User Session Identification Based on Strong Regularities in Inter-activity Time

Aaron Halfaker[1] Oliver Keyes[1] Daniel Kluver[2] Jacob Thebault-Spieker[2]
Tien Nguyen[2] Kenneth Shores[2] Anuradha Uduwage[2] Morten Warncke-Wang[2]

[1]Wikimedia Foundation
149 New Montgomery St.
6th Floor
San Francisco, CA 94105
{ahalfaker,okeyes}@wikimedia.org

[2]GroupLens Research
Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455
{kluver,thebault,tien,shores,uduwage,morten}@cs.umn.edu

ABSTRACT

Session identification is a common strategy used to develop metrics for web analytics and perform behavioral analyses of user-facing systems. Past work has argued that session identification strategies based on an inactivity threshold is inherently arbitrary or has advocated that thresholds be set at about 30 minutes. In this work, we demonstrate a strong regularity in the temporal rhythms of user initiated events across several different domains of online activity (incl. video gaming, search, page views and volunteer contributions). We describe a methodology for identifying clusters of user activity and argue that the regularity with which these activity clusters appear implies a good rule-of-thumb inactivity threshold of about 1 hour. We conclude with implications that these temporal rhythms may have for system design based on our observations and theories of goal-directed human activity.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.1.1 [**Coding and Information Theory**]: Formal models of communication

General Terms

Theory, Measurement, Human Factors

Keywords

User session; Activity; Human behavior; Regularities; Metrics; Modeling; Analytics

1. INTRODUCTION

In 2012, we had an idea for a measurement strategy that would bring insight and understanding to the nature of work in an online community. While studying volunteer participation in Wikipedia, the open, collaborative encyclopedia, we found ourselves increasingly curious about the amount of time that volunteer contributors invested into the encyclopedia's construction. Past work measuring Wikipedia editor engagement relied on counting the number of contributions made by a user[1], but we felt that the amount of time editors spent editing might serve as a more appropriate measure.

The measurement strategy we came up with was based on the clustering of Wikipedia editors' activities into "edit sessions" with the assumption that the duration of an edit session would represent a lower bound of the amount of time invested into Wikipedia contributions [9]. Through our ethnographic work in Wikipedia we had found the notion of a work session to be intuitive, yet there did not appear to be a consensus in the literature on how to identify work sessions from timestamped user activities. This led us to look to the data for insight about what might be a reasonable approach to delineating users' editing activity into sessions. The regularities we found in inter-activity time amazed us with their intuitiveness and the simplicity of session demarcation they implied. It is that work that led us to look for such regularities in other systems and to write this paper to share our results.

We are not the first to try our hands at identifying a reasonable way to measure user session behavior in human-computer interaction. User sessions have been used extensively to generate metrics for understanding the performance of information resources [11] – especially in the domain of search [7, 8] and content personalisation [10, 20]. Despite this interest in understanding the nature and manifestation of user sessions, no clear consensus about how to perform session identification has emerged.

Some have have even argued that human behavior is best understood as a series of goal-driven tasks as opposed to activity sessions and that the common strategy of choosing a global inactivity threshold is ineffective at identifying the boundaries of such tasks [12]. We draw from Activity Theory [16] to conceptualize tasks as sub-session activities and argue that both are are important for undertanding goal-oriented human behavior.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741117.

[1]for example, "Wikipedian is first to hit 1 million edits" `http://www.dailydot.com/news/wikipedian-first-1-million-edits`

In this paper, we describe a strategy for identifying user sessions from log data and demonstrate how the results match both intuition and theory about goal-directed human activity. We also show how this strategy yields consistent results across many different types of systems and user activities. First, we summarize previous work which attempts to make sense of user session behavior from log data. Then we discuss theoretical arguments about how goal-directed user behavior ought to manifest in the data. Third, we discuss a generalized version of the inactivity threshold identification strategy we developed in previous work [9] and present strategies for identifying optimal inactivity thresholds in new data. Then, we introduce 6 different systems from which we have extracted 10 different types of user actions for analysis and comparison. Finally, we conclude with discussions of the regularities and irregularities between datasets and what that might imply for both our understanding of the measurement of human behavior and the design of user-facing systems

2. RELATED WORK

2.1 Human activity sessions

The concept of an activity session is an intuitive one, but it's surprisingly difficult to tie down a single definition of what a session *is*. A "session" may refer to "(1) a set of queries to satisfy a single information need (2) a series of successive queries, and (3) a short period of contiguous time spent querying and examining results." [12]

(1) is referred to, particularly in search-related literature [8, 12], not as a session but as a task–a particular information need the user is trying to fulfil. Multiple tasks may happen in a contiguous browsing period, or a single task may be spread out over multiple periods. (2) is unclear. It may refer to a series of contiguous but unrelated queries (in which case it is identical to the third definition), or a series of contiguous queries based on the previous query in the sequence (in which case it is best understood as a sequence of tasks). (3) is the most commonly-used definition in the literature we have reviewed [11, 15, 20, 22]. This contrasts with the notion of *task* and is the definition of "session" that we have chosen for this paper. It's also the definition used by the W3C [21].

