Discussion Paper WI-333

Predicting Users’ Future Level of Communication Activity in Online Social Networks: A First Step towards More Advertising Effectiveness

by

Florian Probst

Paper 39
Predicting Users’ Future Level of Communication Activity in Online Social Networks: A First Step towards More Advertising Effectiveness

Florian Probst
FIM Research Center, Department of Information Systems Engineering & Financial Management, University of Augsburg
Universitaetsstr. 12, Augsburg, Germany
florian.probst@wiwi.uni-augsburg.de

ABSTRACT
Many Online Social Networks do not generate sustainable revenues through advertising, even though active usage has reached enormous scales. To enable more effective advertising strategies in Online Social Networks, it is essential to identify users who can affect a large number of friends, acquaintances, or other users in the network. In this context, especially users’ future level of communication activity in the Online Social Network plays an important role. A highly active past, however, does not guarantee high levels of future communication activity. Thus, approaches for the prediction of users’ future level of communication activity are needed. Therefore, we transfer a probability-based method that has been primarily developed to forecast purchasing behavior of customers to the context of users’ communication activity in Online Social Networks. In addition, we demonstrate the method’s applicability and suitability by using a publicly available dataset of Facebook.com.

Keywords
Online Social Networks, advertising effectiveness, communication activity, probability-based method, forecasting.

INTRODUCTION
In the last couple of years, Online Social Networks (OSN) have become popular Internet platforms that connect people around the globe. Thereby, the active usage of OSN has reached enormous scales: In March 2010, the OSN Facebook.com surpassed Google.com to become the most visited website of the week in the US (Dougherty, 2010), while a few months later, the number of active users of Facebook.com exceeded 500 million (Facebook, 2010). According to a recent study, two-thirds of the US Internet users already visit OSN each month, with 43% of them using OSN more than once a day (Alison, 2010). Thus, the phenomenon OSN has evolved into a global mainstream medium that generates an increasing social and economic impact.

However, many OSN face the question of how to leverage on their fast growing popularity to achieve sustainable revenues (Heidemann, Klier and Probst, 2010). Nowadays, the majority of OSN relies on an advertisement-based business model (Gnyawali, Fan and Penner, 2010). Nevertheless, many OSN do not generate sustainable revenues through advertising: Even though the worldwide advertisement spending on OSN is expected to grow from US$ 2.0 billion in 2008 to US$ 3.5 billion in 2013 (Williamson, 2009), OSN often do not know how to unleash this potential (Clemons, 2009; Lu and Hsiao, 2010). Therefore, more effective advertising strategies in OSN are needed in order to remain financially viable (Wen, Tan and Chang, 2009; Xu, Zhang, Xue and Yeo, 2008).

To tap the enormous potential originated by the dramatic increase in the popularity of OSN, particularly the identification of users who can affect a large number of friends, acquaintances, or other users in an OSN is essential (Heidemann et al., 2010; Hill, Provost and Volinsky, 2006). Such users can for example be addressed in marketing campaigns to achieve a high awareness of a product or service. This strategy is often referred to as network-based marketing, word-of-mouth marketing, or viral marketing (cf. e.g. Brown, Broderick and Lee, 2007; Hill et al., 2006; Lee, Lee and Lee, 2009). The underlying key assumption is that users propagate “positive” information about a product or service after they have either been made aware by traditional marketing techniques or experienced it by themselves (Hill et al., 2006). In this context, Ray, Bernoff and Wise
(2010) found “that people in the US generate more than 500 billion online impressions on each other regarding products and services”. However, as only “16% of online consumers generate 80% of these impressions” (Ray et al., 2010), only a subset of users is particularly valuable for marketers (Trusov, Bodapati and Bucklin, 2010). For the identification of these users, especially users’ future level of communication activity, i.e. each user’s number of future communication activities, plays an important role (Cheung and Lee, 2010; Hoffman and Fodor, 2010; Willinger, Rejaie, Torkjazi, Valafar and Maggioni, 2010; Xu et al., 2008). But even though some users might have been highly active in the past, high levels of future communication activity cannot be taken for granted (Cummings, Butler and Kraut, 2002; Viswanath, Mislove, Cha and Gummadi, 2009). Hence, approaches for the prediction of users’ future level of communication activity are needed, which might serve as a first step towards more effective advertising strategies in OSN. However, to the best of our knowledge, approaches for predicting users’ future level of communication activity in OSN are missing. Therefore, we transfer a probability-based method, which has been primarily developed by Fader, Hardie and Lee (2005), to forecast purchasing behavior of customers, to the context of users’ communication activity in OSN. In addition, we demonstrate the practical applicability of the method and evaluate its suitability for predicting users’ future level of communication activity in OSN by using a dataset of Facebook.com.

