Mining Online Training Log Data

Susan Mehringer*
Cornell University
Center for Advanced Computing
Ithaca, New York
shm7@cornell.edu

Christopher R. Myers*
Cornell University
Center for Advanced Computing
Ithaca, New York
c.myers@cornell.edu

Jennifer Houchins
The Shodor Education Foundation, Inc.
Durham, NC
jhouchins@shodor.org

Lorna Rivera
Georgia Institute of Technology
Center for Education Integrating Science, Mathematics, and Computing
Atlanta, GA
lorna.rivera@gatech.edu

ABSTRACT

Online training has been growing in popularity, and offers many advantages for both trainers and learners. Assessing the usage and impact of online material can be difficult, especially if content is made available to anyone and is not part of a course requiring formal enrollment. The Cornell Virtual Workshop (CVW) first offered online training on topics in high-performance computing and computational science in 1994, and ten years ago we began logging usage. We are now performing our first in-depth analysis of those log data to identify patterns in usage, so that we can better understand how users access the material, which types of topics and materials result in the greatest impact, how topic usage changes over time, and what types of presentation format might be preferred. While the CVW is built around a cohesive, sequential narrative for each training topic, we find that many users access our content in a more targeted fashion, suggesting that we rethink how we package our material. We anticipate that ongoing analysis using data science and machine learning methods will enable us to produce more useful training materials, and provide the educational community with valuable information about patterns in online material usage.

CCS CONCEPTS

• **Social and professional topics** → **Computing education programs**; **Computational science and engineering education**; *Informal education*; • **Computing methodologies** → *Cluster analysis*; *Dimensionality reduction and manifold learning*.

*Both authors contributed equally to this research.

PEARC '19, July 28-August 1, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-7227-5/19/07…$15.00
https://doi.org/10.1145/3332186.3332235

KEYWORDS

Training, Usage Statistics, HPC, XSEDE

ACM Reference Format:
Susan Mehringer, Christopher R. Myers, Jennifer Houchins, and Lorna Rivera. 2019. Mining Online Training Log Data. In *Practice and Experience in Advanced Research Computing (PEARC '19), July 28-August 1, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3332186.3332235

1 INTRODUCTION

Online training has been growing in popularity over the past couple of decades. From the learner's perspective, the materials can be accessed at any time, without travel costs, and it is possible to go directly to the specific content desired, often for free. From the provider's view, offering material in this way is scalable and enables reaching a wider audience. Online training is used in accredited courses and MOOCs, and students taking classes for credit are motivated to be engaged and complete coursework whether it is offered online or in person. But how are these online materials used in more informal contexts, where the motivation is unknown by the developer? The content might be used by someone to learn a new concept, to find a quick fix to a programming problem, or to get started using a new computing resource.

The Cornell Virtual Workshop (CVW) first offered online training in 1994 (cvw.cac.cornell.edu). Our target population includes anyone who wants instruction on using high performance computing to enable or improve their computational science efforts; this group can include researchers, professionals, and students who are typically more interested in finding the instruction they need to continue a project or research endeavor, rather than completing a course for credit or other recognition. Topics range from general material such as Introduction to Python to more complex topics such as Vectorization and Checkpoint/Restart. The web-based format includes use of video, pop-up glossary terms, simulation codes, exercises, and self-tests. The CVW is free, offers no credit, and sign-in is not required.

CVW tutorials are focused around particular topical subject matter, each organized as something like a mini-course. While we have designed the tutorials around the general notion that users would work through the material from start to finish, we expected to find

different use patterns. Some users probably start working through a tutorial, but then lose interest or decide that the associated content is not really what they were looking for. Other users may come looking for particular information, find (or not) what they are looking for, and move on. Ten years ago, we began logging usage, both in a local database and using a commercial tracking tool. We are interested generally in the degree to which, and patterns by which, users access CVW content. We address the questions below through different lenses and filters, using both deeper dives into smaller portions of the data and broader characterizations using tools and methods from data science and machine learning.

How can we know if learners are benefiting from the material? We are motivated to examine the data on usage patterns so that we can understand:

- How does usage differ among different groups of visitors?
- How are the materials used? Do learners step through in order, or go directly to the desired materials and leave?
- How have usage patterns changed over time?
- How does topic interest change over time after release of material? For example, resource-specific materials are clearly most popular when the resource is first available, but what about other types of topics like programming or performance optimization?
- Which content types are most useful or most widely accessed?
- Which topics should we target for updates and new development?
- Can we draw conclusions about learning outcomes?

