Tiago Filipe Teixeira dos Santos, Dipl.-Ing.

Activity Dynamics in Peer Production Systems

DOCTORAL THESIS

to achieve the university degree of
Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

Supervisor

Denis Helic, Assoc.Prof. Dipl.-Ing. Dr.techn.
Institute of Interactive Systems and Data Science

Graz, May 2020

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date, Signature

Abstract

Peer production systems are digitally embedded socio-economic systems for content creation, curation and sharing [BN06; Wil08], and they enable large-scale collaborations. Prominent examples of peer production systems include Wikipedia, an online encyclopedia, or Stack Exchange, a network of online communities dedicated to questions-and-answers (Q&A) on topics which range from academia to writing. Growing and sustaining large-scale peer production systems like Wikipedia poses a complex challenge [VWD04; Gil05], and not all comparable efforts succeed. However, successful systems stand to capitalize on the "wisdom of the crowds" [Gal07; Sur05], i.e. positive synergies arising from mass collaboration. Therefore, understanding how and why some systems grow and thrive, while others decline and shut down, is an important endeavor. In particular, understanding and modeling the activity dynamics of peer production systems poses a stepping stone towards shaping the activity dynamics. The main objective of this thesis is to model aspects of activity dynamics in peer production systems. This thesis first derives activity dynamics characterizations on the level of users and whole systems. The results from that characterization inform the derivation of Hawkes process-based models for user excitation, a measure for strength of influence between and within user groups. These models uncover a range of excitation effects closely related to the growth trajectory of Stack Exchange Q&A communities. This thesis also contributes methodological insights to Hawkes process-based models, as well as empirical insights on other aspects of activity dynamics, such as the effectiveness of badges as an activity dynamics steering tool, and the performance of the crowds in peer production systems. Overall, the results contained in this thesis are of interest to practitioners and researchers aiming to improve their system managing and activity dynamics modeling efforts.

Kurzfassung

Online-Kollaborationssysteme sind digital eingebettete sozioökonomische Systeme zur Erstellung, Kuration und Weitergabe von Inhalten [BN06; Wil08]. Diese Systeme ermöglichen Gruppenarbeit mit sehr großer Anzahl an Teilnehmer. Zu den namhaften Beispielen der Online-Kollaborationssysteme zählen Wikipedia, eine Internet-Enzyklopädie, oder Stack Exchange, ein Netzwerk von Online-Gemeinschaften, die Fragen und Antworten (Q&A) zu zahlreichen Themen bieten. Die Förderung und Aufrechterhaltung von großen Online-Kollaborationssystemen wie Wikipedia stellen eine komplexe Herausforderung dar [VWD04; Gil05], dessen Erfolg sich nicht leicht nachahmen lässt. Erfolgreiche Systeme können jedoch von der "Weisheit der Vielen" [Gal07; Sur05] profitieren, d.h. von positiven Synergien, die sich im Kontext der Kooperation auf großer Skala ergeben. Daher spielt die Bemühung zu erforschen, wie und warum einige Systeme wachsen und erfolgreich werden, während andere schrumpfen und stillgelegt werden, eine wichtige Rolle. Die Aktivitätsdynamik in den Online-Kollaborationssystemen empirisch zu erforschen und mathematisch zu modellieren, ist ein wichtiger Meilenstein auf dem Weg zur Kontrolle und Steuerung der Aktivitätsdynamik in Online-Kollaborationssystemen. Das Hauptziel dieser Arbeit besteht in der Modellierung bestimmter Aspekte der Aktivitätsdynamik in den Online-Kollaborationssystemen. Diese Arbeit leitet zunächst eine Charakterisierung der Aktivitätsdynamik auf der Benutzerebene und der der ganzen Systeme ab. Die Ergebnisse der Charakterisierung formen die Basis für die Modellierung mit Hawkes Prozessen, die die Benutzeranregung, ein Maß für die Stärke des Einflusses zwischen und innerhalb von Benutzergruppen, erfasst. Diese Modelle zeigen neuartige Benutzeranregungseffekte auf, die eng mit dem Wachstumspfad von Stack Exchange Q&A-Gemeinschaften zusammenhängen. Diese Arbeit liefert sowohl Einblicke in die Theorie und Anwendung von Hawkes Prozessen als auch empirische Einsichten in andere Aspekte der Aktivitätsdynamik, wie die Wirksamkeit von sogenannten "badges" als Steuerungsinstrumente und die Eigenschaften der "Weisheit der Vielen" in Online-Kollaborationssystemen. Zusammenfassend sind die in dieser Arbeit enthaltenen Ergebnisse für im Rahmen der Systemverwaltung und Modellierung der Aktivitätsdynamik fungierende Praktiker und Forscher von großem Interesse.

Acknowledgments

I would like to express much gratitude to my advisor Denis Helic for his academic guidance, for his support, time and openness to discuss all kinds of technical and non-technical questions alike, for his ability to motivate me to push my boundaries further, and, especially, for having the patience to almost always make me overcome my sporadic "Betreuungsresistenz"[1] (here, the one exception is the "MJ is the greatest" discussion, since we are still not quite on the same page).

I would also like to thank (in chronological order) all other "academic fathers" who have taught me so much throughout my years as a PhD student, and with whom I collaborated on many research projects: Roman Kern, Simon Walk, Markus Strohmaier and Florian Lemmerich. As part of this group of collaborators, I would also like to thank Kristina Lerman and Keith Burghardt for my memorable visit at the USC.

Next (and in alphabetical order), I would like to acknowledge all other research group colleagues I have had the pleasure of crossing paths with: Lukas Eberhard, Bernhard Geiger, Niklas Hopfgartner, Andre Kaestner, Aleksandar Karakaš, Patrick Kasper, Philipp Koncar, Georgios Koutroulis, Olivia Pfeiler, Thorsten Ruprechter, Stefan Schrunner, Maximilian Toller, Matthias Wölbitsch and Anja Zernig. I am also thankful to all the colleagues at the institute for having me and for teaching me so much of the Austrian dialect during the coffee breaks.

Finally, my deepest gratitude goes to the ones closest to me: my friends and the Žibret family, who were at my side especially when it mattered most, my mom, who is relentless in her support, and my wife Veronika, who nurtures the kind of love, friendship, and infallible sense of humor I am grateful to wake up every day to.

Institutional acknowledgments. Most of the work comprising this thesis was done during my time as a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

[1] Translating from German, this means something like "resist taking the supervisor's advice".

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Peer production systems, as digitally embedded socio-economic systems for content creation, curation and sharing [BN06; Wil08], have enabled large-scale collaborations on a broad range of topics. Prominent examples of peer production systems include Wikipedia, an online encyclopedia, GitHub, an open source software development platform, Reddit, a social news aggregator, and Stack Exchange, a network of online communities dedicated to questions-and-answers (Q&A) ranging from academia to writing. Such platforms function as knowledge repositories [And+12; Dab+12], spread news [Lih04], and form public spaces for discussion [WZH13].

Growing and sustaining large-scale peer production systems like Wikipedia poses a complex challenge [VWD04; Gil05], and not all comparable efforts succeed: To name an example, Google knol, a website for user-contributed encyclopedic articles, failed to attract users and activity, and, as a consequence, shut down four years after inception. However, peer production systems which attain critical mass [RMJ10; SW14] stand to benefit from the "wisdom of the crowds" [Gal07; Sur05], i.e. positive synergies arising from mass collaboration. As such, understanding and modeling the activity dynamics of peer production systems has attracted much research attention, which lead to macro-level predictive models [Yan+10; Rib14; Wal+16] for the survival and development of entire peer production systems. On the level of individual users, previous research has enabled the automated discovery of user roles [Gei+19] and characterized a newcomer-vs.-veteran dichotomy [Kit+07]. Linking both streams of research, i.e. improving our understanding of how user activity patterns combine to form longitudinal macro-level trends, will bridge both user-level and system-level research. This can be seen as a crucial stepping stone towards better modeling and shaping of activity in peer production systems.

This thesis focuses on modeling user-level and system-level aspects of activity dynamics in peer production systems. Besides deriving empirical results which uncover developmental properties of peer production systems, this thesis also focuses on illustrating and extending tools to facilitate future studies of longitudinal patterns in and beyond peer production systems. Hence, this thesis is relevant for both peer production system builders and managers aiming to capture and optimize the development of the systems they oversee, as well as researchers modeling dynamics of such systems.

In Section 1.2, this thesis addresses challenges and opportunities in understanding and modeling activity dynamics of peer production systems. This serves as the basis for Section 1.3, which comprises the description of the problem statement, objectives and an overview of the approach. The research questions framing this thesis are the subject of

Figure 1.1: **Excitation types between groups of users.** This schema illustrates different kinds of excitation between two groups of users of peer production systems. In this model of excitation, users of a given group may react to each other, as suggested by the gray arrows, or to users of other groups, as indicated by the pink arrows.

Section 1.4. Section 1.5 then lists which publications underlie each research question, and Section 1.7 exposes the main contributions of this thesis. This introduction concludes, in Section 1.8, with a presentation of the structure of the rest of the thesis.

1.2 Activity Dynamics in Peer Production Systems

This thesis focuses on modeling activity dynamics in peer production systems, on the level of single participants as well as whole systems. Though the previously mentioned peer production systems differ starkly in scope, they all consist of large numbers of users pursuing a common goal. The engagement of users towards that shared objective varies along a continuum typically following a power-law relation [Wil08]: The overwhelming majority of users contribute very little, while select few produce large amounts of content. The temporal activity patterns of all users combine to form system-level dynamics.

One particular problem in understanding users' activity dynamics in peer production systems lies in explicitly considering *excitation* (a measure for strength of influence) within and between groups or types of users, how such excitations shape the development of the whole system and how those excitations evolve over time. Consider Figure 1.1: Without loss of generality, assume that there are two groups of users in a peer production system, e.g., a majority of users who contribute little and a minority of very active power users. There are within-group and between-group relationships, whose strength may vary over time and thereby shape activity dynamics of users in a group, and, in turn, of the peer production system as a whole. Such interactions and influence between and within groups are ubiquitous. Yet, because they are often unobserved, incorporating them into models for activity dynamics poses a challenge. While one line of previous research [Mam+11; Fur+13; Gei+19] distilled user groups in successful Q&A communities and the other [Kit+07; Suh+09] uncovered the prominence of certain user groups at different times of Wikipedia's development, there is an opportunity to combine and

extend both lines of inquiry by focusing on how excitations within and between groups evolve to shape system-level dynamics.

Beyond user excitation dynamics, this thesis concentrates on temporal signatures of activity in peer production systems as implicit indicators for the development of user groups as well as the system itself. Doing so facilitates (i) estimations of activity dynamics of a system in general and of users in particular, and (ii) the incorporation of that knowledge in models for the activity dynamics in not only successful but also failing peer production systems.

1.3 Problem Statement, Objectives and General Approach

Problem statement. Peer production systems are omnipresent forms of large-scale collaboration on the web. How and why some systems gain traction and evolve to attract and sustain activity from large numbers of users is a subject of ongoing research. In particular, there is a research gap in linking temporally evolving user behavior and excitations in a peer production system to the system's development as a whole. Partial observability of excitations and influence between users compounds the difficulty in incorporating such factors in models for activity dynamics.

Objectives. First, this thesis aims at improving the modeling of activity dynamics in peer production systems. The second main objective of this thesis consists in, with improved modeling tools, deriving actionable insights for peer production system managers to better understand the systems they oversee. Combined, addressing both objectives poses a first step towards not only better understanding and modeling but also steering and promoting activity dynamics in peer production systems. Further, the methods presented and developed within the scope of this thesis also target generalization in their applicability across peer production system types and beyond.

General approach. I propose leveraging temporal signatures of activity both at the user-level as well as at the level of whole systems to uncover factors linked to growth in activity of peer production systems. To do so while coping with nonlinearity, burstiness and randomness in activity dynamics, I employ a broad range of methods from the temporal analysis toolbox, including nonlinear time series analysis, time series clustering, descriptive statistics and Bayesian inference, and, prominently, Hawkes processes, a class of stochastic point processes.

1.4 Research Questions

This thesis subdivides the study of activity dynamics in peer production systems into three research questions. First, I explore characterizations of user activity patterns and system-level activity dynamics in peer production systems. Leveraging that characterization, the second research question asks how can user-level excitation and activity dynamics be modeled to understand whole systems. The third and final research question focuses on deriving, from the previously uncovered temporal patterns, empirical knowledge and actionable insights about the evolution of peer production systems.

Table 1.1 lists the articles which form this thesis, as well as the corresponding research questions, topics and main contributions.

RQ1: How can we characterize user activity dynamics in peer production systems?

Problem. To model complex patterns in the activity dynamics of peer production systems, it is necessary to first explore and grasp basic characteristics of activity patterns. On a high-level, previous work [Rib14; Wal+16] postulated nonlinear dynamical systems-based descriptions for activity dynamics in several peer production systems. However, the kinds of nonlinear dynamics to capture may need to be a choice best left to the modeler, as there may be system-specific characteristics to account for. Further, while nonlinear dynamical systems may capture global developmental trends of a peer production system, there may be settings where stochasticity dominates perhaps nonlinear chaotic behavior. Distinguishing between both settings is a open problem, and surfacing the caveats of committing to a specific model setting may support practitioners in improving high-level forecasts of activity dynamics. Where stochasticity dominates observed global dynamics of peer production systems, related research [Kle03; Bar05; CFL09] suggests likely causes: burstiness in user behavior, and heavy-tailed distributions of user activity. The resulting heterogeneity in user behavior leads to previous work [Mam+11; Fur+13] grouping users into roles, which typify users as a set of activity patterns. However, the connection between the user type mix and the overall activity dynamics of peer production systems is not well established. In particular, there is a need to dynamically characterize which user type compositions can be linked to systems with thriving levels of user activity.

Approach. To characterize the presence of nonlinear (perhaps chaotic) or rather stochastic activity dynamics, the analysis presented in [SWH17] leverages Takens' theorem [Tak81] and nonlinear time series analysis to reconstruct and study state spaces describing the activity dynamics of 16 Stack Exchange Q&A communities. For the analysis of stochasticity in user activity, the research [San+19a] resorts to clustering features extracted from time series describing user activity in a random sample of 50 Stack Exchange Q&A communities. Repeatedly applying this clustering procedure in regular intervals following the inception of each Q&A community uncovers how user mix relates to the community's development in activity.

Findings and contributions. My detailed exposition on how to tailor nonlinear time series analysis and time series clustering methods to characterize user activity dynamics in peer production systems directly addresses this first research question. The results of the application of these methods reveal some Q&A communities feature rather nonlinear dynamics, for which forecasts from reconstructed dynamical systems work best, and others with rather stochastic dynamics. The characterization of users in Q&A communities uncovers a set of four user types. These results indicate that (i) a parsimonious group of features suffices to characterize bursty user behavior, (ii) the main distinction in user types is on the overall activity levels, which can be coarsely defined as high or low, and (iii) the user mix evolves as the systems mature.

RQ2: How can we model users' evolving excitation in peer production systems?

Problem. Beyond the previous findings that nonlinearity and stochasticity play an important role in user behavior, and that a discrete set of user activity patterns may relate to overall peer production system development, one key ingredient is missing to improve models for activity dynamics in peer production systems: user excitation. Users do not contribute in isolation to a system, but rather react to their peers. However, capturing excitation, as a proxy measure for inner motivation and between-user influences, is a non-trivial task, as the strength of such relationships between users and user types cannot be directly quantified in observational studies, and would require e.g. an interventional or qualitative experimental setup. Further, the presence of exogenous shocks (such as interventions in the form of interface changes), exponential growth or other non-stationary phenomena common to peer production systems should not impair modeling efforts of excitation in particular and temporal patterns in general. Overall, accounting for temporally evolving user excitation in activity dynamics models for peer production systems holds the promise of strengthening the findings mentioned in RQ1, as theoretical work [Ete+16; XFZ16; EDD17] established a link between certain forms of excitation and Granger causality, a form of temporal causality. Therefore, the focus of this research question is to devise an approach to quantify, in observational studies, temporally evolving excitation in users of peer production systems.

Approach. In a first methodological step to address this research question [San+19c], this thesis proposes combining an exponentially decaying Hawkes process, which captures excitation via sequences of event timestamps, with time series structural change estimation models, to ensure theoretical assumptions are met when fitting Hawkes processes longitudinally. To improve the applicability of such Hawkes processes to challenging contexts common in activity dynamics of peer production systems and beyond, this thesis also characterizes and tackles inherent difficulties in Hawkes process estimation within a classical Bayesian framework [SLH20].

Findings and contributions. One key finding of [San+19c] is that estimating periods of stationarity (i.e. distributional translation-invariance) through time series structural change estimation allows to fit Hawkes processes longitudinally, even across non-stationary phenomena such as activity dynamics with exponential growth. That same work also finds that, despite difficulties in fitting the decay parameter of Hawkes processes with exponential kernels, a common off-the-shelf approach suffices for downstream tasks. Follow-up work [SLH20] then characterizes, formulates and mitigates this decay parameter fitting problem for a range of synthetic and real-world settings, thereby contributing a parsimonious approach to estimate Hawkes processes in the context of activity dynamics in peer production systems and beyond.

RQ3: How do users' excitation and temporal patterns shape the evolution of peer production systems?

Problem. Given the first results on activity patterns in peer production systems and previous models of temporally-changing excitation, this thesis is now in a position to address the empirical issue of estimating and understanding how excitation and user temporal patterns longitudinally impact (peer production) system-level activity dynamics. Uncovering the link between user excitation, user type mix and success and failure (in terms of activity dynamics) of peer production systems, and how that link changes over time, is a challenge due to the previously mentioned unobservability of user excitation. Further, if such a link exists, it is not clear whether it is of practical relevance: Can the knowledge of user excitation and user temporal patterns help improve e.g. activity prediction experiments? Further, peer production system managers often perform interventions to steer the systems they oversee towards goals such as improving new user onboarding processes. Measuring how the users react to such external shocks and thereby shape new activity dynamics of the whole system is also of practical relevance. Finally, given that tapping the wisdom of the crowds represents one key promise of creating, developing and maintaining activity dynamics in large-scale peer production systems [Kit+07; Suh+09], there is also the issue of estimating when and under which circumstances the wisdom of the crowds matches or even surpasses that of the few, and how interactions between these two user types evolve and shape each other over time. Further, while previous research [Gil05; Sur05] suggests crowds perform comparably to experts in knowledge production, there is still ongoing debate (cf. e.g. [Che+14]) on comparing wisdom of the crowds vs. the few in the broad class of peer production systems dedicated to matters of personal experience and opinion rather than matters of fact.

Approach. The methods to address this research question build on the Hawkes process models outlined in the previous research question [San+19c]: Longitudinally fitting Hawkes processes to highly and lowly active users in growing and declining Stack Exchange Q&A communities enables the study of excitation effects over time. Statistical and permutation tests establish the significance of excitation effects, and a range of prediction experiments validate their practical usefulness for community managers. To estimate short- and long-term effects of an intervention to improve user onboarding practices in Stack Exchange Q&A communities, [San+20] proposes a difference-in-difference regression to isolate the impact of the intervention on user behavior with respect to seasonality and temporal trends. With a longitudinal descriptive and predictive analysis of video game reviews on Metacritic, the study [San+19d] performs a first step towards understanding (i) how wisdom of the crowds emerges in peer production systems dedicated to an experience good, (ii) how the wisdom of the crowds compares to that of the few, and (iii) how both the crowds and the few change and interact over time.

Findings and contributions. The results of the Hawkes process applied to Q&A communities reflect a relationship between user excitation and future activity developments of said communities. For example, high excitation by power users as a reaction to casual users plays a pivotal role in early developmental stages of successful communities when compared to declining ones. In successful communities, the importance of casual

user excitation increases while that of power user excitation decreases. Given that these excitation effects appear contingent on the topic of the Q&A community itself (where exemplary topical groups may be STEM or humanities), community managers may carefully want to manage and correctly time initiatives to promote participation of different user types. On the topic of promoting user participation, the intervention designed to encourage newcomers was found to have enacted a temporary effect, which, interestingly, appears independent of community topic. This finding may encourage peer production system managers to replicate such "one-size-fits-all" interventions and thereby counter challenges in managing previously mentioned topical differences. On the topic of the comparison of the wisdom of the crowds vs. the few in the context of a peer production system focusing on experience goods, namely video games, there appear to be key discrepancies between the two, especially on the level of temporal reviewing patterns. Nevertheless, prediction experiments for the future reception of video games indicate that combining both kinds of user input yields best predictive performance. This supports the conclusion that both kinds of wisdom combine to form the most complete views, and that catering to both types of users is conductive to positive outcomes in the evolution of peer production systems.

1.5 Main Publications

This cumulative thesis consists of the following publications:

- **Article 1:** [SWH17] Santos, T., Walk, S., and Helic, D. Nonlinear Characterization of Activity Dynamics in Online Collaboration Websites. *WWW Companion.* 2017

- **Article 2:** [San+19a] Santos, T., Walk, S., Kern, R., Strohmaier, M., and Helic, D. Activity Archetypes in Question-and-Answer (Q&A) Websites – A Study of 50 Stack Exchange Instances. *ACM TSC.* 2019

- **Article 3:** [San+19c] Santos, T., Walk, S., Kern, R., Strohmaier, M., and Helic, D. Self- and Cross-Excitation in Stack Exchange Question & Answer Communities. *WWW.* 2019

- **Article 4:** [SLH20] Santos, T., Lemmerich, F., and Helic, D. Estimating the Decay Parameter of Hawkes Processes with Exponential Kernels. *Submitted to ICDM.* 2020

- **Article 5:** [San+20] Santos, T., Burghardt, K., Lerman, K., and Helic, D. Can Badges Foster a More Welcoming Culture on Q&A Boards?. *ICWSM.* 2020

- **Article 6:** [San+19d] Santos, T., Lemmerich, F., Strohmaier, M., and Helic, D. What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic. *PACM HCI (CSCW).* 2019

1.6 Further Publications

In the course of my PhD studies, I also worked on the following publications:

- **Article 7:** [SK18] Santos, T., and Kern, R. Understanding wafer patterns in semiconductor production with variational auto-encoders. *ESANN*. 2018
- **Article 8:** [San+19b] Santos*, T., Schrunner*, S., Geiger, B.C., Pfeiler, O., Zernig, A., Kaestner, A., and Kern, R. Feature Extraction From Analog Wafermaps: A Comparison of Classical Image Processing and a Deep Generative Model. *IEEE TSM*. 2019
- **Article 9:** [Kas+19a] Kasper, P., Koncar, P., Walk, S., Santos, T., Wölbitsch, M., Strohmaier, M., and Helic, D. Modeling User Dynamics in Collaboration Websites. *Book chapter in Dynamics On and Of Complex Networks*. 2019
- **Article 10:** [TSK19] Toller, M., Santos, T., and Kern, R. SAZED: parameter-free domain-agnostic season length estimation in time series data. *Data Mining and Knowledge Discovery*. 2019
- **Article 11:** [Kas+19b] Kasper*, P., Koncar*, P., Santos*, T., and Gütl, C. On the Role of Score, Genre and Text in Helpfulness of Video Game Reviews on Metacritic. *SNAMS*. 2019
- **Article 12:** [RSH19] Ruprechter, T., Santos, T., and Helic, D. On the Relation of Edit Behavior, Link Structure, and Article Quality on Wikipedia. *International Conference on Complex Networks and Their Applications*. 2019
- **Article 13:** [Kou+20] Koutroulis, G., Santos, T., Wiedemann, M., Faistauer, C., Kern, R., and Thalmann, S. Enhanced Active Learning of Convolutional Neural Networks: A Case Study for Defect Classification in the Semiconductor Industry. *Submitted to Discovery Science*. 2020
- **Article 14:** [Hop+20] Hopfgartner, N., Santos, T., Auer, M., Griffiths, M., and Helic, D. Social Facilitation Among Gamblers: A Large-Scale Study Using Account-Based Data. *Submitted to ICWSM*. 2020

The asterisk denotes authors with equal contribution.

1.7 Contributions and Implications

This thesis features two kinds of contributions on understanding and modeling activity dynamics in peer production systems, namely on the empirical level and on the methodological level.

Specifically, the main contributions of this thesis are as follows:

- This thesis characterizes nonlinear and bursty temporal signatures in system-level and user-level patterns of activity in peer production systems, thereby paving the way for extending activity dynamics models. Further, this characterization suggests a temporal link between user mix and developmental stages of whole systems.
- With methodological extensions to Hawkes processes as excitation models, this thesis contributes parsimonious tools for more easily applying them in practice. This can be seen as a stepping stone towards increasing the adoption of Hawkes process models by researchers and practitioners interested in learning from temporal phenomena in a wide range of application scenarios.
- The empirical study of excitation in peer production systems uncovers the importance of correctly timing the user mix to maximize the activity potential of developing systems. This thesis also provides insights into the power of interventions in shaping activity dynamics, and specifies circumstances under which the wisdom of the crowds complements that of the few.

Overall, this thesis provides another stepping stone for modeling and thus improving activity dynamics models for peer production systems, and, by deriving actionable insights from their empirical application to peer production systems, supports managers of said systems to optimize their activity development efforts.

1.8 Structure of this Thesis

Following this introduction, I review related work in Chapter 2. Specifically, I position this thesis against other work in the space of research on peer production systems (cf. Section 2.1) and on Hawkes processes (cf. Section 2.2).

Chapter 3 comprises the main articles which combine to form this thesis. In particular, Section 3.1 describes my contributions to each of the main articles. Refer to Figure 1.2 for a graphical summary of which articles address which research questions.

This thesis concludes, in Chapter 4, with a reiteration of the main findings and contributions of this body of work in Section 4.1, an outline of potential implications in Section 4.2, a discussion of limitations in Section 4.3 and a reflection on avenues for future work in Section 4.4.

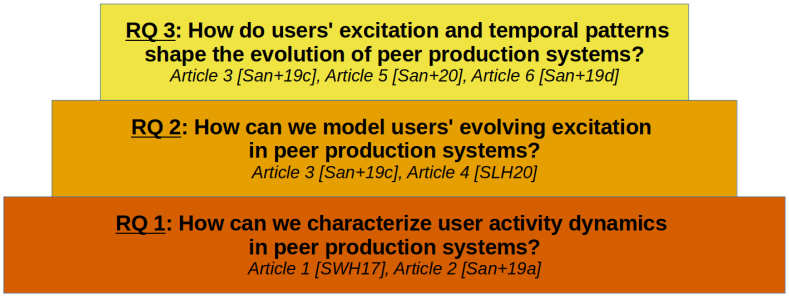


Figure 1.2: **Structure of this thesis.** This figure outlines how this thesis' main articles combine to form answers to the corresponding research questions (RQ). RQ 1 asks about a characterization of system-level and user-level activity dynamics in peer production systems. The answers to RQ 1 underlie RQ 2, which focuses on how to model user excitation as it changes over time. Building on those two research questions, RQ 3 concerns the analysis of how uncovered temporal patterns empirically relate to the evolution of peer production systems.

Table 1.1: **Context of main articles.** This table summarizes the main articles which form this thesis, the research questions they address, and their topics and main contributions.

Article	**RQ**	**Topic**	**Main Contribution**
Article 1 [SWH17]	RQ 1	Characterization of activity dynamics	Uncovering nonlinearity in activity patterns
Article 2 [San+19a]	RQ 1	Characterization of activity dynamics	Clustering users' bursty activity patterns and linking them to system-level developmental stages
Article 3 [San+19c]	RQs 2+3	Modeling and learning from excitation	Fitting Hawkes processes in the presence of non-stationary dynamics, and leveraging them to understand excitation patterns and to establish temporally causal relations to the development of whole systems
Article 4 [SLH20]	RQ 2	Modeling excitation	Characterizing, formulating and mitigating problems in fitting Hawkes processes
Article 5 [San+20]	RQ 3	Assessing interventions	Quantifying the impact of interventions to shape user interactions and activity dynamics of whole peer production systems
Article 6 [San+19d]	RQ 3	Wisdom of the crowds	Juxtaposing the wisdom of the crowds with that of the few and understanding their temporal properties and interactions

2 Related Work

This chapter reviews two main streams of research related to this thesis: empirical studies on the properties and (activity) dynamics of peer production systems (cf. Section 2.1), and theory and applications of Hawkes processes to modeling temporal phenomena in general and in peer production systems in particular (cf. Section 2.2). This chapter intends to provide an overview on both streams of research and how they relate to this thesis; more details on related work can be found in each of the thesis' main articles.

2.1 Peer Production Systems

As the population of Internet users increases remarkably, peer production systems, as socio-economic systems for content creation, curation and sharing [BN06], feature collaborative work on ever-growing scales. Several peer production systems dedicated to collaboration on a broad range of topics achieved massive scales: Wikipedia, an online encyclopedia, GitHub, an open source software development platform, Reddit, a social news aggregator, and Stack Exchange, a network of online communities dedicated to questions-and-answers. These successful systems have been the subject of much research. This thesis relates to studies of activity dynamics of peer production systems on a macro-scale (i.e., on the level of the whole system), and a micro-scale (i.e., on the level of user behavior), as well as to research on the properties of crowd-collaborative efforts, or, in other words, the wisdom of the crowds.

2.1.1 System-level Activity Dynamics

Early work in the research space of broad properties of peer production systems includes Wilkinson's [Wil08] study of power law-like user tenure and lognormal content contribution patterns in Wikipedia and other software-related and social peer production systems. This work can be framed within more general studies [Bar05] of power law-like dynamics of human behavior. This thesis acknowledges that the presence of such universal properties permeates analyses more focused on the success, in terms of activity dynamics, of peer production systems.

Turning to more granular studies of the activity dynamics of peer production systems, previous work in this space leveraged models based on networks [ZAA07; Wan+13; ZCF16] or dynamical systems [Rib14; MSF15; Wal+16] to describe activity in peer production systems. Specifically, Zhang et al. [ZAA07] construct and analyze a network of user interactions in a Java programming forum, relating structural properties of the network to overall activity dynamics and users' topical expertise. Wang et al.'s [Wan+13]

study of the social, topical and knowledge item graphs of Quora, a large Q&A community, reveals that heterogeneity in user characteristics drives users' attention and activity. The NetTide model proposed by Zang et al. [ZCF16] captures the growth of social activity and connections in WeChat, a large social network, and in arXiv, a system for sharing pre-prints of manuscripts. In relation to this stream of work, this thesis recovers many of the same high-level dynamics these previous studies observed. Furthering those observations, this thesis also seeks to identify discrete growth phases in global activity dynamics of peer production systems, and, by characterizing them, identify factors related to the success or failure of a system, such as, prominently, the importance of timing in the user type mix. This thesis' measurements of such factors complement qualitative work on community lifecycles [IL09; You13]. Further, this thesis also intertwines global activity dynamics with other user-level properties, such as, notably, user excitation.

In the literature leveraging dynamical systems descriptions for activity dynamics, Ribeiro [Rib14] devises a model which captures growth and death of a range of membership-based websites in terms of active users. The author characterizes website growth as result of word-of-mouth adoption or of marketing campaigns, and as sustainable or not. With a comparable modeling technique, Matsubara et al. [MSF15] also highlight competition between activities drive user behavior online. The model proposed by Walk et al. [Wal+16] leverages a dynamical system operating on a network to identify collaboration networks which become self-sustainable, i.e., reach a certain level of activity which makes systemic failure unlikely (in other words, critical mass [RMJ10; SW14]). These works share an objective with this thesis, namely to identify and characterize the dynamics of successful and failing peer production systems and identify factors which lead to diverging activity levels and system-level outcomes. The results presented in this thesis agree, on a high level, with those of this stream of work. However, this thesis uncovers that not all kinds of activity dynamics in peer production systems may be well-captured via non-linear dynamical systems, as it appears stochasticity dominates in some cases. Further, the nonlinear characterization presented in this thesis may also support future endeavors to improve dynamical system-based modeling of activity dynamics, as nonlinear time series analysis helps reconstruct properties which a dynamical system-based description of activity dynamics should have.

2.1.2 User-level Activity Dynamics

To understand activity dynamics of users in peer production systems, previous work typically introduces a discrete set of user types, a simplification which counters heterogeneity in user behavior. First, I review such studies mostly in the space of the analysis of Q&A communities, as those are one of the kinds of peer production system this thesis focuses on. One of the first such studies in this space was performed by Mamykina et al. [Mam+11] on Stack Overflow, a Stack Exchange Q&A community dedicated to programming questions. That mixed-methods study proposed a set of four user types to discretize a continuum of user activity: community activists, shooting stars, and low-profile and inactive users. The prevalence of each user type is also inversely proportional to their activity levels, as the community activists comprise only 1% of the total user

population, whereas low-profile users account for more than 94% of the user base. In particular, this study also highlights the importance of recognizing user activity is bursty, a finding which Vázquez et al. [Váz+06] and Kushner and Sharma [KS20] corroborate in human activity in general and, respectively, in mental health communities in particular. This thesis arrives at comparable conclusions when extending an activity-based clustering to more Stack Exchange Q&A communities. However, one key difference to that work is that, as this thesis focus more on differences in burstiness and overall activity levels, it appears that the more predominant distinction between user types can be delineated between the very active but small core of power users and the large mass of lowly active casual users.

Sinha et al.'s [SMG13] user type characterization agrees with the previously mentioned dichotomy, and Gilbert's study of underprovision on Reddit [Gil13] and other work [ZAA07; Yan+14] focused on identifying experts in Stack Exchange communities suggest a similar binary user type distinction. However, there are diverging perspectives: Furtado et al. [Fur+13] offer a more nuanced categorization of users of five Stack Exchange Q&A communities. The ten user types they distill span not only differences in overall activity levels, but also quality of contributions. The authors also arrive at the conclusion that the user mix is stable over time, that is, the percentual composition of user types remains unchanged even as a community develops. This thesis' longitudinal study of user types and user mix in 50 Stack Exchange Q&A communities paints a partially different picture: The user mix changes over time depending on the kind of community (one of which is the one analyzed by Furtado et al.). Further, I also extend those previous findings by connecting the prevalence of each user type to specific developmental stages of a community.

On a high level, this connection of user types to developmental stages of Q&A systems appears to generalize to other peer production systems as well, as Kittur et al. [Kit+07] and Suh et al. [Suh+09] report similar observations in Wikipedia, much like Chung et al. [CPU12] in their study of a forum for offline community building in Australia. Specifically, those authors and this thesis report a reliance on power users in the early developmental stages of a peer production system, which gradually gives way to more casual user-centric activity. In this regard, this thesis is more closely related to Chung et al.'s work, as both cover aspects of user interaction: Those authors find overall disassortative interaction networks whose assortativity increases over time. The user excitation analysis in this thesis also captures higher cross-excitation between power and casual users in early stages, an excitation which later on transitions to more within-group excitation (in particular casual user excitation). This thesis extends previous work by estimating points in time by when it may be expected to observe such transitions, and it also indicates there may be limited generalizability of such results across community topic (e.g. between Q&A communities dedicated to topics in STEM vs. humanities). Last but not least, those previous studies considered only growing communities, while this thesis' analysis shows there are key excitation and user mix differences between Q&A communities with increasing and those with declining levels of activity.

Understanding and measuring differences in activity dynamics enables peer production

system managers to timely assess and react to declining levels of activity, or to promote the growth of thriving systems even further. The tools managers have at their disposal to steer activity include introducing interface or platform changes. Previous research has shown that prominently highlighting a community on Reddit [Lin+17] or software repository on GitHub [Mal+20] may channel an influx of users and contributions to highlighted systems. Another prominent set of tools available to peer production systems are badges, which, according to previous studies [And+13; KGR18; Yan+19b] of Stack Exchange Q&A communities, may incentivize participation. That body of work inspected the effect of badges only on users with a history of activity, and provides initial evidence for the interplay between user-specific effects and community-wide benefits. This thesis addresses this research gap with a study of a badge-like indicator to welcome newcomers to Stack Exchange communities, as supporting and integrating newcomers to a peer production system is crucial to mitigate user churn and support community growth [All06; Yan+10; KR12; See+20]. Similarly to previous studies of badges, and in particular Anderson et al.'s [And+13] and Yanovsky et al.'s [Yan+19b] findings that users increase work towards attaining a badge and then revert to previous behavior, I attribute a temporary benefit in newcomer retention and community reactions to the introduction of the indicator.

2.1.3 Wisdom of the Crowds

In this Subsection, I review work which, in the context of peer production systems, connects activity dynamics of the crowd to the performance of the crowd. One prominent reason for analyzing this connection lies in the promise that peer production systems provide a platform for the wisdom of the crowds to manifest its potential, which seminal work [Gal07; Sur05; Che+14] on general properties of the wisdom of the crowds indicates. In the context of peer production systems, Wilkinson and Huberman [WH07] and Murić et al. [Mur+19] identified a net qualitative benefit to collaboration and contributions by the crowds in Wikipedia and GitHub. However, previous studies [Bur+17; DKS19] of user voting patterns in Stack Exchange provide evidence for user biases such as a preference for answers at the top of a page, and Robert and Romero [RR15] also claim a dependency of crowd performance in certain projects on Wikipedia on the size and diversity of the crowd. Therefore, there is a need to evaluate under which conditions the crowd's input and judgments can be trusted. To that end, previous work proposed comparing the wisdom of the crowds with the wisdom of the few, i.e. experts in a given field. Although there are objective measurements on how the wisdom of the crowds compares to the wisdom of the few in online knowledge communities (cf. e.g. Giles' juxtaposition of Wikipedia with Encyclopædia Britannica [Gil05]), there are a vast number of peer production systems dedicated to matters of opinion, such as discussing and curating experience goods like books, movies or video games. This thesis addresses the research gap in how the crowds perform in comparison to experts in a peer production system dedicated to reviews of video games. I find the activity dynamics of experts differ from that of the crowd, but, in contrast to previous research suggesting that the crowds perform comparably if not better than experts in online knowledge communities [Gil05], my analysis suggests they

are complementary in this setting of views on experience goods.

2.2 Hawkes Processes

I begin with a brief overview of Hawkes process theory, which is followed by a review of applications and recent advances related to this thesis.

2.2.1 Theoretical Background

A point process is a set of points randomly distributed across some mathematical space, which, in the case of temporal point processes, is $\mathbb{R}^+$ and represents continuous time. For more on point processes and temporal point processes in general, please refer to [CI80; DVJ03; DVJ08] and, respectively, to [ABG08].

This thesis employs and studies a particular class of temporal point processes, namely the Hawkes process [Haw71]. Intuitively, Hawkes processes model the arrival times of incoming events of interest, such as, for example, the timing of tweets by a user on Twitter. Further, Hawkes processes impose a dependency of future event timings on past events. In the example of the Twitter user, this corresponds to the assumption that past tweets motivate the user and perhaps increase her future event rate. Mathematically, Hawkes processes are characterized via a conditional intensity function $\lambda^*(t)$, which expresses a dependency of the event rate of a stochastic process on its event history. Given a set of n events occurring at times $t_i \in \mathbb{R}^+$ (which, again, could be the timings of tweets) the conditional intensity of a Hawkes process is

$$\lambda^*(t) = \mu + \alpha \sum_{t_i \in \mathcal{H}(t)} \kappa_\beta(t - t_i). \tag{2.1}$$

The meaning and interpretation of each parameter in that equation is as follows: $\mu \in \mathbb{R}^+$ is the *baseline* intensity, which is independent of the event history. In fact, setting $\lambda^*(t) = \mu$ results in a Poisson process, which is a special case of the Hawkes process. In the example of the Twitter users, baseline intensity corresponds to a minimal predisposition to randomly tweet at any point in time, regardless of how many Tweets a user has previously written. $\alpha \in \mathbb{R}^+$ captures *excitation*, i.e. magnitude of an increase in $\lambda^*(t)$ at each event time t_i. In the one-dimensional case, the excitation is termed *self-excitation*, since each event increases intensity in that same dimension, while in the multi-dimensional case *cross-excitation* may occur, i.e., an event in one dimension may lead to an intensity increase in another. In the ongoing example of the Twitter user, self-excitation captures an increase in motivation to tweet following own tweets, while cross-excitation models a motivation to tweet more after tweets by another user (not necessarily a followee). Self- and cross-excitation are amenable for interpretation as directions of influence among and within dimensions of a Hawkes process, as non-zero excitations in Hawkes processes with an exponential kernel capture Granger causality between dimensions [Ete+16; XFZ16; EDD17], and the magnitude of the excitation corresponds to the strength of influence. Following each event i at time t_i, the conditional intensity decreases according to the

kernel κ_β, which is typically chosen to be an exponential or power-law function, or a mix of basis functions in more complex cases. This thesis focuses on Hawkes processes with exponential kernels. This kernel reflects e.g. forgetting rates of users on Twitter, which, through the exponential kernel assumption, follow an (exponential) curve with decay rate β. Interpreting real-world temporal phenomena through Hawkes process parameters requires estimating their values via maximizing their log-likelihood given event data and a value for β. Note that (fitted) Hawkes processes assume stationarity, i.e. distributional invariance of the process over time. Consult Liniger [Lin09] for a formal treatment of multivariate Hawkes processes.

2.2.2 Applications and Recent Advances

In practical applications, the power of Hawkes processes lies in their ability to surface typically unobservable properties, namely excitation and directions of influence, in temporal phenomena. Moreover, to do so, Hawkes processes parsimoniously require only knowledge of timestamps of events. Both properties underlie the widespread adoption of Hawkes process models to study complex temporal phenomena. One of the first application areas includes seismology: Ogata et al. [OAK82; Oga83] extend Hawkes processes to numerically quantify then-suspected geological relationships between regions in Japan and New Zealand, a step which enabled later predictions of earthquakes in certain regions following earthquakes in given other regions. In similar fashion, research in the economics and finance sought to estimate the strength and timing of follower-followee relationships in e.g. American and European stock markets [ASCDL15] or trade clustering patterns [DFZ14]. Though I also aim at uncovering hidden patterns of interaction and excitation much like that body of research in seismology and finance, this thesis relates more closely to Hawkes process-based studies of temporal phenomena in web communities and peer production systems. Early work in this stream of research includes Gomez-Rodriguez et al.'s [GRLK10] theoretically-grounded algorithm to infer networks of information flow through the web by leveraging the times of adoption of information. Later, in the same application context, research such as that by Zhou et al. [ZZS13a] extended such algorithms to infer hidden networks via a multi-dimensional low-rank Hawkes processes. While that body of work, similarly to this thesis, extracts influence and excitation relationships, that body of work does so at the level of the whole web. The object of study of this thesis are more granular dynamics, namely at the level of single peer production systems and their users. In that sense, this thesis is more closely related to Hawkes process-based studies of user behavior in e.g. online marketplaces [HF20], in health-related apps against the backdrop of offline activities [KAL18] and in Q&A [KGR18] or social network [Jun+19] communities, as those works focus on interpreting user behavior in terms of the learned Hawkes processes. To elaborate on the relation between this thesis and that stream of literature, consider Junuthula et al.'s work [Jun+19], which marries the stochastic block model from network science with Hawkes processes to estimate and interpret excitation and other parameters in user behavior on Facebook. Similarly to their work, I focus on estimating community-level user groups and their excitation relationships, but beyond their work, I extend the estima-

tion of such relationships longitudinally across a peer production system's development. Estimating longitudinal user patterns in peer production systems with Hawkes processes is also the topic of Upadhyay et al.'s [UVGR17] and Mavroforakis's [MVGR17] work, as those authors focus on estimating learning curves of users in Q&A communities. Those authors track a user's progress through her tenure in a community as she gains knowledge on a variety of topics, and they find users can be assigned to two groups with more or less knowledge and propensity to acquire it. In contrast to their work, this thesis focuses not on the content of user actions but rather their activity dynamics, and this thesis also relates those user patterns to the development of the system, rather than studying user patterns in isolation.

However, although Hawkes processes afford to uncover many temporal phenomena of practical interest, there are some important caveats to their application. First, there is no consensus on optimizing the decay parameter of the commonly-used Hawkes process with an exponential kernel, as, in contrast to the other parameters, the Hawkes process log-likelihood function is non-convex in the decay. As such, previous work proposed a broad palette of approaches to estimate the decay parameter: using a given constant [Far+15; Bac+18; Du17; Cho+18; LWK18], cross-validating a range of values [Far+14; Cho+15; Sal+19], or estimating them with a range of different optimization approaches [Oza79; DFZ14; Bac+16; UVGR17; KAL18; Fig+18; San+19c]. This thesis compares them, characterizes difficulties in estimating the decay parameter, and extends those point estimates with a Bayesian framework to provide estimation uncertainty. Further, real-world data does not always fulfill the stationarity requirement, due to phenomena such as exponential growth in the number of events or exogenous shocks to a system under study. Previous work which studied the dynamics of the popularity of YouTube videos [Riz+17], and one of the earlier articles of this thesis [San+19c], resorted to additional assumptions to cope with stationarity violations, such as estimating stationary periods via ad-hoc changepoint models. This thesis' Bayesian framework for estimating the decay parameter copes with such issues by allowing for intractable setups which explicitly capture such phenomena.

Orthogonal to the Hawkes process framework proposed in this thesis, there is a trend towards non-parametric estimation of temporal point process kernels [LM11; ZZS13b; BM16; EDD17; ZWR20], as well as towards fusing deep learning with Hawkes processes [Du+16; Cao+17; TWS19; Sal+19]. Such approaches are powerful alternatives to the parametric approach this thesis focus on, and the Bayesian framework presented in can be understood as an interpretable, parsimonious counterpart to such methods.

3 Papers

3.1 Contributions to the Main Articles

This section details *my* contributions to the main articles of this cumulative thesis.

- [SWH17] Santos, T., Walk, S., and Helic, D. Nonlinear Characterization of Activity Dynamics in Online Collaboration Websites. *WWW Companion*. 2017

All three authors designed the research, i.e., conceptualized the idea to study nonlinear properties and the feasibility of dynamical system reconstruction of activity dynamics. As the primary author of this article, I tailored the nonlinear time series analysis tools to the objectives of this research, compiled the data, analyzed the data, performed the research and interpreted the results. All three authors wrote the paper.

- [San+19a] Santos, T., Walk, S., Kern, R., Strohmaier, M., and Helic, D. Activity Archetypes in Question-and-Answer (Q&A) Websites – A Study of 50 Stack Exchange Instances. *ACM TSC*. 2019

For this article, Denis Helic advanced the idea to study activity dynamics in Q&A communities as a function of user behavior. As the primary author of this article, I tailored the time series extraction and clustering tools to the objectives of this research, compiled the data, and performed the research. All authors analyzed, discussed and interpreted the data and the results, and contributed towards writing the paper.

- [San+19c] Santos, T., Walk, S., Kern, R., Strohmaier, M., and Helic, D. Self- and Cross-Excitation in Stack Exchange Question & Answer Communities. *WWW*. 2019

The main idea for this article, i.e. to understand excitation effects in the activity dynamics of Q&A communities, was proposed by myself and conceptually refined in discussions with Simon Walk and Denis Helic. As the primary author of this article, I designed and developed the Hawkes process-based methods to address the objectives of this research, compiled the data, analyzed the data, and performed the research. My tasks also included open-sourcing the code for this manuscript[1]. The interpretation of the results was done mainly by Simon Walk, Denis Helic and myself. All authors wrote the paper.

[1] `https://www.github.com/tfts/excitation_in_QA`

- [SLH20] Santos, T., Lemmerich, F., and Helic, D. Estimating the Decay Parameter of Hawkes Processes with Exponential Kernels. *Submitted to ICDM*. 2020

I devised the premise of this study, namely to understand and mitigate difficulties in the estimation of the decay parameter of Hawkes processes with exponential kernels. To address the objectives of this research, I designed the approach, a Bayesian framework to learn about the decay parameter, with the help of Denis Helic. As the main author of this article, I compiled the data, analyzed the data, performed the research and interpreted the results. All three authors wrote the paper.

- [San+20] Santos, T., Burghardt, K., Lerman, K., and Helic, D. Can Badges Foster a More Welcoming Culture on Q&A Boards?. *ICWSM*. 2020

Understanding the effects of an intervention to welcome newcomers to Stack Exchange Q&A communities was an idea which stemmed from discussions among all authors. As the primary author of this article, I adjusted interrupted time series analysis methods to address the objectives of this research, compiled the data, analyzed the data, and performed the research. All authors contributed to the tasks of both interpreting the results and drafting the manuscript.

- [San+19d] Santos, T., Lemmerich, F., Strohmaier, M., and Helic, D. What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic. *PACM HCI (CSCW)*. 2019

I first conceived this work as a study of differences in expert and amateur reviews of video games, and all authors discussed and re-framed this idea as a juxtaposition of the wisdom of the few (experts) and the wisdom of the crowds (amateurs). Again, as the main author of this article, I was responsible for all methodological aspects of this article (in particular the statistical analyses and application of machine learning algorithms), and for compiling and analyzing the data, performing the research and interpreting the results. Florian Lemmerich helped to devise baseline algorithms for the prediction experiments, and Denis Helic gave methodological support in the application of the Latent Dirichlet Allocation algorithm. All authors wrote the paper.

