Realizing the Activation Potential of Online Communities

Completed Research Paper

Marios Kokkodis
Boston College
kokkodis@bc.edu

Theodoros Lappas
Stevens Institute of Technology
tlappas@stevens.edu

Abstract

Online communities suffer from the 1-9-90 principle, which states that 1% of the community's user base generates original content, an additional 9% is limited to interacting with existing content, while the remaining 90% of the participants is passively lurking. In this work we present a data-driven stochastic framework that estimates (1) the activation potential (i.e., the users that are currently lurkers but present a high likelihood of becoming heavy contributors) of an online community and (2) when and which users are more likely to become heavy contributors. Our proposed framework captures the transitional evolution of a user by a Hidden Markov Model, and estimates each user's propensity to become a heavy contributor by employing parametric survival models. We build and evaluate our models on a unique large dataset of a specialized online community about diabetes.

Keywords: Empirical analysis, Survival analysis, Data mining, HMM

Introduction

In the past fifteen years online communities have experienced an accelerated growth on par with the Internet boom. Online forums on an ever-increasing variety of topics are now part of the daily routine of millions of online users. The fundamental premise of these communities is their reliance on voluntary content generation. After being introduced to the community by one or more users, a quality piece of content can be shared, evaluated, discussed, and annotated, nurturing the user interaction and engagement that are necessary ingredients for the prosperity of the platform.

Despite the well-documented importance of this process, modern platforms typically fail to motivate the vast majority of their users to generate content. This phenomenon is best captured by the well-documented 1-9-90 principle(van Mierlo 2014)[1], which states that 1% of an online community's user base generates original content, an additional 9% is limited to interacting with existing content introduced by others, while the remaining 90% of the participants is passively lurking. This staggering imbalance is a testament to the tremendous unexploited growth potential that online communities have yet to realize.

A long line of relevant work has focused on understanding different aspects of user participation in online communities, such as collaboration(Lappas et al. 2009; Ransbotham and Kane 2011) and the forging of user relationships(Shi et al. 2009). Other work has studied the features of user generated content (UGC) that tend to attract attention(Weiss et al. 2008) and even found evidence that users are motivated to contribute more and better content when they receive feedback on previous contributions (Arguello et al. 2006; Moon and Sproull 2008).

More relevant to our study, researchers have tried to understand patterns of user participation, coming up with the well established and accepted by the IS community “reader-to-leader" framework(Lave and Wenger 1991; Preece and Shneiderman 2009). Even though these works create a strong theoretical framework regarding user evolution in these communities, they fail to provide concrete evidence on which users and when are more likely to transition from one state (e.g., reader) to another (e.g., leader).

In this work we present a data-driven stochastic framework that answers both the “which" and the “when” questions, but also provides an estimate of the activation potential (i.e., the users that are currently lurkers but present a high likelihood of becoming heavy contributors) of an online community. In particular, we start by acknowledging that an online community dynamically evolves through time. Given the current state of a community, we assume that the behavior of a user is also dynamic: a user might switch from lurker to heavy contributor to intermittent contributor etc. The actual state of the user is unobserved; Instead, we observed sporadic actions for each user (i.e., a new topic creation or a response). Based on these observable actions, and in order to capture the transitional evolution of the user, we propose to build a Hidden Markov Model (HMM) that at each point in time provides a stochastic estimate of the latent state of the user. Once we know the state of each user, we build parametric survival models that estimate the likelihood of each user to transition from a non-contributor state to a heavy contributor state. These survival models identify (1) which users are more likely to make this transition, (2) when they are more likely to do so, and (3) they provide an estimate of the activation potential of an online community at a given time.

We build and evaluate our models on a unique dataset of a specialized online community about diabetes (tudiabetes.org). Specifically, our dataset includes all the actions taken by all the users throughout the eight years that this online community exists. This allows us to draw a catholic picture about both the community and its users' behavior through time. Our findings for this specific community indicate that there are a series of user/community characteristics (e.g., number of users responses, number of created topics, first day of action) that increase the likelihood of transitioning to a heavy-contributor state. On the

[1] https://goo.gl/MRaEcp

[2] We discuss in detail the new features that we used in our models in the “Experimental Setting and Results" section.

other hand, other characteristics appear to decrease this likelihood (e.g., the total number of days between consecutive actions and whether or not the user responded or created a topic within a week after the user joined the platform). Finally, we showcase how the activation potential of tudiabetes decreases over time, even though the total number of new users per year remains constant, indicating that the true activation potential of an online community is not proportional to the community's annual growth.

Our work contributes to the current literature in online communities by (1) introducing the notion of activation potential and (2) by identifying when and which users are more likely to transition to the heavy contributor state. From a methodological perspective, we draw on advanced machine learning and survival analysis to present a novel, completely dynamic framework that captures user behavior. Even more, our work has two major implications for online community platforms. First, by following our proposed framework an online community can realize its true activation potential, which is a new evaluation metric for the community welfare. Second, the platform can identify (and potentially target) which users and when are more likely to transition to a heavy-contributor state.

It is important to note that our work does not study how policy changes (e.g., the introduction of a new feature) might increase participation. As a result, our survival analysis results should not be given a causal interpretation. This does not limit the contribution of our work. Our main goal is to identify the activation potential of a community and study how the community-user interactions appear to be correlated with the user's likelihood of becoming a heavy contributor. These objectives are met by the estimated likelihoods that represent the current state of the community.

Background and Related Work

Our study focuses on understanding the activation potential of an online community. Previous research has dealt with a series of dimensions of user-generated content in online communities. In this section, we cluster these works into studies that focus on (1) why users participate in online communities, (2) identifying patterns of user participation, and (3) evaluating different incentive mechanisms that affect user engagement. We then present for completeness other related work in online communities, and we conclude by clearly pinpointing the contributions of our study.

Why users participate in Online Communities

In the previous years researchers have repeatedly studied the question of why users contribute in online communities from a variety of vantage points. Their findings are very diverse, and they are usually associated with the type and context of the community studied as well as with the employed methodologies. In particular, users have been found to engage in content creation because (1) it enhances their professional reputation (Wasko and Faraj 2005), (2) they have psychological bonds with the community (Bateman et al. 2011), (3) of social ties (Bagozzi and Dholakia 2006; Zeng and Wei 2013), and (4) of cognitive, emotional and social characteristics (Bagozzi and Dholakia 2006; Tsai and Bagozzi 2014).

Our work does not focus on understanding why users join and participate in an online community. Instead, we assume that for all the reasons described above (and for other unobserved causes) users have an intrinsic motivation and interest to join a particular community. Our analysis focus on how the online community evolves conditional on the fact that users who join have an interest in the community. Hence, in this work we study the users' conditional likelihood distribution of engaging in the content generation process.

Patterns of participation

On a different direction, researchers have also investigated participation patterns in a communal setting, both offline and online. In their work on learning processes in communities of practice, (Lave and Wenger 1991) proposed a description of community behavior over time, which has been adopted by the majority of research in online communities. In particular, they proposed that users “become more competent as they become more involved in the main processes of the particular community. They move from legitimate peripheral participation to full participation". Their approach has been further developed in recent years in (1) visitor, novice, regular, and leader differentiation (Kim 2000) (2) a reader to leader framework with emphasis on different needs and values at different levels of participation (Preece and Shneiderman 2009) (3) social technographics profiling (Li and Bernoff 2011) and (4) confidence-based contribution (Brzozowski et al. 2009).