We found inspiration in thinking about how to model user session behavior in two distinct, but related threads: the empirical modeling work of cognitive science and the theoretical frameworks of human consciousness as applied to "work activities".

The lack of purely random distribution in the time between logged human actions has been the topic of recent studies focusing on the cognitive capacity of humans as information processing units. Notably, Barbasi showed that, by modeling communication activities with decision-based priority queues, he could show evidence for a mechanism to explain the heavy tail in time between activities [1] – a pattern he describes as bursts of rapid activity followed by long periods of inactivity. Wu et al. built upon this work to argue that short-message communication patterns could be better described by a "bimodal" distribution characterized by Poisson-based initiation of tasks and a powerlaw of time inbetween task events[23].

In contrast, Nardi calls out this cognitive science work for neglecting context in work patterns, motivation and community membership – thereby inappropriately reducing a human to a processing unit in a vacuum [16] (p21). Instead, Nardi draws from the framework of Activity Theory (AT) to advocate for an approach to understanding human-computer interaction as a conscious procession of *activities*. AT describes an activity as a goal-directed or purposeful interaction of a subject with an object through the use of tools. AT further formalizes an *activity* as a collection of *actions*[2] directed towards completing the activity's goal. Similarly, *actions* are composed of *operations*, a fundamental, indivisible, and unconscious movement that humans make in the service of performing an *action*.

For an example application of AT, let us examine Wikipedia editing. Our ethnographic work with Wikipedia editors suggests that it is common to set aside time on a regular basis to spend doing "wiki-work". AT would conceptualize this wiki-work overall as an *activity* and each unit of time spent engaging in the wiki-work as an "activity phase" – though we prefer the term "activity session".

The *actions* within an activity session would manifest as individual edits to wiki pages representing contributions to encyclopedia articles, posts in discussions and messages sent to other Wikipedia editors. These edits involve a varied set of *operations*: typing of characters, copy-pasting the details of reference materials, scrolling through a document, reading an argument and eventually, clicking the "Save" button.

In this work we draw from both the concepts of the *operation-action-activity* heirarchy of Activity Theory and the empirical modeling strategies of cognitive science as applied to time between events.

2.2 Session identification

User sessions have been used as behavioral measures of human-computer interaction for almost two decades, and for this reason, strategies for session identification from log data have been extensively studied [8].

Cooley et al. [5] and Spiliopoulou et al. [20] contast two primary strategies for identifying sessions from activity logs: "navigation-oriented heuristics" and "time-oriented heuristics".

Time-oriented heuristics refer to the assignment of an inactivity threshold between logged activites to serve as a session delimiter. The assumption implied is that if there is a break between a user's actions that is sufficiently long, it's likely that the user is no longer *active*, the session is assumed to have ended, and a new session is created when the next action is performed. This is the most commonly-used approach to identify sessions, with 30 minutes serving as the most commonly used threshold [8, 20, 17]. Both threshold and approach appear to originate in a 1995 paper by Catledge & Pitkow [4] that used client-side tracking to examine browsing behavior. In their work, they reported that the mean time between logged events 9.3 minutes. They choose to add 1.5 standard deviations to that mean to achieve a 25.5 minutes inactivity threshold. Over time this threshold has simplified to 30 minutes.

[2]We see Jones' conceptualization of tasks [12] as analogous to AT's conceptualization of *action*.

The utility and universality of this 30-minute inactivity threshold is widely debated; Mehrzadi & Feitelson [13] found that 30 minutes produced artefacts around long sessions, and could find no clear evidence of a global session inactivity threshold[3], while Jones & Klinkner [12] found the 25.5 minute threshold performed "no better than random" in the context of intentifying search tasks. Other thresholds have been proposed, but Montgomery and Faloutsos [14] concluded that the actual threshold chosen made little difference to how accurately sessions were identified.

Navigation-oriented heuristics involve inferring browsing patterns based on the HTTP referers and URLs associated with each request by a user. When a user begins navigating (without a referer), they have started a session; when a trail can no longer be traced to a previous request based on the referers and URLs of subsequent requests, the session has ended. This approach was pioneered by Cooley et al in 2002 [5]. While it demonstrated utility in identifying "tasks", and has been extended by Nadjarbashi-Noghani et al. [15], it shows poor performance on sites with framesets due to implicit assumptions about web architecture [3]. Further, the sheer complexity of this strategy and it's developmental focus on *task* over *session* make it unsuitable as a replacement for time-oriented heuristics in practical web analytics of user sessions.