3.2 Nonlinear Characterization of Activity Dynamics in Online Collaboration Websites

This first article addresses the first research question, which focuses on a characterization of user activity dynamics in peer production systems. This article proposes a battery of statistical tests to assess nonlinearity in the activity dynamics of a peer production system, and compares the performance of different time series forecasting methods, namely linear regression with Fourier coefficients, ARIMA, ETS (exponential smoothing with trends) and nonlinear time series methods. This approach thus not only assesses the presence of nonlinearity in the activity dynamics of peer production systems, but it also relates forecasting performance to a measure of nonlinearity in activity dynamics, as indicated by the statistical tests. Finally, I employ techniques from nonlinear time series analysis to illustrate the analysis of activity dynamics in communities exhibiting nonlinear behavior.

Applying this a approach to time series describing global activity dynamics of 16 Stack Exchange Q&A communities, I find that nonlinearity appears to describe activity dynamics better in some communities, and, in those cases, nonlinear forecasting methods outperform other methods. These results underscore heterogeneity in community characteristics, and highlight the need for activity dynamics models to capture stochasticity or to use tools from nonlinear time series analysis to study deterministic but chaotic behavior.

Nonlinear Characterization of Activity Dynamics in Online Collaboration Websites

Tiago Santos
Know-Center
Graz, Austria
tsantos@know-center.at

Simon Walk
Stanford University
walk@stanford.edu

Denis Helic
Graz University of Technology
dhelic@tugraz.at

ABSTRACT

Modeling activity in online collaboration websites, such as StackExchange Question and Answering portals, is becoming increasingly important, as the success of these websites critically depends on the content contributed by its users. In this paper, we represent user activity as time series and perform an initial analysis of these time series to obtain a better understanding of the underlying mechanisms that govern their creation. In particular, we are interested in identifying latent nonlinear behavior in online user activity as opposed to a simpler linear operating mode. To that end, we apply a set of statistical tests for nonlinearity as a means to characterize activity time series derived from 16 different online collaboration websites. We validate our approach by comparing activity forecast performance from linear and nonlinear models, and study the underlying dynamical systems we derive with nonlinear time series analysis. Our results show that nonlinear characterizations of activity time series help to (i) improve our understanding of activity dynamics in online collaboration websites, and (ii) increase the accuracy of forecasting experiments.

Keywords

Nonlinear time series analysis; Q&A online communities

1. INTRODUCTION

Online Question and Answering portals, such as StackExchange or Quora, are immensely popular and helpful online resources with very large communities, amassing millions of users, questions and answers each[1]. However, while some online portals strive and blossom, the majority fails to attract users and never reaches critical mass, requiring them to shut down due to lack of activity, such as Google's knol project[2]. In this paper, we are motivated by the identification of key deciding features of activity time series, which hopefully will provide the foundation to distinguish successful and failing systems. In a first step towards this ambitious goal, we generalize the problem and apply several nonlinear time series analysis techniques to grasp and characterize hidden nonlinear behavior affecting activity dynamics. Current research on the study of dynamics governing such online collaboration websites focuses on model derivation with nonlinear, differential, parametric dynamical systems to describe observed data [14, 22]. However, such approaches are general purpose approaches, designed to fit observed data while minimizing the model's configuration effort to retain interpretability. In particular, these models do not address specificities of different websites or portals (e.g. StackExchange's Math vs. TeX portals) and do not aim to provide more than a general indicator for trends in activity of collaboration networks.

In this paper, we expand on existing work by conducting the following experiments on 16 randomly picked instances of the StackExchange portal: First, we categorize activity time series derived from online collaboration websites by the time series' likelihood to have stemmed from some hidden, nonlinear dynamical system. To that end, we use 9 statistical tests for nonlinearity to assess the adequateness of a nonlinear dynamical system to model activity. Then, we validate the plausibility of this categorization by comparing forecast performance from 3 standard time series models with nonlinear models, reconstructed from the observed activity time series. Finally, we present an exemplary study of nonlinearity properties of 2 datasets.

We find that activity in online collaboration websites may be modeled accurately by underlying, reconstructed dynamical systems to varying degrees, with some online collaboration websites showing more signs of nonlinear behavior than others. We use these differences to characterize the datasets and show how this knowledge may be used to not only improve activity modeling and forecasting efforts, but also better grasp datasets with nonlinear behavior by using tools from nonlinear time series analysis.

Our main contribution is therefore the improvement of the dynamical system modeling process for activity dynamics in online collaboration websites: Instead of postulating a "one-size-fits-all" dynamical system description via parametrized nonlinear equations, as done e.g. by Ribeiro [14] and Walk et al. [22], we reconstruct dynamical system descriptions directly from observed data and assess the feasibility of such a reconstruction. This allows us to tailor time series models to different data origins and thereby improve activity dynamics forecast quality. Furthermore, the use of nonlinear time

[1]See, for example, `http://stackexchange.com/sites?view=list#traffic`
[2]`http://knol.google.com/`

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051117

series analysis techniques, such as Recurrence Plots analysis, further boosts our understanding of nonlinear activity dynamics, for example through the identification of changes in stationarity or chaotic dynamics, leading to more model fine-tuning possibilities, which incorporate such information.

2. RELATED WORK

We review related work from the following two fields of research: nonlinear time series analysis applications and dynamical systems for networks.

Nonlinear Time Series Analysis and its Applications. Nonlinear time series analysis revolves around reconstructing a high dimensional dynamical system from an univariate time series, and studying the properties of the reconstructed dynamical system to derive knowledge on the original, univariate time series [1]. Nonlinear time series analysis enables studies on the *deterministic and chaotic*, rather than stochastic, nature of time series. Chaos means, in this sense, that small differences in a time series' present lead to great changes in its future, despite the dynamical system governing the time series' evolution being intrinsically deterministic.

Nonlinear time series analysis offers theoretical and practical tools to deal with reconstructed dynamical systems [12, 1], and these tools have found application in numerous areas [16, 7, 17]. In one of the most prominent applications of nonlinear time series analysis, Small and Tse [16] discuss how to predict the outcomes of a roulette wheel. A number of authors have also investigated the presence of chaotic behavior in financial markets, for example, by assessing nonlinearity in stock returns with statistical tests for chaos [7] or identifying events in stock returns with Recurrence Plots (RP) and Recurrence Quantification Analysis (RQA) [17].

For a detailed survey of the theory and application of RP and RQA we refer the interested reader to Bradley and Kantz [1] and Marwan et al. [12].

Dynamical Systems for Networks. Dynamical systems are a well studied topic from the standpoint of mathematics and physics [11, 6]. In general, dynamical systems provide mathematical descriptions on the evolution along the time dimension of a set of numeric quantities. They are employed to describe phenomena like the motion of a mass along some path according to Newton's laws, population growth or even macro-economic systems.

Application categories for dynamical systems on networks include, for example, *activity dynamics*. In the context of collaboration on the Web, activity dynamics apply dynamical systems on network theory to study the evolution of activity in different types of networks. The work by Ribeiro [14] introduces a dynamical system to model activity in membership-based community web pages, where activity is a time series representing the number of daily active users in such web pages. The author's model incorporates two main factors, namely web page users becoming spontaneously active and active users influencing inactive ones to become active. With this model, the author explains and predicts when a web page has reached a self-sustaining level of activity. More recently, Walk et al. [22] also applied dynamical systems theory to study activity dynamics in the context of collaboration networks, such as those arising in Question and Answering portals in the web. Here, the authors directly derive their key contributions from the activity dynamics model they propose, which include the self-sustaining level of activity for that type of collaboration network and the robustness of a collaboration network's activity.

In general, the authors of previously mentioned papers all propose a mathematical model, consisting of parametrized equations for a dynamical system, as a means to describe observed behavior. In contrast, we do not postulate parametrized equations describing a dynamical system on a network. Instead, we interpret the observed activity data, in the form of time series, as one dimensional projections of a hidden, complex and higher dimensional dynamical system. We study the feasibility of reconstructing the dynamical systems underlying the activity time series, characterize these activity time series by their propensity to have originated in such complex dynamical systems, and inspect the reconstructed dynamical system's properties.

3. METHODOLOGY

3.1 Forecasting univariate time series

Time series are sequences of numerical values (or observations), indexed and ordered by time. We consider discrete univariate time series, where each time index is uniquely associated with one observation. Moreover, we assume the time series observations are equally spaced in time.

Assessing nonlinearity in univariate time series. Not all univariate time series are equally suited for the reconstruction of a dynamical system; the presence of e.g. noise or randomness greatly influence the embedding. Therefore, we assess nonlinearity of univariate time series via the 9 following statistical tests: *Broock, Dechert and Scheinkman test* [2]; *Teraesvirta's neural network test* [19]; *White neural network test* [10]; *Keenan's one-degree test for nonlinearity* [9]; *McLeod-Li test* [13]; *Tsay's test for nonlinearity* [21]; *Likelihood ratio test for threshold nonlinearity* [4]; *Wald-Wolfowitz runs test* [4]; *Surrogate test - time asymmetry* [15].

We apply these tests without configuration changes, except for the *Broock, Dechert and Scheinkman* and *Wald-Wolfowitz runs* tests. As described in Zivot and Wang [23, p. 652], we compute the test statistic of *Broock, Dechert and Scheinkman* on the residuals of an autoregressive integrated moving average (ARIMA) model, a class of linear models basing on auto regression, to check for nonlinearity not captured by the ARIMA model. For the *Wald-Wolfowitz runs* test, since a run represents a series of similar responses, we define a positive run as the amount of times the time series value was greater than the previous one [20].

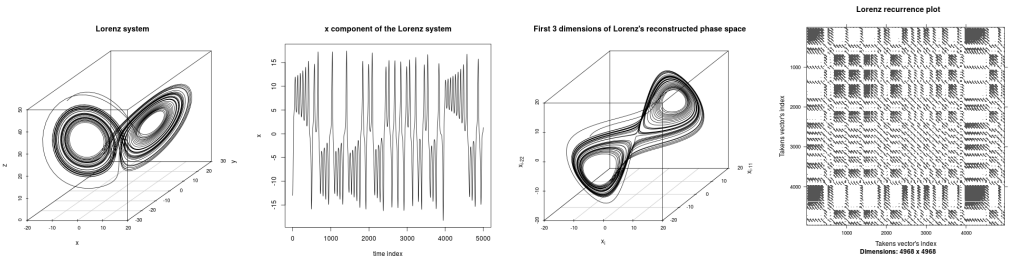
Reconstructing state space from univariate time series. Nonlinear time series analysis studies dynamical systems reconstructed from univariate time series. Takens [18] presents an embedding function, which, under certain conditions, maps an univariate time series to the higher dimensional phase space the reconstructed dynamical system lives in, and restores the topological characteristics of the dynamical system's reconstructed state space.

We briefly present theory on the embedding map required to reconstruct the state space of a dynamical system.

If x_t denotes the value of a time series x at time t, then an embedding of x can be obtained with a reconstruction vector of the form

$$R_t = (x_t, x_{t-\tau}, x_{t-2\tau}, \ldots, x_{t-(m-1)\tau}) \in \mathbb{R}^m. \qquad (1)$$

There are two free parameters in equation 1: τ and m. τ is the time lag, representing a distance in time between time



(a) Lorenz system has two attractors

(b) Time series of the Lorenz system's first component

(c) The reconstructed state space plot shows the attractors

(d) The recurrence plot of the Lorenz system reflects overall dynamics of the system

Figure 1: **Illustration of nonlinear time series analysis with the Lorenz dynamical system.** Figure 1a depicts the Lorenz system with parameters $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$. We extract the Lorenz system's first component, shown in Figure 1b, and then reconstruct its state space with the embedding described in the embedding theory part of Section 3.1. The results of that embedding, with parameters $\tau = 11$ and $m = 4$, are the subject of picture 1c (showing only 3 dimensions). The reconstructed state space captures the original structure of the Lorenz system and its two attractors remarkably well. The structure of the Lorenz system can also be observed in the corresponding recurrence plot (RP) (see 1d). The RP shows the Lorenz attractors prominently around time indexes 1600 and 4200. Note also the large number of short diagonals around the main diagonal of the plot: These reflect the chaotic behavior of this Lorenz system.

series observations. m is the embedding dimension, i.e. the size of the vectors R_t in the space of the reconstructed dynamical system.

To estimate the embedding parameters, we start with the time lag τ. Bradley and Kantz [1] stress that τ should be large enough to encompass one full cycle of a time series' periodic dynamics. To estimate such a cycle's length (and thus τ too), the same authors propose different measures of independence between time series observations. We use the first minimum of average mutual information between observations as a measure of independence to estimate τ. The estimation of the embedding dimension m is an iterative process, which consists of computing some invariant of the reconstructed dynamical system for $m = 1, 2, \ldots$. We stop the process when the value of the invariant stabilizes, which indicates that the reconstructed dynamical system has been properly unfolded. We employ the commonly used iterative procedure [3] for the estimation of the embedding dimension.

Forecasts from linear models. To forecast an univariate time series, often used models include linear, ARIMA and ETS models.

In a linear model, a target variable is expressed as a linear combination of explanatory variables. We choose Fourier coefficients as explanatory variables, to account for seasonality effects of the type we encounter in the data described in 4.1.

The ARIMA class of models comprises auto-regressor models, which express the target univariate time series as a linear combination of its own past values and some lagged moving average error terms as well. This class of models assumes weak stationarity of the time series, so differencing—a technique to make a time series stationary—may be applied.

The ETS class of models includes exponential smoothing models, which—similarly to ARIMA—define the value of the target time series as a linear combination of lagged terms, such as level, trend, seasonality and error.

There are many variations of ARIMA and ETS time series models, and we automate the choice of model parameters and configurations with the algorithm devised by Hyndman and Khandakar [8].

Forecasts nonlinear models. Forecasts from nonlinear models require, first, the embedding map to reconstruct state space dynamics from the target time series, as described in the embedding theory part of Section 3.1. Given an embedding, nonlinear models forecast a target time series first by searching for nearest neighbor (with respect to the target time series) trajectories in the reconstructed state space. Then, the forecast from the nonlinear model is the arithmetic mean of future values of those near trajectories.

3.2 Recurrence Analysis

We analyze recurrences in reconstructed state space trajectories with Recurrence Plots (RPs), which give insights into both the behavior (e.g. stationary or drifting) and type (e.g. periodic, deterministic chaotic or random) of reconstructed dynamical systems, so we aim to use RPs to help with the nonlinear characterization of our data.

The RP is associated with a recurrence matrix—a square matrix which shows reconstructed state space trajectories $\vec{x}_i$ close to each other:

$$R_{i,j}(\epsilon) = \Theta(\epsilon - \|\vec{x}_i - \vec{x}_j\|), \tag{2}$$

where Θ is the Heaviside function and ϵ is the recurrence threshold establishing closeness between reconstructed state space trajectories $\vec{x}_i$. Thus, the RP is a scatter plot, simply showing points where the recurrence matrix is equal to 1. In Equation 2, we use the Euclidean norm and for ϵ we take the standard deviation of the distance matrix of all reconstructed state space trajectories.

Figure 1 shows an example of nonlinear time series analysis, complete with an RP characterization of the reconstructed state space. Starting with the standard chaotic Lorenz system, we extract its first component to reconstruct its state space and we analyze the reconstruction. We observe that the reconstructed system's topology accurately resembles the original system's one, and that the RP, with

Table 1: The table shows, per dataset, activity time series length in weeks, embedding parameters τ and m, nonlinearity test results (nonlin. tests), i.e. number and reference of statistical tests indicating nonlinearity with a significance level of 95% out of the 9 applied tests, normalized root mean squared error (RMSE) of a 1 year forecast per model. We also show the ranking the Friedman test assigns to the models' forecast RMSE for datasets with 5 or more tests indicating nonlinearity and the rest. Nonlinear models show best prediction performance on datasets with more than five statistical tests indicating nonlinearity.

Dataset	Weeks	τ	m	Nonlin. test score	Positive nonlin. tests	ARIMA	ETS	Linear	Nonlin.
english[b]	240	2	9	2/9	[2] [13]	0.6794	0.4452	0.3329	0.3080
unix[b]	239	1	7	2/9	[2] [13]	0.2091	0.2092	0.2418	0.2074
chemistry[b]	158	2	7	3/9	[2] [13] [4]	0.4982	0.2539	0.3247	0.4610
webmasters	244	1	8	3/9	[9] [13] [15]	0.2313	0.2528	0.3341	0.2346
chess	148	2	8	4/9	[2] [9] [13] [15]	0.2545	-[a]	0.5622	0.5110
history	177	1	9	4/9	[2] [9] [13] [4]	0.3503	0.2368	0.3044	0.4052
linguistics	181	2	6	4/9	[2] [9] [13] [15]	0.2512	0.2704	0.3009	0.3280
sqa	200	3	9	4/9	[2] [9] [13] [15]	1.8136	0.2531	0.6549	0.3903
tex[b]	241	1	7	4/9	[13] [21] [4] [15]	0.1589	0.1580	0.2767	0.2751
tridion	107	1	7	4/9	[19] [10] [9] [13]	0.2717	-[a]	0.6144	-[a]
Friedman test rank of models' forecast RMSE on datasets with nonlin. test score $< 5/9$						2	1	4	3
arduino	56	1	10	5/9	[2] [19] [10] [9] [13]	0.3489	-[a]	-[a]	-[a]
sports	159	1	7	5/9	[2] [9] [13] [4] [15]	0.2442	0.3348	0.4019	0.3323
ux	239	2	8	5/9	[2] [10] [9] [13] [21]	0.3479	0.1743	0.3491	0.1374
bitcoin	182	4	11	6/9	[2] [19] [10] [9] [13] [15]	0.6099	0.5549	0.5938	0.5781
math[b]	242	2	8	6/9	[2] [19] [13] [21] [4] [15]	0.1327	0.2314	0.3521	0.2912
bicycles	235	2	7	7/9	[2] [19] [10] [9] [13] [4] [15]	0.2971	0.3097	0.3252	0.2805
Friedman test rank of models' forecast RMSE on datasets with nonlin. test score $\geq 5/9$						2[c]	2[c]	4	1

[a] This activity time series is too short for a 1 year forecast with this model.
[b] This activity time series had a strong linear trend, so the results above concern the activity time series detrended with linear regression.
[c] These models achieved the same rank in the Friedman test for this group of datasets.

its large number of small diagonals and its clusters of points depicting the Lorenz attractors, reflects the overall chaotic behavior of the Lorenz system.

4. EXPERIMENTAL SETUP

4.1 Datasets

For our analysis, we gathered data from 16 randomly picked StackExchange[3] questions and answers portals.

We follow the procedure described by Walk et al. [22] to derive univariate time series describing activity in these online collaboration websites: First, we measure a user's activity in such online collaboration websites as the user's number of questions, answers and comments per day. Then, we smooth these daily activities with a rolling mean over a 7 day window, to account for and remove outliers, and aggregate activity over all users per week. Finally, we require the weekly activity time series to have at least one unit of activity (i.e. one post, reply or comment) per day. This implied a burn-in of initial phases of inactivity or very low activity from the activity time series. For more details, see Table 1.

4.2 Predicting Activity

To assess if the activity time series show signs of nonlinear behavior, we apply all 9 statistical tests (described in the nonlinearity assessment part of Section 3.1) on the datasets (see 4.1), with a significance level of 95%. We then build a ratio, per dataset, of the number of tests indicating nonlinearity out of all 9 tests. That ratio serves as an indicator for hidden nonlinear dynamics, or not, in a given activity time series: We conjecture that higher values of that ratio will likely indicate hidden nonlinear dynamics, while datasets, which score lower on that ratio, are less likely to have such dynamics.

To test this approach for distinguishing time series with nonlinear behavior, we benchmark the performance of nonlinear forecasts on all datasets against those from other models. For datasets, characterized as nonlinear by the nonlinearity tests, we expect time series forecasts from nonlinear models to compare favorably against other models. The other models we benchmark nonlinear models against are linear, ARIMA and ETS models. For each of the datasets, we train those four models on a shorter version of the activity time series, excluding the last year of activity. We predict that last year with each of those 4 models and, finally, we compare the models' forecast results with the empirically observed values. We use the root mean squared error, normalized by the range of the activity time series, for the forecast performance comparison.

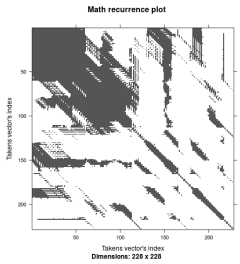
Since the nonlinearity tests (see nonlinearity assessment in Section 3.1) focus on the distinction between possibly chaotic determinism and randomness, activity time series with a strongly increasing (or decreasing) linear trend will be recognized as non-random. Strong linear trends may mask hidden nonlinear dynamics, which we aim to inspect.

Therefore, we first assess the strength of the trend of an activity time series by inspection of both the time series' plot and relative weight of a LOESS decomposition assigns to the trend component of that time series. For time series with a strong linear component, we estimate the linear trend with a simple linear regression, minimizing the weighted least squared error. Finally, we subtract that fitted linear trend from the time series, and perform nonlinearity tests and forecast computations on the detrended activity time series.

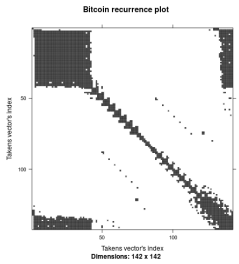
5. RESULTS & DISCUSSION

Findings on nonlinearity assessment and activity forecast models. We have listed all results of the nonlinear characterization via nonlinearity tests and activity forecast benchmarks in Table 1. The nonlinearity test scores indicate some disparity in the presence of nonlinear dynamics for the activity time series. Out of the 16 datasets we analyze, 6 datasets test five or more times positive for nonlinearity, and the other 10 datasets below five times. We interpret this split as an hint at some differences in nonlinear behavior of these datasets, and compare modeling and forecasting performance of the nonlinear, linear, ARIMA and ETS models for each of those two groups of activity time series.

[3] http://stackexchange.com/



(a) Recurrence Plot for the math activity time series



(b) Recurrence Plot for the bitcoin activity time series

Figure 2: **Recurrence Plots (RPs) give insights into activity time series dynamics.** Although both datasets, "Math" and "Bitcoin", have the same amount of statistical tests indicating nonlinearity, their RPs look quite different. The "Math" RP in figure 2a shows a higher density of recurrence points in the upper left corner, which gradually diminishes towards the lower right corner; this is a sign of a drift in the activity time series, still present after linear detrending [12]. Note both the diagonal as well as vertical structures present in Math's RP. The former, prominent around time indexes 100 to 175, could be a sign of chaotic dynamics, while the latter points towards states in the reconstructed state space which are (very) slowly changing. In contrast to Math, Bitcoin's RP in Figure 2b prominently features one strong main diagonal, with some remarkable periodicity around it. Another interesting aspect of Bitcoin's RP are the white bands around the main diagonal and the cluster of recurrence points in the lower left (and by symmetry of the RP also upper right) corner. These both hint at non-stationary transitions in the activity time series [12].

We assess the performance of these four models by calculating the normalized root mean squared error of a one year activity forecast with the Friedman test, as described by Demšar [5]. The Friedman test ranks nonlinear model performance highest for the group of datasets with more than five statistical tests indicating nonlinearity. In contrast, the nonlinear models only rank third for the other datasets, where less than five tests indicated nonlinear behavior. This result suggests a distinction in the degree of hidden nonlinear behavior in these activity time series.

We reason that activity time series, which were characterized as less likely to be driven by hidden nonlinear dynamics, were also better modeled by approaches other than nonlinear models due to their strong stochastic behavior. In such cases, we believe the role noise and external factors such as events play should not be underestimated.

The nonlinearity tests [10] and [19] appear to be more sensitive to the presence of nonlinear dynamics than other tests, since they test positive for nonlinearity 4 times more often in the dataset group with 5 or more tests indicating nonlinearity than in the other dataset group. Since [10] and [19] apply neural networks to assess linearity in mean, we attribute the usefulness of these two tests to the well-studied ability of neural networks to model nonlinear behavior.

We observe that the choice of appropriate models for activity dynamics should incorporate this characterization of activity time series according to evidence found for nonlinear behavior. Therefore, we find that a set of parametrized dynamical system equations to describe activity dynamics for all these StackExchange datasets at once, while easier to grasp and interpret, will likely fail to accurately reflect dataset specificities and thus perform poorer overall than the tailoring of time series models and reconstruction, where appropriate, of nonlinear dynamical system descriptions of the observed data.

Recurrence Plot analysis. Due to limitations in space, we perform an exemplary RP analysis on two activity time series. In Figure 2, the RPs of the datasets "Math" and "Bitcoin", two datasets with 6 statistical tests indicating nonlinearity, suggest differences in their underlying nonlinear dynamical systems, despite the apparent resemblance afforded by similar nonlinearity test results.

Math's RP shows, even after linear detrending, a drift pattern, which is conveyed by the reduction in recurrence point density from the RP's top-left to its bottom-right. We can observe other properties in Math's RP: There are some signs of chaotic behavior, apparent by the numerous short diagonals towards the lower-right corner and alongside the RP's main diagonal, and there are also some signs of slowly changing states in activity, as the long vertical line along time index 150 indicates. Armed with this knowledge we could tailor any type of time series model better to the data: The knowledge of drift enables us to introduce some parameter describing it. Slowly changing states transitioning to chaotic behavior suggest the choice of some threshold model, addressing those characteristics separately.

The main features of Bitcoin's RP are the periodically repeating structures around the main diagonal, the prominent white bands around the main diagonal and the point cluster in the lower left corner (and, by symmetry of the RP, in the upper right corner too). The latter two features indicate strong stationarity changes, while the regularity along the main diagonal hints at deterministic behavior. Again, these observations help with activity dynamics model design: We could introduce some periodic component to address the observed regularities, and we could include some exogenous variable to deal with the stationarity affecting events indicated by the RP's point clusters and white bands.

6. CONCLUSIONS & FUTURE WORK

We set out to explore a new and important issue on modeling activity dynamics: to recognize and characterize different online collaboration websites by the plausibility of

hidden nonlinear dynamical systems governing them, and thereby understand, model and forecast them better.

To address these open issues, we proposed using 9 different statistical tests for the nonlinear characterization of activity time series, and to validate this characterization with a comparison of the performances of different forecasting models. We also provided a sample RP analysis of activity time series characterized as nonlinear, to showcase the utility of these methods.

Our results can be summarized as follows. Firstly, a characterization of nonlinearity in activity time series by statistical tests gauges the plausibility of an activity time series being accurately described by dynamical systems (in contrast to, for example, some stochastic process), thus influencing model choice and helping discern driving forces of activity in our datasets. Secondly, nonlinear models seem adequate for forecasting activity time series, deemed nonlinear by statistical tests, more so than classical forecasting models (and vice-versa), a distinction which improves overall activity dynamic forecast quality. Thirdly, nonlinear modeling enables, via Recurrence Plots, a more granular study and deeper understanding of nonlinear dynamics governing activity time series, allowing for finer customization of time series models to explain activity in online collaboration websites.

This paper's limitations are a direct consequence of those of nonlinear time series analysis and the Friedman test's conservative estimations: Less noise, longer time series and more datasets should make results more conclusive.

With the hope of understanding *why* we see the observed activity dynamics, we believe that one of the most promising avenues for future work on nonlinear analyses of activity dynamics to be the connection between network science and the reconstructed dynamical systems. We speculate that hidden connections between statistics on these reconstructed dynamical systems, given for example by Recurrence Quantification Analysis, and properties of the underlying collaboration networks of websites will deliver further insights into the dynamic processes driving activity.

7. ACKNOWLEDGMENTS

The work has been performed in the project Power Semiconductor and Electronics Manufacturing 4.0 (SemI40), under grant agreement No 692466. The project is co-funded by grants from Austria, Germany, Italy, France, Portugal and - Electronic Component Systems for European Leadership Joint Undertaking (ECSEL JU).

8. REFERENCES

[1] E. Bradley and H. Kantz. Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097610, 2015.

[2] W. Broock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron. A test for independence based on the correlation dimension. *Econometric reviews*, 15(3):197–235, 1996.

[3] L. Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena*, 110(1):43–50, 1997.

[4] K. S. Chan. Percentage points of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 691–696, 1991.

[5] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[6] J. Guckenheimer and P. J. Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42. Springer Science & Business Media, 2013.

[7] D. A. Hsieh. Chaos and nonlinear dynamics: application to financial markets. *The journal of finance*, 46(5):1839–1877, 1991.

[8] R. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for r 7. *URL: https://www.jstatsoft.org/article/view/v027i03 [accessed 2016-02-24]*, 2007.

[9] D. M. Keenan. A tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, 72(1):39–44, 1985.

[10] T.-H. Lee, H. White, and C. W. Granger. Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics*, 56(3):269–290, 1993.

[11] D. G. D. G. Luenberger. Introduction to dynamic systems; theory, models, and applications. Technical report, 1979.

[12] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5):237–329, 2007.

[13] A. I. McLeod and W. K. Li. Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, 4(4):269–273, 1983.

[14] B. Ribeiro. Modeling and predicting the growth and death of membership-based websites. In *Proceedings of the 23rd international conference on World Wide Web*, pages 653–664. ACM, 2014.

[15] T. Schreiber and A. Schmitz. Surrogate time series. *PhysicaD:NonlinearPhenomena*, 142(3):346–382, 2000.

[16] M. Small and C. K. Tse. Predicting the outcome of roulette. *Chaos: an interdisciplinary journal of nonlinear science*, 22(3):033150, 2012.

[17] F. Strozzi, J.-M. Zaldívar, and J. P. Zbilut. Application of nonlinear time series analysis techniques to high-frequency currency exchange data. *Physica A: Statistical Mechanics and its Applications*, 312(3):520–538, 2002.

[18] F. Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[19] T. Teräsvirta, C.-F. Lin, and C. W. Granger. Power of the neural network linearity test. *Journal of Time Series Analysis*, 14(2):209–220, 1993.

[20] A. Trapletti and K. Hornik. *tseries: Time Series Analysis and Computational Finance*, 2016. R package version 0.10-35.

[21] R. S. Tsay. Nonlinearity tests for time series. *Biometrika*, 73(2):461–466, 1986.

[22] S. Walk, D. Helic, F. Geigl, and M. Strohmaier. Activity dynamics in collaboration networks. *ACM Transactions on the Web (TWEB)*, 10(2):11, 2016.

[23] E. Zivot and J. Wang. *Modeling financial time series with S-Plus®*, volume 191. Springer Science & Business Media, 2007.

3.3 Activity Archetypes in Question-and-Answer (Q&A) Websites – A Study of 50 Stack Exchange Instances

This article also tackles the first research question, which regards the characterization of user activity dynamics in peer production systems, but, in contrast to the previous work, this research focuses on a user-level characterization of activity dynamics and how that relates to longitudinal system-level properties. Specifically, this work extracts time series descriptions of user activity in 50 Stack Exchange Q&A communities, and proposes a parsimonious set of features which both captures their burstiness and is amenable for typifying user behavior via K-Means clustering. With linear regression analyses, I measure if a community's user mix is connected to its developmental stage. Repeating this clustering procedure across windows of six months also uncovers trajectories and timings of activity growth.

The results suggest four user archetypes capture user behavior across levels of activity and of recurrence of participation, and the user mix, i.e. the proportion of total activity contributed per user type, relates to the developmental stage of the Q&A community. These results may support Q&A community managers to quantify the development of the communities they oversee, as well as researchers to model the activity dynamics of users and the system.

Activity Archetypes in Question-and-Answer (Q&A) Websites—A Study of 50 Stack Exchange Instances

TIAGO SANTOS, Graz University of Technology, Austria
SIMON WALK, Detego GmbH, Austria
ROMAN KERN, Graz University of Technology & Know-Center, Austria
MARKUS STROHMAIER, RWTH Aachen University & GESIS, Germany
DENIS HELIC, Graz University of Technology, Austria

Millions of users on the Internet discuss a variety of topics on Question-and-Answer (Q&A) instances. However, not all instances and topics receive the same amount of attention, as some thrive and achieve self-sustaining levels of activity, while others fail to attract users and either never grow beyond being a small niche community or become inactive. Hence, it is imperative to not only better understand but also to distill deciding factors and rules that define and govern sustainable Q&A instances. We aim to empower community managers with quantitative methods for them to better understand, control, and foster their communities, and thus contribute to making the Web a more efficient place to exchange information. To that end, we extract, model, and cluster a user activity-based time series from 50 randomly selected Q&A instances from the Stack Exchange network to characterize user behavior. We find four distinct types of user activity temporal patterns, which vary primarily according to the users' activity frequency. Finally, by breaking down total activity in our 50 Q&A instances by the previously identified user activity profiles, we classify those 50 Q&A instances into three different activity profiles. Our parsimonious categorization of Q&A instances aligns with the stage of development and maturity of the underlying communities, and can potentially help operators of such instances: We not only quantitatively assess progress of Q&A instances, but we also derive practical implications for optimizing Q&A community building efforts, as we, e.g., recommend which user types to focus on at different developmental stages of a Q&A community.

CCS Concepts: • **Mathematics of computing** → **Cluster analysis**; *Time series analysis*; • **Information systems** → **Answer ranking**; • **Human-centered computing** → *Web-based interaction*; *Computer supported cooperative work*;

Additional Key Words and Phrases: Question-and-Answer (Q&A) websites, User types in Q&A websites, temporal activity patterns in Q&A websites, Sustainability of Q&A websites

ACM Reference format:
Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. 2019. Activity Archetypes in Question-and-Answer (Q&A) Websites—A Study of 50 Stack Exchange Instances. *ACM Trans. Soc. Comput.* 2, 1, Article 4 (February 2019), 23 pages.
https://doi.org/10.1145/3301612

Authors' addresses: T. Santos and D. Helic, Institute of Interactive Systems and Data Science, Graz University of Technology, Inffeldgasse 16c/I, 8010 Graz, Austria; emails: tsantos@iicm.edu, dhelic@tugraz.at; S. Walk, Detego GmbH, Hans-Resel-Gasse 17a 8020 Graz, Austria; email: s.walk@detego.com; R. Kern, Institute of Interactive Systems and Data Science, Graz University of Technology & Know-Center, Inffeldgasse 13/VI, 8010 Graz, Austria; email: rkern@tugraz.at; M. Strohmaier, HumTec Center, RWTH Aachen University, Theaterplatz 14, 52062 Aachen, Germany; email: markus.strohmaier@cssh.rwth-aachen.de.

2469-7818/2019/02-ART4
https://doi.org/10.1145/3301612

1 INTRODUCTION

Question-and-answer (Q&A) websites (e.g., Stack Exchange[1] or Quora[2]) are publicly accessible platforms, which are used by millions of users to discuss a variety of topics and problems. For example, the StackOverflow[3] instance of the Stack Exchange website deals with topics related to programming and hosts a flourishing community of more than 6 million users. Another prominent example is the Math Stack Exchange[4] instance, where a thriving community of mathematical professionals and other users with shared interests pose and solve mathematical questions.

Problem. However, not all Q&A instances exhibit the same kind of vibrant, self-sustaining community activity. In fact, the majority of Q&A instances fail to attract and engage enough users to reach self-sustainability in terms of activity. Typically, instance operators provide incentives for users in the form of badges or reputation scores. Although several studies analyzed the effects of such endeavors [4, 36, 42], our research community still lacks the tools to understand, measure, model, and predict key factors that influence and drive Q&A communities to sustainable levels of activity. However, without a proper understanding of users, the structures inherent in the communities, as well as the driving mechanisms behind successful Q&A instances, we can not hope to remedy the problems of less successful sites.

Approach. In this article, we set out (i) to characterize user activity profiles, (ii) to reveal the compositions of those profiles in various Q&A communities, and (iii) to analyze similarities and differences between highly and less successful Q&A instances.

Although current research on users of online Q&A communities partially uncovers different user roles in these communities [13, 20, 36, 56], we identify a research gap on (i) the composition of activity profiles for communities at different stages of maturity and (ii) specific compositions that ultimately make thriving communities successful.

Specifically, we characterize temporal activity patterns of users of Q&A instances, analyze and compare the activity composition and development of whole instances, and provide actionable information for instance operators to assess maturity, improve activity, and manage their instances more efficiently. To that end, we randomly pick a total of 50 Stack Exchange instances, from which we derive time series and features that describe commonly occurring temporal activity patterns. We represent user activity as their total count of posts and replies. Subsequently, we apply K-Means on the extracted features to group users with similar activity profiles and find optimal numbers of clusters by calculating and comparing silhouette coefficients for different values of K. Additionally, we analyze the composition of activity across the obtained clusters for all Q&A instances.

Contributions. The main contributions of our work are as follows: First, we find that activity-based time series can be described by the following two quantities: (i) the characteristics of its peaks and (ii) the uniqueness of its non-zero activity values.

Second, we identify typical user activity profiles to describe *Activity Archetypes*, which represent distinct user engagement levels across all analyzed Stack Exchange instances. This result helps not only to better understand the different user profiles that operators of Q&A instances need to cater to, but also which profiles to include when modeling activity for these instances.

Third, we analyze, compare, and categorize the *Activity Archetype* composition of various Stack Exchange instances, which allows us to assess the level of maturity in a Stack Exchange instance's

[1] http://www.stackexchange.com.
[2] http://www.quora.com.
[3] http://stackoverflow.com/.
[4] http://math.stackexchange.com/.

development toward activity-based self-sustainability. To give an example, we find that thriving instances feature substantial amounts of infrequently active users posing questions. If this group of users is underrepresented, then this affects the instance's overall activity and development.

We believe that our analyses represent an important step toward a better understanding of the factors that define and foster success in Q&A instances. With our analyses, we enable Q&A instance operators not only to gauge, quantify, and model the status of their communities in comparison to other communities, but, with our discussion of practical implications, also to pinpoint what user groups to focus activity improvement measures on, on the path toward a thriving, self-sustaining community.

2 RELATED WORK

Dynamical systems for modeling activity. Dynamical systems are systems of parametrized equations describing the evolution of numerical quantities over time. They provide a mathematical formalization for activity dynamics models.

Perra et al. [38] model activity in collaboration networks such as publications and references in the Physical Review Letters journal. The authors measure an empirical probability distribution over interactions of agents in a network, model the formation of dynamic networks based on this activity distribution, and study resulting dynamical processes. This work influenced other authors modeling activity dynamics as explicit dynamic processes on networks, such as Laurent et al. [32]. Those authors propose an activity-driven model for time varying networks to analyze mobile call records from an European telecom. Building on the work by Perra et al. and Laurent et al., Wölbitsch et al. [54] extended an activity-driven network model with a peer-influence mechanism to study peer-influence and its effects on the network in a controlled setting.

Other approaches to model activity in Q&A instances and networks with dynamical systems focus on a few key variables that drive overall activity dynamics. Ribeiro [39] models activity in membership-based community websites as time series counting the number of active users in such websites. The model considers two main factors, namely active users spontaneously becoming inactive and active users spurring inactive ones to become active. These factors are sufficient to distinguish self-sustaining from non-self-sustaining online communities and to forecast their daily active user numbers. Walk et al. [51] proposed a dynamical system description for online Q&A instances such as Stack Exchange instances or Semantic MediaWikis.[5] Their dynamical system equations allow for (i) forecasting activity levels in those online Q&A communities, and for (ii) assessing if an online community reached self-sustaining levels of activity. In an extension of Walk et al.'s [51] models, Koncar et al. [31] recently studied the implications of trolling behavior on various Stack Exchange and Reddit communities.

Similarly to our previous work [41] on nonlinear characterization of Q&A instances, we contribute a data-driven approach to this body of work, which uses mathematical formalization to describe activity in online Q&A instances. In an extension of our previous work, however, we go beyond our analysis of time series of Q&A activity totals by focusing on more granular activity-based time series. Specifically, our objects of study in this work are time series describing user activity in Q&A forums. We thus empirically identify user behavior patterns as key driving forces of activity and thereby pave the way for new models, which take into account users' roles in shaping total Q&A activity as it changes over time.

Characterization of activity in Q&A instances. Literature dealing with dynamics of Q&A instances such as Stack Exchange focuses on many different aspects of these types of online

[5] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki.

communities. Anderson et al. [3] quantify and uncover temporal characteristics of questions, which bring (long-term) value to the community. Burel and He [8] measure the maturity of the ServerFault Stack Exchange instance by its ability to cope with complex questions. In contrast to that study on complex questions, Correa and Sureka [11, 12] measure, via characterization studies and prediction experiments, the properties and impact of closed and deleted questions on Q&A quality maintenance. Srba et al. [45] aim to encourage activity on new questions in Stack Exchange instances with improvements on linking users to unanswered questions by analyzing a larger pool of data sources other than the Stack Exchange data itself (e.g., Twitter). Our work also derives policy suggestions for Q&A community managers, but from a user-based analysis, rather than based on questions and their properties. This enables our focus on macro-level aspects of Q&A community growth and management, complementing these more granular studies on the impact and value of questions.

Other authors, however, have, similarly to us, focused on user types and engagement as their fundamental object of study. Danescu-Niculescu-Mizil et al. [13] characterize user participation in online communities by the evolution of their language, allowing the authors to predict when users depart their communities. In another study of user types in an online Q&A community, Gazan [21, 22] highlights different types of questioners and answerers, namely seekers and sloths and, respectively, specialists and synthesists. Zhang et al. [59] and, more recently, Yang et al. [56] tackle the problem of expert user identification and characterization in, respectively, a help-seeking forum for Java programming and StackOverflow. Put into a broader context, our work also relates to feature-based characterizations of user behavior online in general, such as Lehmann et al.'s work [33] on typifying online forums by their users' activity and Chan et al.'s work [9] on user types and temporal aspects of user engagement in 80 websites.

Early work by Adamic et al. [1] and Nam et al. [37] on understanding knowledge sharing behavior in the Yahoo Answers and Naver Q&A communities explained user behavior as a product of users' interests and motivation. More recently, the analysis of the StackOverflow design by Mamykina et al. [36] combines a statistical investigation of StackOverflow usage patterns with interviews with StackOverflow's designers. The goal of their procedure is to understand which user behavior leads to the site's success. In particular, the authors find a set of three different types of user activity behavior (plus a lurker, non-active type), which base on their activity frequency.

Sinha et al. [42] study participation and participation incentives in Stack Exchange communities. In their work, the authors underline the relevance of a core of highly active users and of participation incentives for less active users in Stack Exchange communities. Our work shares most commonalities with Furtado et al. [20]. In that study, the authors extract metrics measuring quality and quantity of activity in Stack Exchange instances. With those metrics, they describe a set of 10 different user profiles obtained with K-Means clustering on those extracted metrics. The authors then study the composition and activity dynamics of users in five Stack Exchange instances broken down by the user profiles they found. They show that, although users change profiles over time, the overall composition of user profiles of those five instances mostly does not.

Our comprehensive analysis of 50 Stack Exchange instances yields comparable, but, as we discuss later, slightly but crucially different user profile characterizations than those by Mamykina et al. [36], Sinha et al. [42], and Furtado et al. [20]. That work provides the basis for our article to expand on as follows: A temporal analysis of our user characterization enables us to uncover previously overseen patterns regarding the development and maturity of Stack Exchange instances of varying sizes, ages, and activity profiles. In particular, our results, which highlight an instance's evolving activity composition over time, do not contradict the findings by Furtado et al. [20]. We rather extend the results by Furtado et al. [20], as they analyzed only five similarly sized Stack

Exchange instances, one of which (*programmers*[6]) we find to be of one of multiple types we identify. Our Q&A user and instance characterization thus generalizes their work, as we uncover also a relation between not just one but several user compositions of Stack Exchange instances and their evolving activity growth.

We find that the works by Iriberri and Leroy [25] and by Young [58] qualitatively corroborate our findings. Those authors identify four main life-cycle phases of online communities, namely inception, establishment or growth, maturity, and death or self-sustainability or mitosis, which are comparable to the Stack Exchange instance characterization we derive. In particular, Young [58] also derives a set of recommendations for online health community managers to adapt to their communities' different life-cycle stages. Similarly to Young, we also propose measures for boosting activity in Stack Exchange instances at different maturity stages. In the context of the work by these authors, our work complements theirs with quantitative empirical results and with the application domain of online Q&A communities.

We refer the interested reader to the survey by Srba and Bielikova [44] on previous work on community questions and answers websites for more literature on these topics.

Time series clustering. In the task of time series clustering, one aims to group time series with similar shapes or properties, to ultimately categorize time series, find representative patterns and uncover hidden structures in time series.

A number of authors [16, 18, 23, 34, 50, 55] have applied time series clustering techniques to domains such as finance, online content spread, sensor data, or even warfare analysis. These authors share a common time series clustering approach, which begins with the choice of time series representation to feed to different clustering algorithms. Authors such as Hautamaki et al. [23] consider time series without any transformation, while others extract features [16, 18] or apply transformations to the time series, such as Discrete Wavelet Transforms [50, 55] and Symbolic Aggregate ApproXimation [34]. The time series clustering approach continues with the selection of a distance metric, which very often is the Euclidean [16, 34, 50] or the Dynamic Time Warping distance [23]. Finally, authors settle on a time series clustering algorithm, with popular choices being K-Means and variations thereof [23, 34, 50, 55], self-organizing maps [18], and hierarchical clustering [23]. We select time series features and apply Euclidean K-Means on them to cope with the challenge that discrete valued time series data presents and which is, according to Aghabozorgi et al. [2], rarely dealt with in time series clustering literature. We encourage readers interested in more time series clustering methods and applications to acquaint themselves with the review by Aghabozorgi et al.

3 MATERIALS AND METHODS

3.1 Dataset Characterization

We analyze questions and answers from 50 Stack Exchange Q&A instances on many diverse topics, such as *tex*,[7] *english*,[7] *gardening*,[7] or *buddhism*.[7] The observation periods for all instances vary between 4 and 80 months, depending on the inception date of each instance. The final observation month is February 2017.

As different instances originate at different points in time, the communities in each of those instances naturally exhibit different levels of activity and maturity. For example, *english* started in June 2009 and attracted a total of 37,125 users until February 2017. In contrast, *earthscience*[7]

[6] As of February 2017, the *programmers* Stack Exchange instance is termed *softwareengineering* and programmers.stackexchange.com redirects to softwareengineering.stackexchange.com.

[7] All instances have a corresponding *.stackexchange.com website, where * denotes the instance's name.

Table 1. Dataset Characteristics

Dataset group	Size	Users	Activity	Months
Area 51	25	[473, 6309]	[5023, 47421]	[4, 47]
Non-Area 51	25	[1953, 37125]	[8137, 624166]	[11, 80]

We present value ranges for the number of users, activity (i.e., aggregated questions, answers, and comments), and observation periods (in months) of all datasets (i.e., stack exchange instances) per dataset group. Instances listed on Area 51 are typically smaller and younger than those outside Area 51.

managed to attract only 578 users between April 2014 and February 2015. To foster the development of young instances, such as *earthscience*, the Stack Exchange community submits, incubates, and evaluates proposals for new Q&A instances at a dedicated website called Area 51.[8] If an Area 51 Q&A instance reaches a significant level of activity, the Area 51 community deems it ready for a live test. Then, its live deployment ensues and the Area 51 community monitors its progress until it reaches a sustainable level of activity.

In this article, we analyze a total of 50 Stack Exchange instances consisting of 25 randomly chosen Area 51 datasets and another 25 randomly chosen non-Area 51 datasets (see Table 1).

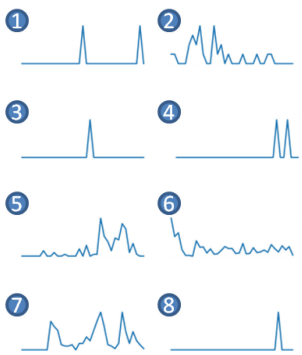
3.2 Feature Engineering

Modeling user activity as time series. We model user activity in Stack Exchange online Q&A instances as two activity-based time series per user. The first one comprises question counts, and the second one reply and comment counts for a given user per month. We stipulate that a user has zero activity if the user did not post a single question (or answer) in any given month of an instance's existence. In all of the following, we treat questions-based activity time series separately from answers-based ones.

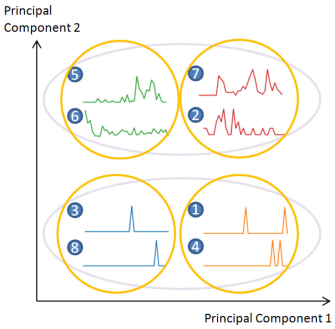
Comparing users' activity-based time series directly. We aim to group users with similar activity profiles by clustering similar activity-based time series.

