Is Editing More Rewarding Than Discussion?

A Statistical Framework to Estimate Causes of Dropout from Wikipedia

Ulrik Brandes
University of Konstanz, Germany
Ulrik.Brandes@uni-konstanz.de

Patrick Kenis
TiasNimbas Business School & Tilburg University, Netherlands
p.kenis@tiasnimbas.edu

Jürgen Lerner
University of Konstanz, Germany
lerner@inf.uni-konstanz.de

Denise van Raaij
Tilburg University, Netherlands
D.P.A.M.Korssen-vanRaaij@uvt.nl

ABSTRACT

In this paper we address the question: what causes formerly active Wikipedians to stop contributing? Seen from a different angle, we estimate characteristics of users, pages, or the whole system that increase or decrease the probability of dropout. We propose a general statistical method with which hypothetical causes of dropout can be tested. With this method it can be analyzed whether the emerging structures in Wikipedia function as incentives preventing Wikipedians to stop contributing. Applying this method to a selection of active users reveals, among others, that participation in discussion pages, as well as editing controversial pages, increases the dropout hazard, whereas editing general content pages has an attenuating effect on dropout. Although our method is solely illustrated on Wikipedia, it can be easily applied to other Web 2.0 applications.

Keywords

Wikipedia, lifetime-analysis, missing Wikipedians, motivation, frustration

1. INTRODUCTION

As any Web 2.0 application, Wikipedia needs, in order to grow and improve, a large number of motivated contributers. Given this fact, it is crucial and insightful for Web 2.0 researchers to learn about the causes to contribute and, as the other side of the coin, learn about the causes to stop contributing. Here we are interested in emerging mechanisms in Wikipedia that either motivate and reward contributors or frustrate users making them to leave as Wikipedians. Although these mechanisms can have an implicit nature (i. e., have not been designed as systematic feedback systems that aim at rewarding contributors [12]), increased knowledge in their functioning could be a first step in helping system designers and administrators to sustain enthusiastic users. In this paper, we focus on active users (i. e., users who performed a certain minimum number of contributions) and attempt to find factors that influence the probability whether such a user *survives as a Wikipedian* (i. e., continues to contribute) or *dies as a Wikipedian* (i. e., not contributes anymore). The restriction to active users is mostly due to statistical reasons (for inactive users we do not have sufficient data) but, arguably, the active users are also the more interesting ones.

Causes for dropout can be manifold and we distinguish between factors that are *exogenous* and factors that are *endogenous* to Wikipedia. Exogenous factors include demographic variables such as age, gender, education level, marriage status, profession, or occupation as well as external events such as getting a new job or getting children. Endogenous factors include everything that can be determined from the history of Wikipedia, i. e., information about edits, discussion, elections for administrator status, featured article voting, user blocking, page blocking and so on. While many exogenous factors may strongly influence the decision to not contribute anymore (in some cases, simply for the reason that the user does no longer have time to spend days or nights editing Wikipedia), we do not use them in this paper. The major reason for this decision is that we are attempting to uncover which features that are endogenous to the system function as incentives for sustained contribution and, vice versa, which endogenous features trigger dropout of Wikipedians. Such information can (to some degree) be used to design and shape Web 2.0 applications in order to enhance motivation of contributors.

Since we do not use exogenous factors—although they might influence the dropout probability—it seems to be obvious that there will be cases of dropout that are not well described by our model. We emphasize that we do not attempt to maximize the precision of predicting dropouts; rather, the goal of our analysis is to test statistically whether specific endogenous factors do, yes or no, increase or decrease the probability of leaving Wikipedia—thereby getting a better understanding which emerging and often implicit mechanisms contribute to sustain users. Such results are very useful because designers or administrators of Web 2.0 applications might use them to mitigate causes for dropout or add features that decrease dropout probability—even if the empirical time-to-dropout data contains unexplained variance due to exogenous factors. An additional consequence of our approach is that we are able to better understand the social collaboration process in Wikipedia by detecting characteristics that distinguish high-quality collaboration from low-quality collaboration; while an obvious quality dimension would be the quality of the encyclopedic entries, we claim that keeping contributors motivated is another very

WWW2009, April 20-24, 2009, Madrid, Spain.
.

important aspect of quality of the system (also see Sect.2.1).

Even if we do not use exogenous predictors for dropout in this paper, we emphasize that the general statistical method presented in Sect. 3 is applicable to all kinds of predictors, independent on whether they stem from log-data, demographic data, or questionnaire-based surveys.

In Sect. 2 we put the topic of this paper into the context of a broader research project, provide background on statistical methods for lifetime analysis, and review related work on Wikipedia research. Section 3 presents our statistical framework to model dropouts from Wikipedia. In Sect. 4 we report on the results of an empirical analysis using this model and Sect. 5 indicates future work.

2. BACKGROUND

2.1 Dropout Hazard as a Proxy for Quality

The long-term goal of this project is to gain insight into the social collaboration process in community forms of organizations (in contrast to formal or hierarchical organizations) that rise at the Internet and that we refer to as *webbased information communities* (*WebICs*). "WebICs are defined as work systems facilitated by the Internet infrastructure and composed of voluntary actors that attempt to produce a product or service such as software or encyclopedic information [2]." WebICs are organized in an informal way and are governed and coordinated by flows and linkages between actors [11]. Based on existing knowledge in the field of organization studies we argue that one of the success factors of WebICs is this implicit and emergent governance and coordination structure. However, since WebICs are not successful by definition, our research attempts to find out the characteristics of high-quality and low-quality collaboration structures.

