Paying attention to each other in visible work communities: Modeling bursty systems of multiple activity streams

Jamie F. Olson
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15214
jfolson@cs.cmu.edu

James Howison
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15214
jhowison@cs.cmu.edu

Kathleen M Carley
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15214
kathleen.carley@cs.cmu.edu

***Abstract*—Online work projects, from open source to wikipedia, have emerged as an important phenomenon. These communities offer exciting opportunities to investigate social processes because they leave traces of their activity over time.**

We argue that the rapid visibility of others' work afforded by the information systems used by these projects reaches out and attracts the attention of others who are peripherally aware of the group's online space, prompting them to begin or intensify their participation, binding separate individual streams of activity into a social entity.

Previous work has suggested that for certain types of bursty social behavior (e.g. email), the frequency of the behavior is not homogeneously distributed but rather can be divided into two generative mechanisms: active sessions and passive background participation. We extend this work for the case of multiple conditionally independent streams of behavior, where each stream is characterized by these two generative mechanisms. Our model can characterized by a double-chain hidden markov model, allowing efficient inference using expectation-maximization. We apply this model to visible work communities by modeling each participant as a single stream of behavior, assessing transition probabilities between active sessions of different participants. This allows us to examine the extent to which the various members of the community are influenced by the active participation of others. Our results indicate that an active session by a participant at least triples the likelihood of another participant beginning an active session.

I. Introduction

Novel information technologies have given rise to new organizational forms based on volunteer contributions, such as open source and Wikipedia. Understanding these organizations is an important challenge both because they are increasingly important in the world and for the lessons they might provide about human behavior generally. In addition better understandings may help these projects as they seek to sustain their success.

Volunteer participation immediately raises the question of motivation, which has been extensively studied [14], [15], [9], [8], [24], [4], [11]. This work cites motivations like a) the product itself, b) learning by doing, c) intellectual stimulation, d) self-efficacy, e) building reputations and f) learning from others. This work speaks of motivations in a very general form and is almost always based on surveys or interviews [5]. The motivations share the characteristic that they are all very rational explanations of behavior and we believe that the possibility exists that these explanations are influenced by post-hoc justifications of the time spent on these projects. In any case they don't explain an interesting question: why do participants choose to participate at a particular time, and what might that tell us about the interesting success of these new organizational forms?

A potential participant's attention can be drawn to the project and its work in different ways. A first is that they may wish to accomplish something in particular and *intend* to focus their attention primarily on the project's work. A second is that their attention may be drawn by intersection with the project as an *incidental consequence* of other activities, such as when one is using a reference manager to write a paper and finds an annoyance, or when one is researching a city for a trip and reads a Wikipedia page.

These two mechanisms appear to relate to the temporal patterns of work in these communities. We argue (and provide evidence) that there are two main such patterns of work: small numbers of events sporadically dispersed in time and short periods of high intensity, where many events come quickly, which we call active sessions. Sporadic participation seems likely to result from attention being drawn to the project, but not held for a substantial period of time, instead returning to other competing activities, such as firing off a quick bug-report while writing an academic paper. Active sessions, on the other hand might be generated by intentional, focused work, where participants are pursuing specific goals.

We argue that this socio-technical characteristic of visible work systems means that participants are likely to synchronize their attention to the project. Not only does such visible work attract attention but it also signals that other participants are awake, online and paying attention to the project (This point has been made in the "social presence" literature, particularly in the context of distance learning, e.g. [12], [21], [20]). This seems likely to create a particularly likely time for a participant to turn their primary attention to the project, taking them out of

978-0-7695-4211-9/10 $26.00 © 2010 IEEE
DOI 10.1109/SocialCom.2010.46

their incidental work pattern and into an active session. Once multiple participants are actively working, they may continue to prompt each other to continue to work. This could occur through obvious mechanisms like directly talking and asking and answering questions. Answered questions might solve problems which would otherwise have blocked a participant's work causing them to turn their attention to other non-project work. Other, less obvious, mechanisms might also operate, such as being annoyed by another participant's work (see Figure 1) and working to correct it. In short, the mechanism of visible work might operate to move a project from a collection of individuals towards a social entity, where the individual's work patterns are affected by and relate to those of others.