We first tried to base our clustering approach on a direct measure of similarity between users' activity-based time series with the Euclidean distance. However, using the Euclidean distance fails to discern users with different activity profiles, as it does not account for the misalignment of activity bursts and other activity-affecting events. For example, notice the misalignment in the time axis of the activity peak in time series three and eight of Figure 1(a). As a counter measure to compare misaligned time series, we employed Dynamic Time Warping (DTW). DTW aligns time series over the time axis before computing their similarity with some measure such as Euclidean distance. However, DTW led to no improvements in activity-based time series clustering. As Yang and Leskovec [55] point out, time series of comparable shape but overall varying volume would be considered similar by DTW, hence, making the assignment of different time series into meaningful clusters harder. We discarded other well-established time series similarity and clustering including Symbolic Aggregate Approximation-based (SAX [28]) and matrix profile-based [57] approaches. These and other times series clustering algorithms we reviewed do not specifically address the clustering of sparse count time series problem we face, and thus do not extract meaningful clusters. Furthermore, we were also careful not to apply clustering to segmented time series (e.g., around user activity peaks) with distance-based metrics, as this may be problematic in practice [27]. The need for caution arises from clusters of time series subsequences being essentially random, i.e.,

[8] http://area51.stackexchange.com/.



Time series #	Peaks of equal height?	More than 5 peaks?	Ratio of unique non-zero values
1	Y	N	0.057
2	Y	Y	0.142
3	N	N	0.057
4	Y	N	0.057
5	N	Y	0.4
6	N	Y	0.743
7	Y	Y	0.343
8	N	N	0.057



(a) Extract time series (b) Compute features (c) Search for optimal clustering

Fig. 1. **Identifying user archetypes as a time series clustering problem.** We start by extracting the questions-based (and, separately, answers-based) activity time series as the monthly sum of posted questions (respectively, answers) of each user in a Stack Exchange instance (cf. Figure 1(a)). We then extract three features from these time series: two Boolean features, describing if an activity time series has peaks of equal maximal height and if it has more than five peaks; and the ratio of unique non-zero values to time series length, a continuous feature varying between zero and one (cf. Figure 1(b)). Finally, we cluster the extracted features with K-Means for $K = 2, \ldots, 10$ and save K^*, the value of K, which maximizes the average silhouette coefficient. Graphical inspection of the clusters via PCA projection to two-dimensional space (cf. Figure 1(c)) yields well-separated and cohesive clusters for $K = 2, 3, 4$. However, in this example, for $K = 2, 3, 4$, we get average silhouette coefficient values of 0.423, 0.563, and 0.871, respectively. Hence, K^* equals four. In Stack Exchange instances, we observe varying K^*, which hints at different activity compositions in terms of user archetypes.

independent of input time series subsequence types, if care is not taken to extract time series motifs rather than (often trivial) time series sliding windows.

Extracting features describing temporal activity patterns. Hence, we devised a different approach: We extract time series features summarizing key aspects of temporal user activity patterns occurring in the 50 Stack Exchange instances.

Feature selection. To select time series features, we started with a list of more than 400 different features [10], which comprise descriptive statistics (e.g., mean, auto-correlation, or kurtosis), time series models (e.g., auto-regressive coefficients), and other time series transformations (e.g., Fourier). Starting with those features, we manually searched for a smaller and simpler set of features, which nicely capture the activity distributions as well as a given user activity behavior. Thereby, we focused our search on the user behavior that we observed in our data or that was identified in the previous studies.

In general, we observe large numbers of users with sporadic peaks in activity as well as fewer users who are more active and contribute fluctuating amounts of questions and answers over longer periods of time (cf. examples in Figure 1(a)). Thus, we observe so-called bursty patterns in user activity-based time series. Such bursty patterns have been frequently observed on the Web [36, 49].

Although the majority of our activity-based time series exhibited such patterns consistently, they varied in their temporal location. In other words, rarely-active users with activity bursts in the beginning of the lifetime of a Stack Exchange instance behave similarly to rarely active users with activity bursts in the tail end of the same Stack Exchange instance. This motivated our first

filtering criterion for reducing the original set of more than 400 time series features: We excluded locality based features, such as the locations of a time series' minimal values, since these types of features differentiate user behavior, which we intuitively believe belongs grouped together.

Then, we excluded features unfit for modeling sparse count time series (i.e., our activity-based time series), such as continuous wavelet transformations, auto-regressive models, and other descriptive statistics inadequate in a high sparseness context. Instead, we focused on descriptive statistics which capture activity bursts, such as peak-related features (such as peak height or the number of peaks) or unique value counts. These two filtering steps left us with a set of 15 features.

We then empirically evaluated this set of 15 features with respect to parsimony, their descriptiveness of sparse and bursty user activity along the dimensions frequency and amplitude, and their contribution in a clustering experiment. After this final feature selection step, we chose three activity-based time series features, which capture exactly these kinds of behavior: the ratio of unique non-zero values to time series length, a Boolean feature describing if the activity time series has more than five peaks, and another Boolean feature measuring if the time series has maximum peaks of equal height.

Feature computation. We compute activity peaks as values that are larger than other values in their direct neighborhood of the previous and next observations of the activity time series (cf. Figure 1(b)). Note that, with this definition of a peak, we do not impose a minimum peak height, both in absolute terms, as well as relative to the peak's neighboring values. For example, assume a user posts a question once in January, responds twice to some other question in February, and then asks one other question in March. The activity-based time series corresponding to this user would thus feature one peak in February. For the binary feature related to the number of peaks, we settled on the threshold five (i.e., the feature measures if an activity-based time series has (or not) more than five peaks). The reason for the threshold five is that this threshold corresponds to the average 90th quantile of the number of peaks per activity-based time series. Thus, this feature separates a minority of users with high volumes of contributions (as measured by peaks in activity) from the majority of other users.

The other binary feature, checking if a user's activity-based time series has a duplicate maximum peak, captures regularity in behavior pattern, both of sparsely active users posting just two questions (or answers) in separate occasions or of regularly and frequently active users with consistently regular activity patterns. Hence, this binary feature combines with the other one to separate users along the dimensions of activity volume over time and activity regularity.

Finally, the third feature, ratio of unique non-zero values to time series length, allows for finer shades of distinction between highly and regularly active from less and irregularly active users, as this continuous feature encompasses both sides of this spectrum. On the one hand, frequently highly (sporadically less) active users will tend to have a low (high) such ratio, but variations of these two extremes are possible (e.g., regular and highly active users) and interesting for later analysis. We use a ratio (and not the absolute count of unique non-zero values) to ensure all our features are defined over the interval $[0, 1]$, which allows for better comparison in Euclidean distance-based clustering methods.

Alternative feature extraction methods for clustering. We note here that other authors, namely Witten and Tibshirani [53] and Fulcher et al. [19], propose a couple of alternatives to our feature selection approaches for clustering and, respectively, time series analysis in general (with applications to time series comparison and clustering). Although the former authors do not focus on time series explicitly and the latter do not specifically address count time series and value sparseness, we believe those approaches could be used with sparse count time series and, in particular, to the activity time series we observe. Hence, we applied Witten and Tibshirani's K-Means and hierarchical-based sparse clustering approaches to our activity-based count time series, thereby

taking care to adjust the hyperparameters to our data. We report the results of the application of their method in comparison to our own later on. Fulcher et al.'s approach, however, would result in the same feature filtering approach we outlined above, as their proposed features include much of the information-theoretic, model-based, and locality-focused features we explicitly excluded from our analysis. To sum up, the features we find lead to clearly separated, interpretable clusters, as we see in the upcoming section.

3.3 Clustering Process

The combination of these three features we propose allows us to derive cohesive, well-separated and interpretable clusters. Each one of the binary features partitions the space of activity-based time series in two sets. Those two features thus yield, when combined, four clusters, since they do not capture the same properties of activity-based time series. The third feature, ratio of unique non-zero values, is continuous and takes values in the interval [0, 1]. This continuous feature measures more granular variations in activity frequency than those afforded by having just the two binary features. Using this continuous feature by itself, however, does not separate the space in clusters.

We employ the commonly-used unsupervised clustering algorithm K-Means [35], with k-means++ cluster center initialization [5], to group similarly active users. We measure time series similarity with the Euclidean distance on the extracted features. We briefly explain K-Means: The algorithm begins with a random initialization of K cluster centers, so-called centroids, as K randomly chosen vectors from an input space. The algorithm labels input vectors with the centroid most similar to each of them. Then, it reassigns all K cluster centroids to each cluster's mean vector. These two steps are repeated until convergence [35]. We also experimented with both variations of K-Means such as bisecting K-Means [47] as well as with other clustering algorithms such as Ward hierarchical clustering [52] and Density-Based Spatial Clustering of Applications with Noise (DBSCAN [15]), but those efforts yielded similar results, as we see below.

Selecting the number of clusters. The main hyperparameter of K-Means is K, representing the number of clusters, which is often a function of expert knowledge or other external factors. However, we aim to learn a suitable number of clusters directly from the data. Therefore, we automate the estimation of K. The *elbow method* [29] executes K-Means clustering for a range of values of K and stores the mean distance of centroids to the clustered input, which is termed the cost function, for each K. With the elbow method, one then graphically identifies the optimal K as the value K^* where the cost function, plotted as a function of K, results in the best tradeoff between low cost and maximum cost reduction with respect to $K^* - 1$'s cost. Intuitively, this description of K* matches the point where the cost function forms an "elbow," hence the method's name.

We employ a purely numeric method to choose the value for K, since we aim to automatize the search for K^* for a large number of time series. Similarly to the elbow method, we estimate a statistic on the quality of the clustering for a range of values of K. Thus, we pick the value K^* that maximizes the *silhouette coefficient* [26, 40], which combines statistics on the cluster cohesion (intra-cluster) and separation (inter-cluster) into a single value. Cluster cohesion, represented by a_i, captures the mean distance of an element i in a cluster to other elements in the same cluster. Cluster separation, represented by b_i, denotes the mean distance of an element i in a cluster to other elements in the closest neighboring cluster. These two factors form the equation for the silhouette coefficient $s_i = (b_i - a_i)/max(a_i, b_i)$, where $-1 \leq s_i \leq 1$. A high silhouette coefficient implies that the cluster distance of i to other elements in its cluster is low, relative to the mean distance to elements in the next nearest cluster, suggesting the correct assignment of i. The opposite holds for low silhouette coefficient values.

With the application of K-Means for $K = 2, \ldots, 10$ on the extracted features, we look for K^*. We validate separation and cohesion of the K^* clusters graphically with Principal Component Analysis (PCA) projections into a two-dimensional space (cf. Figure 1(c)). To check the validity of the clustering obtained with K-Means, we compare its performance with a random clustering baseline, which randomly assigns each input vector to one of K clusters. Furthermore, as previously mentioned, we compared K-Means with other clustering algorithms: bisecting K-Means, Ward hierarchical clustering, and DBSCAN.

Measuring clustering performance. To measure the clustering performance, we first perform random clustering as a baseline. The random clustering yields $K^* = 2$ with average silhouette coefficient values in the interval $[-0.05, 0.02]$. We then cluster activity-based time series of our datasets and obtain significantly better results. For all 50 Stack Exchange instances, we obtain average silhouette coefficient values of at least 0.9 for K^*. $K^* = 4$ for 39 of our 50 Stack Exchange instances. The remaining 11 Stack Exchange instances feature a strictly higher optimal number of clusters between 6 and 10.

Our experiments with other clustering approaches yielded very similar results: Bisecting K-Means and Ward hierarchical clustering return the same average silhouette coefficient values up to a factor of 10^{-3} and agree on K^* for all 50 datasets. DBSCAN, however, had lower average silhouette coefficient values of at least 0.89, but also yielded $K^* = 4$ for the same 39 datasets as before. However, on those 11 Stack Exchange instances with $K^* > 4$, disagreement in both average silhouette coefficient values as well as K^* was highest in the comparison with the other clustering algorithms.

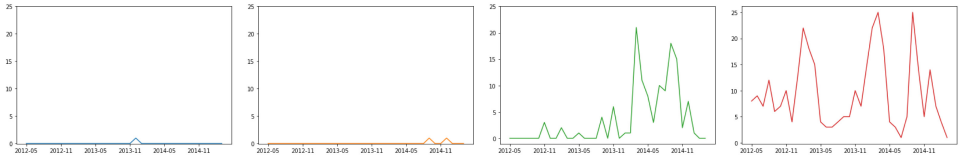
We attribute the similarity of results for different clustering algorithms to the two binary features strongly influencing the distribution of user activity-based time series features in the three-dimensional feature space. We observe lower silhouette values and disagreement in K^* between the clustering algorithms in cases where the continuous feature plays a larger role in the distribution of a Stack Exchange instance's user activity-based time series features. DBSCAN seems most sensitive to these feature distributional changes, as its density region-based clustering approach consistently groups points with different binary feature values but similar continuous feature values in one cluster. This clustering behavior, in turn, leads to lower silhouette and lower K^* than other clustering algorithms agree upon. We stress that other binary features might have led to the same high silhouette coefficient results, but they led to ultimately different results and interpretation of user behavior.

The best results we achieved with the K-Means and hierarchical sparse clustering approaches by Witten and Tibshirani yielded $K^* = 2$ and silhouette coefficient values of a maximum of 0.88 and significantly lower for all $K > 2$. We believe tailoring these algorithms to find more granular structure in sparse count time series data such as ours to be an interesting avenue of future work.

Finally, in two-dimensional projections of the clusters with PCA, we observe clear graphical separation for the K^* clusters in most Stack Exchange instances.

3.4 Analyzing Cluster Properties

We analyze the clusters we obtain to better understand the activity composition captured by K-Means. To that end, we start by computing basic descriptive statistics on the clusters, such as their size, as measured by the number of users per cluster. Further, we plot the activity-based time series closest to each centroid and thereby visualize typical activity profiles for each cluster. We then visually inspect the sum of the activities in each of the clusters to discern overall cluster group dynamics. We corroborate this visual inspection with a quantification of the relative sizes of the clusters as the fraction of a cluster's activity in total activity. Finally, we look for commonalities in these patterns between Stack Exchange instances, and assess and discuss their practical relevance in Q&A community building efforts.



(a) Non-Recurring (b) Sporadic (c) Frequent (d) Permanent

Fig. 2. **Activity Archetypes.** We illustrate typical profiles of the activity-based time series nearest to K-Means centroid for $K^* = 4$. Users of the *Non-Recurring Activity Archetype* (a) often feature one single, isolated peak of activity. Users of the *Sporadic Activity Archetype* (b) typically exhibit a few isolated activity peaks of equal height. Users of the *Frequent Activity Archetype* (c) show varying but regular activity over time. Finally, repeatedly high levels of activity over time characterize users of the *Permanent Activity Archetype* (d). In short, we observe that user activity can be grouped into these four activity profiles, which mainly capture different degrees of frequency in user activity.

4 RESULTS

4.1 Activity Archetypes

For all Stack Exchange instances with $K^* = 4$, we observe four commonly occurring types of temporal user activity patterns, which we term *Activity Archetypes* (see Figure 2). The patterns of these time series are representative of the four *Activity Archetypes*, which we describe in ascending order of activity frequency and volume.

In general, users of the ***Non-Recurring Activity Archetype*** (see Figure 2(a)) exhibit one prominent peak of activity. Taking the median value over all Stack Exchange instances, we find typical users of the *Non-Recurring Activity Archetype* post 1 question and write 1.27 answers or comments in a median of 1 active month (i.e., the number of months in which a user posted at least one question or answer in a given Stack Exchange instance). Furthermore, their median tenure length, as measured by the difference between their first and last dates of activity (i.e., writing a question or answer), is less than 1 month. Users of the *Non-Recurring Activity Archetype* comprise on average 88.4% of total user count of a Stack Exchange instance. This majority of users thus typically posts a question, follows up on it with discussion with the rest of the community in a short, concentrated period of time, and does not return. This suggests users of the *Non-Recurring Activity Archetype* type have one or two concrete asking needs, which, after some discussion, are satisfied, completing the user's participation in the community. User "Amit Kumar Gupta"[9] of the Stack Exchange instance *cogsci*[10] and his question on types of memory and ensuing discussion exemplifies this behavior by users of the *Non-Recurring Activity Archetype*. For better comparison with other *Activity Archetypes*, we use the Stack Exchange instance *cogsci* for further user examples.

The ***Sporadic Activity Archetype*** (see Figure 2(b)) features higher activity levels than the *Non-Recurring Activity Archetype*. Users of the *Sporadic Activity Archetype* write a median of 2.09 questions and 2.26 answers or comments in a median of 2.06 active months. In contrast to the *Non-Recurring Activity Archetype*, the median tenure length of the *Sporadic Activity Archetype* is 6.08 months and they comprise on average 10.1% of total user count of a Stack Exchange instance. Hence, in comparison with the *Non-Recurring Activity Archetype*, not only do users of the *Sporadic Activity Archetype* pose more questions (and answer and discuss them slightly more), but they also

[9] https://psychology.stackexchange.com/users/7338/amit-kumar-gupta.

[10] In December 2017, *cogsci* was renamed to *psychology* (source: https://biology.meta.stackexchange.com/questions/3779/cogsci-has-changed-its-name), but we refer to it by its old name for the sake of consistency with our dataset, which includes data up to February 2017.

do so throughout a remarkably longer period of time. This suggests they lurk and engage more with the Stack Exchange instance community as a whole. For an example of such user behavior, refer to user "201044"[11] of *cogsci*.

We observe significantly more activity from users of the ***Frequent Activity Archetype*** (see Figure 2(c)). Such users have a median of 19.26 questions and 28.23 answers or comments and are active in a median of 12.09 months out of median tenures of 32.31 months. The *Frequent Activity Archetype* is notably less numerous, as it accounts for an average 1.3% of total user count of a Stack Exchange instance. We observe a large gap in the activity profile of the *Frequent Activity Archetype* and the previous two, as users of the *Frequent Activity Archetype* participate in Stack Exchange instance communities with greater quantity and higher frequency. Their remarkably long tenures suggest they accompany community development, despite not being active every month. Average users of the *Frequent Activity Archetype* behave like user "Greg McNulty"[12] of *cogsci*.

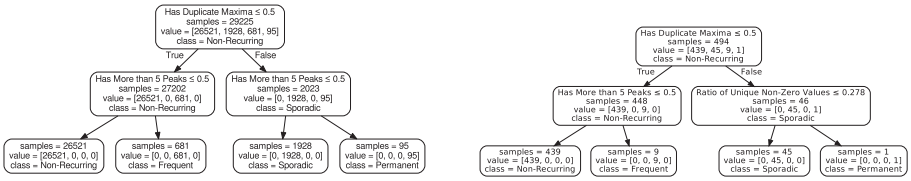
The most active group of users we identified belongs to the ***Permanent Activity Archetype*** (see Figure 2(d)). As such, this group of users posts a median of 26.12 questions and 56.68 answers or comments. They are active in a median of 13.99 months and their median tenure is 32.86 months. On average, users of the *Permanent Activity Archetype* represent just 0.2% of the total user count of a Stack Exchange instance. Although users of this archetype feature tenures comparable to those of the *Frequent Activity Archetype*, the fact they are the most active overall, combined with the fact there are very few of them, could indicate users of the *Permanent Activity Archetype* lead activity in the community. Users such as "Alex Stone"[13] of *cogsci* exemplify and could cement our reading of the role users of the *Permanent Activity Archetype* play in Stack Exchange instances.

Feature importance analysis. To support these descriptions with a quantitative assessment of the four *Activity Archetypes* in terms of our three features, we evaluated, separately on the questions and answers of all 50 Stack Exchange instances, the power of the features in explaining the four *Activity Archetypes* with ANOVA [17] and the distribution of the feature's values over the *Activity Archetypes* with random forests [6] and, in particular, also decision trees [7].

For the ANOVA approach, we fitted a generalized linear model of the three features per user as independent variables and the *Activity Archetypes* resulting from the clustering as dependent variables. As the *Activity Archetypes* represent a discrete dependent variable, we assume it is binomially distributed and we use a logit link function. The ANOVA measure of each feature's effect in such a regression model suggests every feature is significant in explaining the *Activity Archetypes*, as the corresponding p-values (for H_0: dependent variable's coefficient is 0 tested with an F-test) are all smaller than $8.47 \cdot 10^{-6}$. These results hold for both questions and answers datasets of each of the 50 Stack Exchange instances.

In a similar experiment, we fitted random forests on the three user features over all 50 Stack Exchange instances (again, separately for questions and for replies) to explain the *Activity Archetypes*. One of the outputs of random forests is estimated feature importance. In that regard, the random forests' output agrees with ANOVA's: All three features are important to classify *Activity Archetypes* in both their questions and answers activity. Moreover, random forests output a numeric estimation of feature importance for classification on a scale from zero to one: for questions-based (answers-based) activity, 0.695 (0.187) is the feature importance of the ratio of unique non-zero values to time series length, 0.271 (0.614) the one of the feature capturing if the time series has more than five peaks, and 0.034 (0.199) the one of the feature regarding duplicate maxima.

[11] https://psychology.stackexchange.com/users/7340/201044.
[12] https://psychology.stackexchange.com/users/849/greg-mcnulty.
[13] https://psychology.stackexchange.com/users/953/alex-stone.



(a) English Answers Decision Tree (b) Sustainability Answers Decision Tree

Fig. 3. **Decision trees fitted to user activity-based time series features.** We depict the result of applying decision trees to fit the extracted user answer activity-based time series clusters as a function of the three features we propose. The pictures show the number of users (samples) per decision tree node out of all clusters, i.e., *Activity Archetypes*. The decision tree for the *English* Stack Exchange instance (a) shows the defining feature values per *Activity Archetypes*, which are dominated by the two binary features, "duplicate maxima" and "more than 5 peaks." The *Sustainability* Stack Exchange instance (b) features the *Activity Archetypes* as a function of all three features we presented, as there is enough user behavior variability in *Sustainability* for the continuous feature to offset the dominance by the two binary features. Hence, all three features are important to characterize user behavior, but their importance varies with the Stack Exchange instance.

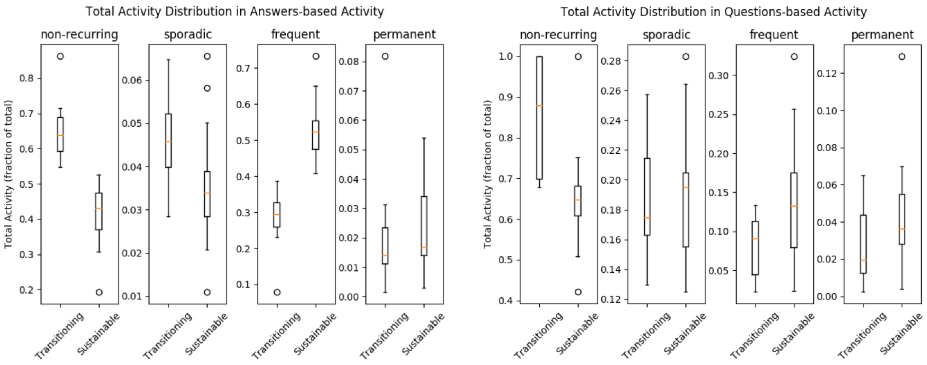
Breaking down this high-level view of all 50 Stack Exchange instances by instance allows us to visualize the feature values composing each of the *Activity Archetypes* and clusters we find. To that end, we visualize, in Figure 3, a decision tree fitted to answers-based activity of the *English* (Figure 3(a)) and *Sustainability* (Figure 3(b)) Stack Exchange instances. Furthermore, we show the number of users at each node (in total, as given by "samples;" and per class, as given by "values") in the decision tree's path, and the resulting class, i.e., Activity Archetype. We observe *English* clearly distinguishes the four *Activity Archetypes* along the values of the two Boolean features "duplicate maxima" and "more than 5 peaks." The decision tree for the Stack Exchange instance *Sustainability* makes use of the two Boolean features, as well as the continuous feature "ratio of unique non-zero values," to classify the four *Activity Archetypes*. We relate this fact to this Stack Exchange instance already including *Activity Archetypes*, but with slightly more variability in them (as captured by the continuous feature).

Note, however, that not all Stack Exchange instances feature such temporal user activity patterns as given by the four *Activity Archetypes*. The structure of the decision trees of such instances included more levels and a number of nodes on the feature "ratio of non-zero unique values." When $K^* > 4$, the temporal user activity patterns we observe represent more granular variations of the four *Activity Archetypes* we highlight, as exemplified in the legend of Figure 5(a). As we find more than 10 different variations of this kind, we do not characterize them in more detail.

4.2 Composition of Stack Exchange Instances

A total of 39 Stack Exchange instances exhibit $K^* = 4$, i.e., the four *Activity Archetypes*, in our clustering experiment. First, we categorize these Stack Exchange instances with a breakdown of their total question and answer-based activity by *Activity Archetypes*. We call this breakdown of activity by *Activity Archetypes* the *activity composition* of a Stack Exchange instance and we find two types of *activity composition*. Then, we analyze how the *activity composition* changes over time.

Derivation and analysis of the *activity composition* of Stack Exchange instances. Among those 39 instances, we observe two distinct *activity compositions* with respect to the contribution to total answer activity by users of the *Non-Recurring Activity Archetype*. Recall users of the *Non-Recurring Activity Archetype* represent the majority at an average fraction of 88.4% of total user count. Interestingly, in some instances, they do not account for the majority fraction of total answer



(a) Total activity distribution of answers-based activity time series

(b) Total activity distribution of questions-based activity time series

Fig. 4. **Distinction between *Transitioning* and *Sustainable* Stack Exchange instances.** For all Stack Exchange instances of the types *Transitioning* and *Sustainable*, we depict the fractions of total answers-based (Figure 4(a)) and questions-based (Figure 4(b)) activity generated per *Activity Archetype*. In *Transitioning* (*Sustainable*) instances, the answer-based activity is dominated by *Non-Recurring Activity Archetype* (*Frequent Activity Archetype*) users, which contribute a median fraction of 0.63 (0.52) of total answer activity. Overall, we observe stark contrasts in contributions of *Activity Archetypes* to total activity of different Stack Exchange instances.

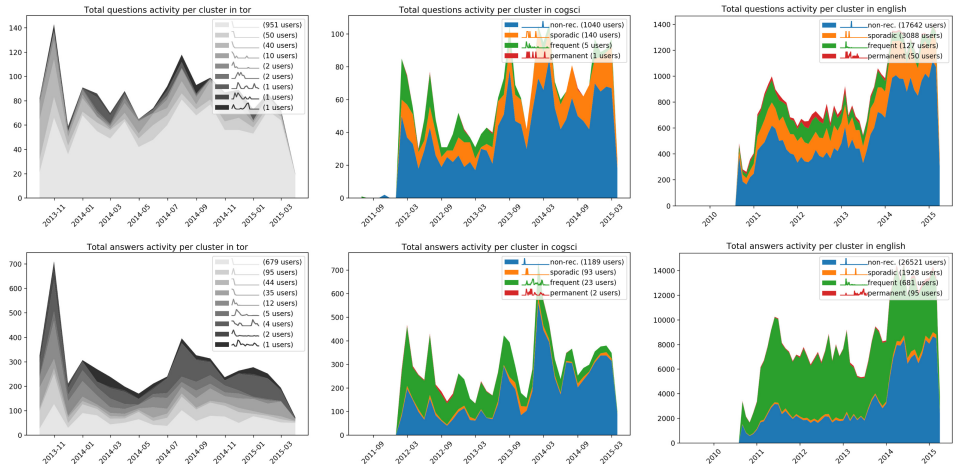
activity and play a less prominent role in total question activity. However, as might be expected, users of the *Non-Recurring Activity Archetype* dominate the *activity composition* in other instances.

Therefore, we derive two distinct groups of Stack Exchange instances by setting the following threshold: In a given Stack Exchange instance, if users of all *Activity Archetypes* except for the *Non-Recurring Activity Archetype* account for 90% or more of answer-based activity by the *Non-Recurring Activity Archetype*, we categorize the instance as *Sustainable*, otherwise as *Transitioning*. We experimented with variations of the 90% threshold in the range [85%, 95%], but we did not arrive at remarkably different results and conclusions. Recall our dataset includes, besides the 39 Stack Exchange instances with $K^* = 4$, a total of 11 instances with $K^* > 4$. We name this group of instances *Emerging*, but, for now, we focus on *Sustainable* and *Transitioning* instances. As we discuss later, this naming choice correlates with key developmental characteristics of the two types of Stack Exchange instances. Using this criterion, we identify 26 *Sustainable* Stack Exchange instances (of which 5 still are in the Area 51 incubator)[14] and 13 *Transitioning* Stack Exchange instances (with 8 of them still in the Area 51 incubator).[15]

We compare the *activity composition* of the *Transitioning* and *Sustainable* Stack Exchange instance types in more detail in Figure 4. We observe the highest proportion of answers-based activity in *Sustainable* Stack Exchange instances comes from the *Frequent Activity Archetype*, whereas the *Non-Recurring Activity Archetype* generates most (questions and) answers-based activity in *Transitioning* Stack Exchange instances. We draw these conclusions from the relatively

[14]The 26 *Sustainable* Stack Exchange instances are *english*, *unix*, *softwareengineering*, *gaming*, *tex*, *stats*, *wordpress*, *physics*, *mathoverflow*, *sharepoint*, *scifi*, *ux*, *webmasters*, *graphicdesign*, *workplace*, *salesforce*, *cs*, *bicycles*, *skeptics*, *christianity*, *sound*, *history*, *gardening*, *linguistics*, *outdoors*, and *tridion*, with *history*, *gardening*, *linguistics*, *outdoors*, and *tridion* still being in the Area 51 incubator as of 02/13/2017.

[15]The *Transitioning* group of Stack Exchange instances consists of *bitcoin*, *chemistry*, *chess*, *codereview*, *cogsci*, *music*, *opendata*, *philosophy*, *poker*, *reverseengineering*, *space*, *sports*, and *sustainability*. As of 02/13/2017, *chemistry*, *codereview*, *music*, and *philosophy* have left the Area 51 incubator.



(a) *Emerging* (b) *Transitioning* (c) *Sustainable*

Fig. 5. **Temporal dynamics of the three types of *activity composition* of Stack Exchange instances: *Emerging*, *Transitioning*, and *Sustainable*.** We plot the total count of questions-based (top) and answers-based (bottom) activity of three Stack Exchange instances over time, and break these activity totals down by *Activity Archetype*. A high value of K^*, indicating the four user archetypes do not prevail, characterizes *Emerging* instances like *tor* (Figure 5(a)). Notably, we observe similarly low levels of activity in *tor* and *Transitioning* instances like *cogsci* (Figure 5(b)), with *tor* having overall declining activity and *cogsci* oscillating around positive growth. The *Sustainable* instance *English* (Figure 5(c)), however, exhibits high activity levels, and pronounced growth in activity. These aspects hint at a link between *activity composition* and overall activity development.

higher (lower) median total activity fraction values for users in the *Frequent Activity Archetype* (*Non-Recurring Activity Archetype*) in *Sustainable* instances compared to *Transitioning* instances (Figure 4(a)). Furthermore, we highlight the relative importance of the *Non-Recurring Activity Archetype* and the *Sporadic Activity Archetype* in questions-based activity (see Figure 4(b)): Although more so in *Transitioning* Stack Exchange instances, both still play a significant role in the *Sustainable* instance type. Differences between instance types in the role of the *Permanent Activity Archetype* are qualitatively the same as in the *Frequent Activity Archetype* but to a lesser degree, as the *Permanent Activity Archetype* accounts for a median fraction of only 0.014 (respectively, 0.018) of total questions and 0.02 (0.037) of total answers in the *Transitioning* (*Sustainable*) instance types.

Contextualization of the *activity composition* of Stack Exchange instances. We feature a graphical comparison of total activity volume and the *activity composition* in instances representative of the three instance types in Figure 5. We draw a connection between the *activity composition* and key developmental statistics of the instances as summarized in Table 2.

We address the ***Emerging*** group of Stack Exchange instances[16] first. *Emerging* Stack Exchange instances do not exhibit the *Activity Archetypes* defined in Section 4.1, but instead feature more variations thereof. In general, *Emerging* instances are among the newest, least active, and smallest out of the 50 instances we consider: three out of five smallest instances listed in Table 2 belong

[16]The *Emerging* group of Stack Exchange instances consists of *arduino*, *buddhism*, *earthscience*, *ebooks*, *freelancing*, *ham*, *joomla*, *lifehacks*, *puzzling*, *startups*, and *tor*. Only *puzzling* has left the Area 51 incubator as of February 13, 2017.

Table 2. Statistics on Largest and Smallest Stack Exchange Instances

Instance name	Instance type	Users	Activity	Months	Trend slope
english	*Sustainable*	37,125	522,128	70	0.013
unix	*Sustainable*	36,397	390,930	80	0.012
softwareengineering	*Sustainable*	35,816	467,234	80	0.006
gaming	*Sustainable*	34,641	321,857	68	0.007
tex	*Sustainable*	31,039	624,166	80	0.014
poker	*Transitioning*	594	5,185	39	−0.002
earthscience	*Emerging*	578	5,981	12	−0.040
sustainability	*Transitioning*	555	5,274	27	−0.015
ebooks	*Emerging*	501	3,094	16	−0.041
ham	*Emerging*	473	5,023	18	−0.037

For the top and bottom five Stack Exchange instances with most and respectively least users, we list a number of statistics, sorted by the number of users: Instance type, number of users, total activity (i.e. sum of questions and answers), age in months and the slope of the trend of total activity (dependent variable) per month and year (independent variables). The top five Stack Exchange instances are all of the *Sustainable* type, and feature a positive growth trend. In contrast to those instances, the bottom five Stack Exchange instances are either *Emerging* or *Transitioning* and have dwindling growths (negative trend slope).

to the *Emerging* group. Moreover, these instances feature an overall negative activity growth,[17] i.e., these instances' activity levels drop on average. Furthermore, 10 out of 11 *Emerging* Stack Exchange instances are still in the Area 51 incubator. Figure 5(a) illustrates a typical activity profile of *Emerging* instances, as exemplified by the Stack Exchange instance *tor*.

As previously discussed, in ***Transitioning*** Stack Exchange instances, users of the *Non-Recurring* and *Sporadic Activity Archetypes* generate the most activity, with the *Sporadic Activity Archetype* acting more prominently in questions-based than answers-based activity. The activity dynamics of *Transitioning* Stack Exchange instances exhibit strong oscillations over time, as exemplified by the Stack Exchange instance *cogsci* in Figure 5(b). We note that some of the *Transitioning* Stack Exchange instances are among the five smallest datasets in our analysis, as Table 2 indicates. Other *Transitioning* instances vary considerably in numbers of users and age, and the Stack Exchange instance *codereview* has one of the largest user bases with 19,140 users and features very high activity levels at a total of 157,593 questions and answers. Overall, however, the average activity growth of all *Transitioning* instances is about 0. In other words, these instances' activity levels oscillate (and stagnate) over the course of their existence.

On the other hand, in ***Sustainable*** Stack Exchange instances such as *english*, users of the *Frequent Activity Archetype* generate the most answers-based activity, despite, again, representing a reduced percentage of total user base. In general, *Sustainable* Stack Exchange instances are among the oldest, most active ones, feature with the highest number of users (cf. Table 2), and exhibit high activity levels and a steady growth of activity (cf. Figure 5(c)). Furthermore, average activity growth of all *Sustainable* instances is positive.

Instance type evolution over time. Now, we analyze how a Stack Exchange instance's type (i.e., its *activity composition* in terms of *Activity Archetypes*) changes over time. To do so, we count the number of instances per type at different points of their existence in Figure 6. Specifically, starting with the first 6 months after inception, we categorize each instance over the course of its existence

[17] Note that we estimate activity growth as the slope of a linear regression on total activity (dependent variable) per month and year (independent variable) fitted with ordinary-least-squares and normalized with a min-max transformation for comparing instances (see Table 2).

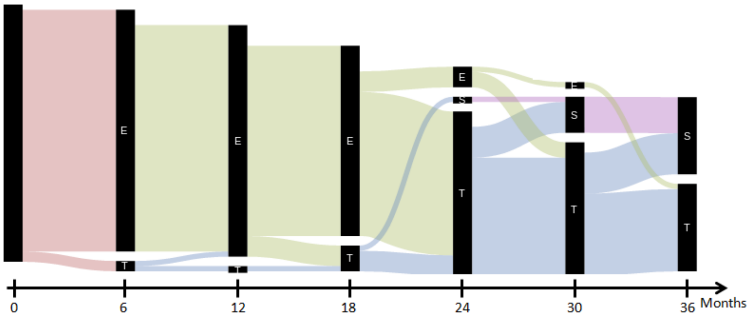


Fig. 6. **Temporal evolution of Stack Exchange instance types.** We count the number of Stack Exchange instances per type (with "E" standing for *Emerging* (in green), "T" for *Transitioning* (in blue), and "S" for *Sustainable* (in pink)) every 6 months until the first 3 years of the instances' existence. We highlight that *Sustainable* instances take at least a couple of years to develop, and *Emerging* instances typically grow to the *Transitioning* type in less than 2 years. This temporal process suggests *activity compositions* of Stack Exchange instances shift and mature with time.

(in increments of 6 months until 3 years) as *Emerging* (or "E" in Figure 6), *Transitioning* ("T"), or *Sustainable* ("S"). Note that 49 out of 50 Stack Exchange instances are at least 6 months old, but only 32 are at least 3 years old.

We note that, after the first 6 months, almost all instances are of the *Emerging* type. We observe that there is no Stack Exchange instance that immediately transitions from the *Emerging* to the *Sustainable* type. Most instances need at least 18 to 24 months before moving from the *Emerging* type to the *Transitioning* type. Roughly 47% or 15 out of 32 Stack Exchange instances evolve to the *Sustainable* type by their third year.

This developmental process suggests the *activity compositions* we propose correspond to maturity stages of Stack Exchange instances.

5 DISCUSSION

We discuss the *Activity Archetypes* with respect to their impact, comparable user behavior identified in related literature, and their role in Stack Exchange instance development. Basing on that discussion, we derive practical implications for Q&A community managers to optimize their community development efforts.

Impact of *Activity Archetypes* in the context of user characterizations in related work. We first discuss similarities and key differences in *Activity Archetypes* and the two or three types of temporal user activity patterns other authors typically mention in their studies of online Q&A instances [20, 36, 42]. The main difference between our *Activity Archetypes* of user behavior and others lies in the *Non-Recurring Activity Archetype*. When describing least active users, namely low-profile [36], low activity [20], and less participatory [42] ones, these authors group users with both short- and long-term lurking behavior. By splitting such lurking behavior into the *Non-Recurring Activity Archetype* and the *Sporadic Activity Archetype*, we uncover a distinction in lurking behavior with respect to a fundamental aspect of users' participatory interest in a Q&A community. Specifically, users of the *Non-Recurring Activity Archetype* seem to join a Q&A community with a specific question or purpose and leave after it is fulfilled. This behavior poses a contrast to users of other *Activity Archetypes* as well as of the low activity types found by other authors: These other users participate in a Q&A community for longer periods of time, suggesting a higher interest in the Q&A community itself or at least in more of its topics. We argue for our granular

characterization of low profile user behavior due to the high impact users of the *Non-Recurring Activity Archetype* have: (i) They represent the majority of total user base and (ii) their role remains important throughout the development of Q&A communities' activity dynamics. We believe other *Activity Archetypes* could correspond more directly to user behavior described by these other authors: The *Sporadic Activity Archetype* could correspond to the occasional [20] and partially shooting star [36] user profiles, the *Frequent Activity Archetype* to the answer activist [20] and the more participatory users [42], and the *Permanent Activity Archetype* to the community activist [36] and hyperactivist [20]. In agreement with descriptions by these other authors, we see a prominent role by users of the *Sporadic Activity Archetype* as the least active type of users that at least engages with the community and thereby provides questions and, to a lesser extent, answers, spread out over time. A large gap in the activity profiles of the *Sporadic Activity Archetype* and the *Frequent Activity Archetype* makes the difference between them obvious. We argue for a distinction between *Frequent Activity Archetype* and *Permanent Activity Archetype* due to the former's high activity profile and non-negligible user count as a backbone of the community, effectively balancing the workload with the relatively few users of the *Permanent Activity Archetype*. Interestingly, Furtado et al. [20] describe users with the hyperactivist role as those that even participate in community moderation, supporting our view that users of the *Permanent Activity Archetype* act as community leaders. Note, however, that we do not claim any of these correspondences are a perfect match, as each of those authors and ourselves focus on different facets of user activity and behavior in online Q&A communities.

Regarding our feature choice to describe *Activity Archetypes*, our analysis reveals that a simple, small set of features is sufficient for separating temporal user activity patterns. We see these facts as a promising result for future Q&A activity dynamics modeling efforts—by using only a small number of parameters, models can be kept simple and interpretable (e.g., we may model user activity as a simple Poisson process), but still effective and accurate. Moreover, empirical estimation of parameters for simple models is typically easy and efficient.

Dynamics of *activity compositions*. Having reiterated the significance of our *Activity Archetypes* characterization, we discuss the importance of their roles at different stages of a Stack Exchange instance's development.

We observed that Stack Exchange instances of the *Transitioning* and *Sustainable* type exhibit an oscillating and respectively growing flow of question activity coming mostly from the *Non-Recurring* and *Sporadic Activity Archetypes*. We thus believe initially setting low entrance barriers and providing incentives for one-time and infrequent external impulses, in the form of participation by the users of the *Non-Recurring Activity Archetype* and the *Sporadic Activity Archetype*, as they form the basis for successful activity development of Q&A communities. Research [30, 43, 48] on the roles played by novice users and their activity dynamics in online collaborative communities such as Wikipedia supports our reading regarding the importance of these less active users. We believe the second ingredient for successful activity development lies in the community's reaction to activity by both less active user archetypes, since both expect answers and comments from the communities they engage with. We observe, in *Sustainable* instances, that users of the *Frequent* and *Permanent Activity Archetypes* bear the bulk of this workload, whereas in *Transitioning* it's the users of the *Non-Recurring* and *Sporadic Activity Archetypes* themselves. We note that publicly available statistics from Area 51 datasets [46] and previously mentioned work [20, 36] stress the importance of this core community from the *Frequent* and *Permanent Activity Archetypes*. Other studies focusing on knowledge sharing dynamics of other online Q&A communities [1, 37] even correlate higher activity levels with question answering performance, thus reinforcing the key role the most active users play. To summarize, our results suggest activity growth-inducing structures prominently feature a core of recurring users, experts, and community leaders of the *Frequent*

Activity Archetype and *Permanent Activity Archetype*, and a steady, numerous stream of users of the *Non-Recurring* and *Sporadic Activity Archetypes*.

In *Emerging* Stack Exchange instances, the dynamics of the four main *Activity Archetypes* have not formed yet, so other clusters of activity types, not belonging to any of the four archetypes, dominate. We reason that this is a direct consequence of *Emerging* instances simply lacking users and time required to establish structures to support these user activity dynamics.

Practical implications for growing Q&A communities. Based on our analysis, we propose a series of measures for operators of Q&A instances to focus on as they grow and foster their communities.

For an operator starting a Q&A community, our analysis indicates her first priority should be to gather users interested in the community she oversees, possibly via integration with other topic related online communities [37]. To do so, we suggest the community operator initializes the community in a controlled beta phase, as previously proposed ([36], [46]), with the intent of establishing simple sets of rules, which ease the load on operators and moderators and ensures newcomers feel welcome. Newbie corners and close monitoring of this initial phase, to, e.g., continuously improve ease-of-access and not introduce counterproductive overregulation [48], should help the community improve activity levels beyond beta status. Although Kittur et al. [30] suggest experts were crucial to bring content and utility to the early days of Wikipedia, our results indicate young Q&A instances, i.e., those less than 2 years old (cf. Figure 6), also benefit strongly from bursty activity by users of the *Non-Recurring Activity Archetype* and *Sporadic Activity Archetype*.

This does not imply, however, that the *Frequent Activity Archetype* and *Permanent Activity Archetype* should be neglected, as developing and rewarding recurrent participation in a Q&A becomes more important over the mid-term of 18 to 36 months (again, cf. Figure 6). In this phase, community operators could invest in a badge and gamification system to elicit more participation and community spirit by users of the *Frequent Activity Archetype* and *Permanent Activity Archetype*, as these badges and gamification elements have been shown to enhance participation by these types of users [4, 14, 24, 37, 60]. Furthermore, community question routing systems, such as the one proposed by Srba et al. [45], should help matching questions to answerers, thus ensuring the needs of the users of the *Non-Recurring Activity Archetype* and *Sporadic Activity Archetype* are met. Finally, operators should gather feedback from their user base continuously [36] and engage leaders, potentially such as those of the *Permanent Activity Archetype*, to help with community moderation [20].

6 CONCLUSIONS

In this article, we uncover temporal activity patterns in 50 Stack Exchange Q&A instances at both the user and instance levels. To achieve this, we start by representing user activity in those instances as time series, which comprise the total count of users' questions and answers over time. We extract representative features from these time series to better cluster them and to derive an optimal numbers of clusters. These clusters represent a set of four *Activity Archetypes*, which characterize users mainly according to the frequency of participation in a Q&A community. Then, we break down activity in Stack Exchange instances by the different *Activity Archetypes*, which allows us to recognize three instance types: *Sustainable*, *Transitioning*, and *Emerging*. *Sustainable* instances have the highest levels of activity and the largest number of active users. Their success correlates with a small but strong backbone of users of the *Frequent* and *Permanent Activity Archetypes*, reacting to a steady flow of users from the *Non-Recurring* and *Sporadic Activity Archetypes*. We find that *Emerging* and *Transitioning* Stack Exchange instances either completely lack or are in the process of establishing such activity profiles. Our *Activity Archetypes* and Stack

Exchange instance characterization allow us to measure online Q&A instance health and success. We provide a methodology for community managers of Q&A instances to detect the maturity stage of their communities, and we recommend activity composition structures for them to aim for, as well as concrete steps to take to help their communities mature from one stage to the next.

Besides the aforementioned limitation regarding feature selection and corresponding clustering quality and interpretation (as other binary features might yield equally good clustering quality but other interpretations), we reflect on the generalization and practical implications of our approach with respect to other Q&A datasets. Although our proposed features are suitable for capturing general bursty types of activity found in Q&A instances, these features might need tailoring in the application to Q&A communities besides Stack Exchange instances. In particular, the threshold for the feature based on the number of activity peaks will vary depending on the Q&A platform, which is why we defined it as a data-dependent percentile value. Moreover, the choice of granularity of time series aggregation, in our case monthly, must be taken with care, since too coarse a temporal resolution will hide burstiness and activity peaks, and too granular a resolution will lead to time series with longer periods of inactivity and thus a less distinguishable "ratio of unique non-zero values" feature. However, once time series granularity and our proposed features have been adjusted for a potentially new dataset, we expect the clustering to yield comparable results, since our proposed features yield clear-cut separated clusters. Therefore, we expect practitioners working with our proposed approach to be able to gauge their extension to their datasets, in particular, in case of modifications to our proposed features, by evaluating the resulting clustering quality and checking if it is comparable to the one we report. One last noteworthy limitation regards the fact the Stack Exchange instances we analyzed do not become completely inactive. As such, we refrain from discussing the generalization of our proposed approach in the case of "death" of Stack Exchange instances.

Naturally, empirically verifying the generalization of our method to other Q&A platforms would be of great interest. Moreover, conducting small-scale real experiments would further cement our argumentation on this work's practical implications. Other future work includes mathematical modeling of activity in online Q&A communities based on the *Activity Archetypes* and their activity compositions with the aim of deriving further recommendations for operators to assess and optimize their online presence. Finally, enhancing our analysis to include quality-related aspects of activity in Q&A communities would be of great interest.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback on the manuscript. Tiago Santos is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

REFERENCES

[1] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge sharing and Yahoo answers: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, 665–674.

[2] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. Time-series clustering—A decade review. *Inf. Syst.* 53 (2015), 16–38.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 850–858.

[4] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 95–106.

[5] David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

[6] Leo Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.

