# User sessions identification based on strong regularities in inter-activity time

Aaron Halfaker
Wikimedia Foundation
ahalfaker@wikimedia.org

Oliver Keyes
Wikimedia Foundation
okeyes@wikimedia.org

Daniel Kluver
GroupLens Research
University of Minnesota
kluver@cs.umn.edu

Jacob Thebault-Spieker
GroupLens Research
University of Minnesota
thebault@cs.umn.edu

Tien Nguyen
GroupLens Research
University of Minnesota
nguy1749@umn.edu

Anuradha Uduwage
GroupLens Research
University of Minnesota
uduwage@cs.umn.edu

## ABSTRACT

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

## General Terms

Theory

## Keywords

User session

## 1. INTRODUCTION

In 2012, we had an idea for a measurement strategy that would bring insight into understand a online community. While studying the nature of 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. While past work had relied on counting the number of contributions made by a user[1] as a measure of investment, we felt that the amount of time a user invested into editing might more accurately measure investment. This supposition was inspired by past work attempting to estimate the amount of time invested into Wikipedia as a whole[**?**].

The measurement strategy we came up with is 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[**?**]. While we found the notion of work sessions to be intuitive from our ethnographic work in Wikipedia, we did not find a consensus in the literature for how to identify sessions from timestamped user activities. So, we looked to the data itself for insight on what might be a reasonable approach splitting users' editing activity into sessions. The regularities we found in inter-activity time amazed us with their intuitiveness and the simplicity of session demarcation implied. It is that work that lead 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 trying to identify 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[9] – especially in the domain of search[cite cite cite] and content personalisation[15, 8]. Despite this interest in understand the nature and manifestation of user sessions, no clear consensus has emerged. In fact, some work has gone as far as arguing that sessions don't actually exist as a useful divide for user activity[10] and that the strategy of choosing a global inactivity threshold is arbitrary[12].

In this paper, we will propose and demonstrate a strategy for identifying user sessions from log data and demonstrate that the strategy works consistently across many different types systems and user activity. We'll start by summarizing previous work trying to make sense of user session behavior from log data. We'll also discuss theoretical arguments about how intuitive user behavior (e.g. tasks and sessions) ought to manifest in the data. Next, we'll discuss a generalized version of the session threshold identification strategy we developed in [**?**] and present strategies for fitting this strategy to new data. Then, we'll introduce 6 different systems from which we have extracted 12 different logged user action types for analysis and comparison. Finally, we'll conclude with discussions of the regularities and irregularities between datasets and what that might imply for both our understanding of human behavior and the measurement of it.

## 2. RELATED WORK

### 2.1 Human activity sessions

The concept of a session is an intuitive one, but it's sur-

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


*WWW* '15 Florence, IT

prisingly difficult to tie down a single definition of what a session is, and how it can be demarcated. 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."[10] (1) is referred to, particularly in search-related literature,[10, 7] 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[15, 17, 9, 13] This contrasts with tasks (you can have multiple tasks in a session, or multiple sessions to a single task) and is the definition of "session" that we have chosen for this paper. It's also the W3C definition.[16]

We found inspiration in thinking about how to model user session behavior in both 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 human behavior has been the topic of recent studies focusing on the cognitive capacity of humans as information processing units. Notably, Barbasi showed that, by modeling the 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 sort-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[18].

The framework of Activity Theory(AT) approaches the conscious process of human work with tools (computer systems in an HCI context) by focusing on activities. AT describes activities 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 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 example application of AT, let's 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 manifest as 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 hope to draw together 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 over a decade, and for this reason, strategies for session identification of log data has been the extensively studied[7].

Cooley et al.[5] and Spiliopoulou et al.[15] contast two primary strategies for identifying sessions from activity logs: "navigation-oriented heuristics" and "time-oriented Heuristics". Time-oriented heuristics refers to the assignment of a inactivity threshold between logged activites to serve as a session delimiter. The assumption here 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 as the most commonly used threshold[15, 7, 14]. Both threshold and approach appear to originate in a 1995 paper by Catledge & Pitkow[4] that used client-side tracking to identify browsing behaviour. In their work, they reported that the mean time between user observed user events in their data was 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 proposed inactivity threshold has gradually been smoothed out to 30 minutes. The utility of the 30-minute threshold is widely debated; Mehrzadi & Feitelson (2012) [11] found that 30 minutes produced artefacts around long sessions, and could find no clear evidence of a natural session inactivity threshold[2], while Jones & Klinkner[10] found the 25.5 minute threshold "no better than random" in the context of search tasks. Other thresholds have been proposed, but Montgomery and Faloutsos[12] concluded that the actual threshold chosen made little difference to its accuracy.

