," *Academy of Management Review* (15:2), pp. 366–395.
- Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., and Gorski, S. 2011. "Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success," *Journal of Applied Communication Research* (39:1), pp. 19–37.
- Lending Club 2011. *Statistics Lending Club*. Retrieved 28. November, 2011, from <http://www.lendingclub.com/>.
- Logsdon, J. M., and Patterson, K. D. W. 2010. "Deception in business networks: Is it easier to lie online?," *Journal of Business Ethics* (90:Suppl 4), December, pp. 537–549.
- Marett, L. K., and George, J. F. 2004. "Deception in the case of one sender and multiple receivers," *Group Decision and Negotiation* (13:1), pp. 29–44.
- Maxwell, S. E. 2000. "Sample size and multiple regression analysis," *Psychological Methods* (5:4), December, pp. 434–458.
- McMahon, J. M., and Harvey, R. J. 2006. "An analysis of the factor structure of Jones' moral intensity construct," *Journal of Business Ethics* (64:4), pp. 381–404.
- Milgrom, P., and Roberts, J. 1992. *Economics, organization and management*.
- Morris, R. 1994. "Computerized content analysis in management research: A demonstration of advantages & limitations," *Journal of Management* (20:4), pp. 903–931.
- Moulton, L. 2007. "Divining value with relational proxies: How moneylenders balance risk and trust in the quest for good borrowers," *Sociological Forum* (22:3), pp. 300–330.
- Neter, J. 1996. *Applied linear statistical models*, 4. ed., OH, USA: McGraw-Hill.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. 2003. "Lying words: Predicting deception from linguistic styles," *Personality and Social Psychology Bulletin* (29:5), p. 665.
- O'Brien, R. M. 2007. "A caution regarding rules of thumb for variance inflation factors," *Quality & Quantity* (41:5), March, pp. 673–690.
- Papke, L. E., and Wooldridge, J. M. 1996. "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates," *Journal of Applied Econometrics* (11:6), pp. 619–632.
- Pavlou, P. A., and Dimoka, A. 2006. "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research* (17:4), pp. 392–414.
- Pavlou, P. A., Liang, H., and Xue, Y. 2007. "Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective," *MIS Quarterly* (31:1), pp. 105–136.
- Pennebaker, J., Booth, R., and Francis, M. 2006. "Linguistic Inquiry and Word Count (LIWC), 2006 versions," *Austin, TX: LIWC*.
- Petter, S., Straub, D., and Rai, A. 2007. "Specifying formative constructs in information systems research," *Management Information Systems Quarterly* (31:4), p. 623.

- Qin, T., Burgoon, J. K., Blair, J., and Nunamaker, J. 2005. "Modality effects in deception detection and applications in automatic-deception-detection," p. 23b–23b.
- Shi, N.-Z., and Tao, J. 2008. *Statistical hypothesis testing: theory and methods*, World Scientific.
- Sonenshein, S., Herzenstein, M., and Dholakia, U. M. 2011. "How accounts shape lending decisions through fostering perceived trustworthiness," *Organizational Behavior and Human Decision Processes* (115:1), May, pp. 69–84.
- Stigler, G. J. 1961. "The economics of information," *The Journal of Political Economy* (69:3), pp. 213–225.
- Toma, C. L., and Hancock, J. T. 2010. "Reading between the lines: Linguistic cues to deception in online dating profiles," pp. 5–8.
- Utz, S. 2005. "Types of deception and underlying motivation: What people think," *Social Science Computer Review* (23:1), pp. 49–56.
- Wang, H., Greiner, M., and Anderson, J. 2009. "People-to-people lending: The emerging e-commerce transformation of a financial market," pp. 182–195.
- Weber, R. P. 1990. *Basic content analysis*, Thousand Oaks, CA, USA: Sage.
- Webster, J., and Watson, R. T. 2002. "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quarterly* (26:2), p. xiii – xxiii.
- Williamson, O. E. 1981. "The economics of organization: The transaction cost approach," *American Journal of Sociology* (87:3), pp. 548–577.
- Zheng, B., and Agresti, A. 2000. "Summarizing the predictive power of a generalized linear model," *Statistics in Medicine* (19:13), July, pp. 1771–1781.
- Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D. 2004a. "Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication," *Group Decision and Negotiation* (13:1), pp. 81–106.
- Zhou, L., Burgoon, J. K., and Twitchell, D. P. 2003. "A longitudinal analysis of language behavior of deception in e-mail," *Intelligence and Security Informatics* (2665), pp. 102–110.
- Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., and Nunamaker Jr, J. F. 2004b. "A comparison of classification methods for predicting deception in computer-mediated communication," *Journal of Management Information Systems* (20:4), pp. 139–166.
- Zhou, L., Yongmei Shi, and Dongsong Zhang 2008. "A statistical language modeling approach to online deception detection," *IEEE Transactions on Knowledge & Data Engineering* (20:8), pp. 1077–1081.
- Zhou, L., and Zhang, D. 2008. "Following linguistic footprints: Automatic deception detection in online communication.," pp. 119–122.