Other researchers slightly diverged from the ladder-based approach. For example, (Lanamäki et al. 2015) proposed that participation occurs with uncertainty, involving trial and error, unknown risks and rewards while (Ray et al. 2014) defined engagement and showed that it explains both knowledge contribution and word of mouth. Furthermore, (Ren et al. 2012) found that member attachment is strengthened by group identity and interpersonal bonds.

Our work builds on the information from these past studies and provides a novel sequential model that accurately describes the dynamic behavior of users across different levels of participation. Extending the “reader to leader" framework, at any given time, our approach provides a stochastic estimate of each user to be in any of the available states/levels of participation.

In a parallel stream of work regarding user participation, researchers have focused on explaining how/why leaders (heavy contributors) emerge in these communities. Heavy contributors constitute the most valuable demographic of an online community, since they generate the vast majority of the website's content (Cassell et al. 2006; Yoo and Alavi 2004). As a result, understanding why and how some users become leaders is very important for the community. In online reviewing platforms, (Lu et al. 2013) observed that preferential attachment in a network of users and the number and quality of reviews written are important drives for a leader to emerge. In addition, leaders have been found to (1) use multiple discourse channels to broadcast their messages (Forte and Bruckman 2005), (2) produce more reviews and more objective reviews (Goes et al. 2014), and (3) produce positive feedback that generates local network effects in content generation (Shriver et al. 2013).

Our paper builds on these ideas on heavy contributors in three ways: first, we are interested in answering when users are more likely to become heavy-contributors. Second, we are able to identify which users are more likely to become heavy contributors. Finally, we delve into understanding how the user-community characteristics at a given time correlate with the probability of a user to become a heavy contributor.

Incentives for increased participation

Because of the importance of generating content in online platforms, a stream of work has focused on understanding which and how different community characteristics increase content creation. There is an overall consensus from previous works that providing feedback to a user who creates some content increases participation. In particular, community feedback has been found to increase (1) a newcomer's probability of returning to a site (Lampe and Johnston 2005), (2) the quantity and quality of content that a user subsequently uploads (Arguello et al. 2006; Burke et al. 2009; Moon and Sproull 2008), and (3) the duration of participation (Moon and Sproull 2008).

Researchers have pushed further towards identifying other community characteristics that increase content generation. Specifically, they found that (1) users whose posts attract more attention subsequently contribute more (Huberman et al. 2009), (2) the properties of the replies that a user receives affect the user's likelihood of re-engagement (Joyce and Kraut 2006), and (3) the type of discussion affects engagement--information-only discussions lead to less participation (Ozturk and Nickerson 2015). Part of our work focuses on understanding how community characteristics correlate with observed engagement. Some of the characteristics we use have been studied before (e.g., responses to a post), while others are new (e.g., the number of responses before and after a response[2]. As a result, we extend the discussion on how the characteristics of a community, without any exogenous shocks (e.g., introduction of a new feedback mechanism) at a given time correlate with the likelihood of engagement.

[2] We discuss in detail the new features that we used in our models in the “Experimental Setting and Results" section.

Table 1: Our work is the only one to study when is more likely for a user to become a heavy-contributor. Furthermore, it studies how different characteristics of an online community are correlated with increased engagement. Finally, it introduces a new methodology, which has never been used in the past for describing the dynamic behavior of users in online platforms.

Paper	Data	Incentives	When	Community	Objective	Methodology
Lu et al. 2013	E-pinions (reviews)	No	No	Yes	Leader emergence	Stochastic Net. Growth
Wasko and Faraj 2005	Network of Practice	No	No	No	Why users contribute	PLS
Bateman et al. 2011	Q&A (Survey,n=324)	No	No	No	Why users contribute	PLS
Zeng and Wei 2013	Flickr	No	No	No	Social ties and content	Regression
Tsai and Bagozzi 2014	VC (Survey, n=982)	No	No	No	Why users contribute	SEM
Ray et al. 2014	Q&A, WoM (Survey, n=301)	No	Yes	No	Understand Engagement	SEM, EFA
Oestreicher-Singer and Zalmanson (2013)	Last FM	No	No	No	Willingness to Pay and participation.	Logit, PSM Cox Model
Preece and Shneiderman (2009)	No Data	No	No	No	Understand participation	NA
Bagozzi and Dholakia (2006)	Linux User Groups (Survey, n=401)	No	No	No	Understand participation	SEM
Moon and Sproull 2008	Q&A	Yes	No	Yes	Feedback and UGC	Cox Model
Burtch et al. 2015	Online Reviews	Yes	No	No	Social & Monetary incentives	Randomized Experiment (RA)
Ren et al. 2012	MovieLens	Yes	No	Yes	Member attachment	RA
Goes et al. 2014	Epinions (reviews)	No	No	No	User Popularity & reviews	Panel, Matching
Shriver et al. 2013	Soulrider.com	No	No	No	Social ties & UGC	Regression, IV
Ghose and Han 2011	Mobile UGC	No	No	No	Content usage/generation	SE Panel
Our Work	**tudiabetes.org**	**No**	**Yes**	**Yes**	**Activation potential**	**HMM, AFT**

Previous studies have also looked at the effect of the introduction of new features (as exogenous shocks) in online communities on user engagement. In particular, ideas that have been tested through experiments and have been found to increase user participation are (1) adaptive rewards incentive mechanisms, such as status change and privileges based on user engagement (Cheng and Vassileva 2006), (2) badges (Anderson et al. 2013), (3) social and monetary incentives (Burtch et al. 2015), (4) social comparison of community users (Chen et al. 2010), and (5) imposing additional work (Drenner et al. 2008). Our work does not study how external shocks increase participation, or even the question of what the community should change in order to attract (create) more (new) heavy contributors. In other words, we are not interested in policy changes, but instead, we want to understand the activation potential of a community along with how the community characteristics at a given state-time correlate with user engagement.

Other research in online communities and user participation

Other related work in online communities focused on the relationship between participation and willingness to pay (Oestreicher-Singer and Zalmanson 2012), the type of posting and readership (Huang et al. 2015), the temporal relationship between content generation and content usage (Ghose and Han 2011) as well as the type of content that attracts attention (Weiss et al. 2008). Furthermore, researchers have also worked on understanding and increasing collaboration in these communities (Lappas et al. 2009; Ransbotham and Kane 2011; Ransbotham et al. 2012).

Finally, (Luca 2015) studied the causal impact of user generated content on economic and social outcomes, while (Chaturvedi et al. 2011; Kohler et al. 2011) proposed frameworks for designing virtual worlds. We mention these works here for completeness, since they are not closely related with the focus of this paper.

Contribution of our work

Our work extends the current literature in online communities by contributing mainly in four dimensions. First, we introduce the concept of the activation potential of an online community, given the community's current state. Second, we are the first to focus on when and which users present high likelihood of becoming heavy contributors. Third, we extend the study on community features that have been found in the past to be correlated with increased likelihood of engagement. Finally, from a methodological perspective, we are the first to present a complete dynamic framework that captures the evolution of a user (HMM), as well as the first to employ an appropriate parametric survival analysis on the users likelihood of becoming heavy contributors. **Table 1** clarifies the differences between our work and a series of related works described in this section.