Referer-based reconstruction involves taking the referers and URLs associated with each request by a user, and chaining them together. When a user begins navigating without a referer, they have started a session; when a trail can no longer be traced to that 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.[13] it shows poor performance on sites with framesets due to implicit assumptions about web architecture[3]. 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.

In our this work, we'll challenge the conclusions of prior works' assertions (1) that no reasonable cutoff is from the empirical data and (2) that 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 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 appropriate.

# 3. METHODS

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

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 dataset. First, we'll discuss how we recommend applying our methodology for identifying inter-activity type component clusters to a dataset. Next, we'll describe the origin of our datasets and the cleanup we performed in order to generate inter-activity times to fit.

## 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've 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 per-user inter-activity times, we plot a histogram based on the logarthmically scalled iter-activity time and look for evidence of a valley. Given the observations we have seen (and report in section **??**), we expect to see a valley around about 1 hour with peaks around 1 minute and 1 day. It's 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 absurd 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[3] 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

[**FIXME**: *Oliver writes up dataset descriptions*]

# 4. RESULTS & DISCUSSION

In this section, we present and discuss the result of the application of our proposed inactivity threshold identification analysis on the datsets. First we start off with the common, simple cluster fits. 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.

## 4.1 Simple bimodal fits

[3]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.

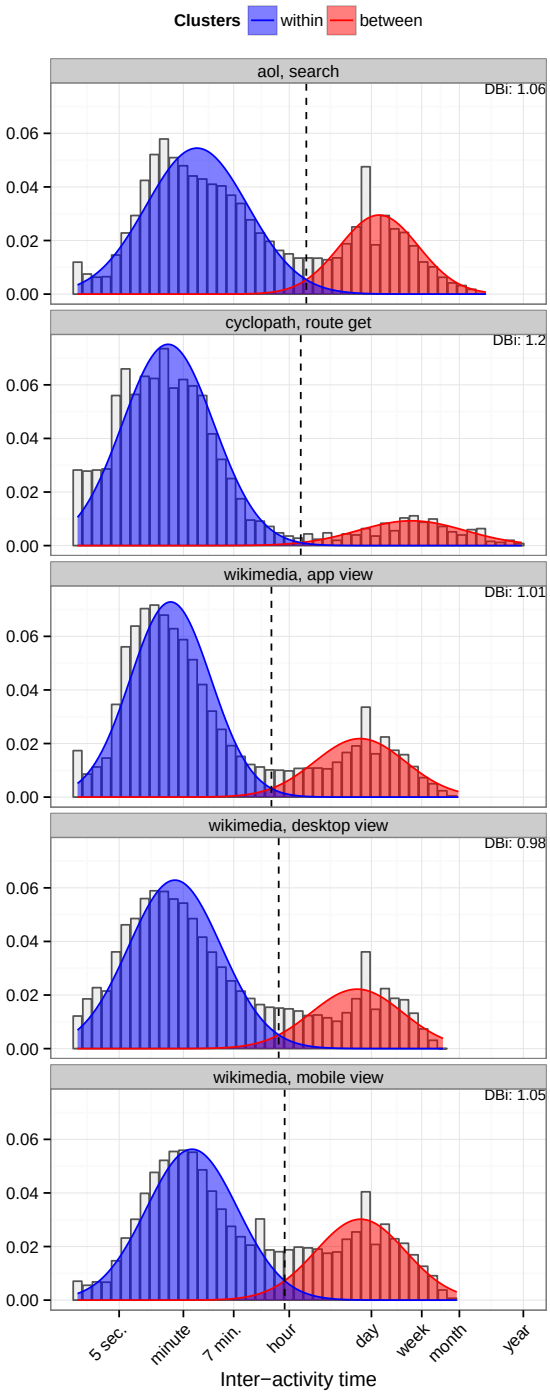


**Figure 1: Bimodal clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for datasets where two clusters appeared to sufficiently explain the observed data.**

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 **??** 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[11], there does seem to to be a clear location for an inactivity cutoff in this dataset – at approximately one hour.