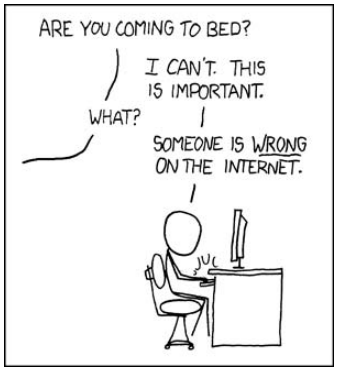


Fig. 1. "Duty Calls" XKCD provides a humorous look at social attention extracting more time than potential participants might otherwise have contributed. Reprinted with permission, see http://xkcd.com/about/

Motivated by this theory, this paper builds a model to analyze temporal data from Wikipedia, which is a leading example of a visible work community. The overall intention of the model is to explore the proposition that participant's temporal work patterns are responsive to one another. We first examine the issues in modeling human behavior in the temporal domain, identifying an existing model as a base and extending it to match our domain more closely. Our extensions are described in detail. We then turn to our dataset, providing relevant contextual background on Wikipedia and descriptive statistics on temporal patterns overall and for individual participants. We then present the results of the model and discuss their interpretation and limitations in the context of the theory above. Finally we conclude and discuss appropriate future work, both for improving our exploration of this theory and alternative applications of the new model introduced in this paper.

II. Modeling temporal behavior

A large body of work suggests that many human behaviors are heavy-tailed and bursty in the temporal domain [2], [22], [18], [13], [1], [7], [19], [23], [16]. However, several different mechanisms have been used to explain these properties. Some propose a priority-queue where individuals choose high priority tasks over low priority tasks [2], [22], similar to preferential attachment in network evolution [3]. Although these models reproduce several of the aforementioned heavy-tail and burstiness of human temporal behavior, they are inconsistent with several important properties of real world human behavior, notably circadian rhythyms and infrequent "sessions" of high activity [18], [1]. Recently, certain nonhomogeneous Poisson processes were shown to be able to produce the heavy-tails and burstiness that have been empirically observed [18]. The nature of these cascading Poisson processes allow researchers to include mechanisms like "session" and circadian rhythm directly in the model.

Our work is heavily influenced by the model proposed by Malmgren and colleagues [18], which they subsequently simplified [17]. They propose a Markov mixture of Poisson processes cast as a double-chain hidden Markov model. Specifically, they use a mixture of two Poisson processes, represented by the hidden states (Figure 3) in their double-chain hidden Markov model. During the active state, events are generated by a homogeneous Poisson process with a high rate ρ_a. In the passive state, events are generated by a nonhomogeneous Poisson process with rate $\rho_p(t)$ dependent on the current time. The passive state is intended to represent a simple version of circadian rhythyms and is defined by two square pulse distributions p_d, p_w and a rate parameter ρ_0.

$$\rho_p(t) = \rho_{p0} W p_d(t|\tau_{d0}, \tau_{d1}, \epsilon_d) p_w(t|\tau_{w0}, \tau_{w1}, \epsilon_w) \quad (1)$$

Where the square pulse distribution with period τ is defined as

$$p(t|\vec{\tau}, \epsilon) = \begin{cases} w & (t \text{ modulo } \tau) \in [\tau_0, \tau_1) \\ \epsilon w & \text{otherwise} \end{cases} \quad (2)$$

$$w = (\epsilon\tau + (1-\epsilon)(\tau_1 - \tau_0))^{-1} \quad (3)$$

such that probability density between τ_0 and τ_1 is elevated relative to rest ($\epsilon < 1$). In order to represent circadian rhythyms, p_d represents the activity during the hours of the day with $\tau_d = 24$ and p_w represents elevated activity during some portion of the week with $\tau_w = 7$.