After the discussion of the general relevance of the problem and its motivation within this introduction, we specify the problem context and review prior research on users’ communication activity in OSN. Thereby, we identify our research gap. Afterwards, we propose our artifact as a probability-based method. In the penultimate section, we illustrate the method’s practical applicability and suitability to predict users’ future level of communication activity in OSN. Finally, we conclude with a summary of the results and an outlook on future steps.

PROBLEM CONTEXT AND RELATED WORK

OSN are a particular type of virtual communities1 (Dwyer, Hiltz and Passerini, 2007). According to 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”. Current OSN are primarily used for maintaining existing relationships from an offline context, although they also allow for creating pure online relationships (Ellison, Steinfield and Lampe, 2007; Lampe, Ellison and Steinfield, 2006). Especially the visibility and searchability of the users’ relationships is a distinctive feature of OSN. Thus, OSN can “create substantial value for the individuals who participate in them, the organizations that sponsor them, and the larger society in multiple ways” (Agarwal, Gupta and Kraut, 2008, p. 243). The majority of OSN that rely on the advertisement-based business model, however, face the challenge to tap the enormous potential originated by the dramatic increase in the popularity of OSN in order to generate sustainable revenues (Clemons, 2009; Lu and Hsiao, 2010). Therefore, more effective advertising strategies are needed (Wen et al., 2009; Xu et al., 2008). Word-of-mouth marketing, for example, “is known to be the most effective form of advertising, but, until recently, was the most expensive” (Jacks and Salam, 2009, p. 2). In contrast, word-of-mouth and targeted marketing can be much more efficient and cost-effective in OSN (Trusov, Bucklin and Pauwels, 2009; Zhang, Wang and Xia, 2010). However, only a small subset of users has actually significant influence on other users (Trusov et al., 2010). Thus, the identification of the most influential users is necessary to enable more effective advertising strategies in OSN (Heidemann et al., 2010; Trusov et al., 2010; Zhang et al., 2010). In this context, literature indicates that particularly the identification of users with high levels of future communication activity is essential (Cheung and Lee, 2010; Hoffman and Fodor, 2010; Willinger et al., 2009; Xu et al., 2008).

Communication activity in OSN can be any sort of interaction among users facilitated by methods provided by OSN, such as messages or wallposts (cf. Wilson, Boe, Sala, Puttaswamy and Zhao, 2009). Prior work emphasizes the importance of users’ communication activity: “No matter what resources are available within a structure, without communication activity those resources will remain dormant, and no benefits will be provided for individuals” (Butler, 2001, p. 350). Recent work supports that the value of OSN lies in the communication activity between users (Krasnova, Hildebrand and Günther, 2009; Willinger et al., 2009). Latest studies further show that user’s communication activity is highly relevant for advertising effectiveness in OSN (e.g. Cheung and Lee, 2010; Ganley and Lampe, 2009). The record of communication activities between users in OSN can be used to identify users’ with high levels of communication activity in the past (Xiang, Neville and Rogati, 2010). However, prior research found that users who have been highly active in the past are not necessarily highly active in the future (Cummings et al., 2002; Viswanath et al., 2009). Hence, approaches for the prediction of each user’s future level of communication activity need to “abandon the traditional treatment of OSNs as static networks” (Willinger et al., 2010, p. 49) and incorporate the dynamic of communication activity in OSN.

