Collective Exposure: Peer Effects in Voluntary Disclosure of Personal Data

Rainer Böhme[1] and Stefanie Pötzsch[2]

[1] Department of Information Systems, University of Münster, Germany
rainer.boehme@wi.uni-muenster.de
[2] Department of Computer Science, Technische Universität Dresden, Germany
stefanie.poetzsch@tu-dresden.de

Abstract. This paper reports empirical evidence for peer effects in privacy behavior using field data from online social lending. Our content analysis and regression models show that individuals copy observable behavior of others in decisions on a) how much to write about oneself, b) whether to share custom pictures, c) what personal data to disclose, and d) how identifiable to present oneself. We frame this finding in the theory of descriptive social norms and analyze moderating effects, such as similarity of context, social proximity, and mimicry of success factors. The presence of peer effects in disclosure behavior can explain the formation and change of apparent social norms and attitudes towards privacy.

1 Financial Privacy and Human Behavior

Information technology has created an age of perfect memory, which raises issues of information privacy and informational self-determination. With legal and technical means available that in principle empower individuals to control the distribution of their personal data, researchers of all disciplines still lack a good understanding of *how* individuals make use of this control.

Scholars of the *economics of privacy* (see [1] for a survey) assume rational individuals who consider all costs and benefits when making a decision to disclose personal data. Works in this tradition largely draw on analytical economic models to study the efficiency of privacy regimes in specific market situations [2–4]. Yet it remains doubtful if individual decisions to disclose personal data can be explained sufficiently well with models of rational agents.

Instead, it has been suggested to approach the subject with theories borrowed from *social psychology* [5] and *behavioral economics* [6]. In this spirit, a number of behavioral biases affecting the decision to disclose personal data has recently been identified empirically: a general discrepancy between stated preferences and behavior [7], present-based biases and discounting [8], anchor effects [9], social norms [10], perceived control [11], and contextual primes [12]. All these results have been obtained from laboratory experiments. While experiments are the method of choice for exploring causality under controlled conditions, they oftentimes suffer from small samples and questionable ecological validity.

The contribution of this paper is to complement the laboratory studies with new evidence from field data. More specifically, we explain voluntary disclosure of personal data with *peer effects*, that is, the tendency of individuals to mimic other peoples' disclosure behavior. Our data are loan applications by real users of *Smava.de*, the largest German platform for *online social lending*.

Also known as "P2P lending", "Ebay for loans", or "crowd-sourcing of finance", online social lending has grown rapidly over the past couple of years [13]. Drawing on concepts of (offline) micro-finance, the idea of social lending is to provide a marketplace for unsecured loans between individuals: an online platform lets borrowers advertise loan applications to lenders, who decide in which loan they invest. Each lender funds only a small share of the financed amount so that the credit risk is shared in loan-specific pools of lenders. Lenders receive interest as a compensation for taking risk, whereas the platform operators typically charge fixed (i. e., risk-free) fees. Market mechanisms differ between platforms, a fact that led to research in mechanism design [14].

Online social lending is an ideal data source for the study of behavioral aspects of financial privacy. By their very nature, loan applications contain a lot of personal details, which enable lenders to assess the associated risk [15]. Data requirements and sharing arrangements already raise privacy concerns in the traditional banking industry [16]. These issues are further exacerbated in online social lending where personal data of loan applications does not remain within heavily regulated institutions, but is accessible to all Internet users [17]. Another feature of this data source is that loan applicants disclose their personal data to this audience *voluntarily*. (In fact, financial regulators require the platform operator to collect additional personal data, which is not disclosed to the public though.) This enables us to look for pattern that explain the influence of peers, i. e., other applicants, on the decision to disclose personal data.

There is a clear link between peer orientation and the notion of *herd behavior*. The latter is an active field of research in finance, often entangled with the question of rationality [18]. At some level of sophistication, models can be found that explain herding and resulting bubbles as rational action. For a first cut on the topic of peer effects in personal data disclosure, we spare us the discussion of whether peer effects are rational or not (the former requires a model of competition between borrowers). We rather see our contribution in the description and rigorous measurement of the phenomenon based on longitudinal field data.

This paper is organized straight. The following Section 2 develops seven hypotheses and introduces the data and analysis method. Results are presented in Section 3, and then discussed in Section 4.

2 Approach

The design of online social lending websites, including *Smava.de*, was inspired by other online marketplaces. A list of current loan applications is displayed right on the homepage, and details of each applications are accessible to everybody by following a single link. Likewise, an archive of completed loan applications —

successful and unsuccessful loans alike | is only a few clicks away. Common sense suggests that new applicants who complete their own loan application seek inspiration for filling the various text fields with information about themselves and their project. Most likely, they will take recent loan application on the platform as examples. This way, we expect to see serial similarities in the longitudinal analysis of all loan applications, which can be interpreted as peer effects.