[7] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees*. CRC press.
[8] Grégoire Burel and Yulan He. 2013. A question of complexity: Measuring the maturity of online enquiry communities. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, 1–10.
[9] Jeffrey Chan, Conor Hayes, and Elizabeth M. Daly. 2010. Decomposing discussion forums and boards using user roles. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*. 215–218.
[10] Maximilian Christ, Andreas W. Kempa-Liehr, and Michael Feindt. 2016. Distributed and parallel time series feature extraction for industrial big data applications. *Arxiv Preprint Arxiv:1610.07717*.
[11] Denzil Correa and Ashish Sureka. 2013. Fit or unfit: Analysis and prediction of 'closed questions' on stack overflow. In *Proceedings of the 1st ACM Conference on Online Social Networks*. ACM, 201–212.
[12] Denzil Correa and Ashish Sureka. 2014. Chaff from the wheat: Characterization and modeling of deleted questions on stack overflow. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 631–642.
[13] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 307–318.
[14] David Easley and Arpita Ghosh. 2016. Incentives, gamification, and game theory: An economic approach to badge design. *ACM Trans. Econ. Comput. (TEAC)* 4, 3 (2016), 16.
[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, 226–231.
[16] Christos Faloutsos, Mudumbai Ranganathan, and Yannis Manolopoulos. 1994. *Fast Subsequence Matching in Time-series Databases* 23, 2 (1994).
[17] John Fox. 2015. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications.
[18] Tak-chung Fu, Fu-lai Chung, Vincent Ng, and Robert Luk. 2001. Pattern discovery from stock time series using self-organizing maps. In *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*. AAAI Press, 26–36.
[19] Ben D. Fulcher, Max A. Little, and Nick S. Jones. 2013. Highly comparative time-series analysis: The empirical structure of time series and their methods. *J. R. Soc. Interface* 10, 83 (2013), 20130048.
[20] Adabriand Furtado, Nazareno Andrade, Nigini Oliveira, and Francisco Brasileiro. 2013. Contributor profiles, their dynamics, and their importance in five Q&A sites. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, 1237–1252.
[21] Rich Gazan. 2006. Specialists and synthesists in a question answering community. In *Proceedings of the Annual Meeting on Association for Information Science and Technology*. 1–10.
[22] Rich Gazan. 2007. Seekers, sloths and social reference: Homework questions submitted to a question-answering community. *New Review of Hypermedia and Multimedia* 13, 2 (2007), 239–248.
[23] Ville Hautamaki, Pekka Nykanen, and Pasi Franti. 2008. Time-series clustering by approximate prototypes. In *Proceedings of the 19th International Conference on Pattern Recognition*. IEEE, 1–4.
[24] Nicole Immorlica, Greg Stoddard, and Vasilis Syrgkanis. 2015. Social status and badge design. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 473–483.
[25] Alicia Iriberri and Gondy Leroy. 2009. A life-cycle perspective on online community success. *ACM Comput. Surv. (CSUR)* 41, 2 (2009), 11.
[26] Leonard Kaufman and Peter J. Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344. John Wiley & Sons.
[27] Eamonn Keogh and Jessica Lin. 2005. Clustering of time-series subsequences is meaningless: Implications for previous and future research. *Knowl. Inf. Syst.* 8, 2 (2005), 154–177.
[28] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. 2004. Towards parameter-free data mining. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 206–215.
[29] David J. Ketchen Jr. and Christopher L. Shook. 1996. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal* 17, 6 (1996), 441–458.
[30] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *WWW* 1, 2 (2007), 19.
[31] Philipp Koncar, Simon Walk, Denis Helic, and Markus Strohmaier. 2017. Exploring the impact of trolls on activity dynamics in real-world collaboration networks. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1573–1578.
[32] Guillaume Laurent, Jari Saramäki, and Márton Karsai. 2015. From calls to communities: A model for time-varying social networks. *The Eur. Phys. J. B* 88, 11 (2015), 1–10.

[33] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. 2012. Models of user engagement. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*. Springer, 164–175.
[34] T Warren Liao. 2004. Clustering of vector time series with fuzzy c-means. In *Proceedings of the IIE Annual Conference*. Institute of Industrial Engineers-Publisher, 1.
[35] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* 28, 2 (1982), 129–137.
[36] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest Q&A site in the west. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2857–2866.
[37] Kevin Kyung Nam, Mark S. Ackerman, and Lada A. Adamic. 2009. Questions in, knowledge in?: A study of Naver's question answering community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 779–788.
[38] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. 2012. Activity driven modeling of time varying networks. *Sci. Rep.* 2 (2012).
[39] Bruno Ribeiro. 2014. Modeling and predicting the growth and death of membership-based websites. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 653–664.
[40] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.
[41] Tiago Santos, Simon Walk, and Denis Helic. 2017. Nonlinear characterization of activity dynamics in online collaboration websites. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1567–1572.
[42] Vibha Singhal Sinha, Senthil Mani, and Monika Gupta. 2013. Exploring activeness of users in QA forums. In *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 77–80.
[43] Jacob Solomon and Rick Wash. 2014. Critical mass of what? Exploring community growth in wikiprojects. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*. AAAI Press, 476–484.
[44] Ivan Srba and Maria Bielikova. 2016. A comprehensive survey and classification of approaches for community question answering. *ACM Trans. Web (TWEB)* 10, 3 (2016), 18.
[45] Ivan Srba, Marek Grznar, and Maria Bielikova. 2015. Utilizing non-QA data to improve questions routing for users with low QA activity in CQA. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 129–136.
[46] StackExchange Area 51 Community. 2017. Tridion area 51 report. Retrieved February 13, 2017 from http://area51.stackexchange.com/proposals/38335/tridion.
[47] Michael Steinbach, George Karypis, Vipin Kumar, et al. 2000. A comparison of document clustering techniques. In *Proceedings of the KDD Workshop on Text Mining*, Vol. 400. Boston, 525–526.
[48] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. 2009. The singularity is not near: Slowing growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM, 8.
[49] Alexei Vázquez, Joao Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. 2006. Modeling bursts and heavy tails in human dynamics. *Physical Review E* 73, 3 (2006), 036127.
[50] Michail Vlachos, Jessica Lin, Eamonn Keogh, and Dimitrios Gunopulos. 2003. A wavelet-based anytime algorithm for k-means clustering of time series. In *Proceedings of the Workshop on Clustering High Dimensionality Data and Its Applications*. 23–30.
[51] Simon Walk, Denis Helic, Florian Geigl, and Markus Strohmaier. 2016. Activity dynamics in collaboration networks. *ACM Trans. Web (TWEB)* 10, 2 (2016), 11.
[52] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 301 (1963), 236–244.
[53] Daniela M Witten and Robert Tibshirani. 2010. A framework for feature selection in clustering. *J. Amer. Stat. Assoc.* 105, 490 (2010), 713–726.
[54] Matthias Wölbitsch, Simon Walk, and Denis Helic. 2017. Modeling peer influence in time-varying networks. In *Proceedings of the International Workshop on Complex Networks and their Applications*. Springer, 353–364.
[55] Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. ACM, 177–186.
[56] Jie Yang, Ke Tao, Alessandro Bozzon, and Geert-Jan Houben. 2014. Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*. Springer, 266–277.
[57] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2016. Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 1317–1322.

[58] Colleen Young. 2013. Community management that works: How to build and sustain a thriving online health community. *J. Med. Internet Res.* 15, 6 (2013), e119.
[59] Jun Zhang, Mark S Ackerman, and Lada Adamic. 2007. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 221–230.
[60] Jiawei Zhang, Xiangnan Kong, and S Yu Philip. 2016. Social badge system analysis. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 453–460.

Received June 2017; revised September 2018; accepted December 2018

3.4 Self- and Cross-Excitation in Stack Exchange Question & Answer Communities

This article proposes a partial answer to both the second and the third research questions. The second research question concerns models for user excitation, which this article addresses via Hawkes process models. In particular, this work outlines a method to cope with technical difficulties when applying Hawkes processes in practice: To ensure stationarity in activity dynamics even in the presence of exponential growth, I proposed a changepoint detection model supported by constant linear regressions, and to estimate decay parameter values, I employ a Bayesian hyperparameter optimization approach. These technical advances facilitate addressing the third research question, which concerns the link between user excitation and temporal patterns to the evolution of peer production systems. To measure this link, this work fits Hawkes processes to activity dynamics of prominent user types in Stack Exchange Q&A communities with growing and declining activity levels and with different topical focuses. Further, I also validate any differences in user excitation via statistical (permutation) tests as well as prediction experiments.

Summarizing the results, this article uncovers differences in user excitation at different developmental stages in the Q&A communities, and underscores excitation differences between communities with similar growth trajectories but different topical focus. In particular, excitation in the early stages of growing communities appears to be more power-user centric than in declining ones, and the reliance by growing communities on power user excitation gives way to a higher dependency on casual user excitation in the long-run. For communities with comparable growth trajectories but different topics, I find higher casual user excitation in communities devoted to STEM-related topics than in those dedicated to the humanities, where excitations between power and casual users dominate. As the excitation effects I uncovered with respect to the comparison of growing-vs.-declining communities are also predictive of future activity dynamics, these results ascribe an important role of timing the user mix correctly and of promoting certains kinds of user interactions, in order to optimize growth in activity dynamics.

Self- and Cross-Excitation in Stack Exchange Question & Answer Communities

Tiago Santos[1], Simon Walk[2], Roman Kern[1,3], Markus Strohmaier[4], Denis Helic[1]
[1]Graz University of Technology, [2]Detego GmbH, [3]Know-Center, [4]RWTH Aachen University
tsantos@iicm.edu,s.walk@detego.com,rkern@tugraz.at,markus.strohmaier@cssh.rwth-aachen.de,dhelic@tugraz.at

ABSTRACT

In this paper, we quantify the impact of self- and cross-excitation on the temporal development of user activity in Stack Exchange Question & Answer (Q&A) communities. We study differences in user excitation between growing and declining Stack Exchange communities, and between those dedicated to STEM and humanities topics by leveraging Hawkes processes. We find that growing communities exhibit early stage, high cross-excitation by a small core of power users reacting to the community as a whole, and strong long-term self-excitation in general and cross-excitation by casual users in particular, suggesting community openness towards less active users. Further, we observe that communities in the humanities exhibit long-term power user cross-excitation, whereas in STEM communities activity is more evenly distributed towards casual user self-excitation. We validate our findings via permutation tests and quantify the impact of these excitation effects with a range of prediction experiments. Our work enables researchers to quantitatively assess the evolution and activity potential of Q&A communities.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing*; • **Mathematics of computing** → *Stochastic processes.*

KEYWORDS

Q&A communities, Excitation effects, Hawkes processes

ACM Reference Format:
Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, Denis Helic. 2019. Self- and Cross-Excitation in Stack Exchange Question & Answer Communities. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313440

1 INTRODUCTION

Why and *how* some Question & Answer (Q&A) communities gain traction and attract activity from large numbers of users—while others do not—are questions of theoretical and practical relevance [31, 38]. Understanding how users become active in such systems, and how user activity evolves over time, can be considered an important stepping stone towards better modeling and shaping of online Q&A communities. This will allow to devise novel approaches to guide and encourage activity [12] and to support community managers in their community building efforts [2, 20].

User excitation. In this paper, we investigate *self-excitation* and *cross-excitation* of users in Q&A communities. Self-excitation reflects how a user's *own* past activity shapes her future activity, while cross-excitation reflects how *other users'* past activity influences future activity of a given user. Modeling temporal traces of user activity informs the inference of excitation effects and thus provides a first step towards deeper causal analysis of user excitation.

In the present work, we adopt point processes [7–9]—in particular Hawkes processes [14]—to leverage temporal traces of user activity as latent indicators of self- and cross-excitation. We empirically analyze 69 Stack Exchange Q&A instances where we fit a multivariate Hawkes process model[1]. With that, we are able to analyze self- and cross-excitation of users across: (i) communities with growing and declining activities, (ii) the topics of the conversations, (iii) activity types (e.g., question, answers), and (iv) activity source (e.g., power user, casual users). Subsequently, we characterize self- and cross-excitation as a function of community age. We then validate, with a range of statistical tests, the excitation effects we uncover, and we quantify their relative importance in the evolution of Q&A communities with a prediction experiment. We illustrate various types of excitation and how they generate user activity in Figure 1.

Findings. Our empirical findings emphasize the need for Q&A communities to maintain a steady core of highly cross-excited power users (i.e. very active users) reacting, particularly in a community's early stages, to the community as a whole. In thriving communities, casual users (less active users) shape each others' activity levels via cross-excitation. This suggests that growing communities are, in general, facilitating and embracing less active and casual users, thereby offering low barriers of entry. Additionally, we observe late-stage domination of self-excitation over cross-excitation, meaning that self-driven activity becomes a crucial factor in successful communities. This effect may serve as a long-term growth indicator, as this self-excitation dominance is most prominent in growing communities. Finally, we observe differences in user participation across distinct topics: Q&A communities dedicated to topics in the humanities (such as languages) are more driven by cross-excitation of power users, whereas those in STEM-related fields are not.

With our work we make the following contributions. First, we model self- and cross-excitation effects in successful and unsuccessful Q&A communities. Second, we empirically show how self- and cross-excitation manifests in communities defined by different levels of success and different topics. Third, our validation provides a foundation for building further predictive models of user activity in Q&A communities. Finally, we provide and illustrate an approach

[1]We make our code available at https://github.com/tfts/Excitation_in_QA.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313440

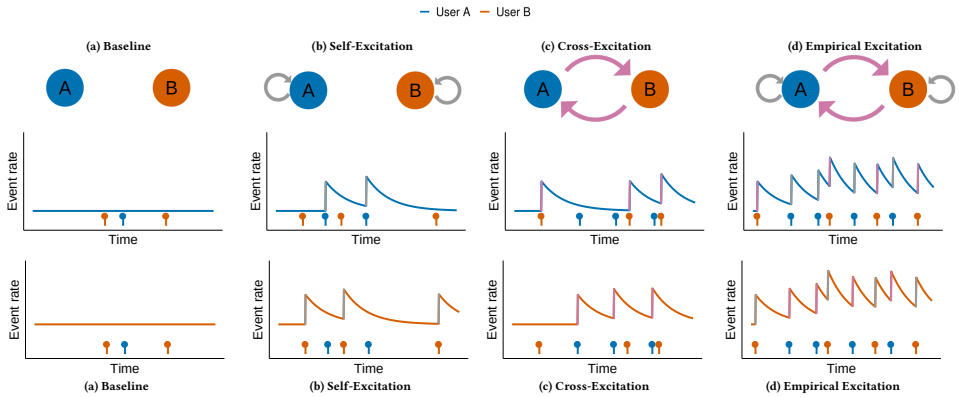


Figure 1: Excitation types. We distinguish between three drivers of user activity excitation: (a) Baseline, a constant base event rate level, (b) Self-Excitation, a proxy for increased propensity by a user to be active in the future following her past activity, and (c) Cross-Excitation, a boost to event rate triggered by activity of other users. The upper row of this Figure depicts the links between users A and B per excitation type, and the two lower rows the corresponding event rate as it reacts to the user's own and others' activity events, marked by the trees below each event rate line and colored by the corresponding user. The first three excitation types cover excitation components which combine to form (d) empirical excitation. This work characterizes and quantifies how each type of excitation manifests itself in Q&A communities and how excitation strength changes over time as a Q&A community develops.

that allows community managers to quantitatively compare long-term dynamics of their online Q&A communities—in terms of user excitation—to well-established ones.

2 HAWKES PROCESSES

A point process can be broadly defined as a collection of points randomly located in some mathematical space. Temporal point processes employ the real line, representing time, as the underlying mathematical space. For the interested reader, the work by Daley and Vere-Jones [8, 9] and by Cox and Isham [7] are comprehensive references on point process theory.

In practice, temporal point processes model the arrival of discrete events over time with the help of a *conditional intensity function* λ^*, a stochastic model for the arrival of the next event given the event history. Hawkes processes [14] are a particular class of temporal point processes, which assume a particular functional form for the intensity function. Specifically, the intensity function of Hawkes processes is in itself a stochastic process and it explicitly encodes *self-excitation*, the increase in intensity caused by past events:

$$\lambda^*(t) = \mu + \sum_{t_i < t} \alpha e^{-\beta(t-t_i)}, \tag{1}$$

where $\mu > 0$ is the baseline intensity independent on the event history, and $\alpha, \beta > 0$ establish the dependence on previous events. In particular, each previous event at time t_i increases the intensity by α, the self-excitation factor. We choose to let the intensity jumps exponentially decay at the rate β, which is a commonly used functional form of intensity decay called an *exponential kernel*.

Equation 1 describes *univariate* Hawkes processes, as they consider only the effect of past events in future event times of the same event stream. The *multivariate* generalization of univariate Hawkes processes includes not only self-excitation but also *cross-excitation*. Cross-excitation is the intensity increase that an event in one event stream implies in another event stream. More formally, let $\mathbf{N}(t) = (N^1(t), N^2(t), \ldots, N^M(t))$ be a simple multivariate point process, where each of the $N^i(t)$ is a counting process in the i-th dimension. An M-variate Hawkes process with an exponential kernel is defined by the following intensity function:

$$\lambda^{*m}(t) \quad = \quad \mu_m + \sum_{n=1}^{M} \sum_{t_i^n < t} \alpha_{mn} e^{-\beta_{mn}(t-t_i^n)}. \tag{2}$$

We write μ_m as the baseline intensity in dimension m, α_{mn} as the cross-excitation on dimension m caused by an event in dimension n and the corresponding decay rate as β_{mn}. In matrix notation, we write $\boldsymbol{\mu} \in \mathbb{R}^M$, $\boldsymbol{\alpha} \in \mathbb{R}^{M \times M}$ and $\boldsymbol{\beta} \in \mathbb{R}^{M \times M}$. Linniger [23] provides a more detailed treatment of multivariate Hawkes process theory.

Samples from multivariate Hawkes processes, which can be obtained via Ogata's thinning algorithm [27], generate self-excitation and cross-excitation effects of the kind depicted in Figure 1, where users A and B correspond to two dimensions of a Hawkes process and their event rate to the Hawkes process intensity. We observe, in both dimensions, intensity peaks corresponding to the sampled events, and we note the intensity decays exponentially until converging to the baseline intensity level $\boldsymbol{\mu}$.

Fitting Hawkes processes. Given an observed sequence of events $\{t_i\}$, we fit the parameters of Hawkes processes by maximizing its *log-likelihood*. Closed form expressions for the log-likelihood can be derived for many different types of intensity function kernels, including the exponential kernel that we assume. For a given $\boldsymbol{\beta}$, all other parameters of the process may be estimated by maximizing the log-likelihood via well-known convex optimization methods

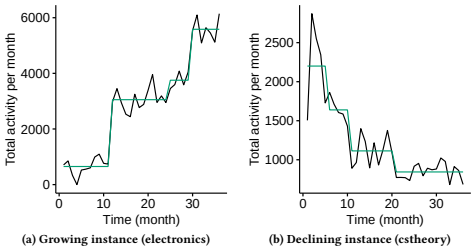
such as Levenberg-Marquardt [21]. However, fitting $\boldsymbol{\beta}$-s is a challenging task, since the likelihood functions of Hawkes processes with exponential kernels are either flat around the optimal $\boldsymbol{\beta}$ (see, for example, Upadhyay et al. [36]) or, in some other formulations of the kernel function, even non-convex in $\boldsymbol{\beta}$. In this paper, we propose an effective Bayesian hyperparameter optimization step, which allows fitting the decay-related and then the remaining parameters of a Hawkes process. Assuming $\beta_{m,n} = \beta, \forall_{1 \leq m,n \leq M}$ [12, 35], we apply the Tree of Parzen Estimators approach, as described by Bergstra et al. [4], on the convex optimization routine of log-likelihood for a given set of event sequences to estimate β. We perform 15 runs of the Bergstra et al. algorithm and keep the β yielding highest likelihood, since this effectively allows for convergence even in the presence of flat plateaus around local maxima of Hawkes likelihood as a function of β. Finally, using the learned β, we fit $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$.

Furthermore, practical fitting of Hawkes processes requires the fits to be done on *stationary* [23] periods of the corresponding count time series. Stationarity, in this context, refers to translation-invariance in the Hawkes process distribution, which implies a linear growth in the associated time series of event counts over time. However, in the time series representing activity in the Q&A communities we work with, we observe a range of non-stationary phenomena: exponential growth and decline and other sudden structural changes, such as level jumps. Therefore, we need to restrict the fitting procedure to stationary subsequences of an observed event stream. To that end, we use the time series structural change detection algorithm devised by Zeilis et al. [45]. Given a linear regression model, this algorithm returns optimal points in time for structural change in the regression model's fit to a given input time series. Using a constant regression model allows us to detect level changes in an input event count time series, and thus to segment it into stationary subsequences.

3 EXPERIMENTAL SETUP

We study user self- and cross excitation by empirically analyzing Stack Exchange instances. We distinguish activity in these datasets by two aspects: (i) activity content, which we define as questions and answers (with the latter including answers and comments), and (ii) activity source, meaning whether it originated from highly active (power) or less active (casual) users. Further, we explore the differences in excitation across communities, which we group according to two criteria: (i) growth pattern, and (ii) topical focus.

Datasets. Stack Exchange encompasses several Q&A communities, termed Stack Exchange instances, with each dedicated to Q&A on a single topic, such as computer science, the English language or movies. We extract user activity in all 159 Stack Exchange instances[2] (as of June 2017) as the timestamps of users' activity events: posts (i.e. questions) and replies (i.e. answers and comments). In a first step, we consider these instances' complete history, which spans the period from August 2008 to June 2017 and comprises a total of 22 million events. However, our analysis is independent from the calendar date a Stack Exchange instance originated, as we map the inception of each instance to a time scale starting at zero.



(a) Growing instance (electronics) (b) Declining instance (cstheory)

Figure 2: Exemplary *growing* (left) and *declining* (right) Stack Exchange instances. These two curves exemplify the monthly total activity time series of a *growing* (*declining*) instance, electronics (cstheory). Following Zeilis et al., we indicate stationarity in subsequences of both time series with the green lines. Note that *growth* (*decline*) in electronics (cstheory) was, at 757.62% (-61.63%)**, one of the highest (lowest) among Stack Exchange instances we analysed.**

[2]The Stack Exchange dataset is available at https://archive.org/details/stackexchange.

Table 1: Dataset characteristics. We show the datasets per group sorted by activity growth (top and bottom three growth percentages per group shown in parenthesis), the total number of datasets per group (#), as well as ranges for a number of descriptive statistics: the activity total as the sum of all questions and answers, the age in years and the total growth as a percentage of the level of the first subsequence found by Zeilis et al.'s algorithm. We observe a clear separation in strongly positive and negative growths (and thus also total activity) in the major distinction we draw between datasets, *growing* vs. *declining*. This distinction is remarkably less pronounced in *STEM* vs. *humanities* instances, which both feature positive and negative growths.

Dataset Group	Datasets	#	Activity total	Age (years)	Growth (%)
Stack Exchange Growing	electronics (757.62%), ru (736.42%), codegolf (510.06%), chemistry, sharepoint, academia, puzzling, tex, codereview, blender, unix, money, gis, ux, crypto, security, stats, salesforce, dba, wordpress (182.28%), opendata (174.69%), askubuntu (169.29%)	22	[7987, 1489384]	[3.08, 7.83]	[169.29, 757.62]
Stack Exchange Declining	boardgames (−28.53%), fitness (−34.56%), sound (−35.01%), productivity, tridion, parenting, pets, craftcms, webapps, spanish, cooking, ham, bricks, gardening, cstheory, expressionengine, pm, skeptics, sustainability, genealogy (−80.26%), ebooks (−81.52%), stackapps (−82.7%)	22	[3301, 117474]	[3, 7.75]	[−82.7, −28.53]
Stack Exchange STEM	electronics (757.62%), chemistry (473.48%), stats (199.18%), biology, datascience, physics, astronomy, cs, space, cogsci, earthscience, engineering, reverseengineering (0.00%), softwareengineering (−21.28%), sound (−35.01%)	15	[15759, 745674]	[2.41, 8.75]	[−35.01, 757.61]
Stack Exchange Humanities	philosophy (122.45%), english (117.76%), chinese (23.17%), music, german, mythology, portuguese, christianity, esperanto, arabic, russian, writers, buddhism (−26.62%), french (−27.91%), spanish (−50.10%)	15	[87, 896631]	[0.17, 6.83]	[−50.10, 127.47]

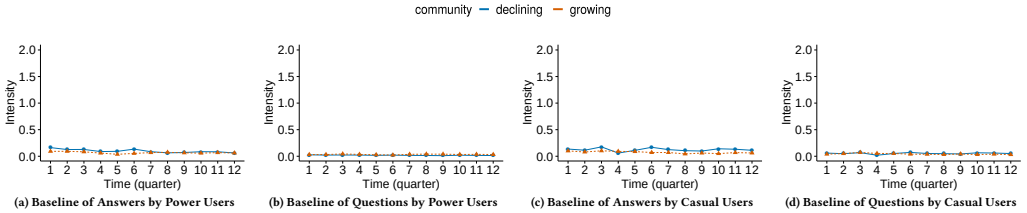


Figure 3: Low baseline excitation in *growing* and *declining* communities. Given the Hawkes process dimensions questions and answers per power and casual users, we depict the baseline parameters μ of the Hawkes processes fitted every three months over three years of *growing* (orange lines) and *declining* (blue lines) Stack Exchange instances. Error bars in this Figure and Figures 4 and 5 show bootstrapped 95% confidence intervals, many of which are too small to be visible. Note our use of the same scale in Figures (a)-(d) and throughout Figure 4: The relatively low baseline intensities in comparison to the effects depicted in Figure 4 stress that overall activity is driven by self- or cross-excitation rather than baseline intensity.

Before we group Stack Exchange instances according to our two criteria, growth patterns and topical focus, we discard all datasets with less than ten events in any period of three months, which ensures that we have enough events for the fitting procedure.

Growth pattern. In our first comparison by growth pattern, we analyze the first three years of existence of Stack Exchange instances, so we exclude datasets with durations shorter than three years. To better distinguish excitation effects driving overall activity increase or decrease in these communities, we then extract, from the remaining datasets, two groups of strongly *growing* and strongly *declining* datasets. The extraction criterion stems from our application of Zeilis et al.'s algorithm to find level structural changes in the time series of total activity count per month: We define a dataset as strongly growing (declining), if the percentage change in structural level from the first fitted window to the last fitted window is in the 80^{th} (20^{th}) percentile over all datasets. The grouping of Stack Exchange instances into *growing* and *declining* yields two groups of 22 datasets each, of which we provide descriptive statistics in Table 1. Note that the *growing* (*declining*) group only includes instances with strongly positive (negative) growth. We plot the total monthly event counts for a selected dataset from each dataset group in Figure 2 to exemplify their activity curves and the detected structural level changes. Often, there are prolonged periods of stagnancy in one structural level, both in *growing* as well as *declining* datasets. Typically, such periods vary in length.

Topical focus. For the topical comparison, we study Stack Exchange instances dedicated to *STEM* (i.e. science, technology, engineering and mathematics) and *humanities* topics. To that end, we randomly picked a set of 15 Stack Exchange instances we manually classified as STEM topics, and another 15 as humanities. The instances in these two groups vary in size and age, and feature no distinctive growth patterns, although some *humanities* instances are smaller and have shorter overall durations than *STEM* instances (cf. Table 1). In this comparison, we also analyze instances' first three years, but we do not impose a minimum duration, which leads to fewer than 15 instances per group later in time. However, the number of instances per group remains comparable over time and reaches a minimum of nine instances per group by the third year.

Hawkes process application. In the Stack Exchange instances, we distinguish between more active and less active users, which we term *power users* and *casual users*. This definition mainly distinguishes a *core* of remarkably engaged power users typically found in Q&A communities [13, 24, 41] from casual users. Thus, for each dataset individually, we count the total activity per month per user, and per activity type (question or answer) and postulate that power users are those in the 90^{th} percentile of most active users for that month. This implies that this monthly group of power users is ever-changing, as users join and leave the communities or as the users' intrinsic motivation to contribute content rises and falls over time. Note that our results changed only minimally with different percentile thresholds (i.e., 85^{th} and 95^{th}) for power user classification. To measure self- and cross-excitation per user and activity event type, we map the event stream of question and answer activity to four Hawkes process dimensions: questions by power users, questions by casual users, answers by power users and answers by casual users. For each such dimension, we work with the corresponding event timestamps at the resolution of a second.

We then follow the procedure outlined in Section 2 to fit four-dimensional Hawkes processes to each dataset group (Stack Exchange instances in the groups *declining* vs. *growing* and *STEM* vs. *humanities*). For each dataset group comparison, we begin by fitting overall β for all datasets over the first three years of their existence. Then, we perform structural level change fits on the total monthly event count, and observe a minimal window length of five months. According to our experimentation with different window lengths, specifically two to six months, we find a window length of three months is long enough to ensure we have enough events per window and do not overfit a particular window, while also short enough to capture granular changes in the evolution of the underlying Hawkes process distribution. Hence, we set the constant window length to three months (a quarter).

To measure variability in the evolution of the fitted models, we bootstrap, with 100 repetitions, the fitting procedure of all Stack Exchange instances per dataset group per window. From the resulting bootstrap distribution, we compute 95% confidence intervals for the mean value of each fitted Hawkes process parameter. We display the confidence intervals as error bars in Figures 3, 4 and 5.

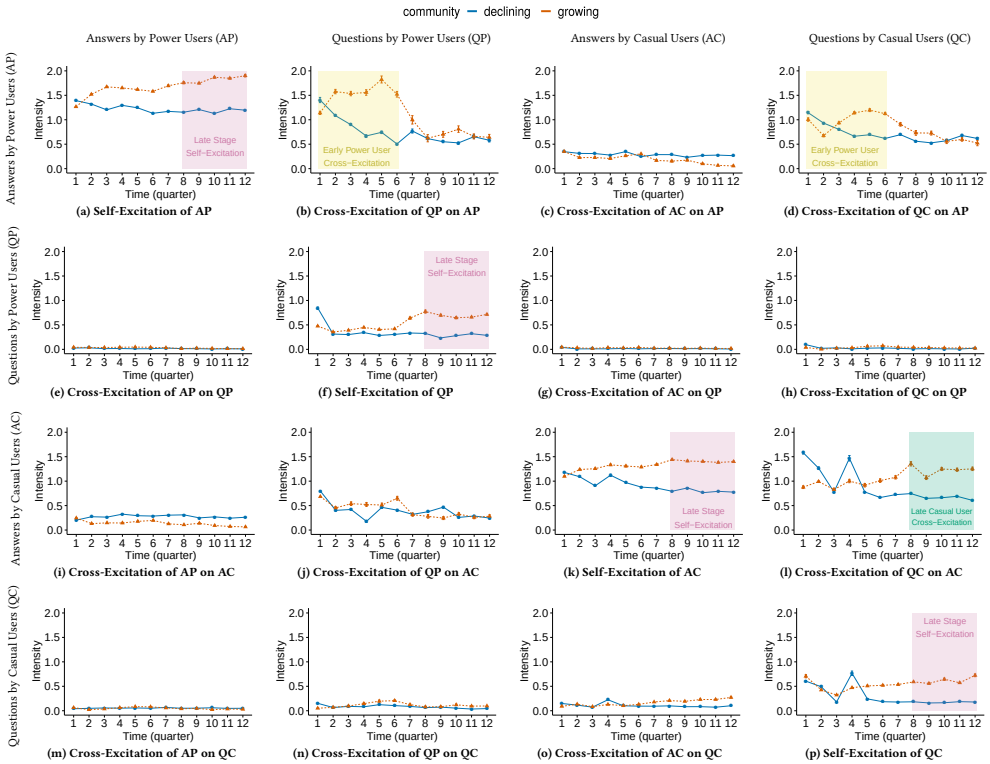


Figure 4: Excitation in *growing* vs. *declining* communities. Given the Hawkes process dimensions questions (Q) and answers (A) per power (P) and casual (C) users, we depict the α matrix of self-excitation (diagonal plots) and cross-excitation (off-diagonal plots) parameters of the Hawkes processes fitted every three months over three years of *growing* (orange lines) and *declining* (blue lines) Stack Exchange instances. The first and second row, respectively, show how answer and question intensities by power users are self- and cross-excited, and the third and fourth the influence on answer and question intensity by casual users as a result of self-excitation and cross-excitation. The yellow highlighted regions (■) of Figures (b) and (d) depict the effects of questions by power and casual users on answers by power users and thus the crucial importance of power user cross-excitation in driving early stage dynamics of *growing* instances. In Figure (l) we observe another difference between the groups: *growing* communities also thrive off interaction between casual users, as shown by long-term cross-excitation of questions by casual users on answers by casual users (cf. green region ■ of Figure (l)). In the long-term, self-excitation (pink highlighted regions ■ of diagonal entries, i.e. Figures (a), (f), (k), and (p)) is the most dominant form of excitation in all four Hawkes process dimensions.

4 EXCITATION EFFECTS

4.1 Comparison by growth pattern

The β value, fitted over 44 datasets in the *growing* vs. *declining* comparison and the whole three year period, is 2.288, corresponding to an intensity half-life of about 0.3 hours (meaning that the intensity jump of magnitude α caused by either self-excitation or cross-excitation decays to $\alpha/2$ after about 18 minutes). With a single constant β, we restrict our model to capture distributional changes in terms of baseline, self- and cross-excitation intensities, allowing us to focus on these factors as direct proxies for the role of different user groups in overall activity intensity over time.

We visualize the evolution of all baseline parameter values for questions and answers by power users and casual users in *growing* (orange) and *declining* (blue) instances in Figure 3. We depict the corresponding self- and cross-excitation parameter values in Figure 4. Note that we employ the same scales throughout both Figures for better comparison.

Low baseline intensities. The Figures 3a through 3d show the baseline intensities ($\boldsymbol{\mu}$) fitted over time. We observe roughly constant baseline intensities throughout the whole period, for both *growing* and *declining* instances. Furthermore, the baseline intensities are rather low, especially in comparison with the self-excitation and cross-excitation effects ($\boldsymbol{\alpha}$) depicted in Figure 4.

Finding: Constant and low baseline intensities suggest Q&A communities thrive off self- and cross-excitation, representing interaction between different user (power and casual) and activity types (questions & answers), rather than featuring constant levels of activity over time, independent from other activity dimensions.

Early Power User Cross-Excitation (■). We continue our analysis with the Hawkes process dimension with the highest intensity values: intensity in answer activity by power users (*first row* in Figure 4, i.e. Figures 4a through 4d). We observe, in the early stages of *growing* instances, high impact of questions by both power as well as casual users on answer activity by power users (see yellow region highlighted in Figures 4b and 4d). In contrast, in *declining* instances, especially in the yellow highlighted region, questions by both types of users elicit declining numbers of answers by power users over time, and self-excitation in answers by power users dominates over all temporal windows. Regarding question activity by power users (*second row* in Figure 4, i.e. Figures 4e through 4h), there is a clear prevalence of self-excitation intensity with respect to other cross-excitation intensities. However, we observe, albeit minor, differences in the short and medium terms between *growing* and *declining* instances, as power users in *growing* instances are more encouraged to participate with new questions as a response to questions by casual users and answers by both (see inlines of Figures 4e, 4g, 4h).

Finding: These observations suggest strong activity by power users, as a response to questions by both power users as well as, crucially, casual users, is related to increased growth in the early stages of Q&A communities. This finding suggests the importance of an active core of users to jumpstart Q&A community development.

Late Casual User Cross-Excitation (■). In Figure 4's *third row* (Figures 4i through 4l), we highlight another type of effect: Answers and discussion by casual users is driven strongly by questions also from casual users, especially in the long-term as highlighted by the green region of Figure 4l. The main difference between *growing* and *declining* instances in this dimension is, besides the intensity magnitude difference, that this cross-excitation effect loses importance in the long-term in the *declining* instances, while overall it does not in *growing* instances. We point out one interesting effect in the *fourth row* (Figures 4m through 4p), which depicts the question intensity dimensions of casual users: In the long term, there is a small increase in questions by casual users after answers also by casual users in *growing* Stack Exchange instances (cf. quarters eight through twelve of Figure 4o).

Finding: Long-term cross-excitation from questions on answers by casual users is a key factor present in *growing* Stack Exchange instances and lacking in *declining* ones. We find contributions by casual users thus attract more participation by casual users, likely helping to sustain and even enhance activity levels. Hence, we identify openness from the community towards casual users in the form of healthy interaction between them as a sign of enduring community growth.

Late Stage Self-Excitation (■). In the diagonal of Figure 4, consisting of Figures 4a, 4f, 4k and 4p, we observe strong and growing self-excitation effects, which dominate over cross-excitation effects in the long-term. We indicate long-term with the pink region marking quarters 8 to 12, the last five quarters we fit. We note this effect is most predominant in *growing* communities. Further, for *growing* instances, notice that, for a given dimension (e.g. answers by power users, Figures 4a through 4d), the timing of the surge in self-excitation coincides in general with a decline in cross-excitation.

Finding: In *growing* Stack Exchange instances, we attribute the phenomenon of higher long-term self-excitation to steadily growing arrivals of questions and answers from power and casual users. As users react to a constantly and regularly growing pool of questions and answers, this makes distinction of direct interaction between single questions and corresponding answers harder over time. The timing of this self-excitation surge may be of particular interest for Q&A community managers, who may be concerned about growth should they not observe this effect by the community's third year.

4.2 Comparison by topic

The comparison between *STEM* and *humanities* instances of Figure 5 shares a few commonalities with our previous findings on the *growing* and *declining* instance comparison: roughly constant and relatively low baseline intensities (not depicted due to limitations in space) and comparatively high long-term self-excitation. In this comparison, we obtained $\beta = 2.067$, which corresponds to an intensity half-life of 0.33 hours. These values are comparable to the ones we obtained previously, which may indicate a universal pattern of user activity decline across Stack Exchange instances.

Power User Cross-Excitation (■) vs. Casual User Self-Excitation (■). In the light blue highlighted region of Figure 5d, we stress the notable role of answer activity by power users in *humanities* instances. We observe answers by power users after questions from casual users is notably higher in *humanities* instances than in *STEM* instances (see Figure 5d). With the light orange region of Figure 5k, we underline the counterpart in casual user activity: There are higher long-term intensities in self-excitation of answers by casual users in *STEM* instances as compared to *humanities*.

Finding: In comparison with *STEM* Stack Exchange instances, *humanities* Stack Exchange instances are more reliant on cross-excitation by power users to address questions by both types of users. We observe more power user centric interactions in Stack Exchange instances in the *humanities*, while activity in *STEM* Stack Exchange appears more focused on casual users. Higher long-term self-excitation by casual users in *STEM* instances indicates stronger interactions between casual users. In turn, casual user activity is less dependent on power users in these instances. Overall, this finding suggests the existence of topic-dependent user type structures, which can be cast as measurable goals for community managers.

5 EVALUATION

In this section, we assess whether differences in excitation effects we observe in the evolution of *growing* vs. *declining* (*STEM* vs. *humanities*) instances result by chance or if there is some causal link between excitation, as measured with the Hawkes processes, and community growth (topical focus). Moreover, we evaluate the sizes of the observed effects by quantifying their impact on the future user activity.

Comparison of activity distributions. While the comparison of *growing* vs. *declining* instances aims to distinguish excitation effects in instances of increasing vs. decreasing and thus different total

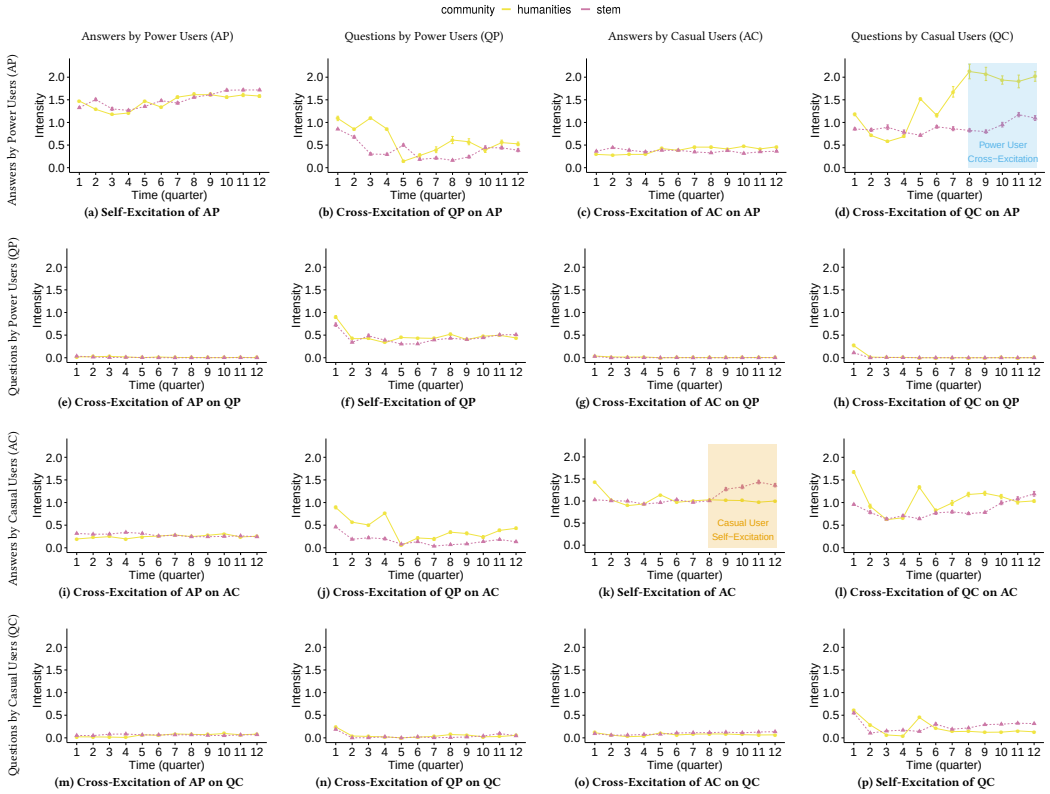
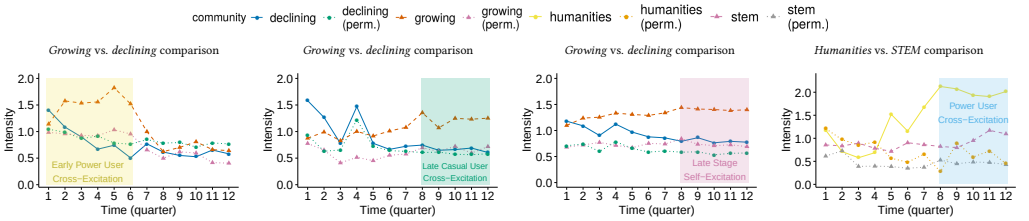


Figure 5: Excitation in *humanities* vs. *STEM*. Using the same notation and format as Figure 4, we depict the parameters of the Hawkes processes fitted every three months over three years of Stack Exchange instances dedicated to *STEM* (purple lines) and *humanities* (yellow lines) topics. We observe a more prominent role by power users in *humanities*, as indicated by by the more important role of power user cross-excitation originating from power users answering questions by casual users (cf. blue highlighted region ■ of Figure (d)). Furthermore, regarding casual user activity in *STEM* vs. *humanities* instances, we note the former's casual users feature more prominently in the long-term in the form of higher answer self-excitation (cf. orange highlighted region ■ of Figure (k)).

activity volumes, the *STEM* vs. *humanities* comparison is intended towards providing decoupled effects, which ideally should not be confounded with the excitation effects of *growing* vs. *declining* instances. However, in the *STEM* vs. *humanities* instance comparison, we highlight long-term self-excitation in answers by casual users, an effect which could be similar to the late stage self-excitation of *growing* vs. *declining* instances. Furthermore, if *humanities* instances simply featured overall higher answer-based activity levels by power users than in *STEM* instances, power users would also likely react stronger to questions by casual users, as opposed to them being an inherently more important backbone to questions by casual users.

Hence, we verify if the total answer-based activity distributions of both user types are similar in *STEM* and *humanities* instances. We compare the sample distributions of answers-based activity by power (and separately casual) users in *STEM* vs. *humanities* instances with the Kolmogorov-Smirnov two-sample test for their equality. As this test results in a p-value of 0.3855 (0.2305) for power (casual) users' activity distributions, we conclude there is not enough evidence to reject the null hypotheses of the probability distributions being equal at all usual significance levels. In turn, this test result indicates that the power and casual users' activity distributions are comparable, which supports our finding regarding the importance of the role power (casual) users play in *humanities* (*STEM*) Stack Exchange instances.

Permutation tests. To assess the significance of the excitation effects we conduct the following permutation test. First, we randomly permute the association of event types (questions by power users, questions by casual users, answers by power users and answers by casual users) to the corresponding time stamps per time window.



(a) Permuted Cross-Excitation of Questions by Power Users on Answers by Power Users (b) Permuted Cross-Excitation of Questions by Casual Users on Answers by Casual Users (c) Permuted Self-Excitation of Answers by Casual Users (d) Permuted Cross-Excitation of Questions by Casual Users on Answers by Power Users

Figure 6: Permuting event sources destroys observed excitation effects. We illustrate the temporal evolution of selected cross-excitation effects of multivariate Hawkes process models fitted to both the original (solid lines) as well as the permuted event streams (dashed lines) of the *growing* vs. *declining* and *STEM* vs. *humanities* comparisons of Stack Exchange instances. The permuted event streams in the colored regions of this Figure feature only few of the differences of the original event streams, and, for differences that remain (e.g. of Figure (c)), they are of perceptibly lower magnitude. Hence, the absence of effects in permuted event streams strengthen the significance of our main findings.

This procedure keeps the amount of events per event type constant, but destroys the temporal connection between event types. Then, we refit the multivariate Hawkes processes over windows of these permuted event streams, and we repeat these two steps 100 times. Finally, we compare the difference in mean Hawkes process parameter values fitted on the permuted event streams to the original ones. If there is a notable difference between them, then this indicates that growth (or the evolution of the topical instances) does not come about by chance, but that differences in self-excitation and cross-excitation between *growing* vs. *declining* (*STEM* vs. *humanities*) communities play an important role in their temporal evolution. We depict the result of these permutation tests in Figure 6, in which previously described differences between *growing* vs. *declining* and *STEM* vs. *humanities* Stack Exchange instances are all either remarkably weaker or non-existent. We arrive at similar results with the permutation tests on the other effect not included in Figure 6 (namely casual user self-excitation, in the *STEM* vs. *humanities* comparison). Inspired by Chandrasekharan et al.'s [6] quantification of differences in permutation test distributions, we numerically summarize our permutation tests with a comparison of the absolute difference of *growing* vs. *declining* (*STEM* vs. *humanities*) Hawkes process parameter values fitted on the original event streams with the distribution of absolute differences in parameter values obtained on permuted event streams. If the difference in original values is extreme in relation to the distribution of permuted values in the effects' time spans, then this is further evidence the effects we observe are unlikely to arise by chance. We quantify "extreme" with the p-value, in this case the proportion of values from the permuted difference distribution greater than the original difference. Over the quarters per effect time span, almost all p-values are smaller than or equal to 0.01[3]. Thus, we find the existence of a weak causal link between excitation effects in Stack Exchange instances and their temporal evolution in terms of activity volume.

Prediction experiment. To quantify the impact of the observed excitation effects on future activity, we design the following prediction experiment. For each three-month time window (quarter) and for each *growing* and *declining* Stack Exchange instance, we fit three variants of the Hawkes process, with the same four dimensions as previously: answers by power users, questions by power

[3]Single exception: Early Power User Cross-Excitation in quarter one (p-value 0.04).

Table 2: Kolmogorov-Smirnov (K-S) distance between predicted and real interevent times, per predicted quarter, effect type, Hawkes process dimension and model variant. Lower K-S values are better. Distance values marked with an asterisk correspond to not rejecting equality of simulated and real interevent times. The Full model produces forecasts with lowest K-S distances and highest number of non-significant distances. Thus, as the Excitation Effects Removed model features higher K-S distances than the Full model, we find all excitation effects are important for prediction. To quantify the importance of each excitation effect, we observe removing Late Stage Self-Excitation (■) in the Excitation Effects Removed model is most detrimental for prediction performance (cf. high values in mid-section of four rightmost columns). Hence, Late Stage Self-Excitation is most important for prediction, followed by Early Power User (■) and Late Casual User (■) Cross-Excitation.

