12-2017

Data Clustering Techniques to Identify User Groups and Resource Grouping in nanoHUB

Mugdha Gogte
Purdue University

DATA CLUSTERING TECHNIQUES TO IDENTIFY USER GROUPS AND RESOURCE GROUPING IN NANOHUB

by

Mugdha Gogte

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

December 2017

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF THESIS APPROVAL

Dr. John A. Springer, Chair

Department of Computer and Information Technology

Dr. Michael Zentner

Network for Computational Nanotechnology

Dr. Gerhard Klimeck

Department of Electrical and Computer Engineering

Dr. Eric Dietz

Department of Computer and Information Technology

Approved by:

Dr. Eric T. Matson

Head of the Departmental Graduate Program

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. John A. Springer, Associate Professor in the Department of Computer and Information Technology at Purdue University for pointing me in the right direction with my work and research. I would also like to thank Dr. Michael Zentner, Senior Research Scientist and Analytics Lead at the Network for Computational Nanotechnology (NCN), for his able guidance and providing the required experimental space to explore different scientific approaches.

I am very grateful to Professor Gerhard Klimeck, Director, NCN, for his vision of nanoHUB and Professor Eric Dietz, Department of Computer and Information Technology, for their encouragement to undertake this study. Without their encouragement and involvement, this thesis would not have been possible.

I wish to express my sincere appreciation to key personnel from the team at nanoHUB who have made enormous contributions to this study by sharing their previous work and experience. Special thanks to Dwight McKay, Senior Data Scientist at NCN, for helping me at every step in the study right from the initiation to the end. His astute observations and constructive criticism have been crucial in driving this study.

I greatly appreciate the inputs provided by Nathan Denny, Senior Data Science Engineer at NCN and Gustavo Valencia, Research Assistant at NCN, during the course of my research. I also appreciate the guidance given by Dr. Bruno Ribeiro, Assistant Professor in the Department of Computer and Leonardo Vilela Teixeira, Research Assistant in the Department of Computer Science at Purdue University.

Finally, I would like to give thanks to many peers and fellow students at Purdue University for being sounding boards for new ideas during this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Gogte, Mugdha. MS
Institution: Purdue University
Degree Received: May 2017
Title: Data Clustering Techniques to Identify User Groups and Resource Grouping in nanoHUB
Major Professor: John Springer

With a massive increase in the number of online resources for education and research, it is important to study their usage by target audience comprised mainly of students, educators and researchers. This study explores the application of data clustering techniques on user access data of online science platforms in order to detect user groups and categorize resources with the aim of finding evidence that nanoHUB, the largest science gateway in the field of nanotechnology, aids educational advancement and research. Several algorithms are examined to find the best-suited algorithm for the data set in question. The study uses a two-stage methodology to find classroom like user groups with the help of clustering and further evaluates categorization of the set of resources used by such groups based on a limited set of available features. The techniques used in the methodology are Spatio-Temporal Density Based Scan to detect groups of similar users and Jaccard index to find resource categories by monitoring continued usage of nanoHUB by these groups of users. The resulting user groups and resource sets are evaluated to understand the utility of nanoHUB in a classroom-like group. From the resulting grouping, we can say that spatiotemporal clustering based on a limited number of features reveals group usage patterns of nanoHUB across the globe.

1. INTRODUCTION

The advent of efficient data storage and data delivery through the internet has impacted the field of education and research just like it has impacted many other fields. Apart from structured research and academic education, the scientific community today relies on the power of online platforms for learning, communication, and computation in which science gateways play a big part. Science gateways are remotely-accessible, browser-based platforms that connect researchers from around the globe to powerful, high-performance computing resources. Users of science gateways have access to massive computational resources that would otherwise be expensive and difficult to set-up, use and maintain for individuals. Users also gain access to a number of resources and datasets hosted on the gateways along with the knowledge pool of fellow researchers and a platform to connect with peers. The resources of these online platforms can be broadly classified into two categories. First, resources of shared learning like lectures, research papers and articles. Second, resources that complement experimental analysis such as simulation tools, graphical visualizations, and animations. Science gateways exceed the scope of Massive Online Open Courseware (MOOC) platforms by building a community of contributors and consumers associated with a specific field of science. The purpose of science gateways is much broader than merely hosting lectures. They support virtual experimentation through simulation tools and information sharing with the help of articles and papers.

The vast variety of resources, wide-spread accessibility to global users via the internet, and lack of thorough categorization of resources and target audience make it interesting to study different aspects of the user behavior of science gateways. Some insightful aspects of user behavior are the interactions between different users and association between different resources hosted on the gateways. It is also interesting to study group behavior of users and common resources being used by different groups of users. This study attempts to take an inch forward in answering the research question of detecting classroom-like groups of users, that have a common interest area and tend to access resources that have a common underlying thread with limited available information derived without requiring user log-in. We use the context of nanoHUB, the largest science gateway in the field of nanotechnology, to study such group behavior on science gateways. The term “classroom-like groups” is used to refer, either to actual students participating in an academic course possibly associated with an educational institute or to a group of users with common research interests, possibly associated with a research institute or laboratory. This study is directed towards finding associations between various resources, the interaction between different users and relationships between users and resources on a mixed-media platform like science gateways using data clustering techniques.

1.1 Scope

The motivation behind this study comes from the need to assess and reinforce the usability of large science gateways like nanoHUB by showing evidence of nanoHUB's usage in classroom-like settings across the globe. nanoHUB is the largest science gateway in the field of nanotechnology. It was started in 2002 as part of NSF's National Nanotechnology Initiative and is hosted by Network for Computational Nanotechnology (NCN) (Wilkins-Diehr et al., 2008). Other prominent science gateways are, Cyberinfrastructure for Phylogenetic Research (CIPRES), Cancer Biomedical Informatics Grid (CaBIG), Neuro Science Gateway (NSG) and Linked Environments for Atmospheric Discovery (LEAD).

nanoHUB hosts over 4500 resources including 400 simulation tools at present. According to Google Analytics, there were 32,259 annual users of nanoHUB in 2008, which increased to 217,574 in 2010. In 2016, the number of nanoHUB users rose to 260,543. This increasing trend clearly depicts that online science gateways are becoming more and more popular over time. This gives us the motivation to study the behavior of the annual users in detail. We have chosen to study the user behavioral trends from 2011 to 2015 to analyze a larger dataset as is evident from the upward trend in the number of annual users.

The detection of user groups and resource associations is based on clustering parametrically similar users into groups. The parameters that determine the similarity of users are the proximity of users accessing nanoHUB with respect to time and geographic location. We adopt this approach based on our aim to find classroom-like groups and the assumption that such groups may be associated with an educational or research institute leading to the geographic proximity of users of the group and time proximity of the course or research project in which the users are involved. Further, similar resources are detected by analyzing resources accessed by these sets. The similarity of resources is determined by the overlap of resources accessed by different user groups.

1.2 Assumptions

This study uses geographic coordinates that are derived from the IP address. nanoHUB uses a third-party conversion software, IP2Location, to do so. We assume that all data collection by nanoHUB and conversion of IP address to geolocation is accurate. IP2Location ensures accurate conversion of IP addresses to locations by regularly performing tests of accuracy on reported location and actual location (data accuracy IP2Location, 2017). Another assumption which was mentioned earlier is that groups of users are likely to be associated with an educational or research institute.

1.3 Limitations

The user behavior analysis performed in this study uses IP addresses to distinguish unique users. We cannot distinguish between two users accessing nanoHUB from the same IP address because users can by dynamically assigned different IP addresses in different sessions. Multiple users accessing nanoHUB through a common computer might have the same IP address. If a user accesses nanoHUB via a Virtual Private Network (VPN), the IP address of the VPN will be recorded in the dataset and not the original IP address.

1.4 Delimitations

This study looks at ways of finding user-user, user-resource and resource-resource relationships with the help of data clustering techniques. The features used in the implementation of these clustering techniques are resource ID, time of access and geographic location which can be derived without user log-in requirement. Additional features which require user login like user profile information, ratings and comments are not considered for this study.

1.5 Definitions

GPS coordinates: This study uses the latitude and longitude information derived from IP addresses of users instead of actual geographical coordinates.

DBSCAN: Density Based Spatial Clustering of Applications with Noise. A partitioning based clustering algorithm.

ST- DBSCAN: Spatio-temporal Density Based Spatial Clustering of Applications with Noise. A partitioning based clustering algorithm which is a variation of DBSCAN.

Core Point: A data point with a threshold number of points in scanning radius.

Border Point: A data point with less than a threshold number of points in scanning radius but which lies within the scanning radius of a core point.

Noise: A data point with less than threshold number of points in scanning radius and which does not lie within the scanning radius of a core point.

2. LITERATURE REVIEW

In this chapter, we give an overview of some other science gateways and nanoHUB as a model science gateway to be used as the data source of this study. We will take a closer look at nanoHUB and discuss studies conducted on this online information platform to establish the need for data clustering analysis based on location and time.

2.1 Science Gateways

This section establishes nanoHUB as a prominent science gateway and emphasizes the detailed user analytics being performed by nanoHUB. We give an overview of the online platform in comparison to other similar science gateways. We start by defining science gateways and the purpose of them. Then we look at some well-documented and highly used gateways and compare some defining features of science gateways like the architecture, purpose, target audience, year of origin and number of users.

> A science gateway has to fulfill the following requirements: it has to comply with the specific demands, it needs to support data sharing and multi-user data management, it needs to hide (completely) the complexity of the grid infrastructure. As the end-users, probably don’t have knowledge about grids, and they focus on their own research area, the creation of new domain-specific applications, or the usage of existing ones must be supported within their research area domain. (Balasko et al., 2010).

The target audience for science gateways is a global community of students, researchers, and educators who are members of the scientific community. Science gateways are designed to bridge the knowledge gap for researchers who are experts in their research and domain but have limited knowledge of application development to port applications to distributed systems. Due to this knowledge gap, it is important to assess whether the resources provided by the platform are indeed catering to the needs of the audience. It can be difficult to collect direct feedback from each user. But science gateways, like any other online platform, collect a large set of data originating from each web session. In the case of users who create an online profile, more information volunteered by the user is available for analysis. These data can be utilized by science gateways to indirectly gain feedback through user analytics. nanoHUB has several user analytics related studies published. In the following passages, we compare and contrast nanoHUB with some prominent gateways and later inspect the applications of user analytics for each.

The first example we discuss is Cyberinfrastructure for Phylogenetic Research, also known as CIPRES. It was established in 2009 and is intended for providing a remote parallel computing infrastructure for storing and enabling the use of phylogenetic data for experiments and studies using different genetic datasets. CIPRES is one of the XSEDE (Extreme Science and Engineering Development Environment) group of gateways. It is funded by the National Science Foundation (NSF) as part of the efforts to increase the accessibility of science through online platforms. The architecture of CIPRES consists of a browser-based platform easily accessible globally; a powerful parallel computing framework called the workbench framework that allows submission of analytics jobs which utilize data from a relational database like MySQL used for data storage. This architecture is common for many other science gateways including nanoHUB.

Some other examples of gateways in various fields of science with similar architecture are Linked Environments for Atmospheric Discovery (LEAD), established in 2003, and Neuroscience Gateway (NSG), established in 2012. (Miller et al., 2015) (Towns et al., 2014). LEAD and NSG also receive funding from NSF. Cancer Bioinformatics Grid (caBIG), which was established in 2003 as a part of the TeraGrid group of science gateways, follows a similar architecture (Wilkins-Diehr et al., 2008). The architecture of nanoHUB similarly allows researchers to run analytical computations and simulation jobs through a browser-based front-end using a model based framework and a MySQL relational database at the back-end.

Each of these gateways is focused on the research communities in their respective fields. nanoHUB is the largest gateway in the field of nanotechnology. Apart from the heavy focus on research scientists, nanoHUB also serves students to a large extent by hosting several academic resources. This sets it apart from previously mentioned examples, CIPRES and NSG, which are centered around hosting datasets and tools. nanoHUB was established in 2002 and is among the older gateways to have been established which is heavily accessed to date. nanoHUB started as a part of NSF’s earlier TeraGrid science gateways program which also included LEAD and caBIG. TeraGrid has now evolved into XSEDE group of gateways. XSEDE consists of over thirty gateways among which are more recently founded gateways like CIPRES and NSG (XSEDE Gateways Listing, 2017).

While LEAD and caBIG have not continued onwards into XSEDE, nanoHUB has become a member. The example of caBIG is a notable one in emphasizing the importance of user analytics studies for science gateways. The number of users impacted, in the case of caBIG, has been admittedly low as per the annual report submitted to the scientific advisory board of caBIG in 2011 (NCI report, 2011). According to the NCI report submitted, caBIG started with a vision of providing easily accessible and usable software toolkits to scientists working in the field of cancer bioinformatics. But most users found the software cumbersome and used other commercial systems instead. As the project grew in scope, the system overheads could not outweigh the usefulness of the toolkit. Adequate analysis of system usage could have helped identify weaknesses of the online platform. The user statistics for older gateways like caBIG and LEAD are not widely published which is evidence of the lack of importance given to user analytics studies in these cases.

Newer science gateways lay more emphasis on collecting and assessing user data. The total number of users accessing CIPRES has been studied and published in a paper by Miller and his team (Miller et al., 2016). The total users of CIPRES recorded in the year 2014-2015 were 5,663. NSG has published the total number of jobs run on their online page (NSG metrics, 2017), which was approximately 1000 in 2013. But most of these studies are limited to preliminary metrics. Even in the case of preliminary metrics, the method of calibrating the metrics is unclear from these publications. For a holistic understanding of the impact on users, it is important to categorize users and study independent and group behavior. nanoHUB, on the other hand, has several studies that review user analytics from different perspectives. We give a brief overview of these studies in the next subsection.

nanoHUB emerges to be widely used with a global impact among some comparable gateways examined here. nanoHUB also emerges as a leading science gateway in terms of collecting and analyzing usage metrics. From this overview of other science gateways, we can say that nanoHUB is an exemplary sample set for studying user analytics of science gateways. It helps us establish that the methods used in this study can also be applied to the data obtained from other gateways.

2.2 Previous Studies on nanoHUB

For this study, we look at nanoHUB and prior analysis performed on its user base. The nanoHUB environment serves scientists and students who are primarily nanotechnologists, by hosting a variety of resources for experimentation, publication, teaching, and learning. As mentioned earlier, nanoHUB has a large user base and analyzing prior work gives an understanding of areas and perspectives from which user analysis studies have been done. It has been used by researchers in several other studies related to user categorization and cloud infrastructure. We mention some of the studies here and discuss the information they provide for building this study.

We learn from previous studies that nanoHUB has several characteristics that make it similar to, but not exactly like, a Massive Online Open Course (MOOC) platform. The similarities are that it hosts course material and lectures that are open to all users. It has been used in over 1,000 academic courses by over 20,000 learners. However, it goes beyond being a MOOC platform as it also promotes virtual experimentation and research. It is a stable and reliable online infrastructure that hosts a community of researchers, educators, and students who consume, contribute and share resources (Madhavan et al., 2013).

nanoHUB hosts over 4500 resources available in different formats like lectures, videos, slides, presentations, notes, simulations, animated concept illustrations. Due to its wide user base and variety of online resources, nanoHUB provides a rich data source for the study of the behavior of members of an online scientific community. In 2011, a study was published by a team of researchers at nanoHUB that used network analysis and graph analysis tools to study the social networks formed among researchers and educators who contribute and consume content from nanoHUB (Klimeck et al., 2011). As a result of this study, the development of research networks was quantifiably established by studying collaborations between authors to reveal author sub-groups and communities.

In 2013, Omid Nohadani and his team conducted data science based research on nanoHUB's user base to analyze user behavior. Nohadani analyzed nanoHUB user access data for the categorization of cloud users using a deductive approach for classification. The conclusion of the study was the detection of five distinct groups of users which were detected based on data-intrinsic metrics like frequency, diversity, and intensity. The five categories are namely

Undergraduate, Graduate, Faculty, Non-university and uncategorized users. In this research, Nohadani uses a nested series of zero and infinity norms to overcome relying on hypothesizing the possible outcomes (Nohadani et al., 2013). This conclusion was further solidified in a dissertation by Mingyang Qi on anomaly detection in user categories using principal component analysis (Qi, 2014).

According to these studies, users of nanoHUB can be broadly classified into two categories - individual users and group users (Nohadani et al., 2013). We use this conclusion in our study on analyzing the behavior of group users by monitoring continued engagement of groups of similar users and the resources they use. In this study, data clustering techniques are used to identify the groups of similar users. Unlike previous studies, this study is based on features obtained from user access data instead of the distribution of length of user sessions. The parameters used for clustering users into groups are location, resources accessed and time of access.

The behavior of users of simulation tools was analyzed by Madhavan, Zentner, and Klimeck (Madhavan et al., 2013). An interesting aspect of this study is that all users of simulation tools are required to log-in to nanoHUB. This gives analysts access to profile information about each user. However, users are not required to log-in to access other type of resources.

A key distinction of our study is that it uses information about users that can be derived without logging in, i.e., the IP address, identifier of the resources accessed and time of access. The study maximizes the data used and is inclusive of resources that do not require log-in. In the next section, we discuss different data clustering techniques used in user segmentation and cohort detection as well as best suited techniques for the choice of parameters mentioned above.

2.3 Summary

In this chapter, we discussed the nature of science gateways, their target audience and the type of user analytics performed to find out more about users of the nanoHUB science gateway. Through the literature review of prior user analytics of nanoHUB, we establish the need for using data clustering techniques to study group behavior by detecting groups of users and resource associations.

3. OVERVIEW OF ALGORITHMS

The objective of this chapter is to, first, give an overview of data clustering techniques and their classification and secondly, to provide a justification of the techniques chosen for implementation in following chapters. We describe the basis of classification of clustering algorithms, shortfalls of each class of clustering algorithms and the techniques that are ideal for the type of data used in this study. A detailed description of the working of selected techniques is also provided.