Figure 2 shows how the square pulse distributions, representing "out-of-session" activity, and the high "in-session" rate are combined to describe human attentional patterns by taking into account individual variation in circadian rhythms. Each individual has certain days of the week in which they are generally active as well as certain times of the day in which they are generally active. People may also vary as to how much how much more active they are during those times of the day/week(how much above/below the dotted-line they go). Separately, people vary as to the intensity with which they contribute when they are actively contributing (red line). These things can all be estimated from data and together they define the model.

The EM algorithm is used to jointly estimate the hidden states and the parameters, $\theta = \{\varepsilon, \rho_a, \rho_{p0}, \tau_{p0}, \tau_{p1}, \epsilon_d, \tau_{w0}, \tau_{w1}, \epsilon_w\}$. The complexity of these nonhomogeneous Poisson processes means that direct update formulas for the M-step are not available. However, because the likelihoods are convex with respect to the

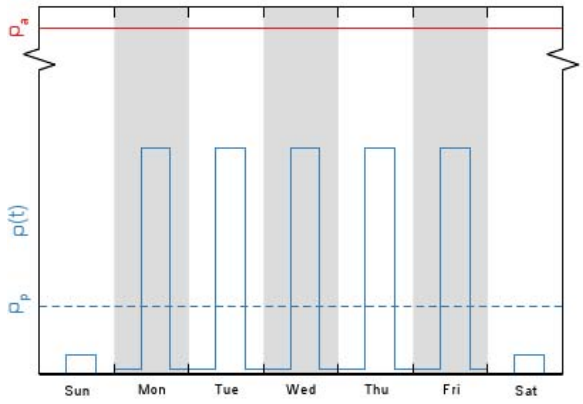


Fig. 2. Circadian rhythms are approximated by combining a "week" and "day" square pulse distribution.

parameters, Powell's method can be used to obtain maximum likelihood estimates for the parameters.

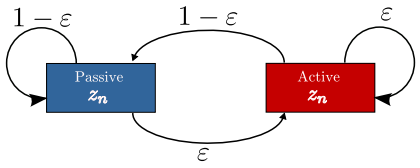


Fig. 3. A single activity stream is characterized by two states: Active and Passive

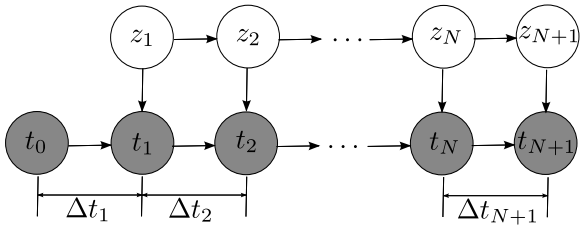


Fig. 4. Double-chain Hidden Markov Model where the observed event times, t, are conditioned on the hidden states, z.

Although this hidden markov model may seem somewhat arbitrary and complex, it has a number of advantages over alternative models. Traditional autoregressive models are not appropriate due to correlations among the outcome variables as well as the burstiness of the processes. The whole purpose of this analysis is exploring the complex inter-dependencies between the different activity streams, which are generally assumed not to exist under the autoregressive econometric models, undermining the validity of their statistical power. This multicolinearity is likely to exist regardless of the actual amount of interactivity between streams as a result of overlapping circadian rhythms. Furthermore, the error/noise term is almost surely bursty and not Gaussian or homogeneous Poisson as would generally be assumed. In contrast, the hidden markov model has been validated against both general properties of human social behavior [18] as well as specific datasets [17], and can be interpreted as a branching process, a general class of point processes that has been used successfully to model such behavior [7], [13].

III. Model

We extend the work of [17] to model multiple streams of activity. The goal is to model interactions between these streams while preserving the computational properties of the DCHMM. We assume that these K streams are not fully independent but are conditionally independent given the state information. We use the two "in-session" and "out-of-session" states but duplicate those hidden states for each activity stream with the assumption that each stream is conditionally independent of all other streams given its current state.