		Early Power User Cross-Excitation					Late Casual User Cross-Excitation				Late Stage Self-Excitation			
	Prediction Quarter	**2**	**3**	**4**	**5**	**6**	**9**	**10**	**11**	**12**	**9**	**10**	**11**	**12**
Baseline	Answers by Power	0.26	0.26	0.29	0.27	0.24	0.24	0.25	0.23	0.24	0.24	0.25	0.23	0.24
	Questions by Power	0.3	0.28	0.25	0.25	0.24	0.23	0.24	0.23	0.24	0.23	0.24	0.23	0.24
	Answers by Casual	0.24	0.22	0.25	0.21	0.21	0.2	0.18	0.17	0.18	0.2	0.18	0.17	0.18
	Questions by Casual	0.23	0.16	0.18	0.13	0.12	0.12	0.14	0.13	0.12	0.12	0.14	0.13	0.12
Excitation Effects Removed	Answers by Power	0.15	0.13	0.14	0.12	0.13	0.1*	0.11*	0.11*	0.12	0.43	0.42	0.41	0.42
	Questions by Power	0.28	0.27	0.27	0.26	0.25	0.22	0.24	0.23	0.25	0.3	0.33	0.3	0.31
	Answers by Casual	0.22	0.12	0.16	0.15	0.16	0.15	0.14	0.13	0.14	0.42	0.4	0.39	0.4
	Questions by Casual	0.24	0.19	0.21	0.17	0.17	0.15	0.16	0.16	0.16	0.190	0.19	0.19	0.19
Full	Answers by Power	0.16	0.12	0.12	0.11*	0.11*	0.1*	0.1*	0.1*	0.11*	0.1*	0.1*	0.1*	0.11*
	Questions by Power	0.28	0.24	0.23	0.23	0.22	0.22	0.24	0.22	0.24	0.22	0.24	0.22	0.24
	Answers by Casual	0.23	0.11*	0.15	0.14	0.15	0.12	0.11*	0.11*	0.11*	0.12	0.11*	0.11*	0.11*
	Questions by Casual	0.25	0.16	0.18	0.13	0.13	0.12	0.13	0.13	0.12	0.12	0.13	0.13	0.12

users, answers by casual users and questions by casual users. The Hawkes process variants we consider are (i) a multivariate baseline model (i.e., a multivariate Poisson process), consisting of only baseline excitation $\boldsymbol{\mu}$ (Baseline in Table 2), (ii) a reduced model where we fit a full Hawkes process model but set the model parameters corresponding to the observed excitation effect to zero for quarters in which we observe a given excitation effect (i.e., we set cross-excitation of power users to zero for quarters one to five to remove the effects of Early Power User Cross-Excitation effect, then we set cross-excitation of casual users to zero for quarters eight to eleven to remove Late Casual User Cross-Excitation effect, and finally we set self-excitation of all users to zero also for quarters eight to eleven to remove the effects of Late Stage Self-Excitation effect) (Excitation Effects Removed), and (iii) a full Hawkes process model as defined in Equation 2 (Full), which we fit in the same manner as when uncovering excitation effects.

Overall, Hawkes process-based models such as ours are suited to forecast event timings, as classical machine learning approaches cannot make such time predictions (cf. e.g. Kurashima et al. [19]). Hence, for each variant of the Hawkes process, each Stack Exchange instance, and for each quarter that we fit, we predict the next quarter's event times by simulating the fitted process 100 times. To assess the model's performance we first compute the distribution of interevent times as well as event counts in all dimensions for each simulated and for given observed event sequences. Then, for each simulated quarter, we compute the mean of the Kolmogorov-Smirnov (K-S) test statistic to compare distributions of interevent times (and the root mean squared error (RMSE) to compare the event counts) between simulated and real events. We list the K-S test distance values for each predicted quarter in Table 2. We highlight, with an asterisk, values of the K-S test distance which correspond to not rejecting the hypothesis of equality of simulated and real interevent times at all usual significance levels.

We observe best K-S distance values overall for the Full model, indicating the overall importance of observed excitation effects at every developmental stage of a Stack Exchange instance for prediction experiments. Moreover, the very poor performance of the Excitation Effects Removed model, at times worse than the Baseline model, reinforces the importance of the effects we found for modeling and prediction. However, there seems to be a difference in the impact of different observed effects on the prediction performance. In particular, removing two cross-excitation effects (i.e., Early Power User Cross-Excitation and Late Casual User Cross-Excitation effect) from the models does not impair the performance of those models as strongly as the removal of Late Stage Self-Excitation effect. In Table 2's columns corresponding to the two Cross-Excitation effects, a comparison of predictions by the Excitation Effects Removed model with the corresponding predictions by the Full model reveals their differences in K-S distances lie in the interval $[-0.01, 0.04]$. In these cases, the Full model has only slightly better performance. On the other hand, the impact of the Late Stage Self-Excitation effect dramatically impairs the performance of the Excitation Effects Removed model. The differences between K-S distances in this case (cf. predictions by the Full model and the Excitation Effects Removed model in the Late Stage Self-Excitation columns of Table 2) range from 0.06 to 0.32, indicating a larger effect size of self-excitation than that of cross-excitation.

To further validate these results we perform another prediction experiment with a fourth variant of Hawkes processes. In this variant, we fit self-excitation only models by setting all cross-excitation parameters to zero. These additional experiments with a self-excitation model confirm previous observations: A model with only self-excitation achieves performances (as measured by the K-S distance and by the event count RMSE) in general on par with those of the Full model and, in the Late Stage Self-Excitation effect, even surpassing its performance slightly.

For all model variants we come to comparable conclusions when measuring the RMSE between simulated and real event counts. Limited by space, we summarize these results: The average RMSE of the Full model is 638.17 events, an improvement of 59.91% (43.09%) upon the Excitation Effects Removed (Baseline) model.

6 LIMITATIONS

Although we experimented with slightly different percentiles in the user type distinction and the instance characterization and obtained qualitatively similar results, we recognize those are arbitrary thresholds, which impact the results if changed significantly. Our results are more robust to changes in window size of the activity event stream of the Q&A communities (e.g. to two, four or five months), since this hyperparameter controls for the granularity of our results. Nevertheless, the Hawkes process model itself could include time-varying parameters, as an alternative to this repeated fitting procedure we apply over fixed time windows.

The effect time spans we propose, namely a one-and-a-half-year-long early stage and a late stage starting in the last quarter of the second year of a Q&A community, stem from our empirical observations of large differences in excitation in specific temporal segments in the community comparisons. Pinpointing exact transition dates is beyond the scope of this work, as we focus on learning temporal user excitation effects.

We acknowledge that mapping each user's questions and answers event streams to a Hawkes process dimension may be a more realistic model. However, we argue that such a model would suffer from sparsity, high dimensionality and higher computational cost. Further, such a model might also not improve the excitation effect characterization, as it would also struggle with distinguishing sources of self- and cross-excitation in high-activity regimes.

Note that we avoid a discussion of how casual users become power users with our characterization of power users as the most active each month, regardless of their histories. We believe engagement reward systems such as badges play an important role in casual user's development in particular and user excitation in general [20], but we leave a detailed investigation of the role of reward systems on excitation effects for future work.

We also caution that our work indicates a *temporal* link between (i) specific community structures in terms of user types and their excitation and (ii) the overall development of activity volume in a Stack Exchange community. This work does not establish causality.

7 RELATED WORK

Research on Q&A communities. There is a considerable amount of authors [1, 10, 13, 24, 33, 41] analyzing the roles different types of users play in Web communities such as Q&A websites. In addition,

several authors surveyed the motivation and behaviour of individual users [17, 26] of Q&A communities. While Mamykina et al. [24] and Furtado et al. [13] concentrate on uncovering and studying the roles of user types present in thriving Q&A communities, Danescu-Niculescu-Mizil et al. [10] and Yang et al. [41] explicitly focus on specific user types in Web communities and the user types' static and temporal characteristics. More broadly, Yang et al.'s work is part of a larger body of literature [28, 30, 46] on identifying experts in Q&A websites and characterizing their behavior. Our work leverages a comparable user type characterization to infer properties about the temporal evolution of communities themselves.

In approaches methodologically related to Zhang et al.'s [46], multiple authors [3, 25, 31, 38, 39, 43] study evolution dynamics of Web communities by relying on an explicit description of networks underlying a given Web community, and these networks often serve as a basis for dynamical systems models of the communities. In their study of Quora, another Q&A website, Wang et al. [39] analyze the role different social network structures play in Quora's community growth. Ribeiro [31] and Walk et al. [38] model users and activity in diverse Web communities including Q&A communities, with the former focusing on growth and decline of communities and the latter on the model's implications for self-sustainability in a community's activity. Matsubara et al. [25] and Zang et al. [43] study information diffusion and growth dynamics of Web communities.

Similarly to Matsubara et al.'s, Walk et al.'s and Zang et al.'s work, in this paper we also model growth and interaction dynamics of Q&A communities, but we do not assume an underlying network. We focus rather on excitation between groups of users, which we distinguish not on their expertise but on their overall activity levels. Furthermore, by encoding community lifecycles in Hawkes processes fitted to sequences of time windows, we extend the empirical discussion of Web community lifecycles [15, 42] and the critical mass literature [29, 32] to the Q&A community domain with measurable results.

Applications of Hawkes processes. Hawkes processes and their variations, as models for event streams with unequally spaced events in time, have found wide application in literature on different aspects of Web phenomena [12, 16, 36, 37, 44, 47, 48]. One such topic regards content popularity dynamics, in particular how to predict the influence of internal and external aspects of activity in social networks [12] and reshare popularity of items on the Web [47]. To infer causal links between users and user influence from user activity in social networks, Ver Steeg and Galstyan [37], Iwata et al. [16] and Zhou et al. [48] propose point process-related approaches, which cope with high dimensionality in number of users. Further, Upadhyay et al. [36] model the crowdlearning process of Stack Overflow users and characterize different user types by their expertise and learning curves. The work by Zang et al. [44] models and predicts the growth dynamics of individuals' ties in social networks and predicts its evolution.

Our work draws inspiration and methodological know-how from all above mentioned papers to expand on a topic closely related to Zang et al.'s: the development of not just the relatively small circle of an individual's social ties, but of excitation and interaction of user groups in Q&A communities. Furthermore, we contribute, to the growing body of work on fitting Hawkes process kernels [5, 22, 40, 49], a parsimonious Bayesian hyperparameter optimization method for fitting the decay parameter of exponential kernels in Hawkes processes. Finally, our extension of this fitting method to non-stationary multivariate event streams enables the extraction of temporal excitation effects from Q&A communities.

8 CONCLUSIONS

Summary. In this work, we modeled self- and cross-excitation in Q&A communities along several dimensions, including activity type, user engagement level, growth path and topical focus of a given community. We approached this task by fitting multivariate Hawkes processes to stationary temporal segments of Q&A communities' activity volumes. We found stronger cross-excitation of power (casual) users in early (late) stages of growing communities when compared to communities with declining total activity. Further, in growing communities, we observed self-excitation dominates in the long-term. Moreover, we uncovered strong long-term cross-excitation by power (casual) users in Q&A communities dedicated to topics in the fields of the humanities (STEM). We validated the presence of these excitation effects with statistical and permutation tests and we quantified their strength via prediction tasks.

Implications. Our work can support Q&A community managers in their ambition to promote sustainable community structures. To jumpstart community growth in its first six months, engaging a core of power users, for example in community building efforts, appears to be of crucial importance. In the medium- to long-term, we find community developers should carefully monitor and foster participation rather by casual users. While literature on critical mass in Web communities [29, 32] and studies on the user mix in Wikipedia [18, 34] also support this recommendation, we can afford further advice, as our casual cross-excitation analysis specifically underlines the importance of interaction between casual users. In practice, we believe adjusting reward or badge systems to encourage contributions by casual users, perhaps by welcoming newcomers or by easing their adjustment to community rules, would be of value to community development. Furthermore, community managers, which have not observed a surge in self-excitation by the third year of their communities, may have reason to concern over growth. Such excitation effects should be carefully monitored, as Q&A community growth may come at the cost of other community parameters [11]. Furthermore, our results indicate concrete implementations of these suggestions should depend on community topic, as it impacts excitation effects. Overall, our findings thus highlight the impact of timing in the user mix development.

Future work. Comparing other Q&A communities would allow to further generalize the results we obtained on Stack Exchange communities. Our work can be extended to uncover excitation effects in other domains, such as of Q&A instances in other languages or of other contribution types (e.g. open-ended vs. focused question), as our proposed approach is generic and can be readily extended.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback on the manuscript. Tiago Santos is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

REFERENCES

[1] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, 665–674.

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering User Behavior with Badges. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, 95–106.

[3] Roy M Anderson, Robert M May, and B Anderson. 1992. *Infectious Diseases of Humans: Dynamics and Control*. Vol. 28. Wiley Online Library.

[4] James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*. 115–123.

[5] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. DeepHawkes: Bridging the Gap between Prediction and Understanding of Information Cascades. In *Proceedings of the ACM on Conference on Information and Knowledge Management (CIKM'17)*. ACM, 1149–1158.

[6] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of the 2017 Conference on Computer Supported Cooperative Work (CSCW'17)*. ACM, 31–53.

[7] David Roxbee Cox and Valerie Isham. 1980. *Point processes*. Vol. 12. CRC Press.

[8] Daryl J Daley and David Vere-Jones. 2003. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer Science & Business Media.

[9] Daryl J Daley and David Vere-Jones. 2008. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media.

[10] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, 307–318.

[11] Himel Dev, Chase Geigle, Qingtao Hu, Jiahui Zheng, and Hari Sundaram. 2018. The Size Conundrum: Why Online Knowledge Markets Can Fail at Scale. In *Proceedings of the 2018 World Wide Web Conference (WWW'18)*. ACM, 65–75.

[12] Mehrdad Farajtabar, Nan Du, Manuel Gomez Rodriguez, Isabel Valera, Hongyuan Zha, and Le Song. 2014. Shaping Social Activity by Incentivizing Users. In *Advances in Neural Information Processing Systems (NIPS'14)*. 2474–2482.

[13] Adabriand Furtado, Nazareno Andrade, Nigini Oliveira, and Francisco Brasileiro. 2013. Contributor profiles, their dynamics, and their importance in five q&a sites. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW'13)*. ACM, 1237–1252.

[14] Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* (1971), 83–90.

[15] Alicia Iriberri and Gondy Leroy. 2009. A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)* 41, 2 (2009), 11.

[16] Tomoharu Iwata, Amar Shah, and Zoubin Ghahramani. 2013. Discovering Latent Influence in Online Social Activities via Shared Cascade Poisson Processes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. ACM, 266–274.

[17] Jiahua Jin, Yijun Li, Xiaojia Zhong, and Li Zhai. 2015. Why Users Contribute Knowledge to Online Communities: An Empirical Study of an Online Social Q&A Community. *Information & management* 52, 7 (2015), 840–849.

[18] Aniket Kittur, Ed Chi, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. 2007. Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. In *Proceedings of the Alt.CHI Conference on Human Factors in Computing Systems (Alt.CHI'07)*.

[19] Takeshi Kurashima, Tim Althoff, and Jure Leskovec. 2018. Modeling Interdependent and Periodic Real-World Action Sequences. In *Proceedings of the 2018 World Wide Web Conference (WWW'18)*. ACM, 803–812.

[20] Tomasz Kusmierczyk and Manuel Gomez-Rodriguez. 2018. On the Causal Effect of Badges. In *Proceedings of the 2018 World Wide Web Conference (WWW'18)*. ACM, 659–668.

[21] Kenneth Levenberg. 1944. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics* 2, 2 (1944), 164–168.

[22] Sha Li, Xiaofeng Gao, Weiming Bao, and Guihai Chen. 2017. FM-Hawkes: A Hawkes Process Based Approach for Modeling Online Activity Correlations. In *Proceedings of the ACM on Conference on Information and Knowledge Management (CIKM'17)*. ACM, 1119–1128.

[23] Thomas Josef Liniger. 2009. *Multivariate Hawkes Processes*. Ph.D. Dissertation. ETH Zurich.

[24] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design Lessons from the Fastest Q&A Site in the West. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, 2857–2866.

[25] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. 2012. Rise and Fall Patterns of Information Diffusion: Model and Implications. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. ACM, 6–14.

[26] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. What Do People Ask their Social Networks, and Why?: A Survey Study of Status Message Q&A Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, 1739–1748.

[27] Yosihiko Ogata. 1981. On Lewis' Simulation Method for Point Processes. *IEEE Transactions on Information Theory* 27, 1 (1981), 23–31.

[28] Aditya Pal, Shuo Chang, and Joseph A Konstan. 2012. Evolution of Experts in Question Answering Communities. In *Proceedings of the Sixth International Conference on Web and Social Media (ICWSM'12)*. 274–281.

[29] Daphne R Raban, Mihai Moldovan, and Quentin Jones. 2010. An Empirical Study of Critical Mass and Online Community Survival. In *Proceedings of the 2010 Conference on Computer Supported Cooperative Work (CSCW'10)*. ACM, 71–80.

[30] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. 2012. Finding expert users in community question answering. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, 791–798.

[31] Bruno Ribeiro. 2014. Modeling and Predicting the Growth and death of membership-based websites. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. ACM, 653–664.

[32] Jacob Solomon and Rick Wash. 2014. Critical Mass of What? Exploring Community Growth in WikiProjects. In *Proceedings of the Eight International Conference on Web and Social Media (ICWSM'14)*. 476–484.

[33] Ivan Srba and Maria Bielikova. 2016. A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Transactions on the Web (TWEB)* 10, 3 (2016), 18.

[34] Bongwon Suh, Gregorio Convertino, Ed H Chi, and Peter Pirolli. 2009. The Singularity Is Not Near: Slowing Growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM.

[35] Behzad Tabibian, Isabel Valera, Mehrdad Farajtabar, Le Song, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2017. Distilling Information Reliability and Source Trustworthiness from Digital Traces. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*. ACM, 847–855.

[36] Utkarsh Upadhyay, Isabel Valera, and Manuel Gomez-Rodriguez. 2017. Uncovering the Dynamics of Crowdlearning and the Value of Knowledge. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM'17)*. ACM, 61–70.

[37] Greg Ver Steeg and Aram Galstyan. 2012. Information Transfer in Social Media. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, 509–518.

[38] Simon Walk, Denis Helic, Florian Geigl, and Markus Strohmaier. 2016. Activity Dynamics in Collaboration Networks. *ACM Transactions on the Web (TWEB)* 10, 2 (2016), 11.

[39] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. 2013. Wisdom in the Social Crowd: An Analysis of Quora. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, 1341–1352.

[40] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. 2016. Learning Granger Causality for Hawkes Processes. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16)*. 1717–1726.

[41] Jie Yang, Ke Tao, Alessandro Bozzon, and Geert-Jan Houben. 2014. Sparrows and Owls: Characterisation of Expert Behaviour in StackOverflow. In *User Modeling, Adaptation, and Personalization*. Springer, 266–277.

[42] Colleen Young. 2013. Community Management that Works: How to Build and Sustain a Thriving Online Health Community. *Journal of Medical Internet Research* 15, 6 (2013).

[43] Chengxi Zang, Peng Cui, and Christos Faloutsos. 2016. Beyond Sigmoids: The Nettide Model for Social Network Growth, and its Applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, 2015–2024.

[44] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu. 2017. Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. ACM, 565–574.

[45] Achim Zeileis, Christian Kleiber, Walter Krämer, and Kurt Hornik. 2003. Testing and Dating of Structural Changes in Practice. *Computational Statistics & Data Analysis* 44 (2003), 109–123.

[46] Jun Zhang, Mark S Ackerman, and Lada Adamic. 2007. Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, 221–230.

[47] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. ACM, 1513–1522.

[48] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning Social Infectivity in Sparse Low-Rank Networks Using Multi-Dimensional Hawkes Processes. In *Artificial Intelligence and Statistics*. 641–649.

[49] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning triggering kernels for multi-dimensional hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*. 1301–1309.

3.5 Estimating the Decay Parameter of Hawkes Processes with Exponential Kernels

Expanding upon the previous article's answer to the second research question, this manuscript presents a study of the properties of the Hawkes process with exponential kernels as a function of its decay parameter. Specifically, given the fact that point estimates of decay parameter values by many different methods still yield significantly worse excitation estimates, I inspect reasons for this result and proposes an approach to mitigate this problem. This work pinpoints the noisy, non-convex log-likelihood of the Hawkes process as a function of the decay parameter as a motivation to not rely on point estimates of the decay but to rather quantify its uncertainty, and to surface and empirically validate any (perhaps implicit) hypotheses on the decay parameter value and on the Hawkes process itself. To do so, I develop a classical Bayesian framework to address these issues while estimating decay parameter values. Across a range of experiments with synthetic and real-world data, this approach provides uncertainty estimates, helps surface and diagnose hypotheses on the decay parameter value, and is also applicable even in the presence of non-stationary phenomena such as exogenous shocks to a system. The empirical insights and the framework may be useful for practitioners and researchers aiming to apply Hawkes processes to learn about not only the activity dynamics of peer production systems, but also other application contexts.

Estimating the Decay Parameter of Hawkes Processes with Exponential Kernels

Anonymous

***Abstract*—Hawkes processes with exponential kernels are a ubiquitous tool for modeling and predicting event times. However, estimating their decay parameter is challenging, and there is no consensus in previous work for this estimation task. In this work, we formulate, empirically characterize and mitigate the problem of estimating the decay parameter of Hawkes processes with an exponential kernel. In particular, we find that many common decay fitting approaches perform similarly well, but do not exactly recover true parameter values. These estimation difficulties relate to the noisy, non-convex shape of the Hawkes process' log-likelihood as a function of the decay. To learn more about likely decay values, we propose a Bayesian framework. We demonstrate that our Bayesian framework alleviates the decay estimation problem across a range of experiments with synthetic and real-world data. Our work supports researchers and practitioners in their applications of Hawkes processes.**

***Index Terms*—Hawkes process, decay rate, Bayesian inference**

I. Introduction

As a method for modeling and predicting temporal event sequences (henceforth *event streams*), Hawkes processes have seen broad application, ranging from modeling earthquakes [1], through measuring financial market movements [2] to estimating social dynamics in online communities [3]. Researchers and practitioners derive utility from Hawkes processes due to their flexibility in capturing history-dependent event streams. Hawkes processes model event streams via the conditional intensity function, the infinitesimal event rate given the event history. Events cause *jumps* in the conditional intensity function, which *decays* to a *baseline* level following a pre-defined functional form, the so-called *kernel*. Frequently, this kernel is chosen to be an exponential function. The reasons for this choice are manifold, as Hawkes processes with an exponential kernel are (i) efficient to simulate and estimate [4], [5], (ii) parsimonious, and (iii) realistic in many practical applications [6].

Hawkes processes are usually fitted via convex optimization. However, their likelihood is convex only in the baseline and jump parameters, leaving estimating the decay parameter of the exponential kernel an open problem. We observe this is of practical consequence, as there is a need to interpret exact Hawkes process parameter values [1], [6]–[8] and directions of temporal dependency [9]–[11], which may be incorrect for (remarkably) wrong values of the decay parameter. Previous work and frameworks simply assumed the decay parameters are given constants [5], [12], [13], cross-validated decay parameter values [3], [14], [15], or estimated them with a range of different optimization approaches [6], [16]–[21]. Qualitatively, such point estimates are sufficient for simulating and predicting event streams. However, there is a research gap in quantifying the *uncertainty* of decay parameter estimates, as well as diagnosing mis-estimation. Further, there are only initial studies [21], [22] on how exponential growth and exogenous shocks to a system under study may compromise key stationarity assumptions and aggravate estimation errors.

This Work. We empirically characterize, formulate and mitigate the problem of estimating the decay parameter of Hawkes processes with an exponential kernel. To that end, we first uncover that common decay fitting approaches all perform comparably well, but do not very accurately recover true decay values. We explain this difficulty in obtaining accurate estimations with the non-convex and noisy shape of the Hawkes process log-likelihood as a function of the decay parameter. Here, we identify an opportunity to address uncertainty in the estimation of the decay. In particular, we call for an approach to (i) quantify the consequences of that uncertainty, (ii) diagnose estimation errors, and (iii) address breaks of the crucial stationarity assumption. We propose to fulfill those three conditions by integrating the estimation of the decay parameter in a Bayesian framework. In this framework, we formulate and evaluate closed-form and intractable hypotheses on the value of the decay parameter. Specifically, we encode hypotheses for the decay as parameters of a prior distribution, and we consider estimations of the decay across Hawkes process realizations as samples from a likelihood. These likelihood samples form the data that we combine with the prior to perform Bayesian inference of posterior decay values.

We demonstrate that our Bayesian inference procedure for fitting the decay parameter fulfills the three previously mentioned conditions in synthetic and real-world settings. In particular, the diversity of the real-world domains that we study demonstrates the broad applicability of our approach. In our first application, a study of earthquakes in Japanese regions [1], we uncover low uncertainty in certain geographical relationships. Second, we validate Settles and Meeder's [23] supposition that vocabulary learning effort correlates to the estimated difficulty of the learned words [24]. Further, we diagnose difficulties in numerically capturing learning progress. Our final real-world study concerns a stationarity-breaking exogenous shock: Leveraging a dataset of Tweets before and after the Paris terror attacks of November 2015, we find evidence for the hypothesis advanced by Garcia and Rimé [25] that Tweet timings reflect collective effervescence.

Overall, our work sheds light on fitting a widely used class of Hawkes processes, i.e., Hawkes processes with exponential kernels. Better understanding these models and explicitly

surfacing uncertainty in their fitted values facilitates their use by practitioners and researchers. We expect the impact of our study to be broad[1], as our results influence the application of a key analysis approach for studying time-dependent phenomena across most diverse domains.

II. Background on Hawkes Processes

We briefly discuss temporal point processes, a set of mathematical models for discrete events randomly arriving over time. Temporal point processes capture the time of an upcoming event given the times of all previous events via the so-called *conditional intensity function* (or simply *intensity*) $\lambda^*(t)$. Mathematically,

$$\lambda^*(t)dt = \text{P(event in } [t, t+dt)|\mathcal{H}_t), \tag{1}$$

where $\mathcal{H}_t$ represents the event history up to (but excluding) time t. Dividing Equation 1 by dt (in the Leibniz notational sense), we see $\lambda^*(t)$ equals the conditional probability of an event in an interval of (infinitesimal) length dt per such interval dt. In other words, $\lambda^*(t)$ models the probability or, in frequentist terms, relative frequency of an event per time interval, and we interpret $\lambda^*(t)$ as a history-dependent *event rate*. Such temporal point processes are often termed doubly-stochastic, as events occur randomly over time, and the model for these events $\lambda^*(t)$ is a random process too. We also note $\lambda^*(t)$ characterizes temporal point processes as counting processes $N(t)$ for the number of events up to time t.

Hawkes processes [26] assume $\lambda^*(t)$ follows a certain functional form. Specifically, given a Hawkes process *realization*, i.e., a set of n events occurring at times $t_i \in \mathbb{R}^+$, the conditional intensity of a Hawkes process is

$$\lambda^*(t) = \mu + \alpha \sum_{t_i \in \mathcal{H}(t)} \kappa_\beta(t - t_i), \tag{2}$$

where $\mu \in \mathbb{R}^+$ is the *baseline* intensity and $\alpha \in \mathbb{R}^+$ the *self-excitation*, i.e., magnitude of an increase in $\lambda^*(t)$ at each event time t_i. Immediately after each t_i, the intensity decreases according to the kernel κ_β. A common choice [3], [5], [6], [10], [21], [27] for the kernel is an exponential function parametrized by the *decay* rate β, i.e., $\kappa_\beta(t) = e^{-\beta t}, \beta \in \mathbb{R}^+$. Plugging this kernel in Equation 2 we obtain the Hawkes process intensity

$$\lambda^*(t) = \mu + \alpha \sum_{t_i \in \mathcal{H}(t)} e^{-\beta(t-t_i)}, \tag{3}$$

which we illustrate in Figure 1.

Multivariate Hawkes processes with an exponential kernel generalize univariate ones by introducing parameters for self-excitation and for the decay per dimension. Beyond self-excitation, they also capture *cross-excitation*, the intensity jump an event in one dimension causes in another. Formally, the intensity of dimension p of an M-variate Hawkes process is

$$\lambda^{*p}(t) \quad = \quad \mu_p + \sum_{q=1}^{M} \sum_{t_i^q < t} \alpha_{pq} e^{-\beta_{pq}(t - t_i^q)}. \tag{4}$$

[1]We make our code available at *URL anonymized*.

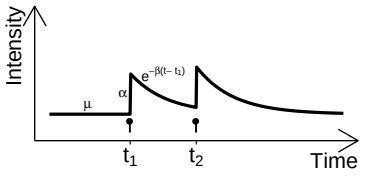


Fig. 1: **Illustration of Hawkes Process Intensity λ^*.** This two event sample from a Hawkes process exemplifies how its intensity λ^* changes over time. Starting with a minimal level of intensity μ, λ^* jumps by a constant α at each event time t_i, and then decays exponentially at rate β over time.

Notice that we index each dimension's intensity function with λ^{*p}, and its baseline with μ_p. This generalization also includes an excitation matrix with self-excitation and cross-excitation entries α_{pp} and respectively α_{pq}, as well as, analogously, a matrix of decay rates β_{pq}. Note that α_{pq} captures the increase in intensity in dimension p following an event in dimension q. In matrix notation, we write $\boldsymbol{\mu} \in \mathbb{R}^M$, $\boldsymbol{\alpha} \in \mathbb{R}^{M \times M}$ and $\boldsymbol{\beta} \in \mathbb{R}^{M \times M}$. We point to the work by Linniger [4] for more on multivariate Hawkes process theory.

We now introduce the notions of *stationarity* and *causality* in the multi-dimensional Hawkes process context. Stationarity implies translation-invariance in the Hawkes process distribution, and, in particular, it also implies that the intensity does not increase indefinitely over time and therefore stays within bounds. More formally, a Hawkes process with an exponential kernel is stationary if the spectral radius ρ, i.e., the largest eigenvalue of the L^1-norm of $\boldsymbol{\alpha}/\boldsymbol{\beta}$, satisfies $\rho < 1$. Note that assessing stationarity of a one-dimensional Hawkes process with an exponential kernel reduces to evaluating $\alpha/\beta < 1$.

Recent work [28] established a connection between Granger causality and the excitation matrix: In the particular case of our exponential kernel, dimension q Granger-causes dimension p if and only if $\alpha_{pq} > 0$. Beyond this result, we interpret the magnitude of excitations α_{pq} as the strength and direction of temporal dependency between dimensions p and q: For example, we say dimension q influences dimension p more strongly if $\alpha_{pq} > \alpha_{qp}$.

Finally, we define the Hawkes process likelihood function. We work with the log-likelihood function due to its mathematical manipulability and to avoid computational underflows. The equation for the one-dimensional log-likelihood of the Hawkes process with an exponential kernel is as follows [16]:

$$log\, L(\{t_i\}_{i=1}^n) = -\mu t_n - \frac{\alpha}{\beta} \sum_{i=1}^{n} (1 - e^{-\beta(t-t_i)}) + \sum_{i=1}^{n} log(\mu + \alpha A(i)), \tag{5}$$

with A(1) = 0 and, for $i > 1$, $A(i) = \sum_{t_j < t_i} e^{-\beta(t_i - t_j)}$.

Ozaki [16] also proposes a (computationally less intensive) recursive formulation for Eq. 5. We refer to Daley and Vere-Jones [29] for a general formulation of the log-likelihood of multivariate temporal point processes.

Henceforth, as we focus on Hawkes processes with exponential kernels, we refer to them simply as Hawkes processes.

III. Decay Estimation Problem

To learn from streams of events, practical applications start by fitting Hawkes processes, i.e., optimizing the log-likelihood given in Equation 5 to a set of event timestamps. Practitioners then inspect fitted parameters to understand inherent temporal dependencies, and perform downstream tasks such as prediction via simulation of the fitted processes [19], [21].

As inferring and interpreting (all) fitted Hawkes process parameter values is crucial in many real-world applications [1], [6]–[9], [11], [21], we turn our attention to the challenges in fitting Hawkes process parameters, and especially, in estimating the decay parameter. In the following, we first observe that previous work proposes a wide range of approaches to fit the decay parameter. We then empirically study the performance of commonly used decay parameter fitting methods, and we highlight their limitations. Inspecting the reasons for such limitations inspires our formulation of the decay fitting problem.

A. Theoretical & Empirical Observations

To fit a Hawkes process, we typically maximize its log-likelihood (cf. Eq. 5). Previous research has shown that the baseline μ and excitation jump α can be efficiently computed since the log-likelihood is amenable for convex optimization of these parameters (cf. Bacry et al. [7] as well as a proof of a more general case in Farajtabar et al.'s [5] Theorem 3). However, that is not the case with the decay β, in neither the univariate nor the multivariate case.

Therefore, previous work suggested a wide range of methods to address the decay estimation problem with approaches which provide point estimates. These approaches include setting β to a given constant value [5], [12], [13], cross-validation over a range of values [3], [14], [15], or one of different optimization methods. Those methods comprise non-linear optimization [16], [17], Bayesian hyperparameter optimization [20], [21], expectation-maximization [19], [30] or visual inspection of the log-likelihood function [6], [18]. This variety of decay fitting approaches motivates the following empirical comparison of their performance with respect to recovering a given decay parameter β.

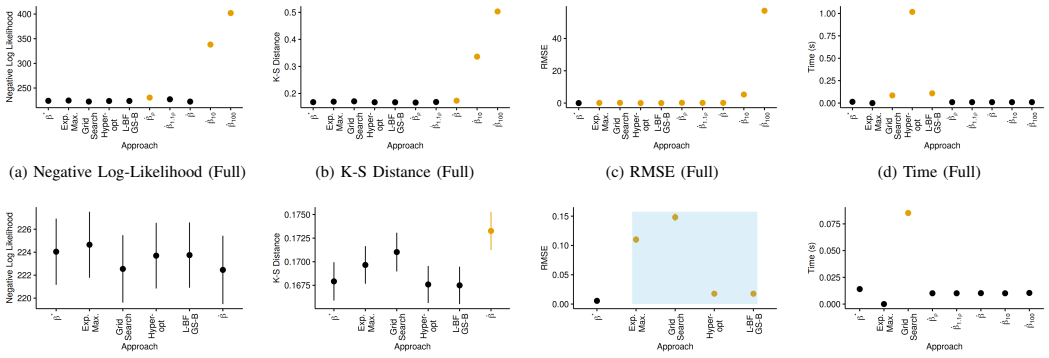
Common β Fitting Approaches Perform Comparably Well, But Do Not Exactly Recover True Parameter Values. Given the previously mentioned heterogeneity in β fitting approaches, we first compare their performance on synthetic data across a range of configurations and metrics. For that, we generate 100 realizations from a univariate Hawkes process with random parameters. To generate these random parameters, we start by randomly sampling $\mu \sim U(0.02, 0.05), \alpha \sim U(0.1, 0.3)$ and $\beta \sim U(0.7, 1.5)$, a set of values which clearly satisfies the stationarity constraint. As we also aim to fit a reasonable number of events, which we define as $10 \leq n \leq 1000$, we then increase the α and β parameter values closer to the stationarity constraint. Specifically, we keep increasing α and respectively β by a randomly chosen percentage between 1% and 20% until ρ is neither too small nor too large ($0.7 \leq \rho \leq 0.9$). While these bounds are conservative, we find that this setup produces enough events and alleviates unnecessarily long simulations of processes with ρ (very) close to 1. We simulate each such Hawkes process for a total of 1000 time units. This simulation horizon satisfies the constraint we place on the number of events n often, but, if it does not, we perform another random initialization of μ, α and β and repeat the previously described process of increasing α and β values. Changes to the aforementioned thresholds do not qualitatively alter our results.

For each such set of simulated Hawkes process realizations, we perform alternating maximization of the log-likelihood[2] with respect to μ, α and then β. To optimize β, we employ the methods *L-BFGS-B* [31] as a non-linear optimization routine, *Bayesian hyperparameter optimization (Hyperopt)* [32], *Expectation-Maximization (Exp. Max.)* [30], *Cross-Validation (Grid Search)* across 10 evenly distributed values on a log scale in $[-1, 2]$ (similarly to Salehi et al. [15]), as well as a range of constant values. These constant values include β^*, i.e., for reference, the true β value used in simulation, $\dot{\beta}$, a constant value which, like previous work [5], [15], we set to 1 as a (probably) "wrong" value in the same order of magnitude of β^*, as well as other wrong β values: $\dot{\beta}_{10}$ and $\dot{\beta}_{100}$, 1 and respectively 2 orders of magnitude greater than the true value, and finally β values "close" to the stationarity constraint, i.e., $\dot{\beta}_{1.1\rho}$ and $\dot{\beta}_{\rho}$, which are $1.1\,\alpha^*$ and respectively α^*, i.e., the true α used in simulation.

We measure each method's performance with the following metrics: negative log-likelihood (*Negative Log-Likelihood*; cf. Eq. 5), root mean squared error (*RMSE*) of all fitted Hawkes process parameters, mean Kolmogorov-Smirnov distance between 100 realizations of the fitted process and 100 realizations from the true process (*K-S Distance*) and mean run-time of the fitting procedure using commodity hardware (*Time (s)*). We repeat this whole process 1000 times to derive bootstrapped 95% confidence intervals for each metric's mean per method.

We depict the results of this comparison in Figure 2. We observe all approaches perform comparably well along most metrics, except for the $\dot{\beta}$ variations (and especially $\dot{\beta}_{10}$ and $\dot{\beta}_{100}$). In particular, most optimization approaches attain log-likelihood (cf. Figs. 2a and 2e) and K-S distance values (cf. Figs. 2b and 2f) statistically indistinguishable from those of β^*. Note that log-likelihood is not always minimal at β^* due to noise, as we illustrate later in Figure 3. Figures 2d and 2h may indicate a trade-off between RMSE and runtime, as all fitting approaches are slower than simply setting a constant.

[2]For computational reasons, we utilize the μ, α optimization routine and the log-likelihood implemented in the tick library [12] (version 0.6). The log-likelihood shown in Eq. 5 differs from that implemented in tick in essence by constant factors. Both log-likelihoods have the same qualitative properties.



(a) Negative Log-Likelihood (Full) (b) K-S Distance (Full) (c) RMSE (Full) (d) Time (Full)

(e) Negative Log-Likelihood (Excerpt) (f) K-S Distance (Excerpt) (g) RMSE (Excerpt) (h) Time (Excerpt)

Fig. 2: **Most Approaches for Fitting the Decay Perform Comparably Well, Except in the RMSE Metric.** We simulate one-dimensional Hawkes processes with random parameter values, and we compare several beta estimation approaches with constant baselines (the true value $\beta^* \in [0.7, 1.5]$, and values far and close to the stationarity constraint, $\dot{\beta}, \dot{\beta}_{10}$ and $\dot{\beta}_{100}$ and respectively $\dot{\beta}_{1.1\rho}$ and $\dot{\beta}_{\rho}$). The upper row of Figures depicts all approaches and the lower row selected ones for visualization purposes. We measure performance along a series of metrics, where lower values are better. Error bars throughout this paper indicate 95% confidence intervals (often too small to be visible). The orange color highlights metric values which are significantly worse in comparison to those obtained with β^*. Most approaches perform comparably well across metrics, with the notable exception of RMSE (cf. blue region in Fig. 2g). Note that log-likelihood is not always minimal at β^* due to noise (cf. Figure 3).

More importantly, however, we stress that there are remarkable differences in RMSE values (as highlighted in blue in Fig. 2g): L-BFGS-B and Hyperopt attain the lowest values, and both still perform significantly worse than the β^* optimum[3] (orange points in Fig. 2 indicate significant differences at the Bonferroni-corrected bootstrapped $p < 0.01$). In Figure 2c, we report even higher RMSE values with all other approaches.

Generalizing to multivariate Hawkes processes, we qualitatively confirm the previous observations. The main difference to the univariate case lies in the magnitude of estimation errors, as e.g. RMSE values of all methods become remarkably larger.

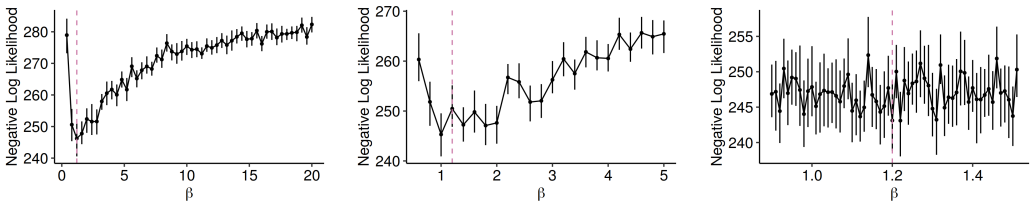
A Noisy, Non-Convex Log-Likelihood in β Explains Optimization Difficulties. To understand why most approaches appear to optimize log likelihood but still underperform in terms of parameter RMSE, we turn our attention to the shape of the log-likelihood as a function of β.

In the following illustration, we consider a univariate Hawkes process with $\mu = 0.1, \alpha = 0.5$ and $\beta^* = 1.2$, and we then compute the negative log-likelihood for different values of β. We generate three sets of 100 realizations from that Hawkes process. In Fig. 3, we evaluate the negative log-likelihood with one set of realizations per each of three ranges of β around β^*, namely a large (cf. Fig. 3a), a medium (cf. Fig. 3b) and a small (cf. Fig. 3c) range. In the large range, it appears there is a convex basin around β^* (which we annotate with a pink dashed line), but this function's shape shifts to a concave curve with increasing decay values. The function then converges on the right, as $\lim_{\beta \to +\infty} log\, L = -\mu t_n + n log(\mu)$. Inspecting the "convex" region more closely uncovers a wide and noisy basin around β^*, where β^* does not always feature minimal negative log-likelihood (cf. Fig. 3b). This explains difficulties in obtaining correct estimations (regardless of the optimization strategy). We note that these observations are robust to choosing other parameter values corresponding to stationary processes, and they generalize to multivariate Hawkes processes as well.

[3]Note that the RMSE of the β^* method does not equal 0 due to RMSE in the estimation of μ and α.

B. Problem Statement

As previously mentioned, the inherent properties of the Hawkes process' log-likelihood as a function of β hinder the estimation of (very) accurate values for β, and, consequently, for the other Hawkes process parameters as well. Also, recall that there are numerous applications which rely on interpretation of Hawkes process parameter values [1], [6]–[11]. Therefore, we believe there is a currently unmet need to surface β estimation uncertainty and the magnitude of potential estimation errors, across β fitting approaches (i.e., point estimates). Given that previous work [1] also calls for encoding and validating hypotheses on the decay value, we anticipate that diagnosing not only estimation errors but also misaligned hypotheses is of practical interest too. Further, current methods return only decay values which fulfill the stationarity constraint $\rho < 1$. However, previous work [21], [22] studied applications with non-stationary changes such as exponential growth and exogenous shocks. Hence, we see an



(a) Large Range (b) Medium Range (c) Small Range

Fig. 3: **The Negative Log-Likelihood of Hawkes Processes as a Function of β Is Non-Convex And Noisy.** We simulate three separate sets of 100 realizations from a one-dimensional Hawkes process with $\beta^* = 1.2$ (marked with a pink dashed line), and we evaluate the process' log-likelihood with one set of realizations per each of three ranges of β around β^*: a large (cf. Fig. 3a), a medium (cf. Fig. 3b) and a small (cf. Fig. 3c) range. In Fig. 3a, it is apparent the negative log-likelihood is neither concave nor convex in β. Zooming in around β^* reveals a wide and noisy basin and, in Fig. 3b, the negative log-likelihood is not minimal at β^*. Both observations explain difficulties in optimizing the log-likelihood for the decay parameter.

opportunity for an extension of current decay fitting methods to address uncertainty surrounding the stationarity assumption.

More formally, given a set of event timestamps $\{t_i\}_{i=1}^n$ which we model as a stationary Hawkes process, we call for a principled approach to extend current β estimation approaches while fulfilling the following conditions:

1) Quantify the uncertainty of estimates and potential consequences of estimation uncertainty,
2) Diagnose mis-estimation and misaligned hypotheses,
3) Address potential estimation errors due to unmet assumptions, such as, notably, breaks in stationarity.

IV. Decay Estimation Approach

We propose a parsimonious Bayesian inference procedure for encoding and validating hypotheses on likely values for β. In our Bayesian framework, we sequentially collect a series of univariate Hawkes process realizations, one by one. With each realization that arrives, we fit β with a given optimization method and obtain an ever-increasing set of estimated decay values that we denote as $\{\hat{\beta}\}_{k=1}^K$[4]. After this collection of $\hat{\beta}$, we apply Bayes' theorem to make inferences about the true β.

Before formalizing our Bayesian inference approach, we intuitively explain one key difference between our approach and typical applications of Bayesian inference. The classical Bayesian inference setup typically places a prior distribution for an unknown parameter of interest in a probability distribution which captures the likelihood of given data. Then, applying Bayes' theorem allows for inferring likely values of the unknown parameter given the data. In our Bayesian framework, we assume the unknown parameter of interest, i.e., the decay parameter, is also the data, which consists of the aforementioned sequence of decay parameter estimations. This setup enables the freedom to choose between computing posterior (i.e., in parameter space) or posterior predictive (i.e., in data space) distributions to learn about the decay. We find that this flexibility is useful and can improve performance in applications.

More formally, given data $\{\hat{\beta}\}_{k=1}^K$ and a model $\mathcal{M}_H$ parametrized by hypothesis H for the parameter of interest β, we propose computing the Bayesian posterior

$$P(\beta|\{\hat{\beta}\}_{k=1}^K, \mathcal{M}_H) \propto P(\{\hat{\beta}\}_{k=1}^K|\beta, \mathcal{M}_H)P(\beta|\mathcal{M}_H), \quad (6)$$

where $P(\{\hat{\beta}\}_{k=1}^K|\beta, \mathcal{M}_H)$ is the likelihood and $P(\beta|\mathcal{M}_H)$ the prior encoding a hypothesis. Again, note the accordance between the model parameter and the data in our Bayesian framework: β is the model parameter and, at the same time, our data $\{\hat{\beta}\}_{k=1}^K$ contains the estimates of that same parameter. Having derived or estimated a posterior distribution density, we have multiple inference possibilities for β. In particular , we can –as introduced above– (i) obtain a point estimate for β directly from the posterior (e.g., mean, median, or maximum), or (ii) compute a new estimate $\tilde{\beta}$ or a statistic (e.g., mean) from the posterior predictive distribution, which is given by (omitting for simplicity $\mathcal{M}_H$):

$$P(\tilde{\beta}|\{\hat{\beta}\}_{k=1}^K) = \int P(\tilde{\beta}|\beta, \{\hat{\beta}\}_{k=1}^K)P(\beta|\{\hat{\beta}\}_{k=1}^K)d\beta. \quad (7)$$

Thus, this setup leaves us with the task of deciding on appropriate distributions for the prior and the likelihood, as well as mapping hypotheses to specific parametrizations of the prior. While practitioners are free to exhaust the broad spectrum of available techniques for Bayesian inference, we propose simply assuming the likelihood $P(\{\hat{\beta}\}_{k=1}^K|\beta) \sim Exponential(\lambda)$ and a conjugate prior, namely $P(\beta) \sim Gamma(a_0, b_0)$. These assumptions lead to the following practical advantages: (i) there is a closed form solution to the posterior predictive density mean, $\beta' = E[\tilde{\beta}|\{\hat{\beta}\}_{k=1}^K] = (b_0 + \sum_{k=1}^K \hat{\beta}_k)/(a_0 - 1)$, and (ii) users of this framework only have to encode hypotheses on

[4]Note that the set $\{\hat{\beta}\}$ does not contain independent observations (as e.g. the realizations used to obtain $\{\hat{\beta}\}_{k=1}^2$ are also in $\{\hat{\beta}\}_{k=1}^3$). Anticipating that practitioners use as much data as available, we consider $\{\hat{\beta}\}$ as previously defined. However, repeating all our experiments using only iid $\hat{\beta}_k$ (i.e., each fitted only on a single realization), we obtain similar though noisier results.

β as the prior's b_0 parameter, as they can set a_0 simply to K. We note that assuming other distributions of the exponential family such as Pareto for the log-likelihood and the conjugate prior is also a valid choice, for which we experimentally obtain equivalent results. Note another advantage to our choice of Bayesian over frequentist inference: We show, in our experiments, that complex inference setups, which require the application of e.g. Markov chain Monte Carlo (MCMC), are of practical interest, in particular to address breaks in stationarity. Finally, in our approach, we restrict Bayesian inference to the decay and estimate confidence intervals for other parameters via (frequentist) bootstrap. This enables the estimation of other parameters via more efficient convex optimization routines.

We focused the exposition of our approach on univariate Hawkes processes. To generalize to the multivariate case, we set $\beta_{pq} = \beta \; \forall_{p,q}$, a common simplification of the decay estimation problem [3], [8], [15], [21], and we then proceed as previously.

V. Experiments

We proposed to address the decay fitting problem via a Bayesian estimation approach, where we aim to learn more about plausible values for β through encoding and evaluating hypotheses. We illustrate that our Bayesian approach fulfills the three conditions we outlined in Section III-B, as it allows for estimating the uncertainty in the decay and other Hawkes process parameter values, diagnose mis-estimation and address breaks in stationarity. For each of those three conditions, we (i) illustrate the condition with a synthetic dataset and (ii) present a real-world application in which the condition arises. Hence, we demonstrate that, besides fulfilling the required conditions, our approach is broadly applicable across practical scenarios.

A. Quantifying the Uncertainty of Decay Estimates

We begin by addressing the problem of estimating the uncertainty of fitted decay values $\hat{\beta}$, as well as potential consequences of mis-estimation. In particular, note that one prominent application of multi-dimensional Hawkes processes consists in the estimation of directions of temporal dependency, e.g., when studying influence in online communities [9], [21], or in approximating complex geographical [1] or cortical [10] relationships. Recall that inferring such relations between a pair of Hawkes process dimensions may be framed as a problem of estimating which cross-excitation between the two dimensions is higher. Using synthetic and real-world data, we demonstrate how our Bayesian procedure helps in estimating the uncertainty in such inferred relationships, and how to quantify the impact of potential errors. In this setting, we remark that uncertainty may also be estimated via bootstrap. Thus, we view our approach as a Bayesian alternative to such frequentist techniques.