Quality of Wikipedia most often refers to the quality of its encyclopedic entries: For instance it has been suggested that various forms of vandalism are indicators of (low) quality of articles [18]. Others have applied self-assessment criteria developed in Wikipedia, such as distinctions between excellent featured pages and worth-reading featured pages [15], or featured versus controversial pages [2]. Another way to assess the quality of articles is to present a number of selected Wikipedia entries to scientific experts [6].

However, quality of Wikipedia does not only mean quality of its encyclopedic articles; instead we argue that the dropout hazard of Wikipedians can also function as a proxy for quality of the system or certain parts of it (needless to say that a high dropout hazard is interpreted as pointing to low quality). This approach is based on the observation that: "Wikipedia operates from the presumption that any individual's knowledge is by definition incomplete and that ongoing revisions enabled by mass collaboration tools and involving a large group of eyeballs will produce a reliable yet continually evolving knowledge repository [5, p.361]." As a consequence, the ability of Wikipedia to prevent turnover and motivate Wikipedians to continue to contribute can be understood as a quality indicator of its governance and coordination structure. Turnover in formal, hierarchical organizations is associated with the loss of human capital and thus the loss of hiring and training investments [14]. Turnover in the context of Wikipedia can be associated with the loss of work force, their skills and knowledge and consequently, the decrease of production of encyclopedic knowledge. While formal, hierarchical organizations can manage employee commitment through, among others, economic incentives, formal training, contracts, and formal supervision procedures, Wikipedia can only rely on non-economic incentives to sustain contributors commitment [12]. Hence, if Wikipedia is able to preserve large numbers of highly contributing users, it is likely to produce higher outcome quality than if it lacks the ability to motivate contributors.

2.2 Statistical Methods for Lifetime Analysis

Lifetime analysis (also referred to as *time-to-event analysis*, *failure analysis*, or *survival analysis*) is an area of statistics that is concerned with modeling the elapsed time until a specific event happens; a general reference is given by Lawless [9]. Using a customary vocabulary, lifetime analysis models the time until a certain object *dies*, where *death* is sometimes meant in a metaphorical way. Lifetime analysis is frequently used in medicine, engineering, social science, and political science, among others. For instance, in medicine researchers are interested in how long a patient suffering a certain illness survives; engineers might be concerned with how long it takes until a manufactured item (e. g., a computer) breaks down. In this paper we are interested in the dropout of Wikipedians, i. e., in the events in which formerly active Wikipedia users stop contributing.

Besides estimating the actual survival times, another goal of lifetime analysis is to discover factors that increase or decrease the probability to die. Returning to the above examples, a specific pharmaceutical treatment may or may not empirically increase the survival time of patients; the lifetime of a computer may be dependent on the specific machine that manufactured it (potentially pointing to faults of machines). As already mentioned, we are attempting to uncover the reasons for dropout in Wikipedia, i. e., which factors increase or decrease the probability of dropout.

Lifetime analysis is often confronted with specific properties of the data that require special care. In many cases (and also in our case) lifetime analysis is faced with so-called *right-censoring*, meaning that some of the selected instances have not died at the time of data collection. Ignoring these survivors would introduce a serious bias into the analysis (intuitively, it would be hard to learn about the causes of survival, if surviving instances were discarded). Instead our model has to deal with the fact that for one part of the instances (namely those that died, later in this paper referred to as *dropouts*) we know the time when the individual died and for the other part of instances (later in this paper referred to as *survivors*) we only know that they survived beyond a certain point in time. See Sects. 3.3 and 3.5 how these instances are treated differently. Another issue to take care of is the definition of when a specific individual enters the *risk set* (i. e., the set of individuals that have a non-zero probability to die). We note first that in our case individuals (i. e., contributors of Wikipedia) enter the risk set at different time points, namely at the time of their first edit. However, since we restricted our analysis to *active* users (see Sect. 3.2 for a definition of an active user) we introduced a further bias: by discarding inactive users the probability of reaching the active state is artificially set to one (if a user died before, it would not be in our set of instances). Section 3.5 shows how to correct for this bias. Nevertheless, we stress that even with this correction it would not be valid to generalize findings to inactive users: those that dropout

quickly might do so for totally different reasons than those that reach the active state.

2.3 Further Related Work

Wikipedia—besides being a popular Web page—has become a popular case in academic research. Several papers visualize certain aspects of the history of Wikipedia pages, i. e., the development of their content over time. The *history flow* visualization [18, 19] shows how sentences persist over time or get deleted at later revisions. Other researchers constructed and visualized networks encoding how users interact with the edits of others, e. g., [8, 16, 3, 2]. The revision history of Wikipedia articles has been further used to distinguish the edit behavior of different user groups [7], to define reputation or Wikipedians [1], to estimate the impact of vandalism [13], and to identify controversial articles [20]. We are not aware of any work that quantitatively analyzes causes for dropout of Wikipedians, which is the topic of the current paper. However, Lento et al. [10] examined causes for continued participation in the Wallop Weblogging system; a difference to their approach is that our method takes the effects of *time-varying* explanatory variables into account.