Because Poisson processes are memoryless (equation 5), we can easily construct likelihoods for each observation in the K streams. For an observation, o_n occuring at time t_n in activity stream s_n, the likelihood of the system being in each state is shown in equation 6.

$$T \sim \text{Poisson}(\lambda) \tag{4}$$

$$P(T = t) = P\left(T < t_0\right) P\left(T = (t - t_0)\right) \quad \forall t_0 < t \tag{5}$$

$$P(o_n | Z_n^* = k) = P(T_{s_n} = t_n | Z_n^*) \prod_{l \neq s_n} P(T_l > t_n) \tag{6}$$

In order to represent this unconstrained system using a hidden markov model, we would be required to explicitly represent each of the 2^K possible configurations of "session" engagement (in a session vs not in a session) across the K individuals. Inference of such a model would require estimation of the 2^{2K} transition probabilities between these states and even using some smoothing strategy e.g. pseudocounts, we are unlikely to have sufficient data for robust parameter estimates. Alternatively, it may be useful to model interactions between the different individuals as a time-dependent markov random field. However, although some work has explored unsupervised inference of markov random fields [6], the statistical and computational properties of such inference are not well understood and as such they may not be able to provide reliable parameter estimates.

We choose instead to limit configurations of the system to those where at most one of the K streams is active. This leads to $K + 1$ states for the system, with the total number of transition probabilities, and ultimately, parameters, growing with the square of the number of activity streams. This limits the model to scenarios where only one activity stream can be active at each event but where transitions between streams are of interest. We discuss limitations deriving from this decision below.

IV. Data

We applied this model to a WikiProject in Wikipedia. A WikiProject is a group of Wikipedians who work to improve

a section of Wikipedia. Projects include topics such as Music, Sports and Geographical regions. The Project identifies Articles which are considered "in scope" and tags them as associated with the WikiProject. We assume that participants are observing portions of these pages in their Watchlists, and discussions of work done for the WikiProject in the Project's collaboration pages. For this reason we argue that it is reasonable to believe that participants are aware of when others are actively working, making this an example of visible online work.

We chose to study WikiProject Oregon both for its high level of activity and because, being geographical in nature, participants likely to be in the same time-zone, making real-time coordination more likely. The Project Oregon "About Us" page states that it was founded in March 2005 and "experienced a lot of growth in late 2007 and 2008."

We accessed a March 12, 2008 dump of English Wikipedia and downloaded all revisions to Articles, and their associated Talk pages, marked as in scope for the Project. A datapoint consists of a user-id and a timestamp; we do not use data about which specific page was edited, since we wanted to capture the idea that participants could also be motivated to work on nearby pages, or indeed anywhere else in the Project's scope.

Overall the dataset consists of 354,793 revision events by 25,780 different users and 5622 articles. Because we seek to model interactions within a community, we limit ourselves to those we define as community members. In this case, we consider a participant to be a member of WikiProject Oregon if over the history of the project they engage in Talk activities and Edit activities at least 100 times. This reduced the dataset to 55,104 (15% of the total) revisions and 24 users (0.09%) (a highly skewed distribution of contributions).

V. Results

We apply our model to the WikiProject Oregon data described previously. The outputs of the model are parameters associated with the temporal distribution of revisions and a transition matrix, showing the likelihood estimates for transitions between the state of the system.

Figures 5-7 shows the parameters learned for temporal distribution of revisions across the three years. Most of the revision parameters are quite stable across the three years. This is consistent with previous results on email patterns [17] which found consistent characteristic temporal patterns in individual behavior. Of potential interest is that the user population generally begins their WikiProject "wiki-week" on Thursday and works over the weekend. This means that for the most actively involved members of the WikiProject, the bulk of their contributions came not during the traditional work-week, but on week-ends.

The transition matrix shows transition probabilities between the states of the model; there are $K+1$ states, one indicating an active session for each participant and one indicating that no participant is currently in an active state (system passivity). For each event (i.e. a revision) the model estimates the probability that the system is in each state.