3.1 Data Clustering Techniques

Clustering is a technique of grouping data points based on similarity or dissimilarity. It is an unsupervised learning method, meaning that there are no predefined classes or class labels. There are several types of clustering methods, and there are a few different ways in which these methods can be categorized. In one classification, they can be broadly categorized as hierarchical and partitioning techniques. Hierarchical clustering techniques divide the data space into smaller subsections with each step. An example of this technique is a decision tree which divides the sample space into subsets with each level of the tree. A major disadvantage of hierarchical techniques is the ambiguity in deciding the termination cut-off which determines at what stage to stop clustering. Due to this disadvantage, hierarchical clustering techniques are not ideal for the dataset used in this study.

On the other hand, partitioning clustering techniques decompose a dataset into disjoint subsets. Some popular examples of this type of clustering are k-means and k-medoids. While k-means is the most popular partitioning clustering technique, it is not ideal for the dataset used in this study as it requires knowing or determining the value of “k” where “k” is the number of expected clusters. In the context of this study, “k” is the number of user groups for each resource on nanoHUB. Since there are over 4500 resources, it is not ideal to find a value of “k” for each resource. Therefore, we look at other partitioning clustering algorithms which are independent of the value of “k” such as DBSCAN which determines the number of groups based on the density of the sample space. This is explained in detail in the next subsection.

3.2 DBSCAN

DBSCAN is an acronym for Density Based Spatial Clustering of Applications with Noise. It was first published by Martin Ester, Hans-Peter Kriegel, Jrg Sander, Xiaowei Xu in the second International Conference on Knowledge Discovery and Data Mining in 1996. The main principle of density-based clustering is to recognize a group of points as a cluster if the density of points is considerably greater than the area outside the cluster. More importantly, the density of the areas considered as noise is significantly less than the density of points in the clusters (Ester et al., 1996).

A brief explanation of the working of DBSCAN is as follows. DBSCAN initializes at a random point (S) in the dataset and scans the radius defined in the input parameters as “Eps,” which is a distance measure like Euclidean distance, Manhattan distance, etc. If the algorithm finds a threshold number of points or more, defined by another input parameter “MinPts,” then the algorithm classifies S as a core point, and all the scanned points as neighbors of S. DBSCAN proceeds to iteratively perform the same steps for all neighbors of S.

If MinPts are not found in the radius of Eps, the point is simply not classified as a core point, and DBSCAN moves on to the next point in the dataset. For detailed definitions of the algorithm, we refer to the original paper by Ester et al. (Ester et al., 1996).

A drawback of this method is that it only allows the definition of one scanning radius “Eps,” meaning that data can only be scanned as spherical clusters regardless of the number of dimensions of the dataset. This drawback can be circumvented either by rescaling various dimensions to accommodate a single scanning radius for cluster detection or using multiple scanning radii as described in a variation of DBSCAN in the following section. The different dimensions used in this study are time and distance which can be difficult to rescale because of the two types of measures, time and distance, are difficult to compare as they are of very different nature. Moreover, the proximity we are interested in with respect to time (1 week out of 52 weeks), widely varies from the proximity of interest with respect to distance (50 km diameter). Therefore, we consider the implementation of ST-DBSCAN which allows giving multiple radii as input as an alternative approach.

3.3 ST-DBSCAN

Spatio-Temporal Density Based Spatial Clustering of Applications with Noise (ST-DBSCAN) is a variation of the DBSCAN algorithm. It was proposed by Birant and Kut in 2006 (Birant Kut, 2006). It overcomes the main drawback of DBSCAN by allowing the definition of one scanning radius (Eps) for one dimension. Therefore, an n-dimensional data set will have n distinct scanning radii (Eps1, Eps2, ..., Epsn).

The working of ST-DBSCAN is similar to DBSCAN. The algorithm gets initialized to a random data point (S) in the dataset and scans a radius of Eps1 in one dimension to check for MinPts. If MinPts are found in the first dimension, then the algorithm scans a radius of Eps2 in the second dimension. For a detailed explanation, we refer to the original paper (Birant & Kut, 2006).

This is an ideal solution for the dataset being analyzed in this study which has three dimensions. Two of these dimensions, Latitude, and Longitude, based on which we determine the distance between two locations. The distance measure can be clustered using one radius, Eps1. The third dimension is the time dimension, which can be clustered using a separate scanning radius Eps2. The following three-dimensional figure, Figure 2.1, illustrates the application of ST-DBSCAN on nanoHUB user access data. Each cluster is depicted by a unique color. The data points that do not fall into any cluster are shown in black. We will describe the implementation of this algorithm for our dataset in the next chapter.

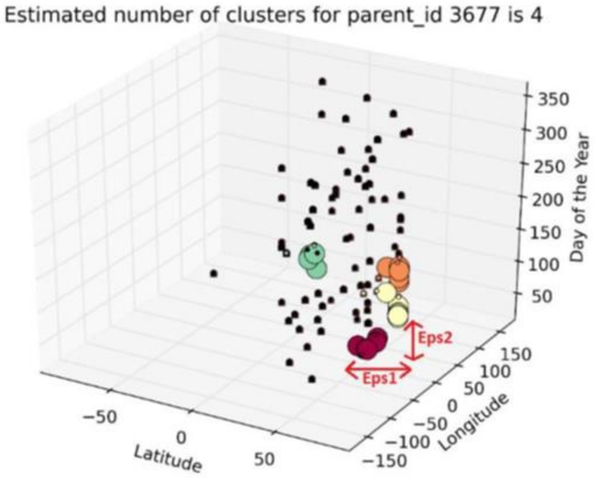


Figure 3.1. Depiction of ST-DBSCAN using two scanning radii, Eps1 for distance dimension and Eps2 for time dimension.

3.4 Tribeflow

Apart from the above-mentioned classes of clustering techniques: hierarchical and partitioning, there is another group of clustering algorithms based on the distribution of the sample set and probabilistic models. An example of this category of clustering techniques is a recently proposed algorithm called Tribeflow which can be a suitable fit for the type of dataset used in this study. According to the introductory paper on Tribeflow (Figueiredo et al., 2016), it was successfully implemented for other online platforms on the basis of features that are similar to the features used in this study. The objective of Tribeflow is to build a recommendation and trajectory prediction system based on user access data of online platforms. This algorithm takes user access records as input. Each record contains three key features, the IP address of each user accessing the website, time of access and the resource accessed. Each user, identified by the IP address, is treated as a random surfer, traversing from one resource to the other in a random walk. The Tribeflow algorithm builds several latent environments based on the probability of the random walks resulting from the input data. Based on the probability distribution of a user picking one latent environment over others, Tribeflow makes a prediction about the next “n” resources that will be accessed by a particular user, where “n” is an integer. We will describe the implementation and shortcoming in the implementation of this algorithm in the following chapter.

3.5 Summary

This chapter gives an overview of different types of clustering techniques – hierarchical, partitioning and distribution based. We focus on density-based and distribution based clustering techniques with a detailed explanation of density-based techniques, DBSCAN and ST-DBSCAN, and distribution based technique – Tribeflow. The chapter serves as an explanation of the choice of the methodology used in this study.

4. FRAMEWORK AND METHODOLOGY

In this chapter, we describe the details of implementing the selected clustering technique in finding user groups and resource associations for nanoHUB's user data. The aim of the methodology is to first detect groups of similar users and based on the detected groups, find the association between different resources being accessed by these groups.

Science gateways are an amalgamation of various types of resources. As mentioned earlier, nanoHUB hosts a large number of resources and has a wide audience across the globe. nanoHUB's resources are spread across multiple disciplines of nanotechnology like nanotechnology in biotechnology, nanomedicine, nanophotonics to name a few. It is interesting to find links between resources which may or may not have explicit associations. For instance, resources related to electron microscopy are associated with both nano-biotechnology and nanoelectronics but are not always categorized under a common topic. We hope to uncover such underlying association through this study to not only show nanoHUB's application in classrooms but also discover seemingly unrelated connections between different resources. The approach used in this study can be summarized into two stages. In the first stage, ST-DBSCAN is used to cluster the users based on three dimensions - time, geographic location and resources accessed. The clusters found as a result of the first stage depict different user groups for a period of one year. In the second stage, we collapse the time dimension and compare common resources accessed by user groups at each location. Collapsing the time dimension allows us to compare different resources accessed by the same group in a year, such as a series of lectures that gets accessed by a user group every consecutive week. Each step of the above-mentioned process is described in detail in the following sections.

4.1 Data source

nanoHUB is a platform open to registered and non-registered users. This means that users need not necessarily log in to access resources on nanoHUB. Therefore, data related to user ID, name and email ID of the users are not available for the users who do not log in before utilizing nanoHUB's infrastructure. The best features to track users and user behavior in this scenario are IP address, time of access, and resources accessed.

Data from web log sessions of each user of nanoHUB is stored in the nanoHUB's central database, which is a relational database. The database is used to store several attributes received from the web logs of which the key attributes relevant to this study are the IP address and the time of access of every URL on nanoHUB. This information is extracted from the nanoHUB database into five comma-separated value (CSV) files, one for each year from 2011 to 2015, which are used as the data source for this study. Each file has 858,366 records on an average These key attributes can be transformed suitably into features that can be given as input to the selected clustering algorithm.

4.2 Data Transformation

In this section, we explain how we have handled three aspects of the knowledge discovery (Fayyad et al., 1996) process in our study: preprocessing, cleaning and feature extraction. Relevant attributes from the weblog - IP address, time of access and the resource identifier are extracted and translated into GPS coordinates and day of the year respectively. The resource identifier is extracted without any transformation[1] . Table 4.1 shows the attributes before transformation and table 4.2 shows the attributes after transformation. The IP address is transformed to the geographic location using a third-party mapping service IP2Location which is updated every month. The transformation process also translates the time of access to a more abstract unit of time, Day of Year. Day of year refers to the count of the day from January 1 of that year.

Table 4.1. Original Data

IP Address	**Resource ID**	**Timestamp**
128.211.253.139	5469	2015-04-06 00:02:23
69.180.129.110	5469	2015-01-06 00:02:23
69.180.129.110	5544	2015-01-23 00:02:52
69.180.129.110	5469	2015-01-27 00:03:10
70.145.178.140	5544	2015-05-06 00:03:20
96.219.203.100	5845	2015-04-23 00:03:20
155.69.128.186	5544	2015-04-16 00:03:22

[1] nanoHUB follows a hierarchy of grouping similar resources and resources IDs under a single parent ID. An approach to cluster user accesses for each parent ID was tested, but did not yield accurate results. The likely reason for unfavorable results is that in the case of a special category of resources on nanoHUB known as collections the grouping of a wide range of unrelated resources under a single parent ID.

Table 4.2. Transformed Data

Latitude	Longitude	Resource ID	Day of the Year
37.567	127	5469	97
19.0158	72.8599	5469	18
19.0158	72.8599	5544	23
19.0158	72.8599	5469	27
28.6329	77.2195	5544	125
3.033	101.717	5845	118
3.167	101.7	5544	106

Raw data pose several challenges like missing values and null values. In our study, the IP addresses that cannot be mapped to any GPS location are stored as null values. For the scope of this study, we disregard such records as they make up less than three percent of over eight hundred thousand records available to us. Of the 810,470 records considered for years 2014 and 2015, 2.99% are null values. We also do not take into consideration the URLs that cannot be linked to a unique resource identifier number. The data for the year 2015 after processing has 12,129 unique IP addresses and 3,410 unique resource identifier numbers. The data for 2014 has 16,593 unique IP addresses and 3,605 unique resource identifier numbers.

4.3 Study Design

In this section, we describe the implementation of the selected clustering model in the context of this study. As concluded from the survey of literature in the previous chapter, ST-DBSCAN is the clustering model found to be suitable for our dataset. The aim of the clustering technique is to find groups of similar users based on three features: time, location and resources accessed. The idea of using these three mentioned features stems from the assumption that users of nanoHUB are associated with an educational institute or research institute and are likely located near such institutes.

ST-DBSCAN is an unsupervised learning algorithm that is well suited for this dataset as it does not require the specification of an expected number of clusters. This algorithm decides the number of clusters based on the density of the sample space. It takes three input parameters for the data set used in this study. The parameters are, the scanning radius in distance dimension

(Eps1), the scanning radius in time dimension (Eps2) and the minimum density of points (MinPts) which needs to be met for a point to be classified as a core point of the cluster. Using these parameters, ST-DBSCAN is performed for each resource ID individually. The result of each ST-DBSCAN performed for individual resource IDs is a set of user groups that have accessed the respective resource ID. In the next stage, which is discussed later in this chapter, the user clusters for multiple resource IDs are compared. In the following paragraphs, we discuss the basis for selecting values of input parameters in this study and the format of the input and output for ST-DBSCAN.

4.3.1 Parameters

The values set for these parameters in our study are also based on the assumption that nanoHUB is mostly accessed by users centered around an educational or research institute. The value of the geographic scanning radius, Eps1, is based on the average distance traveled per person for each household as recorded in the National Household Travel Survey (NHTS) in 2009 (Santos A. et al., 2009). Eps2 depicts the scanning radius in the time dimension. Given the assumption that users are associated with academic courses or research projects, we set Eps2 to 7 days based on the interval commonly used for class assignment in courses. In our methodology, we have used the above assumptions as a baseline, and fine tuning of the hyperparameter values is done using a randomly selected sample set. The error is calibrated through manual analysis of cluster purity in which we check whether the data points clustered as one group of users is a credible grouping. The value of Eps1, the geographic scanning radius, was set to 50 kilometers based on the NHTS survey and results of the control set. The value of Eps2, the time radius is set to 7 days.

To understand the scale of the distance scanning radius value used in this study, we compare the distances with equivalent geographic co-ordinate measures. The distance between latitudes and longitudes varies from the equator to the poles. The distance between latitudes does not vary much and is equal to 111.132 kilometers approximately. The distance between two longitudes varies greatly from 0 kilometers at the poles to 111.32 km at the equator. At 45° N, the distance between two longitudes is 78.847 kilometers. 95% data points lie between 60° N and 60° S latitudes, where the approximate distance between longitudes is 55 kilometers. To give a real-time example, the main campus of Purdue University (West Lafayette) and a remote campus of Purdue University (Calumet) are separated by a distance of 131.94 kilometers. This example helps in demonstrating the judicious separation between user clusters facilitated by the distance scanning radius selected in this study.

4.3.2 Input

The source data from the CSV files, after transformation, is given as input to the clustering algorithm. Each data point is the record of a user and contains information about the geographic location of the user, IP address, the resource number accessed by the user and timestamp. The geographic location is a pair of coordinates given by the latitude and longitude. The distance between locations is calculated using Haversine distance given by the following equation. Haversine distance is a widely-used measure of geographic distance equal to the greater circle distance between two points on a sphere. The Haversine equation assumes the earth to be a perfect sphere and θ is the angle between the lines connecting the two locations to the center of the sphere.

$$\text{hav}\ (\theta) = \sin^2\ (\theta\ /\ 2) = (1 - \cos\ (\theta))\ /\ 2 \qquad \text{Eqn 4.1}$$

The calibration of Haversine distance is implemented with the help of an existing package in the scikit-learn library of Python as seen in the code specified in the appendix.

4.3.3 Output

The output of each iteration is a set of clusters of users for each resource identifier. They are stored as CSV files. The resulting files store the IP address, resource ID and GPS coordinates of the records from the source file that were found to be a part of a cluster. A new attribute, cluster ID, is also added to each record. The cluster ID is a number used to identify to which cluster each record belongs and is common for all data points belonging to the same cluster. The resulting set of files is used as input for the second stage in our methodology, which is the comparison of user groups of different identifiers.

4.4 Comparing Clusters

Once we have the results from our clustering algorithm, we perform a one-to-one comparison of clusters of each pair of resources. A one-to-one comparison will help us identify which resources can be grouped together into categories. We will now discuss the approach used in this study for one-to-one comparison of user groups of different resources.

4.4.1 Jaccard Index

Jaccard index is a statistical measure used to find the similarity or diversity of two sets. It is the ratio of all the intersecting elements of the two sets to the total number of elements in both sets as illustrated in Fig. 4.1. For more details, we refer to Das and his team's work in Google news personalization for an example of this common technique (Das et al., 2007).

$$J\ (A,B) = |A \cap B| / \ |A \cup B| \qquad \text{Eqn 4.2}$$

$$= |A \cap B| / |A\ | + |B| - |A \cap B| \qquad \text{Eqn 4.3}$$

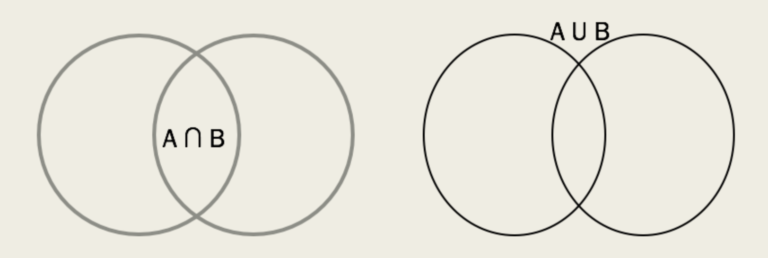


Figure 4.1. Intersection of two sets and union of two sets

In our study, we implement this using a similarity matrix in which the columns represent all unique locations found in user clustering and rows represent unique resources, as illustrated in Table 4.3. The cells in the matrix are assigned a value of 1 if the respective resource ID has been accessed at the location represented by the corresponding column and the cell is assigned a value of 0 if otherwise. Each pair of rows is compared to find interesting elements. Jaccard index is computed for each pair of resources and the pairs of resources with an index of greater than 0.5 are considered to be similar. A similar usage of the Jaccard coefficient can be observed in the paper that proposes a procedure to construct a social network based on web search engine by Kubota and his team (Kubota, 2014).

Table 4.3. Similarity Matrix

	Location 1	Location 2	Location 3	…	Location n
Resource ID 1	1	0	0	…	1
Resource ID 2	1	0	0	…	1
Resource ID 3	0	1	0	…	0

The figures below show a visual representation of clusters of two resources being considered in Jaccard similarity. Figure 4.2 is a three-dimensional plot the locations and time of access of user groups of resource ID 192, depicted by blue data points, and resource ID 8, depicted by red data points. When we collapse the time dimension, as seen in Figure 4.3, we notice that there is an overlap of several data points of the two resources. Since resources 192 and 8 are being accessed by user groups in the same geographic locations, they are considered similar. Figures 4.5 and 4.5 depict the latitude vs. time and longitude vs. time plots of resources 192 and 8. Figures 4.5 and 4.5 convey information about pattern resource accesses as time progresses. The shading of the data points indicates depth with respect to the three-dimensional plot. Darker data points are closer to the point of perception and vice-versa.