1 A definition of virtual communities can be found in Leimeister, Sidiras and Krcmar (2004). While Dwyer et al. (2007) and Boyd and Ellison (2007) use the term social networking site, we are using the term OSN throughout the paper synonymously.
Plenty of research addressing the dynamic nature of OSN can be found with respect to network structures in OSN. Thereby, previous studies focus on the evolution of network structures in general (for an overview cf. Dorogovtsev and Mendes, 2003) and the establishment of static social links, i.e. friendship relationships, between users in particular (e.g. Liben-Nowell and Kleinberg, 2007). However, Xu et al. (2008, p. 14) emphasize “that interaction information is invaluable to marketers, more important than the static links”. Consequently, some studies take into account the dynamic nature of OSN and users’ communication activity: De Choudhury, Sundaram, John and Seligmann (2007), for example, determine the intent to communicate and the communication delay between users based on several contextual factors in OSN, such as the relevance of a topic. Therefore, it is first assumed that a user receives a message. Second, the likelihood that the receiver will communicate with the sender on a particular topic and the delay in communication are predicted. However, to the best of our knowledge existing approaches do not allow for determining conditional expectations about users’ future level of communication activity in OSN on an individual-level, i.e. making predictions about each user’s future level of communication activity given information about his or her past communication activity. Thus, approaches for predicting users’ future level of communication activity in OSN are missing.

METHOD

For predicting users’ future level of communication activity, we draw on a probability-based method that has been primarily developed by Fader et al. (2005) to address a very similar problem, i.e. to forecast purchasing behavior of customers. This beta-geometric/negative binomial distribution (BG/NBD)-based method goes back to the highly regarded “counting your customers” framework introduced by Schmittlein, Morrison and Colombo (1987). In the following, we discuss the BG/NBD-based method as possibility to predict users’ future level of communication activity in OSN.

First, it is assumed that a user $i \in \{1,2,\ldots,N\}$ is active at $t_0 = 0$. This will generally be satisfied if we take $t_0$ as the point of time at which the user’s initial communication activity occurred. Furthermore, the BG/NBD-based method requires three pieces of information about each user, represented by $(X_i=x_i,t_x,T_i)$. Thereby, $x_i \in \mathbb{N}$ denotes the “frequency”, i.e. the number of communication activities after the initial communication activity within the observation period $(0,T_i)$, $t_x \in \mathbb{R}^+$ (with $0 \leq t_x \leq T_i$) is the “recency”, i.e. the point of time of the last communication activity (if $x_i=0$ then $t_x=0$), and $T_i \in \mathbb{R}^+$ represents the length of the observation period, i.e. the time between the initial communication activity and the end of the observation period. Notice, that the relation between the period $(0,T_i]$ and calendar time will vary from user to user depending on when the user’s initial communication activity occurred. Based on this information, we aim to predict each user’s future level of communication activity during the forecasting period of length $t_i$ (cf. Figure 1).

To predict the future level of a user’s communication activity, assumptions about the communication activity process and the time that users stay active are needed (cf. Fader et al., 2005; Schmittlein et al., 1987):

**Assumption 1 (Users’ communication activity):**

While being active, each user $i$’s communication activity during a forecasting period of length $t_i$ follows a homogeneous Poisson process with rate $\lambda_i \in \mathbb{R}^+$. This is equivalent to assuming that the time between communication activities is distributed exponentially with rate $\lambda_i$ and mean $\frac{1}{\lambda_i}$ (Fader et al., 2005).