2 DATASET

CVW web log data from our local database were filtered and cleaned by various means to arrive at a dataset that we could analyze further. Two large XSEDE-sponsored meta-topics (Applications of Parallel Computing, and Engineering Parallel Software) that are online renditions of university courses requiring user authentication were removed from this analysis due to their rather different access patterns. All page visits by CVW developers in the Cornell Center for Advanced Computing (CAC) were removed, and we removed to the extent possible those page views that appeared to arise from bots, spiders, web crawlers, and other automated forms of access. This was done through a combination of filtering out IP ranges/addresses known to be associated with such agents and identifying sessions associated with aberrant temporal signatures in their page access patterns (e.g., those with highly repetitive or regular rates of page access). The cleaned dataset analyzed here amounts to roughly 1.3 million page views over a nearly 10-year period (March 2008 through December 2018); counts of unique page visits and total page views are comparable in magnitude to the summary statistics provided by the commercial tracking tool we use. Each page view entry in our local database logs the time of access, an identifier (SessionID) for the user/agent visiting the page, the CVW page visited, and the referring URL, that is, the previous page visited (although in many cases that is not available). The SessionID is either an identifiable username (logged in to some portal or authentication mechanism, e.g., CVW, XSEDE, OAuth) or – in the absence of authentication – a guest identifier that encodes the IP Address of the user/agent.

3 RESULTS

3.1 Named and Guest Users

As noted, some visitors login with an identifiable user name, although the vast majority do not and are thus identified as guests. We are interested whether these two types of users exhibit different behaviors in visiting CVW pages. Filtering our full dataset just for the year 2018, we identified page visits from 671 unique named users and 52177 unique guest visitors. The degree of page visits by these two groups is considerably different, as shown in Figure 1. Shown here is the fraction of users in each group who visited just 1 page, 2-10 pages, 11-100 pages, or 101-1000 pages. While guest users mostly visit one page, with decreasing visits for increasing number of pages, named users engage with the material substantially more. Perhaps this is not so surprising since logging in both represents a form of commitment and offers additional resources to CVW visitors that are not available to guests (e.g., the ability to make and save page notes and bookmarks).

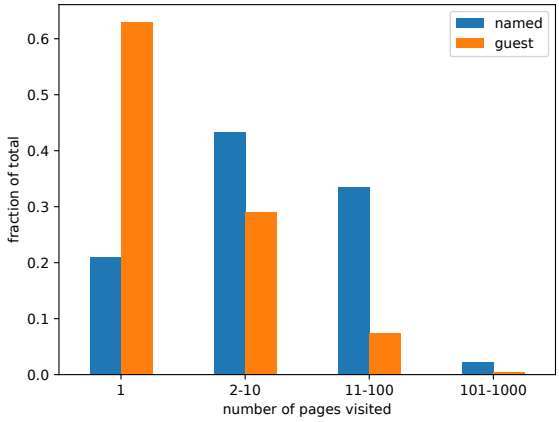

Figure 1: Distribution of the number of CVW pages visited during the year 2018, for named and guest users.

3.2 Visits to a single page

Clearly, many visitors viewed only one page. We explored this data to see if we could learn which pages were visited heavily in this mode, and which pages brought them to the site in the first place.

Of these visits, 4339, or 14% of the 31563 single page visits, were to the CVW home and topic pages; we can guess that seeing the general CVW description or topic list was enough for that searcher to see that the CVW was not the particular content they were looking for. Unfortunately, the referring page for about three quarters of these records were not available. Of the 872 referring pages we do know, 367 were from Coursera [1], 236 were from other locations on the Cornell site, 159 were from search engines (primarily Google), 52 were from partner portals, 33 were from academic and HPC sites, and there were a handful of other sources.

Named / Guest	Page Visits Logged	Unique Pages Visited	Time Span	Topics Visited	Visit Description
Named	7	6	3 min	1	Visited the first six pages of the topic, but not in the order suggested.
Guest	7	1	1 month	1	Visited the MapReduce title page seven distinct dates/times.
Named	10	10	3 min	3	Visited five pages in Parallel I/O, not in order, mostly introductory pages and one content-specific page. Then went to Advanced SLURM, finding the contents list and visiting three specific example pages near the end. Ended on the title page of Large Data Visualization.
Guest	10	1	3 months	0	Only visited the site home page, each visit on a different day.
Named	22	17	21 hours	6	Visited the title page of three different topics. Then visited a few pages in six different topics, usually mid-content pages, returning to the topic list in between. Visits began late one night and continued for about an hour the following day.
Guest	22	4	7 months	4	Visited 22 pages on 22 different days, alternating evenly between the exercise in Introduction to Python and the quiz in OpenMP. There were a few single page visits to GPU, and Parallel Programming Concepts.
Guest	22	22	1 hour	2	Visited 7 pages in order in Intro to Linux, followed by 15 pages in Python for High Performance, randomly after a few introductory pages.
Named	24	22	4 minutes	3	Visited 22 pages in Python for High Performance (in no particular order), Introduction to GPUs (the first three pages in order), and MPI-One-Sided Communication (the title page then two mid-topic pages).
Guest	206	130	8 months	7	Visited a few pages in MPI Point to Point Communication in April, then returned to visit the pages in August for about an hour. MPI Collective Communications was next, visiting and revisiting the pages over the course of 8 hours. Two days later, 45 minutes were spent on MPI One-Sided Communications. The following day, over the course of 7 hours, Parallel IO and Parallel IO Libraries pages were visited and revisited.
Named	234	115	4 days	9	Over 6 minutes, skimmed most of Python for High Performance (PyHP). 45 minutes later, visited 8 pages in Profiling and Debugging (first section). The GPU topic was then visited over 30 minutes, with frequent revisits. After a 30 minute break, Parallel Concepts was skimmed for 10 minutes, mostly in order. Then a return to PyHP, visiting multiple pages over 4 hours, and returning to the same topic 2 days later, frequently moving between pages for about 4 hours. After a 10 minute break, a few pages in Relational Databases, Introduction to C, and GPU were visted, finally returning to PyHP.