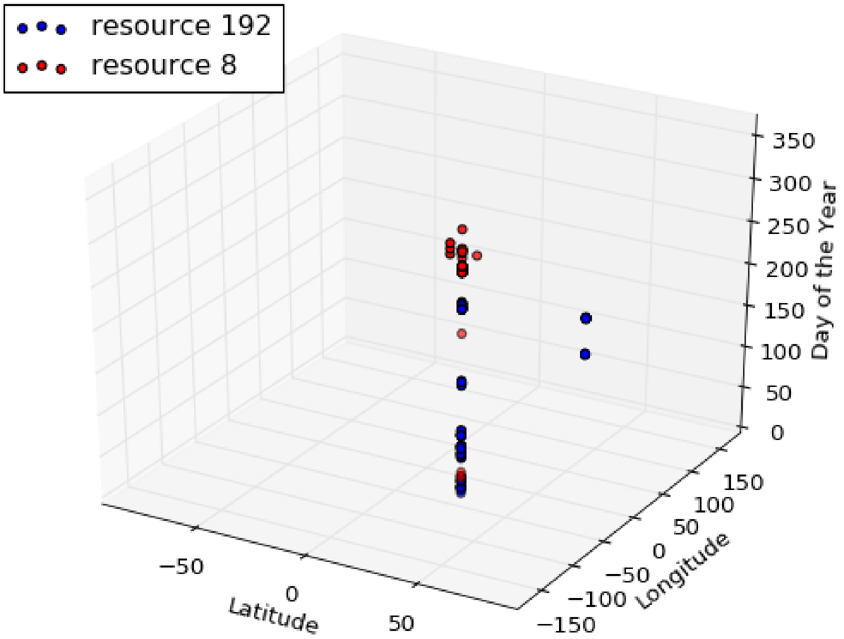


Figure 4.2 Three-dimensional plot of group user accesses of resource ID 192 (blue) and resource ID 8 (red)

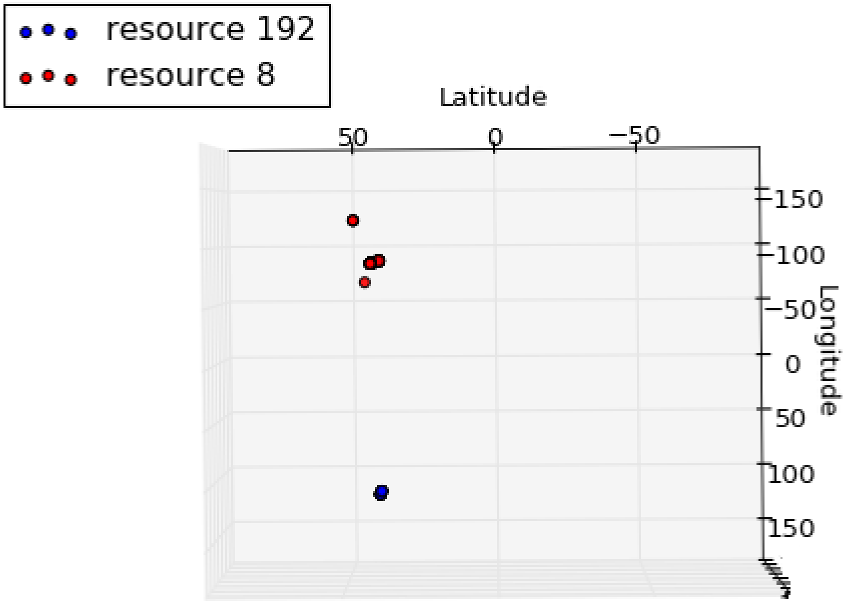


Figure 4.3 Latitude vs. Longitude plot of group user accesses of resource ID 192 (blue) and resource ID 8 (red)

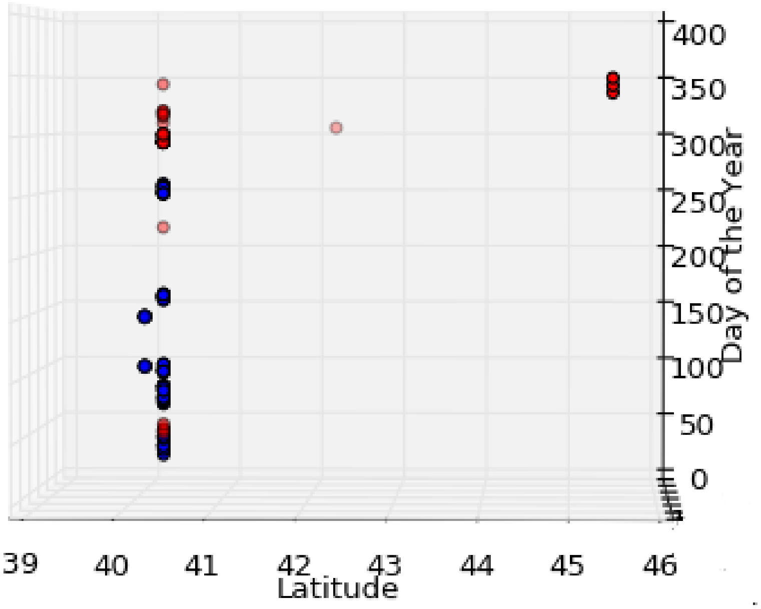


Figure 4.4 Latitude vs. Time plot of group user accesses of resource ID 192 (blue) and resource ID 8 (red). It shows the progressive annual access of the two resources.

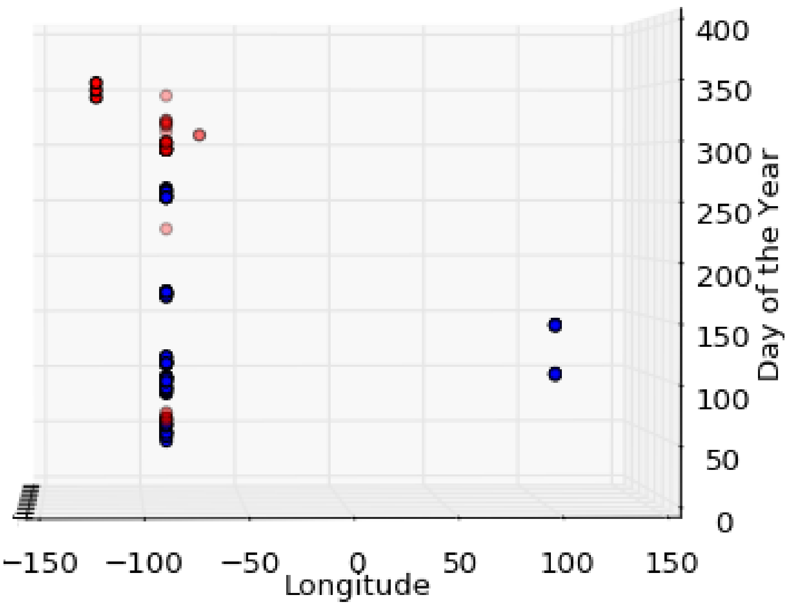


Figure 4.5 Longitude vs. Time plot of group user accesses of resource ID 192 (blue) and resource ID 8 (red)

4.5 Tribeflow

In the previous sections, we discussed the implementation of the methodology which was used to obtain results in this study. In this section, we will talk about the implementation of Tribeflow. Although Tribeflow is not the primary focus of this study, we give a brief overview of its implementation and useful lessons learned that can be used as a foundation for later studies. As mentioned in the previous chapter, Tribeflow is intended to predict the resources that a particular user might access next. The algorithm works by building a set of latent environments based on input access records and find the probability of a user choosing each latent environment. Tribeflow takes user access records as input, where each record consists of the IP address of the user, time of access and the resource identifier. Four parameters need to be defined during the implementation of Tribeflow. The first three parameters are used to define the order of the columns in the input file. The fourth parameter, m, is used to define the number of expected latent environments. In our study, we implemented the algorithm for the data of years 2014 and 2015 for varying values of the number of expected latent environments as there is no recommended technique to determine "m." Value of "m" was set to 15, 50, 100 and 200 in different iterations of the experiment. The algorithm takes “m” as a baseline and dynamically determines the actual value which is close to the given value.

The generated output file is in the form of a “a” x “b” matrix where “a” is the number of users, and “b” is the number of resources. Each cell, C_{ab}, contains the probability of a^{th} user accessing the b^{th} resource in user’s next step. Of the given values of m, 15 was found to be the most optimal as it gives the highest probability values for a single resource for each user. However, for all given parameters values of m, none of the resources were assigned a probability which was significantly higher than all other resources. Due to this shortcoming, it is difficult to process the results of Tribeflow, with the current set of parameters, to get a meaningful outcome. Tribeflow was introduced in 2016 and has very few papers published except for the introductory paper by Figueiredo, Ribeiro, and team. With an availability of more literature and case studies in the future, implementation of this algorithm can be made more effective for nanoHUB’s data.

5. RESULTS

The objective of the methodology used in this study is to find classroom like groups that use nanoHUB and find common resources that each group uses. In this chapter, we present the results of two stages of the process explained in the previous chapter. The first stage results are user groups obtained from clustering users on the basis of the time of access and location using ST-DBSCAN. The user access data are derived from the preprocessing of web logs of nanoHUB in the year 2014 which is chosen as the representative year to illustrate all results in this chapter. Next, we look at the results of the Jaccard similarity to detect resource categories based on user clustering results. We also look at some examples of resource categorization using Jaccard index. Before presenting the results, we look at the distribution of user access data before and after clustering.

5.1.Data Distribution

Figure 5.1 shows the distribution of total number of accesses for each resource in year 2014. The number of accesses ranges from 1 to 17,435. Figure 5.2 shows the distribution of accesses per resource by users after clustering. The total number of records is filtered down to 44,574 after clustering as against the initial 427,357 records. This means that approximately 10.43 % (44,574 out of 427,357) of the total accesses were made by users belonging to groups that show collective behavior. This is the fraction of data which will be processed further in the second stage to identify resource associations. The total accesses for 2013, 2012 and 2011 are 516881, 890630 and 2058925 which gets filtered down to 67360, 229476, 642423.

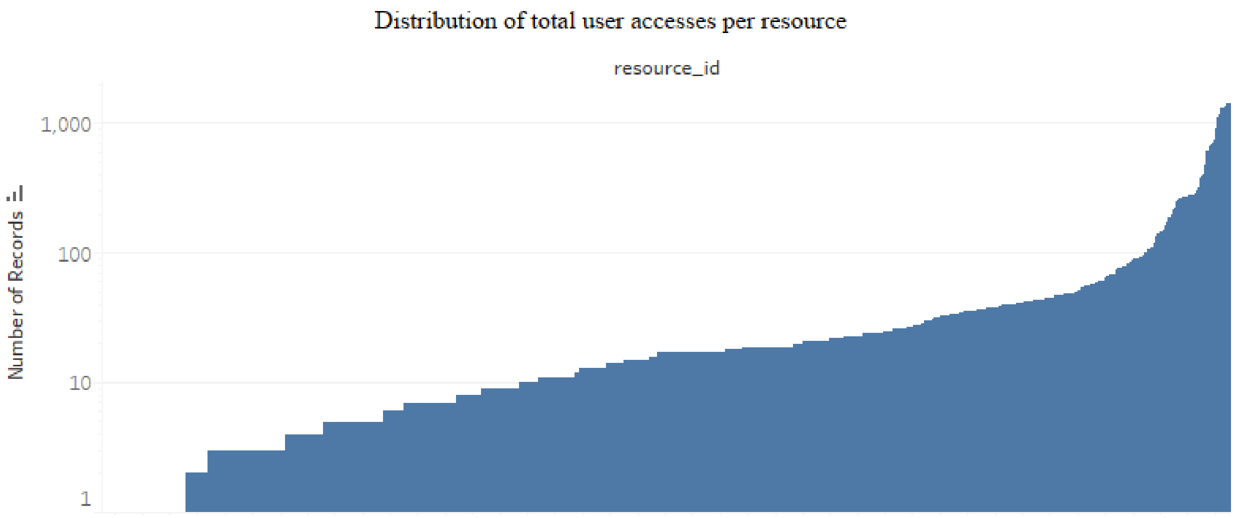


Figure 5.1 Distribution of total user accesses per resource in 2014

On the x-axis, we have the unique identifier of each resource and on the y-axis, we have the total number of accesses.

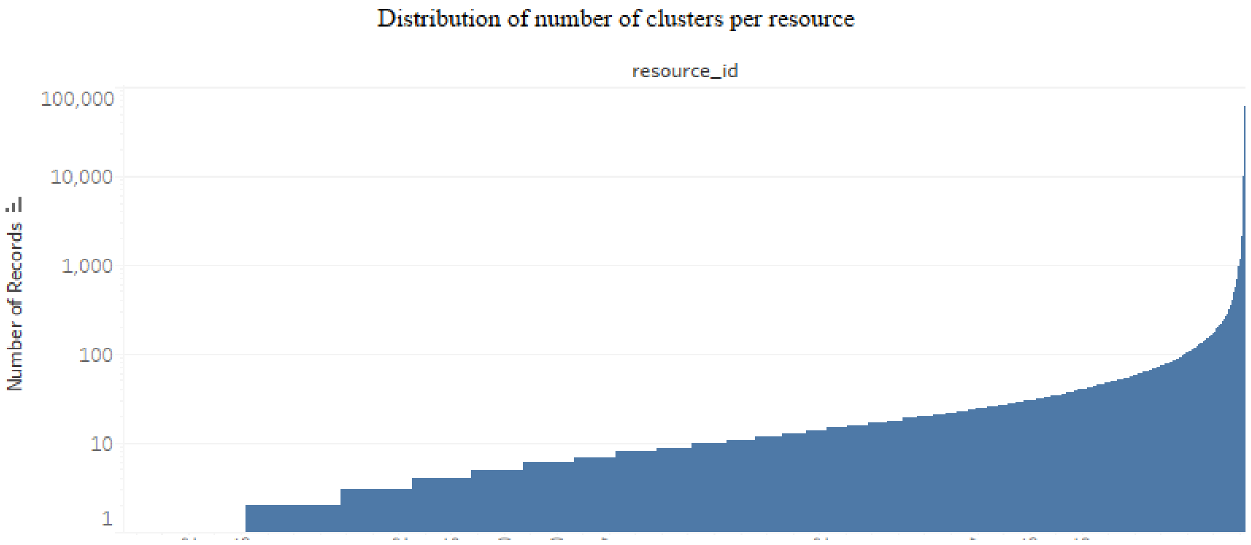


Figure 5.2. Distribution of number of clusters per resource in 2014

Unique resource identifiers are on the x-axis and number of accesses are on the y-axis. The distribution of data points before and after the first stage of processing is depicted geographically in the following paragraphs.

Figure 5.3 shows the distribution of all accesses worldwide of the 3605 unique resource identifiers that were accessed in 2014. Of these, 592 resources were accessed by group users shown in Figure 5.4. The distribution of total user access points and points accessed by group users shows that approximately 10% of the total users belong to a group.

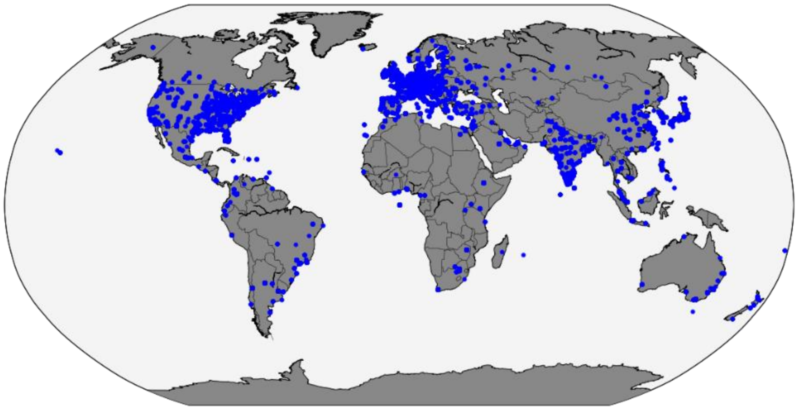


Figure 5.3. Global distribution of nanoHUB accesses during 2014

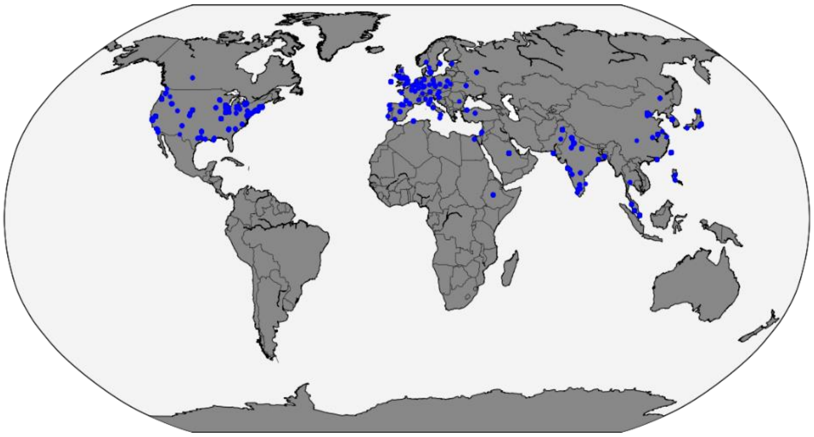
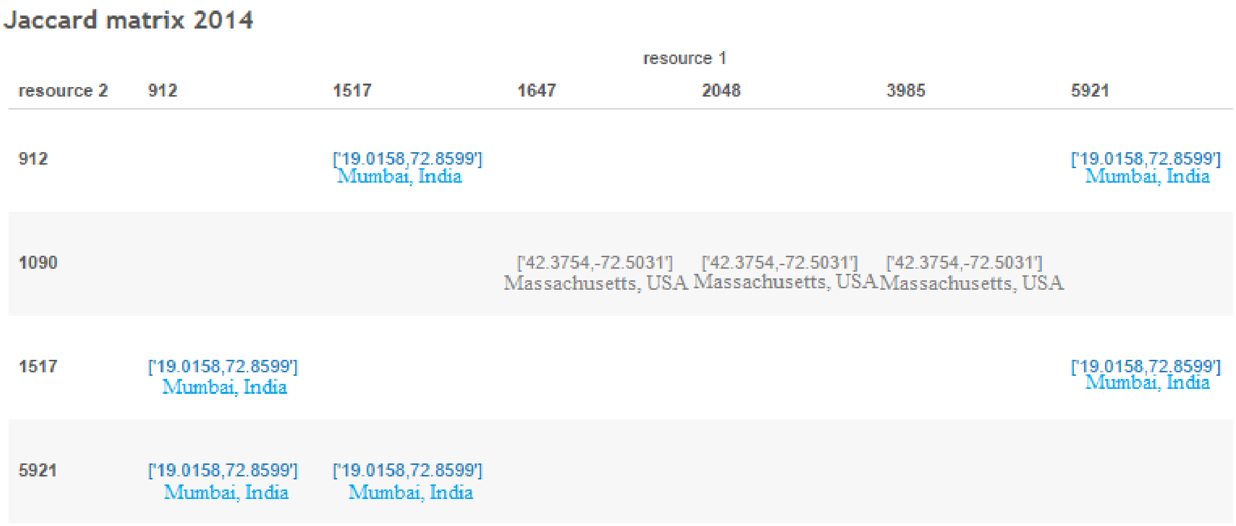


Figure 5.4. Cluster distribution of nanoHUB accesses during 2014

5.2 Jaccard-index Based Categorization

In this section, we present and explain the results of the second stage of the methodology used in our study. As mentioned in the framework, in the second stage of our process, we compare the user groups detected for various resources to find commonalities. If multiple resources are being accessed over time by the same user group, it indicates an association among such a set of resources. The results are presented using a resource-to-resource association matrix. Cells depicting resources that do not meet a Jaccard index threshold are left blank. Other cells which depict resources that are found to be similar based on Jaccard index contain the values of the geographic coordinates where user groups have accessed these resources. Figure 5.5 shows a sample portion of the resource-to-resource association matrix.



