Clustering Download Events to Identify Classrooms

Dwight D. McKay
HUBzero
Purdue University
West Lafayette, IN 47907
mckay@purdue.edu

Michael Zentner
HUBzero
Purdue University
West Lafayette, IN 47907
mzentner@purdue.edu

Gerhard Klimeck
Network for Computational Nanotechnology
Purdue University
West Lafayette, IN 47907
gekco@purdue.edu

***Abstract*—The Network for Computational Nanotechnology's (NCN) [1] nanoHUB site uses the HUBzero® platform [2] to offer a variety of content, simulation tools, and collaboration methods to an international community of students, teachers and professionals. Understanding and identifying educational usage of nanoHUB to form communities around nanotechnology education and improve education content is a long term objective of nanoHUB. While certain users log into nanoHUB, providing us with an identity with which to associate their usage, the majority of activity is from unidentified users who download content and come to the site from outside references such as search engine results. This paper describes a method to detect classroom usage from content download events with no additional information, identifying classroom usage by any user of nanoHUB material and providing insights into content usage.**

I. INTRODUCTION

To further NCN's educational mission, the nanoHUB site provides a venue for educational content and simulation tools. We are aware of classes that directly use nanoHUB through the creation of user accounts and the use of simulation tools. However, we suspect there are other classes that use nanoHUB indirectly, making use of content that can be accessed without creating a user account. Are there classes that use the education materials on nanoHUB that we should be reaching out to? Are there groupings of content, implied by actual usage, that we might discover and offer to users as suggestions, or to content contributors as feedback? There is significant potential to connect educators and students with similar interests and improve educational content if classrooms can be detected.

Users of simulation tools are required to create an account before using nanoHUB. Account registration information is generally incomplete and not reflective of actual usage subsequent to the registration. Some information on institutional or University associations can be gathered [3], however registration information on user roles such as undergraduate student, graduate student, researcher or educator are not predictive of actual user behavior and better role categorization can be derived from user behavior analysis. [4] Earlier work [5, 6, 7] used clustering methods to detect where tools were in use in classrooms. Tool contributors then received usage feedback, including the identified classrooms using their tool. This detection method grouped the users and tools into classrooms using several signals, including the user account, close synchrony of simulation tool execution events, and similarity of simulation tool input. Simulation tools are often used in group settings for other events such as tutorials and demonstrations. These groups are also captured by this clustering method.

However, over 95% of the 1.4 million visitors to nanoHUB do not log in. nanoHUB only requires user registration for the use of simulation tools or to join online classes. Visitors who come to nanoHUB to download content (e.g. papers, tutorials, videos, etc.) can do so without registering. These visitors often arrive from search engines or other sites that reference nanoHUB materials. Little data is available about these visitors: only the time when they interacted with nanoHUB, their IP address, and what they viewed or downloaded. Unlike the case of simulation tool usage, there is little synchrony present in download events. Classroom participants may download a homework assignment or other content over a wide span of time, not synchronously with other members of their class. Is it possible to cluster these events in a way that detects class use while ignoring other downloads of the same material?

II. METHODOLOGY

Three features of the download event data were considered for clustering: content item ID, location and time. First, the download events were grouped by the internal identification number of the content item. The initial clustering considers one item at a time. We later looked across content items for possible combinations of items.

A. Location, location, location

The second feature used was location. The event log entries for user downloads record the Internet Protocol (IP) address of the browser the user used to download the content item. There are at least two ways to make use of the IP address as an indicator of location. An IP address to geographic location mapping provided by a service, such as IP2Location [8], can give an approximate geographical location. However, such mappings are often inaccurate [9]. There are numerous sources of inaccuracy: locations are not verified; unvalidated sources such IP address registry information may be used; network features such as virtual private networks may obfuscate location; and proxies and relays can obscure a user's actual IP address. Further, mobile devices are continually assigned new IP addresses as they move.