Table 1: Details of selected user paths and session lengths.

In terms of topics visited, 1121, or 3.5% of the total, were to the title page of 5 general topics: Linux, C, GPUs, Python, and MATLAB. On this landing page, the visitor sees a general introduction, table of contents, and when the topic was last updated. There is only enough information here to assume the visitor did not see the information they were looking for.

The remainder of the single page visits, 26103 (82.5%), visited a content page (not the landing page) in one of the topics. The top three most-visited single pages were:

- An Introduction to Linux: How does Linux work?: 3857 visits
- MPI Collective Communications: Scatter vs. Scatterv: 815 visits
- How to Make the Most of MIC: Intel Xeon vs. Xeon Phi: 304 visits

From this brief look at single-page visitors, it is interesting to note that the great majority visited a content page in the middle of a topic. If we had also collected a time stamp of when a visitor leaves the page, we might be able to comment on whether their visit was useful. As online training developers, this quantity of single-page visitors encourages us to design each page in a more self-contained fashion, potentially as a stand-alone topic in its own right.

3.3 User visit paths

To get a more in-depth look at the order in which users access content within the CVW, we next looked at four examples from the larger page visit categories:

- 2 Named and 2 Guest visitors who viewed about 10 pages
- 2 Named and 2 Guest visitors who viewed about 25 pages
- 1 Named and 1 Guest visitor who viewed over 200 pages

This small sample size indicates that movement between topics seemed to be driven by returning to topic listings, rather than by following inter-topic links. Movement to a topic is typically by either going to the title page if the previous page visited was a topic listing, or going directly to a content page via search. Once in a

topic, the order of page visits seems to be almost random, and are probably driven by the table of contents list, which is visible on every page. People visiting the largest number of pages seem to be paging back and forth frequently, revisiting the same page multiple times over the course of a session. Further detail on page visit paths and session lengths can be viewed in Table 1.

3.4 Frequency of visits within topics

Why are some pages visited more frequently than others? Is the type of content on a page somehow predictive of that frequency? To answer this question, we looked at the 2018 data for the topic on Vectorization, as this topic is more directly focused on serving our target audience. The first few pages are in standard order: a title page, goals, prerequisites, and the overview. The topic ends with an exercise page followed by a quiz page.

We categorized each page according to several criteria: (1) Order: its numerical order in the page sequence, (2) Contains an Image or Diagram, (3) Contains a Video, (4) Contains Example Code, (5) Is a main Topic heading page, (6) Is a Subtopic page, (7) Contains an Exercise, (8) Contains a Quiz. For every page in the Vectorization tutorial, we constructed a table of such characteristics, and have ordered the rows of the table by how frequently they are visited, from highest to lowest, shown in Figure 2. With the exception of the title page, the number of visits has very little correlation with page order; people are not viewing the online topics as a book that is to be read from beginning to end. The presence of diagrams, video, and example code seem to have no effect on how often the page is visited. However, whether a page is a main topic page – such as "Vector Hardware" – or a Subtopic page – such as "Registers" or "Instruction Sets" – does seem to have an effect, with Subtopic pages being more interesting, presumably to visitors who are choosing where to click next based on reading the table of contents for the topic. The large majority of the most highly visited pages are Subtopic pages, with the exception of the common landing page and the first few introductory pages. From this, we might conclude that people are looking for specific information via text, either by using search engines or the table of contents.

3.5 CVW access patterns over time

Having examined some page access patterns, we are interested how those patterns may have changed over time. We identified, for every unique combination of user and topic during some time period of interest (where the user visited that topic at least once), what fraction of web pages in the topic were visited by that user. Those results are summarized in Figure 3 for the years 2010, 2014 and 2018, where we have plotted the distribution of visit fractions over all user+topic pairs.