Jaccard matrix 2014

				resource 1		
resource 2	912	1517	1647	2048	3985	5921
912		['19.0158,72.8599'] Mumbai, India				['19.0158,72.8599'] Mumbai, India
1090			['42.3754,-72.5031'] Massachusetts, USA	['42.3754,-72.5031'] Massachusetts, USA	['42.3754,-72.5031'] Massachusetts, USA	
1517	['19.0158,72.8599'] Mumbai, India					['19.0158,72.8599'] Mumbai, India
5921	['19.0158,72.8599'] Mumbai, India	['19.0158,72.8599'] Mumbai, India				

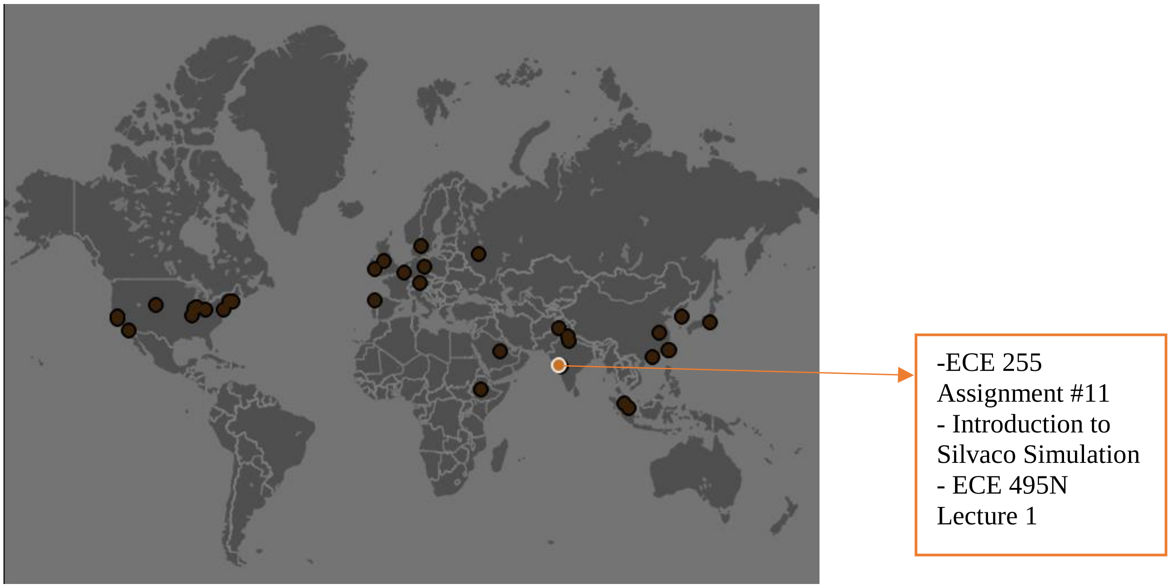
Figure 5.5 Section of the resource-resource association matrix that illustrates the location usage of resources on x-axis and resources on y-axis overlaps.

In the example shown in Figure 5.5, the location depicted by blue is where resources 912, 1517 and 5921 are used by classroom like groups of users. However, resource 912 and 1090 do not have any locations in common. Therefore the cell representing their Jaccard index is left blank. Some examples of groups that emerged from this analysis are described in detail in the tables below. Each example represents a set of resources found to be accessed by a common user group. The evaluation of results is done qualitatively based on tags, which are added by contributors to each resource. The evaluation is done qualitatively as the tags can vary greatly and need not come from a limited set of tags.

Table 5.1, shows the titles and tag categories of resources 912, 1517 and 5921. This example shows resources of seemingly unrelated categories being used by common user groups. Figure 5.6 shows the location of this resource grouping. Although these resources might be accessed by individuals at other geographic locations, this is the only geographic location where resources 912, 1517 and 5921 are accessed by user groups which were detected through the clustering methodology.

Table 5.1. Resource Grouping using cluster-based Jaccard indexing - Example 1

Title	Child ID	Tags
Homework for Circuit Simulation: ECE 255 Assignment #11 - High Frequency	912	Circuits Education Nanoelectronic
Introduction to Silvaco Simulation Software	1517	ACUTE Course Lecture Nanoelectronics
ECE 495N Lecture 1: What Makes Current Flow? Lecture Notes	5921	Course Lecture iTunes U Nanoelectronics Transistors



-ECE 255 Assignment #11 - Introduction to Silvaco Simulation - ECE 495N Lecture 1

Figure 5.6 Geographic location of accesses of resources in Table 5.1.

Figure 5.7 shows comparative three-dimensional graph of user groups that accessed resources 912, 1517 and 5921 in 2014. Figure 5.8 shows the latitude vs. longitude plot of the same.

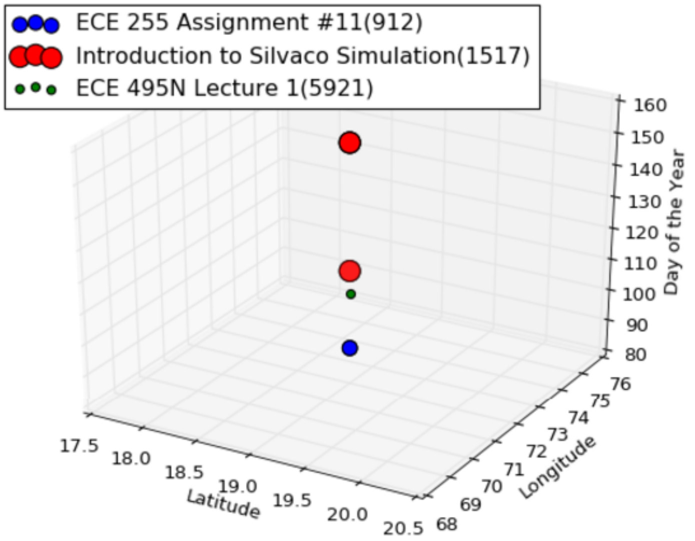


Figure 5.7 Three-dimensional user group access plot of resources 912, 1517 and 5921

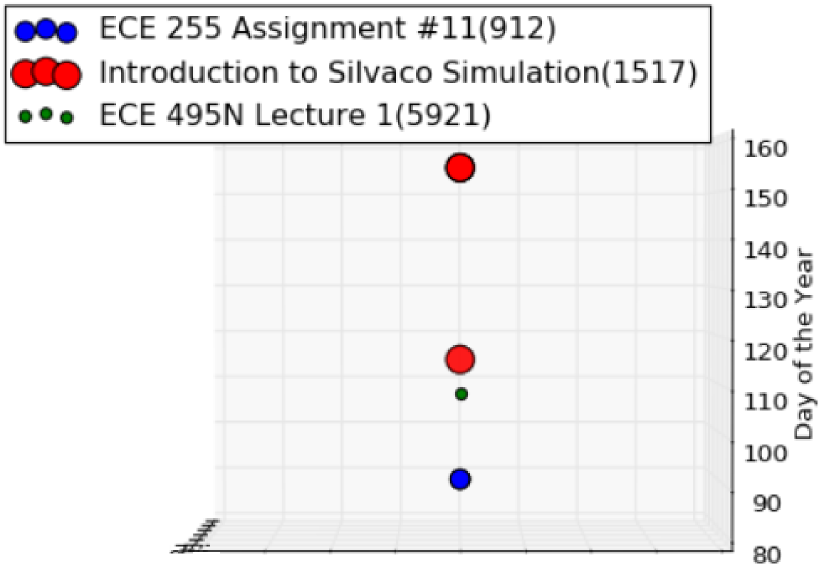
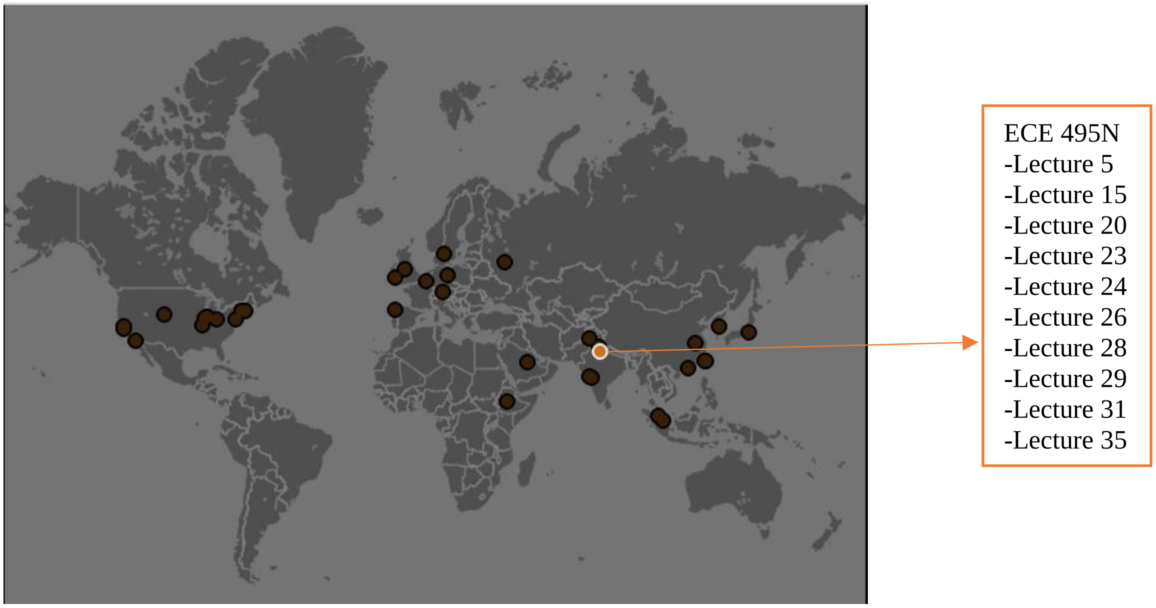


Figure 5.8 Time plot of user groups accessing resources 912, 1517 and 5921 depict the geographic overlap of usage of the three resources

We now look at some other examples of resource sets that were detected through the one to one comparison of user groups accessing each pair of resources. In Table 5.2, each resource was found to have a one-to-one similarity with every other resource at the location shown in Figure 5.9.

Table 5.2. Resource Group using cluster-based Jaccard indexing - Example 2

Title	Child ID
ECE 495N Lecture 5: Quantitative Model for Nanodevices II	5427
ECE 495N Lecture 15: Covalent Bonding	5648
ECE 495N Lecture 20: Bandstructures III	5657
ECE 495N Lecture 23: Density of States II	5730
ECE 495N Lecture 24: Subbands	5732
ECE 495N Lecture 26: Ballistic Conductance	5969
ECE 495N Lecture 28: Reciprocal Lattice	5974
ECE 495N Lecture 29: Landauer Formula	5991
ECE 495N Lecture 31: Coherent Quantum Transport	5993
ECE 495N Lecture 35: NEGF Continued II	6023



ECE 495N
-Lecture 5
-Lecture 15
-Lecture 20
-Lecture 23
-Lecture 24
-Lecture 26
-Lecture 28
-Lecture 29
-Lecture 31
-Lecture 35

Figure 5.9 Geographic location of accesses of resources in Table 5.2

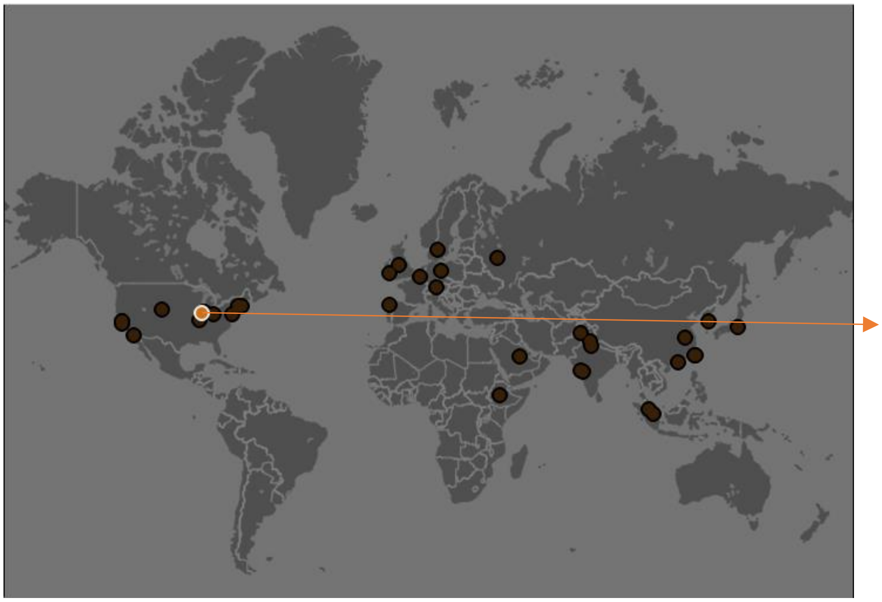
In this example, it is easy to identify that all resources belong to a single category, depicted by tag value 1, which refers to nanoelectronics and electrostatics. All the above lectures are contributed by nanoHUB contributor Dr. Supriyo Dutta, which makes verification easy. The group of resources seen in this example is expected to be grouped together as it belongs to the same lecture series. Therefore, the findings of such expected groups of resources serve as proof of concept of the efficiency of the method in this study and draw attention to the resources in the series that were absent from the grouping.

In the next example shown in Table 5.3, there are three different categories of resources. Tag value 1 represents tags related to crystals, crystal viewer tools. Tag value 2 represents nano/bio and self-assembly tags. Most resources have a tag value of 3, which represents Diagnostics, Illinois, nano/bio and therapeutics. The analysis of tags illustrating the difference between the two types of results is further evaluated by observing the cardinality of tags detailed in the appendix.