Aside from common sense, peer effects in disclosure decisions can also be derived from established behavioral theory, notably the influence of descriptive social norms [19]. Following descriptive as opposed to injunctive norms is facilitated if the disclosure decision is made rather heuristically than systematically [20]. The finding that data disclosure is extremely sensitive to contextual cues [12, 21] supports the assumption of heuristic decisions and thus the dominance of descriptive social norms. Moreover, determinants of *self-disclosure* have been studied in social psychology for decades [22], however with focus on relationship building and therapy rather than on the desire or need to protect one's privacy. Studies of self-disclosure typically include disclosure of personal thoughts and feelings, unlike our study, which defines personal data primarily as facts about individuals. These theoretical considerations lead us to the following hypotheses.

2.1 Hypotheses

We postulate four hypothesis on the existence of positive peer effects in voluntary disclosure of personal data:

Hypothesis 1 *The total lengths of all descriptions associated with a loan application is positively correlated with the lengths of descriptions of recent loan applications.*

Hypothesis 2 *A loan application is more likely illustrated with a custom project picture if recent loan applications include a custom project picture.*

Hypothesis 3 *The probability of disclosure of personal data of a specific type increases with the disclosure of personal data items of the same type in recent loan applications.*

Hypothesis 4 *Borrowers present themselves more identifiable in loan applications if the borrowers in recent loan applications appear more identifiable.*

The hypotheses so far predict that peer effects exist and are observable by using different indicators of disclosure as dependent variable. Length of description (H1) and provision of a custom picture (H2) were chosen as objective indicators. The divulgence of specific personal data items (H3) and the overall identifiability (H4) operationalize our research question better. Testing the latter requires subjective decisions by expert coders in a content analysis (see Sect. 2.2).

In addition, we postulate three hypotheses on factors that moderate the strengths of peer effects.

Hypothesis 5 *The peer effects predicted in Hypotheses 1–4 are reinforced for recent loan applications which are similar to the newly drafted loan application.*

We test this hypothesis by measuring the additional explanatory power of loan applications in the same category.

Hypothesis 6 *The peer effects predicted in Hypotheses 1–4 are reinforced for recent loan applications which share borrower characteristics with the borrower of the newly drafted loan application.*

We test this hypothesis by measuring the additional explanatory power of loan applications of which a) the borrowers' credit grades match and b) borrowers are of the same sex.

Hypothesis 7 *The peer effects predicted in Hypotheses 1–2 are reinforced if recent loan applications were successful.*

We test this hypothesis by measuring the additional explanatory power of directly preceding loans which had been completely funded at the time when the new application was drafted.

Hypotheses 5–7 are motivated by different definitions of the peer group. They were formulated to shed more light on the individuals' motivation to select specific loan applications as examples: similarity of context (H5), perceived social-psychologic proximity (H6), and mimicry of apparent success factors (H7).

2.2 Data

This study is a secondary data analysis in a broader research effort on privacy in online social lending [17, 23]. Our data consists of 4701 loan applications posted on the largest German social lending site *Smava.de* between March 2007 and July 2010, representing a total asked amount of 41.9 million euro (about US$ 55 million). German borrowers are said to be particularly privacy-aware, reflecting a long tradition of comprehensive data protection regulation as well as high public interest and participation in debates on privacy issues.

Smava.de lets potential borrowers propose the basic credit conditions (amount, interest rate, and maturity of 36 or 60 months), checks their identity and publishes on its website verified demographic information (age, gender, state) along with a credit grade, a rough debt service-to-income ratio, an assignment to one of 19 categories[3], as well as a user-provided project description and optional user-provided pictures. Lenders can review this information and contribute to its funding in step sizes of 250 euros. When the loan is fully funded or after two

3 *Smava.de* defines the following categories: debt restructuring; liquidity; home, gardening & do-it-yourself; cars & motorbikes; events; education & training; family & education; antiques & art; collection & rarity; electronics; health & lifestyle; sports & leisure; travel; pets & animals; volunteering; commercial; business investment; business extension; miscellaneous