Figure **??** demonstrates the striking regularity of inter-activity time between systems. All of the systems presented show a clear fit for a theoretical *within-session* cluster with a mode around 1 minute and a theoretical *between-session* cluster with a mode at 1 day. Each fit intersects at approximately one hour – with Wikimedia app views display 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 1 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 datasets. 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 encyclopedia 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 **??** shows the fits for the inter-activity time between Open Street Map's changesets and English Wikipedia edits. Note that, like the bimodal fits above, we again see modes for the *within-session* cluster around 1 minute and modes for the *between-session* cluster around 1 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 [**?**], 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"[4]. We suspect that the reason for the tiny scale of this cluster is two-fold: (1) contributors who work on Wikipedia or Open Street Map for long enough to to take an extended break are rare compared to other, higher frequency activity and (2) breaks often result in total abandonment of participation in the project.

## 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

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

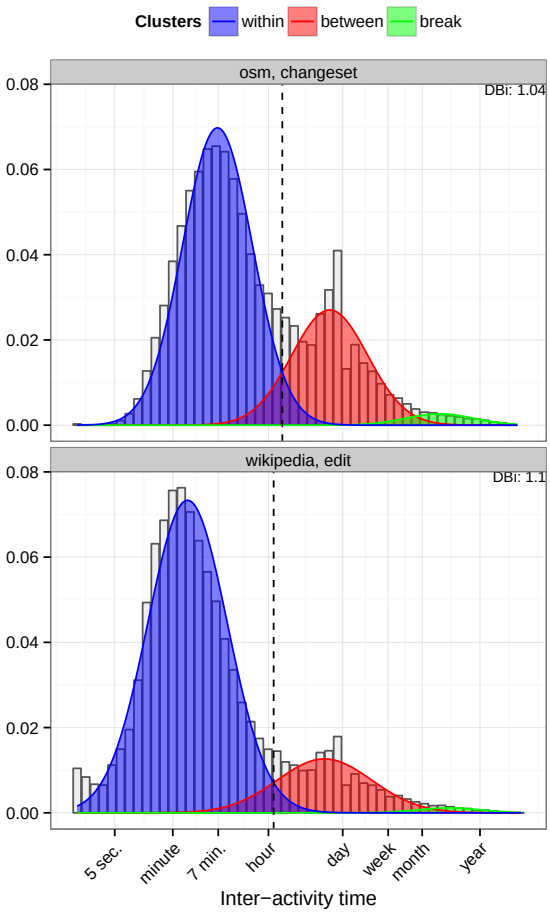


**Figure 2: Trimodal clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for datasets where an additional, "break" cluster was needed to fit the data.**

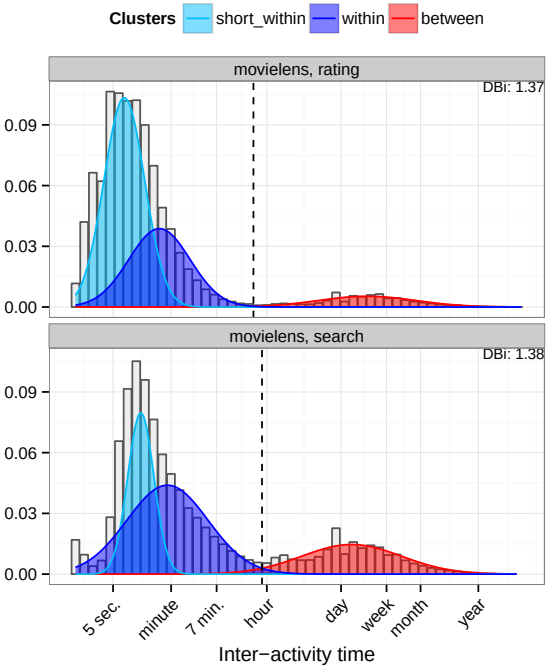


**Figure 3: High frequency activity clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for datasets where an additional, high-frequency inter-activity cluster was needed to fit the data.**

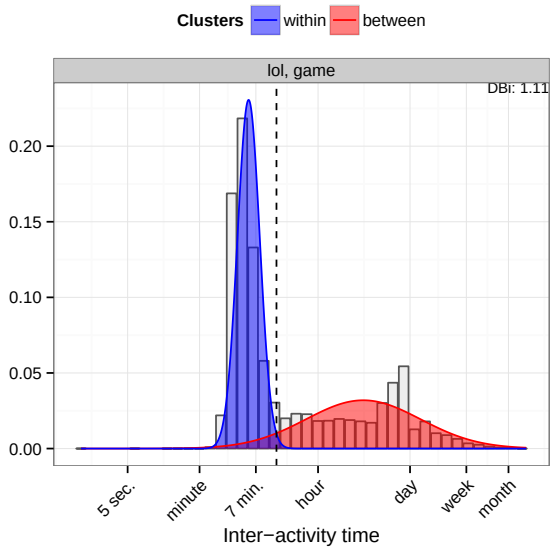


**Figure 4: Inter-game clusters. Empirical inter-activity density (bars) and fitted mixture models of gaussians are plotted for time between League of Legends games.**

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 **??** 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 1 day. However, in these datasets we also observed a pattern in inter-activity times that suggested a faster component with a mode around 30 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're less sure of 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, the other datasets we observed suggest that the this strategy for identifying session thresholds is not universally suitable for all user-intiated events.

