University of Augsburg
Prof. Dr. Hans Ulrich Buhl
Research Center
Finance & Information Management
Department of Information Systems
Engineering & Financial Management

Universität Augsburg
University

Discussion Paper

Data Quality Goes Social: What Drives Data Currency in Online Social Networks?

by

Quirin Görz, Florian Probst

in: Proceedings of the 21st European Conference on Information Systems, ECIS, Utrecht, Netherlands, June 2013

Universität Augsburg, D-86135 Augsburg
Visitors: Universitätsstr. 12, 86159 Augsburg
Phone: +49 821 598-4801 (Fax: -4899)
www.fim-online.eu

DATA QUALITY GOES SOCIAL: WHAT DRIVES DATA CURRENCY IN ONLINE SOCIAL NETWORKS?

Görz, Quirin, FIM Research Center, University of Augsburg, Universitätsstraße 12, 86159 Augsburg, Germany, quirin.goerz@wiwi.uni-augsburg.de

Probst, Florian, FIM Research Center, University of Augsburg, Universitätsstraße 12, 86159 Augsburg, Germany, florian.probst@wiwi.uni-augsburg.de

Abstract

Despite the increasing acceptance and usage of Online Social Networks (OSN), particularly the currency of user data has been identified as a crucial success factor for both, users as well as providers. However, while research on Data Quality (DQ) in traditional information systems is considered to be relatively mature, findings from prior research on currency have not yet been confirmed for the context of OSN. Moreover, neither theories taking into account the specific social characteristics of OSN nor real-world evaluations have been applied to learn more about potential factors influencing the currency of user data in OSN. By using publicly available data of the business network XING, we therefore empirically investigate factors influencing the currency of user data. Thus, the paper makes three contributions: First, we confirm factors influencing currency from prior research in the field of DQ for the context of OSN. Second, we find that social characteristics such as social pressure induced by the number of direct contacts and users' activity have a significant impact on the currency of user data. Third, we show that taking into account these additional factors when estimating currency leads to significantly better results than relying solely on previously known factors.

Keywords: Online Social Network, Business Network, Data Quality, Currency.

1 Introduction

While research on Data Quality (DQ) in traditional information systems is considered to be relatively mature, DQ in Online Social Networks (OSN)[1] has hardly been researched so far (Chai et al., 2009). This is surprising, as the increasing social and economic impact of OSN is largely based on the enormous amount of shared personal information and user-generated content (Heidemann et al., 2012). The quality of this user data, however, varies considerably (Chai et al., 2009; Gross and Acquisti, 2005) and especially the currency of user data, that is, the question whether it is still up-to-date, is a critical success factor for OSN (Lin, 2008; Lin and Stasinskaya, 2002; Leimeister et al., 2006). From a professional user perspective, especially marketers depend on up-to-date user data when applying targeted advertising campaigns in OSN (Evans, 2011) and recruiters rely on the currency of user data within business networks such as XING or LinkedIn (Lin and Stasinskaya, 2002). To spend their resources efficiently, professional users of OSN could thus benefit from an increasing currency of user data itself or the possibility to identify outdated user data. At the same time, OSN providers can grow their advertising or membership revenues from professional users only, if they are able to guarantee the currency of the provided user data. Taken together, increasing the currency of user data (e.g., by incentivizing users deliberately to keep their data up-to-date), is essential for the success of OSN.

However, for huge datasets such as OSN with millions of users, it is neither feasible nor economically reasonable to quantify the currency of user data manually and by real-world tests (Heinrich and Klier, 2009). Thus, currency needs to be quantified by an automated process (Chai et al., 2009), for instance, using one of the multiple metrics that have been proposed in DQ literature (cf. e.g., Ballou et al., 1998; Heinrich and Klier, 2011). These metrics have often been based on probabilistic theory assuming that the probability that a data attribute is up-to-date declines over time. Furthermore, it has been shown that additional information indicating how long an attribute value is usually valid, allows for estimating currency more precisely (Heinrich and Klier, 2009).

However, these findings from prior research on currency in traditional information systems have not been confirmed for the context of OSN yet. Moreover, neither theories considering the specific social characteristics of OSN nor real-world evaluations have been applied to learn more about potential further factors influencing the currency of user data in OSN (Chai et al., 2009). Therefore, we first draw on theories and concepts of both, DQ and OSN literature to identify factors that might influence the currency of user data in OSN. Second, we empirically investigate their impact on the currency of user data and derive theoretical as well as practical implications. Here, we use the case of the business network XING for three reasons: First, business networks gain increasing revenues from specialized membership offerings for recruiters (XING, 2011) and the success of recruiters and their willingness to pay heavily depend on the currency of user data (Lin and Stasinskaya, 2002). Second, business networks allow for collecting and statistically analyzing rich data about their users' behavior. Third, the currency of user data provided in business networks can be verified without drawing on surveys, which enables an empirical analysis based on objective data. Even though our results may not be generalized for all types of OSN and are explorative in nature, we believe that the lack of prior research at the interface of DQ and OSN, the given relevance, and the facilitated rigor justify our focus on business networks. We hope that our results help to contribute to a better understanding of DQ in general and currency in particular in contexts such as OSN, which can be extended and generalized in future research.

The remainder of the paper is structured as follows: In the next section, we define the problem context, discuss relevant literature, and outline our research gap. To address this gap, we propose our research model based on literature on DQ and OSN in the subsequent section. Afterwards, we test this model empirically through objective data collected from the business network XING and highlight our findings. Finally, we summarize the results, discuss limitations, and point out areas for future research.

[1] While some authors use the term Social Networking Site (SNS), we use the term OSN throughout the paper synonymously.