Modeling the User's Dynamic Evolution

In this section we define the underlying model that describes both the community's and the user's dynamic evolution through time. We then formulate the user evolution through an appropriate Hidden Markov Model. Finally, we describe our parametric survival formulation that estimates the likelihood of each user to transition to a heavy contributor state.

The online community as a funnel

Online communities can be seen as dynamic organisms that evolve with time. For example, the demographics of a community constantly change: new users join the community, current users disable their accounts, lurkers become contributors and vice a versa, etc. Even further, communities implement (or retract) policy changes (i.e., introduction of a new feedback mechanism or a design upgrade) that also alter the dynamics and interactions among the community members. At any given point in time, an online community has activation potential. This potential is a function of the community's new and old users, and their likelihood of generating content. Given that a community's goal is to maximize the content generation (or user engagement), we are interested in estimating (1) which users and when are more likely to become contributors and (2) how the community-user interactions correlate with the likelihood of new users to create new content.

We can imagine an online community as a funnel: All users enter the funnel when they join the community. The vast majority of these users become lurkers; hence they remain inside the funnel. Some users generate content and as a result they exit the funnel. The funnel size, input and output streams are dynamic. We show this funnel visualization at different random points in time, $t_1, \ldots, t_n$ in **Figure 1**. Based on this diagram, we refine the problem we study as follows:

Problem Definition: *Given the state of the online community at some time t_k we focus on understanding (1) when and which users are more likely to exit the funnel, (2) what characteristics of the community correlate with the users' likelihood of exiting the funnel and (3) what is the community's activation potential.*

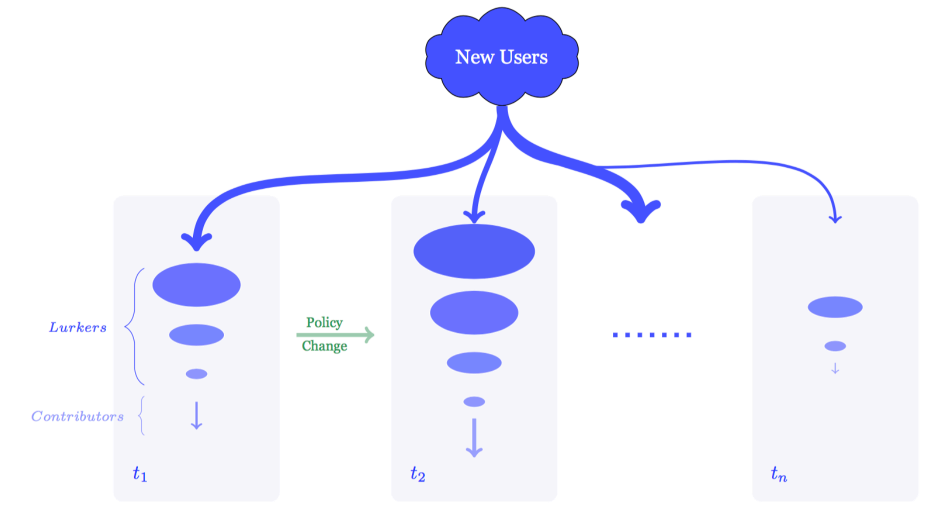


Figure 1: We visualize an online community as a funnel. The size of the funnel represents the total number of registered users. The exit arrow widths represent the number of contributors. The entry arrow widths represent the flow of new users. At discrete points in time (e.g., t1, t2, ..., tn) the community undergoes exogenous shocks (e.g., a policy change). These shocks transition the community to a new state.

Activation Potential: We define as “activation potential" the stochastic estimate of the number of users that are likely to become heavy contributors. Assuming that each user i has a probability p_i of becoming a heavy contributor, then a sequence of users follows a Poisson binomial distribution. Hence, the activation potential is the expected number of successes of this distribution:

$$Activation\ Potential = \sum_i p_i$$

The underlying model of a user's evolution

Based on our earlier discussion, and in order to answer our main questions for the rest of this section we assume that the state of a community is fixed at some time interval t_k. During this timespan, we assume that there are no exogenous shocks that affect the community behavior. As a result, we can focus now on modeling the behavior of the users given the characteristics of the community.

Our dataset (presented in the next section) contains complete information of all the observable actions that community users have taken since their joining date. The underlying model for a new user who joins a community works as follows:

- A user decides to join the online community. We assume that a user joins a community to (1) ask a question, (2) answer a question or (3) find information that already exists in the platform. The user's intrinsic motives are unobserved.
- The user reveals some of her initial intrinsic motives during the first week. A user might decide to ask a new question, respond to a current thread, or lurk.
- Every user evolves dynamically: the objectives of each user change over time. Users who joined the community to respond to a question might lose interest, or they might start answering other questions. Users who joined the community to learn and collect information about a topic might start answering questions, etc.
- At each point in time the internal state of a user is unobserved. What we observe is the user's actions, i.e., a new post, a new response, or the absence of any action.

User evolution as a Hidden Markov Model

Based on this underlying model we naturally propose to model the user behavior as a Hidden Markov Model (HMM). HMMs assume that users transition across a series of underlying latent (unobserved) states. Simply put, HMMs capture the unobserved dynamic behavior of a user through time.

To build our HMM, we first assume a set of unobserved user states ($\boldsymbol{S}$). In particular, we define $\boldsymbol{S}$ = {lurker, intermittent contributor, and heavy contributor}. These underlying states represent different probability distributions across a set of observable actions. The observable actions in our case are (1) a user creates a new topic (***C***), (2) a user responds to a question (***R***) and (3) a user lurks (***L***). Each day, a user emits one observation in the set ***X*** = {**C**,**R**,**L**}.

Let a sequence of observations of a user at a given time t be $X_{\{1:t\}}$. We estimate the most likely path for each user (i.e., $S_{\{\boldsymbol{1}:\boldsymbol{t}\}}$ $S \in \boldsymbol{S}$) by using the Viterbi algorithm (Viterbi 1967).

The algorithm takes as input some prior probability distribution across all the available hidden states, a transition matrix among the available states, and some local evidence, which is a vector of the conditional probability distributions over the emitted observations. Based on these input distributions, the algorithm efficiently searches through all the possible transitions in the state-time continuum and returns the path that maximizes the joined likelihood $\Pr\left(X_{\{1:t\}}, S_{\{1:t\}}\right)$ (Bishop 2006). In our scenario, the Viterbi algorithm is an efficient way to estimate the most probable sequence of latent states for each user in our dataset.

Up to this point we assumed that the input parameters for the Viterbi algorithm we have just described are known. In practice we estimate these parameters by using a version of the Baum-Welch algorithm (Baum et al. 1970).