**League of Legends.** Figure **??** shows the two cluster fit for League of Legends game playing. Here, we see a very high density component with a mode around 5 minutes and a very wide component with a mode around 5 hours. The intersection of these component places the threshold at approximately 14 minutes. It's important to note that

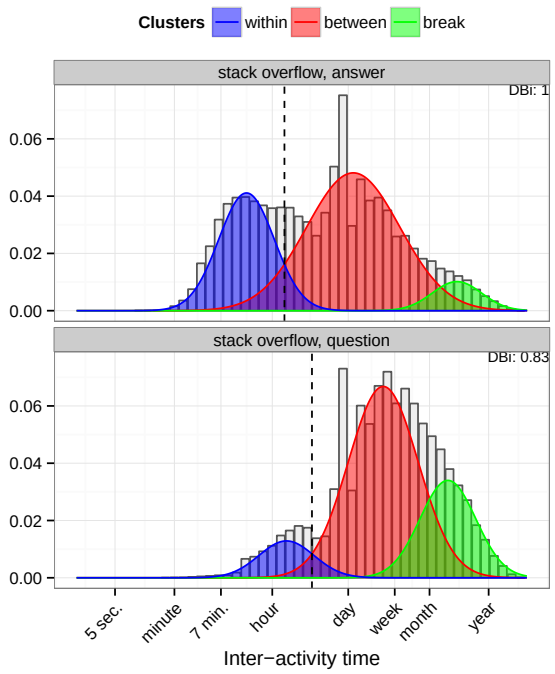


**Figure 5: Low frequency clusters. Empirical inter-activity density (bars) and (non-convergent) fitted mixture models of gaussians are plotted for time between posts on Stack Overflow.**

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 – not the time between game-start or game end.

There also may be constraints placed on the potential time spans in which a user could possibly act by the software itself. For example, League of Legends employs a queueing mechanism for managing the number of active games that takes approximately 5 minutes to complete at most times. Our own experience with the game suggests that many users will jump right from one game to the queue for another. If this is the cause of the dense component, it could be that, in this case, we not modeling human behavior so much as the constraints placed on human bahavior by software.

**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 **??** 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 StackOverflow is largely dependent on how other users rate the quality of their posts[5]. It's seems likely that producing a high quality post would take a substantial amount of time and that this time investment would make posting with a high enough frequency to produce a short inter-activity time component like we saw in other systems difficult. 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

Our work stands in the face of previous work ...

* We propose a rule of thumb and a simple methodology. Our analysis suggests that setting an inactivity threshold at one hour may be robust to new datasets. However, we still advise that any new application of session identification using an inactivity threshold is preceeded by a plot of a log histograph of inter-activity times and visual inspection for a natural valley between 1 minute and 1 day.

* On the nature of human activity, strong regularities of inter-activity time. ** Activity Theory? Maybe these clusters represent operation, action and activity session. * What this might imply for the design of systems. ** If human behavior is well represented by a sequence of conscious activities the regularities in time between events hold... ** System designers can take advantage of this by designing systems that afford operations, actions and activity at timescales that humans will feel to be natural. ** Indeed, we suspect that activities that force users to devate from these patterns may be frustrating or may otherwise limit their ability to function at full capacity.

# 6. ACKNOWLEDGMENTS

# 7. ADDITIONAL AUTHORS

Additional authors: Morten Warncke-Wang (GroupLens Research, email: `morten@cs.umn.edu`) and Kennith Shores (GroupLens Research, email: `shores@cs.umn.edu`).

# 8. 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.

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

[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] 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.

[8] S. Gomory, R. Hoch, J. Lee, M. Podlaseck, and E. Schonberg. Analysis and visualization of metrics for online merchandizing. In *Proceedings of WEBKDDâĂŹ99*, 1999.

[9] 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.

[10] 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] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.