2 Problem Context and Related Work

Following Boyd and Ellison (2007, p. 211), we define OSN as "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system". Business networks such as XING or LinkedIn are a particular category of OSN (Heidemann et al., 2012). In contrast to common private networks such as Facebook, business networks "specialise in maintaining professional contacts and searching for new jobs" (Bonneau and Preibusch, 2010, p. 125). Therefore, business networks provide a lot of relevant professional information about potential employees that can be utilized by recruiters (Bonneau and Preibusch, 2010). As in the case of general OSN such as Facebook, this information includes identifying information (e.g., name, photo), interests (e.g., subscribed interest groups), and personal contacts (e.g., list of connected users) (Heidemann et al., 2012). Additionally, business networks contain a curriculum vitae (e.g., current position, employer) and additional information, such as their registration date or an index indicating a user's activity within the business network (Strufe, 2010).

Recent studies provide evidence for the popular usage of business networks for recruiting purposes and that recruiters prefer business networks such as XING or LinkedIn over general OSN such as Facebook (Weitzel et al., 2011). Business models in business networks are mainly two-tiered instead of marketing based, meaning that basic services are offered for free and premium services are provided for a fee (Riggins, 2003). While in 2010, XING still yielded the majority of its earnings with such so-called "premium members" (16% of its users), revenues from specialized membership offerings for recruiters rose by more than 65% in 2011, almost doubling the network's advertising revenues (XING, 2011). However, particularly DQ is a major concern (Gross and Acquisti, 2005), as the quality of the provided user data is a critical factor for online recruiting success (Lin and Stasinskaya, 2002). Therefore, both recruiters and the providers of business networks are highly interested in a high currency of user data.

In general, high-quality data is needed to perform all kind of decisions and business processes within and across companies properly (Even and Shankaranarayanan, 2007; Falge et al., 2012). Since insufficient DQ may lead to wrong decisions and consequently high costs (Heinrich and Klier, 2009), both the benefit and the acceptance of information systems depend significantly on the quality of data processed and provided by these systems (Ballou et al., 1999). In traditional information systems, numerous DQ dimensions such as accuracy, completeness, and currency[2] have been found to be essential (cf. e.g., Falge et al., 2012; Lee et al., 2002) and their measurement has been intensively researched (cf. e.g., Ballou et al., 1998; Even and Shankaranarayanan, 2007; Heinrich and Klier, 2011; Pipino et al., 2002). In the context of OSN, prior work indicates that user data is quite accurate (Donath and Boyd, 2004; Gross and Acquisti, 2005) and accuracy in business networks is even higher than for instance in average cover letters (Davison et al., 2011). The completeness of user data (e.g., information provided within the users profiles such as gender, age etc.) has been investigated under the label of (self-)disclosure (Nosko et al., 2010; Schrammel et al., 2009). Thereby, several studies found that users of OSN intensively share private information (Gross and Acquisti, 2005) and that (self-)disclosure in business networks is even higher than in general OSN (Schrammel et al., 2009; Strufe, 2010). Moreover, the completeness of user data in business networks can be assessed easily by observing if an attribute value is publicly available or not. However, currency of (user) data and its quantification has been found to be crucial not only in DQ literature (Falge et al., 2012; Klein et al., 2007) but also in the context of OSN in general (Lin, 2008; Leimeister et al., 2006) and business networks in particular (Lin and Stasinskaya, 2002).

According to DQ literature, currency can be defined as a probability that an attribute value is still up-to-date (for an overview cf. e.g., Heinrich and Klier, 2009; Heinrich and Klier, 2011). In other words, we consider data provided by users in business networks to be up-to-date, if it corresponds to its real-world representation at the point in time when DQ is quantified and has not become outdated. However, for

[2] Currency is also referred to as timeliness, freshness, or up-to-dateness. We use the term currency throughout the paper synonymously.

huge datasets, such as business networks with millions of users, it is not feasible to quantify currency by comparing attribute values to their real-world counterparts, as such real-world tests are "by far too time- and cost-intensive and not practical at all" (Heinrich and Klier, 2009, p. 2652). Prior research found that it could be useful to consider so-called *attribute metadata* when estimating currency of data instead (cf. e.g., Ballou et al., 1998; Heinrich and Klier, 2011). Such attribute metadata can be for example information that indicates when an attribute value's corresponding real-world counterpart has been created (Heinrich and Klier, 2011). In the context of business networks, this could be for example the date of a research assistant's enrolment, at which the attribute value "Research Assistant" of the data attribute "Current Position" in his or her user profile became valid. Other relevant attribute metadata is an attribute value's shelf-life, representing how long an attribute value is usually valid. However, this information is often unknown. Heinrich and Klier (2009, p. 2651) indicate that an attribute value's shelf-life can be instead estimated by using so-called *supplemental data*, that is, "additional data attributes that allow drawing conclusions about the [currency] of the data attribute considered". For instance, taking into account that the attribute value of the data attribute "Current Position" is "Research Assistant" (with a relatively shorter shelf-life) and not "Professor" (with a relatively longer shelf-life), allows for predicting currency more precisely. These findings have been derived in the context of traditional information systems with a quite strict focus on attribute values. A user perspective including assumptions about the probability that a user keeps his or her user data up-to-date and the role of social influence among users as in the context of OSN has been neglected so far. Therefore, our research gap is twofold: First, findings from research on currency in traditional information systems have not been confirmed for the context of business networks yet. Second, neither theories taking into account the specific social characteristics of business networks nor real-world evaluations have been applied, in order to learn more about factors beyond established attribute metadata and supplemental data, which might potentially influence the currency of user data in contexts such as OSN.

