

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.



This work is licensed under a [Creative Commons Attribution International 4.0 License](https://creativecommons.org/licenses/by/4.0/).

© 2019 Copyright held by the owner/author(s).

2469-7818/2019/02-ART4

<https://doi.org/10.1145/3301612>

1 INTRODUCTION

Question-and-answer (Q&A) websites (e.g., Stack Exchange¹ or Quora²) are publicly accessible platforms, which are used by millions of users to discuss a variety of topics and problems. For example, the StackOverflow³ 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 Exchange⁴ 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

⁵https://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 Iriberry 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

⁶As of February 2017, the *programmers* Stack Exchange instance is termed *softwareengineering* and [programmers.stackexchange.com](https://stackoverflow.com) redirects to softwareengineering.stackexchange.com.

⁷All 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 time 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/>.

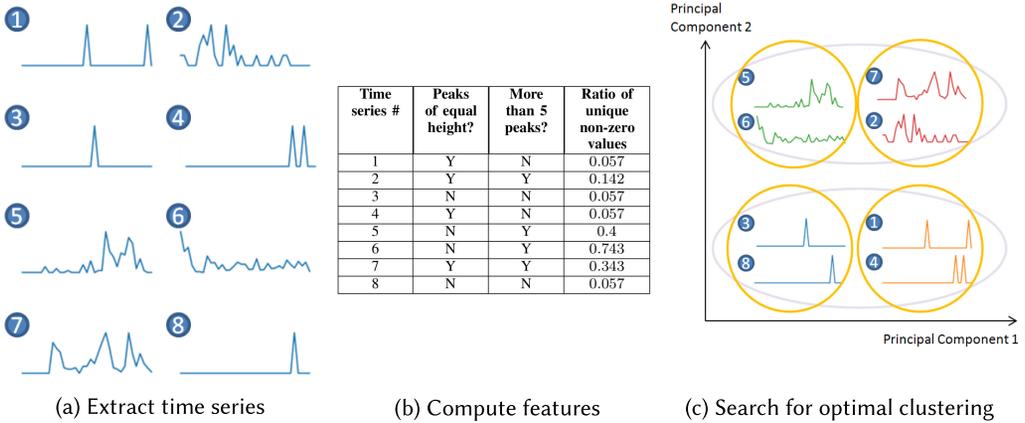


Fig. 1. **Identifying user archetypes as a time series clustering problem.** We start by extracting the questions-based (and, separately, answers-based) activity time series as the monthly sum of posted questions (respectively, answers) of each user in a Stack Exchange instance (cf. Figure 1(a)). We 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 $K = 2, \dots, 10$ 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 $K = 2, \dots, 10$ 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.

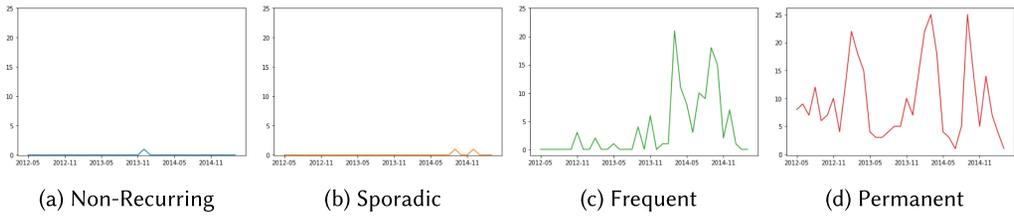


Fig. 2. **Activity Archetypes.** We illustrate typical profiles of the activity-based time series nearest to K-Means centroid for $K^* = 4$. Users of the *Non-Recurring Activity Archetype* (a) often feature one single, isolated peak of activity. Users of the *Sporadic Activity Archetype* (b) typically 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>.

¹⁰In 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].

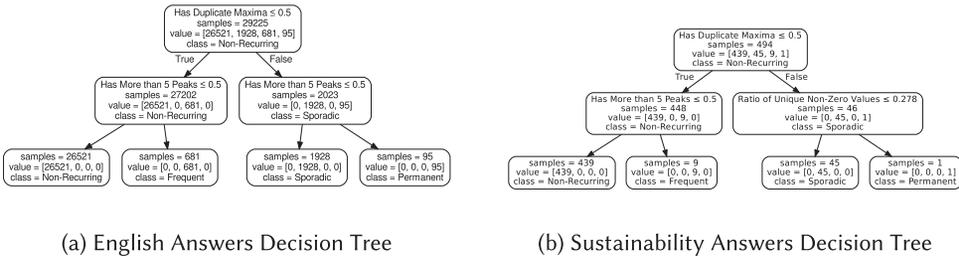
For the ANOVA approach, we fitted a generalized linear model of the three features per user as independent variables and the *Activity Archetypes* resulting from the clustering as dependent variables. As the *Activity Archetypes* represent a discrete dependent variable, we assume it is binomially distributed 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 (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