3. METHOD

3.1 Data

The selection of instances and the extraction of the explanatory variables is mostly based on the so called stub files from the latest available database dump of the English Wikipedia (see `http://download.wikimedia.org`). These stub files contain metadata (most notable page title, username, and timestamp) of every revision on every page (including talk pages etc.) since the launch of Wikipedia. The dump that we used for this paper has been started on October 8th, 2008. Although the file contains edits with later timestamps, we ignore these and take October 8th, 2008 as the day of data collection. The uncompressed XML-file has a size of 66 gigabytes. Although this is quite large, it is nevertheless manageable since the needed information can be extracted in a sequential manner.

Besides the history stub file, the content of two additional Wikipedia pages have been used: The list of users on the page `Wikipedia:Missing Wikipedians` is helpful for selecting dropout instances (see Sect. 3.2 for details) and the page `Wikipedia:List of controversial issues` is used for the computation of one of the explanatory variables (see Sect. 3.4 for details).

3.2 Selection of Instances

As already noted in the introduction, we restrict our analysis to active users which are defined as users that performed a given minimum number of edits. These active users are later partitioned into *dropouts* (those who are known to have stopped contributing at a certain moment) and *survivors* (those who are known to continue editing beyond the time of data collection). We note that some active users fall between these two categories, i. e., for those users we do not have sufficient information to decide whether they are dropouts or survivors; those users are discarded.

More precisely, the dropouts are (a subset of) users listed on the page `Wikipedia:Missing Wikipedians`. This page has been mentioned in *The Economist* in an article about Wikipedia stating that "It serves as a reminder that frustration at having work removed prompts many people to abandon the project [4]." The first lines of the missing Wikipedian page already give an intuitive definition of what is a missing Wikipedian:

> This is a list of Wikipedians who are no longer an integral part of the community. [. . .] Wikipedians who no longer edit due to confirmed death should instead be added to `Wikipedia:Deceased Wikipedians`.
>
> [. . .]
>
> Please do not add people to this list who were never an integral part of the community. Don't add users with fewer than about 1,000 edits. Do not add people unless you are certain they have left, do not add anonymous users identified by their IP address (they could have created an account and still be contributing, or they might have a roaming IP address) and do not add yourself.

To make things precise we define (motivated by the above quotation) an *active Wikipedian* to be a logged-in user (in contrast to anonymous users identified by IP addresses) who is not a robot (i. e., not a software program that performs routine tasks) and who has performed at least 1,000 edits. From the database dump we derive that slightly more than 19,000 users qualify as active Wikipedians.

Dropout instances. To define the set of dropouts we start with all users listed on `Wikipedia:Missing Wikipedians`, yielding 501 users. From this set we deleted all those that made fewer than 1,000 edits, leaving us with 465 users. In order to not just trust the editors of the missing Wikipedian page we further delete all those that edited on or later than September 1st, 2008 (a bit more than one month before data collection). This gives us our final set of dropouts containing 413 users.

Survivor instances. For the survivors we start with the active Wikipedians, delete all those that are listed on the page of missing Wikipedians, and further delete all those that performed less than 30 edits in the time from July 1st, 2008 until the day of data collection. With the last step we want to exclude Wikipedians that do not qualify as dropouts but that are nevertheless not very active anymore; these users are simply harder to interpret. However, we suggest that formerly active Wikipedias that have not been listed on the page of missing Wikipedians (the *un-missed* dropouts) are an interesting population for future research. Altogether, the set of survivors contains 10, 454 users.

3.2.1 Notes on the Selection of Instances

We have chosen to select dropouts via the list of missing Wikipedians since this gives us some confidence that those users have indeed decided to stop participating, rather than just taking a break. However, it should be noted that this selection strategy implies that, strictly spoken, we estimate the causes for ending up on the page of missing Wikipedians, rather than the causes for dropout. Since only Wikipedians that are (well) known to at least one other user are put on this page, this selection procedure could introduce a bias in the analysis. We will analyze in future work the pros and

cons of alternative ways to divide active users into dropouts and survivors.

3.3 Statistical Model for Time-to-Dropout

While the procedure to select dropouts and survivors from Sect. 3.2 reflects a particular choice—giving emphasis to users that are recognized as missing by others—the model that is presented now is independent on the particular selection strategy and is (with a slight adaption in notation) also not restricted to Wikipedia.

3.3.1 Notation

Let $U = \{u_1, \ldots, u_n\}$ denote the selected users, where for an n_0 between one and n the set $D = \{u_1, \ldots, u_{n_0}\} \subseteq U$ contains exactly the dropouts. Let $u \in U$ be any selected Wikipedian. The random variable encoding u's dropout time is denoted by $T_u^{(drop)}$. The actual value of $T_u^{(drop)}$ is only observed if $u \in D$; in this case the observed dropout time of u is denoted by $t_u^{(drop)}$. Each user potentially starts (i. e., makes her first edit) at a different time point, denoted by $t_u^{(start)}$. By definition, selected users have performed at least a thousand edits; the time when u performed her thousandth edit is denoted by $t_u^{(1000)}$. Finally, the time point of data collection (i. e., October 8th, 2008) is denoted by $t^{(end)}$; it is equal for all users.