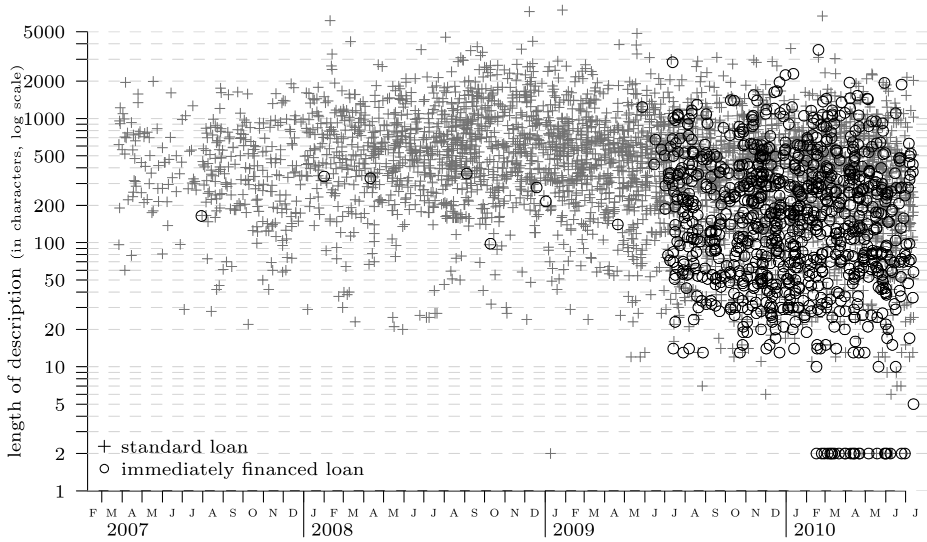

Fig. 1. View on the data: variation in the length of the user-provided description for 4701 loan applications published on the largest German social lending platform *Smava.de* between March 2007 and July 2010. Loans which got immediate funding were excluded for the analysis. This reduces the sample size to 3786 loans. Immediate loans picked up from July 2009 after a change in *Smava.de*'s bidding mechanism

weeks, whatever is earlier, the (partial) loan is granted via a commercial bank, who partners with *Smava.de* to comply with the local financial supervision regulations. Borrowers and lenders can appear on the platform under self-chosen nick names, however their full identity is known to and verified by *Smava.de*.

We enrich this data by conducting a content analysis [24] to measure the amount of personal data in loan applications. Variation in personal data disclosure can be found in textual project descriptions, voluntary categories of the borrower profile page, and possibly associated pictures. Three trained coders independently rated the textual descriptions and rated the disclosure of personal data without knowing our hypotheses. The underlying code book distinguishes between ten types of personal data, namely borrower's name, financial situation, education, profession, special skills and qualifications, housing situation, health situation, hobbies and memberships, contact details (address, phone, e-mail, etc.), and information about close relatives (family or partner). Each type has several sub-types that encode in which detail borrowers disclose personal data of the respective type.

Orthogonal to the disclosure of facts, privacy can also be measured by identifiability. If individuals are identifiable, it is easier to complete a profile by linking facts from other sources. To measure identifiability, we asked the coders to rate the likelihood that a borrower can be identified on 7-point scales. Individual ratings were collected for several levels of prior knowledge, i. e., presumed identifiability by relatives, neighbors, colleagues or arbitrary persons with access to a search engine. For the purpose of this study, we add these four ratings to an *identifiability index*. This also helps to reduce measurement error. Expectedly, this index correlates with the raw count of disclosed data types, but it is still sufficiently distinct to be interpreted as additional source of information.

2.3 Method

Figure 1 shows the length of the user-provided description in all 4701 loan applications between March 2007 and July 2010 over time. The variance in the length of description (measured on a log scale) remains substantial throughout time. So, in principle, there is room to explain part of this variance with peer effects. Note that *Smava.de* changed the market mechanism in July 2009 by introducing so-called "immediate loans". Instead of waiting for a posted loan application to be funded, the platform suggests an interest rate high enough so that the loan can immediately be financed by lenders who pre-commited offers in a kind of order book. Obviously, voluntary disclosure of personal data does not affect the credit decision for immediate loans. As immediate loans are not distinguishable from other loans in *Smava.de*'s public archive, we use a proxy and exclude from the analysis all loans where the maximum latency between two subsequent bids is less than two minutes. The high density of loans matching this criterion just after the introduction of immediate loans (see Fig. 1) confirms the validity of this procedure.

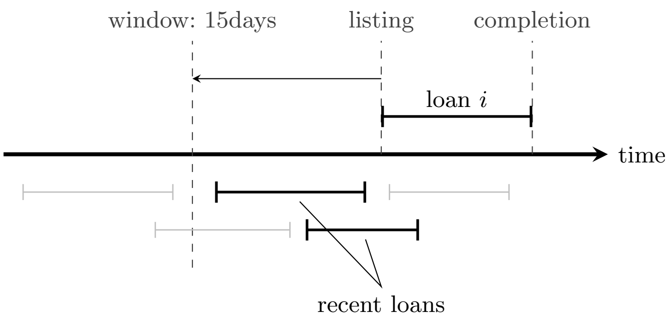

Fig. 2. Definition of *recent loan applications* for loan i: all other loans *listed* between the beginning of the window and the listing of loan i

Next we have to define the peer group, more specifically, what qualifies *recent loan applications*, as referred to in the hypotheses. For each loan, we determine the specific set of recent loans as depicted in Fig. 2. We take a window of 15 days before the initial listing of the loan and consider all loan applications listed in this window as recent. To test Hypotheses 5{7, the set of recent loans is further reduced to the subset of loans which share a particular property. An exception arises for the property of successful funding (H7). Successful loans can only be more in uential if the fact that the loan is successful had been observable at the time when the new application was drafted. Hence we de ne the subset of successful recent loans as all loans listed within the windows *and* fully funded before the listing of the new loan. For example, even if fully funded, the second (lower) recent loan in Fig. 2 would not be considered as recent successful loan.