One trend evident in Figure 3 is that early on (2010), a greater fraction of users tended to access topics in greater depth: the green 2010 histogram bottoms out once it reaches a fraction of approximately 0.6 and then grows again for larger fractions, suggesting that once users reached that level of involvement in a topic, they tended to continue to work through the remaining material. In contrast, in the later years (2014, 2018), the histograms continue to drop within increasing fraction, indicating that we are continuing to lose visitors as more pages in a topic are considered. We note

Hits	Order	Image or Diagram	Video	Example Code	Topic	Subtopic	Exercise	Quiz
402	1							
233	19			■		■		
122	26	■				■		
113	11			■		■		
105	10	■				■		
95	5		■			■		
87	23	■				■		
76	2							
70	15			■		■		
68	6					■		
66	18					■		
65	8	■				■		
64	20			■		■		
61	4				■			
60	3							
58	7					■		
58	12					■		
51	13	■				■		
50	27						■	
47	9				■			
46	21			■		■		
41	25	■				■		
40	16			■		■		
40	17				■			
36	14				■			
34	28							■
33	24	■				■		
30	22				■			

Figure 2: Visitation rate to pages in Vectorization topic, based on page characteristics.

that in this analysis, we are not tracking the order in which pages in a topic are being accessed, so we cannot really say whether users are working through the content in the order presented, although that could be investigated further by stitching together user paths through a topic. It is also worth noting that even though the distributions for 2014 and 2018 are generally decreasing with fraction of pages visited, that drop off is slower than exponential, since an exponential decay on this semilog plot would appear as a straight line with negative slope, whereas the observed histograms tend to flatten out to some degree. An exponential decay would be consistent with user loss occurring at a constant rate per page. Instead, it would appear that users are remaining within the topic to a greater degree than that implied by the simple exponential decay model.

Another type of access pattern involves the referring pages from which users access CVW pages. These are summarized in Figure 4, where we highlight the top four sources of page referrals, as well as all others. By far the largest source of referral is from another page hosted by the Cornell CAC (which produces the CVW), i.e., indicating that users accessing a CVW page tend to have come from another CVW page (either within that same topic, or a different CVW topic). With the widespread growth in the use of search engines, it could have easily been the case that CVW traffic was dominated by specific referrals from search engines to users looking for specific information. We have not yet determined whether a user's persistence within the CVW over a relatively short time window differs based on how it enters the CVW initially. For example, does a user

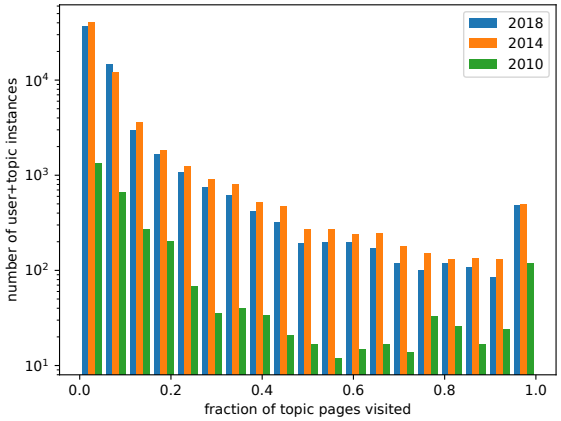

Figure 3: Distribution of the fraction of pages visited by a particular user to a particular topic, within the time period specified. In 2010, topics were being visited in greater depth, with more users visiting a larger fraction of pages as compared to 2014 and 2018.

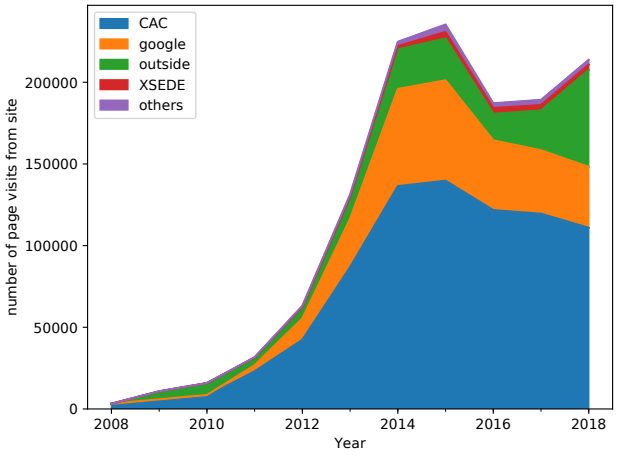

Figure 4: Number of CVW page views originating from different sites, over time.

who connects to the CVW from a link to that material in the XSEDE Training Course Catalog (https://portal.xsede.org/training/course-catalog.org) engage in more subsequent page views internally with CVW/CAC than a user who enters originally via a Google search?