3 Research Model

As already briefly discussed in section 2, attribute metadata such as the age of an attribute value plays an important role when quantifying the probability that an attribute value still corresponds to its real-world representation (Heinrich and Klier, 2011). Heinrich and Klier (2009) consider the age of an attribute value at the point in time when estimating its currency. The age of an attribute value is thereby defined as the difference between the instant of quantifying DQ and the instant of the attribute value's real-world counterpart's creation. We follow their perception and consequently expect that an attribute value in a user's profile is more likely to become outdated, the higher its age is:

H1: *The higher an attribute value's age in a user's profile is, the lower is the likelihood that this attribute value is up-to-date.*

Since attribute metadata beyond the age of an attribute value is often unknown, Heinrich and Klier (2009) suggest using also supplemental data when estimating currency (cf. section 2). Especially supplemental data that determines a data value's shelf-life allows for predicting currency more precisely (Heinrich and Klier, 2009). Therefore, we hypothesize:

H2: *In a user profile, supplemental data that indicates an attribute value's longer (shorter) shelf-life has a significant positive (negative) influence on the likelihood that this attribute value is up-to-date.*

To identify further potential factors that might influence the currency of user data in business networks, particularly literature on the usage and adoption of OSN seems to be promising (cf. e.g., Boyd and Ellison, 2007; Hu et al., 2011; Sledgianowski and Kulviwat, 2009). Prior research in this context indicates that especially constructing and maintaining a personal profile to present oneself to other users and managing contacts are major motives to use OSN in general (Heidemann et al., 2012; Richter et al., 2011) and business networks in particular (Schaefer, 2008). Thus, users aim at increasing their social capital, which is defined as the "sum of the actual and potential resources embedded within, available through, and derived from the network of relationships" (Nahapiet and Ghoshal, 1998, p. 243). This is particularly relevant in business networks, as social capital has been found to be very beneficial with respect to potential employment opportunities (Ellison et al., 2007).

When explaining and predicting users' goal-directed behavior in OSN, for instance, towards building and maintaining social capital by participating in OSN, it has been "conceptualized as an intentional social action where users regard themselves as part of the social fabrics" (Cheung and Lee, 2010, p. 25). Building on Social Influence Theory, prior research found that particularly subjective norms, that is, perceived social pressure induced by a user's "perception that most people who are important to him think he should or should not perform the behavior in question" (Fishbein and Ajzen, 1975, p. 302), plays a decisive role (Cheung and Lee, 2010; Hu et al., 2011; Sledgianowski and Kulviwat, 2009). For instance, Page and Kobsa (2010, p. 174) state that users feel pressured to not only adopt OSN but also to "strictly adhere to the established social etiquette". This is in line with Utz and Kramer (2009), who emphasize "that perceived norms play an important role when it comes to the appropriate use of [OSN]". In the context of business networks, perceived social pressure can be induced by the common agreement that information provided in a curriculum vitae need to be up-to-date (cf. e.g., Provoost, 2009). Therefore, we hypothesize that users being aware of a large number of other users visiting their profiles are more likely to keep their user data up-to-date due to a higher level of perceived social pressure:

H3: *The higher the number of page impressions of a user profile is, the higher is the likelihood that an attribute value within this profile is up-to-date.*

Besides the number of partly randomly visiting users, which also includes total strangers, especially directly connected users including acquaintances, friends, or colleagues, are more likely to induce social pressure (Haythornthwaite, 2002; Lewis et al., 2008). In the context of business networks, this is also due to the fact, that directly connected users have a higher probability of knowing if the presented user data is actually up-to-date (Provoost, 2009). This is in line with Livingstone (2008), who highlights that profile information in OSN is not only updated because it no longer represented users' identity but also due to the influence of direct contacts. Furthermore, profiles of users with a high number of direct contacts have been found to receive even more attention by other users (Strufe, 2010). Taken together, we consequently hypothesize that the number of direct contacts is related to the perceived level of social pressure and thus the probability that users keep their user data up-to-date:

H4: *The higher a user's number of direct contacts is, the higher is the likelihood that an attribute value within his or her profile is up-to-date.*

In addition to users' connectivity in business networks, users' activity can be expected to influence the currency of user data. According to Lewis et al. (2008, p. 82), "more active users may have more elaborate profiles". Thus, it can be assumed that particularly active users that engage vividly in business networks, for instance, in order to increase their social capital, care about their profiles and keep them up-to-date. This is in line with Utz and Kramer (2009), who indicate that especially active users engage in OSN to present themselves to others. Furthermore, perceived social pressure may be amplified by being very active in a business network, as users become more aware of the prevalence of subjective norms (Lewis et al., 2008). For instance, by extensively browsing other users' profiles and reading their curriculum vitae, active users may become more sensitive and more likely to keep their own profiles up-to-date. Taken together with the significant impact of user activity on the attention that the corresponding profiles attract (Strufe, 2010), we therefore hypothesize:

H5: *The higher a user's activity is, the higher is the likelihood that an attribute value within his or her profile is up-to-date.*

Along with social influence characterized by subjective norm, a second mode of social influence derived from Social Influence Theory, that is, identification, has been found to be significantly related to users' intentional social actions in OSN (Cheung and Lee, 2010). By joining interest groups in business networks, users achieve a social identity consisting of self-categorization, group-based self-esteem, and affective commitment (Bagozzi and Dholakia, 2002). Thereby, particularly affective commitment has been found to be an important factor of influence, which "implies a sense of emotional involvement with the group", which "fosters loyalty and citizenship behaviors in group settings" (Dholakia et al., 2004, p. 245). Thus, memberships in interest groups should lead to a higher probability that users keep their user data up-to-date. Taken together with the significant impact of the number of subscribed interest groups on the attention a user's profile receives (Strufe, 2010), we hypothesize:

H6: *The higher a user's number of subscribed interest groups is, the higher is the likelihood that an attribute value within his or her profile is up-to-date.*

Besides these hypothesized relationships, we believe that the following factors are important enough to be controlled for: First, the propensity to pay for premium memberships might influence users' activity and engagement in OSN (Oestreicher-Singer and Zalmanson, 2009). Second, multiple studies found that users' activity and engagement in OSN declines over time (Wilson et al., 2009). Third, women and men might react differently when being exposed to social pressure in OSN (cf. e.g., Lewis et al., 2008). Our complete research model is presented in Figure 1 and tested empirically in the next section.