An alternative way to use IP address space as an indicator of location is to make use of the structure of the IP address space itself. This has several factors to recommend it over mapping an IP address to a location. IP address space is allocated in blocks of addresses to Internet Service Providers (ISPs), businesses and institutions. Schools and Universities typically have blocks of IP address space allocated to their institutions. This allocation methodology groups IP addresses by organization, which aids in detecting clusters of events occurring in classrooms and, in some cases, allows for detection of the specific building in which a class occurred.

Presented at Gateways 2018, University of Texas, Austin, TX, September 25–27, 2018.
https://gateways2018.figshare.com/

The ability to use IP addresses in this way is due to the design of the Internet v4 address space and its allocation to entities requesting addresses [10]. Initially, the IP address space was allocated in fixed-size blocks of three different sizes. The blocks of addresses are contiguous. For example, one of the blocks allocated to Purdue University is 128.46.00.00/16, which contains over 64,000 IP addresses, from 128.46.00.00 to 128.46.255.255. Events that were recorded with IP addresses in this range, such as 128.46.19.99 or 128.46.19.160, are certain to have originated at Purdue University.

Another aspect of the structure of IP address allocation is the implication it has for routing information across the Internet. The Internet is a collection of numerous small networks tied together by devices that route between networks, referred to as “routers”. The IP address is used by these routers to determine which physical connection to use in sending information on to its final destination. Information moves along from router to router from source to destination. By breaking a large space of addresses into contiguous blocks, the task of deciding which direction to move from one router to the next towards the destination is simplified.

Routing also takes place inside assigned address blocks. Network engineers divide institutional networks into sub-networks that align to physical boundaries such as campuses and buildings. This corresponds to the physical placement of interconnecting cabling and local routers. Smaller slices of an allocated block of IP addresses are, in turn, set aside by network engineers for each sub-network. This subdivision of a block of IP addresses by physical boundaries provides a fine-grained location for an individual IP address and a strong assumption that events recorded as originating from numerically "close" IP addresses are originating from physically "close" devices and their users. Given the strength of likelihood that IP addresses related in this way indicate close geographical proximity, we chose to use IP addresses over IP address geolocation as an indication of location in the analysis that follows. We will discuss the problems with this choice later in this paper.

B. *Time*

The third feature used was time. In the work cited earlier [5], close synchronicity provided a signal that a group of tool simulation events was part of a classroom activity. However, downloads of classroom materials are rarely synchronized by activities such as demos or hands-on training, and may occur over several days surrounding a class. Several experiments were run with known classroom events to estimate a reasonable period of time to consider. A further decision was made to discretize the continuous time variable into buckets. In this analysis, the events are gathered into calendar-week buckets.

Taken together, the events for a single content item can be viewed as 2 dimensional heat map. Each cell in the map is colored for the number of events that occurred from that IP address (x axis) during that week of the year (y axis) [Figure 1].

C. *Density-Based Clustering*

Visual inspection of the chosen features, pre-processed as described previously, suggested a density-based clustering algorithm. On visual inspection, it could be seen that the data included a high degree of noise in the form of other, probably unrelated, download events. Density-based clustering algorithms perform well in the presence of noise datums not associated with any cluster. Other clustering algorithms such as K-means were not considered since the expected number of clusters is required, and that is not known in this case. Density-based methods work well in situations where the number of expected clusters is unknown.

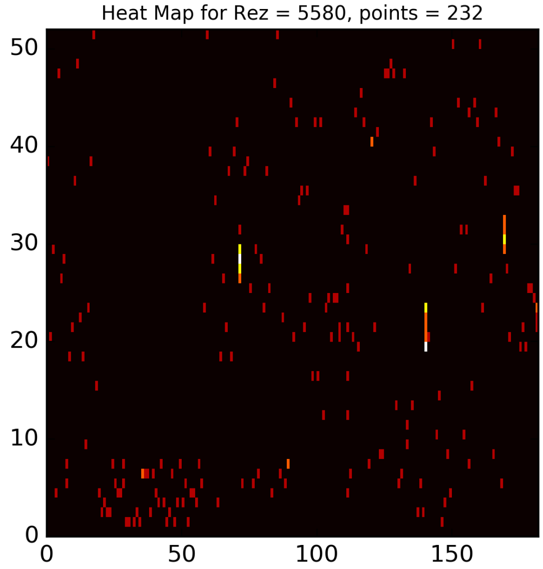

Figure 1, Heatmap of download events. X axis is list of IP addresses in IP address order renumbered from 1 to 185. Y axis is week of the year.