The window size was chosen with regard to the expiry of loan applications after two weeks, which sets an upper bound. A lower bound is given by the number of recent loans in the subsets when testing Hypotheses 5{7. Several unreported robustness checks also revealed that the choice of the window size does not a ect the direction of the e ects for window sizes between 5 and 60 days. For 15 days, the number of recent loans varies between 0 and 104, following a unimodal distribution with mean 63.2 and median 67. Loans listed later in the sample tend to have a higher number of recent loans due to the raising popularity of the platform. Note that alternative de nitions of the peer group are conceivable (e. g., a xed number of most recent loans), but were not explored so far.

We use regression models to conduct hypothesis tests by statistical inference while controlling for several intervening factors. In general, the models for Hypotheses 1 and 4 are speci ed as follows:

$$y_i = \beta_0 + \beta_1 \bigodot_{j \in P_i} y_j + \beta_2 \bigodot_{j \in P_i \cap \{k \mid x_k = x_j\}} y_j + \cdots + \beta_3 \log a_i + \beta_4 x_i + \cdots + \beta_{(\cdot)} f(t_i) + \varepsilon_i,$$

where

- y_i is the dependent variable of loan i;
- P_i is the set of recent loans of loan i;
- $\bigodot$ is an aggregation operator that calculates the arithmetic mean of the argument over a speci ed set of loans;
- x_i is an auxiliary property of loan i, which can appear to build subsets of P_i to test Hypotheses 5{7. In this case, we also have to include the property as a control variable or | for multinomial properties | as xed e ect to avoid spurious results from an unbalanced sample;
- a_i is the amount of loan i, for which we have to control as people might disclose more personal data if they ask for more money;
- $f(t_i)$ is a function generating time dummies to control for low frequency uctuations over time (annual dummies unless otherwise stated);
- $\boldsymbol{\beta} = (\beta_0, \dots)$ is a coe cient vector that can be estimated with ordinary least squares to minimize the squared residuals ε_i, i. e., $\sum_i \varepsilon_i^2 \to \min$.

We estimate several variants of this general model by including terms for different dependent variables y and auxiliary properties x. We report the estimated coe cients along with empirical standard errors and a signi cant level for the two-sided test of the null hypothesis $\beta_k = 0$. For the interpretation, estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are of most interest. If $\hat{\beta}_1$ is positive and significant, the dependent variable can be explained by the aggregate realizations of the dependent variable in the respective sets of recent loans. This indicates the existence of peer effects. If both $\hat{\beta}_1$ and $\hat{\beta}_2$ are positive and significant, there is evidence for additional explanatory power of loans in the subset sharing the auxiliary property, hence peer effects are stronger if the property matches.

Hypotheses 2 and 3 concern binary indicators as dependent variables. In these cases, we use logistic regression analyses to regress the logit-transformed odds ratio of the dependent variable on the same set of predictor terms as above. These coefficients are estimated with the maximum likelihood method.

3 Results

3.1 Length of Description

Table 1 shows the estimated coefficients for the regression models specified to test Hypothesis 1 in conjunction with Hypotheses 5–7. Each specification (A–G) is reported in a column. The dependent variable is given by the log of the number of characters in all text field descriptions provided by the loan applicant. For brevity, we do not report all coefficients for multinomial controls and fixed effects, but indicate whether the respective terms are included by the label "yes".

Model A estimates a plain peer effect. The term for all recent loans (i. e., $\hat{\beta}_1$) is positive and highly significant. With a coefficient value of almost 0.9, a change by one order of magnitude in the lengths of description of recent loans translates to about 2.5 times longer descriptions for the average new loan application. Most likely, this specification overestimates the size of the peer effect because other relevant predictors of the verbosity are not controlled for. Model B resolves this by introducing a control for the loan amount (log transformed) and time dummies to capture long-term trends. The relevant coefficient remains positive and highly significant. Therefore our data supports Hypothesis 1. The coefficient for the loan amount is positive and significant, too. People tend to write more if they ask for more money.