Turning to the explanatory variables, for a time point t let W_t denote the *history of Wikipedia* up to time t, i. e., information about every edit, discussion, voting, blocking (and so on) that took place on or before t. Later we let the risk of dropout at time t depend on W_t—more precisely, on particular *statistics* computed from W_t, see Sect. 3.4—and on nothing else. With $W = W_{t^{(end)}}$ we denote the history at the time of data collection, i. e., the entire data that we use to compute explanatory variables.

3.3.2 Survival, Hazard, and Probability Density

The methodology outlined in this section is not restricted to model dropouts; it is rather standard methodology for lifetime analysis in general, see [9].

As before, let $u \in U$ be any selected Wikipedian. The function

$$f_u(t) = \lim_{\Delta t \to 0} \frac{Pr\left(t \leq T_u^{(drop)} < t + \Delta t\right)}{\Delta t} \qquad (1)$$

is the probability density for u's dropout time being equal to t; f_u is defined on the real interval $[t_u^{(start)}, \infty[$.

At a first glance it seems that we could use f_u to test hypothetical causes of dropout by specifying f_u as parametrically dependent on covariates (encoding the potential causes) and testing whether those covariates show the predicted effect: covariates that empirically increase f_u (i. e., the risk to drop out) would then be interpreted as causes of dropout. However, this approach would not take into account an intrinsic dependency in lifetime data: an instance that dies at time t must necessarily survive up to this time point. To illustrate this on a simple example, assume that we were modeling the lifetime of humans. It is plausible that only a small percentage of people dies at the age of 100 years. However it would be wrong to conclude that people in their hundredth year are at a low risk of dying; the low percentage is rather due to the fact that very few people ever survive up to their hundredth year.

Returning to the case of Wikipedians but keeping the above example in mind, we see that we should rather model the *conditional* probability of users dropping out at time t, under the precondition that they survived up to t. This conditional probability density

$$h_u(t) = \lim_{\Delta t \to 0} \frac{Pr\left(t \leq T_u^{(drop)} < t + \Delta t \middle| t \leq T_u^{(drop)}\right)}{\Delta t}$$

is called the *hazard function* [9]; h_u is defined on the real interval $[t_u^{(start)}, \infty[$.

The hazard to drop out at time t is modeled as a function of various *statistics* $s_i(u; W_t)$, $i = 1, \ldots, k$ (characterizing certain aspects of the Wikipedia history at time t around user u) and parameters $\theta = (\theta_1, \ldots, \theta_k)$ that encode whether the respective statistics have a decreasing or increasing effect (or none) on the hazard to drop out. More precisely, we model the dropout rate in the following functional form:

$$h_u(t) = h_u(W_t; \theta) = \exp\left(\sum_{i=1}^{k} \theta_i \cdot s_i(u; W_t)\right) \qquad (2)$$

The estimated parameter values give information about the causes of dropout: if, for instance, a statistic $s_i(u; W_t)$ encodes how much u participates in discussion and if the associated parameter θ_i is significantly positive (negative), then participation in discussion is correlated with a higher (lower) probability to drop out. (Actually, it turns out that discussion is correlated with a *higher* probability to drop out, see Sect. 4.)

The general model outlined so far can be applied to test hypotheses about the interplay between characteristics of the Wikipedia system and the dropout hazard of Wikipedians. The model is specialized to test concrete hypotheses by plugging appropriate statistics into Eq. (2). The statistics that we take in this paper are defined in Sect. 3.4.

Equation (2) formalizes the assumption that the time dependence of the dropout hazard is completely captured in W_t. In other words, we assume that only endogenous factors are responsible for triggering dropout and, given the history of Wikipedia W_t, the hazard is conditionally independent of time.

While the hazard rate is convenient for parametric modeling, we nevertheless need for parameter inference (Sect. 3.5) the probability density f_u, see Eq. (1), and the *survivor function*

$$S_u(t) = Pr\left(t \leq T_u^{(drop)}\right);\ t \in [t_u^{(start)}, \infty[$$

(denoting the probability to survive as a Wikipedian beyond time t). However, specifying the hazard function h_u is sufficient since it determines both, the survivor function S_u and the probability density f_u by (cf. [9])

$$\begin{aligned} S_u(t) &= \exp\left(-\int_{t_u^{(start)}}^{t} h_u(x)\, dx\right) \text{ and} \\ f_u(t) &= h_u(t) \cdot \exp\left(-\int_{t_u^{(start)}}^{t} h_u(x)\, dx\right) . \end{aligned}$$

3.4 Explanatory Variables

In this section we define the concrete statistics that we take in this paper as the determinants of the dropout hazard, see Eq. (2). Each statistic corresponds to a hypothetical factor that might increase or decrease the hazard to drop out.

The estimation of the associated parameter (see Sects. 3.5 and 4) reveals whether such a hypothetical dependency can be empirically validated.

The statistics that we take in this paper are quite simple from a computational point of view. Other more involved statistics will be treated in future research (also see Sect. 5).