Supervised algorithms were not considered due to the limited amount of labelled training data. While we knew of a few classroom uses of nanoHUB that do not involve user registration or the use of simulation tools prior to this study, such data were sparse and anecdotal. A useable training data set of representative data from classrooms of varying sizes, locations and activity was not available.

This analysis used the DBSCAN algorithm [11] to identify clusters of events in a 2D grid with the numerical IP address as one dimension and the week of the year as the other. DBSCAN takes two parameters, a minimum number of "nearby" points used to declare a point a "core" point and a distance the defines "nearby". A cluster contains one or more “core” points. In this analysis, the definition of "distance" was an ellipse with the long axis aligned along the IP address dimension and the short axis aligned along the week of the year axis. We chose the minimum number of points and the ellipse shape by observation and experimentation on data from known classroom usage situations.

The DBSCAN algorithm proceeds from a random starting point to enumerate any nearby points. When a point is shown to have at least the minimum number of nearby neighboring points, it is declared a "core" point, marked in red [Figure 2: point A]. Points which do not reach this minimum may be part of a cluster, [Figure 2: points B, C], or they may be "noise" points that have no nearby neighbors [Figure 2: point N]. Points in a cluster are mutually density-connected, that is, nearby to at least one other point in the cluster.

If DBSCAN is run on the heat map in figure 1, the following clusters with core, fringe, and noise points are identified [Figure 3]. Clusters are brightly colored, with core points labeled with large dots and fringe points as smaller dots. Groups of events which did not meet the minimum size for a cluster, or are noise points, are colored grey.

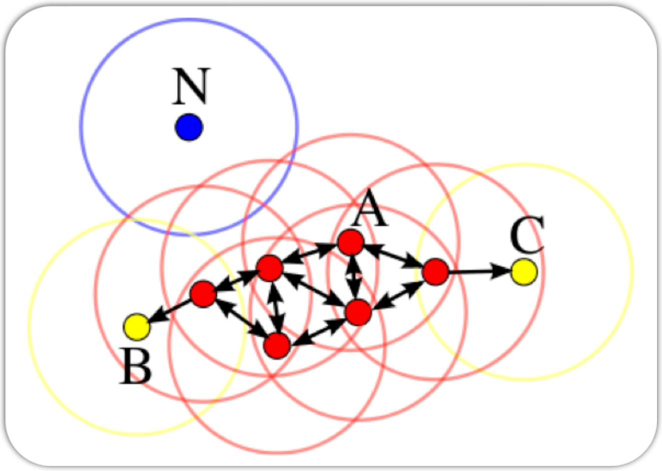

Figure 2, DBSCAN algorithm visualization. Core points are red, fringe points that are part of a cluster are yellow, and noise points are blue.

After clustering with DBSCAN, a secondary analysis of the density of clusters with respect to the season of the year was performed. Content items with more clusters in the spring, fall, or both spring and fall, were categorized as being academic-year classes [Figure 4].