In this work, we will challenge the assertion by prior works that (1) no reasonable cutoff is identifiable from the empirical data and (2) a global inactivity threshold is inappropriate as a session identification strategy. To our knowledge, we are the first to apply a general session identification methodology to a large collection of datasets and to conclude that not only are global inactivity thresholds an appropriate strategy for session identification, but also that, for most user-initiated actions, an inactivity threshold of 1 hour is most appropriate.

3. METHODS

This section is intended to both serve as a description of our methodology as well as to instruct readers on how to apply the same methods to their own datasets. First, we will discuss how we recommend applying our methodology for fitting interactivity clusters to a dataset. Then, we describe the origin of our datasets and the cleanup we performed in order to remove artifacts.

3.1 Fitting inter-activity times

First, we must gather a dataset of user-initiated actions with timestamps of at least *seconds* resolution. We generate inter-activity times on a per-user basis, so a relatively robust user identifier is necessary. While a persistent user identifier such as one associated with a user account is preferable, we have found that in the case of request logs, a fingerprint based on the request's IP and User-agent seems to be sufficient.

Once we have generated per-user inter-activity times, we plot a histogram based on the logarithmically scaled inter-activity time and look for evidence of a valley. Given the observations we have seen (and report in section 4), we expect to see a valley around 1 hour with peaks around 1 minute and 1 day. It is at this time that anomalies in the data should be detected and removed. For example, we found that the time between Wikimedia Mobile Views (described in the next section) had an unreasonably large spike at exactly 18 minutes of inter-activity time caused by a few (likely automated) users and removed their activities from the dataset.

Next, we try to fit a two component gaussian mixture model using expectation maximization [2] and visually inspect the results[4]. When the simple bimodal components did not appear to fit the data appropriately, we explored the addition of components to the mixture model with careful skepticism and repeated visual inspection.

Finally, if we have found what appears to be an appropriate fit, we identify a theoretically optimal inter-activity threshold for identifying sessions by finding the point where inter-activity time is equally likely to be within the gaussians fit with sub-hour means (within-session) and gaussians fit with means beyond an hour (between-session).

3.2 Datasets

To test this approach to session identification, we used a variety of datasets covering multiple sites, user groups, and types of action.

Wikimedia sites. One of the broadest groups of datasets comes from the Wikimedia websites (such as Wikipedia) and covers both page views (read actions) and edits. For the page views, we gather three datasets, each consisting of randomly-sampled page view events from the Wikimedia request logs. These covered app views (page views from the Wikimedia's official mobile app), mobile views (page views to the mobile site) and desktop views (page views to the desktop site). 100,000 IP addresses (or UUIDs, in the case of the mobile app) were selected, and all requests from those IPs/UUIDs for the month of October 2014 were retrieved. For desktop and mobile views, a UUID was produced by hashing the IP address, the User agent, and the accept_language provided with each request. After filtering out known crawlers and automata using tobie's ua-parser[5], we arrived at three page view datasets consisting of 2,376,891(app), 932,754(mobile) and 2,285,521(desktop) pageviews. These came from 100,000, 235,067 or 247,269 UUIDs. We also extracted inter-edit times from the English Wikipedia using the methodology we employed in previous work [9] – randomly selecting 1 million edits from 157,342 registered users.

AOL search. Contrasting with the Wikimedia datasets we used the (now infamous) AOL search logs[6] (aol, search)

[3]Note that this conclusion was reached using the same AOL search dataset that we analyze in this paper

[4]Note that we tried several strategies for statistically confirming the most appropriate fit – of which we found Davies–Bouldin index(DBI) [6] to be most reasonable – but none were as good as a simple visual inspection, so we employ and recommend the same.

[5]`https://github.com/tobie/ua-parser`

[6]These logs are controversial due to their inclusion of search terms containing private information, and there has historically been an ethical debate about their use. We have modified the dataset to strip search terms so that it consists solely of unique IDs and timestamps, as has been used in the past.[13] See `https://en.wikipedia.org/wiki/AOL_search_data_leak` for more discussion.

Table 1: Fit and threshold information for clusters. Note that fits correspond to logarithmically scaled (base 2) seconds between events. For example, $2^{6.7} = 104$ seconds. It's important to report these values in log scale because, while the mean can be re-exponentiated, the standard deviation of log values doesn't make sense that way.

dataset	theshold (min)	short within			within			between			break		
		μ	σ	λ	μ	σ	λ	μ	σ	λ	μ	σ	λ
aol search	115				6.7	2.9	0.70	16.8	2.2	0.30			
cyclo. route	89				5.0	2.5	0.87	18.6	3.1	0.13			
wiki. app	29				5.2	2.3	0.74	15.7	2.5	0.26			
wiki. mobile	50				6.4	2.6	0.65	15.8	2.5	0.35			
wiki. desktop	46				5.5	2.6	0.75	15.7	2.5	0.25			
osm change	101				8.6	2.1	0.68	15.5	2.5	0.30	22.7	2.0	0.02
wiki. edit	80				6.8	2.5	0.83	15.4	2.7	0.16	22.6	1.9	0.01
mov. rating	33	3.0	1.3	0.58	5.2	1.9	0.34	18.0	3.0	0.07			
mov. search	52	4.0	0.8	0.30	5.7	2.5	0.50	17.1	3.1	0.20			
lol game	14				8.3	0.5	0.59	14.1	2.8	0.41			
s. o. answer	91				10.2	1.7	0.30	16.6	2.9	0.63	23.0	1.5	0.06
s. o. quest.	335				12.7	1.7	0.10	18.5	2.1	0.63	22.4	1.7	0.26

consisting of 36,389,567 search actions from 657,427 unique IDs. These actions span from March through May of 2006.