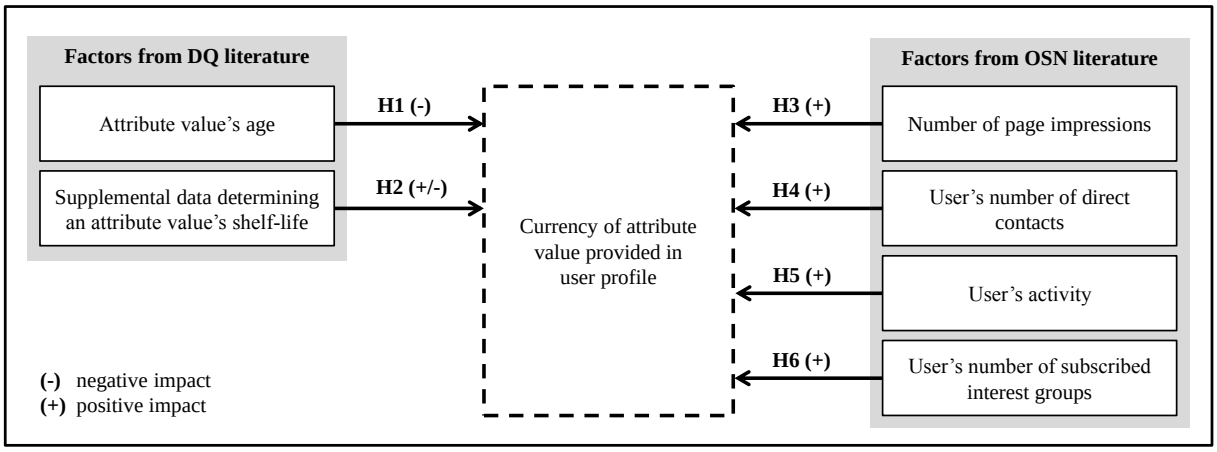

Figure 1. Research model

4 Method and Analysis

4.1 Data Collection and Variables

For the empirical analysis of our hypotheses, we collected data from November 2011 to June 2012 by randomly selecting 1,383 user profiles of university employees in Germany within the business network XING, containing profiles of 1,285 research assistants, 43 assistant professors, and 55 full professors. One major challenge when investigating the currency of data is verifying whether an attribute value is up-to-date or not. Since extensive internal company data about their employees can hardly be collected, university employees have been chosen for this study. In contrast to most companies, information about current university employees as well as alumni and emeriti are usually publicly available on university websites and can thus be used to validate the information provided within XING. For our analysis, we used the data attribute "Current Position", which is one of the most relevant data attributes in business networks. By performing a real-world test for each user profile, we tested if the current position is actually up-to-date. Therefore, we first compared manually the current positions stated in the user profiles with the respective universities' websites. If the current position could not be validated this way, we secondly also used search engines such as Google and Yahoo. If the current position could also not be validated by an extensive and standardized search procedure, we classified the current position to be out-of-date. To further improve the reliability of the classification, we visited all user profiles containing an out-of-date data attribute "Current Position" again in September 2012. Thus, we were able to check if a user updated his or her profile in the meantime and if his or her position at the university actually ended or changed before or after the date, at which his or her profile information had been collected first. Finally, our dataset consisted of user profiles containing 1,211 up-to-date and 172 (12.44%) out-of-date data attributes "Current Position". In order to check, whether the ratio of 12.44% is reasonable, we collected a second dataset in cooperation with an international management consultancy to account for a potential selection bias. Thereby, the currency of the data attribute "Current Position" of 130 user profiles could be analyzed by a real-world test using the consultancy's internal employee database. As we found that 16% of the data attributes "Current Position" within the analyzed user profiles had been out-of-date, we considered our dataset to be reasonable.

Besides currency, the following variables could be constructed for 1,150 user profiles of our dataset by collecting publicly available information (233 user profiles could not be further considered, as the accessibility has been limited, for instance, by privacy settings): *duration of the current employment* (time span between the date when the current enrolment began and data collection in years), *professional status* (research assistant, assistant professor, or full professor), *number of page impressions* (measured by unique page hits), *number of direct contacts*, *activity meter* (publicly available or not), *number of subscribed interest groups*, *premium membership* (yes or no), *membership duration* (time span between the date of joining XING and data collection in years), and *gender* (male or female). To operationalize supplemental data that determines a data value's shelf-life (cf. section 2), we used the manifestations of the variable *professional status* (i.e., research assistant, assistant professor, or full professor). We believe this is reasonable, as in Germany full professors have a longer duration of employment (about 23 years on average) than assistant professors (about 8 years on average) and research assistants (about 4.2 years on average) (Janson et al., 2006). Thus, we are able to test, if supplemental data indicating a longer (shorter) shelf-life of the data attribute "Current Position" has significant positive (negative) influence on the data attribute's currency. To approximate users' activity, the actual values of the *activity meter* implemented in XING needed to be approximating by binary values indicating whether the activity meter is publicly available or not. This is due to two reasons: First, providing public information about the activity meter is optional within XING and many users opt-out (59% of the users in our dataset). Thus, relying on the actual values would lead to a stark reduction of the dataset. However, we believe that the fact whether a user allows the activity meter to be public itself already reveals a lot of relevant information. For instance, if the activity level is low, many guides for a successful usage of business networks suggest opting-out in order to avoid disadvantageous signaling. Therefore, it can be assumed that users who opt-out have on average a lower activity level than users who publicly present the activity meter on their profile. Second, the activity meter within XING itself provides only a rough indication of the frequency of a user's utilization of the business network. It is neither very fine grained, nor does it support any conclusion on the included factors (e.g., times of a user's logins) (Strufe, 2010). Hence, using the actual values of the activity meter would also be only an approximation. We therefore believe that using binary values is reasonable and a good approximation for the level of users' activity.