Even though the Poisson assumption has a long validated history for frequently purchased consumer goods (Ehrenberg,
1972) and is often used to describe aspects of human activity (Malngren, Stouffer, Motter and Amaral, 2008), the transferability to the context of online communication behavior is also critically discussed (Malngren, Hofman, Amaral and Watts, 2009). Particularly circadian cycles, i.e. for example some people’s habit to answer e-mails only in the morning and the evening, are reported to lead to heavy-tailed power-law distributions of inter-communication times (cf. Malngren et al., 2008). However, many studies confirm the applicability of a homogeneous Poisson process for modeling repeated communication behavior within these cycles, i.e., referring to our example stated above, while being in one session of answering e-mails (e.g. in the morning) (Malngren et al., 2009). In a first approximation, we assume that users in OSN in particular (cf. Ahmed, Berchmans, Neville and Kompella, 2010). As there are users with high and with low levels of communication activity as well as users with high probability to drop out and vice versa, some heterogeneity assumptions are mandated. The gamma distribution is a flexible distribution and can capture the spirit of most of the reasonable distributions on λ and p (cf. Fader et al., 2005; Schmittlein et al., 1987):

Assumption 2 (Users’ probability of becoming inactive):

After any communication activity, user i becomes inactive with probability p_i ∈ [0;1]. The point at which the user “drops out” is distributed across communication activities according to a (shifted) geometric distribution with probability mass function:

\[ P(\text{inactive immediately after } j\text{th communication activity}) = p_i(1-p_i)^j, \text{ with } j=1,2,3,... . \]

This assumption is supported by prior work on customer retention in general (cf. Fader and Hardie, 2007) and the dropout process of users in OSN in particular (cf. Ahmed, Berchmans, Neville and Kompella, 2010). As there are users with high and with low levels of communication activity as well as users with high probability to drop out and vice versa, some heterogeneity assumptions are mandated. The gamma distribution is a flexible distribution and can capture the spirit of most of the reasonable distributions on λ_i and p_i (cf. Fader et al., 2005; Schmittlein et al., 1987):

Assumption 3 (Heterogeneity in users’ communication activity):

Heterogeneity in λ_i follows a gamma distribution with shape parameter r ∈ IR^+, scale parameter α ∈ IR^+, and probability density function:

\[ f(λ_i|α, r) = \frac{α^r}{Γ(r)} λ_i^{r-1} e^{-α λ_i}, \text{ with } λ_i > 0. \]  

(1)

Assumption 4 (Heterogeneity in users’ probability of becoming inactive):

Heterogeneity in p_i follows a beta distribution with probability density function:

\[ f(p_i|a, b) = p_i^{a-1}(1-p_i)^{b-1}/B(a, b), \text{ with } 0 ≤ p_i ≤ 1, \]

(2)

where B(a,b) with a ∈ IR^+ and b ∈ IR^+ is the beta function, which can be expressed in terms of gamma functions, i.e.:

\[ B(a,b) = Γ(a)Γ(b)/Γ(a+b). \]

Finally, there is no a priori reason to favor a positive correlation between users’ future communication activity and their probability to drop out over a negative correlation. On the one hand, users with high levels of communication activity may have more frequent opportunities to be disenchanted by the OSN (e.g. due to privacy concerns, system malfunctions) and drop out. On the other hand, these users are probably more strongly attached to the OSN and hence less easily disenchanted. Hence, neglecting interdependencies seems to be a reasonable first approximation:

Assumption 5 (Independence of users’ communication activity and probability of becoming inactive):

The transaction rate λ_i and the dropout probability p_i vary independently across users.