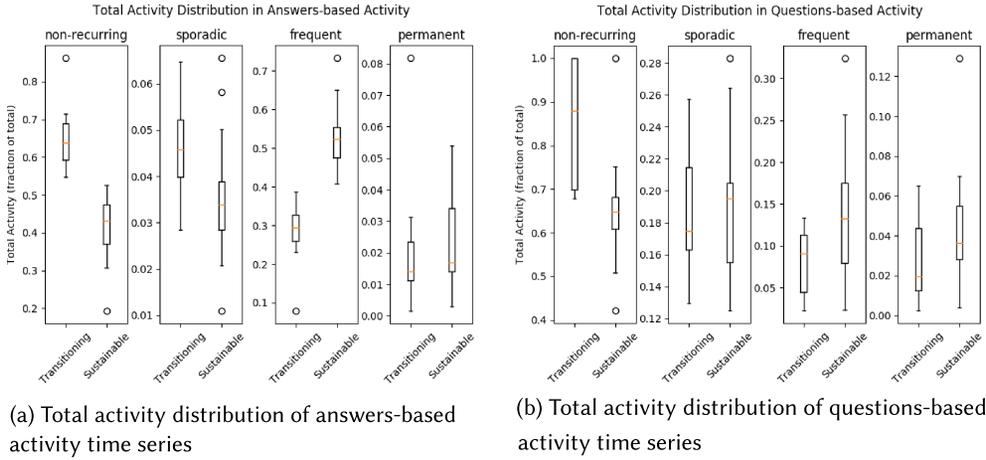


Fig. 4. **Distinction between *Transitioning* and *Sustainable* Stack Exchange instances.** For all Stack Exchange instances of the types *Transitioning* and *Sustainable*, we depict the fractions of total answers-based (Figure 4(a)) and questions-based (Figure 4(b)) activity generated per *Activity 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

¹⁴The 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.

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

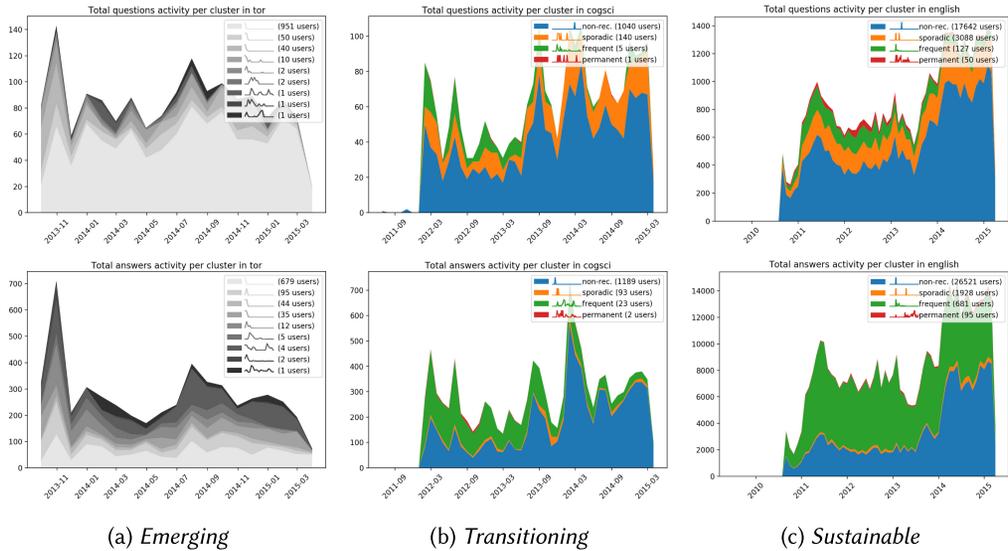


Fig. 5. **Temporal dynamics of the three types of activity composition of Stack Exchange instances: Emerging, Transitioning, and Sustainable.** We plot the total count of questions-based (top) and answers-based (bottom) activity of three Stack Exchange instances over time, and break these activity totals 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 instances¹⁶ 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

¹⁶The *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	<i>Sustainable</i>	37, 125	522, 128	70	0.013
unix	<i>Sustainable</i>	36, 397	390, 930	80	0.012
softwareengineering	<i>Sustainable</i>	35, 816	467, 234	80	0.006
gaming	<i>Sustainable</i>	34, 641	321, 857	68	0.007
tex	<i>Sustainable</i>	31, 039	624, 166	80	0.014
poker	<i>Transitioning</i>	594	5, 185	39	-0.002
earthscience	<i>Emerging</i>	578	5, 981	12	-0.040
sustainability	<i>Transitioning</i>	555	5, 274	27	-0.015
ebooks	<i>Emerging</i>	501	3, 094	16	-0.041
ham	<i>Emerging</i>	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

¹⁷Note 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).

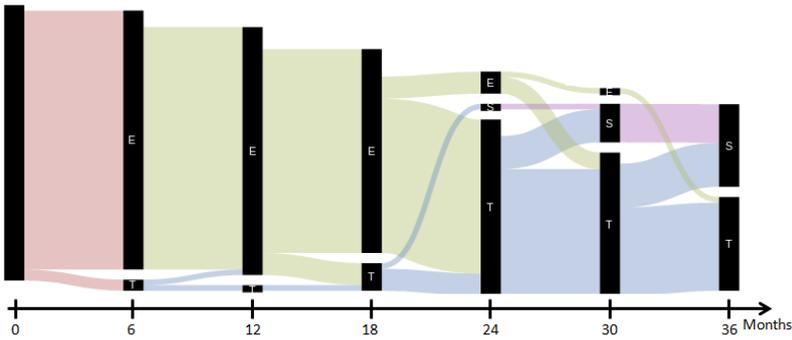


Fig. 6. **Temporal evolution of Stack Exchange instance types.** We count the number of Stack Exchange instances per type (with “E” standing for *Emerging* (in green), “T” for *Transitioning* (in blue), and “S” for *Sustainable* (in pink)) every 6 months until the first 3 years of the instances’ existence. 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 Shirshorshidi, 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 Iriberry 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 Grzmar, 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 Dezső, 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