3.4.1 Editing, Discussing, and Organizing

The first family of statistics is constructed to answer the question: do users become more robust against dropout when they accumulate a growing number of contributions? A positive answer to this question would imply that users are more likely to drop out at the beginning of their career than at later stages. A negative answer would imply that users wear out and their dropout hazard increases with a growing number of contributions. However, since users can contribute to Wikipedia in different ways, we distinguish between three different kinds of contributions: (1) editing encyclopedic entries, (2) discussing, and (3) performing organizational work in Wikipedia.

To provide some background on this distinction, we recall that the set of Wikipedia pages is partitioned into various *namespaces* representing different types of pages (see the page `Wikipedia:Namespaces`). The *main namespace* comprises the set of encyclopedic articles. In the following, we denote contributions to the main namespace as *editing*. Besides the articles pages—whose creation is the main purpose of Wikipedia—there are pages which are concerned with various kinds of organizational work. These include pages in the namespaces `Wikipedia (Project)`, `Portal`, `User`, `File`, `MediaWiki`, `Template`, `Category`, `Help`, `Media`, and `Special`. In the following, we denote contributions to these namespaces as *organizing*. Finally, pages of all namespaces except `Media` and `Special`, but including the main namespace, have associated *talk pages* providing space for discussion. In the following, we denote all contributions to the talk pages as *discussing*.

Several researchers, including [19, 8], pointed out that discussion and organization work increased more rapidly over the last years than editing main articles. In this paper we analyze whether contributions to these three types of pages have different implications for the dropout hazard.

To define the statistics encoding how much a particular user u contributed to these three types of pages up to a timepoint t, let $E_{u,t}$ denote the set of revisions that u performed on pages of the main namespace on or before time t; let $T_{u,t}$ denote u's revisions to discussion pages on or before t; and let $O_{u,t}$ denote u's revisions to pages in all other namespaces (listed above) on or before time t. The respective statistics, to be used in Eq. (2), are defined by

$$\begin{aligned} \texttt{edit}(u; W_t) &= \log(1 + |E_{u,t}|) \\ \texttt{discuss}(u; W_t) &= \log(1 + |T_{u,t}|) \\ \texttt{organize}(u; W_t) &= \log(1 + |O_{u,t}|) \ . \end{aligned}$$

The logarithmic scaling of the number of revisions has been chosen due to the extremely skewed distribution (there are users who performed more than 100,000 revisions, while most of the selected users have a count of only slightly more then 1,000).

The interpretation of the associated parameters is as follows. A significantly positive (negative) parameter associated with `edit` implies that users with a higher number of revisions to the main namespace have a higher (lower) hazard to drop out. The interpretation for the parameters associated with `discuss` and `organize` is analogous.

3.4.2 Feedback

Another likely determinant of the dropout probability is the feedback that a user receives from others. Positive feedback is likely to have a motivating effect and, thus, might reduce the dropout hazard. On the other hand, negative feedback is likely to be frustrating and might increase the dropout hazard. Feedback can be provided to a user via her *user talk page* (see the page `Wikipedia:User talk page`). Since we want to rely in this paper only on automatic (and simple) methods, we do not evaluate whether feedback is positive or negative but only count the number of revisions made to the talk page of a particular user. Additionally we count how many contributions to the talk page of user u are made by u herself; thereby we can distinguish between users who reply to feedback given to them and users who do not (or less) reply.

More precisely, let $T_t^{(u)}$ denote the set of revisions to the user talk page of user u that are performed by any user on or before time t. Similarly, let $T_{u,t}^{(u)}$ denote the revisions made by u to her own user talk page on or before t. The respective statistics, to be used in Eq. (2), are defined by

$$\begin{aligned} \texttt{getFeedback}(u; W_t) &= \log(1 + |T_t^{(u)}|) \\ \texttt{replyFeedback}(u; W_t) &= \log(1 + |T_{u,t}^{(u)}|) \ . \end{aligned}$$

A significantly positive (negative) parameter associated with `getFeedback` implies that users with a higher number of revisions made to their user talk page have a higher (lower) hazard to drop out.

3.4.3 Controversy

Another reason for dropping out might be that Wikipedians are frustrated from ongoing controversies or edit wars with other users. To analyze this we look at how much a certain user edits *controversial* pages, i. e., pages mentioned on `Wikipedia:List of controversial issues`. Similar as above, let $C_{u,t}$ denote the set of revisions that a user u made to any controversial page on or before time t and define the respective statistic by

$$\texttt{editControversial}(u; W_t) = \log(1 + |C_{u,t}|) \ .$$

A significantly positive (negative) parameter associated with `editControversial` implies that users with a higher number of revisions made to controversial articles have a higher (lower) hazard to drop out.

3.5 Parameter Inference from Observations

This section provides details about how the parameters θ_i in Eq. (2) are computed from a set of observed dropout users and survivors. Readers not interested in this may directly continue with Sect. 4 (note that the parameters can be interpreted without knowledge of the estimation algorithm).