Models C–F test Hypotheses 5–7. We do find support for Hypothesis 5. Recent loans of the same category have additional explanatory power in predicting the length of description even if base effects of individual categories are controlled for (model C). The picture is more mixed for Hypothesis 6, which can only be supported if the same credit grade is taken as indicator of social proximity (model D). Loan applicants apparently do not prefer peers of the same sex when mimicking their disclosure behavior (model E). This is most likely not an artifact of an unbalanced sample, as about one quarter of all loan are requested by women. Nor do we find support for Hypothesis 7: borrowers do not seem to copy from successful recent applications more often than from pending or unsuccessful applications (model F). Model G serves as robustness check to see if any coefficient changes its sign when all predictors are included at the same time.

Table 1. Factors influencing the verbosity of descriptions

	Dependent variable: length of description (log)						
Terms	A	B	C	D	E	F	G
all recent loans	0.88 *** (0.047)	0.63 *** (0.078)	0.53 *** (0.088)	0.53 *** (0.087)	0.60 *** (0.126)	0.58 ** (0.220)	0.35 (0.245)
recent loans with ...							
same category			0.08 * (0.038)				0.07 † (0.038)
same credit grade				0.12 ** (0.041)			0.15 *** (0.042)
borrower same sex					0.03 (0.100)		0.00 (0.110)
complete funding						0.06 (0.203)	0.06 (0.224)
controls:							
log amount		0.21 *** (0.021)	0.18 *** (0.023)	0.21 *** (0.022)	0.21 *** (0.021)	0.21 *** (0.022)	0.19 *** (0.024)
fixed effects:							
category			yes				yes
credit grade				yes			yes
gender					yes		yes
complete funding						yes	yes
annual		yes	yes	yes	yes	yes	yes
(constant)	0.67 * (0.277)	0.49 (0.497)	0.75 (0.521)	0.24 (0.504)	0.42 (0.499)	0.37 (0.518)	0.47 (0.550)
(number of cases)	3784	3784	3553	3750	3777	3780	3531
(adjusted R^2 [%])	8.5	11.1	12.9	11.5	11.1	11.0	13.4

Std. errors in brackets; stat. significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

This does not happen, which indicates robustness. However, some estimates lose statistical signi cance supposedly due to collinearity between the higher number of predictors.

Overall, the models explain between 8 and 14 % of the variance in the dependent variable. Note that the bulk of explained variance is due to the direct peer e ects (model A) and not, unlike in many other eld studies, due to the inclusion of xed e ects. The number of cases varies somewhat between models because missing values appear if subsets of recent loans turn out to be empty.

3.2 Provision of a Picture

Table 2 shows the estimated coe cients for the logistic regression models speci- ed to test Hypothesis 2 in conjunction with Hypotheses 5{7. We take the odds that a loan applicant has uploaded a custom picture as dependent variable. Overall, 22.6 % of all loan applications contain a custom project picture.

Table 2. Factors influencing the decision to publish a project picture

	Dependent variable: prob. of custom picture (logit link)						
Terms	H	I	J	K	L	M	N
all recent loans	2.27^{***} (0.421)	1.90^{***} (0.530)	1.48^{**} (0.519)	2.20^{***} (0.505)	1.80^{**} (0.669)	2.36^{*} (1.099)	$2.36^{\dagger}$ (1.357)
recent loans with ...							
same category		$0.34^{\dagger}$ (0.207)	0.82^{***} (0.195)				0.34 (0.209)
same credit grade				−0.07 (0.237)			0.01 (0.253)
borrower same sex					0.30 (0.486)		0.18 (0.554)
complete funding						−0.31 (0.998)	−0.70 (1.202)
controls:							
log amount	0.14^{**} (0.050)	0.17^{**} (0.055)	0.14^{**} (0.052)	0.15^{**} (0.050)	0.14^{**} (0.050)	0.15^{**} (0.052)	0.19^{**} (0.058)
fixed effects:							
category		yes					yes
credit grade				yes			yes
gender					yes		yes
complete funding						yes	yes
annual	yes	yes	yes	yes	yes	yes	yes
(constant)	-2.94^{***} (0.447)	-3.43^{***} (0.508)	-2.98^{***} (0.473)	-3.25^{***} (0.473)	-2.95^{***} (0.450)	-3.01^{***} (0.497)	-3.91^{***} (0.592)
(number of cases)	3784	3553	3553	3750	3777	3780	3531

Std. errors in brackets; stat. significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Model H tests Hypothesis 2 and nds highly signi cant peer e ects. The interpretation of coe cients is less straightforward for logistic models. A value of $\hat{\beta}_1$ = 2.3 denotes that the odds of providing a custom picture are exp($\hat{\beta}_1$) $\approx$ 10 times higher if *all* recent loans contained a custom picture than if *none* of the recent loans contained one. Model I nds weakly signi cant support for the strict test of Hypothesis 5 including category xed e ects. The e ect remains robust and (unsurprisingly) stronger if the category xed e ects are omitted (model J). This demonstrates the importance of including the respective xed e ects, as di erent o sets between categories (e. g., positive sign for cars & motorcycles; negative sign for debt restructuring) otherwise feed into the \same category" terms and overstate the true e ect size. Models K{M test Hypotheses 6 and 7 and do not nd support for any of them.