Cyclopath. We also gathered a dataset from Cyclopath, a computational geowiki leveraging cyclists' local knowledge to collabortively build a map [18]. The dataset consists of HTTP requests to the Cyclopath server that are automatically labelled by type. We filtered these requests to include only those that represent a request for a cycle route between two points (cyclopath, route get). This came to 6,123 requests from 2,233 distinct registered users.

Movielens. To explore different types of search and contributory behavor, we also extracted logs from the MovieLens movie recommender system, which has been in use since 1997. As of November 2014 there are 225,543 unique users who have provided more than 21 million movie ratings for more than 25,000 movies. From MovieLens, we extracted two datasets: (movielens, rating) consists of movie rating actions from between 1997 until 5 November 2014, and (movielens, search) consists of search actions from 19 December 2007 to 1 January 2014.

StackOverflow. This popular question/answer system relating to programming and software engineering regularly releases public data dumps. For our analysis, we extracted questions asked and answers posted between July 2008 and September 2013. The question dataset (stack overflow, question) consists of 6,397,301 questions from 1,191,748 distinct users, while the answer dataset (stack overflow, answer) consists of 11,463,991 answers from 790,713 distinct users.

OpenStreetMap (OSM). This open-source alternative mapping service also publishes regular database dumps. We downloaded a full history dump of OSM contributions as of 24 February 2014, restricting this to the North American region as defined by Geofabrik[7], which consists of the United States, Canada and Greenland. OSM groups individual changes to the map into *changesets*[8] when an editor saves their work. We used the timestamp of the last revision in a changeset as the time that the user saved the changeset. The resulting dataset (osm, changeset) contains 13,388,923 million changesets from 46,595 distinct users. We found that more than 75% of changesets occured with less than 5 seconds of inter-activity time and assume that this is only possible in the case of a data import – not human behavior – and filtered them from the dataset.

League of Legends. This widely-played online multiplayer game supports an extension tha logs game data and play times for all users of the extension. Notably, we used this dataset in previous work to study the effect of deviant behaviour on player retention [19]. The dataset consists of roughly 2.5 million unique players participating in almost 166 million games. We extracted the time between when a user finished a game and started playing the next game as an inter-activity time (lol, game). Though not all games were captured in the dataset provided via this extension (see [19] for more details), missing data is believed to be most prevalent around newer players with less consistent play habits.

Taken together, these datasets represent seven different systems and include different interaction mechanisms (mobile apps, mobile devices, desktop devices and a video game interface), and different classes of interaction (web search & route finding, contributions to collaboratively edited artifacts, page reads, and games played).

4. RESULTS & DISCUSSION

In this section, we present and discuss the result of applying our proposed inactivity threshold identification method to the datasets described in the previous section. We start with datasets that were well fit with two clusters. Then we move to more complicated fits and discuss the implications of additional clusters. Finally, we demonstrate datasets with less suitable fits and discuss what this implies about the nature of participation in these systems. Reference table 1 for fitted values and thresholds.

[7] `http://download.geofabrik.de/north-america.html`
[8] `http://wiki.openstreetmap.org/wiki/API_v0.6#Changesets_2`

4.1 Simple bimodal fits

Most of the datasets of user-initiated inter-activity times that we observed display a simple bimodal distribution when their histograms are plotted on a logarithmically scaled X axis. Figure 1 plots a log inter-activity time histogram overlaid with expectation maximization fits of a mixture of two log-normal cluster components. Notably, the AOL search logs represent one of the most clear fits to this bimodal distribution. This suggests that, counter to Mehrzadi & Feitelson's conclusions [13], there *does* seem to to be a clear location for an inactivity cutoff in this dataset – at approximately one hour.

Figure 1 also demonstrates the striking regularity of inter-activity time clusters between systems. All of the systems presented show a clear fit for a theoretical *within-session* cluster with a mode around one minute and a theoretical *between-session* cluster with a mode at one day. Each fit intersects at approximately one hour – with Wikimedia app views displaying the lowest intersection at 29 minutes while AOL searches display the highest intersect at 115 minutes – nearly two hours. Despite this variance in the intersection points, a visual inspection of the empirical distribution does not suggest that the choice of a one hour cutoff for either of these datasets would be inappropriate. Indeed, many of the *between-session* clusters appear to be left shifted due to a lack of longitudinal data, and it is only in these cases that the intersection falls below the one hour mark.