4.2 Analysis

To test our hypothesis (cf. section 3), we apply the following binary logistic regression (logit) model:

$$\log \frac{\Pr(\text{up-to-date})}{1-\Pr(\text{out-of-date})} = \beta_0 + \beta_1\, \textit{duration of the current employment} + \beta_2\, \textit{professional status} + \beta_3\, \textit{professional status(1)} + \beta_4\, \textit{professional status(2)} + \beta_5\, \textit{number of page impressions} + \beta_6\, \textit{number of direct contacts} + \beta_7\, \textit{activity meter} + \beta_8\, \textit{number of subscribed interest groups} + \beta_9\, \textit{premium membership} + \beta_{10}\, \textit{membership duration} + \beta_{11}\, \textit{gender}$$

Before estimating the model, we had to adjust our dataset owing to the fact that only about 9.7% of the attribute values "Current Position" within the user profiles containing all required information are out-of-date (112 out of 1,150). If we used this dataset for estimating our model, the occurrence of out-of-date attribute values would be rare and the model would predict users with an up-to-date attribute value for the data attribute "Current Position" with 99.6% accuracy but users with out-of-date attribute values with only 4.5% accuracy. A common technique to overcome this problem of misclassification is choice based sampling (Scott and Wild, 1986; King and Zeng, 2001), which is often applied in the context of so-called "rare events" (Ben-Akiva and Lermann, 1985). Thus, we estimate our model while under-sampling the up-to-date user profiles. In our case, we accordingly use the 112 profiles containing out-of-date attribute values and 150 randomly selected profiles containing up-to-date attribute values for our analysis, which is a commonly chosen ratio (cf. e.g., Oestreicher-Singer and Zalmanson, 2009). To obtain a split-sample validation, we further divide the choice based sample ($N = 262$) randomly into two subsamples. One subsample, the analysis sample ($N = 167$), which approximates 62% of the choice based sample, is used to estimate the logit model (cf. Table 1). The second subsample, the holdout sample ($N = 95$), is afterwards employed to test the model's classification accuracy (Hair et al., 2006).

Independent Variable	Coefficient	Standard Error	Wald	df	p-value	Exp(B)
Duration of the current employment	-.503	.115	19.147	1	.000***	.604
Professional status	-	-	10.540	2	.005***	-
Professional status(1)	-3.626	1.523	7.672	1	.017**	.027
Professional status(2)	-6.548	2.020	10.502	1	.001***	.001
Number of page impressions	.000	.000	3.157	1	.076*	1.000
Number of direct contacts	.014	.005	7.470	1	.006***	1.014
Activity meter	1.855	.518	12.813	1	.000***	6.390
Number of subscribed interest groups	-.022	.062	.128	1	.720	.978
Premium membership	-.083	1.070	.006	1	.938	.920
Membership duration	-.205	.130	2.506	1	.113	.815
Gender (male)	.194	.540	.129	1	.719	1.214
Constant	5.460	1.791	9.296	1	.002***	235.083
Revised constant	3.904 after estimation intercept adjustment					
* Significant at the 0.10 level; ** Significant at the 0.05 level; *** Significant at the 0.01 level *N* (up-to-date) = 95; *N* (out-of-date) = 72 *Professional status* = full professor, *N* = 6; *Professional status(1)* = research assistant, *N* = 154; *Professional status(2)* = assistant professor, *N* = 7 Omnibus test: Chi-Square = 77.923, df = 10, p = 0.000 Cox & Snell R-Square: 0.373; Nagelkerke R-Square: 0.500						

Table 1. Results of the binary logistic regression model

The correlation between any of the independent variables included in our dataset is in the range between -0.166 and 0.798[3] and the maximum variance inflation factor (VIF) is 2.94, which is below the critical threshold of 10. Thus, we assume that multicollinearity can be neglected and all independent variables have been included in the model. The Omnibus test for our model is highly significant ($\chi^2 = 77.923$, $p < 0.001$) and the Hosmer and Lemeshow measure of overall fit is not significant ($\chi^2 = 6.404$, $p = 0.602$), which jointly indicate that the goodness of fit of the model is acceptable. The Cox and Snell (0.373) and the Nagelkerke R-Squared (0.500) suggest a satisfactory explanatory power of the proposed model. Since choice based sampling leads to inconsistent intercept estimations, the intercept needs to be adjusted after estimating the model. Therefore, we subtract the constant $\ln(S_i/P_i)$ from the intercept's exogenous maximum likelihood estimates (cf. e.g., Manski and Lermann, 1977; Scott and Wild, 1986; King and Zeng, 2001). Here, S_i denotes the percentage of observations for alternative *i* in the sample (i.e., 0.57) and P_i denotes the percentage of observations for alternative *i* in the whole dataset (i.e., 0.12). All results are summarized in Table 1.