Let $U = \{u_1, \ldots, u_n\}$ denote the selected users, where for an n_0 the set $D = \{u_1, \ldots, u_{n_0}\} \subseteq U$ contains exactly the dropouts. Any observation of a $u \in U \setminus D$ (i. e., each survivor) gives us the information that u survived beyond time $t^{(end)}$. Since all selected users have at least thousand edits, the probability for surviving up to $t_u^{(1000)}$ is equal to

one. Thus, the probability for observing $u \in U \setminus D$ is

$$\begin{aligned}
&Pr\left(t^{(end)} \leq T_u^{(drop)} \middle| t_u^{(1000)} \leq T_u^{(drop)}; W; \theta\right) \\
&= \frac{S_u(t^{(end)}; W; \theta)}{S_u(t_u^{(1000)}; W; \theta)} \\
&= \frac{\exp\left(-\int_{t_u^{(start)}}^{t^{(end)}} h_u(W_x; \theta)\, dx\right)}{\exp\left(-\int_{t_u^{(start)}}^{t_u^{(1000)}} h_u(W_x; \theta)\, dx\right)} \\
&= \exp\left(-\int_{t_u^{(1000)}}^{t^{(end)}} h_u(W_x; \theta)\, dx\right) \\
&= \texttt{survive}_u(W, \theta)
\end{aligned}$$

For each $u \in D$ (i. e., for each dropout instance) we know that u dropped out at $t_u^{(drop)}$. As above, we have to correct for the fact that we selected only users with at least thousand edits. Thus, the probability density for observing $u \in D$ is

$$\begin{aligned}
&f_u(t_u^{(drop)} \big| t_u^{(1000)} \leq T_u^{(drop)}; W; \theta) \\
&= \frac{f_u(t_u^{(drop)}; W; \theta)}{S_u(t_u^{(1000)}; W; \theta)} \\
&= h_u(W_{t_u^{(drop)}}; \theta) \cdot \frac{S_u(t_u^{(drop)}; W; \theta)}{S_u(t_u^{(1000)}; W; \theta)} \\
&= h_u(W_{t_u^{(drop)}}; \theta) \cdot \exp\left(-\int_{t_u^{(1000)}}^{t_u^{(drop)}} h_u(x; W_x; \theta)\, dx\right) \\
&= \texttt{dropout}_u(W, \theta)
\end{aligned}$$

The joint probability density to observe the complete set of selected users U is

$$f(U, \theta) = \left(\prod_{i=1}^{n_0} \texttt{dropout}_{u_i}(W, \theta)\right) \cdot \left(\prod_{i=n_0+1}^{n} \texttt{survive}_{u_i}(W, \theta)\right)$$

(Here we assumed that dropouts are conditionally independent, given the history of Wikipedia W, i. e., we assume that W captures all the necessary information that determines dropout. For instance, an agreement between two users of the kind "I drop out, if you drop out" would violate this independence assumption; nevertheless, if two users drop out due to the same endogenous factor these dropout events are *conditionally* independent, although not independent.)

For a fixed observation U, we obtain a likelihood function L on the space of parameters $\Theta = \mathbb{R}^k$ by

$$L\colon \Theta \to \mathbb{R};\ \theta \mapsto f(U, \theta)$$

and we estimate those parameters $\hat{\theta} = \operatorname{argmax} L$ that maximize L (maximum likelihood principle, cf. [21]).

Computational simplification. We note that the state of Wikipedia W_t changes only when an edit is performed, i. e., only at finitely many time points (albeit a lot). Hence, if the statistics $s_i(u; W_t)$ have no explicit time-dependency, they are piecewise constant functions and the integrals in the equations above are equal to weighted sums (where the weights correspond to the lengths of the time intervals during which the state of Wikipedia remains unchanged). For practical and computational reasons we will simplify this further and approximate the state of Wikipedia in the sense that we let W_t change only once a day. Thus, the statistics $s_i(u; W_t)$ are constant for each day and the integrals reduce to a manageable number of summands.

Thus, from now on we assume that time is given by integer numbers denoting a counter for days. In particular, $\sum_{x=t_1}^{t_2} h_u(x; \theta)$ denotes the sum over $h_u(x; \theta)$, where the day counter x goes from t_1 to t_2.

Estimation algorithm. The maximum likelihood estimates of the parameters are computed by the established Newton-Raphson algorithm. First, we note that parameters $\hat{\theta}$ maximize L if and only if $\hat{\theta}$ maximize $\log L$; however, $\log L$ has a simpler functional form. It is

$$\begin{aligned}
&\log L(\theta) = \\
&\left(\sum_{i=1}^{n_0} \log \texttt{dropout}_{u_i}(W, \theta)\right) \\
&+ \left(\sum_{i=n_0+1}^{n} \log \texttt{survive}_{u_i}(W, \theta)\right) = \\
&\left(\sum_{i=1}^{n_0} \log h_{u_i}(W_{t_{u_i}^{(drop)}}; \theta) - \int_{t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} h_{u_i}(W_x; \theta)\, dx\right) \\
&+ \left(\sum_{i=n_0+1}^{n} -\int_{t_{u_i}^{(1000)}}^{t^{(end)}} h_{u_i}(W_x; \theta)\, dx\right) \quad ,
\end{aligned}$$

where $h_{u_i}(W_x; \theta) = \exp\left(\sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_x)\right)$. With the convention that we make changes to W_t only once a day (see above) we obtain

$$\begin{aligned}
&\log L(\theta) = \\
&\sum_{i=1}^{n_0} \sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_{t_{u_i}^{(drop)}}) \\
&- \sum_{i=1}^{n_0} \sum_{x=t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} \exp\left(\sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_x)\right) \\
&- \sum_{i=n_0+1}^{n} \sum_{x=t_{u_i}^{(1000)}}^{t^{(end)}} \exp\left(\sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_x)\right)
\end{aligned}$$