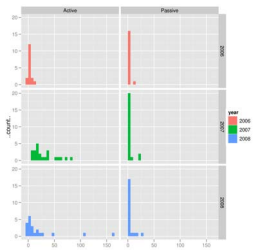
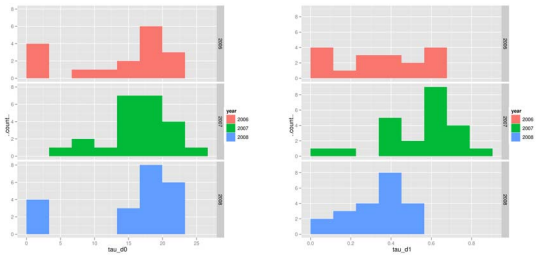


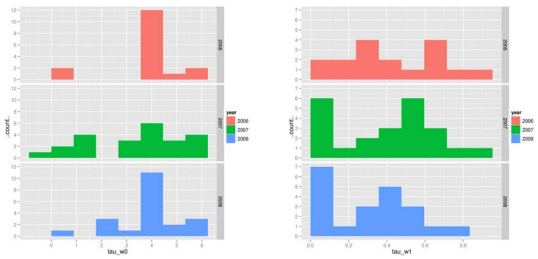
Fig. 5. Active(ρ_a) and Passve(ρ_p) rate per hour



(a) Begin Day(τ_{d0}) (b) Day Length(τ_{d1})

Fig. 6. Elevated activity "wiki-day"

Two types of transitions are particularly relevant to the social theory motivating this study: the first is the probability of a transition to system passivity, given an active session (by any participant), informally written as P(Active—No Active). The second is the probability of a transition to an active session of any participant, given an active session by a different participant, informally written as P(Active—Other Active). A t-test comparing these transition probabilities, (means shown in Table I), indicates that in all years P(Active—Other Active) is significantly greater ($p < 0.05$) than P(Active—No Active),



(a) Begin Week(τ_{w0}) (b) Week Length(τ_{w1})

Fig. 7. Elevated activity "wiki-week"

Year	P(Active—Other Active)	P(Active—No Active)
2006	0.0195	0.004
2007	0.0385	0.012
2008	0.0271	0.002

TABLE I
THE CONDITIONAL PROBABILITIES LEARNED FOR EACH YEAR

with P(Active—Other Active) averaging 6.21 times greater than P(Active—No Active).

We can also construct a network based on the state transition matrix. Using the probability of any user spontaneously becoming active we dichotomize the state transition network, removing all transitions that have a probability less than twice the baseline P(Active—No Active). Using the 2007 parameters, this results in the sparse network shown in Figure VI.

VI. Discussion

The central question of this paper is the extent to which participants' temporal work patterns are responsive to one another. Table I and the associated hypothesis test give confidence that transitions between active sessions of different participants are significantly more likely than transitions to an active session when no one is active. These results also indicate a substantial effect size, suggesting that an active session at least triples the likelihood of another user beginning an active session. We interpret this as evidence that visible activity by WikiProject members increases the probability that other members will begin an active session of work. The magnitude of this effect appears to be stronger in 2007 and 2008 than in 2006; this matches the statement on the WikiProject homepage that they became more active and organized beginning in 2007.