Also of note in these results is the spike of probability of a 24 hour inter-activity time for all but the cyclopath dataset. This suggests that, for reading Wikimedia sites and searching in AOL, there is a strong tendency to return on a daily basis. The curious lack of such a day-spike for cyclopath route searches could be explained by the type of usage the site sees. Bicycle route searching may be less of a daily information need than web search and Wikimedia's reference content.

4.2 Fits with extended breaks

In some cases, we found that the data were fit better by adding a third component to the mixture model that represents very low frequency events. Figure 2 shows the fits for the inter-activity time between OpenStreetMap's changesets and English Wikipedia edits. Note that, like the bimodal fits above, we again see modes for the *within-session* cluster around one minute and modes for the *between-session* cluster around one day. However, we found that we could more cleanly fit these datasets with an additional cluster with a mode of around 2.5 months.

As we noted in our past work [9], we believe that this low frequency cluster represents an extended break from contributing that corresponds to a life event – like getting married, buying a house, going to school or getting a job. Wikipedia editors refer to this phenomena in volunteer participation as a "wikibreak"[9]. We suspect that the reason for the tiny scale of this cluster is two-fold: (1) contributors who work on Wikipedia or OpenStreetMap for long enough to take an extended break are rare compared to the rate of higher frequency activity and (2) breaks at the scale of 2-3 months often result in total abandonment of participation in the project.

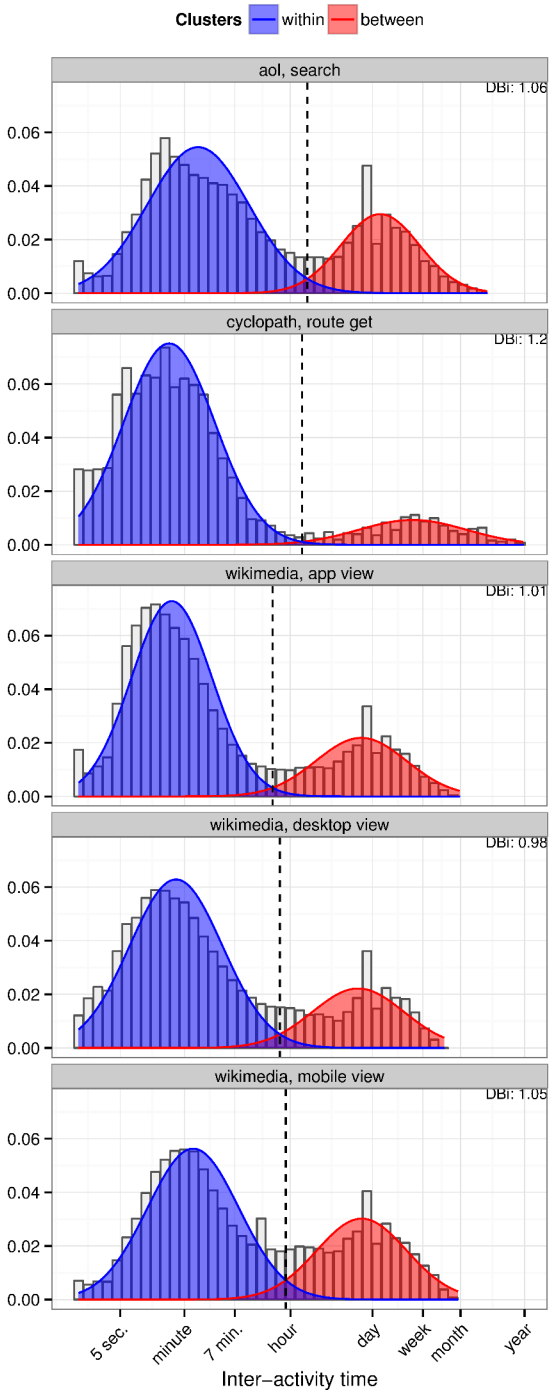

Figure 1: Bimodal clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for datasets where two clusters appeared to sufficiently explain the observed data.