This algorithm is a variant of Expectation Maximization (EM): It assumes that we have a total of N users, each of which has a lifespan of T_i days ($i \in \{1, .., N\}$). In addition it considers $K = |\boldsymbol{S}|$ different unobserved states ($\boldsymbol{S} = \{S_1, S_2, \dots, S_K\}$). At the E-step the algorithm estimates the expected log-likelihood of some previously estimated parameter vector $\theta^{\{old\}}$. At the M-Step the algorithm finds the new parameters that maximize the estimated expected log-likelihood. Note that the parameter vector $\theta = [\Pr(S_i), A(j,k), \Pr(X_t | S_t)]'$, where $\Pr(S_i)$ is the initial probability for a user to be in state S_i, $A(j,k)$ is a matrix that stores the transitional probabilities from state S_j to state $S_k$and $\Pr(X_t | S_t)$ is the emission probability of state S_t to emit the observation X_t, where $X_t \in \boldsymbol{X}$. The algorithm keeps iterating until the parameter vector θ converges.

The HMM learns through the users' previous actions and allows them to move across latent states over time. Once learnt, the HMM provides daily estimates regarding the probability that (1) a user will respond to a topic, (2) a user will create a new topic, and (3) a user will lurk.

However the HMM does not consider how community-user characteristics interplay with the transitional probabilities from one state to another. In order to study how these characteristics correlate with the transitions to a new state, but also in order to estimate when such a transition is more likely to occur we employ a survival analysis on top of the learnt HMM.

Survival analysis on the HMM latent states

Survival models associate the time before some event occurs to a set of covariates that might be correlated with both the event and the time lapsed. These models are typically used to answer questions such as "which portion of a given population will survive past a certain time?" Because of the nature of the problems that survival analysis has been applied to, the actual occurrence of the expected event is referred to as death. In our setting, death corresponds to the first time that a user makes a transition from states lurker, intermittent contributor to heavy contributor[3].

The most commonly used survival model in the literature is the Cox model of proportional hazards (Cox and Oakes 1984). The Cox model assumes that the effect of a unit increase in a covariate is multiplicative with respect to the hazard function[4]. By testing for this proportionally (Grambsch and Therneau 1994) we found that our set of variables violate this assumption, suggesting that the Cox model is not appropriate in our setting.

For the purposes of our study we employ an Accelerated Failure Time (AFT) model (Cleves 2008). AFT models are parametric, and the survival probability is assumed to follow a given distribution. In particular, let t_j be the actual time, then:

$$\tau_i = \exp(-\boldsymbol{\beta X_i})\, t_i \quad (1)$$

In the previous equation τ_i is a random quality that it is assumed to follow an underlying distribution. We can re-write this in the following form:

$$t_i = \exp(\boldsymbol{\beta X_i})\, \tau_i$$

It's easy to notice that the quantity $\exp(\boldsymbol{\beta X_i})$ is an accelerating factor: If this factor is positive then time moves slower than the baseline τ_i (positive $\boldsymbol{\beta}$ decelerates) and as a result failure occurs later. Otherwise, time moves faster than the baseline (negative $\boldsymbol{\beta}$ accelerates), and failure occurs sooner. We can further choose to take the logs of the previous equation and transform it into the linear relationship:

$$\log(t_i) = \boldsymbol{\beta X_i} + \log(\tau_i) \quad (2)$$

The choice of an underlying model comes down to our belief regarding the distribution of τ_i. We discuss the specific distribution we chose for our scenario in the next section.

Experimental Setting and Results

In this section we describe the dataset (tudiabetes.org) and the experimental setting we have used to build and evaluate our models. We conclude with a discussion regarding the evolution of tudiabetes.org through the years and its activation potential.

[3] In this work we only present an analysis of transitions to the heavy contributor state, however we can easily extend this analysis to any type of transition--we further discuss this in our last section.

[4] The hazard function is defined as $\lambda(t) = -\frac{S'(t)}{S(t)}$, where $S(t)$ is the survivor function which denotes the probability of an instance surviving after some time t, and S'(t) is its density function.

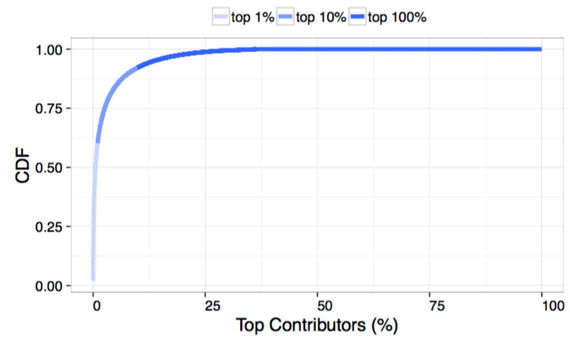


Figure 2: The 1-9-90 rule in tudiabets.org.

Data

For our analysis we use real data from a major online forum about diabetes (tudiabetes.com). In particular, we analyze a total of 38,791 users who joined the forum between June 2007 and October 2015. These users generated a total of 45,181 topics with 409,973 responses. Tudiabetes is an ideal online community for the purpose of our study because (1) it has been around for enough time (8 full years) to go through multiple phases, (2) its relatively small size gives us the opportunity to analyze the complete set of interactions (i.e., all actions of all users throughout the years) and (3) the highly specialized topic meets our earlier assumption that every user who joined the platform had an intrinsic interest in the community's subject (diabetes). In **Figure 2** we show the cumulative distribution of the users' contributions on tudiabetes.org. This distribution verifies that the 1-9-90 rule applies in our dataset.

Learning the HMM

Our first step in implementing our approach is to learn the HMM that best describes the behavior of our community. As we mentioned earlier, the parameters of the HMM are learned through the Baum-Welch algorithm. However, the convergence of this expectation maximization algorithm depends on the initial input distributions that we provide. In order to select the best HMM model for our data, we perform an exhaustive grid search that learns different HMMs for a total of 100,000 initial input distributions. We then choose the model that yields the lowest AIC scores (Akaike 1974). Note that this approach has been proposed by multiple outlets, including (Bishop 2006; Koller and Friedman 2009; Murphy 2012).

The resulting learned HMM[5] is presented in **Figure 3**. To be consistent with the standard representation of probabilistic graphical models, the latent states are transparent (i.e., lurker, intermittent contributor and heavy contributor) while the observable actions are shaded. Each latent state presents a different probability distribution across the three emitted observations ($\boldsymbol{X}$={$\mathbf{C}$,$\boldsymbol{R}$,$\boldsymbol{L}$}). For example, a user in the state heavy contributor has 68% probability to lurk, 6% probability to create a new topic and 26% probability to post a new response. These emission probabilities might appear very low at first. However, note that our HMM is built for daily observations, and that a user's observable actions follow a binomial distribution. As a result, a user that is in the heavy contributor state, in a span of 30 days will create on expectation 30 * 0.06 = 1.8 new topics, and will post 30 * 0.26 = 7.8 responses.

Beyond the emission probabilities our HMM presents both the estimated initial probabilities for each new user to land in any of the three latent states $\Pr(S_i)$ as well as the transition probabilities between states

[5] We use the complete dataset here to build a global HMM that captures the overall behavior of the community users throughout the eight years that the platform has been operating.

$A(j,k)$. By looking at these values we observe that (1) with 67% probability a new user will be a lurker (2) the next most probable state to land is heavy contributor (23%) and (3) that once you are in one state, chances are that you will remain in that state for a sequence of days (i.e., probabilities to remain in the same state range between 92% and 99.9%).

Survival Analysis

Now that we have learnt the HMM that best describes the users' dynamic behavior, we are ready to deploy our survival analysis. As we discussed before, we will follow a parametric (AFT) approach. To reiterate, a death in our study is defined as the first time that a user transitions from any other state (lurker or intermittent contributor) to the heavy contributor state.