Table 5.3. Resource Grouping using cluster-based Jaccard indexing - Example 3

Title	Child ID
Illinois ABE 446 Lecture 5: Self-Assembly and Bioconjugation	8547
viewer.swf	8554
[Illinois] BioNanotechnology Seminar Series Spring 2012: DNA Mediated Synthesis of Novel Gold Nanoflowers for Diagnostic and Therapeutic Applications	13757
[Illinois] BioNanotechnology Seminar Series Spring 2012: Exploring Academic Collaboration with the University of Cape Coast, Ghana with the Global Health Initiative at UIUC	13758
[Illinois] BioNanotechnology Seminar Series Spring 2012: Mesenchymal Stem Cells Contribute to Vascular Growth in Skeletal Muscle in Response to Eccentric Exercise	13760
[Illinois] BioNanotechnology Seminar Series Spring 2012: Direct Write Assembly of 3D Microperiodic Hydrogel Scaffolds for Stem Cell Culture and Tissue Engineering	13761
[Illinois] ECE 416 Introduction to Biosensors II	16708
[Illinois] Rational Design of MegaDalton-Scale DNA-Based Light Harvesting Antennas	19481
[Illinois] DIY BIOSENSORS Day 1 Summer 2014 Workshop	21192

[Illinois] DIY BIOSENSORS Day 2 Summer 2014 Workshop	21194
[Illinois] DIY BIOSENSORS Day 3 Summer 2014 Workshop	21196
[Illinois] DIY BIOSENSORS Day 4 Summer 2014 Workshop	21198
[Illinois] DIY BIOSENSORS Day 5 Summer 2014 Workshop	21201
[Illinois] Mechanobiology in Neuronal Development	21348
[Illinois] Fundamentals of Nano-Optics and Plasmonics for the Biomedical Researcher	21351
[Illinois] Gold Nanostars as Tiny Hitchhikers for Cancer Therapeutics	21353
[Illinois] Translational Nanomedicines Using Biomedical Nanomaterials	21361
A link	22330



-ABE 446 Lecture 5
- DNA Mediated Synthesis
- Exploring Academic Collaboration
- Mesenchymal Stem Cells
- Tissue Engineering
- ECE 416 Introduction to Biosensors
- Design of MegaDalton
- DIY BIOSENSORS Day 1,2,3,4,5
- Mechanobiology in Neuronal Development
- Plasmonics for the Biomedical Researcher
- Tiny Hitchhikers for Cancer Therapeutics
- Translational Nanomedicines

Figure 5.10 Geographic location of accesses of resources in Table 5.3

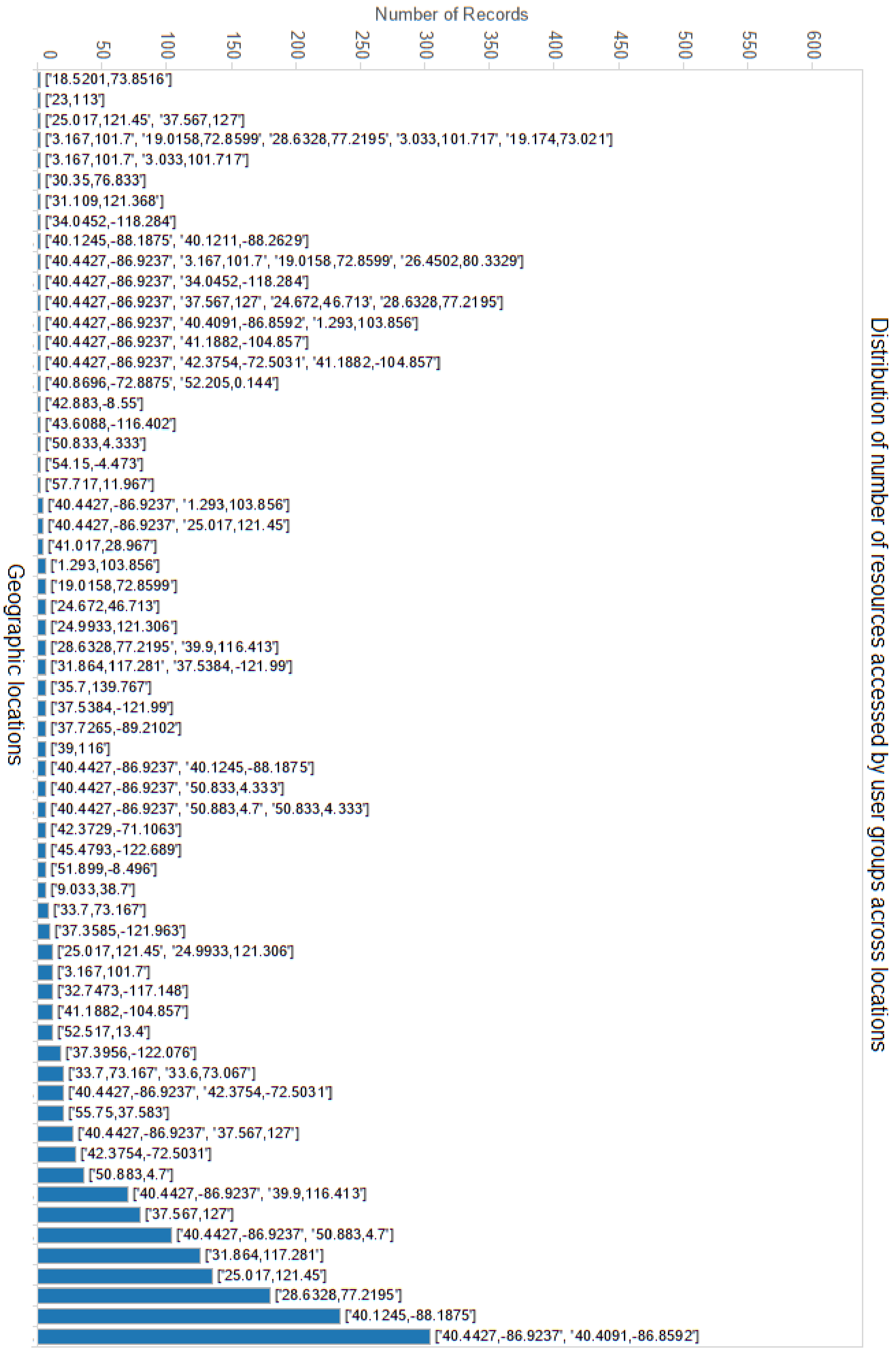


Figure 5.11 Distribution of resources per location

Figure 5.11 shows the distribution of resources for each location. It shows that the number of resources that users in West Lafayette (where Purdue University is located) access are 263. The second highest number of resources accessed are from a location near Beijing with 224 resources accessed. 65 resource sets were found in the year 2014 across 43 locations some of which have been discussed above. A detailed listing of the remaining resource sets is provided in the appendix. Resource sets obtained at the conclusion of the second stage of the methodology used in this study yield insightful results for most locations. For sites like Purdue University, where there are over 200 resources viewed by user groups, it is difficult to distinguish the association between all resources. With the exception of the two mentioned locations, all other locations have less than twenty individual resources in a set of resources. The distributions and results of other years studied are detailed in the appendix.

In chapter 5 we presented the initial distribution of the data and the outcomes of each stage in the process of user clustering leading to resource grouping. It shows some examples of types of resource groups with different tag-based purity. In Chapter 6, we will present a discussion of the results and analysis of the types of resource groups obtained as a result of this process.

6. CONCLUSIONS

In the concluding chapter, we summarize the work done in this study and discuss the implications of the results obtained. A detailed explanation of each type of result and an analysis of the approach used, along with its strengths and weaknesses is provided in this chapter. We also look at the value of current results in aiding future user analytics studies for science gateways.

The methodology applied in this study uses features related to geographic proximity and time, to explore relationships between different users and resources of nanoHUB. The resulting grouping captures important information about the impact of nanoHUB for education and regions of popular usage. This information is useful in analyzing sections across the globe where nanoHUB has emerged as an influential platform. The results depict a continued trend of popularity in several regions in Europe, USA, India, China and North America among many others that are mentioned in detail in the appendix.

Our objective throughout this study has been to analyze the type of influence and impact nanoHUB has in these regions. This is achieved using the framework described in earlier chapters, which is intended to detect “classroom” like groups of users and find associations between resources used by these groups. The method uses features which do not require users to log in and succeeds in optimizing the amount of usable input data. However, it is worth noting that the result of clustering users into user groups contains approximately ten percent of the total data. This implies that majority of the users of nanoHUB are independent users who do not belong to a “classroom-like” group. From these results, we can say that a study on individual, non-group users is required and will be useful in understanding more about user behavior.

We now focus on the resource sets detected based on the clustering of users into groups. In the results, it was found that density-based clustering techniques used in this study lead to two types of resource groups.

We refer to the first category of resource sets as resource sets with structured content. An example of this category of resource sets is the lectures of series ECE 495N listed in Table 5.2.

All resources in this set are known to belong to the same series or set. Such resources have already been categorized under a common entity by nanoHUB. The content of this category of resource sets, detected through the methodology described, is easy to verify as it is already structured, titled and tagged under a common topic. The results allow us to extract and interpret several important aspects of indirect feedback from the detected resource sets. It is possible to analyze the geographic regions and even specific institutes where each of the resource sets has a strong impact. We can further extract information about the content creators behind the detected resource sets and provide feedback to them about the regions where their content is being viewed. This information can help content creators establish research connections and grow their research network to specific institutes, in turn helping nanoHUB expand and add to its global community.

Apart from geographic information, another useful analytics aspect is the structure of the detected resource sets. As seen in Table 5.2, not all lectures under lecture series ECE 496N are a part of the resultant resource set. Partial use of a lecture series, if occurring multiple times, would indicate an inclination of the users to skip some of the content that the creator deemed fit to be a part of the series. Another reason for such discontinuous usage pattern could be that some resources in the structured content are difficult to use or understand. The methodology used in this technique can help detect users' inclination to skip some parts of series and can help contributors to restructure elements of their content to make it more usable. It would help draw attention to resources that are working incorrectly and need corrective maintenance from nanoHUB. Evidence can be found which indicates that some resources of such a series that appear in the detected resource sets are pre-selected into groups called collections by some users.

The second category of resource sets can be defined as resource sets with content that does not have an easily identifiable/verifiable link. This includes resources that belong to different topics, or lecture series, or different disciplines of nanotechnology, that do not have an easily verifiable link and have not been categorized by nanoHUB under the same group. An example of this category of resource sets is defined in Table 5.3, wherein a workshop on DIY biosensors and a paper on cancer research appear under the same resource grouping. Resources in the given examples appear in a single cluster but are not tagged under one topic by nanoHUB. However, there could be an underlying relationship between the two. It is also possible that the grouping of vastly unrelated resources by our methodology is a fundamental failure of the clustering technique. Our methodology is based on location and time which could lead to unconnected resources, which are accessed by several different departments at an institute, to be grouped as a single resource set leading to the merging of multiple classroom-like groups. Since our methodology is blind to the context and content of the resources, it is difficult to verify the credibility of such a grouping. However, if the clustering is accurate, then the observation that links two or more resources, belonging to unrelated categories, could not have been achieved without user analytics studies focused on geographic and time features. The implication of seemingly unrelated resources being grouped together points us to investigate the link between these resources and draws attention to new knowledge area involving disparate resources which can be supported on nanoHUB.

6.1 Future Work

A context based method like the use of natural language processing (NLP) or deep learning can be a good approach to evaluate the finding of the seemingly unrelated resource sets, wherein the title and contents of the resources can also be taken as input data. Since NLP and deep learning approaches are computationally very intensive, the findings of this study (i.e. the detected resource sets) can be used as a basis to narrow down the amount of data given as input.

6.2 Conclusion

The findings of the resource grouping are relevant in two ways. First, is validating the usage of already structured content and reinforcing its usefulness with the help of data and the second is laying the groundwork for further analysis, by identifying the type of resources which cannot be detected using ST DBSCAN.

We conclude by stating that the results of the study help us to evaluate information about the usability of science gateways. The methodology followed in this study is effective in establishing the use of nanoHUB in education and research by identifying classroom like groups with the help of user access data. Due to the low level of dependence on platform specific data, the methodology used in this study is applicable to other science gateways as well. Our approach maximizes usage of available data. The findings of the study point to a need for future research on individual user trajectory analysis and context-based group behavior analysis.

REFERENCES

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227.

Kut, A., & Birant, D. (2006). Spatio-temporal outlier detection in large databases. *CIT. Journal of computing and information technology*, *14*(4), 291-297.

Nohadani, O., Dunn, J., & Klimeck, G. (2016). Categorizing Users of Cloud Services. *Service Science*, *8*(1), 59-70.

Klimeck, G., McLennan, M., Brophy, S. P., Adams III, G. B., & Lundstrom, M. S. (2008). nanohub.org: Advancing education and research in nanotechnology. *Computing in Science & Engineering*, *10*(5), 17-23.

Madhavan, K., Zentner, M., & Klimeck, G. (2013). Learning and research in the cloud. *Nature nanotechnology*, *8*(11), 786-789.

Klimeck, G., Adams III, G. B., Madhavan, K. P., Denny, N., Zentner, M. G., Shivarajapura, S., ... & Beaudoin, D. L. (2011, May). Social Networks of Researchers and Educators on nanoHUB. org. In *Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (pp. 560-565). IEEE Computer Society.

Das, Abhinandan S., et al. "Google news personalization: scalable online collaborative filtering." *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.

Figueiredo, F., Ribeiro, B., Almeida, J. M., & Faloutsos, C. (2016, April). TribeFlow: Mining & Predicting User Trajectories. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 695-706). International World Wide Web Conferences Steering Committee.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, *2*(2), 169-194.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Huang, Z., & Ng, M. K. (2003). A note on k-modes clustering. *Journal of Classification, 20*(2), 257-261.

Gomez-Uribe, Carlos, A., & Hunt, Neil. (2015). *The Netflix Recommender System: Algorithms, Business Value, and Innovation*. 1-19.

Rahman, A., & Verma, B. (2013). Cluster-based ensemble of classifiers. *Expert Systems*, 30(3), 270-282.

Gomide, J., Veloso, A., Meira Jr, W., Almeida, V., Benevenuto, F., Ferraz, F., & Teixeira, M. (2011, June). *Dengue surveillance based on a computational model of spatio-temporal locality of Twitter.* In Proceedings of the 3rd international web science conference (p. 3). ACM.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Bradley, P. S., & Fayyad, U. M. (1998, July). Refining Initial Points for K-Means Clustering. In *ICML* (Vol. 98, pp. 91-99).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

Qi, M., Nohadani, Omid, Klimeck, Gerhard, & Landry, Steven. (2014). NanoHUB Usage Analysis: Using Anomaly Detection and Principal Component Analysis, ProQuest Dissertations and Theses.

Wedel, M., Kamakura, W.: Market Segmentation: *Conceptual and Methodological Foundations.* Springer, Heidelberg (1999)

Ng, & Jiawei Han. (2002). CLARANS: A method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 14(5), 1003-1016.

Ankerst, M., Breunig, M., Kriegel, H., & Sander, J. (1999). OPTICS: *Ordering points to identify the clustering structure. ACM SIGMOD Record*, 28(2), 49-60.

Kubota, Y., Suzuki, R., & Arita, T. (2014). A procedure for constructing social network using Web search engines: The case for Japanese automotive industry. *Artificial Life and Robotics*, *19*(1), 103-108.

John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, Nancy Wilkins-Diehr, "XSEDE*: Accelerating Scientific Discovery",* *Computing in Science & Engineering*, vol.16, no. 5, pp. 62-74, Sept.-Oct. 2014, doi:10.1109/MCSE.2014.80

XSEDE | Gateways Listing. (n.d.). Retrieved April 08, 2017, from https://www.xsede.org/gateways-listing

ScienceGateways.org(n.d.). Retrieved April 08, 2017, from http://sciencegateways.org/

NSG Metrics. (n.d.). Retrieved June 11, 2017, from http://www.nsgportal.org/metrics.html

"National Cancer Institute (NCI)." *National Institutes of Health*. U.S. Department of Health and Human Services, 30 Jan. 2017. Web. 30 May 2017.

Christie, M., Marru, S., & Wilkins-Diehr, Nancy. (2007). The LEAD Portal: A TeraGrid gateway and application service architecture. *Concurrency and Computation: Practice and Experience*, *19*(6), 767-781.

Droegemeier, K. K., Chandrasekar, V., Clark, R., Gannon, D., Graves, S., Joseph, E., ... & Leyton, T. (2005, January). Linked environments for atmospheric discovery (lead): Architecture, technology roadmap and deployment strategy. In *21st Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*.

Miller, M., Schwartz, T., & Pfeiffer, W. (2016). User behavior and usage patterns for a highly accessed science gateway. *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale,* 1-8.

Miller, M. A., Schwartz, T., Hoover, P., Yoshimoto, K., Sivagnanam, S., & Majumdar, A. (2015, July). The CIPRES workbench: a flexible framework for creating science gateways. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure* (p. 39). ACM.

APPENDIX

SQL QUERY

Below is the SQL query used to extract the data into Comma Separated Value (CSV) files. The script is a sample for extraction of data for the year 2011. The resultant file is title “in_STDB_2011_1.csv”. For extraction of data of consequent years, the datetime filtering criteria needs to be edited appropriately.

```
select amu.id, amu.ip, jra.child_id, amu.datetime ,ig.ipLATITUDE, ig.ipLONGITUDE,
jra.parent_id #
 INTO OUTFILE '/var/lib/mysql-files/in_STDB_2011.csv'
 from nanohub_metrics.andmore_usage amu,
 nanohub_metrics.ip_geodata ig,
 nanohub.jos_resource_assoc jra
 where http_method = 'GET'
 and amu.cms_action_name = jra.child_id
 and ig.ip = amu.ip
 and http_return_code = 200
 and amu.datetime between '2011-01-01' and amu.datetime <= '2012-01-01'
 and amu.cms_action_name = jra.child_id
order by dayofyear(amu.datetime)     ;
```

SOURCE CODE

The source code for the implementation of the methodology used in this study is made available at the following links.

https://purr.purdue.edu/projects/mugdhathesis/files/browse
https://github.com/mgogte/nanoHUB-user-access-data-clustering

The link consists of two iPython Notebooks.

1) ST_DBSCAN.ipynb
2) Jaccard_sim.ipynb

1) ST_DBSCAN.ipynb: This iPython notebook contains the script for the implementation of ST_DBSCAN clustering algorithm and gives the user clusters in the form of core cluster points and neighboring cluster points as output into two CSV files – core_XXX.csv and neigh_XXX.csv, where XXX represents the respective year

2) Jaccard_sim.ipynb : This iPython notebook contains the implementation of the comparison of user clusters to detect resource sets. Resource sets are given as output into files titled jacc_sim_2011.csv, where XXX represents the respective year.