With the random variable Y(t_i) denoting the number of a user i’s future communication activities initiated in the forecasting period of length t_i, we finally aim to derive the expected number of a user’s future communication activities by computing the conditional expectation E(Y(t)|X_i=x_i,t_i,T_i) for a user i with observed behavior (X_i=x_i,t_i,T_i). However, λ_i and p_i (cf. assumptions 1 and 2) are unobserved. While there is usually not enough observed user-specific behavior to reliably estimate these parameters for each user, there is generally enough information to estimate the distribution of (λ_i, p_i) over all users. Hence, we can derive the desired probabilities for a randomly chosen user by taking the expectation of the individual-level results over the mixing distributions for λ_i and p_i as given in formulas (1) and (2) (cf. Fader et al., 2005). Thus, the four BG/NBD parameters (r, α, a, b) (cf. assumptions 3 and 4) can be estimated via the method of maximum likelihood. For N users, the sample log-likelihood function is given by (cf. Fader et al., 2005):

\[ LL(r, α, a, b) = \sum_{i=1}^{N} \ln \left[ \frac{1}{r, α, a, b} X_i = x_i, t_i, T_i \right], \]

(3)
which can be maximized by using standard numerical optimization routines. Thereby, \( L(r, a, a, b) \) represents the likelihood function for a single user \( i \), which can be derived according to formula (6) in Fader et al. (2005). Afterwards, we can calculate the expected number of each user’s communication activities in the forecasting period of length \( t_i \) by (for a detailed derivation of the formula cf. Fader et al., 2005):

\[
E(Y_i | X_i = x_i, t_i, r, a, a, b) = \frac{\alpha + b + x_i - 1}{a - 1} \frac{\alpha + T_i}{\alpha + T_i + t_i}^{r + x_i} 2F_1 \left( r + x_i, b + x_i; a + b + x_i - 1; \frac{t_i}{\alpha + T_i + t_i} \right) \\
= \frac{a + b + x_i - 1}{b + x_i - 1} \frac{\alpha + T_i}{\alpha + t_i}^{r + x_i} 1 + \delta_{x_i>0} \frac{\alpha}{b + x_i - 1} \frac{\alpha + T_i}{\alpha + t_i}^{r + x_i},
\]

where \( 2F_1(.) \) is the Gaussian hypergeometric function, which can be closely approximated with a polynomial series, and \( \delta_{x_i>0} = 1 \) if \( x_i > 0 \), 0 otherwise (cf. Fader et al., 2005). By calculating the expected number of each user \( i \)’s future communication activities, users’ future level of communication activity in OSN can be predicted. In the following, we demonstrate the method’s practical applicability and evaluate its suitability for predicting users’ future level of communication activity.

DEMONSTRATION AND EVALUATION

As many other OSN, Facebook.com allows users to set up personal profiles and to establish undirected social links by entering virtual friendship relationships. One of the most popular mechanisms for communication activity within OSN in general and Facebook.com in particular is a message board called “wall” that is included in each profile (Wilson et al., 2009).

For the demonstration and evaluation of the BG/NBD-based method for predicting users’ future level of communication activity, we use a publicly available dataset provided by Viswanath et al. (2009) that has also been used and described in detail in Heidemann et al. (2010). It contains 63,731 users of the Facebook.com New Orleans Network connected by 817,090 undirected social links and exhibits the OSN specific characteristics (cf. Heidemann et al., 2010). The dataset also contains information on users’ communication activity in terms of 876,687 wallposts initiated and received by the users covered by the dataset. Each wallpost includes information about the initiator, the receiver, and the time at which the wallpost was made. Overall, the wallposts span 227 weeks from September 14, 2004 to January 22, 2009. In the following, we use these wallposts to represent users’ communication activity. To account for the potential bias induced by the strong growth of the number of active users after week 175, we chose two scenarios, i.e. a low volatility scenario spanning nine weeks from week 150 to 158 (scenario 1) and a high volatility scenario spanning from week 216 to 224 (scenario 2). Figure 2 displays the development of the number of wallposts and the number of distinct active users covered by the dataset over time.