Choosing the underlying distribution

Our first step is to determine which parametric family fits best our data. Recall that essentially we are looking for the distribution that τ_i follows, as presented in Equation (1).

The empirical distribution of our dataset (including right-censored instances) is shown in **Figure 4**. This distribution seems to have a lognormal hue: the likelihood of death increases in the beginning and up to a point, and then starts to decrease.

In order to be statistically meticulous about our underlying distribution choice, we estimate the AIC (Akaike 1974) scores for four different distributions: the “exponential", “lognormal", “loglogistic" and "“weibull". Note that this is a standard procedure when choosing among different parametric models (Cleves 2008). The results are shown in Table 2**.** The AIC scores verify our initial intuition, showing that the best choice (from a statistical perspective) is the lognormal distribution (lowest AIC score).

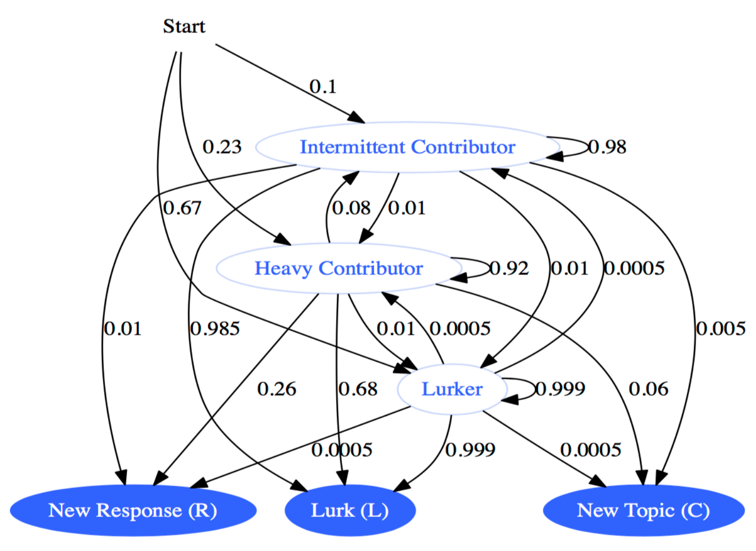


Figure 3: The HMM that best explains the behavior of our users.

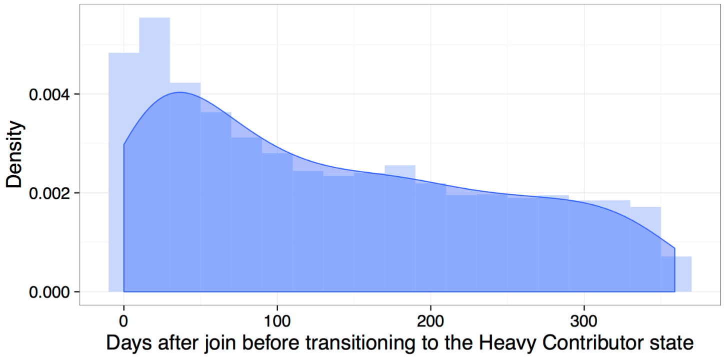


Figure 4: The empirical distribution of "deaths" in our dataset.

Table 2: Comparison of Different Parametric Models

Distribution	AIC
Weibull	916
Lognormal	911
Exponential	924
Loglogistic	915

The Log-Normal specification: In the log-normal case, we assume that τ_i follows a log-normal distribution, i.e., $\tau_i \sim logN(\beta_o, \sigma^2)$ and as a result $\log(\tau_i) \sim N(\beta_o, \sigma^2)$. This allows us to further re-write Equation (2) as:

$$\log(t_i) = \beta_0 + \boldsymbol{\beta}\mathbf{X_i} + u_i$$

The previous transformation converts the problem into a linear regression problem where the error u_i is distributed normally with mean 0 and standard deviation σ.

The Covariates Vector X

To create our set of covariates we draw on previous works on user engagement in online communities. In particular, we consider a set of eleven covariates. As we mentioned earlier, we assume that every user joins the community with different objectives. Each user's observable actions unveil the user's intentions, and as a result might accelerate or decelerate the user's survival probability estimate. In this direction, and in sync with previous literature (Huffaker 2010) we include in our covariate set (1) the number of times that a user responds and (2) the number of topics that a user creates. These two attributes represent the basic actions that a user takes, and intuitively both should accelerate the user's death (transition to the heavy contributor state).

Researchers in the past have found that being active for a greater number of weeks correlates positively with the post's quality, while being more intermittent presents a weak but negative correlation (Nam et al. 2009). To control for this expected effect of different activity patterns, we further consider in our covariate set the (3) average number of days between consecutive actions and (4) the standard deviation between consecutive actions.

Previous research has also highlighted the importance of early engagement for community users (Arguello et al. 2006; Burke et al. 2009). In this direction, and to control for users who join with an objective to ask something or respond to a single post, we further include in our set of covariates the (5) first day of action (new topic creation or new response), (6) whether or not the user has created a topic in the first week after joining the platform and (7) whether or not the user has responded to a thread in the first week after joining the platform. Intuitively, the greater the first day of action, the higher is the chance for a user to become a heavy contributor since it shows that the user did not join the platform with a given objective, but instead, the user built up confidence and now feels comfortable enough to participate. Similarly, if the dummy variables (6) and (7) are true then there is a high likelihood that these users have joined the platform with a single objective (ask a question or reply to a question), and hence these variables should decelerate the death of a user (i.e., delay the transition to the heavy contributor state).

Furthermore, researchers in the past have also established that users who have broader interests play a key role in information dissemination within the community (Hecking et al. 2015). In sync with this work, we further control for the (8) average number of responses per topic and (9) the entropy of the user across different topics. These two capture whether or not the user is focused into a specific set of threads/topics or whether the user is interested in a broader sense in diabetes. Next, and since one of our goals is to correlate community-user interactions with the likelihood of a user to become a heavy contributor, we include the (10) number of responses in a thread before a user's response and (11) the number of responses in a thread after a user's response.

For our survival analysis we use time-varying covariates. In particular, we assume that a new measurement is taken every time a user takes an action (i.e., responds to a post or creates a new topic). Most of our instances never die (i.e., they never become heavy contributors) and as a result, our datasets consist mostly of right-censored observations. Finally, we standardize our non-binary variables to facilitate faster convergence time of our models but also to make the interpretation of our resulting coefficients more straightforward.

Table 3: The distribution of our users at risk and heavy contributors throughout the eight years we are analyzing.

Year	New Users (at risk)	Heavy Contributors (died)
1	1864	67
2	4172	114
3	5090	174
4	4252	168
5	3459	100
6	4038	90
7	5485	100
8	3462	51

Results

In this work we are arguing that an online community dynamically evolves through time. Because we are not aware of any exogenous shocks/policy changes that our community has undergone through the years, we showcase our approach by splitting our data into 8 annual slices. This way we are able to study different instances of this community even though these instances do not correspond to specific changes. ***Table 3*** shows the statistics of each year. The first column shows the number of new users that joined the community during that year, while the second column shows the number of users that became heavy contributors. The difference between the two columns captures the right-censored instances in our survival analysis.