3.6 Patterns in user/topic interactions

We are interested broadly in understanding how users are accessing our topical material, although the raw log data is a complex, interleaving time series of visits from different users to different pages. A useful way of summarizing this data at least partially is through the construction of *user-topic visit matrices*, where each column constitutes a unique user (SessionID), each row reflects a CVW Topic, and the matrix contains information about page visits of each user to each topic. For example, we can compute the total number of visits by a user to a topic, the number of unique pages in the topic visited by the user, or the fraction of pages in the topic visited by the user. Typically, we construct such a matrix within a specified time window. An example of a user-topic visit matrix for the year 2018 is shown in Figure 5, where each matrix entry indicates the fraction of pages in the topic visited by the user (i.e., the same data summarized in Figure 3 above). During that year, there were 46104 unique visits scattered over 50 CVW topics. We can see that some topics are visited by more users and in some cases at a greater depth of coverage (e.g., GPU, LINUX, MPICC), and that overall, the matrix is quite sparse: only approximately 2.6% of possible user-topic visits are observed.

We consider here a pair of related questions based upon the 2018 user-topic visit matrix: (1) the identification of topic similarities and (2) the identification of user similarities. Because there are many more visitors than topics, there are different kinds of questions we would like to ask for each, and hence, different approaches for analyzing each type of data.

Each CVW topic (row in Figure 5) is associated with a high dimensional vector of user visits. We can identify how similar those visit vectors are between any two pairs of topics, and use that pairwise similarity as the basis for topic clustering. Clustering is an unsupervised machine learning method, aiming to find groups of similar items. To address topic clustering, we compute pairwise similarities between topic vectors using a cosine similarity metric: this measures the cosine of the angle between two topic vectors embedded in our 46104-dimensional user space. For two vectors a and b, the angle θ between them has cosine given by $\cos\theta = a \cdot b/(|a||b|)$, where $a \cdot b$ is the dot product of a and b and $|v|$ is the L2-norm of a vector v . The quantity $s = \cos\theta$ is a measure of similarity between two vectors, and $d = 1-s$ is a measure of dissimilarity, such that two identical vectors have angle $\theta = 0$ and dissimilarity $d = 0$. This pairwise dissimilarity matrix then forms the basis of an agglomerative (hierarchical) clustering algorithm, carried out using the Python SciPy library (scipy.cluster.hierarchical and scipy.spatial) [3]. The results of this topic clustering are summarized in Figure 6.

Figure 6 plots the pairwise dissimilarity matrix between topics, where topics have been ordered to reflect the result of the clustering, and clusters have been identified by cutting the cluster tree at a level $d = 0.9$ (dendrogram not shown). For each cluster so defined, the topics comprising that cluster are labeled in a contiguous block and with the same text color: e.g., the set {OPENMP, HYBRID, MPIP2P, MPICC, MPI, MPIONESIDED, MPIADVTOPICS} make up one cluster, the set {R,MATLAB,SCRIPTING, FINTRO, CINTRO,PYHTONINTRO,PYTHON} another, etc. If we cut the tree a bit more stringently, e.g., at $d = 0.8$, some of the larger clusters break up into groups of disjoint subclusters. Topic names listed in black text are isolated in a cluster by themselves.

This topic clustering – based only on user-visit data – reveals various patterns. Some subjects that we would expect to show similar visitation patterns do in fact do so. For example, the MPI-related topics are all clustered together, along with related material on OpenMP and hybrid computing. The topics of performance optimization on advanced architectures (profiling, code optimization, vectorization,