[9] https://en.wikipedia.org/wiki/Wikipedia:Wikibreak

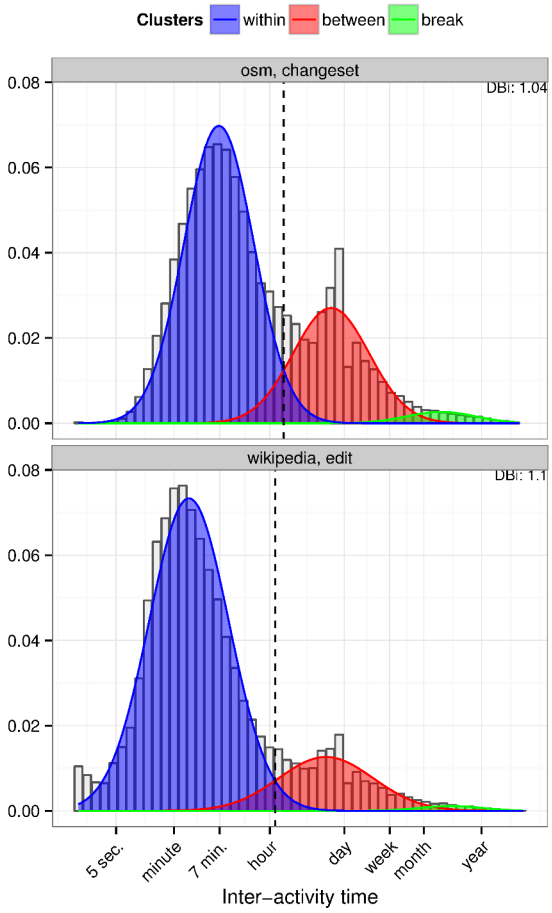

Figure 2: Trimodal clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for datasets where an additional, "break" cluster was needed to fit the data.

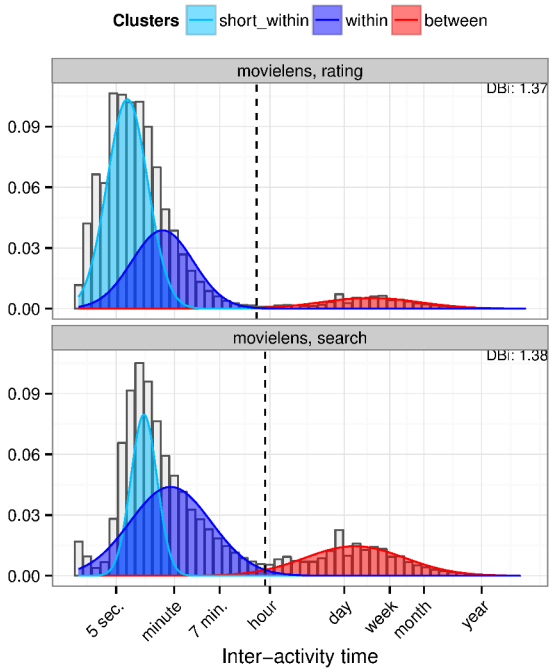

Figure 3: High frequency activity clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for datasets where an additional, high-frequency inter-activity cluster was needed to fit the data.

4.3 Fits with a high frequency component

When observing the distribution of inter-activity times for ratings and searches in Movielens, we found that both these events occurred with higher frequency than the other datasets. This made us suspect that there could be an additional cluster component at a high frequency time interval. Figure 3 shows how the two datasets lent themselves to this additional "short within" component. Like in previous mixture models, we see a within-session cluster with a mode around one minute and a between-session cluster with a mode around one day. However, in these datasets we also observed a pattern in inter-activity times that suggested a faster component with a mode around 15 seconds.

Given that this component occurs at shorter intervals than the within-session component, we assume that it also represents within-session activity. In the case of rating, this high frequency component could represent the rapid rating behavior that the MovieLens interface affords – a user can rate several movies from a list without leaving a page. However, we are less sure on how to explain the high frequency component of MovieLens searches. It could be that, unlike when performing a web search (AOL) or reading encyclopic content (Wikimedia), users' movie searches are more likely to benefit from more rapid iteration.

4.4 Unusual fits

While the fits described so far follow a clear pattern with somewhat minor nuance as to the nature of the gaussian fitting strategy, other datasets did not seem to fit at all. This

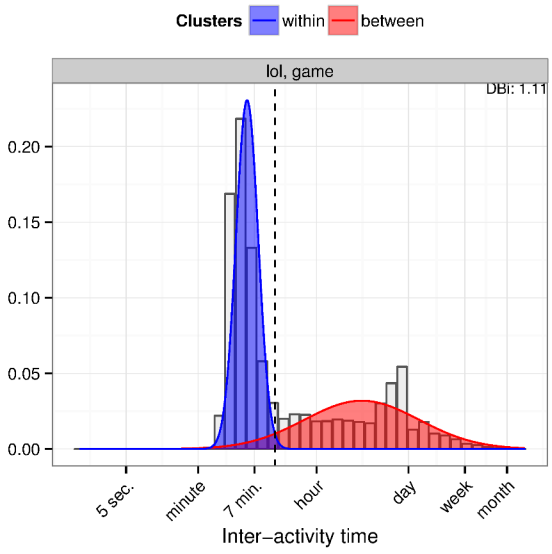

Figure 4: Inter-game clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for time between League of Legends games.