For each of these eight years we build a different lognormal AFT model. The resulting coefficients (along with their p-values) are shown in **Figure 5**. The first observation is that most of the coefficients are sign-consistent throughout the years. For example, (1), (3) and (6) present always the same sign. Furthermore (2), (4), (5), (7),(11) are also consistent in the majority of the eight years (especially if we only focus on the significant coefficients). (8) and (9) are always insignificant, while (10) changes from negative and significant in the fourth year to positive and significant in the 7th year. These observations show that overall our community follows consistent trends throughout the years. On the other hand, the year-to-year fluctuations capture the dynamically evolving behavior of the community.

The y-axis of **Figure 5** shows the value of the coefficients, while the x-axis shows the years. The color of the bars represents the coefficient's significance (the darker the shade, the more significant). The interpretation of these coefficients is as follows. Take the coefficient of the first year of "(3) Days Between Actions". The value of this coefficient is around 1.4. As a result, we can say that an one standard deviation increase of the "(3) Days Between Actions" in year one will decelerate the death of a user by a factor of exp(1.4) = 4.05 (i.e., will decelerate by ~305%).

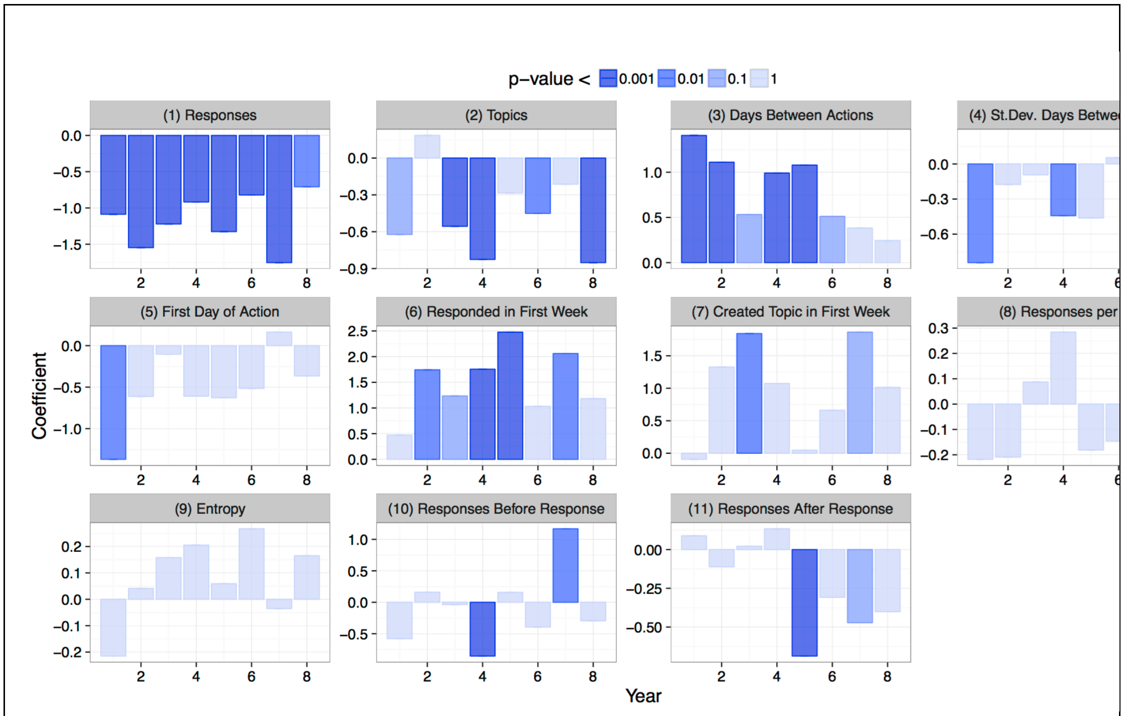


Figure 5: The resulting coefficients for the eight years we are considering in this study.

Similarly, in the first year, an one standard deviation increase of the number of “(1) Responses" will accelerate the death of a user by a factor of exp(-1.1) = 0.33 (i.e., will accelerate by 67%). As we mentioned earlier, the negative coefficients accelerate death while the positive ones decelerate it. Generally speaking we see that higher number of responses (1), higher number of topics (2) higher standard deviation between actions (4), greater fist day of action (5), higher responses after response (11) accelerate failure-- or at least they do accelerate failure when they are significant. On the other hand, a larger number of days between actions (4) and whether or not the user responded to a question or created a topic in the first week (6 and 7) decelerate death.

The differences between the annual coefficients can be interpreted by the variation in our annual datasets. To dig a bit deeper into this, let us focus on the inconsistency of “(2) Topics", i.e., what happens in the second year and the coefficient becomes positive and insignificant?[6]. To understand this we draw the relationship between the number of topics created (normalized) and the censored time in our data, for year 2 and for all the other years grouped together. The results are shown in *Figure 6***.**

In this figure we observe a clear positive relationship between the number of topics and time in the dead instances during the second year. On the other hand, there is an unclear-close to negative relationship between the number of topics and time in the dead instances for all the other years. Even though this observation explains the inconsistency in our coefficients, it does not answer the question of why it happens. To do so, we draw **Figure 7**. In this Figure we see that in the second year, proportionally more people that were heavy contributors did not create new topics (shaded column at 0 in the “Deaths" graph of **Figure 7**), while many users that were not characterized by our HMM as heavy contributors created only one topic. This indicates that during the second year proportionally more heavy contributors were responding to current threads than creating new topics.

As we mentioned in the beginning, our coefficient analysis should not be interpreted as causal. What we have shown here are pure correlations between a set of community characteristics and the likelihood of a user to transition to the heavy contributor state. On the other hand, the fact that the relationship between our independent variables and the dependent variable is not causal does not affect the predictive validity of our probability estimates, which are the epicenter of this study.

“When", “which", and the activation potential

To answer the “when" and the “which users" we need to estimate the survival probability of each user.

Since an AFT model has the characteristic of accelerating the event by a constant factor over a baseline model, we can estimate this baseline model by zeroing the variables of vector $\boldsymbol{\beta}$ of Equation (2). The baseline survivor function of t now becomes:

$$S_0(t) = 1 - \Phi\ ((\log t - \beta_0)/\sigma)$$

where Φ is the cumulative normal distribution. Now we can compute the conditional survival cumulative distribution, $S(t\ |\boldsymbol{X_i})$ for each user i: We know that the proposed AFT model will accelerate the previous baseline by a factor of $\exp(-\boldsymbol{\beta X_i})$. Hence we get:

$$S(t_i|\boldsymbol{X_i}) = S_0\left((\exp(-\beta X_i)\ t_i\right) = 1 - \Phi\ (\ (\log\ (\exp\ (-\beta X_i) t_i\right) - \beta_0)/\sigma)$$
$$= 1 - \Phi((\log\ (t_i) - (\beta_0 + \beta\ X_i)\ )/\sigma)$$

By estimating the conditional probability of a user to die we are estimating the likelihood of a user to become a heavy contributor. Hence, at each point in time, we know which users are more likely to make that transition.

[6] A similar type of analysis can be performed for the rest of the observed inconsistencies but it is omitted here due to space limitations.