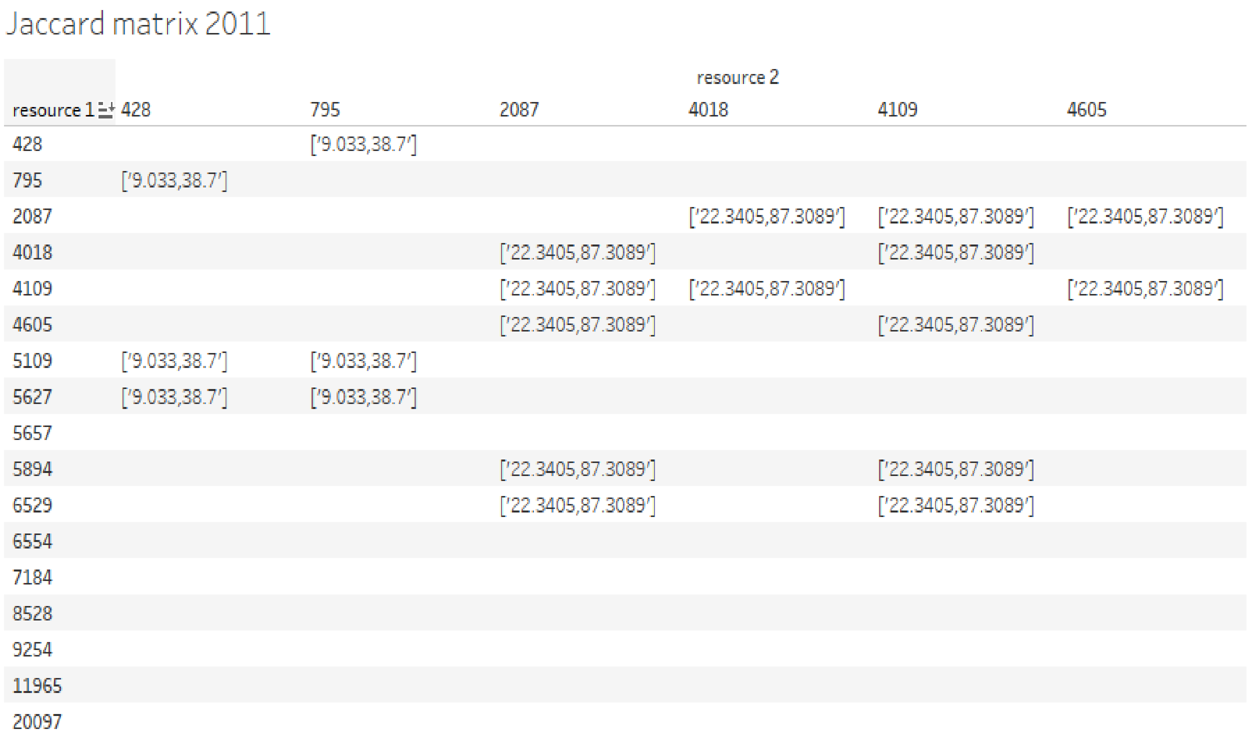
Jaccard Matrix Result Examples

The user groups derived as a result of clustering and the resource to resource comparison using Jaccard index are listed for all years from 2011 to 2015 at the link below.

https://purr.purdue.edu/projects/mugdhathesis/files/browse?subdir=Final%20Results

Below are the sample sections of Jaccard matrices for each year to illustrate the resource-to-resource comparison for each year and details of the number of resources compared for each year.

2011



Jaccard matrix 2011

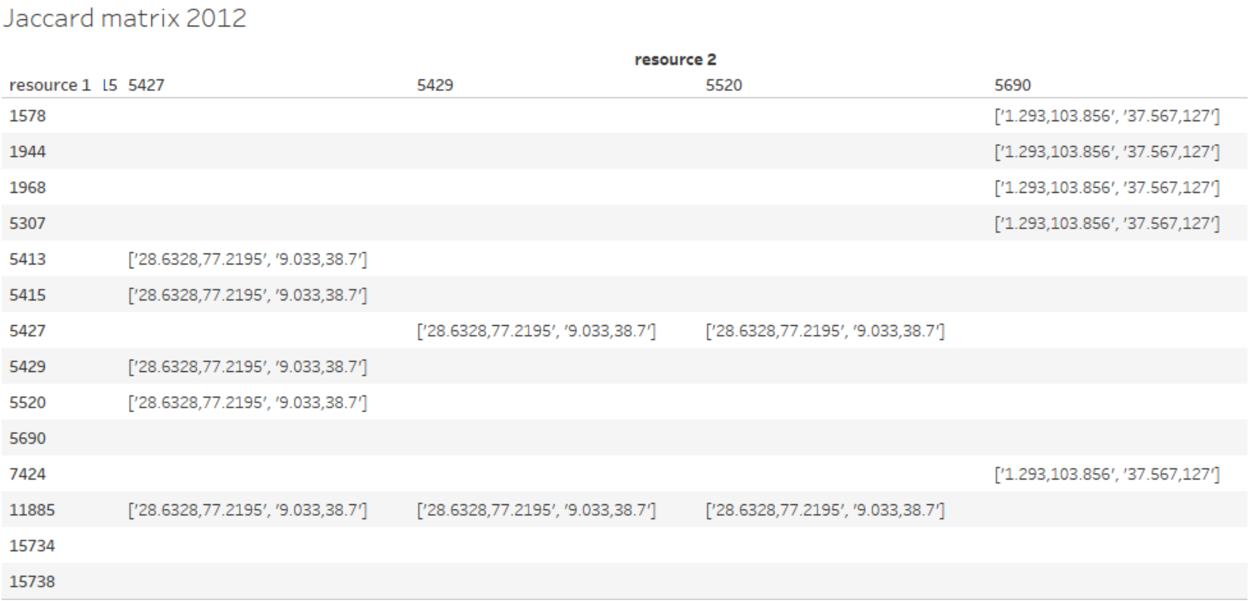
	resource 2					
resource 1	428	795	2087	4018	4109	4605
428		['9.033,38.7']				
795	['9.033,38.7']					
2087				['22.3405,87.3089']	['22.3405,87.3089']	['22.3405,87.3089']
4018			['22.3405,87.3089']		['22.3405,87.3089']	
4109			['22.3405,87.3089']	['22.3405,87.3089']		['22.3405,87.3089']
4605			['22.3405,87.3089']		['22.3405,87.3089']	
5109	['9.033,38.7']	['9.033,38.7']				
5627	['9.033,38.7']	['9.033,38.7']				
5657						
5894			['22.3405,87.3089']		['22.3405,87.3089']	
6529			['22.3405,87.3089']		['22.3405,87.3089']	
6554						
7184						
8528						
9254						
11965						
20097						

Figure A1. Section of resource-to-resource matrix for 2011

number of unique resources in 2011: 1623 number of unique locations in 2011: 1340

dimensions of similarity matrix: (1623, 1340)

2012



Jaccard matrix 2012

		resource 2			
resource 1	l5	5427	5429	5520	5690
1578					['1.293,103.856', '37.567,127']
1944					['1.293,103.856', '37.567,127']
1968					['1.293,103.856', '37.567,127']
5307					['1.293,103.856', '37.567,127']
5413		['28.6328,77.2195', '9.033,38.7']			
5415		['28.6328,77.2195', '9.033,38.7']			
5427			['28.6328,77.2195', '9.033,38.7']	['28.6328,77.2195', '9.033,38.7']	
5429		['28.6328,77.2195', '9.033,38.7']			
5520		['28.6328,77.2195', '9.033,38.7']			
5690					
7424					['1.293,103.856', '37.567,127']
11885		['28.6328,77.2195', '9.033,38.7']	['28.6328,77.2195', '9.033,38.7']	['28.6328,77.2195', '9.033,38.7']	
15734					
15738					

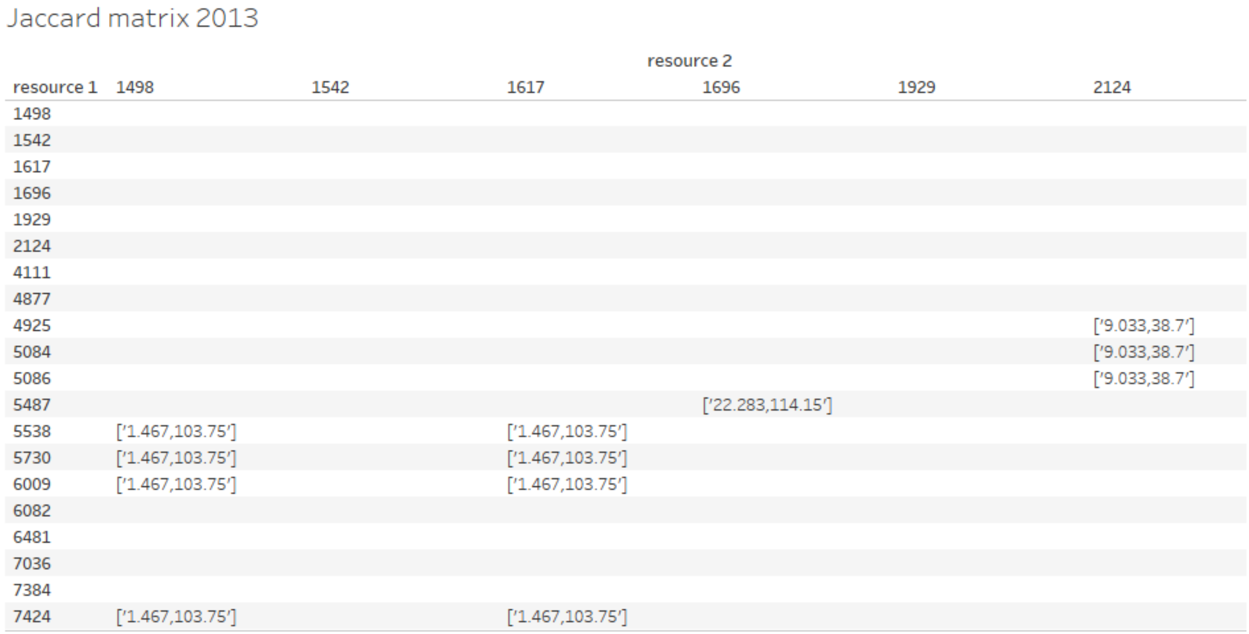
Figure A2. Section of resource-to-resource matrix for 2012

number of unique resources in 2012: 1715

number of unique locations in 2012: 452

dimensions of similarity matrix: (1715, 452)

2013



Jaccard matrix 2013

	resource 2					
resource 1	1498	1542	1617	1696	1929	2124
1498						
1542						
1617						
1696						
1929						
2124						
4111						
4877						
4925						['9.033,38.7']
5084						['9.033,38.7']
5086						['9.033,38.7']
5487				['22.283,114.15']		
5538	['1.467,103.75']		['1.467,103.75']			
5730	['1.467,103.75']		['1.467,103.75']			
6009	['1.467,103.75']		['1.467,103.75']			
6082						
6481						
7036						
7384						
7424	['1.467,103.75']		['1.467,103.75']			

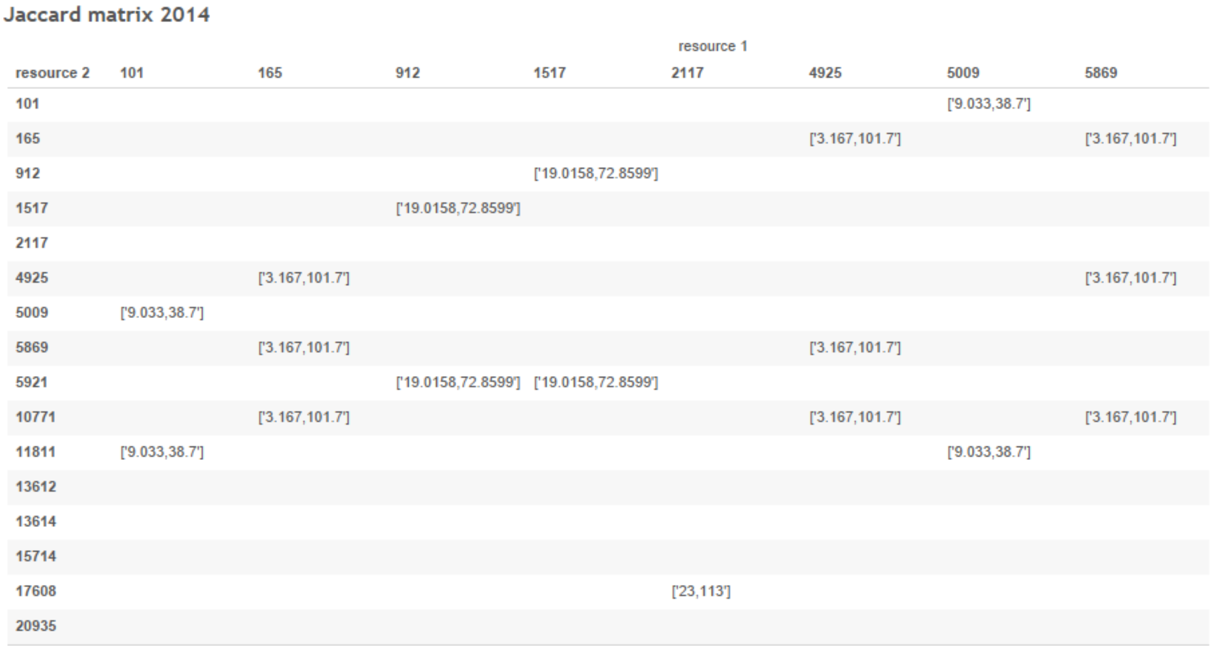
Figure A3. Section of resource-to-resource matrix for 2013

number of unique resources in 2013: 934

number of unique locations in 2013: 354

dimensions of similarity matrix: (934, 354)

2014



Jaccard matrix 2014

	resource 1							
resource 2	101	165	912	1517	2117	4925	5009	5869
101							['9.033,38.7']	
165						['3.167,101.7']		['3.167,101.7']
912				['19.0158,72.8599']				
1517			['19.0158,72.8599']					
2117								
4925		['3.167,101.7']						['3.167,101.7']
5009	['9.033,38.7']							
5869		['3.167,101.7']				['3.167,101.7']		
5921			['19.0158,72.8599']	['19.0158,72.8599']				
10771		['3.167,101.7']				['3.167,101.7']		['3.167,101.7']
11811	['9.033,38.7']						['9.033,38.7']	
13612								
13614								
15714								
17608					['23,113']			
20935								

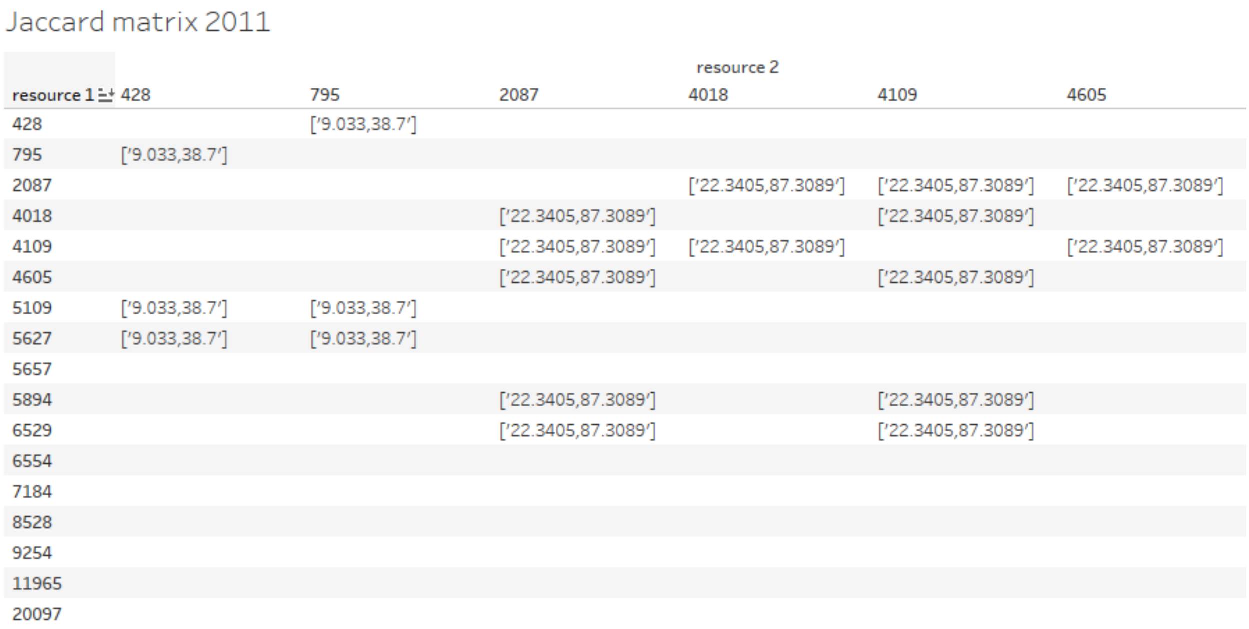
Figure A4. Section of resource-to-resource matrix for 2014

number of unique resources in 2014: 699

number of unique locations in 2014: 276

dimensions of similarity matrix: (699, 276)

2015



Jaccard matrix 2011

	resource 2					
resource 1	428	795	2087	4018	4109	4605
428		['9.033,38.7']				
795	['9.033,38.7']					
2087				['22.3405,87.3089']	['22.3405,87.3089']	['22.3405,87.3089']
4018			['22.3405,87.3089']		['22.3405,87.3089']	
4109			['22.3405,87.3089']	['22.3405,87.3089']		['22.3405,87.3089']
4605			['22.3405,87.3089']		['22.3405,87.3089']	
5109	['9.033,38.7']	['9.033,38.7']				
5627	['9.033,38.7']	['9.033,38.7']				
5657						
5894			['22.3405,87.3089']		['22.3405,87.3089']	
6529			['22.3405,87.3089']		['22.3405,87.3089']	
6554						
7184						
8528						
9254						
11965						
20097						