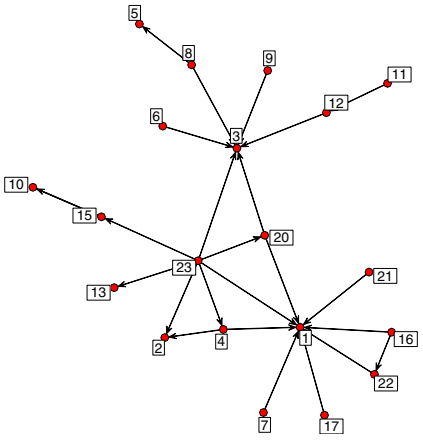


Fig. 8. Potential influence network for 2007

There is also evidence that some participants are more likely to respond to the active sessions of specific others. Figure VI can be interpreted as an implied influence network, with an edge drawn between two nodes, A and B if the estimated value $P(B|A)$ is more than twice the estimated marginal $P(B)$. In other words, individuals are connected is active attention is likely to spread from one to the other. Node 23 has an outdegree of 7, meaning that their visible work may capture the attention of others and motivate their transition into active participation. Nodes 1 and 3 have high indegree meaning that their active sessions tend to follow those of many others, suggesting that they attend to the visible work of a large number of WikiProject Oregon participants.

VII. Limitations

There are several key limitations of this study, relating both to the model and to our interpretation of the results.

The model constrains the system such that only one activity stream can be active at each event, restricting its ability to model simultaneous bursty work. In situations where multiple activity streams are in fact in simultaneous sessions, this model will instead identify numerous transitions back and forth between the active sessions. This will be reflected in increased stream transitions probabilities for overlapping periods. This means that the model cannot distinguish between multiple simultaneous activity sessions and a sequence of non-overlapping activity sessions. We do not believe that this threatens the overall result of the paper, since both patterns are indicative of responsiveness between participants whether it is sequential or coordinated simultaneous active attention.

Furthermore, the Markov assumption in this limited state space means that transitions are now effectively conditioned on the single activity stream that was in an active session (or the latent state) rather than the set of previous session information. For relatively dense streams of activity, this may be problematic. If stream A undergoes an active session which leads to an active session in B, but a third irrelevant stream, C, is active in the time between when A stops and when B begins, the relationship between A and B will not be correctly inferred. Given the asynchronous capabilities of the information system this means we are unable to capture all sources of responsiveness, just those that are immediate and direct.

We have interpreted the social synchronization of participants indicated by the transition probabilities to indicate the operation of the attention effect theorized in the introduction to this paper; that is endogenously to the social system. However it is possible that attention is drawn to the project exogenously, through events occurring in the world and being reported on in the news or blogs. Such events are known to produces bursts of activity in traditional media [13] and so seem particularly likely to affect articles regarding entities currently in the news. This effect is particularly seen in biographies of living persons, where Wikipedia has often had to institute temporary locks on article editing, due to a flood of participants attracted by the currency of the articles topic. There is no reason to believe that Project Oregon's scope would be particularly susceptible to this effect, but it would be desirable to extend the modeling framework to identify exogenous effects from newsworthy events.

VIII. Conclusions and Future Work

This paper argues that the ability to successfully compete for potential participant's moment-to-moment attention con-

tributes to the surprising success of projects like open source software development and Wikipedia. We argue that one reason they successfully compete is the social pull resulting from the visibility of other's work. This mechanism might explain why the experience of work in these projects is described as social rather than simply being experienced as individual work.

We extended a previously validated generative model of human attention to characterize the interactivity and responsiveness between individuals in these visible work communities. Although the restriction of a single activity stream being "in-session" at any point in time may not be entirely appropriate for the modeling of a single type of activity across many different people, it may be both reasonable and desirable when considering the attention of a single individual across multiple different streams of activity. The main focus of this work is to argue for the presence of attentional responsiveness in visible work communities; however, we hope that this model will prove useful in modeling other activity streams, especially those constrained by the cognitive/attentional limits of a single individual.

Throughout the paper we have presented this effect as attracting more participation and suggested that is a reason for the surprising success of visible work communities. It is clear, however, that having too many participants descend on a page at once is likely to lead to coordination problems. In Wikipedia social responsiveness has a negative side too, as participant's attention can be drawn by opposing perspectives, descending into edit wars. It may be that the mechanism of motivation and attracting investigated in this paper has an inflection point, above which its effects become negative, prompting system administrators to dampen its effects by instituting temporary edit locks, slowing down the stream of attention grabbing edits. Despite some concerns regarding some of the modeling assumptions, we believe we have found good support for the hypothesis that the active attention of some members of visible work communities can lead other members to actively devote their attention to the project as well.