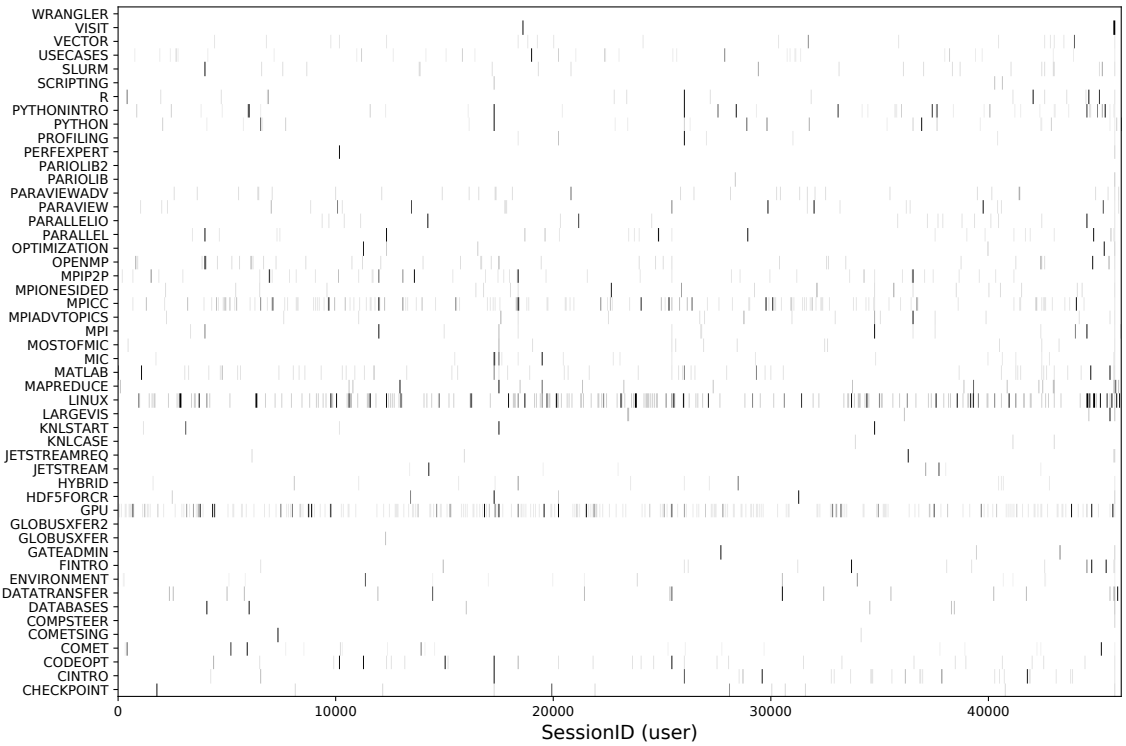

Figure 5: User-topic visit matrix, for the year 2018. For each user and topic, the corresponding entry in the matrix reflects the fraction of topic pages visited by that user during the year 2018. White indicates zero visits, and darker entries reflect larger page visit levels (capped here at 20% for purposes of visualization, i.e, all values larger than 20% are saturated in black).

etc.) are also frequently co-visited, suggesting that users are working through the set of complementary techniques and approaches that comprise that broader goal of performance optimization. Perhaps a bit surprisingly, even the introductory programming material (introductions to R, MATLAB, C, Fortran, Python, etc.) show signs of co-visits; one might expect that a given user would be looking for information about a specific language, not looking for introductory material on several different languages. That being said, the degree of visitation similarity is greater within the clusters targeting more advanced material (e.g., MPI, performance optimization, etc.) than for the introductory programming topics. In addition, there is some degree of co-visitation among some of the clusters that have been highlighted here. The performance optimization cluster (profiling, etc.) shares some amount of co-visitation with another cluster addressing various topics in HPC (VISIT, et al: visualization, data transfer, checkpointing, etc.), and there is a broad grouping of material related to the use of particular computational resources (Jetstream, Comet, KNL nodes, etc.).

Similar to the topic cluster, we can also look for patterns in the user data (i.e., the columns of the user-topic visit matrix). Our interest is not in learning about particular users, but in identifying different patterns of access that might be lurking in the page view data. Rather than clustering, we have taken a related approach in unsupervised machine learning, namely dimensionality reduction. Each user is represented by a 50-dimensional topic-visit vector: dimensionality reduction can be useful both for visualization of high-dimensional data and identification of related entities. Figure 7 represents a projection of the topic visits of all users from 2018 onto a two-dimensional space using t-Distributed Stochastic Neighbor Embedding (t-SNE), a popular dimensionality reduction algorithm [2, 4]. The t-SNE method places items near to each other in this two-dimensional projection if they have similar profiles, i.e., similar topic visit patterns. (The axes of the two-dimensional projection are themselves uninterpretable, unlike some other dimensionality reduction methods such as PCA.) All 46104 users are represented, and are colored orange if they are an authenticated (named) user, or blue if they are a guest identified only through their IP address. The dark blue group in the lower-left quadrant consists entirely of guest users, and will be discussed in further detail below. While t-SNE is not itself a clustering method, it could serve as the basis for one (e.g., by identifying different well-separated groups).