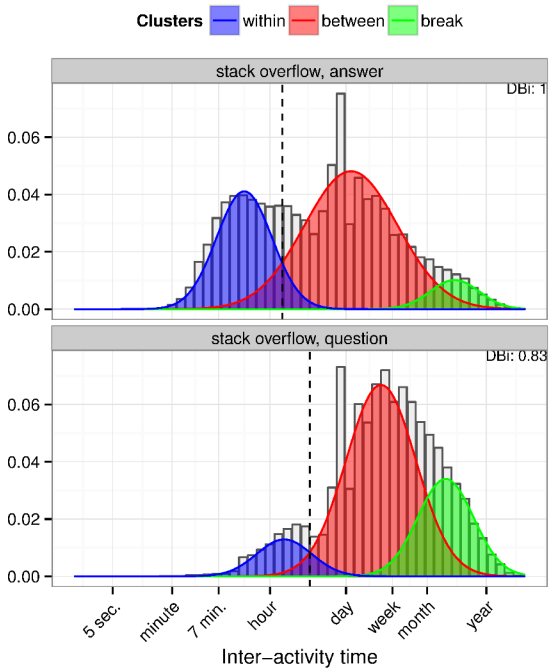

Figure 5: Low frequency clusters. Empirical inter-activity density (bars) and (non-convergent) fitted mixture models of gaussians are plotted for time between posts on Stack Overflow.

suggest that our strategy for identifying session thresholds is not universally suitable for all user-intiated log events.

League of Legends. Figure 4 shows the two cluster fit for League of Legends game playing. Here, we see a very high density component with a mode around five minutes and a very wide component with a mode around five hours. The intersection of these components place the threshold at approximately 14 minutes. It is important to note that the tightness of the dense component may be an artifact of the way that inter-game times differ from the inter-activity times observed in the other datasets. In the case of this dataset, only the time between games is accounted for – the time between the end of one game and the beginning of the next.

There also may be constraints inherent to the system that limit the potential time spans in which a user could possibly act. For example, League of Legends employs a queuing mechanism for matching teammates with opponents that takes approximately 5 minutes to complete most of the time. Our own experience with the game suggests that many users will often finish one game and immediately get into the queue for another. It is likely that these system limitations are the reason for infrequent between-game times under 1 minute. It seems clear from this result that understanding a system's limitations on user behavior is important when interpreting cluster fit.

Stack overflow. Unlike the other datasets observed, the time between Stack Overflow posts does not suggest a clear valley from which to draw intuition about where to draw a session cutoff. Figure 5 shows the (non-convergent) fits of question asking and answering activities. In this case, there is a dramatic reduction in the scale of the higher frequency time components and what appears to be a shift of the within-session component to the right.

If we are to interpret the fit of these clusters as meaningful, the right shift of the within-session component could be due to the time needed to produce a high quality question or answer. Stack Overflow's incentive structure is designed to encourage high quality posts. High quality posts are more likely to be reviewed positively by other users, and a user's score within Stack Overflow is largely dependent on how other users rate the quality of their posts[10]. It seems likely that producing a high quality post would take a substantial amount of time and that this time investment would make it difficult to complete posts with a high enough frequency to produce a short inter-activity time component like we saw in other systems. In this case, it seems that either our strategy for identifying a suitable inactivity threshold is insufficient or that Stack Overflow users rarely post more than one question or answer within an *activity session*.

5. IMPLICATIONS & FUTURE WORK

In this paper, we have challenged previous literature that suggests no apparent global inactivity threshold exists for identifying user sessions from logs. From our results, we propose a simple, yet apparently robust, rule of thumb and a methodology for checking this rule in other datasets. The rule of thumb is easy to apply; our analysis suggests that setting an inactivity threshold to demarcate the end of a session at *one hour* will be appropropriate for most kinds of activity log analysis.

[10]`http://meta.stackexchange.com/help/whats-reputation`

We suspect that this strategy will be robust to new datasets since (1) it is grounded in empirical observations of a natural valley in activity times that corresponds to our intuitions about users' activities and (2) it holds constant across a wide range of systems and activity types. Even when our threshold detection strategy deviated from one hour, the devations were relatively small given the scale of activities, and in some cases, this deviation could be explained by limitations in the data used to fit our models. However, we still advise that any new application of session identification using an hour as an inactivity threshold is preceeded by a plot of a histograph of log-scaled inter-activity times and visual inspection for a natural valley between 1 minute and 1 day.

These results and our recommendations stand in the face of a long and nuanced discussion of the nature of user sessions as can be extracted from logged interactions with a computer system. We place our criticisms of previous work into two categories: (1) previous empirical work did not attempt to look for log-normally distributed patterns and therefore concluded that no obvious separation between within- and between-session inter-activity times exists[13][4] and (2) other work exploring *task driven* behavior conflates "task" with "session". We challenge (1) on the basis of the clear trends represented in the results of this work and (2) by drawing a distinction between goal-directed tasks and activity sessions which often represent a collection of heterogenious goal-directed tasks.