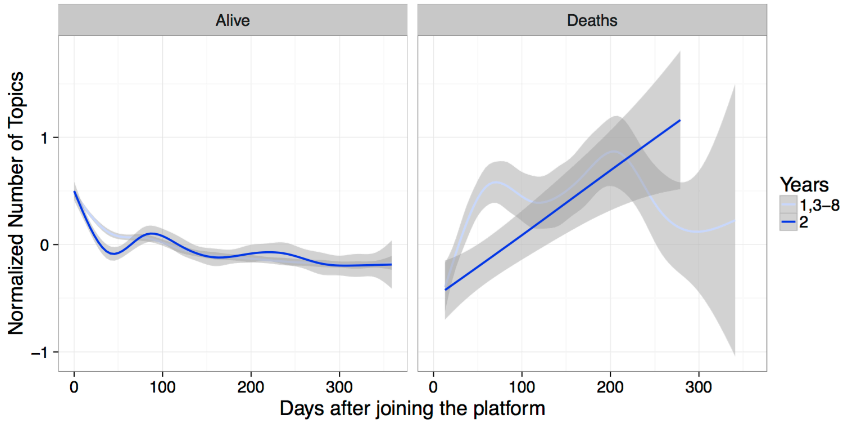


Figure 6: The positive relationship of the normalized number of topics with time in the second year explains the positive sign of the coefficient.

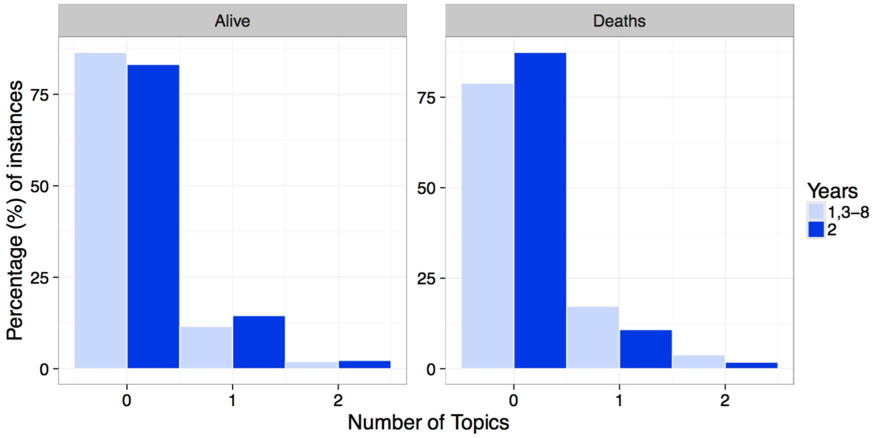
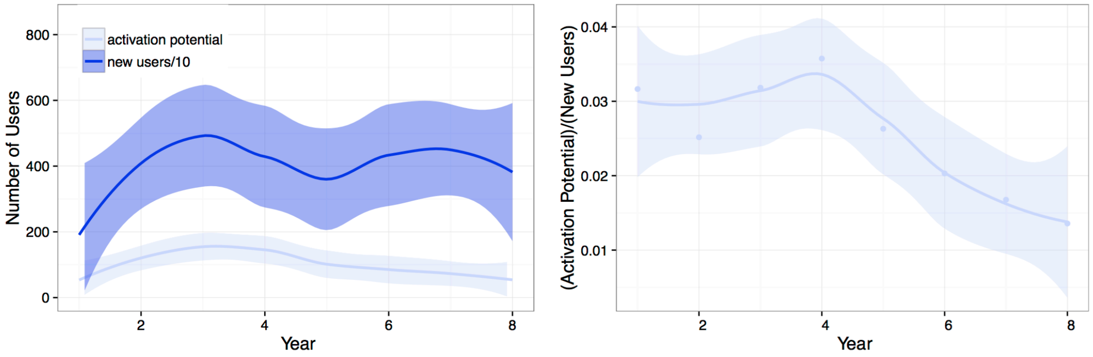


Figure 7: During the second year, proportionally more people that were heavy contributors did not create new topics.

To estimate the annual activation potential of our community we apply its definition. Note that the activation potential is estimated only on users that are currently “alive" (i.e., censored instances).

In **Figure 8a** we show the estimated activation potential as an absolute number of users, and we compare it with the new users that join the platform each year. We see that even though the number of new registrations remains almost constant around 4000 throughout the 8 years, the activation potential increases in the first three years, and then it starts decreasing. This trend is more obvious in **Figure 8b**, where we visualize the activation potential as a percentage of the new users. In that picture the trend is clear: the activation potential seems to be decreasing as the community matures.



(a) Activation potential vs. new users (divided by 10)

(b) Activation potential as a percentage of new users

Figure 8: Activation potential of our community for the 8 years since its inception.

Conclusions

In this work we built a data-driven stochastic framework that provides online communities with the means to estimate their activation potential as well as to understand which users and when are more likely to become heavy contributors. There are a series of straightforward extensions of this work that we intend to pursue in the near future. First, in the current version of this study, we focused only in transitions from any other state to heavy contributor. Of equal interest are the reverse transitions, i.e., from heavy contributor to lurker and from heavy contributor to intermittent contributor. Ideally, we would be interested in community characteristics that are positively correlated (decelerate) with these transitions. Second, we intend to employ a survival analysis with multiple deaths (i.e., multiple state transitions). Finally, we plan to pursue a similar analysis on a completely different type of community (stackoverflow.com) to identify and compare any recurring patterns between the two communities' behavior.

References

Akaike, H. 1974. "A New Look at the Statistical Model Identification," *IEEE transactions on automatic control* (19:6), pp. 716-723.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. 2013. "Steering User Behavior with Badges," *Proceedings of the 22nd international conference on World Wide Web*: ACM, pp. 95-106.

Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., and Wang, X. 2006. "Talk to Me: Foundations for Successful Individual-Group Interactions in Online Communities," *Proceedings of the SIGCHI conference on Human Factors in computing systems*: ACM, pp. 959-968.

Bagozzi, R. P., and Dholakia, U. M. 2006. "Open Source Software User Communities: A Study of Participation in Linux User Groups," *Management science* (52:7), pp. 1099-1115.

Bateman, P. J., Gray, P. H., and Butler, B. S. 2011. "Research Note-the Impact of Community Commitment on Participation in Online Communities," *Information Systems Research* (22:4), pp. 841-854.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The annals of mathematical statistics* (41:1), pp. 164-171.

Bishop, C. M. 2006. "Pattern Recognition," *Machine Learning* (128).

Brzozowski, M. J., Sandholm, T., and Hogg, T. 2009. "Effects of Feedback and Peer Pressure on Contributions to Enterprise Social Media," *Proceedings of the ACM 2009 international conference on Supporting group work*: ACM, pp. 61-70.

Burke, M., Marlow, C., and Lento, T. 2009. "Feed Me: Motivating Newcomer Contribution in Social Network Sites," *Proceedings of the SIGCHI conference on human factors in computing systems*: ACM, pp. 945-954.

Burtch, G., Hong, Y., Bapna, R., and Griskevicius, V. 2015. "What Are Social Incentives Worth? A Randomized Field Experiment in User Content Generation,").

Cassell, J., Huffaker, D., Tversky, D., and Ferriman, K. 2006. "The Language of Online Leadership: Gender and Youth Engagement on the Internet," *Developmental Psychology* (42:3), p. 436.