1) Synthetic Data: We consider a two dimensional Hawkes process with parameters $\boldsymbol{\mu} = \left(\begin{smallmatrix} 0.1 \\ 0.5 \end{smallmatrix}\right)$, $\boldsymbol{\alpha} = \left(\begin{smallmatrix} 0.1 & \alpha_{12} \\ \alpha_{21} & 0.2 \end{smallmatrix}\right)$, and $\beta_{pq} = \beta \; \forall_{1 \leq p,q \leq 2}$. We assume $\beta = 1.2$. For the cross-excitation parameters, we set $\alpha_{21} = 0.7$ and, successively, $\alpha_{12} = \alpha_{21} * c$ for a range of 10 linearly spaced values of $c \in [0.75, 1.25]$. This implies that each configuration encodes a different direction and strength of influence, where dimension 2 dominates dimension 1 for $c < 1$, and vice-versa for $c > 1$. For each such configuration, we simulate $K = 100$ realizations with a stopping time of $T = 1000$. We apply the decay estimation approach L-BFGS-B after the arrival of each such realization to obtain $\{\hat{\beta}\}_{k=1}^{100}$ per configuration. Using our closed-form Bayesian inference framework, we hypothesize that β equals 1.5 by setting $b_0 = 1.5$ in the previously described Gamma prior. We then perform the aforementioned Bayesian inference on each set of $\{\hat{\beta}\}_{k=1}^{100}$ to derive $\beta'_{0.025}$ and $\beta'_{0.975}$, the lower and upper bounds of the 95% credible interval of the posterior predictive density. For 100 linearly spaced values of the decay in $[\beta'_{0.025}, \beta'_{0.975}]$, we then fit the remaining Hawkes process parameters and check the accuracy of the inferred direction of influence between dimensions. This accuracy measure captures how many of the 100 decay values lead to correct recovery of the encoded relation between α_{12} and α_{21}. We bootstrap 95% confidence intervals for this accuracy measure directly from that distribution of 100 decay values. This procedure provides (i) an estimate of the uncertainty surrounding the fitted decay value via the 95% credible interval, (ii) an estimation of the robustness of the temporal dependency between dimension 1 and 2, and (iii) empirical evidence for the consequences of fitting Hakwes processes with misaligned β. We note that similar parametrizations of the Hawkes process and using alternative decay fitting approaches like Hyperopt do not qualitatively alter our results.

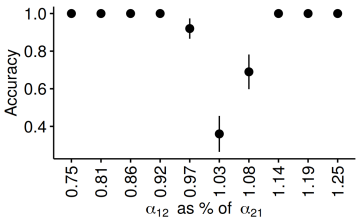


Fig. 4: **Estimating Uncertainty When Inferring Directions of Influence.** We first fit β on realizations from two-dimensional Hawkes processes with cross-excitation α_{12} varying from 75% to 125% of α_{21}. We apply closed-form Bayesian inference to estimate the uncertainty in fitted decay values as the posterior predictive 95% credible interval. For a set of decay values in that interval, we estimate the other parameters, and we measure the accuracy in recovering the encoded direction of influence between dimension 1 and 2. The accuracy is low and features larger error bars when α_{12} is close to α_{21}, where many decay values in the 95% credible intervals lead to wrong estimations of the direction of influence.

Results. Figure 4 summarizes the outcome of this experiment. As expected, we observe lower accuracies in inferring the direction of influence between dimension 1 and dimension 2 for α_{12} close to α_{21}. This implies that many decay values in the 95% credible interval of such configurations lead

to mis-estimations of the direction of influence, and the large error bars reflect this as well. Overall, we believe practitioners may leverage this approach to check the robustness of inferred directions of influence with respect to hypothesized (or estimated) decay values.

2) Earthquakes and Aftershocks: We illustrate the outlined uncertainty estimation procedure with a dataset of earthquakes in the Japanese regions of Hida and Kwanto, as originally studied by Ogata et. al [1]. We consider the data listed in Table 6 of that manuscript, i.e., a dataset of 77 earthquakes from 1924 to 1974. We employ the decay value listed in Table 5 of that manuscript as the prior's parameter in our closed-form Bayesian framework[5]. We assume a two-dimensional Hawkes process with a single β value, where the dimensions represent earthquakes in the Japanese regions of Hida and Kwanto. As pre-processing, we split earthquake occurrences into $K = 4$ equally sized segments which we treat as process realizations, and we convert the event timescale to decades.

Results. We replicate the seismological relationship that earthquakes in the Japanese Hida region precede those in Kwanto, as we obtain $\alpha_{\text{Kwanto Hida}} > \alpha_{\text{Hida Kwanto}} = 0$. Our Bayesian estimation procedure yields the posterior predictive density mean of $\beta' \approx 31.71$, which corresponds to an intensity half-life of about $log(2)/31.71 \approx 0.02$ decades. The inferred relationship between Hida and Kwanto is present for all but one value of the 95% credible interval for β', the lower extremity $\beta'_{0.025}$. This result underscores the low uncertainty and high robustness of the inferred direction of influence.

B. Diagnosing Mis-Estimation & Misaligned Hypotheses

Having derived estimates for the uncertainty in estimated decay values, we now focus on diagnosing errors in point estimates. Specifically, we demonstrate that our Bayesian framework facilitates diagnosing (inevitable) estimation errors and misaligned hypotheses as over- or under-estimates, as well as the magnitude of that error. Hence, we address a need, which previous work [1], [6], [18] implies, to encode, validate and diagnose estimations and hypotheses on the decay parameter value. Again, we illustrate how our approach meets that need with synthetic and real-world data.

1) Synthetic Data: We consider a univariate Hawkes process with parameters $\mu = 1.2, \alpha = 0.6$ and $\beta = 0.8$. Comparable choices of parameters lead to the same qualitative results. Using this parametrization, we successively generate $K = 100$ realizations with 100 events each. We apply the decay estimation approaches L-BFGS-B, Grid Search and Hyperopt after the arrival of each such realization and obtain $\{\hat{\beta}\}_{k=1}^{100}$ per approach. Again using our closed-form Bayesian inference framework, we leverage a Gamma prior with $b_0 = 1$. We then perform the aforementioned Bayesian inference on each set of $\{\hat{\beta}\}_{k=1}^{100}$ and compare the RMSE between the resulting β' estimates and β. We estimate uncertainty per fitting method via 95% credible intervals and via bootstrapped 95% confidence intervals, which in this case are essentially the same.

[5]We choose that prior for demonstration purposes, as that decay value was estimated with a different Hawkes process from the one we study in this work.

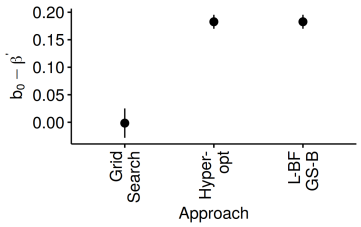


Fig. 5: **Closed-Form Bayesian Inference Helps Diagnose Estimation Errors, Such As Those Caused by Mis-Specified Hypotheses.** We simulate from a univariate Hawkes process, fit β with the L-BFGS-B, Grid Search and Hyperopt approaches, and apply closed-form Bayesian inference to the fitted values of each approach. We depict the hypothesized prior parameter value b_0 minus the posterior predictive density mean β', and we observe mostly positive differences. This discrepancy stems from a b_0 prior parameter value which is larger than true β, thus validating our Bayesian framework as a diagnosis tool.

Results. Overall, this experiment's estimation results agree with those of Figure 2g, as L-BFGS-B and Hyperopt return the best estimates. However, the RMSE values are above what we expect from that Figure, despite the very similar experimental setup. Looking into $\hat{\beta}_k$ reveals that they are consistently below our hypothesized b_0. We suspect this discrepancy arises due to our prior parameter value $b_0 = 1$, which is an over-estimate. Therefore, we look into the difference between that b_0 and the posterior predictive density mean β' per fitting approach, and we depict this comparison in Figure 5. Except for Grid Search (which also has high RMSE), the approaches boast positive differences, which imply the prior parameter is larger than the posterior predictive density mean. These inspections of the direction and magnitude of posterior shifts away from the prior suggest the use of our Bayesian framework as a diagnosis tool, which in this case correctly signals our hypothesis likely over-estimates true β.

2) Vocabulary Learning Intensity: We address future work proposed in Settles and Meeder's [23] study of user behavior on the Duolingo language learning app: The authors speculate that vocabulary learning intensity in Duolingo correlates with word difficulty as defined by the CEFR language learning framework [24]. We complement the Duolingo data with a dataset of English-language vocabulary and its corresponding CEFR level[6], and we build two groups of words: those from the easiest CEFR levels, A1 and A2 (A-level group), and those from the hardest ones, C1 and C2 (C-level group). We observe that there are 28 users with 10 vocabulary learning events in the C-level. To control for total learning events per user, we randomly sample a set of 28 users with 10 learning events in the A-level. Increasing the number of events to 11 or 12 leads to qualitatively similar results, but decreased statistical

[6]http://www.englishprofile.org/american-english

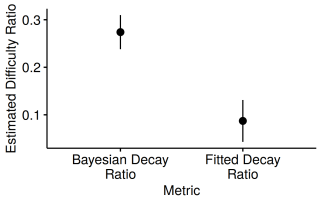
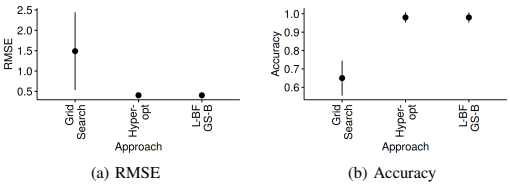


Fig. 6: **In the Duolingo App, Users Learning C-level Words Have Longer Learning Bursts Than Those Learning A-level Words.** We fit two Hawkes processes to users with 10 word learning events on Duolingo: one for users studying "hard" (C-level) words, and the other for users learning "easy" (A-level) words. We posit that the decay value of the C-level process is half as large as the A-level one, and we depict the ratio of the former to the latter. The ratio of fitted values ("Fitted Decay Ratio") is lower than that of posterior predictive density means ("Bayesian Decay Ratio"), due to a conservative prior parametrization. Lower decay values for the same number of events imply longer learning bursts, and this suggests that learning C-level words requires more sustained effort.



(a) RMSE (b) Accuracy

Fig. 7: **Bayesian Inference of the Decay in the Presence of Exogenous Shocks.** Here, the synthetic data setup is similar to that behind Fig. 5, but we now simulate half of the realizations with a Hawkes process with incremented β value. Capturing such a changepoint with Bayesian inference requires an intractable model. We encode a contrary hypothesis (i.e., that β decreases after a certain number of realizations), and we measure not only parameter RMSE (cf. Fig. 7a), but also the accuracy in correctly estimating that β increased (cf. Fig. 7b). We obtain higher RMSE values than in previous experiments, which reflect the misaligned hypothesis. Nevertheless, we almost always recover the correct direction of the change in β, as accuracy is close to 1.

power due to smaller sample size. We repeat this random sampling for a total of 100 times to compute bootstrapped 95% confidence intervals (as a straightforward alternative to deriving a more complex Bayesian statistic to compare the two user groups). We posit that each Duolingo user learning the A-level set of words represents a realization of a univariate Hawkes process, and users learning the C-level words represent realizations of another univariate Hawkes process. Mapping the six CEFR levels (A1, A2, B1, B2, C1 and C2) to a scale from 1 to 6, we naively assume, for illustration purposes, that the C-levels may be more than twice as hard as A-levels. If the data reflects that hypothesis, then we expect that a short learning burst suffices for grasping A-level words, in contrast to the C-level words, which may require perhaps more than two times as much effort over time. We encode this hypothesis in our closed-form Bayesian framework (with L-BFGS-B) as $b_{\text{C-level}} = 1$ and $b_{\text{A-level}} = 2$, since, after controlling for the total event count, we interpret the former as corresponding to longer periods of higher intensity, when compared with the latter.

Results. In Figure 6, we depict the ratio of posterior predictive density means $\beta'_{\text{C-level}}/\beta'_{\text{A-level}}$ ("Bayesian Decay Ratio"), as well the analogue ratio computed on the basis of the mean of actual L-BFGS-B estimations for both levels ("Fitted Decay Ratio"), i.e., $\{\hat{\beta}_{\text{C-level}}/\hat{\beta}_{\text{A-level}}\}_{k=1}^{28}$. Overall, we confirm Settles and Meeder's hypothesis that word difficulty correlates positively with the effort required to learn them: The posterior (and fitted) decay values of the C-level words are lower than those of the A-level words, resulting in more prolonged learning bursts in the former vs. the latter. Moreover, we underscore that this practical example illustrates the usefulness of our framework as a diagnosis tool: We observe a moderate shift away from the hypothesized $1:2$ difficulty ratio and towards a posterior predictive density mean value slightly below $3:10$. This finding indicates that the CEFR language levels may not directly translate to numerical scales for quantifying learning progress. Further, the relatively small size of the user sample highlights the importance of using Bayesian inference to not only diagnose a-priori estimates, but also to quantify estimation uncertainty. Therefore, although we suggest caution in extrapolating our results, we believe that our observations may contribute to ongoing research on the challenges of quantifying language learning progress [33].

C. Addressing Breaks in Stationarity

We turn our attention to the assumption of stationarity and its effect on fitting the decay parameter of Hawkes processes. Recall that stationarity implies that the intensity of a Hawkes process is translation-invariant. In practical applications such as the study of virality of online content [22] or the growth of online communities [21], exogenous shocks or exponential growth break the stationarity assumption. With synthetic data and a real-world example, we show how our Bayesian framework allows for assessing and capturing breaks in stationarity caused by exogenous shocks.

1) Synthetic Data: We start with the same experimental setup as in section V-B1, but we introduce two key differences. First, we assume that there was an underlying change in β at some point during the $K = 100$ realizations. We set the index of that change to $k^* = 50$, but our conclusions also hold for other choices of k^* (such as $k^* \in [30, 70]$). For $k < k^*$, we simulate the Hawkes process as previously, but, for $k \geq k^*$, we increment β by 1 while keeping the other

parameters unchanged, and we simulate from that updated process instead. With β_1 we denote β before the change, and with $\beta_2 = \beta_1 + 1$ the β afterwards. Second, we build a Bayesian inference setup to reflect the hypothesis that the β value changed at some point in the set of realizations. Such models are termed *changepoint models*. We propose the following intractable setup: A prior with $b_0 = \begin{cases} b_1, & k < \kappa \\ b_2, & k \geq \kappa \end{cases}$, where we set $b_1 \sim Exponential(1)$, $b_2 \sim Exponential(0.7)$ and $\kappa \sim U\{1, 100\}$, and an exponentially distributed likelihood. Note that the hypothesis $b_1 > b_2$ contradicts the true second part of simulated realizations. As the metrics for this experiment, we first measure the RMSE between the mean (respectively median) of samples from the posterior, $\bar{\beta}_1$ and $\bar{\beta}_2$ (resp. κ/K), and both true β (resp. k^*/K). Beyond RMSE, we also assess the estimation accuracy as the relative frequency of a correctly inferred ordering $\bar{\beta}_1 < \bar{\beta}_2$. We again repeat this whole process 100 times to derive bootstrapped 95% confidence intervals for the mean RMSE per fitting method.

Results. Figure 7 illustrates that the approaches' performance on RMSE (cf. Fig. 7a) is qualitatively, but not quantitatively, as in previous experiments. Specifically, the RMSE values of L-BFGS-B and Hyperopt are slightly below 0.5, which is one order of magnitude higher than previous RMSE values of below 0.05 (cf. Fig. 2g), despite the comparable range of Hawkes process parameter values in both experiments. This result reflects the higher complexity of the inference problem and the misaligned hypothesis. Further, we report accuracy values close to 1 (cf. Fig. 7b). Although we encoded prior parameter values which contradict the data, we still almost always recover the correct relationship $\bar{\beta}_1 < \bar{\beta}_2$ with the L-BFGS-B and Hyperopt methods. Qualitatively, this suggests identifying the direction of distributional changes in the decay is feasible. Quantitatively, we expect this procedure to yield conservative estimates of the magnitude and timing of the change: Bayesian inference features an inherent "inertia" of a few realizations when updating the posterior after the shock.

2) Strength of Collective Effervescence: Our third real-world scenario concerns the manifestation of collective effervescence on Twitter in response to the 13. November 2015 terrorist attacks in Paris, as studied by Garcia and Rimé [25]. They proposed future work could analyze how tweet timings reflect collective emotions surrounding the attacks. We address this suggestion by fitting the intractable Bayesian framework presented in section V-C1 with the L-BFGS-B method. Specifically, we begin by extracting the timestamps of tweets by users in a two week period centered on the day of the attacks. We model each user's behavior per week as a realization of a univariate Hawkes process, and we also control for tweeting activity per user: We extract all 205 users who tweeted between 20 and 25 times in the week before and in the week after[7]. Lowering the activity bounds yields more users each with less events to fit, while increasing those bounds has the opposite effect. However, by setting the activity bounds to different ranges of 5 tweets (specifically, 10 to 15, 15 to 20, ..., or 45 to 50 tweets), we qualitatively observe the same outcomes. We hypothesize that Twitter users, who partake in the collective emotion as a reaction to the shock, feature a sustained burst of activity. We expect such a burst of activity to translate into a decrease of the decay value after the shock. To numerically capture this hypothesis, we simply set $b_1 = 1.5, b_2 = 1$. However, repeating this experiment with an opposing hypothesis (e.g., $b_1 = 1, b_2 = 1.5$), again leads to the same results.

[7]Note that this extraction process results in a total of 410 realizations, i.e., 205 realizations before the shock and another 205 afterwards.

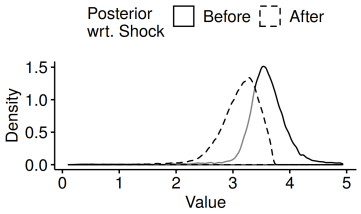


Fig. 8: **Tweet Timing Reflects Collective Emotions.** We depict the result of fitting an MCMC-based changepoint detection model to users' Tweet timings in the two weeks surrounding the November 2015 Paris attacks: The estimated posterior density for the Hawkes process intensity decay assigns more probability mass to higher decay regions before the shock in comparison to afterwards. This suggests that collective effervescence manifests on Tweet timings, as lower decay values after the shock reflect more sustained bursts of activity.

Results. Figure 8 depicts the density of the distribution of the inferred decay posterior before and after the shock. As expected, we confirm the decay value goes down in the week after the attacks, suggesting more sustained bursts of Tweeting activity in response to the attacks. This, in turn, supports the hypothesis that Garcia and Rimé advanced: Reaction timings, in the form of longer bursts of tweets afforded by a 15% lower mean posterior decay after the shock, reflect this collective emotion. We note that this is a conservative estimate of the decrease in the parameter, since activity levels quickly revert back to a baseline within the week after the attacks themselves, as Garcia and Rimé report. Further, this changepoint detection approach also over-estimates the time of the change at realization number 235, i.e., 7.1% later than the first Hawkes process realization after the shock, corresponding to realization number 206.

VI. Further Related Work

One of the first fields to leverage the seminal work by Hawkes [26] includes seismology [1], [29]. Since then, Hawkes process theory and practice emerged in the realm of finance [7], [18], as well as, more recently, in modeling user activity online [6], [8], [9], [11], [19], [21], [22], [27], [34]. More specifically, the latter body of work extended Hawkes processes to predict diffusion and popularity dynamics of online media [22], [27], [35], model online learning [6], capture the

spread of misinformation [8], and understand user behavior in online communities [9], [21], in online markets [11] and in the context of the offline world [19]. As all of those previous references interpreted the parameter values of Hawkes processes (and variations thereof), they may benefit from our study of the decay parameter, especially as we uncover its properties and assess and mitigate estimation issues with Bayesian inference.

Perhaps closest to our work is Bacry et al.'s [18] study of mean field inference of Hawkes process values. In particular, those authors inspected the effect of varying the decay parameter across a range of values: With increasing decay, fitted self- and cross-excitations decrease while baseline intensity increases. We go beyond their study by deepening our understanding of the (noisy) properties of the Hawkes log-likelihood as a function of the decay. Methodologically, our Bayesian framework relates to Hosseini et al.'s [35]. Those authors infer the decay parameter by assuming a Gamma prior and computing the mean of samples from the posterior (as part of a larger inference problem). In our work, we instead focus on the Bayesian framework as a means to quantify estimation uncertainty. Further, as our Bayesian changepoint model captures breaks in stationarity, we simplify previous work [21], [22] which relies on additional assumptions, such as estimating stationarity via the time series of event counts.

VII. Conclusion

In this work, we formalized, empirically characterized and mitigated the problem of fitting the decay parameter of Hawkes processes with exponential kernels. The inherent difficulties we found in accurately estimating the decay value, regardless of the fitting method, relate to the noisy, non-convex shape of the Hawkes log-likelihood as a function of the decay. Further, we identified problems in estimating uncertainty and diagnosing fitted decay values, as well as in addressing breaks of the stationarity assumption. As a solution, we proposed a parsimonious Bayesian framework. We believe our extensive evaluation of that framework across a range of synthetic and real-world examples demonstrates its broad practical use.

Optimization techniques such as constructing convex envelopes or disciplined convex-concave programming may, in the future, help to optimize the Hawkes process likelihood as a function of the decay. We also believe exploring the potential of the vast Bayesian statistics toolbox for learning more from fitted (decay) parameter values is promising future work.

References

[1] Y. Ogata, H. Akaike, and K. Katsura, "The application of linear intensity models to the investigation of causal relations between a point process and another stochastic process," *Annals of the Institute of Statistical Mathematics*, 1982.

[2] Y. Aït-Sahalia, J. Cacho-Diaz, and R. Laeven, "Modeling financial contagion using mutually exciting jump processes," *Journal of Financial Economics*, 2015.

[3] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song, "Shaping social activity by incentivizing users," in *NIPS*, 2014.

[4] T. Liniger, "Multivariate hawkes processes," Ph.D. dissertation, 2009.

[5] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song, "Coevolve: A joint point process model for information diffusion and network co-evolution," in *NeurIPS*, 2015.

[6] U. Upadhyay, I. Valera, and M. Gomez-Rodriguez, "Uncovering the dynamics of crowdlearning and the value of knowledge," in *WSDM*, 2017.

[7] E. Bacry, I. Mastromatteo, and J.-F. Muzy, "Hawkes processes in finance," *Market Microstructure and Liquidity*, 2015.

[8] B. Tabibian, I. Valera, M. Farajtabar, L. Song, B. Schölkopf, and M. Gomez-Rodriguez, "Distilling information reliability and source trustworthiness from digital traces," in *WWW*, 2017.

[9] R. Junuthula, M. Haghdan, K. Xu, and V. Devabhaktuni, "The block point process model for continuous-time event-based dynamic networks," in *WWW*, 2019.

[10] W. Trouleau, J. Etesami, M. Grossglauser, N. Kiyavash, and P. Thiran, "Learning hawkes processes under synchronization noise," in *ICML*, 2019.

[11] T. Hatt and S. Feuerriegel, "Early detection of user exits from clickstream data: A markov modulated marked point process model," in *WWW*, 2020.

[12] E. Bacry, M. Bompaire, P. Deegan, S. Gaïffas, and S. Poulsen, "Tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models," *The Journal of Machine Learning Research*, 2018.

[13] J. Choudhari, A. Dasgupta, I. Bhattacharya, and S. Bedathur, "Discovering topical interactions in text-based cascades using hidden markov hawkes processes," in *ICDM*, 2018.

[14] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive hawkes process," in *ICDM*, 2015.

[15] F. Salehi, W. Trouleau, M. Grossglauser, and P. Thiran, "Learning hawkes processes from a handful of events," in *NeurIPS*, 2019.

[16] T. Ozaki, "Maximum likelihood estimation of hawkes' self-exciting point processes," *Annals of the Institute of Statistical Mathematics*, 1979.

[17] J. Da Fonseca and R. Zaatour, "Hawkes process: Fast calibration, application to trade clustering, and diffusive limit," *Journal of Futures Markets*, 2014.

[18] E. Bacry, S. Gaïffas, I. Mastromatteo, and J.-F. Muzy, "Mean-field inference of hawkes point processes," *Journal of Physics A: Mathematical and Theoretical*, 2016.

[19] T. Kurashima, T. Althoff, and J. Leskovec, "Modeling interdependent and periodic real-world action sequences," in *WWW*, 2018.

[20] F. Figueiredo, G. Borges, P. de Melo, and R. Assunção, "Fast estimation of causal interactions using wold processes," in *NeurIPS*, 2018.

[21] T. Santos, S. Walk, R. Kern, M. Strohmaier, and D. Helic, "Self- and cross-excitation in stack exchange question & answer communities," in *WWW*, 2019.

[22] M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck, "Expecting to be hip: Hawkes intensity processes for social media popularity," in *WWW*, 2017.

[23] B. Settles and B. Meeder, "A trainable spaced repetition model for language learning," in *ACL*, 2016.

[24] C. of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[25] D. Garcia and B. Rimé, "Collective emotions and social resilience in the digital traces after a terrorist attack," *Psychological Science*, 2019.

[26] A. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, 1971.

[27] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes," in *AISTATS*, 2013.

[28] J. Etesami, N. Kiyavash, K. Zhang, and K. Singhal, "Learning network of multivariate hawkes processes: A time series approach," in *UAI*, 2016.

[29] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Elementary Theory of Point Processes*. Springer, 2003.

[30] A. C. Türkmen, "hawkeslib," github.com/canerturkmen/hawkeslib, 2019.

[31] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Scientific Computing*, 1995.

[32] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *NeurIPS*, 2011.

[33] J. Hulstijn, "The shaky ground beneath the cefr: Quantitative and qualitative dimensions of language proficiency," *The Modern Language Journal*, 2007.

[34] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schölkopf, and M. Gomez-Rodriguez, "Enhancing human learning via spaced repetition optimization," *PNAS*, 2019.

[35] S. Hosseini, A. Khodadadi, A. Arabzadeh, and H. Rabiee, "Hnp3: A hierarchical nonparametric point process for modeling content diffusion over social media," in *ICDM*, 2016.

3.6 Can Badges Foster a More Welcoming Culture on Q&A Boards?

This manuscript targets research question three, which aims at uncovering empirical links between the excitation and temporal patterns of users of peer production systems and the evolution of the system itself. In this article, I examine short- and long-term effects of one tool community managers typically employ to steer user behavior: the introduction of a badge. The badge this work studies aimed at improving the onboarding experience of newcomers to Stack Exchange Q&A communities. To measure the outcomes of that badge, I propose measuring newcomer retention and the sentiment of reactions by the community with the VADER sentiment analysis tool [HG14]. This manuscript presents a difference-in-difference regression to control for temporal trends in measurements of the proposed metrics in the short- and long-term before and after the introduction of the badge.

I find that the badge had an ephemeral effect on the sentiment-based metric and did not counter long-term trends in the retention metric, a result which holds for communities dedicated to a wide range of topics. This result may indicate (further) limitations to the power of badges. Also, as the effects appear to not depend on the community topic, this warrants further research into the conditions under which community managers may deploy wide-ranging (rather than tailored) welcoming interventions.

Can Badges Foster a More Welcoming Culture on Q&A Boards?

Tiago Santos*
Graz University of Technology
tsantos@iicm.edu

Keith Burghardt, Kristina Lerman
Information Sciences Institute
University of Southern California
keithab@isi.edu, lerman@isi.edu

Denis Helic
Graz University of Technology
dhelic@tugraz.at

Abstract

Thriving online communities rely on a steady stream of newcomers to contribute new content. However, retaining newcomers has proven challenging. In this paper, we measure the success of an intervention used by Stack Exchange question-answering communities to create a more welcoming environment for newcomers. That intervention consisted in highlighting contributions by new users with a special indicator. We hypothesize that Stack Exchange's new policy would reduce negative reactions to new users and, ultimately, increase new user retention. We leverage causal modeling to assess the introduction of the so-called "new contributor indicator", and we find it did not counter user retention decline in the short- and long-terms. However, our results indicate it did reduce unwelcoming reactions towards newcomers in the short-term. Our work has practical implications for online community managers aiming to improve their onboarding processes.

Introduction

Online communities rely on a steady stream of newcomers to contribute new content in order to thrive. Retaining newcomers, however, is a challenge for many communities (Kraut and Resnick 2012). Better onboarding methods can lower barriers to entry for new users (Yazdanian et al. 2019) and improve their integration in the community (Allen 2006). Both effects serve to mitigate user churn (Yang et al. 2010; Slag, de Waard, and Bacchelli 2015) and allow communities to grow.

However, which onboarding methods are most suitable for a given community and which methods are the most effective in retaining new members? Previous work has shown that user badges can effectively steer individual behavior and incentivize participation in the Stack Exchange online question-answering (Q&A) communities (Anderson et al. 2013; Kusmierczyk and Gomez-Rodriguez 2018; Yanovsky et al. 2019). However, the effect of badges on new users, and their community-wide effect has not been fully investigated.

This work extends the line of inquiry from previous research by studying the impact of a "new contributor" indicator. On August 22, 2018, as part of an initiative to foster a more welcoming community culture (Hanlon 2019), Stack Exchange introduced this badge-like indicator, which appears in all questions and answers a user posts in the first week (and only in the first week) after her first question or answer (cf. Fig. 1). We hypothesize the introduction of this indicator would lead to:

H_n: higher retention of new users, and

H_c: fewer unwelcoming reactions by the Stack Exchange community to their contributions.

Note our hypotheses assess the new contributor indicator in terms of its impact on new users (do they churn less?) and on the community (does it become more welcoming?). We focus on *unwelcoming* reactions, as Stack Exchange identifies those as one of the main contributors to an unfriendly community culture (Hanlon 2019).

We measure the effects of the indicator in the short-term (within one month of the introduction) and long-term (within five months of the introduction). We control for long-term temporal trends in the community behavior by using a difference-in-difference regression.

In our experiments, we do not find evidence for H_n in the short-tern, nor in the long-term, but we do confirm H_c in the short-term. Our work thus sheds light on the causal effects of a new user indicator in counter-acting strong community trends. We conclude by reflecting on how community managers can capitalize on the short-term impact of badges for newcomers and thereby improve user onboarding.

Methodology

Data. We study the Stack Exchange network of community question-answering websites[1], a total of 168 communities with millions of users asking and answering questions on topics ranging from astronomy to writing. Stack Overflow, the largest (with eleven million users) and oldest community (online since 2008), is dedicated to questions related to programming. Although we focus on Stack Overflow, we also analyze all other Stack Exchange communities. We obtained a snapshot of the complete Stack Exchange network

*Research done during an internship at the Information Sciences Institute, University of Southern California.

[1]https://stackexchange.com/sites

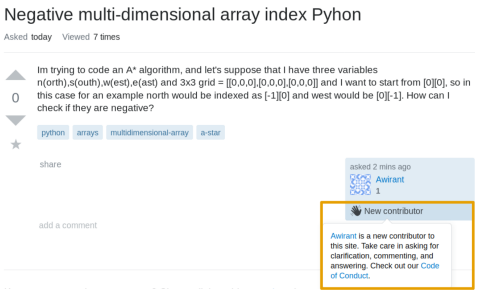


Figure 1: **New contributor indicator example.** In the first week after a user's first question or answer, the new contributor indicator highlighted in orange below her username encourages other users to mind their interactions with her.

as of May 2019[2].

Hypothesis Measurement. In this work, we study the effect of the introduction of a measure aimed at improving the onboarding process in Q&A communities, namely an indicator marking contributions by new users (see Fig. 1). We assess the effect of this new indicator by testing the hypotheses (H_n) new user retention increases, and (H_c) unwelcoming reactions by the community to new users decrease.

To test H_n, we define P_n as the proportion of new users who contribute a question or answer at least once within a fixed time period that extends from three days to three months after the first question. The range of the time period starts at three days to avoid including one-time users, who only return to the site to ask and follow-up on a single question. These users will not become active contributors to the community. Similarly, we set the range of the time period to end at three months to exclude users who are rarely active and are not significant contributors to the community. Our findings are robust to variations of this time window.

To test H_c, we define P_c as the proportion of first questions asked by new users which receive at least one comment with negative sentiment, based on the sentiment analysis tool VADER (Hutto and Gilbert 2014). We concentrate on the language of comments, as they provide a platform known for unwelcoming reactions (Silge and Punyon 2019). Following VADER's recommendation[3], we consider sentiment ≤ -0.05 to be negative.

Preprocessing. For both hypotheses, we focus on the first question (rather than first answer) asked by newcomers, since close to 80% of new users start by first asking questions rather than answering others' questions.

In our analysis of H_c, we consider only comments written within the one-week period after a user's first question, as that is time-frame of the "new contributor" indicator. We also separate first questions by the number of comments they obtained, test each of them separately, and report individual and aggregate results. This is because the probability of a negative comment naturally rises with more comments. We focus on first questions which receive between one and ten comments (97.5% of all first questions).

Experimental Setup. We assess whether our hypotheses hold in the *short-term*—a two-month window centered on August 22, 2018—and in the *long-term*—a ten-month window centered on August 22, 2018. Changing the short-term window to one (resp. four) months slightly reduces (resp. increases) the magnitude but not the statistical significance of our results and interpretations. The long-term window corresponds to the longest time-frame our data affords. In total, our short-term analysis of H_n comprises $60\,222$ new users, and the long-term one $562\,357$, whereas the short- and long-term analysis of H_c contains $36\,047$ and $364\,128$ first questions, respectively.

For the short-term window effects, we report the 95% bootstrapped confidence intervals (CI) for the weighted percent change in the levels of P_n and P_c. However, seasonal and other temporal trends could potentially distort the measurements of the long-term effects (Oktay, Taylor, and Jensen 2010). To control for such trends, we perform a difference-in-difference analysis (Abadie 2005). Specifically, we compare changes around August 22, 2018 to the same period in 2017, since there was no new contributor indicator then. We fit a linear model to the weekly time series of the metric values P_n and P_c:

$$P_i \sim I_{2017} + I_{Intervention} + Week \quad (1)$$

where I_{2017} is an indicator for the year 2017, $I_{Intervention}$ is an indicator for before or after August 22, 2018, and $Week$ stores the week of the year. Borrowing terminology from the causal inference literature, we name the year 2017 the *control* and 2018 the *treatment*. We inspect the magnitude and significance of the model's coefficients to assess the indicator's effects.

We also experimented fitting a logistic regression with the same regressors but a variation of P_i without temporal aggregation. In this alternative model, we let P_i be 1 for each user (resp. first question) which churns (resp. receives an unwelcoming reaction) as previously defined, and 0 otherwise. Although this model has a higher granularity, we obtain very similar quantitative results at a comparable statistical significance. For visualization purposes, we report on the weekly time series model only.

Finally, we extend both short- and long-term analyses of Stack Overflow to the other 168 communities of the Stack Exchange network. In this process, we exclude Stack Exchange communities too young to have the two years of data our long-term analysis requires, as well as those with fewer than an average of 100 new users and new first questions to analyze. That threshold corresponds to excluding communities with less than 0.2% of the newcomer activity we observe in Stack Overflow in the short-term. This leaves us with 50 communities ($\approx 30\%$ of all communities) to analyze H_n and 34 communities ($\approx\ 20\%$ of all communities) for H_c. We

[2]Data source: https://archive.org/download/stackexchange/.
[3]https://github.com/cjhutto/vaderSentiment#about-the-scoring

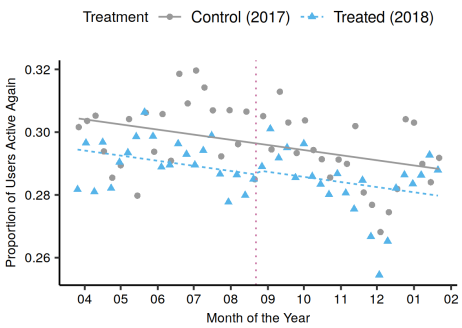


Figure 2: **No significant long-term change in user retention.** Controlling for temporal trends via a difference-in-difference regression on a user retention measure, we observe an overall downwards trend, which the introduction of the indicator (marked with a pink dashed line) did not curb.

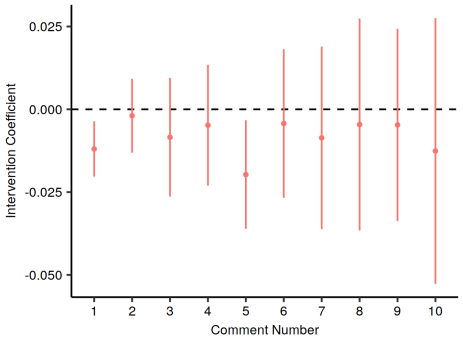


Figure 3: **Few significant long-term changes in unwelcoming reactions.** We control for temporal trends via a difference-in-difference regression on a measure for unwelcoming reactions. In almost all regressions studied, the intervention coefficient $I_{Intervention}$ is not significantly different from zero.

conservatively correct for repeated testing of our hypotheses by considering statistical significance at the Bonferroni-corrected p-value of $0.05/168 = 0.000298$.

Results

We first describe short- and long-term measurements of H_n and H_c on Stack Overflow, and then both time horizons of both hypotheses on other Stack Exchange communities.

Short-Term Effects. We observe small but statistically significant changes. Both user retention (P_n) and unwelcoming reactions (P_c) appear to decrease slightly. Specifically:

> H_n: P_n changed by -0.91% (CI: $[-1.28\%, -0.55\%]$, bootstrapped $p < 0.0001$).
>
> H_c: P_c changed by -1.13% (CI: $[-1.63\%, -0.64\%]$, bootstrapped $p < 0.0001$). Testing this hypothesis separately for first questions with different numbers of comments also resulted in significant changes with mostly the same magnitude and direction: In questions with between one and four comments (78% of all first questions), the weighted change is -1.65% and significantly different from zero (all four bootstrapped $p < 0.002 <$ Bonferroni-corrected $p = 0.05/10 = 0.005$). In questions with between five and ten comments (remaining 22%), the weighted change is 0.0075% and non-significant (all six bootstrapped $p > 0.13$).

When repeating this analysis for data in the year 2017, we find a statistically significant difference of similar magnitude in P_n, but no significant difference in P_c.

Long-Term Effects. We do not observe statistically significant changes as measured by the magnitude and significance of the $I_{Intervention}$ regressor. In the regression for H_n, the statistically significant downward trend of the regression persists in both years, potentially confounding the short-term estimates. On the contrary, the regression for H_c does not feature such a long-term trend. Specifically:

> H_n: The coefficient of $I_{Intervention}$ is 0.13% and it is not statistically significant (t-test $p = 0.72$), in contrast to all other regressors (t-test $p < 0.0006$). We depict this regression in Figure 2.
>
> H_c: Fitting separate regressions for first questions with different numbers of comments yielded mostly no significant coefficients (cf. Fig. 3; almost all t-test $p > 0.35$). The sole exception is the intervention coefficient of $I_{Intervention}$ in the regression for questions with one comment (t-test $p = 0.00499 <$ the Bonferroni-corrected p-value $0.005 = 0.05/10$). In particular, none of the separate regressions has a significant coefficient $Week$ for the temporal trend (all t-test $p > 0.11$ or remarkably larger), as exemplified in Figure 4.

Effects in Other Communities. In the few communities with enough data to analyze H_n, we find evidence supporting the hypothesis (after Bonferroni correction) of the short-term effect in only one of the communities, Software Engineering. The magnitude of the change is 5.36% (CI: $[1.79\%, 8.93\%]$, bootstrapped $p < 0.0001$), and there is no significant short-term change in 2017. However, in Software Engineering, both the effect and trends are absent in the long-term (in contrast to Stack Overflow), as none of the regression coefficients of our difference-in-difference analysis are significant. Regarding H_c, we find five communities (English, Mechanics, Travel, Android, and Worldbuilding) which benefit from the introduction of the new user indicator in the short-term (again, changes are statistically significant after Bonferroni correction and not present in 2017). The

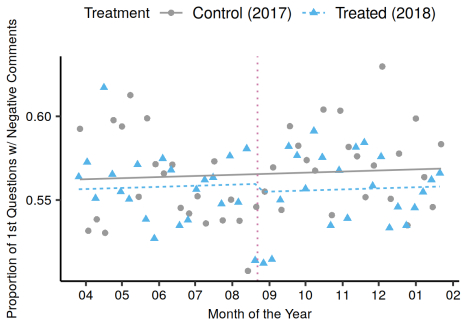


Figure 4: **Example of temporally stable community reactions to newcomers.** This difference-in-difference regression with only first questions which attracted four comments exemplifies the non-significant trend ($Week$) and intervention ($I_{Intervention}$) coefficients we observe across questions with varying numbers of comments.

change magnitudes range from -10.86% to -5.1% (bootstrapped $p < 0.0001$). Here, we again find no significant $Week$ or $I_{Intervention}$ coefficients for all H_c long-term regressions, similarly to our H_c results with respect to Stack Overflow.

Discussion

We inspected the effect of the introduction of an indicator to mark contributions by new users of Q&A communities. We do not find evidence for H_n (the indicator increased retention of new users) in the short- and long-terms, but our results support H_c (the indicator reduced unwelcoming reactions to new users) in the short-term.

On Stack Overflow, there is an ephemeral effect in the community's response to the new contributor indicator, but it does little to stem the long-term decline in new user retention. In particular, notice that P_n declines slightly in the short-term, but, in the long-term (see Fig. 2), there is a small but statistically non-significant positive change associated with the intervention. We reason this discrepancy arises from contextualizing fluctuations of the short-term estimate in the long-term: The decrease in August 2018 is not as strong as the overall long-term decline, and hence we observe a non-significant upwards jump slightly countering the significant downwards trend in new user retention. This result contrasts with our measurements for H_c, where none of the long-term trends are statistically significant. This, in turn, substantiates our finding that there is a reduction in unwelcoming reactions to first questions in the short-term, and that this change subsides in the long-term. This may indicate veteran users become habituated to the new indicator. Therefore, although our results indicate positive short-term changes, enacting long-term changes may require further involvement from all community members. For example, in addition to highlighting the newcomers, community managers may want to also reward veteran users for being particularly welcoming, as well as for mentoring new users (Ford et al. 2018).

Although the smaller volume of data in other communities limits our ability to generalize our findings, our results show heterogeneous effects arising from the introduction of the indicator: While almost all communities do not register an increase in user retention resulting from the indicator, a non-negligible number of them feature significant short-term decreases in unwelcoming reactions. We also note that the topics of the communities where we measured such significant changes span a broad spectrum, indicating reactions to the new indicator may be independent of the topic. This is a surprising finding, given recent work (Dev et al. 2018; Santos et al. 2019) identified topic as a key component of other community development parameters. Thus, although this calls for further research on reactions to new badges, this finding may also encourage practitioners to deploy site-wide (as opposed to community-specific) welcoming initiatives.

Although we believe the metrics we propose to measure user retention and welcoming attitudes capture our hypotheses well, future work may leverage many other metrics. In particular, using VADER to measure comment sentiment reflects only one facet of how comments may be perceived as unwelcoming; future research efforts could aim to characterize other aspects of unwelcoming reactions to newcomers (Silge and Punyon 2019).

As in all causal inference, we cannot rule out that exogenous variables and unmeasured confounders could affect our results. For example, in H_c, the number of comments a question attracts may relate to its quality and need for commenting: Do poor first questions receive more comments? And how does answer quality change and potentially affect this intervention? It would be interesting to study how such factors and other confounding community characteristics, such as age, user mix or strength of norms (Chandrasekharan et al. 2018), may impact our findings.

Finally, extending our approach to measure the impact of badges in welcoming initiatives of websites beyond Stack Exchange would help characterize the potential and limits of badges in steering community culture.

Acknowledgments. We thank the anonymous reviewers for their valuable feedback on the manuscript. Tiago Santos is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology. The work was also supported, in part, by DARPA under contracts HR00111990114 and W911NF-18-C-0011.

References

Abadie, A. 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*.

Allen, D. 2006. Do organizational socialization tactics influence newcomer embeddedness and turnover? *Journal of Management*.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2013. Steering user behavior with badges. In *WWW*.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*.

Dev, H.; Geigle, C.; Hu, Q.; Zheng, J.; and Sundaram, H. 2018. The size conundrum: Why online knowledge markets can fail at scale. In *WWW*.

Ford, D.; Lustig, K.; Banks, J.; and Parnin, C. 2018. We don't do that here: How collaborative editing with mentors improves engagement in social q&a communities. In *CHI*.

Hanlon, J. 2019. Stack overflow isn't very welcoming: It's time for that to change. https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/.

Hutto, C., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Kraut, R., and Resnick, P. 2012. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press.

Kusmierczyk, T., and Gomez-Rodriguez, M. 2018. On the causal effect of badges. In *WWW*.

Oktay, H.; Taylor, B.; and Jensen, D. 2010. Causal discovery in social media using quasi-experimental designs. In *SOMA*.

Santos, T.; Walk, S.; Kern, R.; Strohmaier, M.; and Helic, D. 2019. Self- and cross-excitation in stack exchange question & answers communities. In *WWW*.

Silge, J., and Punyon, J. 2019. Welcome wagon: Classifying comments on stack overflow. https://stackoverflow.blog/2018/07/10/welcome-wagon-classifying-comments-on-stack-overflow/.

Slag, R.; de Waard, M.; and Bacchelli, A. 2015. One-day flies on stack overflow—why the vast majority of stackoverflow users only posts once. In *MSR*.

Yang, J.; Wei, X.; Ackerman, M.; and Adamic, L. 2010. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. In *ICWSM*.

Yanovsky, S.; Hoernle, N.; Lev, O.; and Gal, K. 2019. One size does not fit all: Badge behavior in q&a sites. In *UMAP*.

Yazdanian, R.; Zia, L.; Morgan, J.; Mansurov, B.; and West, R. 2019. Eliciting new wikipedia users' interests via automatically mined questionnaires: For a warm welcome, not a cold start. In *ICWSM*.

3.7 What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic

The last article of this thesis also addresses the third research question, which aims at understanding how user excitation and temporal patterns shape the evolution of peer production systems. This article tackles that research question with a comparison of the wisdom of the few vs. the crowds, to grasp which activity dynamics may support the potential of the wisdom of the crowds. Specifically, this work addresses the research gap in comparing the wisdom of the few to that of the crowd in a peer production system dedicated to reviews and opinions on an experience good, video games. Leveraging descriptive statistical methods to quantify temporal activity dynamics of the few (experts) and the crowds (amateurs), as well as natural language processing and the Latent Dirichlet Allocation topic model to grasp the content of the reviews, this work distills significant discrepancies between experts and amateurs with respect to temporal reviewing dynamics and textual appraisal of video games. These discrepancies lie at the heart of one of this paper's core results, which is: In prediction experiments for the reception of video games among amateurs, leveraging both expert and amateur opinion yields the best prediction performance. These results may help to guide peer production system managers aiming to support and optimize work and collaboration by the crowds and the few.

What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic

TIAGO SANTOS, Graz University of Technology, Austria
FLORIAN LEMMERICH, RWTH Aachen University, Germany
MARKUS STROHMAIER, RWTH Aachen University, Germany
DENIS HELIC, Graz University of Technology, Austria

As video game press ("experts") and casual gamers ("amateurs") have different motivations when writing video game reviews, discrepancies in their reviews may arise. To study such potential discrepancies, we conduct a large-scale investigation of more than 1 million reviews on the Metacritic review platform. In particular, we assess the existence and nature of discrepancies in video game appraisal by experts and amateurs, and how they manifest in ratings, over time, and in review language. Leveraging these insights, we explore the predictive power of early expert vs. amateur reviews in forecasting video game reputation in the short- and long-term. We find that amateurs, in contrast to experts, give more polarized ratings of video games, rate games surprisingly long after game release, and are positively biased towards older games. On a textual level, we observe that experts write rather complex, less readable texts than amateurs, whose reviews are more emotionally charged. While in the short-term amateur reviews are remarkably predictive of game reputation among other amateurs (achieving 91% ROC AUC in a binary classification), both expert and amateur reviews are equally well suited for long-term predictions. Overall, our work is the first large-scale comparative study of video game reviewing behavior, with practical implications for amateurs when deciding which games to play, and for game developers when planning which games to design, develop, or continuously support. More broadly, our work contributes to the discussion of wisdom of the few vs. wisdom of the crowds, as we uncover the limits of experts in capturing the views of amateurs in the particular context of video game reviews.