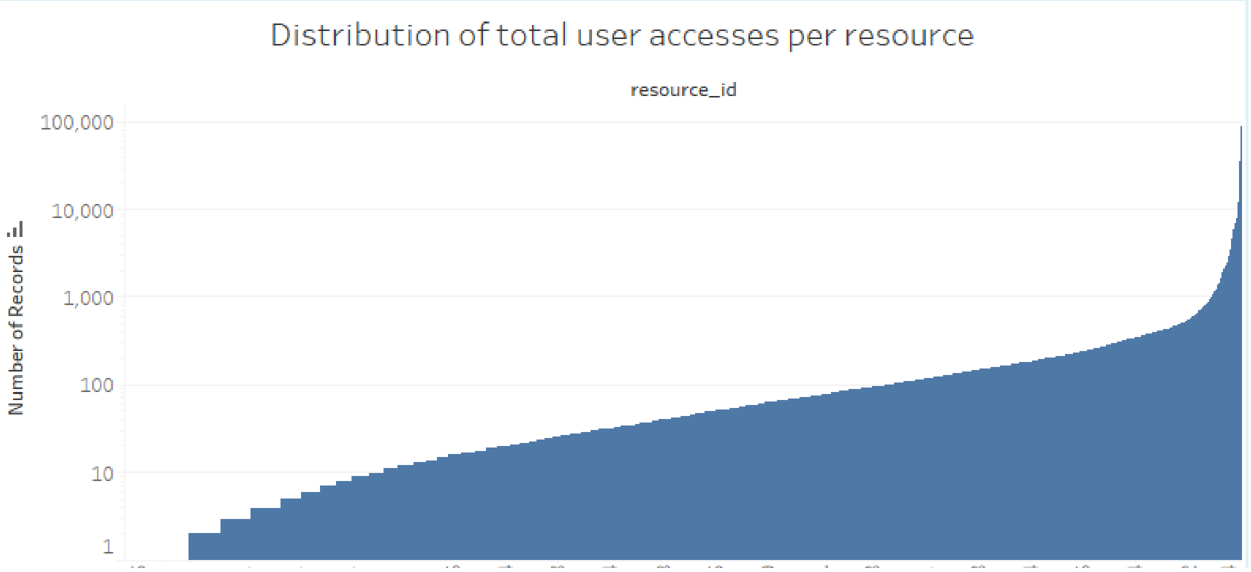
Figure A5. Section of resource-to-resource matrix for 2015

number of unique resources in 2015: 359

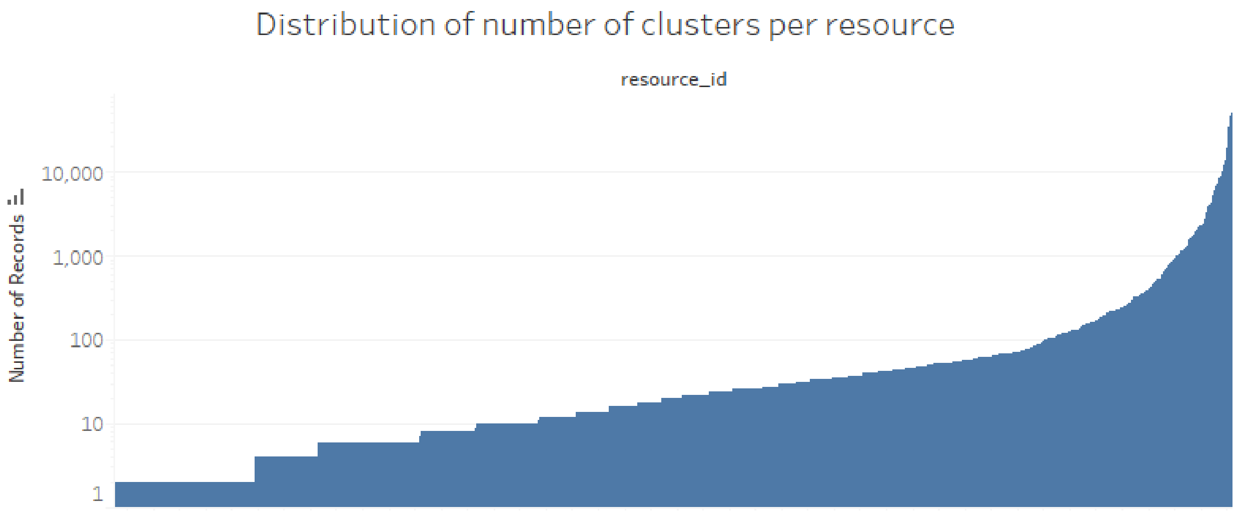
number of unique locations in 2015: 172

dimensions of similarity matrix: (359, 172)

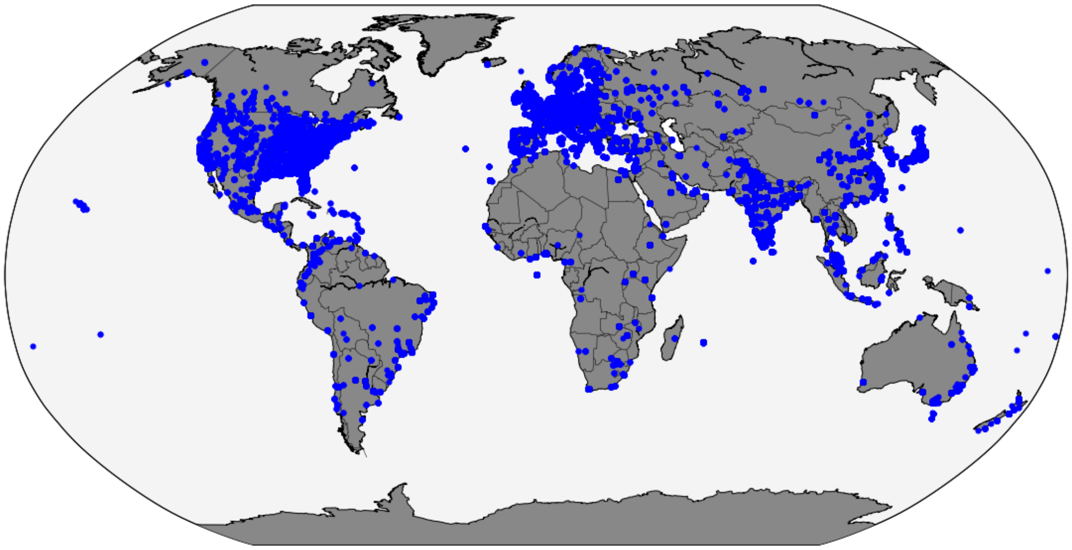
Data Summary 2011

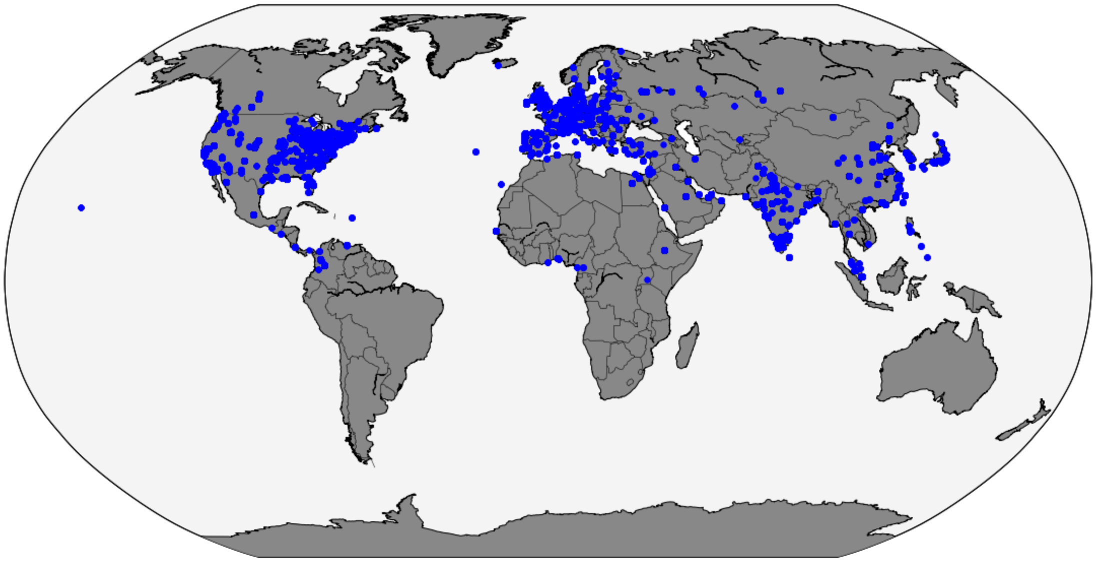


Distribution of total user accesses per resource

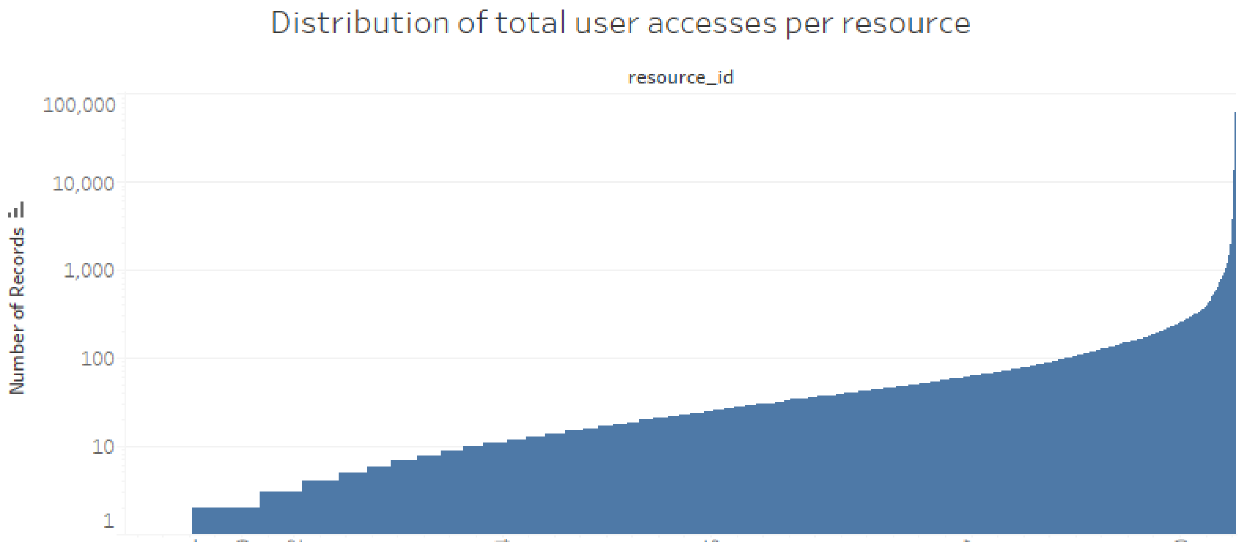


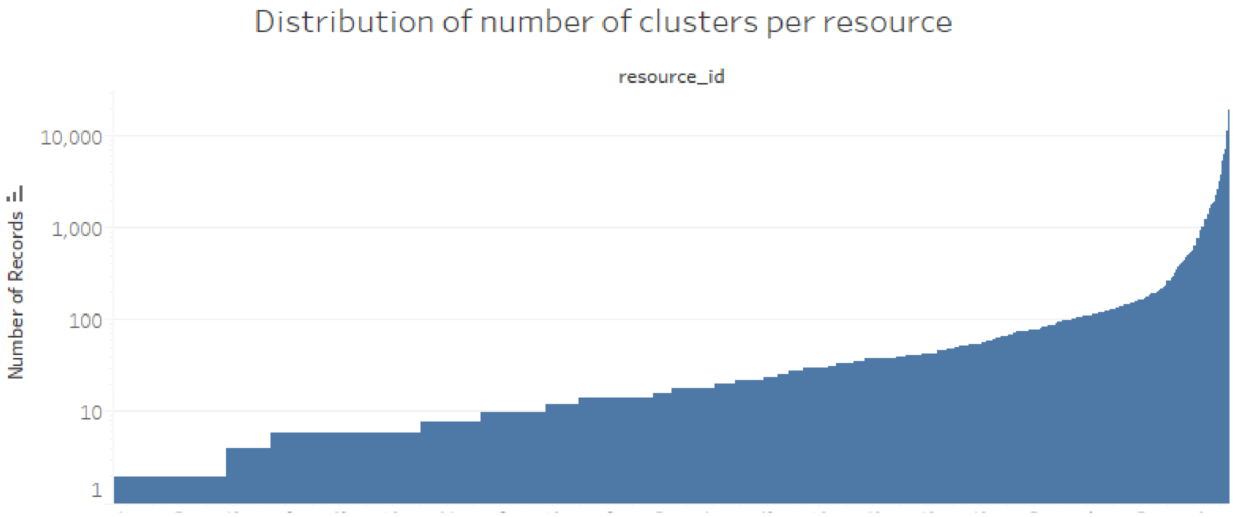
Distribution of number of clusters per resource