The first order partial derivative with respect to $\ell = 1, \ldots, k$ is

$$\begin{aligned}
&\frac{\partial}{\partial \theta_\ell} \log L(\theta) = \\
&\sum_{i=1}^{n_0} s_\ell(u_i; W_{t_{u_i}^{(drop)}}) \\
&- \sum_{i=1}^{n_0} \sum_{x=t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} s_\ell(u_i; W_x) \cdot h_{u_i}(W_x; \theta) \\
&- \sum_{i=n_0+1}^{n} \sum_{x=t_{u_i}^{(1000)}}^{t^{(end)}} s_\ell(u_i; W_x) \cdot h_{u_i}(W_x; \theta)
\end{aligned}$$

The second order partial derivative with respect to $\ell, \ell' = 1, \ldots, k$ is

$$\frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_\ell}\log L(\theta) =$$
$$-\sum_{i=1}^{n_0}\sum_{x=t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} s_\ell(u_i;W_x)\cdot s_{\ell'}(u_i;W_x)\cdot h_{u_i}(W_x;\theta)$$
$$-\sum_{i=n_0+1}^{n}\sum_{x=t_{u_i}^{(1000)}}^{t^{(end)}} s_\ell(u_i;W_x)\cdot s_{\ell'}(u_i;W_x)\cdot h_{u_i}(W_x;\theta)$$

Let

$$\nabla\log L(\theta) = \left(\frac{\partial}{\partial\theta_\ell}\log L(\theta)\right)_{\ell=1,\ldots,k}$$

denote the vector of first order derivatives and let

$$H(\theta) = \left[\frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_\ell}\log L(\theta)\right]_{\ell,\ell'=1,\ldots,k}$$

denote the matrix of second order derivatives. Start with initial parameter values $\theta^{(0)}$ and update for $i = 0, \ldots,$ *max-iter* by setting

$$\theta^{(i+1)} = \theta^{(i)} - \left(H(\theta^{(i)})\right)^{-1}\cdot\nabla\log L(\theta^{(i)}) \ ,$$

until $\nabla\log L(\theta^{(i)})$ is sufficiently close to zero. This $\theta^{(i)}$ is then a good approximation for the maximum likelihood estimate $\hat{\theta}$.

4. RESULTS AND DISCUSSION

We estimated the model outlined in Sect. 3.3 with the six explanatory statistics (editing, discussing, organizing, getting feedback, replying to feedback, and editing controversial articles, defined in Sect. 3.4) plus an additional constant parameter. The main information resulting from this analysis is whether the associated parameters are significantly positive (revealing a tendency for increased dropout hazard) or significantly negative (revealing a tendency for decreased dropout hazard). The constant just normalizes the model to the empirical time scale in which one unit corresponds to the expected time-to-dropout of a (hypothetical) user for which the effects of all other statistics add up to zero. The value of this constant does not provide much information; if we had started with another time unit (the time unit of our model is one day) we would have obtained another value as constant.

The estimated parameter values and estimated standard errors are reported in Table 1. The parameters are significantly different from zero at the 5%-level, if the resulting t-ratio (the absolute value of the parameter divided by the standard error) is at least 1.96, cf. [21]. All six parameters turned out to be significant at this level. The interpretation of the results is below.

The parameter associated with `edit` is negative, indicating that the dropout hazard of a user decreases with a growing number of edits to the main namespace (i. e., the set of encyclopedic articles). Thus, users are more likely to drop out early in their career and gain robustness against leaving Wikipedia while they perform more and more edits to article pages.

Table 1: Estimated parameters, standard errors (in brackets), and t-ratios. Parameters are significantly different from zero at the 5%-level if the t-ratio is at least 1.96. Significantly positive (negative) parameters indicate a higher (lower) hazard to drop out.

statistic	parameter (s.e.)	t-ratio
`edit`	-0.410 (0.061)	6.78
`discuss`	0.137 (0.068)	2.01
`organize`	0.220 (0.060)	3.69
`getFeedback`	0.365 (0.078)	4.66
`replyFeedback`	-0.140 (0.057)	2.44
`editControversial`	0.177 (0.036)	4.98
constant	-10.604 (0.405)	26.18

This is different for participation in discussion: the parameter associated with `discuss` is positive, indicating that users become more likely to drop out when they participated more in discussion pages. This dependency—which lead us, together with the result for the `edit` parameter, to the choice of our title—is not necessarily a causal relationship. It might be the case that users accumulate frustration due to some other unknown reason which, at the same time, has an increasing effect on the frequency of contributions to discussion. To get into the vicinity of causality it will be analyzed in future research whether different forms of discussion (e. g., un-replied threads vs. replied threads, or discussion patterns that resemble a flame war, see [17]) have different effects on the dropout hazard. Thereby we would gain insight into *how* Wikipedians should discuss such that reasons to drop out are attenuated. The participation on pages concerned with the organization of Wikipedia also has an increasing effect on the dropout hazard (positive value of the `organize` parameter).