Further, given the strong regularities we see between different types of human-computer interactions, our results suggest something more fundimental about human activity itself. As discussed in section 2.1, Activity Theory(AT) conceptualizes human consciousness as a sequence of *activitie sessions* which represent a heirarchical relationship with *actions* and *operations*. We suspect that the fact that *operations* and *actions* must be performed in a sequence explains the temporal rhythm we observe. While it's hard to say conclusively, we suspect that the "short_within" clusters we observe represent *operation*-level events, the "within" clusters represent *action*-level events, and the "between" clusters represent *activity*-level events.

If this application of AT to the observed patterns is accurate, this could have substantial implications for the design of systems. System designers may be able to take advantage of the regularities observed by constructing systems that afford operations, actions and activity sessions at timescales that humans will find natural. Our analysis suggests that operations should exist at the timescale of about 5-20 seconds, actions should be completable at a timescale of 1-7 minutes and activities should be supported at daily to weekly time intervals. We suspect that systems that do not allow users to work at these time scales may be frustrating or may otherwise limit the ability of their users to function at full capacity.

These ruminations about human behavior and its manifestation in well designed systems are only speculation at this point. New work will need to be done to explore whether our predictions hold and whether limiting or enabling certain types of activity rhythms substantially affects user experience or performance.

6. ACKNOWLEDGMENTS

We thank Stuart Geiger for his involvement in our previous work and for the inspiration he provided toward pushing for measurements that more accurately represent human activity. This work has been funded in part by the National Science Foundation (grants 0808692, 0964695, 0968483, 1017697, 1111201, 1218826, 1319382). We are grateful to Dror Feitelson for agreeing to share the AOL search dataset with us. Individual authors would also like to thank Katie Horn and Margret Wander for their feedback and inspiration. Our data[11] and source code[12] are openly available.

7. REFERENCES

[1] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[2] T. Benaglia, D. Chauveau, D. R. Hunter, D. S. Young, et al. mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[3] B. Berendt, B. Mobasher, M. Nakagawa, and M. Spiliopoulou. The impact of site structure and user environment on session reconstruction in web usage analysis. In *WEBKDD 2002-Mining Web Data for Discovering Usage Patterns and Profiles*, pages 159–179. Springer, 2003.

[4] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN systems*, 27(6):1065–1073, 1995.

[5] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, 1(1):5–32, 1999.

[6] D. L. Davies and D. W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

[7] D. Donato, F. Bonchi, T. Chi, and Y. Maarek. Do you want to take notes?: identifying research missions in yahoo! search pad. In *Proceedings of the 19th international conference on World wide web*, pages 321–330. ACM, 2010.

[8] C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232. ACM, 2014.

[9] R. S. Geiger and A. Halfaker. Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 861–870. ACM, 2013.

[10] S. Gomory, R. Hoch, J. Lee, M. Podlaseck, and E. Schonberg. Analysis and visualization of metrics for online merchandizing. In *Proceedings of WEBKDD'99*, 1999.

[11] K. Goševa-Popstojanova, A. D. Singh, S. Mazimdar, and F. Li. Empirical characterization of session–based workload and reliability for web servers. *Empirical Software Engineering*, 11(1):71–117, 2006.

[12] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 699–708. ACM, 2008.

[11] `http://dx.doi.org/10.6084/m9.figshare.1291033`

[12] `https://github.com/halfak/Activity-sessions-research`

[13] D. Mehrzadi and D. G. Feitelson. On extracting session data from activity logs. In *Proceedings of the 5th Annual International Systems and Storage Conference*, SYSTOR '12, pages 3:1–3:7, New York, NY, USA, 2012. ACM.

[14] A. L. Montgomery and C. Faloutsos. Identifying web browsing trends and patterns. *Computer*, 34(7):94–95, 2001.

[15] M. Nadjarbashi-Noghani and A. A. Ghorbani. Improving the referrer-based web log session reconstruction. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 286–292. IEEE, 2004.

[16] B. A. Nardi. *Context and consciousness: activity theory and human-computer interaction*. Mit Press, 1996.

[17] J. L. Ortega and I. Aguillo. Differences between web sessions according to the origin of their visits. *Journal of Informetrics*, 4(3):331–337, 2010.

[18] R. Priedhorsky and L. Terveen. The computational geowiki: what, why, and how. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 267–276. ACM, 2008.

[19] K. B. Shores, Y. He, K. L. Swanenburg, R. Kraut, and J. Riedl. The identification of deviance and its impact on retention in a multiplayer game. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1356–1365. ACM, 2014.

[20] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Informs journal on computing*, 15(2):171–190, 2003.

[21] W. W. W. C. (W3C). Web characterization terminology & definitions sheet, May 1999.

[22] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594. ACM, 2010.

[23] Y. Wu, C. Zhou, J. Xiao, J. Kurths, and H. J. Schellnhuber. Evidence for a bimodal distribution in human communication. *Proceedings of the national academy of sciences*, 107(44):18803–18808, 2010.