The dark blue cluster of users in Figure 7 was identified visually as somewhat different because of the fact that it is large but contains no authenticated users. (There are many fewer authenticated users than guests, so the fact that there is a cluster with only guests is by itself not so surprising.) Digging into the makeup of the users in this group (3646 in all), we find the following: almost all users in this group visited just 1 or 2 pages (although sometimes repeatedly) in the CVW tutorial on MPI Collective Communications (MPICC) and had no visits to other CVW pages, with roughly 2/3 coming to that

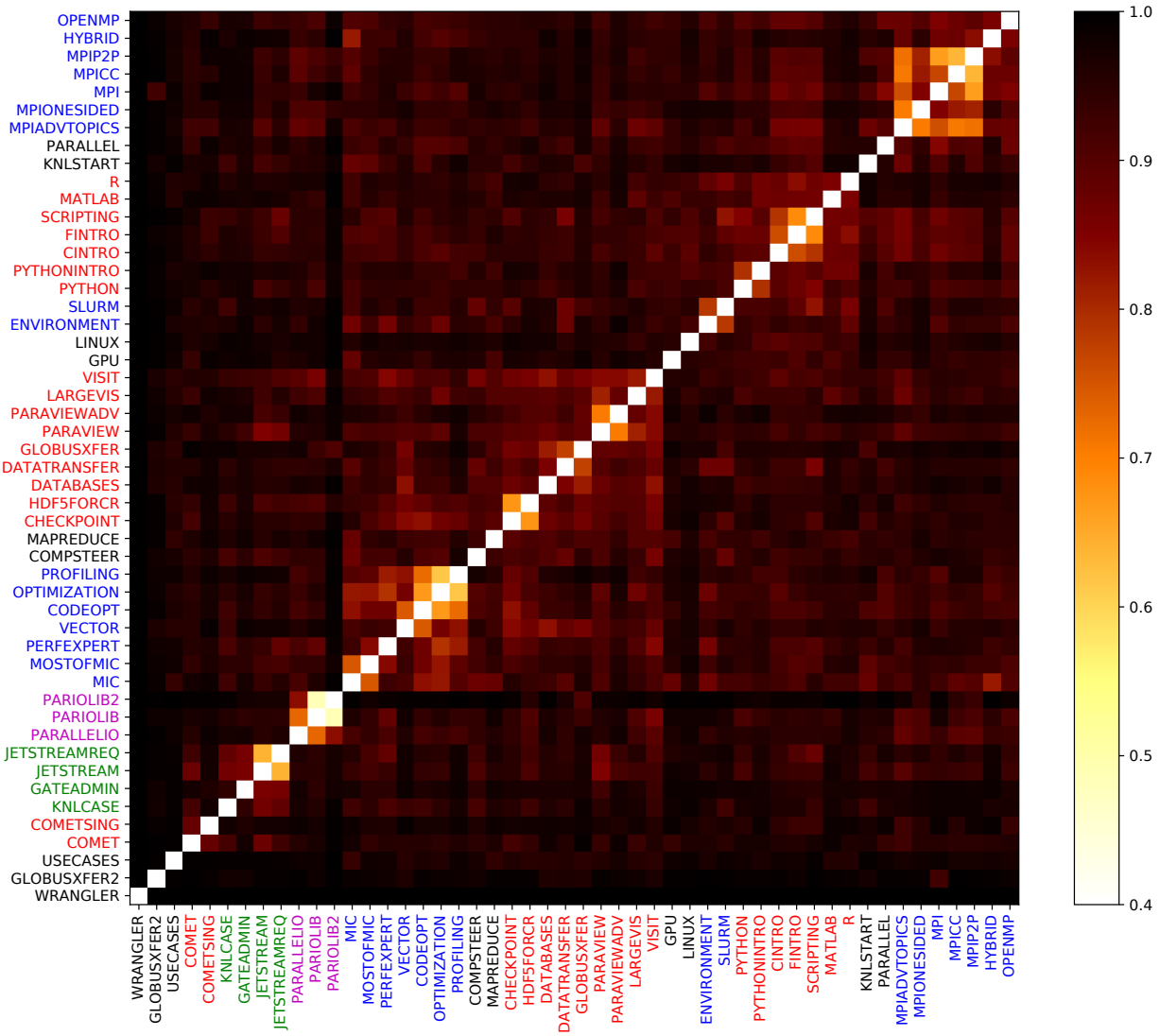

Figure 6: Clustered topic-topic dissimilarity matrix d for the year 2018. Cosine dissimilarity is plotted for each pair of topics: darker entries reflect greater dissimilarity, and lighter ones greater similarity in user visit patterns. Diagonal entries have dissimilarity $d = 0$, i.e., each topic is identical to itself. See text for further description.

CVW page from a web search. There are other users who visited the MPICC topic that are not in this cluster, and are scattered about through the rest of the 2D t-SNE projection, but they tended to visit other topics as well. The dark blue cluster does not represent visits to just a single page in MPICC, but collectively to almost every page in the tutorial. Making sense of all this requires further effort, although it suggests that the dark blue cluster reflects at least partly a subset of users who were looking for particular information, visited a particular CVW page found through a web search, and did not return. (In some cases they did return, but always to the same page.)