With respect to potential factors derived from DQ literature that might influence the currency of user data in business networks, we find that the odds of the *duration of the current employment*, as a measure for the age of the attribute value stated in the data attribute "Current Position", is negatively related to the likelihood that the attribute value is up-to-date (Odds Ratio = 0.604). Since the dependency is statistically significant at the 0.01 level, **H1** can be confirmed. Furthermore, the variables *professional status*, *professional status(1)*, and *professional status(2)* are statistically significant (p = 0.005, p = 0.017, and p = 0.001). Hence, the odds for an out-of-data attribute value are 99.9% higher (Odds Ratio = 0.001) for a research assistant (*professional status(2)*) with an on average shorter shelf-life of his or her current position (cf. section 2) than for a full professor. Thus, **H2** can be confirmed as well. Both findings are consistent to previous studies on measuring currency (Ballou et al., 1998; Heinrich and Klier, 2011), as they reveal that metadata indicating the age of an attribute value as well as supplemental data indicating an attribute value's shelf-life have a significant impact on the currency of user data in business networks.

[3] In the analysis sample (whole dataset in parenthesis) the highest correlation of 0.798 (0.755) exists between the *number of page impressions* and the *number of direct contacts*, followed by a correlation of 0.523 (0.539) between the *number of subscribed interest groups* and the *number of direct contacts*. All other correlations in both the analysis sample and the whole dataset are within]-0.500; 0.500[. Even though the VIFs indicate the absence of multicollinearity, we also estimated the model without including the *number of page impressions* and the *number of subscribed interest groups* and find qualitatively equivalent results.

Concerning potential factors influencing the currency of user data deduced from OSN, our results indicate that the *number of page impressions* is virtually not associated with the likelihood that the data attribute "Current Position" in a user profile is up-to-date (Odds Ratio = 1.000), even though the result is significant at the 0.1 level. **H3** can consequently not be confirmed. This result could be traced back to the fact that in business networks direct contacts induce a higher perceived level of social pressure than randomly visiting users, as they are more likely to know if the presented user data is actually up-to-date (cf., section 3). This is in line with the results for the *number of direct contacts*, which are positively associated with the likelihood that the data attribute "Current Position" in a user profile is up-to-date (Odds Ratio = 1.014). This dependency is significant at the 0.01 level and **H4** can thus be confirmed. The same holds true for *activity meter* (Odds Ratio = 6.390), which is also statistically significant at the 0.01 level. Hence, also **H5** can be confirmed. In contrast, the *number of subscribed interest groups* is not positively related to the likelihood that a data attribute is up-to-date (Odds Ratio = 0.978) and the dependency is not statistically significant (p = 0.720). Thus, **H6** cannot be confirmed. This might be due to the fact that users do not emotionally identify with groups in business networks as assumed by Social Influence Theory (cf. Cheung and Lee, 2010) but rather join to signal their group membership to others and to receive information. Finally, the analysis shows that the control variables *premium membership* (p = 0.938), *membership duration* (p = 0.113), and *gender* (p = 0.719) do not have a significant influence on the likelihood that the data attribute "Current Position" in a user profile is up-to-date.

In order to test the model's predictive power, we analyze the classification accuracy of the logit model. As Table 2 shows, the overall classification accuracy of the analysis and the holdout sample is 74.9% and 69.5%, respectively. Even though the classification accuracy of the analysis sample is slightly higher, the classification accuracy of the holdout sample (69.5%) is significantly ($p < 0.001$) higher than a chance model (Hair et al., 2006).

	Analysis sample (N = 167)			Holdout sample (N = 95)		
	Predicted		Correctly	Predicted		Correctly
	Out-of-date	Up-to-date	predicted	Out-of-date	Up-to-date	predicted
Out-of-date (observed)	51	21	70.8%	24	16	60.0%
Up-to-date (observed)	21	74	77.9%	13	42	76.4%
Overall Percentage			74.9%			69.5% †

† The overall classification accuracy of the holdout sample, at 69.5% (h), relative to the proportional chance criterion π, at 51.2% ($\pi = [(24+16)/95]^2+[(13+42)/95]^2$), is significant ($t = 3.554$, $p < 0.001$), where $t = (h-\pi)/(\pi(1-\pi)/\pi)^{0.5}$.

Table 2. Classification for analysis sample and holdout sample

In order to analyze the robustness of our results, we further checked the model with respect to our choice based sampling. By re-sampling the choice based sample ten times, all results could be confirmed. Moreover, we tested whether considering also social characteristics by including all significant explanatory variables (i.e., *number of page impressions*, *number of direct contacts*, and *activity meter*) is advantageous compared to a model considering significant, DQ specific explanatory variables only (i.e., *duration of the current employment* and *professional status*). We therefore estimated two independent models, one considering both types of explanatory variables and one considering solely DQ specific explanatory variables. Afterwards, we compared the resulting classification accuracy of both models revealing that the model considering both types of explanatory variables has significantly higher classification accuracy ($p < 0.05$). That is, our results provide first evidence that social characteristics should be taken into account when dealing with currency of user data in OSN.

5 Theoretical and Practical Implications

Our analysis first highlights that the attribute value's currency of the data attribute "Current Position" is negatively associated to the attribute value's age. That is, in line with DQ literature (e.g., Heinrich and Klier, 2011), the probability that an attribute value is up-to-date declines with its age. In conformance with further studies on data currency (e.g., Heinrich and Klier, 2009), our analysis depicts that also in