CCS Concepts: • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

Keywords: video game reviews; review aggregator websites; wisdom of the crowds

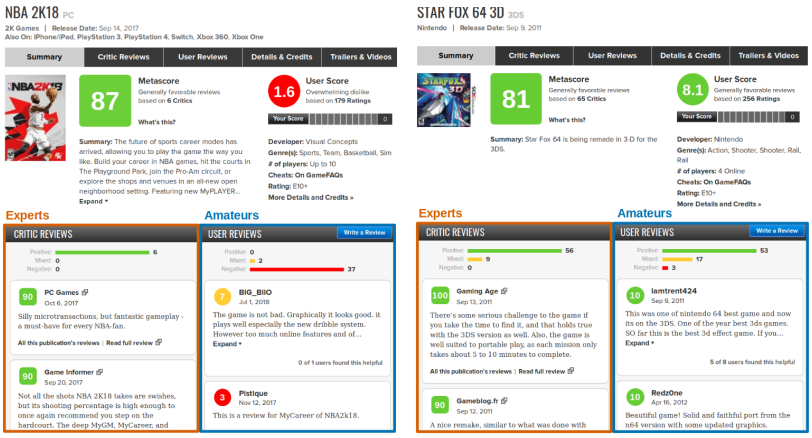
ACM Reference Format:
Tiago Santos, Florian Lemmerich, Markus Strohmaier, and Denis Helic. 2019. What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, CSCW, Article 140 (November 2019). ACM, New York, NY. 22 pages. https://doi.org/10.1145/3359242

1 INTRODUCTION

Understanding the many facets of how expert knowledge compares to the wisdom of the crowds is a fundamental research question [6, 19, 44, 66, 69]. Is expert-produced knowledge more reliable and trustworthy than that by crowds? In particular, in subjective domains, such as experience goods [54], are expert opinions representative of or even needed at all by the crowds?

Authors' addresses: Tiago Santos, Graz University of Technology, Graz, Austria, tsantos@iicm.edu; Florian Lemmerich, RWTH Aachen University, Aachen, Germany, florian.lemmerich@cssh.rwth-aachen.de; Markus Strohmaier, RWTH Aachen University, Aachen, Germany, markus.strohmaier@cssh.rwth-aachen.de; Denis Helic, Graz University of Technology, Graz, Austria, dhelic@tugraz.at.

2573-0142/2019/11-ART140
https://doi.org/10.1145/3359242



(a) Expert-amateur *discrepancies* (b) Expert-amateur *agreement*

Fig. 1. These sample Metacritic webpages display stark contrasts (Fig. 1a) as well as noteworthy agreement (Fig. 1b) in the appraisal of video games by experts and amateurs. This work analyzes differences and similarities in the review ratings and text depicted in the lower part of the Figures, namely the "critic (or expert) reviews" (orange) and "user (or amateur) reviews" (blue) section.

A prominent example of frequent clashes between expert and crowdsourced views on experience goods are video game reviews. As the video game press (henceforth *experts*) and the crowd of gamers (henceforth *amateurs*) produce video game reviews separately while following different motivations, remarkable evaluative discrepancies in their reviews may arise (e.g., see Figure 1a). Such discrepancies influence reputation and ultimately sales of a video game, frequently resulting in controversy surrounding experts and game developers, and occasionally even causing amateur-called boycotts[1]. However, as depicted in Figure 1b, there are examples highlighting the opposite may also occur, as experts and amateurs also express agreeing views on a video game's reputation.

Recently, only a few initial studies [31, 57] analyzed this phenomenon by investigating differences and similarities in video game rating and review language of experts and amateurs, and we still lack a large-scale analysis of video game reviews. The facts that video game reviews impact (i) sales as estimated via numerous regression models [11, 21, 74, 75], (ii) player experience as surveyed by multiple authors [28, 42, 43] and potentially even (iii) developer plans as implicated by the game piracy study of Drachen et al. [17] further support our view that video game reviews pose a relevant and fertile testbed for comparing expert with crowdsourced inputs.

Research Questions. Hence, in this work, we address the following research questions:

RQ1 (*Existence*): Is there substantial disparity between experts and amateurs in their video game appraisal?

RQ2 (*Characterization*): How can we characterize potential disparities, in particular along temporal and textual dimensions?

RQ3 (*Impact*): What are predictive strengths and weaknesses of expert and amateur reviews for short-term and long-term video game reputation?

[1]The video game "Star Wars Battlefront II" is a prominent example of a severe disagreement between experts and amateurs, as indicated by mixed press coverage and overwhelmingly negative amateur reactions, in particular calling for a game boycott (e.g. https://www.metacritic.com/game/pc/star-wars-battlefront-ii/user-reviews#user_review_7754293).

Approach. To address our research questions, we build a large-scale dataset of more than a million video game reviews from the video game portion of Metacritic, a popular online review aggregator. Within the frame our research questions provide, we operationalize video game reputation as amateur review ratings. We empirically analyze reviewing behavior of experts and amateurs, in particular along temporal and textual dimensions. Leveraging this analysis, we predict future short-term and long-term reputation of video games.
Findings. We find a more balanced rating behavior by experts in comparison to amateurs, whose ratings are more polarized and often indicate disagreement with experts. Experts review games shortly after release, on the contrary to amateurs, who still review games years after release and tend to give increasingly positive ratings in those cases. On the textual level, experts write more complex texts and strike a rather detached tone, whereas amateur review language is easier to understand and more emotional. Interestingly, the topic nostalgia arises in reviews by both. Based on these insights, we show video game reputation among amateurs can be accurately predicted with up to 91% and 80% ROC AUC based on amateur and respectively expert reviews in the short-term, and 72% and respectively 71% ROC AUC in the long-term. Prediction models based on reviews from both yield even better performances, suggesting experts and amateurs complement each other.
Contributions. Our extensive investigation of ratings and language in video game reviews by experts and amateurs is of unprecedentedly large scale, and offers actionable insights into similarities and discrepancies between expert and amateur video game appraisal. In particular, by bridging reviews of experts to those of amateurs and vice-versa, we help unify views on overall video game reputation, thereby supporting other video game industry stakeholders (and in particular producers and developers) attain a clearer grasp of the market. Further, we derive suggestions for online commerce platforms to weigh expert and amateur opinions in e.g. recommender systems. For example, video game recommender systems could, when computing recommendations, more appropriately quantify the impact of ratings awarded long ago or more recently, and provide higher weights for expert ratings of older games.

Overall, our work thus contributes to the overarching discussion on wisdom of the few vs. wisdom of the crowds from the particular, nuanced viewpoint of video game reviews.

2 RELATED WORK

Online Reviews. Previous work on online reviews focused on their rating and textual valence, as well as economical impact in sales of books [9, 12, 26], movies [18, 24, 26, 29, 33, 61, 70], restaurants [30, 45] and video games [11, 21, 74, 75]. As these works establish the existence and strength of a relationship between review characteristics and market performance of experience goods [54], we turn our attention to another indicator of public reception of video games in particular: the interplay between expert and amateur review ratings and texts. Multiple authors [9, 12, 26, 70] found that amateur review ratings of books and movies have a bimodal distribution, which may be indicative of a selection bias in items that amateurs review [35]. In the review helpfulness study of Danescu-Niculescu-Mizil et al. [12], the authors discuss how increasingly large variance of amateur ratings translates into a bimodal distribution of review helpfulness as evaluated by amateurs. We replicate these findings in our analysis of the amateur review rating distribution, and extend them via our comparison with the expert review rating distribution.

More closely related to our work are comparative studies of expert and amateur reviews of movies [14, 59, 70] and books [16]. In a survey of 169 college students with varying degrees of exposure to movies, Plucker et al. [59] find a continuum of rating behavior, where more experienced students rate movies more like experts. De Jong et al. [14] perform a textual comparison of 72 expert and amateur reviews of movies, which reveals amateurs evaluate movies from a personal standpoint and experts employ a more informative, contextualizing style. Estimating revenue and

sales of 136 movies and 179 digital cameras via expert and amateur reviews, Wang et al. [70] uncover significant interaction effects between expert and amateur reviews when product ratings have high variances. In their study of 100 book reviews, Dobrescu et al. [16] assess which book characteristics experts and amateurs appreciate more strongly. We go beyond these previous analyses by studying a large-scale dataset of video game reviews, and by exploring temporal rating behavior by experts and amateurs and how both textually manifest (dis-)appreciation for games.

On the topic of video game reviews, previous studies surveyed how players experience playing a game, and how this experience depends on their past exposure to video games [10] and relates to Metacritic critic and user scores [28]. In particular, Johnson et al. [28] call for a textual analysis of amateur reviews to explore further dimensions of player experience. This is a research gap which (i) Thominet [68] explores in a manual study of 180 amateur reviews on the online game shop Steam[2], (ii) Lin et al. [40] inspect in reviews of over 6224 games also on Steam, and (iii) we also address. Past correlational [57] and textual [31] studies of expert and amateur reviews on Metacritic cover the most similar kind of dataset to the one we handle, but those studies do not go beyond a static, small-scale analysis. Again, we expand on these studies with temporal and textual insights from a large-scale dataset of expert and amateur video game reviews on Metacritic. Also, we leverage these insights for short- and long-term prediction of game reputation among amateurs.

In research directions orthogonal to our work, multiple authors [12, 39, 46, 52, 53, 55] focus on fake review detection in a plethora of online marketplaces and review websites. Other authors uncovered the existence of herding effects in the language [20, 51] and rating [35, 36] of online reviews. In this work, we estimate the potential impact of fake reviews as rather negligible. Furthermore, we alleviate the short-term impact herding effects might have by focusing on the long-term horizon as well, as herding effects weaken over time [35].

Recommender Systems. In the literature on recommender systems, multiple authors [1, 7, 8, 15, 37, 41, 48, 49, 62, 73] leverage review text to predict user ratings per item and thereby improve item recommendations in various settings. Previous work [1, 8, 15, 41] utilizes reviews to address, in collaborative filtering and related approaches, common problems such as cold start, data sparsity and noise. Beyond addressing such data issues, Cheng et al. [8] and other authors [7, 37, 48, 49, 62, 73] extract topics [7, 8, 48], user sentiment [37, 73] and viewpoints [62], and categories and attributes [8, 49] of product reviews to provide interpretable recommendations and improve upon baseline recommender systems. Notably, Lei et al. [37] show jointly estimating item reputation and review sentiment of a user and her friends leads to improved recommender system performance. Note we predict game reception among *a group of users* (amateurs), while the previously mentioned recommender system and other natural language processing [60] and deep learning [67] methods predict ratings by *single users*. In particular, Tang et al. [67] extract user-specific word sentiment from movie and restaurant reviews to predict ratings, whereas we contextualize review text sentiment by user group. Overall, as our results advance temporal and textual understanding of video game reviews, our work may contribute to improve recommender systems leveraging review text.

Wisdom of the Few vs. Wisdom of the Crowds. Our work also contributes to the overarching discussion on the wisdom of individual experts vs. that of the crowds. Previous work [6, 19, 66] identified numerous instances of the performance of crowds surpassing that of experts. However, there is also evidence [5, 44, 69] for the deterioration and (e.g. cognitive) biases of crowds as well. With our comparison of video game review production by experts and amateurs, we extend the experts vs. crowds literature to the setting of experience goods [54]. In this setting, subjective views dominate discussion and opinion discrepancies may thus arise. Interestingly, recent work [64] suggests disagreements of the kind we uncover may positively influence (discussion) outcomes.

[2] store.steampowered.com

3 MATERIALS AND METHODS

Background on Metacritic. In this work, we study a dataset covering the video game reviews on the Metacritic platform. Metacritic is a popular online review aggregator which provides review ratings and text snippets by critics of many different experience goods [54], such as movies, TV shows, music records and in particular also video games. Beyond the review information provided by external critics we term *experts*, Metacritic also offers a platform for website users, which we term *amateurs*, to rate and review aforementioned experience goods. As illustrated in Figure 1, a Metacritic video game review website consists of (i) an upper part with aggregated review ratings by experts ("Metascore") and amateurs ("User Score"), and (ii) a lower part with individual expert and amateur reviews. In this work, we focus on this lower part of Metacritic video game review webpages, i.e. on expert and amateur reviews. These reviews consist of a rating on differently graded scales and a review text. We convert all expert ratings, originally mapped by Metacritic from diverse scales to a $0-100$ scale [50], to the amateurs' $0-10$ scale by dividing expert ratings by ten and odd-even rounding the result. Note that the expert review text provided by Metacritic is a brief summary of the actual review written by the expert. Metacritic typically provides a link to the expert's external website (if it exists, which is not always the case).

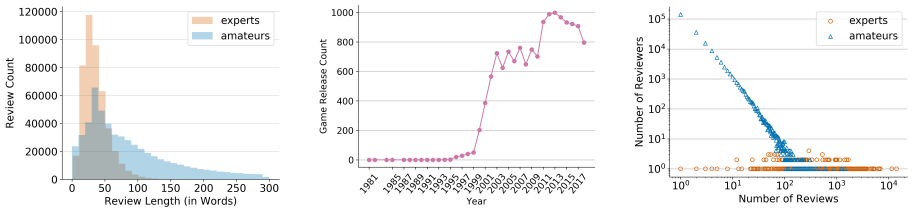
Data Collection. We collected all Metacritic video game reviews by experts and amateurs as of December 2018. Each video game review comprises, besides text and a rating, a review date, the name of the review author and, for amateur reviews only, the number of up- and downvotes. Note that Metacritic does not allow voting on reviews of experts. Besides reviewing data, we gathered video game metadata also available at Metacritic, namely the game platform, genre, developer and release date. In this work, we study rating and reviewing behavior on Metacritic; we thus do not parse experts' full reviews from their external websites.

Pre-Processing and Resulting Dataset. Game release dates in Metacritic may consist only of a quarter and a year. We accommodate such cases (less than 300 in total) by postulating the release date as the first day of the first month in that quarter. Furthermore, we prune all reviews written in languages other than English via Lui and Baldwin's language identification tool [47]. We list the descriptive statistics of our pre-processed dataset in Table 1. Our dataset features more than 400k games released over the course of 34 years and more than a million reviews. Despite the large imbalance in the number of experts (in total 482) compared to the number of amateurs (234,853), we observe a roughly 40-60% split of reviews authored by experts and amateurs.

Preliminary Descriptive Analysis. As expert reviews on Metacritic are only synopses of their full review, their median review length (in words) is considerably shorter than the amateurs' (c.f.

Table 1. Dataset statistics.

Dataset Statistics	
# games	417,391
# game genres	174
# platforms	21
# developers	115,662
Game release years	1984—2018
# experts	482
# amateurs	234,853
# reviews	1,088,731
...thereof # expert reviews	457,228
...thereof # amateur reviews	631,503
Median length of expert reviews (in words)	32
Median length of amateur reviews (in words)	76
# helpful votes in amateur reviews	4,717,255



(a) Review text length distribution (b) Game releases per year (c) Number of reviews

Fig. 2. Dataset overview. We visualize key characteristics of our dataset (from left to right) and assess their impact on our empirical analyses: lengths of review texts, number of games released per year and number of reviews produced. In Figure 2a, we show the distribution of review length in words (amateur review length distribution truncated at the 90th percentile). Metacritic provides only expert review summaries and allows amateurs to produce longer reviews. We thus control for review length in our subsequent analyses. In Figure 2b, we depict the number of games released per year with a minimum of three expert and three amateur reviews on Metacritic. This number increases over time but stabilizes at an average of around 800 games since the year 2003. This period thus allows for isolating rating and reviewing biases with respect to older games. We illustrate the distribution of the number of reviews of experts and amateurs in Figure 2c. The number of reviews by amateurs follows a heavy-tailed distribution, whereas experts have few commonalities in the number of reviews they produce. We thus expect different levels of experience reviewing video games to correspond to complementary (rather than overlapping) viewpoints.

Figure 2a for a summary of these distributional differences). Opposite to the experts, there is a considerable amount of long and detailed amateur reviews, which are substantially different from short, evaluative summaries. Hence, to ensure a fair comparison basis, we control for length in our subsequent analyses of review text. Our control consists of an upper limit in the word count by reviews of both experts and amateurs. To determine a threshold for the word count upper limit, we compute the Standardized Mean Difference (SMD) on the following covariates of the text length: counts of characters, syllables and sentences. SMD is the difference between the means of a covariate in amateur and expert reviews divided by the standard deviation of that covariate in amateur reviews. We set the word count upper limit to the 95th percentile (i.e. 72 words) of the word count distribution of expert reviews, as this value minimizes the median absolute SMD of the three covariates. Changing the word count upper limit to values in the range of balanced covariates (i.e. those with absolute SMD smaller than 0.25 [65]) did not significantly alter our results.

Further, any trends in amateur rating over the years could be confounded by the fact Metacritic features fewer older games (cf. Figure 2b), so the available ones are the more popular ones. However, between 2003 and the end of 2018, the number of games released per year and reviewed at least three times by both experts and amateurs on Metacritic stabilizes at around 800 games. Hence, in our analyses of long-term rating behavior, we focus on this time period, where trends in amateur rating over the years should not be (strongly) confounded by growing numbers of games available.

Finally, we inspect how many experts and amateurs produce how many reviews in Figure 2c. We find a heavy-tailed curve of amateur review counts, as most amateurs write a few reviews only and few of them write more than 1000 reviews. Experts, however, have much less commonalities in the amount of reviews they write. Interpreting review counts per reviewer as an experience indicator, we expect this discrepancy may result in opinion divergences between experts and amateurs.

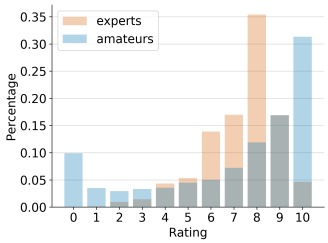


Fig. 3. Rating distribution. The review rating density of amateur and expert reviews differs in the number of peaks and their locations, as amateurs rate games notably often a zero or a ten, while experts an eight. This indicates higher polarization of ratings by amateurs in contrast to experts.

Methodology. To address our research questions, we first resort to a descriptive statistical analysis of rating by experts and amateurs. Specifically, to address our first research question, we examine the probability distribution of ratings by experts and amateurs.

For our second research question, we first perform descriptive statistical, comparative analyses of ratings by experts and amateurs along multiple dimensions, such as game metadata, helpfulness voting behavior by amateurs, and time (in the short- and long-term). Following this numerical study, we inspect the linguistic characteristics of the reviews themselves. To that end, we focus on textual form and style via analyses of textual complexity, readability and part of speech, and we assess review text content via sentiment analysis, subgroup discovery and Latent Dirichlet Allocation [4], an unsupervised topic model.

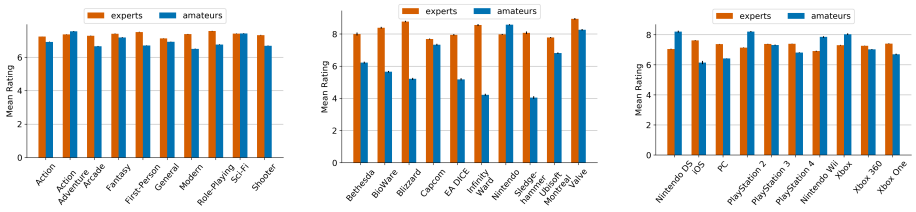
This empirical analysis underlies our answer to the third research question, i.e. our design of prediction experiments to forecast game reputation among amateurs from expert and amateur rating and reviewing data. We present the standard machine learning techniques we employ alongside our exposition of these prediction tasks.

4 EMPIRICAL RESULTS

We structure our analyses of expert and amateur video game reviewing discrepancies in two components corresponding to research questions RQ1 (Section 4.1) and RQ2 (Section 4.2). In the former, we inspect review rating distributions, and, in the latter, we characterize them across game metadata, temporal, and textual dimensions. Then, we leverage those insights to address RQ3 via prediction models for video game reputation among amateurs (Section 4.3).

4.1 RQ1 (*Existence*): Is there substantial disparity between experts and amateurs in their video game appraisal?

We answer **RQ1** positively, i.e. we find substantial disparity in video game appraisal by experts and amateurs. Operationalizing video game appraisal as review rating, we observe, in Figure 3, the rating distribution of expert ratings is unimodal and right-skewed, as experts most often award the grade eight. In contrast, reviews by amateurs form a bimodal rating distribution, with a peak at grade zero and another at grade ten. Distributional differences are significant at $p < 0.001$ according to a Kolmogorov-Smirnov two-sample test. Further, although the median rating of both experts and amateurs is equal to eight, the top quartile of expert ratings is eight and the bottom quartile six, and respectively ten and five in amateur ratings. Differences in mean and standard deviation of ratings underscore the rating discrepancies: The mean and respectively standard deviation of expert



(a) Mean Rating per Genre (b) Mean Rating per Developer (c) Mean Rating per Platform

Fig. 4. Mean rating of genres, developers and platforms of games with most reviews by both experts and amateurs. In this and all subsequent Figures (where applicable), error bars indicate bootstrapped 95% confidence intervals, many of which are too small to be visible. Although genre preferences appear more universal than platform and especially developer ones, we observe significant differences in all of them, signaling further discrepancies in expert and amateur tastes.

ratings is 7.34 and 1.62, whereas that of amateur ratings is 6.97 and 3.4 (bootstrapped differences in mean and in standard deviation are both significant at $p < 0.001$). To frame these results, out of 14,111 games with at least three expert and three amateur reviews, expert rating differs in absolute values from amateurs' by two or more grades in 4,945 games, i.e. in more than 35% of cases.

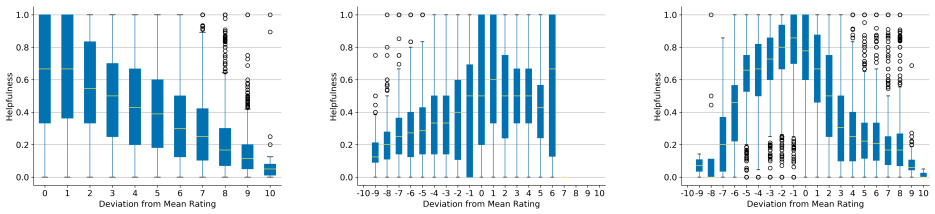
Combined, these findings indicate a strong disparity between expert and amateur video game ratings, and thus their appraisal. Moreover, we observe a more balanced rating behavior by experts, when compared to rather polarized ratings given by amateurs.

4.2 RQ2 (*Characterization*): How can we characterize potential disparities, in particular along temporal and textual dimensions?

In our answer of **RQ2**, we uncover rating discrepancies related to game metadata, amateur review helpfulness, time, and review text. Specifically, amateurs signal clearly different appreciation for top game developers, genres and platforms, when compared to experts. By upvoting critical amateur reviews more highly when disagreeing with experts than when both agree on a rating, amateurs appear to signal the strength of expert-vs.-amateur discrepancies in video game appraisal. Experts award comparable ratings over the years and review games shortly after release. On the other hand, amateurs give lower and more polarized ratings over time, and not only react more slowly to a game release, but also review games years after it and under an increasingly positive light. Finally, experts write rather formal, complex and detached reviews in comparison to amateurs, who exhibit more emotions in their reviews. We also find evidence of nostalgic reminiscence in reviews by both.

Experts Are More Attuned to Game Meta-Properties. We turn our attention to relationships between expert and amateur ratings of game meta properties. Figure 4 highlights differences in mean rating associated with genres, developers and platforms of most reviewed games by both experts and amateurs. Specifically, we plot top ten genres, developers and platforms which gathered the highest number of reviews by both experts and amateurs. All differences between expert and amateur mean ratings across genres, developers and platforms are statistically significant ($p < 0.001$, Welch t-test, except for the Sci-Fi genre ($p = 28.44$)). We see experts rate on average almost all genres, developers and newer platforms notably higher than amateurs. This indicates experts are more attuned to and amateurs more critical of certain meta-properties of games they review.

Amateur Review Helpfulness Signals Discrepancy. Next, we investigate how helpfulness ratios (i.e., the number of "helpful" votes divided by total votes) of amateur reviews relate to



(a) Distribution of helpfulness ratio per absolute deviation of rating from the mean

(b) Distribution of helpfulness ratio per signed deviation from the mean rating of games on which experts and amateurs *agree*

(c) Distribution of helpfulness ratio per signed deviation from the mean rating of games on which experts and amateurs *disagree*

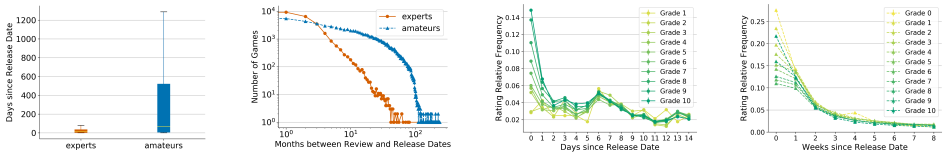
Fig. 5. Helpfulness of amateur review rating relates to its deviation from mean game rating. In Figure 5a, we see helpfulness of amateur reviews declines with higher absolute deviations of review rating from mean game rating. In Figures 5b and 5c, negative x-axis values indicate when a review awards grades lower than a game's average. Median review helpfulness is higher for reviews with above-average ratings when amateurs agree with experts (Fig. 5b), and notably higher for below-average ratings in the case of disagreement (Fig. 5c). This suggests the use of review helpfulness as a metric for the strength of expert-amateur discrepancies in rating, and it also indicates amateurs coordinate to signal distaste for games experts and other amateurs did not sufficiently condone.

discrepancies between expert and amateur reviews. In that direction, Figure 5a shows boxplots with the distribution of helpfulness ratios conditioned on the absolute deviation of the review rating from the overall mean rating of a game. We observe that helpfulness ratios of reviews are in general considerably lower if their rating differs from the overall mean (see [12] for comparable results in another domain). For a more detailed analysis, Figure 5b and 5c show distributions of signed deviations restricted to the 500 games with the lowest and respectively highest differences between expert and amateur ratings[3]. For these two groups, we observe opposite effects. For games where experts and amateurs agree, more positive reviews are generally considered more helpful: Note median helpfulness is overall higher for positive deviations from mean rating than for negative deviations (cf. Fig. 5b). In contrast, when experts and amateurs disagree, moderately more critical reviews are seen as decisively more helpful: The median helpfulness of reviews with ratings one to three grades below the mean is higher than those of reviews with above-average grades (see Fig. 5c). Recalling from Figure 3 that rating discrepancies often occur when experts award a high grade and amateurs a low one, we deduce amateurs give more helpfulness votes to more punishing reviews. The analogous holds for reviews with high grades, as amateurs upvote slightly more positive reviews. Overall, we thus find review helpfulness signals expert-amateur rating disagreement.

Experts Review Games Swiftly After Release; Amateurs (Still Review) Years After. We illustrate the reaction times of experts and amateurs to game release[4], in a time scale of days in Figure 6a and months in Figure 6b. Overall, experts swiftly review games upon their release, posting their reviews in a median of nine days after game release (cf. Fig. 6a). Amateurs, on the other hand, review games in a median of 65 days after release. Further, amateurs review games for remarkably longer periods of time after release than experts, as signaled by the gap in the amount of games amateurs, rather than experts, reviewed as late as ten to 100 months after release (cf. Fig. 6b).

[3]We consider only games with at least three reviews.

[4]Here, we include only reviews from after game release; experts and amateurs may also review beforehand (via e.g. previews).



(a) Distribution of Days Passed between Game Release Date and Reviews

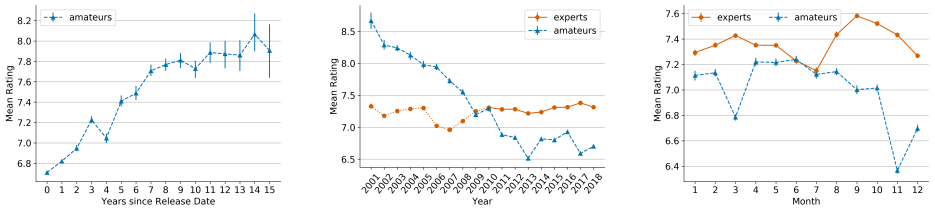
(b) Count of Games With Reviews per Month Since Game Release

(c) Mean Daily Expert Rating Frequency Relative to the Grade

(d) Mean Weekly Amateur Rating Frequency Relative to the Grade

Fig. 6. Timing and rating of reactions to game release. Fig. 6a depicts distributions of days passed between game release dates and reviews produced by experts and amateurs; Fig. 6b the number of games with reviews per month after game release date. Experts review games rather shortly after release, while amateurs produce more reviews for extended periods of time. We illustrate, in Fig. 6c and 6d, daily expert and respectively weekly amateur rating frequency relative to all reviews with a given grade. We choose a daily resolution for expert ratings since experts review games earlier than amateurs. Expert ratings converge faster, and are not as polarized as amateurs': Most frequent expert ratings early on are of rather similar colors (hence overall similar), while high frequencies of grades zero and ten by amateurs reveal polarization shortly after game release.

Amateurs, Moreso Than Experts, Award Polarized Ratings Shortly After Release. We visualize expert rating behavior in the first two weeks after game release in Figure 6c, and amateur ratings in first eight weeks in Figure 6d. Specifically, we plot, over time, the frequency of review ratings relative to all reviews of a given grade. In the plot of expert ratings given shortly after game release (cf. Fig. 6c), we observe not only rather balanced ratings (as grades of similar color are next to each other even early on), but also a rapid convergence of rating frequencies after six days post game release. As depicted in Figure 6d, the highest proportion of amateur grades are given in the week of game release. In contrast to experts, grade zero and grade ten are among the top three grades amateurs most frequently award in the release week. Further, more than 62% of reviews with grade zero and more than 50% of those with grade ten are written in these first eight weeks after release. By week four, grades awarded by amateurs all have roughly the same frequency. Overall, this rating behavior may indicate rather polarized reactions of amateurs shortly after game release.
For Amateurs, Games Were "Better In the Good Old Days". We observe, in Figures 7a and 7d, that, for games released in or after 2003, mean rating of amateur reviews increases, while standard deviation decreases, as the years pass after game release. We remark, again, this trend could be confounded by a relative lack of older games to review in comparison to newer ones, but Figure 2b dispels this concern for the time period spanning 2003 until 2018. Hence, we see this finding as an indicator for a positive rating bias, by amateurs, towards older games. We do not replicate the visualization for expert ratings, since experts review games shortly after release (cf. Fig. 6a).
Amateurs Become Stricter and More Polarized Over Time In Comparison to Experts. In Figures 7b and 7e, we observe (i) constant mean and slightly declining standard deviation of ratings by experts, in contrast to (ii) declining mean and strongly growing standard deviation of ratings by amateurs. Recall expert reviews from before 2010 are not dated on Metacritic. However, as experts review a game rather shortly after release (cf. Fig. 6 and discussion thereof), we estimate review year as the year of game release, and we represent mean estimated expert rating for the years before 2010 as a dotted line, which is in line with the development after 2010. These findings indicate experts seem to rate rather evenly over time, suggesting the existence of expert consensus. We



(a) Mean Ratings per Year After Game Release (b) Mean Ratings per Calendar Year (c) Mean Ratings per Calendar Month

(d) Standard Deviation of Ratings per Year After Game Release (e) Standard Deviation of Ratings per Calendar Year (f) Standard Deviation of Ratings per Calendar Month

Fig. 7. Long-term temporal evolution of ratings by experts (solid lines) and amateurs (dashed lines). The mean of review rating by amateurs increases, while the standard deviation decreases, with time passed since game release (cf. Figs. 7a and 7d). This indicates a positive rating bias of amateurs with respect to older games. Note, in Figs. 7b and 7e, the mean and standard deviation of review rating of experts does not change much over the course of calendar years (expert reviews of games from before 2010 are not dated — the dotted line segments indicate estimated expert review years). Amateurs, however, give lower mean ratings over time, and the standard deviation of their ratings increases. This may indicate amateurs are getting stricter and more polarized over time. As far as monthly ratings are concerned (cf. Figs 7c, 7f), there is a noticeable peak in expert ratings from August until November and a corresponding trough in amateur ratings around November, potentially signaling strong reactions to a surge in game releases before the holiday season.

interpret the declining mean ratings by amateurs as a sign of increasing strictness with games they review, rather than more recent games having lower quality (which expert reviews would have also reflected). Growing standard deviation in amateur ratings suggests increasing polarization among amateurs. Plotting mean and standard deviation of ratings per month (cf. Figs. 7c and 7f), we also observe seasonal differences in mean ratings by experts and amateurs, in particular in the months leading up to the holiday season.

Experts Write More Complex and Detached Reviews, While Amateurs More Accessible and Emotional Ones. After controlling for review text word count (cf. Fig. 2a and discussion thereof), we observe experts use longer words and sentences than amateurs, and the expert reviews are less readable according to the Flesch and Flesch-Kincaid readability scores[5] (differences significant at p-values < 0.001, Welch t-test). The samples of reviews of highly rated games we list in the upper part of Table 2 exemplify the more complex textual formulations employed by experts.

[5]We caution that reading ease scores may not be meaningful measures in this context of short texts and that pre-processing may impact readability score [34, 56].

A part of speech (POS) analysis of expert and amateur reviews reveals how experts and amateurs structure their reviews. Experts use more nouns (POS tag NN) and adjectives (JJ), as well as more conjunctions (IN and CC) and the 3rd person singular (VBZ): They write more like Academic English [3] and employ a more well-structured and distanced style. In contrast, amateurs notably use more personal pronouns like "me" and "myself" (PRP), suggesting they employ a more personally involved narrative in their video game reviews (differences in mean POS tag frequency are significant at $p < 0.001$, Welch t-test). The exemplary amateur review of a lowly rated game at the bottom of Table 2 gives a personal account of the amateur's experience with the game.

We further explore differences in emotional involvement via a sentiment analysis of review text. First, we observe the Spearman-rank correlation between review rating and compound review text sentiment, as estimated by VADER [27], values 0.31 for expert reviews and 0.55 for amateur reviews (both significantly different from zero and from each other, bootstrapped $p < 0.001$). Further, positive sentiment of expert reviews is on average 0.18 while that of amateurs 0.23, and negative sentiment of expert reviews is also on average lower than that of amateurs, at 0.08 vs. 0.09 (all differences significant, bootstrapped $p < 0.001$). This supports our view amateurs write more emotionally charged reviews than experts (cf. examples of Table 2).

Additionally, we employ a subgroup discovery technique with a binomial test interestingness measure [2] to identify words that are characteristic for either expert or amateur reviews, i.e., words that appear significantly more often in one of the review sets[6]. Overall, experts address certain game-related aspects more verbosely and less bombastically, as they mention "narrative" and "presentation" and employ substantives such as "charm" while avoiding adjectives like "horrible". On the other hand, amateurs use simpler, more emotionally charged vocabulary (such as "love" or "suck") in their reviews, which thus appear more exaggerated. Further, both example review comparisons we list in Table 2 support that review vocabulary is in line with our previous findings.

In Their Reviews, Experts Focus on Factual Descriptions of Gaming, While Amateurs Describe Gameplay and Experiences. To assess topic-level differences in expert and amateur reviews we apply a commonly used topic modeling technique, Latent Dirichlet Allocation (LDA) [4]. We adapt the setup outlined by Zhang et al. [72] and also use the Stanford Topic Modeling Toolbox [22] in our LDA application. In Appendix Section A.1, we provide details on our pre-processing procedure for LDA, a procedure which leads us to report on 15 topics. We list, in Table 3, the top ten

[6] As a result, the top 10 words most characteristic to expert reviews are "single player", "likely", "charm", "fails", "narrative", "offering", "certainly", "presentation", "might" and "manages". For amateur reviews, the most characteristic top 10 words are "say", "money", "bad", "think", "love", "people", "bought", "horrible", "play" and "suck".

Table 2. Exemplary reviews [sic] of highly and lowly rated games by experts and amateurs illustrate structural, lexical and emotional discrepancies between their review texts.

Game	Author	Grade	Text
MLB The Show 17	Expert	10	MLB The Show 17 builds on last year's version, which many cited as the best in the series, by adding some key features. [...] All of these additions make the game feel more like real baseball. This authenticity makes games more fun to play, as it's easier than ever to become invested in every pitch.
	Amateur	9	Freaking awesome. Any true ballplayer would appreciate the game they've put together with The Show. Didn't play another game for a full 7 months after it's release... it's THAT good.
Hannah Montana: The Movie	Expert	3	Hannah Montana: The Movie isn't a total flop. There are some good points to the game such as the array of songs, fairground games and the gorgeous Miley Cyrus, while this game isn't worth the price tag it's definitely worth a rent.
	Amateur	0	Worst game on the Wii ever. My little sister played this game and said loudly "I hate this game". So I took a look at the game to fin the reason why my lil' sis hates it. I totally agree with her hating it. The horrible graphic, tedious gameplay, and dumb minigames made this game suck BIG TIME. 0/10

topics (i.e. with greatest weight) we extract from reviews by experts and separately amateurs. The table includes a name for each topic, which we annotate manually, top ten most frequent words per topic and average topic weight per review. We include, in Appendix Section A.2, the bottom five LDA topics. Those topics reflect only video game genres, and we thus do not analyze them further.

These results indicate experts and amateurs discuss general game categories and their feelings about games using different vocabulary registers. Beyond obvious similarities in words used to describe common video game genres, this LDA topic analysis may also suggest game aspects experts and amateurs rather focus on: Experts appear to describe game modalities (words "online", "single-player" in the topic "Shooter Games") and context (topic "Sequels" which is not present in amateurs' topics), while amateurs seemingly rather discuss gameplay itself (words "mission", "combat" in the topic "Shooter Games") and appealing gaming concepts (words "experience", "art" in the topic "Gaming Experience" which is not present in experts' topics).

Finally, we highlight a clear-cut topic containing words related to gaming nostalgia ("Nostalgia" topic) in reviews by both experts and amateurs. Inspecting less frequent words in that topic and selecting those more closely related to nostalgia from the topic's top ten words (by experts and amateurs), we get a clearer picture of gaming nostalgia-related words, namely "old", "classic", "retro", "origin", "collection", "remastered" and the word "nostalgia" itself. As nostalgia appears inherently related with time, we study the temporal dynamics of those nostalgia keywords, as depicted in Fig. 8. Increasing usage of nostalgia-related vocabulary may point towards its growing influence over time, which may be key to grasping the nature and timing of long-term gamer preferences.

Summary of RQ2 Findings. Discrepancies in video game reviews by experts and amateurs manifest themselves on multiple levels. Experts are less critical of top game genres, developers

Table 3. We list, for experts and separately for amateurs, a label we propose for each of the top ten LDA topics, their top ten most frequent words, and average topic weight per review. The LDA analysis reveals not only how experts and amateurs evaluate games, but also which game aspects they discuss about them, such as genre and in particular also nostalgia for older games.

Experts			Amateurs		
Topic	**Top Words**	**Weight**	**Topic**	**Top Words**	**Weight**
Negative Reviews	bad, graphic, lack, repetitive, isn't, frustrating, design, idea, problem, poor	0.1426	Mixed Reviews	control, addictive, puzzle, worth, simple, little, music, boring, challenge, nice	0.1462
Positive Reviews	puzzle, challenge, platform, unique, adventure, level, design, charm, beautiful, visuals	0.1011	Action Games	fight, awesome, fan, mode, control, action, cool, boss, buy, hard	0.1013
Multiplayer Games	multiplayer, online, price, friend, arcade, worth, party, simple, addictive, lack	0.0876	Racing Games	race, car, mode, year, online, drive, buy, sport, control, awesome	0.1009
Horror Games	adventure, episode, character, horror, dark, end, atmosphere, narrow, survival, shadow	0.0718	Negative Reviews	buy, money, worst, bug, terrible, review, release, horrible, boring, waste	0.1007
Sequels	improves, world, content, origin, expansion, franchise, predecessor, sequel, addition, change	0.0685	Shooter Games	shooter, repetition, mission, weapon, bore, action, combat, nice, awesome, pretty	0.0824
Movie Games	movie, character, adventure, kid, humor, entertaining, gamer, ninja, repetitive, film	0.0630	Nostalgia	amazing, platform, control, origin, music, version, fan, classic, adventure, perfect	0.0723
Nostalgia	platform, challenge, classic, level, genre, design, man, old, legend, difficulty	0.0606	Gaming Experience	end, amazing, beautiful, effects, experience, episode, mass, series, choice, art	0.0690
Strategy Games	strategic, war, star, tactic, space, battle, genre, combat, campaign, depth	0.0570	Strategy Games	war, strategy, total, simulation, bug, battle, campaign, unit, space, patch	0.0527
Shooter Games	shooter, multiplay, campaign, online, weapon, co-op, zombie, dead, solid, single-player	0.0568	Adventure Games	rpg, world, combat, quest, fallout, amazing, dragon, mod, explore, fantasy	0.0510
Sports Games	sport, football, improvement, featuring, baseball, soccer, graphics, simulation, online, basketball	0.0564	Positive Reviews	amazing, perfect, year, awesome, life, think, masterpiece, greatest, simply, excellent	0.0503

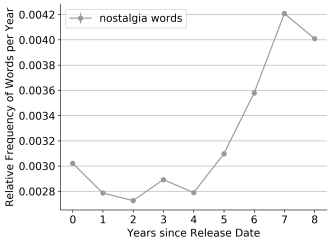


Fig. 8. Relative frequency of usage of nostalgia-related vocabulary in amateur reviews. The usage, by amateurs, of vocabulary related to nostalgia clearly increases over the years, hinting at its importance long after games get released.

and platforms than amateurs, who upvote (very) critical reviews in cases of large disagreement with experts. Experts swiftly review games after their release, whereas amateurs review games years after release and are positively biased towards older games. We observe amateurs (in contrast to experts) award lower, more polarized ratings over the years. Review language reflects a more professional and detached tone by experts, while the amateurs' tone is more casual and emotional.

4.3 RQ3 (*Impact*): What are predictive strengths and weaknesses of expert and amateur reviews for short-term and long-term video game reputation?

To assess the extent to which early expert and amateur reviews capture amateur video game reputation, we design two prediction experiments leveraging and validating our previous findings. Our results show early amateur review data is very predictive of both short-term and long-term video game reputation also among amateurs. Expert review data is on par with amateur review data in long-term predictions, and both combine to produce best forecasts in both prediction tasks.

4.3.1 Task Description. We conduct two binary prediction tasks, where we forecast video game reputation. We operationalize video game reputation as amateur review rating. For our binary prediction tasks, we define class separating thresholds. We exhaustively experiment with variations of these thresholds and while predictive performance of all prediction models decreases slightly with a weaker contrast between classes[7], our results and interpretations remain mostly the same.
Short-Term Prediction Task. Given all review data up to one week after release and video game metadata, we predict if amateur ratings per game, starting one month after game release, will be among the top quartile of amateur ratings (i.e. rating ten) or among the bottom quartile (corresponding to rating five or lower). As the amateur rating distribution is bimodal, a two class prediction target with well separated classes is in line with amateur rating behavior[8]. Figure 6d motivates our choice of the first week after game release as the time window for training prediction models. As amateur review ratings vary the most in this time window, prediction models with good performance would be most useful during this time period.
Long-Term Prediction Task. Given all review data up to one week after release and video game metadata, we predict if the number of reviews a game receives in the time period starting three

[7]We observe that the availability of more data for prediction partially offsets performance decreases due to less contrast between two classes.
[8]Another reason why we do not predict rating grades directly is data sparsity: Amateur grades of e.g. two are very underrepresented (cf. Fig. 3), and effectively predicting such grades would thus require over-sampling or merging into other classes (which would be similar to our setup).

years after release will be in the top or bottom quartile of the distribution of amateur review counts per game. We focus on review counts rather than ratings in the long-term due to the fact most review ratings per game are given in the first two months after game release, and thus average rating per game does not change much in the time period after that (cf. Fig. 6d). Amateur review counts of many games, however, still increase long after game release, as we observed in Figure 6b.

4.3.2 Experimental Setup. The pre-processing routines we use in our prediction tasks involve, as motivated by our discussion of Figure 2a, placing an upper bound on review length in words at 72 words. Further, we impose a minimum amount of reviews per game to include it in a train set. We again set this minimum value at three reviews, to balance between the amount of training instances (i.e. games) available for prediction, and unnecessary noise in the training dataset in the form of games with only two often very disparate (expert or amateur) reviews.

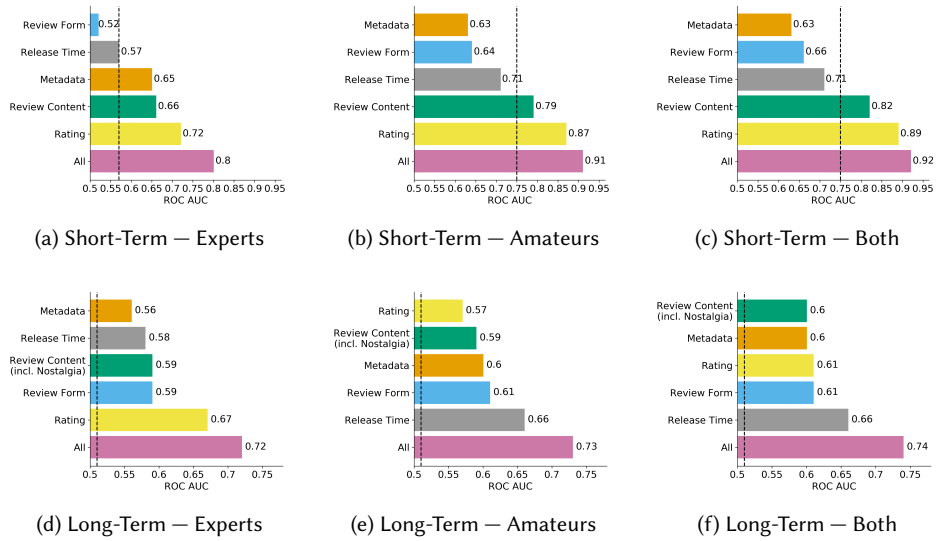
For both of our prediction tasks, we employ standard machine learning configurations provided by Pedregosa et al. [58], i.e., we do not perform hyperparameter search. We measure predictive performance via the area under the receiver operating characteristic curve (ROC AUC), the probability a binary classifier ranks a randomly chosen instance of a given class higher than another randomly chosen instance of the other class. We use a five-fold train-test split and we report on ROC AUC averaged over the folds. For both prediction tasks, we evaluate the classifiers logistic regression, support vector machines and gradient boosted trees. We present only the predictive performance of the latter, as they give the best results.

Prediction Models. Our empirical analysis of rating and reviewing behavior of experts and amateurs informs our choice of features for the two predictive tasks. We create a series of models with these features to assess their individual importance per prediction task. We define the following groups of features:

(i) Rating: Expert/amateur grade (Fig. 3), number of days elapsed between game release and review dates (Fig. 6).
(ii) Metadata: Gaming platform, game genre and developers (Fig. 4). We encode these metadata features as three binary variables indicating if a game's platform, genre or developer is among the top ten most frequently reviewed of its kind (or not).
(iii) Release Time: Year and month of game release date (Fig. 7).
(iv) Review Form: Number of syllables, number of words, number of sentences, readability features (Flesch and Flesch-Kincaid) (Sec. 4.2 "Experts Write More Complex Reviews").
(v) Review Content: Review sentiment, binary variables for the presence of words experts and amateurs use in reviews of top and bottom quartile ratings[9] and, for the long-term task, a binary variable for the presence of nostalgia-related words[10] (Sec. 4.2 "Experts Write More Complex Reviews", Table 3 and Fig. 8).
(vi) All: All the above mentioned features.

Furthermore, per prediction task, we devise two separate sets of models, with one using only review data from experts and the other only from amateurs. With this choice, we assess not only the extent to which expert ratings and reviews are predictive of amateurs', but we also compare their predictive performance with respect to an amateur data-only model. Also, we assess if both models combined lead to better predictive performance, as this would indicate experts and amateurs complement each other to form a fuller picture about future video game reputation. To further contextualize results, we also compare ROC AUC values of our models with a simple rule-based baseline: For a given game and its ratings in the first week, this baseline predicts the game belongs

[9]These word lists include the top ten most frequent words we uncovered in the subgroup discovery analysis and LDA topics. We use separate word lists for words employed by experts and for those by amateurs, i.e. we keep a total of four word lists.
[10]This word list includes words of the LDA topic "nostalgia" of both experts and amateurs.



(a) Short-Term — Experts (b) Short-Term — Amateurs (c) Short-Term — Both

(d) Long-Term — Experts (e) Long-Term — Amateurs (f) Long-Term — Both

Fig. 9. We predict game reputation among amateurs in the short-term (one month after game release) and long-term (three years after). We distinguish our models by their training data: reviews by experts (Figs. 9a, 9d), amateurs (Figs. 9b, 9e) and both (Figs. 9c, 9f). Vertical dashed lines indicate baseline model performance. Our features capture the first week of review data (feature groups "Review Content", "Review Form" and "Rating"), as well as timing ("Release Time") and other metadata ("Metadata") information available at game release. In both tasks, a combination of all feature groups and reviews by both yields highest ROC AUC values. Review rating is among the most predictive features, and the importance of other features varies with prediction horizon.

to the top rating quartile in the short-term (and, in the long-term prediction task, to the top review count quartile) if it was in the top quartile in the first week (i.e. in the training dataset) already.