4 LESSONS LEARNED, AND TO BE LEARNED

While the CVW tutorials, as noted, have been developed with the intention of presenting coherent and sequential mini-courses on particular topics, it is clear from some of this access data that users do not always engage with the material in this intended manner. This is perhaps not surprising, especially as search engines are dominant and environments such as Stack Exchange provide a mechanism for users to find answers to and tips about specific questions of interest. But it does suggest that as we continue to develop and update our CVW material (a never-ending effort), we do so with an eye to making individual pages or small collections of pages more self-contained, so that relevant points are conveyed when

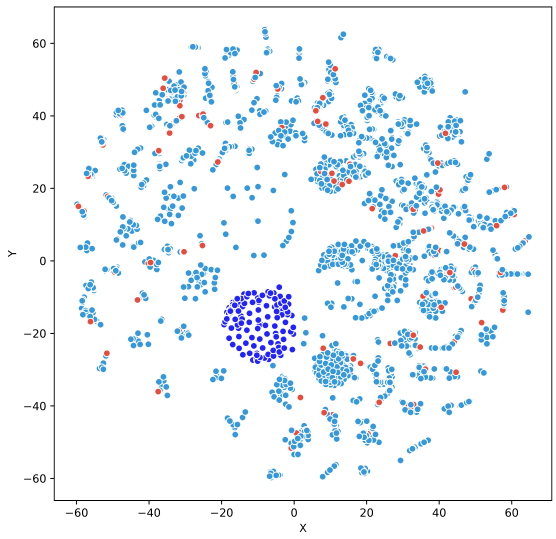

Figure 7: Projection of topic visit patterns for all users during the year 2018, using t-SNE. Each user is represented by a point: authenticated (named) users are orange, and guest users are light blue/dark blue. The dark blue cluster in the lower-left quadrant is discussed further in the text.

necessary without assuming the user has already read through the preceding pages. It also suggests that we might want to prioritize our CVW updates not based just on when a topic was last updated, but based on which topics and/or pages are being most visited through some unexpected routes.

We have only scratched the surface with the use of machine learning (ML) methods to characterize this rich dataset, and there are a variety of additional approaches we could consider. ML methods could be used to develop topic maps to draw in and guide users to material. Some users might find an interactive graphical map of topics, developed via dimensionality reduction techniques of the sort shown here, to be a more compelling gateway to CVW content than the current long and text-heavy list of topics. The topic clustering results suggest that users are finding groups of related material of interest to them. Our Topics page does break out material into thematic groups, but perhaps there are additional cues we can provide to users to facilitate this process. One obvious but more mundane application, only obliquely hinted at near the start of this paper, would be to use ML methods to develop better automated filters to exclude page view activity not associated with interactive human access, as is routinely done in a variety of other web-based and social media sites. And we should note that the user-topic visit matrices shown in Figure 5 are similar to sparse, partially observed user-entity matrices commonly used in other contexts in conjunction with ML algorithms as the basis for recommendation systems (e.g., to suggest items to purchase or movies to watch). But since our topic set is rather small and our users are already apparently locating groups of related content, such a recommendation system is not a priority of ours.

While we have not addressed it in this paper, machine learning could also assist us in developing predictive models of the effectiveness of our training based on the analysis of quiz data. CVW tutorials contain one or more quizzes, typically at the end, that users can take and have autograded. A small fraction of users choose to take these quizzes, but supervised ML methods could be used to predict how well a user will score on a quiz (as a proxy for how well they have learned the material) based on, for example, their patterns of topic access. Such an approach might suggest organizing the material differently in order to promote more effective learning. Supervised ML methods could also be useful in identifying aspects of topic content and organization that are predictive of deeper coverage by users.

Given that the CVW does not require one to log in, it is difficult to determine if learners' educational goals are being met. However, using these data to answer the questions we have posed could identify content gaps that need to be filled, or to support potential learner goals that we infer through usage patterns.

The evaluation of Cornell Virtual Workshop activities is situated within the broader XSEDE project evaluation. During the initial phase of XSEDE, emphasis on formative evaluation activities in the form of 6 month follow up surveys were prioritized, in order to understand the effectiveness and implementation of training activities at all levels. CVW participants were viewed as situated at the beginning of the pipeline of XSEDE training users. The evaluation views this activity as being a gatekeeper to more formal training events requiring larger commitments from trainees such as in person workshops, courses, etc.

In summary, by understanding usage patterns, we can direct development to topics and formats that will better serve the computational science community.

ACKNOWLEDGMENTS

This work was supported by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

REFERENCES

[1] Coursera.org. 2019. Coursera. https://www.coursera.org
[2] Scikit-learn.og. 2019. sklearn.manifold.TSNE. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
[3] Scipy.org. 2019. SciPy. https://www.scipy.org
[4] L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.