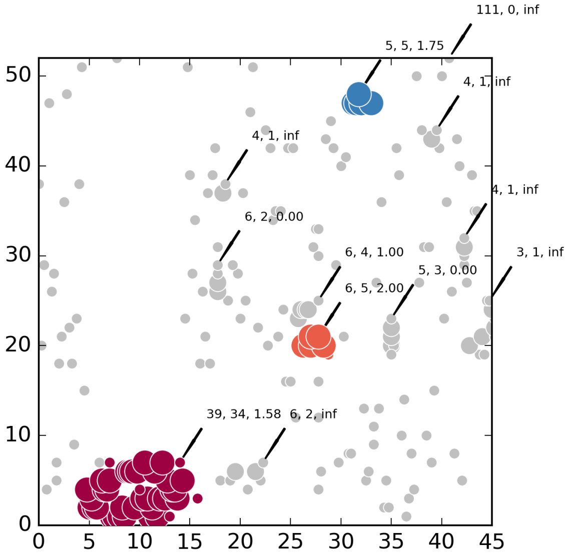

Figure 3, Heatmap processed by DBSCAN with clusters identified. Each cluster is tagged with the total number of points, the number of core points and the ratio of height to width. X axis is renumbered as a result of conversion of IP address space from integer to float for DBSCAN.

III. RESULTS

The pictured heat map [Figure 1], DBSCAN clustering [Figure 3], and seasonality analysis [Figure 4] show a successfully detected classroom. The clustered points in the lower left corner occurred within a single sub-network which, on analysis of the IP addresses is located in an engineering building on the Purdue University campus. Personal conversations with the professor teaching the class confirmed the location, date, time and materials offered. The figure shows the download events for the notes for the lecture presented. For the full year of which this analysis was performed, this method processed 112,290,752 web server event records and detected 2353 classroom clusters which involved 1460 (23.2%) content items, out of a total of 6296 content items on the nanoHUB site at the time the analysis was run.

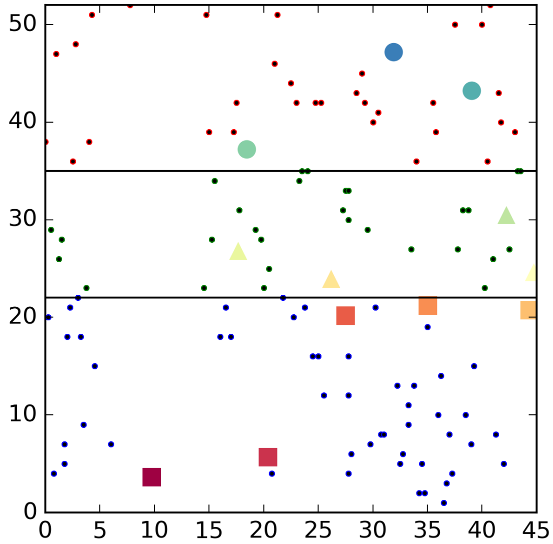

Figure 4, Seasonality test. Identified clusters are labeled with squares, triangles and circles to denote spring, summer and fall semesters. Noise points are color coded blue green and red to denote spring, summer and fall semesters.

Looking across content items, usage patterns that demonstrated an on-going class activity across a semester could be found. Figure 5 shows a series of lecture notes that are downloaded over a series of weeks from the same group of IP addresses in the same building. It is interesting to note the weakening of the cluster as the semester moves on. Also, other materials for the class are not found to cluster, perhaps an indication that they are lightly used or not being used by students in this class in a discernible pattern.

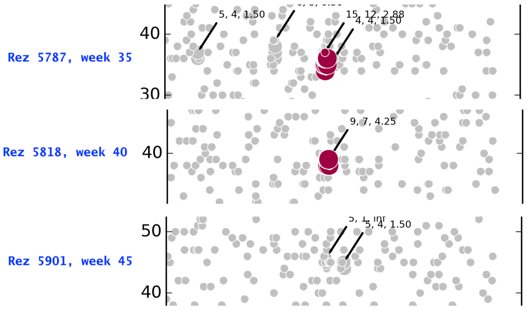

Figure 5, Cross content pattern appearing week after week.

Other phenomena were also detected. A few individual IP addresses with intense usage of a single content item over the entire year were noted. Further investigation showed these IP addresses to be associated with a large number of downloads of video content items. They may be a local caches employed to improve performance for viewing video content.