The provision of a custom picture is probably the crudest indicators for several reasons. Its binary scale is susceptible to noise. The provisioning of a picture largely depends on the external factor whether a suitable picture is available. And the indicator does not di erentiated between pictures of di erent informa-

Table 3. Strength of peer effects on disclosure of personal data (by type)

	Overall	Frequency in recent loans			
Type of personal data	frequency	all	category	credit grade	amount
Profession	62.3 %	2.76*** (0.641)	0.18 (0.277)	0.73* (0.307)	0.28*** (0.080)
Financial situation	33.9 %	3.19*** (0.577)	0.31 (0.273)	0.75* (0.311)	0.20* (0.083)
Family and partner	31.4 %	−0.13 (0.968)	−0.29 (0.316)	0.79* (0.309)	0.11 (0.084)
Hobbies and memberships	29.3 %	2.62*** (0.713)	0.13 (0.271)	0.22 (0.297)	0.11 (0.082)
Housing situation	19.1 %	1.00 (1.568)	0.26 (0.504)	−0.65 (0.518)	0.34** (0.107)
Education	10.8 %	1.68 (1.987)	0.21 (0.474)	0.34 (0.662)	−0.18 (0.122)
Name or first name	5.7 %	8.76** (2.721)	−0.57 (0.956)	0.55 (1.020)	0.25 (0.161)
Health situation	3.5 %	18.14*** (4.590)	−1.66 (1.099)	−1.61 (2.416)	−0.11 (0.230)
Contact details	2.6 %	8.29 (6.773)	1.35 (1.159)	−1.22 (2.282)	0.20 (0.234)
Special skills & qualifications	1.9 %	17.69** (6.664)	−1.70 (2.475)	0.81 (2.480)	0.29 (0.300)

Std. errors in brackets; stat. significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Summary of logit models for the probability of disclosing at least one item of the respective type. $N = 1558$ fully funded loans between Nov 2008 and Jan 2010.

tion content and quality. Nevertheless we decided to report the results because they combine an objective indicator with the same logistic regression model that is used in the following to estimate peer e ects for personal data disclosure. It therefore serves as reference to better interpret the results in Table 3.

3.3 Personal Data Disclosure by Type

Length of description and provision of a picture are imprecise indicators because they do not measure semantic. A high value could either re ect verbosity or extensive disclosure of personal data. Our hand-coded indicators of data disclosure by type do not share this limitation. Due to resource constraints, data of this granularity is only available for 1558 fully funded loans between November 2008 and January 2010. Since partly funded and unsuccessful loans were not included in the coding task, we are unable to test Hypothesis 7 for subjective indicators. This is not a big shortcoming as this hypothesis has been refuted on the basis of objective indicators anyway.

Table 3 shows selected estimated coe cients for the logistic regression models speci ed to test Hypothesis 3 in conjunction with Hypotheses 5 and 6. Unlike in the previous tables, predictors appear in columns and di erent dependent variables in rows. For each row, the dependent variable is de ned by the odds that a loan application contains at least one data item of a specific type (including sub-types). The types are ordered by decreasing marginal probability, as reported in the second column.

Positive and highly significant peer effects can be found for 6 out of 10 types of personal data. Quite surprisingly, while Hypothesis 5 could be retained for objective indicators, it seems that recent loans with the same category have no additional impact on the decision to disclose data of a particular type. Therefore we refute Hypothesis 5 for this indicator. By contrast, recent loan applications with the same credit grade have significant influence in predicting the disclosure of personal data of the three types with the highest marginal probability of disclosure. With 3 out of 10 significant at the 5 % level — 1 out of 20 is the expected value under the null hypothesis — we have to acknowledge weak support for Hypothesis 6. Note that we have also estimated models with matching gender, but none of the relevant coefficients turned out signifiant.

A side-observation in Table 3 is that the disclosure of certain types of personal data, notably data about profession, financial and housing situation, correlated positively with the loan amount, whereas others types of data appear independent of the size of the loan.

3.4 Identifiability

For a sample of 1663 cases, we have sufficient observations of the identifability index (see Sect. 2.2). Table 4 shows the estimated coefficients for the regression models specified to test Hypothesis 4 in conjunction with Hypotheses 5 and 6. Again, we find highly significant evidence for peer effects (model O). This supports Hypothesis 4. As in Sect. 3.3 above, Hypothesis 5 is not supported (model P), whereas Hypothesis 6 can be retained at least for the credit grade as matching property (model Q). Compared to the length of description as dependent variable (see Tab. 1), the ratio of explained variance is lower, but this is not uncommon for noisy measurements from field data that exhibit a lot of unexplained heterogeneity over an extended period of data collection.