Chaturvedi, A. R., Dolk, D. R., and Drnevich, P. L. 2011. "Design Principles for Virtual Worlds," *MIS Quarterly* (35:3), pp. 673-684.

Chen, Y., Harper, F. M., Konstan, J., and Xin Li, S. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on Movielens," *The American economic review* (100:4), pp. 1358-1398.

Cheng, R., and Vassileva, J. 2006. "Design and Evaluation of an Adaptive Incentive Mechanism for Sustained Educational Online Communities," *User Modeling and User-Adapted Interaction* (16:3-4), pp. 321-348.

Cleves, M. 2008. *An Introduction to Survival Analysis Using Stata*. Stata Press.

Cox, D. R., and Oakes, D. 1984. *Analysis of Survival Data*. CRC Press.

Drenner, S., Sen, S., and Terveen, L. 2008. "Crafting the Initial User Experience to Achieve Community Goals," *Proceedings of the 2008 ACM conference on Recommender systems*: ACM, pp. 187-194.

Forte, A., and Bruckman, A. 2005. "Why Do People Write for Wikipedia? Incentives to Contribute to Open–Content Publishing," *Proc. of GROUP* (5), pp. 6-9.

Ghose, A., and Han, S. P. 2011. "An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet," *Management Science* (57:9), pp. 1671-1691.

Goes, P. B., Lin, M., and Au Yeung, C.-m. 2014. "“Popularity Effect” in User-Generated Content: Evidence from Online Product Reviews," *Information Systems Research* (25:2), pp. 222-238.

Grambsch, P. M., and Therneau, T. M. 1994. "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals," *Biometrika* (81:3), pp. 515-526.

Hecking, T., Chounta, I.-A., and Hoppe, H. U. 2015. "Analysis of User Roles and the Emergence of Themes in Discussion Forums," *Network Intelligence Conference (ENIC), 2015 Second European*: IEEE, pp. 114-121.

Huang, Y., Singh, P. V., and Ghose, A. 2015. "A Structural Model of Employee Behavioral Dynamics in Enterprise Social Media," *Management Science* (61:12), pp. 2825-2844.

Huberman, B. A., Romero, D. M., and Wu, F. 2009. "Crowdsourcing, Attention and Productivity," *Journal of Information Science*).

Huffaker, D. 2010. "Dimensions of Leadership and Social Influence in Online Communities," *Human Communication Research* (36:4), pp. 593-617.

Joyce, E., and Kraut, R. E. 2006. "Predicting Continued Participation in Newsgroups," *Journal of Computer-Mediated Communication* (11:3), pp. 723-747.

Kim, A. J. 2000. *Community Building on the Web: Secret Strategies for Successful Online Communities*. Addison-Wesley Longman Publishing Co., Inc.

Kohler, T., Fueller, J., Matzler, K., and Stieger, D. 2011. "Co-Creation in Virtual Worlds: The Design of the User Experience," *MIS quarterly* (35:3), pp. 773-788.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Lampe, C., and Johnston, E. 2005. "Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community," *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*: ACM, pp. 11-20.

Lanamäki, A., Rajanen, M., Öörni, A., and Iivari, N. 2015. "Once You Step over the First Line, You Become Sensitized to the Next: Towards a Gateway Theory of Online Participation,").

Lappas, T., Liu, K., and Terzi, E. 2009. "Finding a Team of Experts in Social Networks," *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 467-476.

Lave, J., and Wenger, E. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge university press.

Li, C., and Bernoff, J. 2011. *Groundswell: Winning in a World Transformed by Social Technologies*. Harvard Business Press.

Lu, Y., Jerath, K., and Singh, P. V. 2013. "The Emergence of Opinion Leaders in a Networked Online Community: A Dyadic Model with Time Dynamics and a Heuristic for Fast Estimation," *Management Science* (59:8), pp. 1783-1799.

Luca, M. 2015. "User-Generated Content and Social Media," *Forthcoming in the Handbook of Media Economics, Simon Anderson, David Strömberg and Joel Waldfogel, eds*).

Moon, J. Y., and Sproull, L. S. 2008. "The Role of Feedback in Managing the Internet-Based Volunteer Work Force," *Information Systems Research* (19:4), pp. 494-515.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT press.

Nam, K. K., Ackerman, M. S., and Adamic, L. A. 2009. "Questions in, Knowledge In?: A Study of Naver's Question Answering Community," *Proceedings of the SIGCHI conference on human factors in computing systems*: ACM, pp. 779-788.

Oestreicher-Singer, G., and Zalmanson, L. 2012. "Content or Community? A Digital Business Strategy for Content Providers in the Social Age," *A Digital Business Strategy for Content Providers in the Social Age (July 01, 2012)*).

Ozturk, P., and Nickerson, J. 2015. "Paths from Talk to Action,").

Preece, J., and Shneiderman, B. 2009. "The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation," *AIS Transactions on Human-Computer Interaction* (1:1), pp. 13-32.

Ransbotham, S., and Kane, G. C. 2011. "Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia," *MIS Quarterly-Management Information Systems* (35:3), p. 613.

Ransbotham, S., Kane, G. C., and Lurie, N. H. 2012. "Network Characteristics and the Value of Collaborative User-Generated Content," *Marketing Science* (31:3), pp. 387-405.

Ray, S., Kim, S. S., and Morris, J. G. 2014. "The Central Role of Engagement in Online Communities," *Information Systems Research* (25:3), pp. 528-546.

Ren, Y., Harper, F. M., Drenner, S., Terveen, L. G., Kiesler, S. B., Riedl, J., and Kraut, R. E. 2012. "Building Member Attachment in Online Communities: Applying Theories of Group Identity and Interpersonal Bonds," *Mis Quarterly* (36:3), pp. 841-864.

Shi, X., Zhu, J., Cai, R., and Zhang, L. 2009. "User Grouping Behavior in Online Forums," *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 777-786.

Shriver, S. K., Nair, H. S., and Hofstetter, R. 2013. "Social Ties and User-Generated Content: Evidence from an Online Social Network," *Management Science* (59:6), pp. 1425-1443.

Tsai, H.-T., and Bagozzi, R. P. 2014. "Contribution Behavior in Virtual Communities: Cogntiive, Emotional, and Social Influences," *Mis Quarterly* (38:1), pp. 143-163.

van Mierlo, T. 2014. "The 1% Rule in Four Digital Health Social Networks: An Observational Study," *Journal of medical Internet research* (16:2), p. e33.

Viterbi, A. 1967. "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE transactions on Information Theory* (13:2), pp. 260-269.

Wasko, M. M., and Faraj, S. 2005. "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS quarterly*), pp. 35-57.

Weiss, A. M., Lurie, N. H., and MacInnis, D. J. 2008. "Listening to Strangers: Whose Responses Are Valuable, How Valuable Are They, and Why?," *Journal of Marketing Research* (45:4), pp. 425-436.

Yoo, Y., and Alavi, M. 2004. "Emergent Leadership in Virtual Teams: What Do Emergent Leaders Do?," *Information and Organization* (14:1), pp. 27-58.

Zeng, X., and Wei, L. 2013. "Social Ties and User Content Generation: Evidence from Flickr," *Information Systems Research* (24:1), pp. 71-87.