IV. ISSUES WITH IP ADDRESSES AND LOCATION

While the use of IP address allocation as an indication of location has several advantages as previously noted, it also has several disadvantages which limit the effectiveness of its use. Three disadvantages, in particular, pose the greatest impediment to using IP address as an indication of location: recent changes in way in which the current IP version 4 address space is allocated; the way institutions use their limited address allocation; and finally, the increasing use of mobile devices.

To divide up the remaining IP address space, classless Internet domain routing (CIDR) [12] has been employed, breaking the earlier pre-defined blocks into smaller, odd-length blocks. This increases the likelihood that an organization might have several, discontinuous blocks of IP addresses, rather than a single block of IP addresses. In such a case, that organization's geographically-close devices are less likely to have numerically-close IP addresses.

Increasingly, network address translation (NAT) [13] is being used to provide a large number of addresses for devices within an organization or internet service provider. NAT uses non-routable addresses [10] within an organization and maps the traffic from those addresses to a smaller number of routable addresses as it leaves the organization. A reverse translation is performed as traffic enters the organization. NAT obscures the actual number of distinct users by mapping multiple users onto a single IP address. This results in fewer "points" from which a cluster might be formed. NAT also removes the relationship of IP address to physical location within an organization's facilities, reducing the usefulness of numerical "closeness" as an indication of physical proximity.

As the internet transitions to the IPv6 addressing standard from the current IPv4 addressing standard, the above two problems for our technique are likely to be mitigated. IPv6 address allocation policies being discussed are aimed at providing sufficient addresses to assignees that they would not need to use address space conservation techniques such as NAT and would be able to assign subnets as needed by their physical infrastructure. [14] That would restore the IP-address-to-physical-location property made we make use of here.

Lastly, users who are geographically near one another but are accessing content via mobile devices will not have numerically close IP addresses — in fact, their IP addresses may potentially be widely spread across the IP address space, making them much less likely to be detected as a cluster. In addition, mobile service providers are employing NAT and mobile devices are frequently reassigned IP addresses as they move. In the past year, mobile users accounted for just over 15% of visitors to nanoHUB, up from 14% the previous year. All of this further reduces the accuracy of IP address as an effective means of detecting locality.

At a broader level, the locality of users, their physical closeness to each other, as a signal for an organized activity such as a class, is fundamentally limited to conventional, in-person classes. Modern techniques such as inverted or flipped classrooms, where content is delivered online and discussion is held in person, and on-line classrooms (e.g. MOOCs, etc.), further spread usage out in both time and IP-address space. This spread is likely to be beyond the limits of what a density-based clustering of time and IP address can detect. Other techniques will be needed.

V. FUTURE DIRECTIONS

Getting past the issues around successfully passively detecting classroom usage of downloaded materials will require combining techniques and perhaps adding additional data where available. Perhaps a multi-dimensional DBSCAN-like algorithm that made use of both IP address locality and IP address geographical location mapping might result in a more reliable source of locality information. Additional data could be gathered by requiring more information before a download, either by requesting location information from the browser or directly from the user. One must weigh the advantage of greatly increased locality accuracy against the privacy concerns such queries might raise, and against the potential that such queries would reduce the number of downloads and reduce the effectiveness of nanoHUB as a source of content.

Another approach is to augment the heat map clustering technique with other data sources. Users who have accounts on nanoHUB have already been clustered into classrooms by their simulation tool usage. Combining those two sources of classroom identification might reveal other patterns of usage and increase the confidence in each technique's identification of a classroom. Combining features of specific institutions' schedule such as their academic calendar when a classroom is detected in their IP address space could boost the confidence that a cluster is in fact a classroom at that specific institution.

VI. CONCLUSION

It is possible to identify classroom usage of downloaded content from download event data. However, the significant challenges of passively inferring a user's location from their IP address limit this approach. Improving this technique will involve combining the available data with other signals or with other pattern detection methods to overcome these limitations.