Recall that we include, in pre-processing for prediction experiments, only games with at least three reviews. As the set of games with at least three expert reviews is different from the set of games with at least three amateur reviews, the expert review-based and amateur review-based models include different games and different quantities of games. However, the dataset sizes are comparable: We have a total of 728 games for the expert review-based and 809 games for the amateur review-based short-term prediction, and 904 games for the expert review-based and 1164 for the amateur review-based long-term prediction. Further, the prediction classes are roughly evenly represented in all experiments, as the class splits range from 39% / 61% to 48% / 52%.

4.3.3 Results. Figure 9 summarizes our results in both prediction tasks. At 0.91 ROC AUC, the model trained with all feature groups and on amateur review data achieves high classification performance values in the short-term prediction task (cf. Fig. 9b). In the long-term and using the amateur review-based models, we get 0.73 ROC AUC (Fig. 9e). The performance of the expert review-based models is lower in the short-term (Fig. 9a; 0.8 ROC AUC) and slightly lower in the long-term (Fig. 9d; 0.72 ROC AUC) prediction tasks. These ROC AUC values are significantly better than random or stratified majority baselines, which reach values of only 0.5. The baseline rule-based model (cf. dashed lines throughout Fig. 9) achieves only 0.57 ROC AUC in the expert review-based

and 0.75 in the amateur review-based short-term prediction, and 0.51 in both variations of the long-term prediction. Therefore, our models strongly outperform such a baseline as well.

We achieve maximal classification performance, at 0.92 ROC AUC in the short-term and 0.74 ROC AUC in the long-term prediction tasks, with models trained on both kinds of reviews (cf. Figs. 9c and 9f).

We now analyze the influence of our feature groups on the prediction performance. In all prediction settings, the combination of all features achieves the best performance. The "Rating" feature group appears to be the most predictive set of features in both prediction tasks and with all variants of the training data, with the exception of long-term prediction from amateur reviews. There, a model using only the "Release Time" feature group attains the second highest performance. However, review text feature groups also carry predictive importance, in particular "Review Content" in the short-term prediction, and "Review Form" in the long-term[11]. Finally, models combining expert with amateur reviews boast classification performance gains in the feature groups "Rating" and "Review Content" ranging from 0.01 to 0.04 ROC AUC.

5 DISCUSSION

In our study of an online video game review aggregator, we uncovered remarkable discrepancies (and few similarities) in video game appraisal by experts and amateurs along temporal and textual dimensions. We leveraged these characteristics for a series of prediction experiments for video game reputation among amateurs, which achieved high predictive accuracy. We now reflect on the implications of these findings for video game market players and the experts vs. crowds discussion.

Existence of Video Game Appraisal Discrepancies. On the level of video game review rating, we found substantial discrepancies, as the rating distribution of experts is unimodal and that of amateurs bimodal. Extrapolating this finding suggests (very) different perspectives of the few vs. those of the crowd on the same experience good may co-exist. Thus, we believe this poses a significant challenge when consolidating expert and amateur opinions of experience goods in general. Consensus finding in such settings requires appropriately understanding and weighing expert and amateur views, a task which benefits recommender systems for online marketplaces.

Characterization of Video Game Appraisal Discrepancies. We observed experts, often video game journalists[12], employ a clinical, detached writing style, which reflects their experience and is reflected in their unimodal rating distribution. Amateurs, on the other hand, often reviewing just a few games, using emotionally charged language and awarding rather extreme ratings, appear dedicated to either glorify games or express their grievances with them, and to at times signal their disagreement with experts. We thus believe gauging overall public opinion of a video game requires adequately discounting the emotion in often extreme video game ratings and reviews by amateurs to distill actionable (or more detached) feedback. This could in turn help video game developers and publishers with video game production. More broadly, our findings suggest reconciling subjective views of the crowd with those of the few entails learning to interpret their signaling mechanisms.

Beyond short-term signals, we also observed how amateurs tend to give lower ratings over the years and better ratings as time passes after game release, a development which may correlate with nostalgia for older games. We believe these trends reflect growing competition between developers in the video game market and increasingly stringent consumer demands, as well as a positive bias towards older classic games and leniency towards their outdated technical prowess.

[11] Adding the nostalgia-based feature to the "Review Content" group in long-term predictions improves expert review-based models by 0.02 (and amateur review-based ones by 0.01) ROC AUC in comparison to "Review Content" without that feature.
[12] The full list of expert publication venues, available at https://www.metacritic.com/faq#item20 as of December 2018, includes many well-known websites in the video game milieu, such as gamespot.com or ign.com, for which teams of video game professionals produce content.

Specifically, our study advances current understanding of nostalgia effects [13, 25, 63, 71] in the popular sense, i.e. as a kind of yearning for one's childhood [13, 25, 63], rather than the clinical and psychological sense [71]. With our large-scale empirical analysis, we show video game reviews can be leveraged to quantify strength and timing of nostalgia-related evaluations of video games. We believe these findings may potentially empower not only video game developers but also experience good producers to assess how and when to recycle elements from older developments.

Prediction of Video Game Reputation. We begin by briefly discussing the "Release Time" feature group: The fact this feature group does not perfectly predict long-term video game reputation simply shows there are older games which fail (and respectively newer games which succeed) at growing long-term reputation. This contradicts the assumption that, as older games have more time to amass more interest and hence reviews, this feature group alone should have great predictive accuracy. Not including this feature group impairs the prediction performance of the expert review-based model by -0.01 and the amateur review-based model by -0.06 ROC AUC. This underscores (i) the strength of expert reviews for assessing long term video game reputation and (ii) the growing importance of the temporal dimension in long-term amateur evaluation of video games.

As experts typically react within a few days of game release, our prediction models facilitate estimations of game reputation shortly after release date. Our prediction experiments also showed early amateur reactions are highly indicative of future reputation. This underlines the potential for game developers to engage with gamers early on and potentially address feedback with updates. Moreover, experts may want to reflect on better capturing short-term concerns by amateurs, as well as focus on their ability to write reviews that stand the test of time. Further, video game recommender systems leveraging review text may obtain performance gains by assigning greater weight to expert opinion of older games. We reason the wisdom of the few may lie herein: Their greater experience, when compared to the crowd, translates into valuable long-term views.

Recall we assess, with models trained on both expert and amateur reviews, if expert reviews can be regarded as complementary to amateur reviews. Using such models, we obtain performance gains in both prediction tasks of 0.01 ROC AUC on the full model and up to 0.04 ROC AUC on the feature groups related to review ratings and content. We find these performance improvements, albeit minor, remarkable, as ROC AUC values were already high with amateur review data only. Hence, we conclude experts (i.e. the few) provide value on their own and in addition to amateurs (i.e. the crowd), as both combine to form a more comprehensive view.

Limitations. In this work, we exclude game ratings of amateurs that do not contribute an accompanying review to their rating. This data on amateur ratings represents, however, another large user population, which silently provides feedback on games. We also expect the full reviews by experts, as written on their own websites and publication venues, to provide further insights into their rating and reviewing behavior. Furthermore, in this work, we characterize the group of experts and amateurs as a whole. We know, however, from Figure 2c and previous work [10, 59], that there are subpopulations of both experts and amateurs with varying degrees of expertise and thus reviewing behavior. Hence, though beyond the scope of the present study, we believe tailoring prediction models to individual amateurs is a promising direction for future work. We also stress our aim in this work is comparing experts and amateurs, rather than optimizing prediction models.

Another issue in review datasets is fake and duplicate review production and identification [12, 39, 46, 52, 53, 55]. As Mukherjee et al. [53] found, behavioral cues, rather than linguistic ones, are most effective in characterizing fake review production. As we do not investigate individual behavior, we also do not control for such fake reviews. However, as noted by Danescu-Niculescu-Mizil et al. [12], fake reviews often feature very similar text. In our dataset and using usual text similarity thresholds of 70% [12, 32] and the Levenshtein distance [38] to match review text (excluding reviews of same game releases in multiple platforms), we estimate a mere 0.1% of all expert reviews and 0.2% of all

amateur reviews are near matches. Both values are well below the usual thresholds of 14% to 16% other studies [46, 53] reported. Therefore, this facet of fake reviews is negligible in our dataset.

Finally, we remark our work is a large-scale study of the video game portion of the website Metacritic. Thus, there is an opportunity to extend our work to other platforms for experts and amateurs, such as MobyGames[13] or even Steam, which recently introduced curators, an expert type of user role. However, Metacritic is one of the largest websites of its kind, as it aggregates reviews from well-known video game experts (which other websites such as Steam do not), and it attracts activity from many amateurs (in particular more than the comparable website MobyGames).

6 CONCLUSION

Summary. In this work, we presented the first large-scale study of video game reviewing behavior by experts and amateurs. We uncovered substantial differences (and some similarities) in short- and long-term rating and reviewing by experts and amateurs. Amateur reviews and ratings are highly polarized as they often rate video games as zero or ten, whereas expert behavior is more balanced, both in rating and over time. The polarization of amateur reviews is corroborated by amateurs often using emotionally charged vocabulary that exhibits stronger sentiment compared to experts. We leveraged our empirical observations to design a series of prediction experiments that demonstrate the feasibility of predicting video game reputation among amateurs in the short- and long-term and that test the predictive qualities of expert reviews.

Implications. Our work thus highlights challenges in bridging subjective views from experts (the few) vs. amateurs (the crowd) and how to overcome such challenges, as we exemplify in the context of video game appraisal. Knowing which mechanisms experts and amateurs utilize to signal (dis)approval and (dis)agreement facilitates the improvement of review aggregators and recommender systems in online marketplaces. We believe combining such signals and capitalizing on the strengths of both the few and the crowd, in particular the former's experience and the latter's emotions, results in the most comprehensive view of experience goods in general.

Future Work. Extending our textual analyses to the temporal dimension [23] would complement our understanding of rating trends. It would also be fruitful to assess, beyond studies of reviewer experience, if there are influential reviewers whose ratings affect others'. We also believe investigating why there appears to be more consensus among experts than among amateurs, ideally via surveys of their motivations, is promising future work. Finally, our approach can be applied to other domains (e.g. book or movie reviews), as our methods are generic and can be easily extended.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback on the manuscript. Tiago Santos is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

REFERENCES

[1] Xavier Amatriain, Neal Lathia, Josep Pujol, Haewoon Kwak, and Nuria Oliver. 2009. The Wisdom of the Few: A Collaborative Filtering Approach Based on Expert Opinions from the Web. In *SIGIR*.
[2] Martin Atzmueller. 2015. Subgroup discovery. *WIREs Data Mining and Knowledge Discovery* (2015).
[3] Douglas Biber. 2010. *Longman Grammar of spoken and written English*. Longman.
[4] David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *JMLR* (2003).
[5] Keith Burghardt, Emanuel Alsina, Michelle Girvan, William Rand, and Kristina Lerman. 2017. The myopia of crowds: Cognitive load and collective evaluation of answers on Stack Exchange. *PLoS ONE* (2017).
[6] Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies* (2014).

[13] mobygames.com

[7] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan Kankanhalli. 2018. A^3 NCF: An Adaptive Aspect Attention Model for Rating Prediction. In *IJCAI*.
[8] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-Aware Latent Factor Model: Rating Prediction with Ratings and Reviews. In *WWW*.
[9] Judith Chevalier and Dina Mayzlin. 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* (2006).
[10] Georgios Christou. 2013. A comparison between experienced and inexperienced video game players' perceptions. *Human-centric Computing and Information Sciences* (2013).
[11] Joe Cox. 2014. What makes a blockbuster video game? An empirical analysis of US sales data. *Managerial and Decision Economics* (2014).
[12] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How Opinions are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes. In *WWW*.
[13] Fred Davis. 1979. *Yearning for Yesterday: A Sociology of Nostalgia*. Free Press.
[14] Ilona De Jong and Christian Burgers. 2013. Do consumer critics write differently from professional critics? A genre analysis of online film reviews. *Discourse, Context & Media* (2013).
[15] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander Smola, Jing Jiang, and Chong Wang. 2014. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). In *KDD*.
[16] Loretti Dobrescu, Michael Luca, and Alberto Motta. 2013. What makes a critic tick? Connected authors and the determinants of book reviews. *Journal of Economic Behavior & Organization* (2013).
[17] Anders Drachen, Kevin Bauer, and Robert Veitch. 2011. Only the Good... Get Pirated: Game Piracy Activity vs. Metacritic Score. In *FDG*.
[18] Wenjing Duan, Bin Gu, and Andrew Whinston. 2008. Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* (2008).
[19] Francis Galton. 1907. Vox Populi. *Nature* (1907).
[20] Eric Gilbert and Karrie Karahalios. 2010. Understanding Deja Reviewers. In *CSCW*.
[21] Adams Greenwood-Ericksen, Scott Poorman, and Roy Papp. 2013. On the Validity of Metacritic in Assessing Game Value. *Eludamos. Journal for Computer Game Culture* (2013).
[22] The Stanford Natural Language Processing Group. 2010. Stanford Topic Modeling Toolbox. https://nlp.stanford.edu/software/tmt/tmt-0.4/. The Stanford Natural Language Processing Group (December 18, 2018).
[23] Will Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*.
[24] Thorsten Hennig-Thurau, André Marchand, and Barbara Hiller. 2012. The relationship between reviewer judgments and motion picture success: re-analysis and extension. *Journal of Cultural Economics* (2012).
[25] Morris Holbrook and Robert Schindler. 1991. Echoes of the Dear Departed Past: Some Work in Progress on Nostalgia. *ACR North American Advances* (1991).
[26] Nan Hu, Paul Pavlou, and Jennifer Zhang. 2006. Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication. In *EC*.
[27] Clayton Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *ICWSM*.
[28] Daniel Johnson, Christopher Watling, John Gardner, and Lennart Nacke. 2014. The Edge of Glory: The Relationship between Metacritic Scores and Player Experience. In *CHI PLAY*.
[29] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah Smith. 2010. Movie Reviews and Revenues: An Experiment in Text Regression. In *NAACL HLT*.
[30] Dan Jurafsky, Victor Chahuneau, Bryan Routledge, and Noah Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* (2014).
[31] Andrew Kehoe and Matt Gee. 2017. "Please please don't buy this game like I did. I feel terrible and wish I could return it!": A corpus-based study of professional and consumer reviews of video games. In *CL*.
[32] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *EMNLP*.
[33] Timothy King. 2007. Does film criticism affect box office earnings? Evidence from movies released in the US in 2003. *Journal of Cultural Economics* (2007).
[34] George Klare. 1976. A Second Look at the Validity of Readability Formulas. *Journal of Reading Behavior* (1976).
[35] Gael Lederrey and Robert West. 2018. When Sheep Shop: Measuring Herding Effects in Product Ratings with Natural Experiments. In *WWW*.
[36] Young-Jin Lee, Kartik Hosanagar, and Yong Tan. 2015. Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings. *Management Science* (2015).

[37] Xiaojiang Lei, Xueming Qian, and Guoshuai Zhao. 2016. Rating Prediction based on Social Sentiment from Textual Reviews. *IEEE Transactions on Multimedia* (2016).
[38] Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* (1966).
[39] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Lauw. 2010. Detecting Product Review Spammers Using Rating Behaviors. In *CIKM*.
[40] Dayi Lin, Cor-Paul Bezemer, Ying Zou, and Ahmed Hassan. 2019. An empirical study of game reviews on the Steam platform. *Empirical Software Engineering* (2019).
[41] Guang Ling, Mike Lyu, and Irwin King. 2014. Ratings Meet Reviews, a Combined Approach to Recommend. In *RecSys*.
[42] Ian Livingston, Lennart Nacke, and Regan Mandryk. 2011. The Impact of Negative Game Reviews and User Comments on Player Experience. In *ACM SIGGRAPH Game Papers*.
[43] Ian Livingston, Lennart Nacke, and Regan Mandryk. 2011. Influencing Experience: The Effects of Reading Game Reviews on Player Experience. In *International Conference on Entertainment Computing*.
[44] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* (2011).
[45] Michael Luca. 2016. Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School NOM Unit Working Paper* (2016).
[46] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* (2016).
[47] Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *ACL System Demonstrations*.
[48] Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *RecSys*.
[49] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning Attitudes and Attributes from Multi-Aspect Reviews. In *ICDM*.
[50] Metacritic. December 18, 2018. How We Create the Metascore Magic. https://www.metacritic.com/about-metascores.
[51] Loizos Michael and Jana Otterbacher. 2014. Write Like I Write: Herding in the Language of Online Reviews. In *ICWSM*.
[52] Arjun Mukherjee, Bing Liu, and Nat Glance. 2012. Spotting Fake Reviewer Groups in Consumer Reviews. In *WWW*.
[53] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Nat Glance. 2013. What Yelp Fake Review Filter Might Be Doing?. In *ICWSM*.
[54] Phillip Nelson. 1970. Information and consumer behavior. *Journal of Political Economy* (1970).
[55] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey Hancock. 2011. Finding Deceptive Opinion Spam by any Stretch of the Imagination. In *ACL HLT*.
[56] João Palotti, Guido Zuccon, and Allan Hanbury. 2015. The Influence of Pre-Processing on the Estimation of Readability of Web Documents. In *CIKM*.
[57] Hyemin Park and Haewon Byun. 2016. Correlation Analysis: Game Professional Score and User Score on Steam. *International Journal of Multimedia and Ubiquitous Engineering* (2016).
[58] Fabian Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR* (2011).
[59] Jonathan Plucker, James Kaufman, Jason Temple, and Meihua Qian. 2009. Do Experts and Novices Evaluate Movies the Same Way? *Psychology & Marketing* (2009).
[60] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In *ACL*.
[61] David Reinstein and Christopher Snyder. 2005. The Influence of Expert Reviews on Consumer Demand for Experience Goods: A Case Study of Movie Critics. *The Journal of Industrial Economics* (2005).
[62] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social Collaborative Viewpoint Regression with Explainable Recommendations. In *WSDM*.
[63] Robert Schindler and Morris Holbrook. 2003. Nostalgia for early experience as a determinant of consumer preferences. *Psychology & Marketing* (2003).
[64] Feng Shi, Misha Teplitskiy, Eamon Duede, and James Evans. 2019. The wisdom of polarized crowds. *Nature Human Behaviour* (2019).
[65] Elizabeth Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* (2010).
[66] James Surowiecki. 2005. The Wisdom of Crowds. *Anchor Books* (2005).
[67] Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. 2015. User Modeling with Neural Network for Review Rating Prediction. In *IJCAI*.
[68] Luke Thominet. 2016. 10/10 Would Review Again: Variation in the Player Game Review Genre. *Technical Communication Quarterly* (2016).

[69] Christian Wagner, Sesia Zhao, Chris Schneider, and Huaping Chen. 2010. The Wisdom of Reluctant Crowds. In *HICSS*.
[70] Feng Wang, Xuefeng Liu, and Eric Fang. 2015. User Reviews Variance, Critic Reviews Variance, and Product Sales: An Exploration of Customer Breadth and Depth Effects. *Journal of Retailing* (2015).
[71] Tim Wildschut, Constantine Sedikides, Jamie Arndt, and Clay Routledge. 2006. Nostalgia: Content, Triggers, Functions. *Journal of Personality and Social Psychology* (2006).
[72] Jason Shuo Zhang, Chenhao Tan, and Qin Lv. 2018. "This is why we play": Characterizing Online Fan Communities of the NBA Teams. In *CSCW*.
[73] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis. In *SIGIR*.
[74] Feng Zhu and Xiaoquan Zhang. 2006. The Influence of Online Consumer Reviews on the Demand for Experience Goods: The Case of Video Games. In *ICIS*.
[75] Feng Zhu and Xiaoquan Zhang. 2010. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing* (2010).

A APPENDIX

A.1 Pre-Processing for LDA

Before running LDA on expert and amateur reviews, we (i) concatenate all expert (and separately amateur) length-controlled reviews per game (with a minimum of 3 reviews), (ii) tokenize the review text, (iii) convert it to lower-case, (iv) remove common stop words, (v) apply the Porter stemmer, and (vi) remove the 30 most common and very rare words (i.e. those occurring in less than 10 reviews), as well as video game-specific words such as Nintendo, Zelda and others (to learn about reviewing practices rather than specific video games). We sanity-check our pre-processing routine for LDA by calculating perplexity scores for numbers of topics ranging from 5 to 50. As expected, perplexity scores decline with more topics: They decrease from 1,960 at 5 topics to 1,722 at 15 topics, and stabilize between 1,722 and 1,640 for numbers of topics between 15 and 50. We also inspect the word distribution of LDA topic models trained on 20 and 25 topics, and we conclude the core topics we present are robust to increasing the number of topics, as they remain even in LDA models with more topics. Manually inspecting LDA models trained with between 5 and 10 topics, we find such models amalgamate many finer concepts. Hence, we set the number of topics at 15.

A.2 Remaining LDA Topics

We list the bottom five LDA topics with a label we manually annotate, their top 10 words, and average weights, for expert and amateur reviews in Table 4.

Table 4. LDA topics eleven through fifteen.

Experts			Amateurs		
Topic	**Top Words**	**Weight**	**Topic**	**Top Words**	**Weight**
Fighting Games	fight, definite, console, fighter, port, graphics, edit, visual, evil, character	0.0563	Multiplayer Games	multiplayer, campaign, map, single, mode, duty, online, shooter, battlefield, buy	0.0453
Adventure Games	rpg, combat, character, fantasy, quest, dragon, battle, final, warrior, dungeon	0.0552	Free2Play Games	dlc, free, pay, community, content, update, friend, money, team, map	0.0386
Racing Games	race, racer, drive, car, speed, track, arcade, graphic, need, simulation	0.0473	Horror Games	horror, evil, zombie, survival, resident, atmosphere, scary, dead, fan, scare	0.0346
Mobile Games	city, super, mobile, level, platform, mission, app, run, world, iphone	0.0430	Open World Games	amazing, world, uncharted, batman, mission, city, open, awesome, gear, everything	0.0294
Music Games	hero, music, rock, song, guitar, band, dance, rhythm, batman, park	0.0242	Online Games	server, online, city, mmo, drm, expansion, content, single, buy, money	0.0263

Received April 2019; revised June 2019; accepted August 2019

4 Conclusions

Peer production systems, as omnipresent venues for collaboration at large on the web, crucially depend on contributions by their users to grow and thrive. Timely identification of slowing growth or acceleration of decline in activity may allow to react to such dynamics and design interventions to counter them. There are, however, remarkable differences in user behavior and unobservable user properties and interactions, which combine to form complex activity dynamics at the system-level activity dynamics of peer production systems. Those characteristics of peer production systems pose a significant challenge for practitioners as well as researchers aiming to understand how and why some systems thrive, i.e. manage to attract sustained levels of activity from large numbers of users, while others do not, and shut down as a consequence. Therefore, identifying and studying factors which relate user patterns to system-level dynamics and which differentiate thriving from declining peer production systems represents an important first step towards supporting successful activity dynamics. Several articles in this thesis proposed a series of tools to longitudinally study a few of those factors, most prominently nonlinear dynamics, user types and user excitation. Beyond user excitation in peer production systems, this thesis also advanced current knowledge on using Hawkes processes, which may be applied beyond the context of excitation in peer production systems. This work also derives actionable insights for practitioners to optimize the systems they oversee, such as, for example, what managers may expect when applying an intervention (e.g. introducing a badge) to steer activity dynamics, or how the crowds in their systems perform in comparison to the few.

In the following, I summarize this thesis' results and contributions in Section 4.1 and implications in Section 4.2. Then, I discuss the limitations of this thesis in Section 4.3, and I present future work in Section 4.4.

4.1 Results and Contributions

In this Section, I subsume the answers to the research questions formulated in Section 1.4.

RQ1: How can we characterize user activity dynamics in peer production systems?

Previous research postulated a dynamical system description for activity in peer production systems, which does not distinguish between nonlinear chaotic and stochastic activity dynamics. The article [SWH17] addresses this research gap by characterizing nonlinear dynamics in the activity of 16 Stack Exchange Q&A communities. I find that

there are noteworthy differences in activity dynamics, which have practical impact as knowing which communities feature more nonlinear dynamics helped improve the accuracy of activity forecasting experiments.

Shifting the focus to the role played by different user types in peer production systems, I present a case study [San+19a] of 50 Stack Exchange Q&A communities where four user types with different levels of activity and burstiness dominate. This analysis revealed that varying proportions of the prevalence of each user type relate to the overall developmental stage of Q&A communities, a finding which gives system managers a target user-base structure to aim for at different points in time.

RQ2: How can we model users' evolving excitation in peer production systems?

I answered this research question in [San+19c; SLH20] by tailoring Hawkes process-based methods to capture user excitation in Stack Exchange Q&A communities and beyond. In particular, in [San+19c] I propose a first solution to the problem of fitting Hawkes processes to non-stationary activity dynamics by leveraging time series structural change models to identify sub-periods of stationarity. In that same article, to solve the issue of estimating the decay parameter of the commonly used Hawkes processes with exponential kernels, I resorted to a Bayesian hyperparameter optimization routine. However, in [SLH20], I showed that that optimization method is only one of many which cannot improve beyond a certain unavoidable estimation error (due to noisy, non-convex Hawkes log-likelihood as a function of the decay). To alleviate this problem with point estimates, I proposed a Bayesian framework to estimate uncertainty in fitted decay parameter values, and I illustrated how to apply this framework to fit Hawkes processes even in the presence of temporal phenomena with stationarity-breaking exogenous shocks.

RQ3: How do users' excitation and temporal patterns shape the evolution of peer production systems?

The previously mentioned study [San+19c] provided an answer to this research question in the form of excitation effects, which ascribed an important role to, e.g., power user (and respectively casual user) excitation in the early (and respectively late) developmental stages of Stack Exchange Q&A communities which grow the most. I also observed remarkable differences in the roles played by power and casual users depending on the community topic, and a prediction experiment indicated that these excitation effects bear predictive qualities with respect to the future development of the activity dynamics of Q&A communities. Taken together, these results underscore the importance of timing the user mix correctly, promoting certain kinds of user interactions and respecting community topical differences, in order to ensure growth-inducing user structures in peer production systems.

Beyond user excitations, my study [San+20] of the effect of a badge-based intervention to improve welcoming culture in Stack Exchange Q&A communities revealed that this indicator did not curb a long-term decline in newcomer retention and, content-wise, it

only enacted a temporary effect on the reactions of veterans to newcomers. Therefore, while individual users' activity dynamics shape system-level dynamics, community managers may need to leverage more far-reaching strategies to more effectively steer their communities and enact long(er)-term changes.

Finally, with a comparative empirical research [San+19d] of the wisdom of the crowds juxtaposed with that of the few on peer production system for video game reviews, this thesis contributes a first step towards understanding conditions under which, in peer production systems, the wisdom of the crowds may reach and perhaps even surpasses that of the few. This study revealed significant discrepancies in the temporal reviewing dynamics of the crowds and the few, which indicates significant effort by peer production system managers may be necessary to reconcile their different views and opinion signaling mechanisms. However, doing so bears the possibility to fully tap the potential knowledge of a peer production system, as a prediction experiment for video game reception among the crowds suggested opinions of both are required to obtain best predictive performance.

4.2 Implications of this Work

The ramifications of this thesis' findings are of consequence especially for practitioners. I also discuss implications of theoretical and methodological nature.

Actionable insights. This thesis distilled a number of empirical findings regarding user mix, user excitation and timing effects, and I strongly believe that these findings may improve the understanding and shaping of activity dynamics on peer production systems. For example, measuring excitation at large-scale poses a cost-effective alternative to survey or other kinds of qualitative research to understand users' intrinsic motivation and unobserved interactions with content and users. Saving resources there may mean they may be allocated elsewhere, and leveraging fine-grained excitation estimations may improve prediction models for both user-level and system-level activity dynamics. Further, early warning systems may be devised to capitalize on such improved prediction models, to maximize reaction time for system managers to devise interventions (which this thesis also studies and whose potential and limitations this thesis also elucidates).

In the context of our assessment of the wisdom of the crowds compared with the wisdom of the few, it is also conceivable that downstream applications leveraging peer production systems based on matters of opinion may adjust according to our nuanced results, and consider mixing crowdsourced knowledge with expert inputs to obtain more holistic views on, e.g., an experience good to integrate in a recommender system.

Future models for activity dynamics, user excitation and beyond. First, it is my hope that the illustration of pitfalls and how to overcome them in the application of Hawkes processes to model and predict activity dynamics (cf. [San+19c; SLH20]) encourages future work to employ these tools too. To support this endeavor and for replication purposes, I share code along with the article [San+19c]. Further, in [SLH20], I hinted at the generalizability of the tools proposed by this thesis to model temporal phenomena in contexts beyond activity dynamics in peer production systems, as improving the applicability of Hawkes processes benefits the study of all kinds of temporal dynamics where

they may be applied (e.g., seismology or finance).

A concern for surfacing implicit assumptions and for parsimony (not unlike *Occam's razor*) guided my design of tools and methods for the research which forms this thesis. I believe future methodological endeavors in the space of activity dynamics models may benefit from my research's implementation of those guiding principles: For example, studying the assumptions behind dynamical systems descriptions of activity dynamics lead to uncovering nonlinear properties of activity in Q&A communities and significant improvements in forecasting performance. As another example, while employing a Bayesian inference setup for fitting all Hawkes process parameters at once would be complex but allow to stay within the realm of Bayesian inference, using the Bayesian framework for only the decay parameter (as done in [SLH20]) enables a simple and efficient closed-form inference of the decay, while leaving the optimization of the rest of the parameters to well-established convex optimization routines. Overall, I hope some of these guiding principles in the body of work of this thesis may serve as inspiration for future efforts to (parsimoniously) model and predict activity in peer production systems.

4.3 Limitations

First, note that this thesis established a temporal link between given excitation effects and activity outcomes of Stack Exchange Q&A communities. Although these effects are predictive of future activity outcomes, and despite the connection between Hawkes processes with exponential kernels and Granger causality, this work does not claim causality in the stricter (e.g. Pearlian) sense. To do so, a first step could be to incorporate propensity score matching on features describing users and communities to control for observed confounders, or, ideally, to devise some kind of natural experiment or quasi-experimental setting to eliminate unobserved confounders too. A similar remark is also valid for both the analysis of the introduction of the badge-like indicator to improve the welcoming culture of Stack Exchange Q&A communities [San+20] and for the empirical comparison of the wisdom of the crowds vs. the few with respect to video game reviews [San+19d].

While my decay estimation research [SLH20] employed a diverse set of datasets, much of the research in this thesis focused on the Stack Exchange Q&A peer production systems. As such, there is an opportunity to extend this research to other peer production systems in the knowledge community space (e.g. Wikipedia) or in the more informal community space (e.g. Reddit). I believe the approaches presented and applied in the context of the Stack Exchange communities may be generalized to those other systems as well, especially since Hawkes processes only require knowledge of timestamps of activities, a feature that is typically available across datasets of peer production systems. Doing this generalization exercise would help to compare peer production system dynamics and to minimize the role that idiosyncrasies of the Stack Exchange platform may play.

One third aspect to note is also the potential cost of parsimony e.g. in the user type categorization: Although the user types aligned well with activity dynamics, clustering users solely basing on their temporal signatures of activity bears the risk of overseeing nuanced kinds of behavior which warrant a closer look. I see an opportunity to design

future research in this space, and I address this aspect of future work in the following section.

4.4 Future Work

I believe that inspecting other definitions of peer production system success beyond activity (cf. e.g. Cunha et al.'s [Cun+19] definitions of success as community survival), and relating those definitions back to user excitation is interesting future work. Analogously, user roles may also be multi-faceted. Thus, extending the activity-based characterization to one inspired by social roles, much like Yang et al. [Yan+19a] propose, is another promising avenue for future work.

Further, future work might also aim at contextualizing peer production systems and users in wider circles: Extending this work beyond Stack Exchange, one might adjust community similarity models like Chandrasekharan et al.'s [Cha+17] to learn about peer production systems with comparable activity dynamics across the web ecosystem. Coming back again to the user context, previous work [TL15; HXL19] indicates users behave in the online world much like in the offline world and also exhibit multi-community engagement. Therefore, studying user excitations and activity dynamics across systems may yield deeper insights into user excitation properties.

In general, this thesis focused on temporal and longitudinal rather than cross-sectional analyses. In doing so, and especially in the case of temporal patterns of video game reviewing behavior, this thesis uncovered interesting long-term effects such as leniency towards older games and nostalgia, the effect of yearning for yesterday. This initial study of such effects warrants a deeper study of which games (or, more generally, objects of interest to users collaborating in peer production systems) attain this long-term interest property. In other words, identifying, characterizing and predicting which such games or more generally objects of interest pass "the test of time" poses an exciting prospect for future work.

Overall, it is my hope this thesis serves as a stepping stone towards further analyses of activity dynamics in peer production systems, and that this work may help advance the future state-of-the-art in understanding and modeling such temporal phenomena.

Bibliography

[ABG08] O. Aalen, O. Borgan, and H. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media, 2008.

[All06] D. Allen. "Do Organizational Socialization Tactics Influence Newcomer Embeddedness and Turnover?" In: *Journal of Management* (2006).

[And+12] A. Anderson et al. "Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow". In: *KDD*. 2012.

[And+13] A. Anderson et al. "Steering User Behavior with Badges". In: *WWW*. 2013.

[ASCDL15] Y. Aït-Sahalia, J. Cacho-Diaz, and R.J.A. Laeven. "Modeling Financial Contagion Using Mutually Exciting Jump Processes". In: *Journal of Financial Economics* (2015).

[Bac+16] E. Bacry et al. "Mean-field inference of Hawkes point processes". In: *Journal of Physics A: Mathematical and Theoretical* (2016).

[Bac+18] E. Bacry et al. "Tick: a Python Library for Statistical Learning, with an Emphasis on Hawkes Processes and Time-Dependent Models". In: *The Journal of Machine Learning Research* (2018).

[Bar05] A.-L. Barabasi. "The origin of bursts and heavy tails in human dynamics". In: *Nature* (2005).

[BM16] E. Bacry and J.-F. Muzy. "First-and second-order statistics characterization of Hawkes processes and non-parametric estimation". In: *IEEE Transactions on Information Theory* (2016).

[BN06] Y. Benkler and H. Nissenbaum. "Commons-based Peer Production and Virtue". In: *Journal of Political Philosophy* (2006).

[Bur+17] K. Burghardt et al. "The myopia of crowds: Cognitive load and collective evaluation of answers on Stack Exchange". In: *PLoS ONE* (2017).

[Cao+17] Q. Cao et al. "Deephawkes: Bridging the Gap Between Prediction and Understanding of Information Cascades". In: *CIKM*. 2017.

[CFL09] C. Castellano, S. Fortunato, and V. Loreto. "Statistical Physics of Social Dynamics". In: *Reviews of modern physics* (2009).

[Cha+17] E. Chandrasekharan et al. "The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data". In: *CHI*. 2017.

[Che+14] H. Chen et al. "Wisdom of crowds: The value of stock opinions transmitted through social media". In: *The Review of Financial Studies* (2014).

[Cho+15] E. Choi et al. “Constructing Disease Network and Temporal Progression Model via Context-Sensitive Hawkes Process”. In: *ICDM*. 2015.

[Cho+18] J. Choudhari et al. “Discovering Topical Interactions in Text-based Cascades Using Hidden Markov Hawkes Processes”. In: *ICDM*. 2018.

[CI80] D.R. Cox and V. Isham. *Point processes*. CRC Press, 1980.

[CPU12] K.S.K. Chung, M. Piraveenan, and S. Uddin. “Community Evolution and Engagement Through Assortative Mixing in Online Social Networks”. In: *ASONAM*. 2012.

[Cun+19] T. Cunha et al. “Are All Successful Communities Alike? Characterizing and Predicting the Success of Online Communities”. In: *WWW*. 2019.

[Dab+12] L. Dabbish et al. “Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository”. In: *CSCW*. 2012.

[DFZ14] J. Da Fonseca and R. Zaatour. “Hawkes Process: Fast Calibration, Application to Trade Clustering, and Diffusive Limit”. In: *Journal of Futures Markets* (2014).

[DKS19] H. Dev, K. Karahalios, and H. Sundaram. “Quantifying Voter Biases in Online Platforms: An Instrumental Variable Approach”. In: *PACM HCI (CSCW)* (2019).

[Du+16] N. Du et al. “Recurrent Marked Temporal Point Processes: Embedding Event History to Vector”. In: *KDD*. 2016.

[Du17] N. Du. *PtPack: The C++ Multivariate Temporal Point Process Package*. `https://github.com/dunan/MultiVariatePointProcess`. 2017.

[DVJ03] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer Science & Business Media, 2003.

[DVJ08] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media, 2008.

[EDD17] M. Eichler, R. Dahlhaus, and J. Dueck. “Graphical Modeling for Multivariate Hawkes Processes with Nonparametric Link Functions”. In: *Journal of Time Series Analysis* (2017).

[Ete+16] J. Etesami et al. “Learning Network of Multivariate Hawkes Processes: A Time Series Approach”. In: *UAI*. 2016.

[Far+14] M. Farajtabar et al. “Shaping Social Activity by Incentivizing Users”. In: *NIPS*. 2014.

[Far+15] M. Farajtabar et al. “COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution”. In: *NeurIPS*. 2015.

[Fig+18] F. Figueiredo et al. “Fast Estimation of Causal Interactions Using Wold Processes”. In: *NeurIPS*. 2018.

[Fur+13] A. Furtado et al. “Contributor Profiles, Their Dynamics, and Their Importance in Five Q&A Sites”. In: *CSCW*. 2013.

[Gal07] F. Galton. “Vox Populi”. In: *Nature* (1907).

[Gei+19] C. Geigle et al. “A Generative Model for Discovering Action-based Roles and Community Role Compositions on Community Question Answering Platforms”. In: *ICWSM*. 2019.

[Gil05] J. Giles. *Internet Encyclopaedias go head to head*. 2005.

[Gil13] E. Gilbert. “Widespread Underprovision on Reddit”. In: *CSCW*. 2013.

[GRLK10] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. “Inferring Networks of Diffusion and Influence”. In: *KDD*. 2010.

[Haw71] A.G. Hawkes. “Spectra of some self-exciting and mutually exciting point processes”. In: *Biometrika* (1971).

[HF20] T. Hatt and S. Feuerriegel. “Early Detection of User Exits from Clickstream Data: A Markov Modulated Marked Point Process Model”. In: *WWW*. 2020.

[HG14] C. Hutto and E. Gilbert. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”. In: *ICWSM*. 2014.

[Hop+20] N. Hopfgartner et al. “Social Facilitation Among Gamblers: A Large-Scale Study Using Account-Based Data”. In: *Submitted to ICWSM*. 2020.

[HXL19] T. Hu, Y. Xia, and J. Luo. “To Return or to Explore: Modelling Human Mobility and Dynamics in Cyberspace”. In: *WWW*. 2019.

[IL09] A. Iriberri and G. Leroy. “A Life-Cycle Perspective on Online Community Success”. In: *CSUR* (2009).

[Jun+19] R. Junuthula et al. “The Block Point Process Model for Continuous-Time Event-Based Dynamic Networks”. In: *WWW*. 2019.

[KAL18] T. Kurashima, T. Althoff, and J. Leskovec. “Modeling Interdependent and Periodic Real-World Action Sequences”. In: *WWW*. 2018.

[Kas+19a] P. Kasper et al. “Modeling User Dynamics in Collaboration Websites”. In: *Book chapter in Dynamics On and Of Complex Networks* (2019).

[Kas+19b] P. Kasper et al. “On the Role of Score, Genre and Text in Helpfulness of Video Game Reviews on Metacritic”. In: *SNAMS*. 2019.

[KGR18] T. Kusmierczyk and M. Gomez-Rodriguez. “On the Causal Effect of Badges”. In: *WWW*. 2018.

[Kit+07] A. Kittur et al. “Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie”. In: *WWW*. 2007.

[Kle03] J. Kleinberg. “Bursty and Hierarchical Structure in Streams”. In: *Data Mining and Knowledge Discovery* (2003).

[Kou+20] G. Koutroulis et al. “Enhanced Active Learning of Convolutional Neural Networks: A Case Study for Defect Classification in the Semiconductor Industry”. In: *Submitted to Discovery Science*. 2020.

[KR12] R. Kraut and P. Resnick. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press, 2012.

[KS20] T. Kushner and A. Sharma. “Bursts of Activity: Temporal Patterns of Help-Seeking and Support in Online Mental Health Forums”. In: *WWW*. 2020.

[Lih04] A. Lih. “Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource”. In: *International Symposium on Online Journalism*. 2004.

[Lin09] T.J. Liniger. “Multivariate Hawkes Processes”. PhD thesis. ETH Zürich, 2009.

[Lin+17] Z. Lin et al. “Better When It Was Smaller? Community Content and Behavior After Massive Growth”. In: *ICWSM*. 2017.

[LM11] E. Lewis and G. Mohler. “A nonparametric EM Algorithm for Multiscale Hawkes Processes”. In: *Journal of Nonparametric Statistics* (2011).

[LWK18] T. Li, P. Wei, and Y. Ke. “Transfer Hawkes Processes with Content Information”. In: *ICDM*. 2018.

[Mal+20] D. Maldeniya et al. “Herding a Deluge of Good Samaritans: How GitHub Projects Respond to Increased Attention”. In: *WWW*. 2020.

[Mam+11] L. Mamykina et al. “Design Lessons from the Fastest Q&A Site in the West”. In: *CHI*. 2011.

[MSF15] Y. Matsubara, Y. Sakurai, and C. Faloutsos. “The Web as a Jungle: Non-Linear Dynamical Systems for Co-evolving Online Activities”. In: *WWW*. 2015.

[Mur+19] G. Murić et al. “Collaboration Drives Individual Productivity”. In: *PACM HCI (CSCW)* (2019).

[MVGR17] C. Mavroforakis, I. Valera, and M. Gomez-Rodriguez. “Modeling the Dynamics of Online Learning Activity”. In: *WWW*. 2017.

[OAK82] Y. Ogata, H. Akaike, and K. Katsura. “The application of linear intensity models to the investigation of causal relations between a point process and another stochastic process”. In: *Annals of the Institute of Statistical Mathematics* (1982).

[Oga83] Y. Ogata. “Likelihood analysis of point processes and its applications to seismological data”. In: *Bulletin of the International Statistical Institute* (1983).

[Oza79] T. Ozaki. “Maximum Likelihood Estimation of Hawkes’ Self-Exciting Point Processes”. In: *Annals of the Institute of Statistical Mathematics* (1979).

[Rib14] B. Ribeiro. “Modeling and Predicting the Growth and Death of Membership-based Websites”. In: *WWW*. 2014.

[Riz+17] M.-A. Rizoiu et al. “Expecting to be HIP: Hawkes Intensity Processes for Social Media Popularity”. In: *WWW*. 2017.

[RMJ10] D.R. Raban, M. Moldovan, and Q. Jones. “An Empirical Study of Critical Mass and Online Community Survival”. In: *CSCW*. 2010.

[RR15] L. Robert and D.M. Romero. “Crowd Size, Diversity and Performance”. In: *CHI*. 2015.

[RSH19] T. Ruprechter, T. Santos, and D. Helic. “On the Relation of Edit Behavior, Link Structure, and Article Quality on Wikipedia”. In: *International Conference on Complex Networks and Their Applications*. 2019.

[Sal+19] F. Salehi et al. “Learning Hawkes Processes from a Handful of Events”. In: *NeurIPS*. 2019.

[San+19a] T. Santos et al. “Activity Archetypes in Question-and-Answer (Q&A) Websites – A Study of 50 Stack Exchange Instances”. In: *ACM TSC* (2019).

[San+19b] T. Santos et al. “Feature Extraction From Analog Wafermaps: A Comparison of Classical Image Processing and a Deep Generative Model”. In: *IEEE TSM* (2019).

[San+19c] T. Santos et al. “Self- and Cross-Excitation in Stack Exchange Question & Answer Communities”. In: *WWW*. 2019.

[San+19d] T. Santos et al. “What’s in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic”. In: *PACM HCI (CSCW)* (2019).

[San+20] T. Santos et al. “Can Badges Foster a More Welcoming Culture on Q&A Boards?” In: *ICWSM*. 2020.

[See+20] J. Seering et al. “Proximate Social Factors in First-Time Contribution to Online Communities”. In: *CHI*. 2020.

[SK18] T. Santos and R. Kern. “Understanding wafer patterns in semiconductor production with variational auto-encoders”. In: *ESANN*. 2018.

[SLH20] T. Santos, F. Lemmerich, and D. Helic. “Estimating the Decay Parameter of Hawkes Processes with Exponential Kernels”. In: *Submitted to ICDM*. 2020.

[SMG13] V.S. Sinha, S. Mani, and M. Gupta. “Exploring Activeness of Users in QA Forums”. In: *MSR*. 2013.

[Suh+09] B. Suh et al. “The Singularity is Not Near: Slowing Growth of Wikipedia”. In: *WikiSym*. 2009.

[Sur05] J. Surowiecki. “The Wisdom of Crowds”. In: *Anchor Books* (2005).

[SW14] J. Solomon and R. Wash. “Critical mass of what? Exploring community growth in WikiProjects”. In: *ICWSM*. 2014.

[SWH17] T. Santos, S. Walk, and D. Helic. “Nonlinear Characterization of Activity Dynamics in Online Collaboration Websites”. In: *WWW Companion*. 2017.

[Tak81] F. Takens. “Detecting strange attractors in turbulence”. In: *Dynamical systems and turbulence*. 1981.

[TL15] C. Tan and L. Lee. “All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement”. In: *WWW*. 2015.

[TSK19] M. Toller, T. Santos, and R. Kern. “SAZED: parameter-free domain-agnostic season length estimation in time series data”. In: *Data Mining and Knowledge Discovery* (2019).

[TWS19] A. C. Türkmen, Y. Wang, and A.J. Smola. “Fastpoint: Scalable Deep Point Processes”. In: *ECML-PKDD*. 2019.

[UVGR17] U. Upadhyay, I. Valera, and M. Gomez-Rodriguez. “Uncovering the Dynamics of Crowdlearning and the Value of Knowledge”. In: *WSDM*. 2017.

[Váz+06] A. Vázquez et al. “Modeling Bursts and Heavy Tails in Human Dynamics”. In: *Physical Review E* (2006).

[VWD04] F.B. Viégas, M. Wattenberg, and K. Dave. “Studying Cooperation and Conflict between Authors with history flow Visualizations”. In: *CHI*. 2004.

[Wal+16] S. Walk et al. “Activity Dynamics in Collaboration Networks”. In: *TWEB* (2016).

[Wan+13] G. Wang et al. “Wisdom in the Social Crowd: An Analysis of Quora”. In: *WWW*. 2013.

[WH07] D.M. Wilkinson and B. Huberman. “Cooperation and Quality in Wikipedia”. In: *WikiSym*. 2007.

[Wil08] D.M. Wilkinson. “Strong Regularities in Online Peer Production”. In: *EC*. 2008.

[WZH13] T. Weninger, X.A. Zhu, and J. Han. “An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community”. In: *ASONAM*. 2013.

[XFZ16] H. Xu, M. Farajtabar, and H. Zha. “Learning Granger Causality for Hawkes Processes”. In: *ICML*. 2016.

[Yan+10] J. Yang et al. “Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities”. In: *ICWSM*. 2010.

[Yan+14] J. Yang et al. “Sparrows and Owls: Characterisation of Expert Behaviour in Stack Overflow”. In: *UMAP*. 2014.

[Yan+19a] D. Yang et al. “Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities”. In: *CHI*. 2019.

[Yan+19b] S. Yanovsky et al. “One Size Does Not Fit All: Badge Behavior in Q&A Sites”. In: *UMAP*. 2019.

[You13] C. Young. “Community Management That Works: How to Build and Sustain a Thriving Online Health Community”. In: *JMIR* (2013).

[ZAA07] J. Zhang, M.S. Ackerman, and L. Adamic. “Expertise Networks in Online Communities: Structure and Algorithms”. In: *WWW*. 2007.

[ZCF16] C. Zang, P. Cui, and C. Faloutsos. “Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications”. In: *KDD*. 2016.

[ZWR20] R. Zhang, C. Walder, and M.-A. Rizoiu. “Variational Inference for Sparse Gaussian Process Modulated Hawkes Process”. In: *AAAI*. 2020.

[ZZS13a] K. Zhou, H. Zha, and L. Song. “Learning Social Infectivity in Sparse Low-Rank Networks Using Multi-Dimensional Hawkes Processes”. In: *AISTATS*. 2013.

[ZZS13b] K. Zhou, H. Zha, and L. Song. “Learning Triggering Kernels for Multi-Dimensional Hawkes Processes”. In: *ICML*. 2013.