IX. Acknowledgments

This work was supported in part by the Office of Naval Research (N00014-06-1-0104) for adversarial assessment and (N00014-08-11186) for rapid ethnographic assessment, the Army Research Office and ERDC-TEC (W911NF0710317) and the National Science Foundation (#0943168). Additional support was provided by CASOS - the center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Army Research Institute, the U.S. Army Engineer Research and Development Center's (ERDC), Topographic Engineering Center, the National Science Foundation or the U.S. government.

References

[1] L. Amaral, D. B. Stouffer, and R. D. Malmgren. Log-normal statistics in e-mail communication patterns. pages 1–18, 2006.

[2] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(May):207–211, 2005.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[4] P. O. Chin and D. Cooke. Satisfaction and coordination in virtual communities. In *Proceedings of the Tenth Americas Conference on Information Systems*, New York, New York, 2004.

[5] K. Crowston, K. Wei, J. Howison, and A. Wiggins. Free/Libre Open Source Software: What We Know and What We Do Not Know. *ACM Computing Surveys*, 2008.

[6] G. B. Davis, J. F. Olson, and K. M. Carley. Unsupervised Plan Detection with Factor Graphs. *Knowledge Discovery from Sensor Data (Sensor-KDD 2008)*, page 33, 2008.

[7] F. Deschatres and D. Sornette. Dynamics of book sales: Endogenous versus exogenous shocks in complex networks. *Physical Review E*, 72(1):16112, 2005.

[8] R. A. Ghosh, G. Robles, and R. Glott. Free/libre and open source software: Survey and study floss. Technical report, International Institute of Infonomics, University of Maastricht: Netherlands, 2002.

[9] A. Hars and S. Ou. Working for free? Motivations of participating in FOSS projects. *International Journal of Electronic Commerce*, 6(3):25–39, 2002.

[10] J. Howison. *Alone Together: A socio-technical theory of motivation, coordination and collaboration technologies in organizing for free and open source software development*. PhD thesis, Syracuse University, School of Information Studies, 2009.

[11] W. Ke and P. Zhang. Participating in open source software projects: The role of empowerment. In *Proceedings of ICIS08 HCI Workshop*, 2008.

[12] B. Kehrwald. Understanding social presence in text-based online learning environments. *Distance Education*, 29(1):89–106, 2008.

[13] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[14] K. Lakhani and E. von Hippel. How open source software works: "free" user-to-user assistance. *Research Policy*, 32:923–943, 2003.

[15] K. Lakhani and R. Wolf. Why hackers do what they do: Understanding motivation efforts in Free/F/OSS projects. Working Paper 4425-03, MIT Sloan School of Management, 2003.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic Evolution of Social Networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.

[17] R. D. Malmgren, J. M. Hofman, and D. J. Watts. Characterizing Individual Communication Patterns. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, volume 15, pages 607–615. ACM, 2009.

[18] R. D. Malmgren, D. B. Stouffer, and A. E. Motter. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158, 2008.

[19] K. Naruse and M. Kubo. Lognormal Distribution of BBS Articles and its Social and Generative Mechanism. *Web Intelligence*, 2006.

[20] L. Rourke, T. Anderson, D. Garrison, and W. Archer. Assessing social presence in asynchronous text-based computer conferencing. *Journal of distance education*, 14(2):50–71, 1999.

[21] J. Short, E. Williams, and B. Christie. *The social psychology of telecommunications*. John Wiley & Sons., New York, 1976.

[22] A. Vázquez, J. a. G. Oliveira, Z. Dezsö, K. I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(036127):1–19, 2006.

[23] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. *International Conference on Management of Data*, 2004.

[24] Y. Ye and K. Kishida. Toward an understanding of the motivation of open source software developers. In *Proceedings of 2003 International Conference on Software Engineering (ICSE)*, Portland, Oregon, USA, May 3-10 2003.