4 Discussion

4.1 Summary and Interpretation

To the best of our knowledge, this paper is the first to report evidence for peer effects in voluntary disclosure of personal data using field data. The existence of this effect has been conjectured before [6, 25], however so far without evidence. Table 5 summarizes the results for all hypotheses.

More specifically, we could find plain peer effects for all four indicators of disclosure — objective proxies, such as the length of description or the provision of custom pictures, and subjective ratings alike. The result pattern is less clear for our additional hypotheses on factors that potentially reinforce peer effects and it is probably too early to generalize from the specific operationalizations made in the context of our data source.

Table 4. Factors influencing the identifiability of a borrower

	Dependent variable: identifiability index			
Terms	O	P	Q	R
all recent loans	0.61 *** (0.127)	0.56 *** (0.143)	0.48 *** (0.141)	0.37 (0.252)
recent loans with ...				
same category		0.00 (0.060)		0.01 (0.060)
same credit grade			0.13 * (0.060)	0.13 * (0.061)
borrower same sex				0.07 (0.200)
controls:				
log amount	0.85 *** (0.205)	0.51 * (0.228)	0.93 *** (0.207)	0.59 * (0.232)
fixed effects:				
category		yes		yes
credit grade			yes	yes
gender				yes
annual	yes	yes	yes	yes
(constant)	−3.64 (2.220)	−0.77 (2.404)	−4.67 * (2.297)	−1.49 (2.501)
(number of cases)	1659	1565	1651	1554
(adjusted R^2 [%])	4.2	3.9	4.3	4.2

Peer effects imply that decisions to disclose personal data are driven by observed behavior of others. Understanding peer effects is relevant in general, because over time, peer effects can create self-reinforcing dynamics that may change the social attitude towards privacy: \if everybody discloses his or her full life on the Internet, it can't be wrong to disclose my information as well." Privacy-friendly interface design and increased user awareness might attenuate this dynamic, but we remain skeptical if those cues can ever be strong enough to reverse dynamics of descriptive social norms, not to mention the missing incentives of market participants to implement effective cues.

A specific relevance of peer effects in social lending emerges for platform designers in the financial industry. If data disclosure is merely a reaction to previous loans, then the fact whether data has been disclosed loses its value as a signal to lenders. A platform providing a better separation of roles, with distinct privileges to access loan applications (e. g., exclusively to registered lenders with positive account balance), could not only increase borrower privacy, but also make data disclosure decisions more individual and thus signal more valuable information to lenders. This might resolve the apparent puzzle that more disclosure does not always translate into better credit conditions, even after controlling for all available hard information [17, 23].

Table 5. Summary of hypothesis tests

	plain	in conjunction with ...		
		Hypothesis 5	Hypothesis 6	Hypothesis 7
Hypothesis 1	retained	retained	partly retained	refuted
Hypothesis 2	retained	weak support	refuted	refuted
Hypothesis 3	partly retained	refuted	weak support	not tested
Hypothesis 4	retained	refuted	partly retained	not tested

4.2 Relation to Related Work

Aside from the theoretical and empirical works referenced in the introduction to narrow down the research question, the following empirical studies are related most closely to this paper. Gross and Acquisti [25] use field data to measure personal data disclosure and identifiability in online social networks. Their approach is quantitative, but largely explorative with emphasis on descriptive statistics. The authors speculate about peer pressure and herding in their conclusions. Acquisti, John, and Loewenstein [10] report positive influence of descriptive social norms on disclosure of *sensitive* personal data. In their laboratory experiments, participants reacted differently if they were told that previous participants revealed sensitive data. In contrast to the case of online social lending, the decision makers had to answer closed-form questions and could not see the disclosed data of others. They had to believe in the abstract information instead. Barak and Gluck-Ofri [26] conducted a self-disclosure experiment in the framework of an online forum. They report peer effects from previous posts. In contrast to our study, the dependent variable was less tailored to privacy. Consistent with the tradition in this literature, the authors also measure disclosure of feelings and thoughts.

Empirical research on online social lending also exists in the economic literature (see [13] for a review). Typical research questions concern market efficiency, signs of discrimination, or the influence of ties in the social networks between borrowers and lenders, respectively. Occasionally, lengths of description is used as an effort measure in these studies [27]. We are not aware of other attempts than our previous works [17, 23] to quantify personal data disclosure and test theories of privacy behavior empirically with data from online social lending.

4.3 Limitations and Future Work

Compared to laboratory experiments, field studies suffer from limited control. In particular, disclosure of false information remains unnoticed in this study (a problem shared with some experiments, e. g., [10]). Our main results are robust to changes in the window size and the use of quarterly instead of annual time dummies, but refinements are needed to check the robustness against different definitions of the peer group. New research efforts are directed to replicate the study in other contexts. A more distant goal is to follow the approach in [28] and unify the collection of individual behavioral effects of data disclosure into a more general theory of planned behavior, which spans effects of attitude, social norms, and self-efficacy.