![Figure 2. Number of Wallposts and Number of Active Users over Time](image)
Table 1 summarizes both scenarios’ characteristics.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Active Users</td>
<td>9,815</td>
<td>25,182</td>
</tr>
<tr>
<td>Total Number of Wallposts</td>
<td>60,138</td>
<td>130,824</td>
</tr>
<tr>
<td>Average Number of Active Users per Week</td>
<td>2,994</td>
<td>6,965</td>
</tr>
<tr>
<td>Standard Deviation of Active Users per Week</td>
<td>100</td>
<td>894</td>
</tr>
<tr>
<td>Average Number of Wallposts per Week</td>
<td>6,682</td>
<td>14,536</td>
</tr>
<tr>
<td>Standard Deviation of Wallposts per Week</td>
<td>296</td>
<td>2,182</td>
</tr>
</tbody>
</table>

Table 1. Overview of Scenarios’ Characteristics

For each scenario, we aim to predict the users’ number of wallposts during a forecasting period of three weeks, i.e. from week 159 to 161 (scenario 1) and week 225 to 227 (scenario 2), respectively. Therefore, we first derive each user $i$’s observed behavior ($X_i = x_i, t_i, T_i$) and estimate the parameters ($r, a, a, b$) by applying formula (3). Table 2 summarizes the results.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>0.485</td>
<td>0.420</td>
</tr>
<tr>
<td>$a$</td>
<td>0.332</td>
<td>0.293</td>
</tr>
<tr>
<td>$a$</td>
<td>0.433</td>
<td>0.742</td>
</tr>
<tr>
<td>$b$</td>
<td>4.346</td>
<td>5.555</td>
</tr>
</tbody>
</table>

Table 2. Estimation Results

Second, we calculate for both scenarios the expected number of each user’s wallposts in the forecasting period according to formula (4). To evaluate the suitability of the BG/NBG-based method for predicting users’ future level of communication activity in OSN, we apply the evaluation approach suggested by Fader et al. (2005). In Figure 3, we report the average of the predicted along with the average of the actual number of wallposts that took place in the forecasting period broken down by the number of wallposts in the users’ observation periods for scenario 1.

Figure 3. Evaluation Scenario 1
The virtually absent deviation between the users’ predicted and actual wallposts highlights that in scenario 1 the BG/NBG-based method provides excellent predictions of the users’ number of wallposts in the forecasting period. Figure 4 displays the evaluation’s results for scenario 2.

![Figure 4. Evaluation Scenario 2](image)

Here, the evaluation reveals that the users’ actual level of communication activity is slightly underestimated. However, the deviation of 0.4 on average is quite small and constant. Thus, we think future work will be able to account for high volatility in the data by slightly adapting the method. Taken together, the BG/NBG-based method seems to be suitable for the prediction of users’ future level of communication activity in OSN.

**CONCLUSION**

Even though active usage has reached enormous scales, the majority of OSN relying on the advertisement-based business model face the challenge of generating sustainable revenues. Particularly the identification of users with high levels of future communication activity plays an important role when developing more effective advertising strategies by addressing users deliberately. For the identification of these users, we transferred a probability-based method, which has been primarily developed to forecast purchasing behavior of customers, to the context of users’ communication activity in OSN. The application and evaluation illustrated that the BG/NBG-based method seems to be suitable for predicting users’ future level of communication activity in OSN. Even though the method seems to be rather complex at first sight, the necessary parameters can be derived easily and the calculation can be automated and even be done in Excel, as pointed out by Fader et al. (2005). Nevertheless, future work is needed to further evaluate the approach (e.g. regarding further sample groups), for instance by using other datasets and by considering also economic aspects. The method’s formal denotation implicates assumptions and limitations, for instance regarding the underlying probability distributions. Future work needs to address these issues either by confirming or by relaxing these assumptions. Moreover, upcoming challenges, for instance due to changing privacy practices, need to be carefully observed and considered when developing approaches for the identification of users with high levels of future communication activity. Finally, in line with prior research on the identification of influential users in OSN (e.g. Trusov et al., 2010) we did not address the question how responsive highly active users are to certain marketing strategies (e.g. word-of-mouth marketing). Even though this question is subject to future research, the BG/NBG-based method serves as a first step towards more effective advertising strategies in OSN.
REFERENCES