Turning to the effects of feedback on user talk pages, we observe that if user u gets revisions on her own user talk page, then the dropout hazard of u increases (positive value of the `getFeedback` parameter); this effect is attenuated, if u herself participates to the discussion on her user talk page (negative value of the `replyFeedback` parameter). A possible explanation for the `getFeedback` parameter is that users might become involved into disputes which could result into the two effects that (1) they get complaints from other users on their user talk page and (2) they become more likely to drop out due to frustration. The negative value of the `replyFeedback` parameter indicates that users who respond to comments on their user talk page have a lower dropout hazard than users who do not respond—potentially being explained that the latter ones do not care anymore since they are already pondering about stop participating. Similar to the `discuss` statistics, it seems to be an important topic for future research to distinguish between positive feedback and negative feedback or, more generally, to find out how conversation on user talk pages should look like such that users are retained in Wikipedia.

The positive value of the `editControversial` parameter indicates that users editing controversial pages have a higher dropout hazard. This relationship seem to be very plausible since editing controversial pages involves confrontation with vandalism or edit wars, which might be a frustrating experience.

5. CONCLUSION AND FUTURE WORK

We presented a statistical framework to assess hypothetical causes for dropout from Wikipedia. The model defined in this paper is generally applicable to all kinds of data that may explain dropout in Wikipedia or other Web 2.0 applications—independent on whether the explanatory data stems from log files, questionary-based surveys, or other sources. The general model can be specialized to test specific hypotheses by plugging appropriate statistics into Eq. (2). The explanatory variables defined in Sect. 3.4 and used in Sect. 4 reflect a particular choice of hypothetical factors for dropout that will be extended in future research.

The most intriguing empirical result obtained in this paper is that participation in discussion seems to cause dropout rather than preventing it. Although several researchers have reported an increase in discussion in Wikipedia (e. g., [19, 8]), we are not aware of any previous quantitative work analyzing the effects of discussion. However, it is obvious that the results obtained in this paper are still very coarse, since only the number of contributions to talk pages has been counted and we did not distinguish between different forms of conversation. It is a promising topic for future research to relate various discussion patterns (see, e. g., [17]) to the dropout hazard, thereby revealing how frustrating discussion and how motivating discussion looks like. Furthermore, as we outlined in Sect. 3.2.1, our analysis is based on a specific selection of dropouts via the page of missing Wikipedians; it will be analyzed in future work whether alternative selection strategies lead to different and potentially more reliable results.

Another promising avenue for future research is to focus more on the effects of collaboration structure on the dropout hazard. We defined in [2] the *edit network* of Wikipedia pages encoding how users contribute to the page and how they interact with each other. It is very likely that certain patterns of users in these edit networks (e. g., getting deleted, getting restored, being a provider of novel content) or patterns of the global collaboration structure (e. g., bipolarity) influence the dropout hazard. If this can be validated we would identify collaboration patterns that are rewarding and motivating for Wikipedians and patterns that frequently lead to dropout and therefore to the loss of human capital.

Acknowledgments. We gratefully acknowledge financial support by the *Netherlands Organization for Scientific Research* (program *Networks of Networks*).

6. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. 16th Intl. Conf. WWW*, pages 261–270, 2007.

[2] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proc. 18th Intl. World Wide Web Conf. (WWW2009)*, 2009, to appear.

[3] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedias. *Information Visualization*, 7:34–48, 2008.

[4] The battle for Wikipedia's soul. *The Economist*, March 6th, 2008.

[5] R. Garud, S. Jain, and P. Tuertscher. Incomplete by design and designing for incompleteness. *Organization Studies*, 29(3):351–371, 2008.

[6] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

[7] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2007.

[8] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: conflict and coordination in Wikipedia. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 453–462, 2007.

[9] J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, 2nd edition, 2003.

[10] T. Lento, H. T. Welser, L. Gu, and M. Smith. The ties that blog: Examining the relationship between social ties and continued participation in the Wallop weblogging system. In *Proc. 3rd Annual Workshop on the Weblogging Ecosystem*, 2006.

[11] P. R. Monge and N. S. Contractor. *Theories of Communication Networks*. Oxford University Press, 2003.

[12] J. Y. Moon and L. S. Sproull. The role of feedback in managing the Internet-based volunteer work force. *Information Systems Research*, 19(4):494–515, 2008.

[13] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proc. Intl. ACM Conf. Supporting Group Work*, pages 259–268, 2007.

[14] S. A. Snell and J. W. Dean Jr. Integrated manufacturing and human resource management: A human capital perspective. *The Academy of Management Journal*, 35(3):467–504, 1992.

[15] K. Stein and C. Hess. Does it matter who contributes? A study on featured articles in the German Wikipedia. In *Proc. 18th ACM Conf. Hypertext and Hypermedia (Hypertext 2007)*, pages 171–174, 2007.

[16] B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *Proc. IEEE VAST*, pages 163–170, 2007.

[17] T. C. Turner, M. A. Smith, D. Fisher, and H. T. Welser. Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, 10(4), 2005.

[18] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 575–582, 2004.

[19] F. B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in Wikipedia. In *Proceedings HICSS*, 2007.

[20] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *Proc. Intl. Conf. Web Search and Web Data Mining*, pages 171–182, 2008.

[21] G. A. Young and R. L. Smith. *Essentials of Statistical Inference*. Cambridge University Press, 2005.