business networks supplemental data have a significant influence on the currency of an attribute value. Consequently, our analysis indicates that findings from DQ literature on the currency of attribute values in traditional information systems can be confirmed in the context of user data in OSN. Second, our analysis reveals that social characteristics such as social pressure induced by the number of direct contacts and users' activity have a significant impact on the currency of user data. By drawing on theories that consider the specific social characteristics of OSN and our real-world evaluation, we find further factors influencing the currency of user data in contexts such as OSN. Third, with respect to classification accuracy, our model also indicates that considering both DQ as well as OSN specific explanatory variables is advantageous when dealing with data currency in the context of OSN. Based on these results, metrics for the currency of attribute values in OSN, similar to the one proposed by Heinrich and Klier (2009), can be developed in future research. For example, the resulting metric for measuring the currency of a user's current position within a business network should take into account the following factors: the age of the value of the data attribute "Current Position" (i.e., the duration of employment), supplemental data (i.e., the professional status), the number of page impressions, the number of direct contacts, and the respective user activity. Based on currency estimates, recruiters could for instance prioritize possibly interesting job candidates in OSN instead of contacting them arbitrarily. That is, recruiters could decide to contact users with the highest estimated currency of their stated current position first. Moreover, the operators of business networks could stimulate and incentivize their users to keep their profiles up-to-date based on a currency estimation to improve DQ in an efficient and economically reasonable manner. For example, users with an estimated currency of their profile below a certain threshold could actively be motivated or even incentivized to update their user profiles. Thus, owing to high DQ, business networks and OSN could actually become even more valuable for recruiters or marketers and revenues from specialized membership offerings and advertising could be increased.

6 Conclusion

Despite the increasing acceptance and usage of OSN, DQ in OSN has hardly been researched so far. Particularly the currency of user data has been identified as a crucial success factor for both, users such as marketers and recruiters as well as providers of OSN. Findings from prior research on currency, however, have not been confirmed for the context of OSN yet. Moreover, neither theories taking into account the specific social characteristics of OSN nor real-world evaluations have been applied to learn more about potential factors influencing the currency of user data in OSN. By drawing on publicly available data of XING, we therefore empirically investigated factors influencing the currency of user data within this business network. Thus, the paper contributes by three means to the field of DQ: First, we confirmed factors influencing currency from prior research in the field of DQ for the context of business networks. Second, we found that social characteristics from OSN literature such as social pressure induced by the number of direct contacts and users' activity have a significant impact on the currency of user data. Third, we showed that taking into account these additional factors when estimating currency leads to significantly better results than relying solely on previously known factors from DQ literature. However, limitations offer areas for further research. First, other types of users than university employees should be considered and the sample size should be extended to enhance model robustness. However, our analysis of 130 user profiles of an international management consultancy already indicates that our dataset could be representative for other occupation groups. Second, our results are currently based on one business network and the data attribute "Current Position". Other business networks such as LinkedIn, general OSN such as Facebook, and further data attributes should be investigated in future research to gain more insights and general validity. Third, also an adapted metric for estimating currency could be designed and tested in order to evaluate the advantages of including the identified additional factors for users such as marketers and recruiters as well as providers of OSN. Finally, even though prior research indicates that currency is the most relevant DQ dimension in OSN, other dimensions such as completeness could be subject to future research. Despite these limitations, our paper constitutes a first step towards a deeper understanding of DQ and currency in contexts such as OSN.

References

Bagozzi, R.P. and Dholakia, U.M. (2002). Intentional social action in virtual communities. Journal of Interactive Marketing, 16 (2), 2-21.

Ballou, D.P. and Tayi, G.K. (1999). Enhancing data quality in data warehouse environments. Communications of the ACM, 42 (1), 73-78.

Ballou, D.P., Wang, R.Y., Pazer, H.L., and Tayi, G.K. (1998). Modeling information manufacturing systems to determine information product quality. Management Science, 44 (4), 462-484.

Ben-Akiva, M. and S.R. Lerman (1985). Discrete Choice Models, MIT Press, Boston.

Bonneau, J. and Preibusch, S. (2010). The privacy jungle: On the market for data protection in social networks. In Economics of information security and privacy (Moore, T., D. Pym, and C. Ioannidis Eds.), Springer, New York, 121-167.

Boyd, D.M. and Ellison, N.B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13 (1), 210-230.

Chai, K., Potdar, V., and Dillon, T. (2009). Content Quality Assessment Related Frameworks for Social Media. Lecture Notes in Computer Science 5593, 791-805.

Cheung, C.M.K. and Lee, M.K.O. (2010). A theoretical model of intentional social action in online social networks. Decision Support Systems 49 (1), 24-30.

Davison, H.K., Maraist, C.C, and Bing, M.N. (2011). Friend or foe? The promise and pitfalls of using social networking sites for HR decisions. Journal of Business and Psychology, 26 (2), 153-159.

Dholakia, U.M., Bagozzi, R.P., and Pearo, L.K. (2004). A social influence model of consumer participation in network- and small-group-based virtual communities. International Journal of Research in Marketing, 21 (3), 241-263.

Donath, J. and Boyd, D. (2004). Public displays of connection. BT Technology Journal, 22 (4), 71-82.

Ellison, N.B., Steinfield, C., and Lampe, C. (2007). The benefits of Facebook "friends": Social capital and college students' use of online social network sites. Journal of Computer-Meditated Communication, 12 (4), 1143-1168.

Evans, S. (2011). CBR Data Governance Survey 2010. Computer Business Review. http://bi.cbronline.com/features/cbr-data-governance-survey-dataflux-140111, accessed 2012-11-26.

Even, A. and Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. The DATA BASE for Advances in Information Systems, 38 (2), 75-93.

Falge, C., Otto, B., and Österle, H. (2012). Data Quality Requirements of Collaborative Business Processes. In Proceedings of the 45th Hawaii International Conference on System Sciences, Hawaii, 2012, 4316-4325.

Fishbein, M. and I. Ajzen (1975). Beliefs, attitudes, intention, and behavior: An introduction to theory and research. Addison-Wesley. Reading.

Gross, R. and Acquisti, A. (2005). Information revelation and privacy in online social networks. In Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, Alexandria, 71-80.