ACKNOWLEDGMENT

The author gratefully acknowledges the Network for Computational Nanotechnology funded by the US National Science Foundation under Grant Nos. EEC-0228390, EEC-1227110, EEC-0228390, EEC-0634750, OCI-0438246, OCI-0832623 and OCI-0721680 for their support of this work. Thanks also to Nathan Denny for editorial assistance and the entire nanoHUB team, especially Lynn Zentner, Krishna Madhaven, and Tanya Faltens for their encouragement and thoughts on presenting this work.

REFERENCES

1. M. Lundstrom, G. Klimeck, "The NCN: Science, Simulation, and Cyber Services," in 2006 IEEE Conference on Emerging Technologies - Nanoelectronics, Singapore, Singapore, January 10-13, 2006, IEEE
2. M. McLennan, R. Kennell, "HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering," Computing in Science and Engineering, Vol. 12, Issue 2, March 18, 2010

3. Krishna Madhavan, Diane Beaudoin, Swaroop Shivarajapura, George Adams III., Gerhard Klimeck, "nanoHUB.org serving over 120,000 users worldwide: it's first cyber-environment assessment" in Nanotechnology (IEEE-NANO), 2010 10th IEEE Conference, Seoul Korea, Aug. 17-20, 2010, Page:90; doi:10.1109/NANO.2010.5697738
4. Omid Nohadani, Jocelyn Dunn, Gerhard Klimeck, "Categorizing Users of Cloud Services" in Service Science, Volume 8 Issue 1, March 2016, In Progress, pp. 59-70; doi:10.1287/serv.2016.0128
5. K. Madhaven, M. Zentner and G. Klimeck, "Learning and Research in the Cloud," Nature Nanotechnology, vol. 8, pp. 786-789, November 2013
6. G. Klimeck, G. Adams, K. Madhavan, N. Denny, M. Zentner, Swaroop S., L. Zentner, D. Beaudoin, M. Jelodar, "Exploring the Impact of nanoHUB.org on Research and Education Users, " September 14, 2011, [Online]. Available: https://nanohub.org/resources/12071. [Accessed: April 29, 2018]
7. Krishna Madhavan, Lynn Zentner, Victoria Farnsworth, Swaroop Shivarajapura, Michael Zentner, Nathan Denny, Gerhard Klimeck, "nanoHUB.org: cloud-based services for nanoscale modeling, simulation, and education" in Nanotechnology Reviews. Volume 2, Issue 1, Pages 107–117, ISSN (Online) 2191-9097, ISSN (Print) 2191-9089, January 2013; doi:10.1515/ntrev-2012-0043
8. ip2location.com, "Traffics Analytics with IP Geolocation," 2018, [Online]. Available: https://www.ip2location.com/articles/traffics-analytics-with-ip-geolocation. [Accessed: April 29, 2018].
9. D. Belson, "Finding Yourself: The Challenges of Accurate IP Geolocation," January 29, 2018 [Online]. Available: https://dyn.com/blog/finding-yourself-the-challenges-of-accurate-ip-geolocation. [Accessed: April 29, 2018]
10. J. Postel, Editor, "RFC: 791, Internet Protocol, DARPA Internet Protocol Specification," September 1981, [Online]. https://tools.ietf.org/html/rfc791. [Accessed: April 19, 2018]
11. M. Ester, H-P Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," in KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, August 2-4, 1996, AAAI Press 1996
12. Y. Rekhter, "RFC: 1518, An Architecture for IP Address Allocation with CIDR," September 1993, [Online]. Available: https://tools.ietf.org/html/rfc1518. [Accessed: April 29, 2018]
13. P. Srisuresh, M. Holdrege, "RFC: 2663, IP Network Address Translator (NAT) Terminology and Considerations," August 1999, [Online]. https://tools.ietf.org/html/rfc2663. [Accessed: April 29, 2018]
14. T. Narten, G. Huston, L. Roberts, "RFC: 6177, IPv6 Address Assignment to End Sites", March 2011, [Online]. https://tools.ietf.org/html/rfc6177. [Accessed: July 11, 2018]