Acknowledgements

These results are based on a larger research project exploring the economics of online lending jointly with Nicola Jentzsch. This paper has also benefited from discussions with Jens Schade. We thank our coders Ulrike Jani, Kristina Schäfer, and Michael Sauer for their excellent research assistance.

References

1. Hui, K.L., Png, I.P.: The economics of privacy. In Hendershott, T.J., ed.: Handbooks in Information System and Economics. Volume 1. Elsevier (2006) 471–493
2. Posner, R.A.: The economics of privacy. American Economic Review **71**(2) (1981) 405–409
3. Calzolari, G., Pavan, A.: On the optimality of privacy in sequential contracting. Journal of Economic Theory **130**(1) (2006) 168–204
4. Jentzsch, N.: A welfare analysis of secondary use of personal data. In: Workshop of Economics and Information Security (WEIS), Harvard University (2010)
5. Margulis, S.T.: Privacy as a social issue and behavioral concept. Journal of Social Issues **59**(2) (2003) 243–261
6. Acquisti, A., Grossklags, J.: Privacy and rationality in individual decision making. IEEE Security and Privacy **3**(1) (2005) 26–33
7. Berendt, B., Günther, O., Spiekermann, S.: Privacy in e-commerce: Stated preferences vs. actual behavior. Communications of the ACM **48**(4) (2005) 101–106
8. Acquisti, A.: Privacy in electronic commerce and the economics of immediate gratification. In: ACM Electronic Commerce, New York, NY (2004) 21–29
9. Grossklags, J., Acquisti, A.: When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In: Workshop of Economics and Information Security (WEIS), Carnegie Mellon University (2007)
10. Acquisti, A., John, L., Loewenstein, G.: The impact of relative standards on concern about privacy. In: Workshop of Economics and Information Security (WEIS), University College London (2009)
11. Brandimarte, L., Acquisti, A., Loewenstein, G.: Misplaced confidences: Privacy and the control paradox. In: Workshop of Economics and Information Security (WEIS), Harvard University (2010)
12. John, L., Acquisti, A., Loewenstein, G.: Strangers on a plane: Context-dependent willingness to divulge personal information. Consumer Research **37**(2) (2011)
13. Berger, S., Gleisner, F.: Emergence of financial intermediaries in electronic markets: The case of online P2P lending. BuR - Business Research **2**(1) (May 2009) 39–65
14. Chen, N., Ghosh, A., Lambert, N.: Social lending. In: Proc. of ACM EC, Palo Alto, CA, ACM Press (2009) 335–344
15. Stiglitz, J.E.: Credit rationing in markets with imperfect information. American Economic Review **71**(3) (1981) 393–410
16. Jentzsch, N.: Financial Privacy. Springer-Verlag, Berlin Heidelberg (2007)

17. Böhme, R., Pötzsch, S.: Privacy in online social lending. In: AAAI Spring Symposium on Intelligent Privacy Management, Palo Alto, CA (2010) 23–28
18. Bikhchandani, S., Sharma, S.: Herd behavior in financial markets. IMF Staff Papers **47**(3) (2001) 279–310
19. Cialdini, R.B., Kallgren, C.A., Reno, R.R.: A focus theory of normative conduct. Advances in Experimental Social Psychology **24** (1991) 201–234
20. Chaiken, S., Yaacov, T.: Dual-process Theories in Social Psychology. Guilford Press, New York (1999)
21. Böhme, R., Köpsell, S.: Trained to accept? A field experiment on consent dialogs. In: Human Factors in Computing Systems (ACM CHI). (2010) 2403–2406
22. Derlega, V.J., Berg, J.H., eds.: Self-Disclosure. Theory, Research, and Therapy, New York, Plenum Press (1987)
23. Pötzsch, S., Böhme, R.: The role of soft information in trust building: Evidence from online social lending. In Acquisti, A., Smith, S., Sadeghi, A.R., eds.: TRUST 2010. LNCS 6101, Berlin Heidelberg, Springer-Verlag (2010) 381–395
24. Holsti, O.R.: Content analysis for the social sciences and humanities. Addison-Wesley (1969)
25. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: Privacy in the Electronic Society (ACM WPES). (2005) 71–80
26. Barak, A., Gluck-Ofri, O.: Degree and reciprocity of self-disclosure in online forums. CyberPsychology & Behavior **10**(3) (2007) 407–417
27. Herzenstein, M., Andrews, R., Dholakia, U., Lyandres, E.: The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities (June 2008)
28. Yao, M.Z., Linz, D.G.: Predicting self-protections of online privacy. CyberPsychology & Behavior **11**(5) (2008) 615–617