Hair, J.F., B. Black, B. Babin, R.E. Anderson, and R.I. Tatham (2006). Multivariate Data Analysis. Prentice Hall, London.

Haythornthwaite, C. (2002). Strong, Weak, and Latent Ties and the Impact of New Media. Information Society, 18 (5), 385-401.

Heidemann J., Klier, M., and Probst, F. (2012). Online social networks: A survey of a global phenomenon. Computer Networks, 56 (18), 3866-3878.

Heinrich, B. and Klier, M. (2009). A Novel data quality metric for timeliness considering supplemental data. In Proceesings of the 17th European Conference on Information Systems, Verona, Italy, 2651-2662.

Heinrich, B. and Klier, M. (2011). Assessing data currency - A probabilistic approach. Journal of Information Science, 37 (1), 86-100.

Hu, T., Poston, R.S., and Kettinger, W.J. (2011). Nonadopters of Online Social Network Services: Is it Easy to Have Fun Yet? Communications of the AIS, 29, article 29.

Janson, K., Schomburg, H., and Teichler U. (2006). Wissenschaftliche Wege zur Professur oder ins Abseits? Study for the German Academic International Network, New York.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. Political analysis, 9 (2), 137-163.

Klein, B.D. and Callahan, T.J. (2007). A comparison of information technology professionals' and data consumers' perceptions of the importance of the dimensions of information quality. International Journal of Information Quality, 1 (4), 392-411.
Lee, Y.W., Strong, D.M., Kahn, B.K., and Wang, R.Y. (2002). AIMQ: A methodology for information quality assessment. Information & Management, 40 (2), 133-146.
Leimeister, J.M., Sidiras, P., and Krcmar, H. (2006). Exploring success factors of virtual communities: The perspectives of members and operators. Journal of Organizational Computing and Electronic Commerce, 16 (3/4), 279-300.
Lewis, K., Kaufman, J., and Christakis, N. (2008). The taste of privacy: An analysis of college student privacy settings in an online social network. Journal of Computer-Mediated Communication, 14 (1), 79-100.
Lin, B. and Stasinskaya, V.S. (2002). Data warehousing management issues in online recruiting. Human Systems Management, 21 (1), 1-8.
Lin, H.-F. (2008). Determinants of successful virtual communities: Contributions from system characteristics and social factors. Information & Management, 45 (8), 522-527.
Livingstone, S. (2008). Taking risky opportunities in youthful content creation: Teenagers' use of social networking sites for intimacy, privacy and self-expression. New Media & Society, 10 (3), 393-411.
Manski, C. and Lerman, S. (1977). The Estimation of Choice Probabilities from Choice-Based Samples. Econometrica, 66 (1), 1977-1988.
Nahapiet, J. and Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage. The Academy of Management Review, 23 (2), 242-266.
Nosko, A., Wood, E., and Molema S. (2010). All about me: Disclosure in online social networking profiles: The case of FACEBOOK. Computers in Human Behavior, 26 (3), 406-418.
Oestreicher-Singer, G. and Zalmanson L. (2009). 'Paying for content or paying for community?' The effect of social involvement on subscribing to media web sites. In Proceedings of the 30th International Conference on Information Systems, Phoenix, USA, paper 9.
Page, X. and Kobsa, A. (2010). Navigating the Social Terrain with Google Latitude. In Proceedings of the iConference, Urbana-Champaign, USA, 174-178.
Pipino, L.L., Yang, W.L., and Wang, R.Y. (2002). Data Quality Assessment. Communications of the ACM, 45 (4), 211-218.
Provoost, L. (2009). Using swarm intelligence to make your corporate social network fly. http://www.capgemini.com/technology-blog/2009/07/using_swarm_intelligence_to_ma, accessed 2012-11-26.
Richter, D., Riemer, K., and vom Brocke, J. (2011). Internet social networking - Research state of the art and implications for enterprise 2.0. Business & Information Systems Engineering, 3 (2), 89-101.
Riggins, F.J. (2003). Market segmentation and information development costs in a two-tiered fee-based and sponsorship-based web site. Journal of Management Information Systems, 19 (3), 69-81.
Schaefer, C. (2008). Motivations and usage patterns on social network sites. In Proceedings of 16th European Conference on Information Systems, Galway, Ireland, 2088-2099.
Schrammel, J., Köffel, C., and Tscheligi, M. (2009). How much do you tell? Information disclosure behaviour indifferent types of online communities. In Proceedings of the 4th International Conference on Communities and Technologies, University Park, USA, 275-284.
Scott, A.J. and Wild, C.J. (1986). Fitting logistic models under case-control or choice based sampling. Journal of the Royal Statistical Society. Series B (Methodological), 48 (2), 170-182.
Sledgianowski, D. and Kulviwat, S. (2009). Using Social Network Sites: The Effects of Playfulness, Critical Mass and Trust in a Hedonic Context. Journal of Computer Information Systems, 49 (4), 74-83.
Strufe, T. (2010). Profile popularity in a business-oriented online social network. In Proceedings of the 3rd Workshop on Social Network Systems, Paris, France, article 2.
Utz, S. and Kramer, N.C. (2009). The privacy paradox on social network sites revisited: The role of individual characteristics and group norms. Journal of Psychosocial Research on Cyberspace, 3 (2), article 1.
Weitzel, T., Eckhardt, A., von Stetten, A., Laumer, S., Kaestner, T.A., and von Westarp, F. (2011). Recruiting trends 2011. Bamberg, Frankfurt am Main, Germany.
Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P.N., and Zhao B.Y. (2009). User interactions in social networks and their implications. In Proceedings of the 4th ACM European Conference on Computer Systems, Nuremberg, Germany, pp. 205-218.
XING (2011). Annual Report 2011.