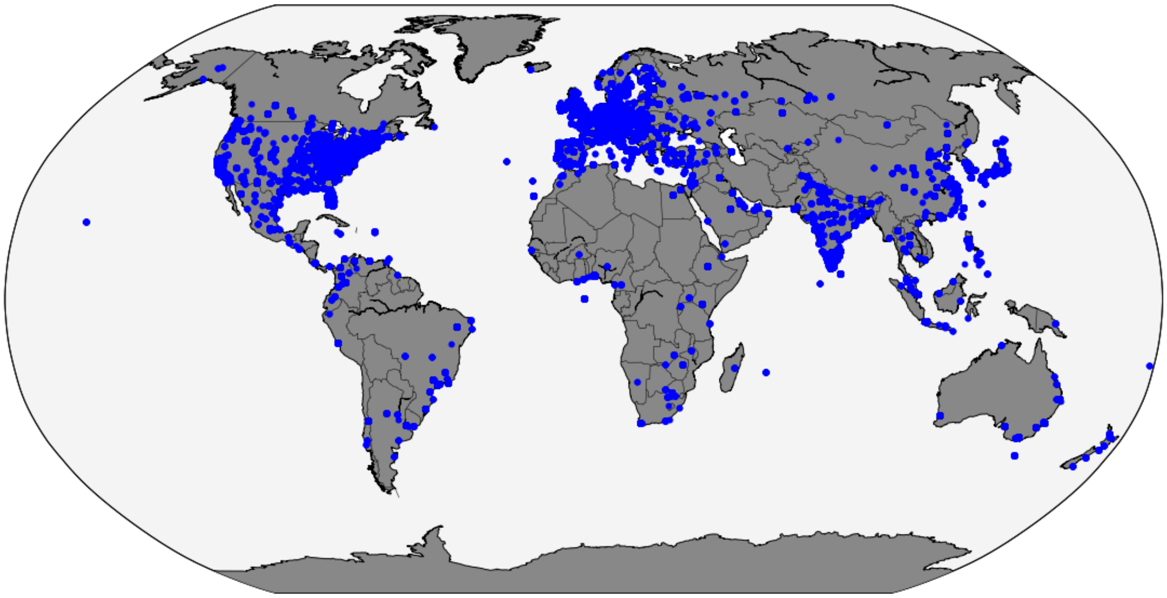


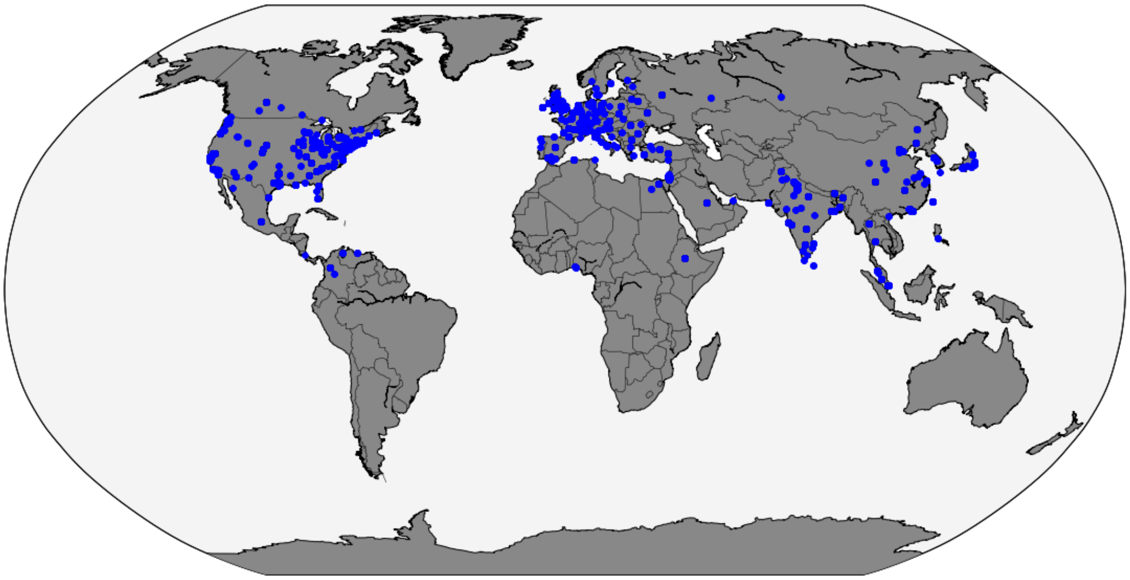


Data Summary 2012

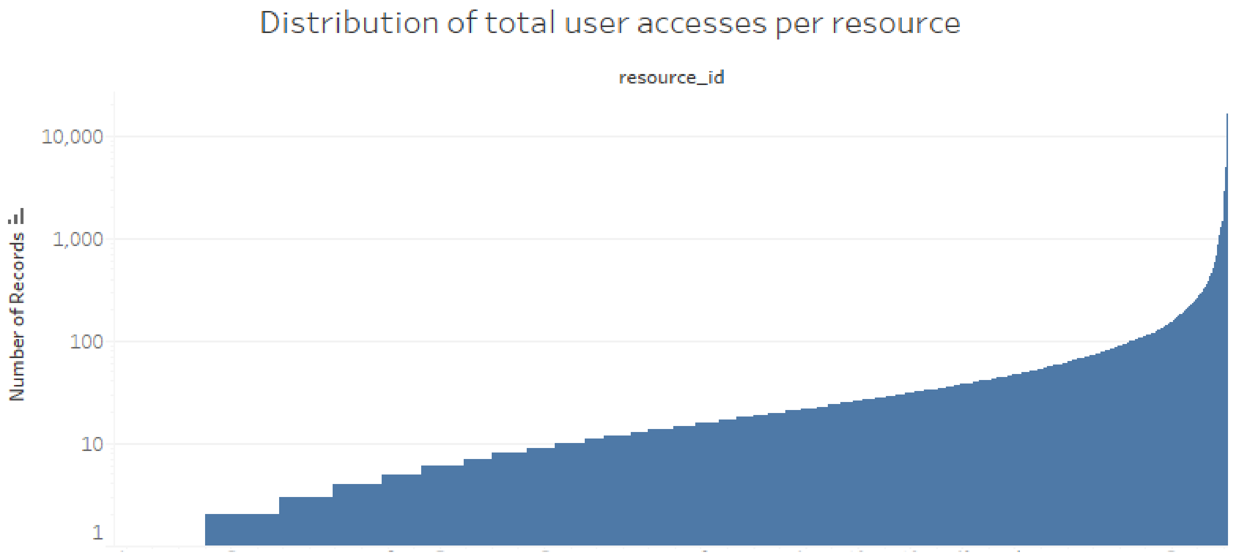




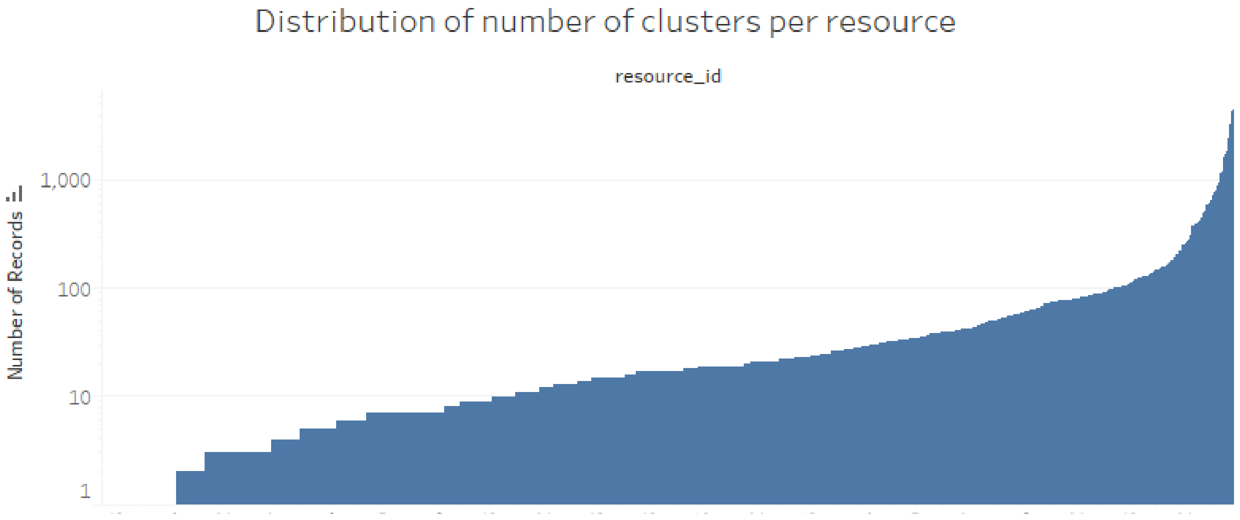




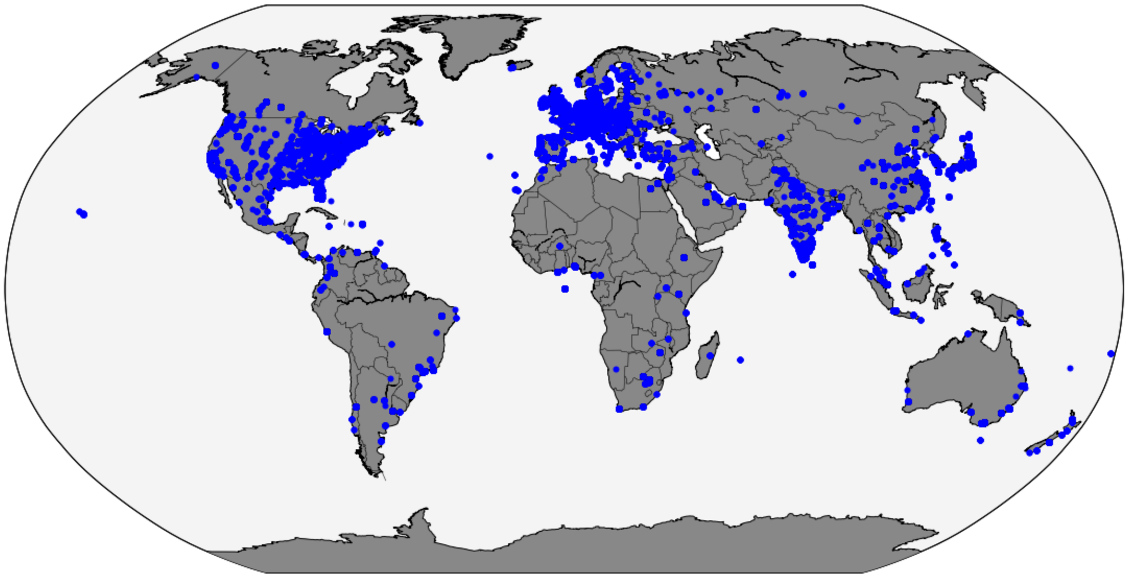
Data Summary 2013

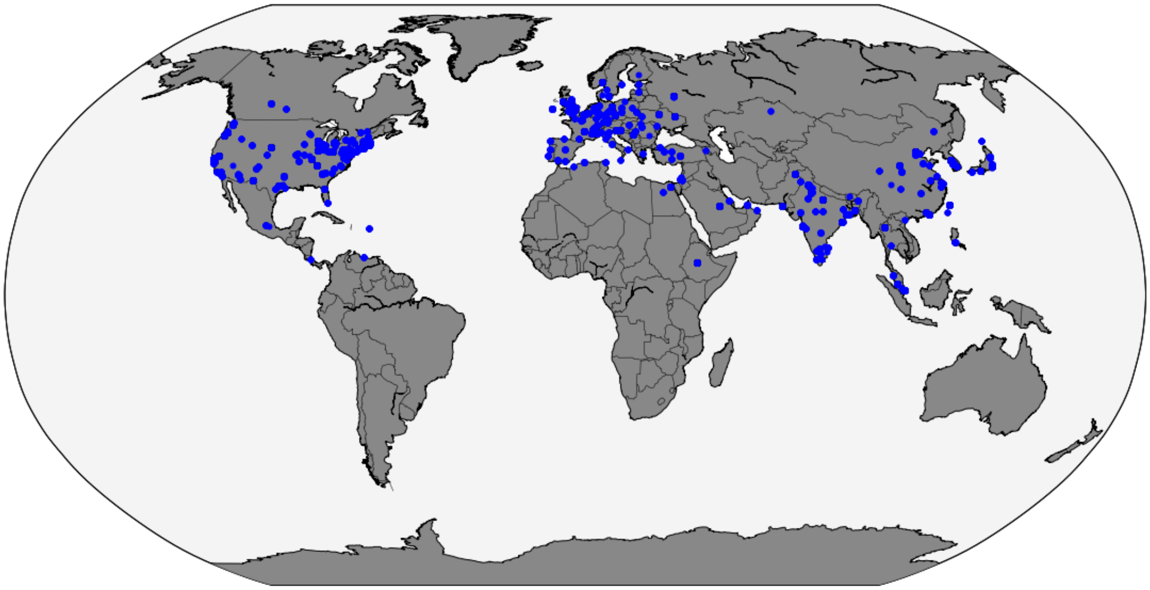


Distribution of total user accesses per resource

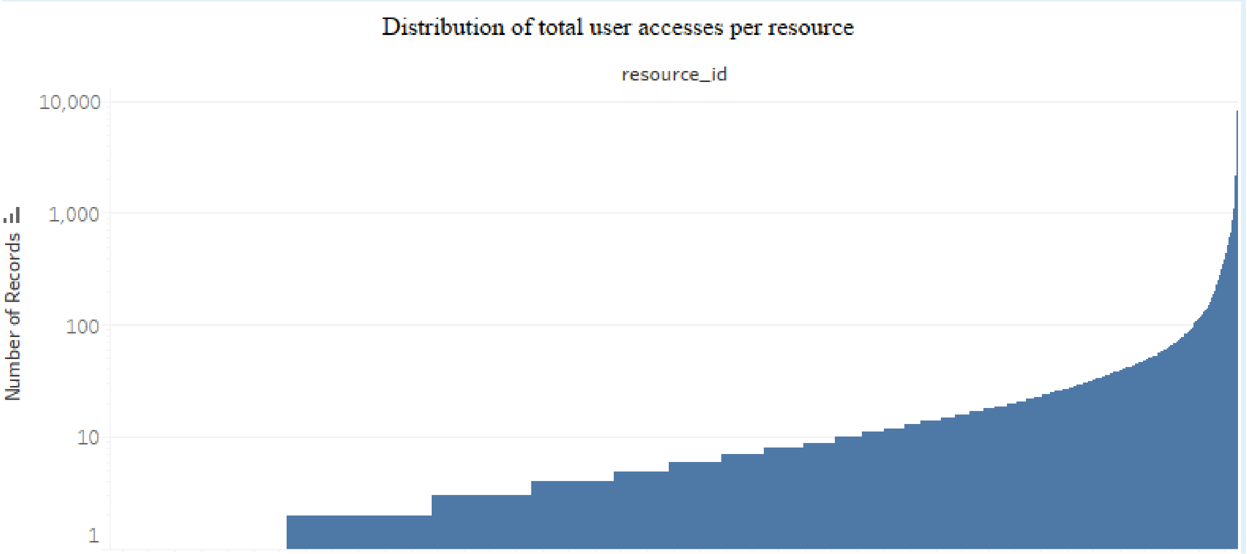


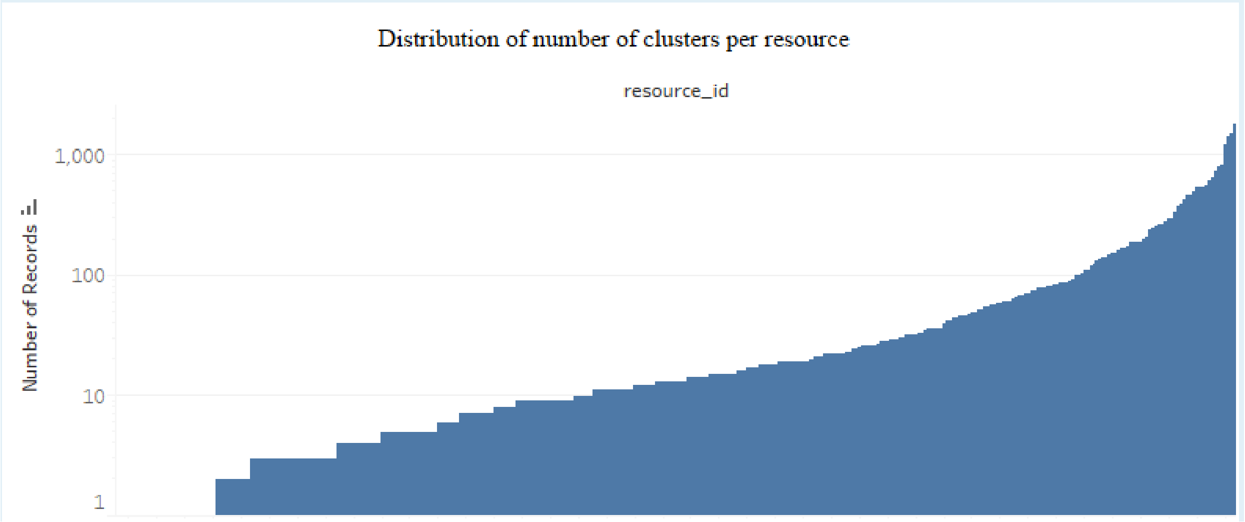
Distribution of number of clusters per resource

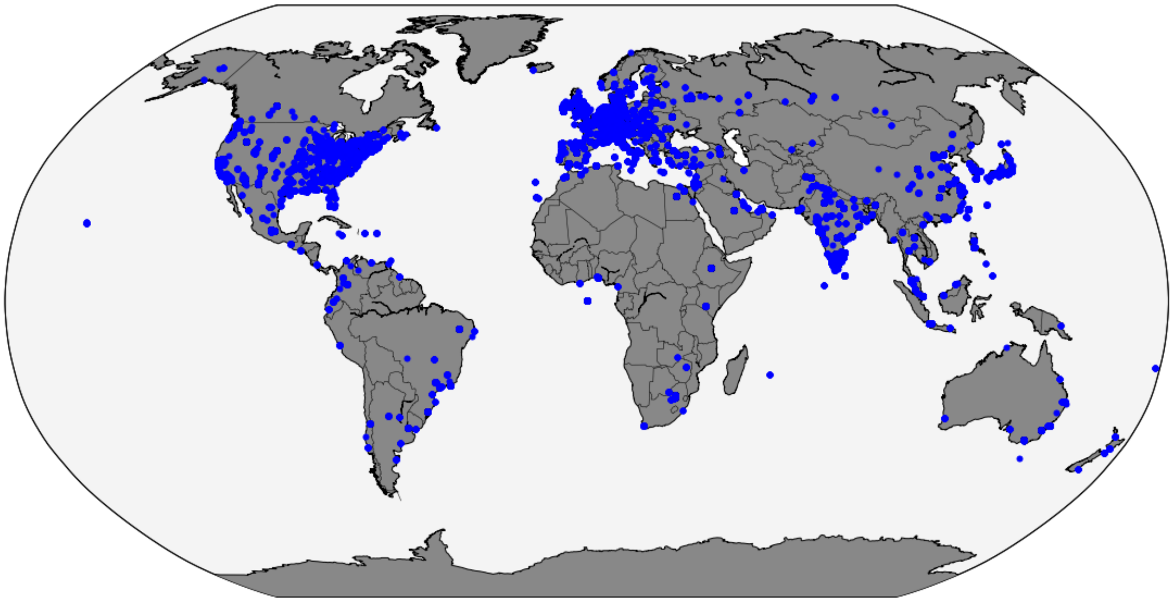


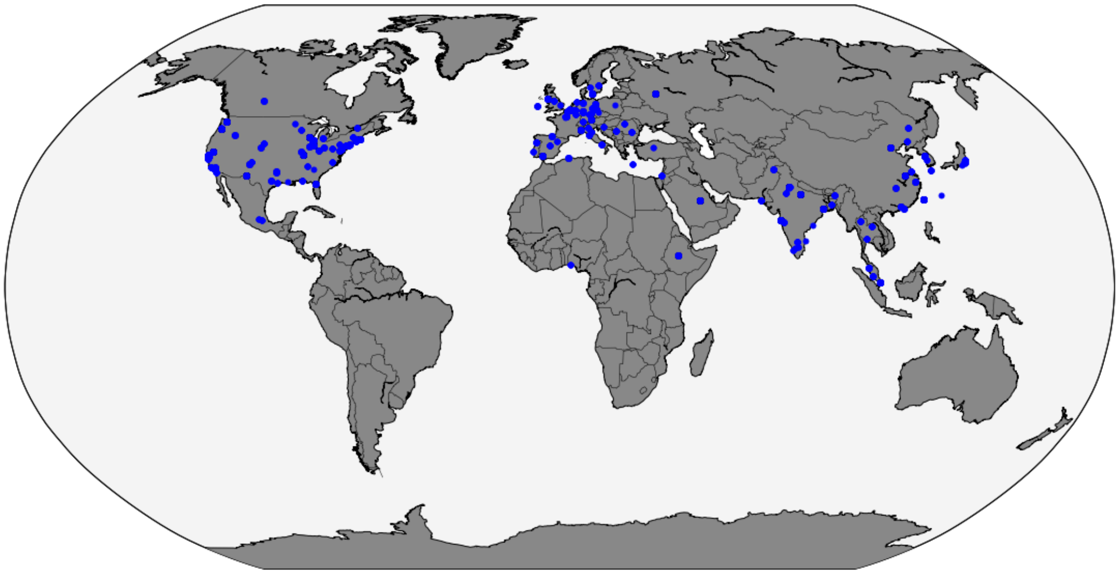


Data Summary 2015









Cardinality Analysis

The following tables each show a pair of resources that were found to be similar as a result of the clustering technique used in this study. The tables list the number of tags and common tags for each pair of resources.

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
[Illinois] DIY BIOSENSORS Day 1 Summer 2014 Workshop	Illinois nano/bio NanoBio Node NIDE workshop	8	1
Illinois ABE 446 Lecture 5: Self-Assembly and Bioconjugation	course lecture nano/bio self-assembly		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
[Illinois] DIY BIOSENSORS Day 1 Summer 2014 Workshop	Illinois nano/bio NanoBio Node NIDE workshop	15	2
[Illinois] BioNanotechnology Seminar Series Spring 2012: DNA Mediated Synthesis of Novel Gold Nanoflowers for Diagnostic and Therapeutic Applications	applications Diagnostics DNA Illinois Mediated MNTL nano/bio Novel Gold Nanoflowers synthesis therapeutics UIUC		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
[Illinois] DIY BIOSENSORS Day 1 Summer 2014 Workshop	Illinois nano/bio NanoBio Node NIDE workshop	12	2
[Illinois] Gold Nanostars as Tiny Hitchhikers for Cancer Therapeutics	cancer gold Illinois NanoBio Node Nanostars therapeutics Tiny		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
[Illinois] DIY BIOSENSORS Day 1 Summer 2014 Workshop	Illinois nano/bio NanoBio Node NIDE workshop	12	2
[Illinois] Gold Nanostars as Tiny Hitchhikers for Cancer Therapeutics	cancer gold Illinois NanoBio Node Nanostars therapeutics Tiny		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
Illinois ABE 446 Lecture 5: Self-Assembly and Bioconjugation	course lecture nano/bio self-assembly	13	1
[Illinois] BioNanotechnology Seminar Series Spring 2012: DNA Mediated Synthesis of Novel Gold Nanoflowers for Diagnostic and Therapeutic Applications	applications Diagnostics DNA Illinois Mediated MNTL nano/bio Novel Gold Nanoflowers synthesis therapeutics UIUC		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
Illinois ABE 446 Lecture 5: Self-Assembly and Bioconjugation	course lecture nano/bio self-assembly	10	0
[Illinois] Gold Nanostars as Tiny Hitchhikers for Cancer Therapeutics	cancer gold Illinois NanoBio Node Nanostars therapeutics Tiny		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
[Illinois] BioNanotechnology Seminar Series Spring 2012: DNA Mediated Synthesis of Novel Gold Nanoflowers for Diagnostic and Therapeutic Applications	applications Diagnostics DNA Illinois Mediated MNTL nano/bio Novel Gold Nanoflowers synthesis	17	1

	therapeutics UIUC		
[Illinois] Gold Nanostars as Tiny Hitchhikers for Cancer Therapeutics	cancer gold Illinois NanoBio Node Nanostars therapeutics Tiny		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
ECE 495N Lecture 5: Quantitative Model for Nanodevices II	course lecture electrostatics nanoelectronics	6	2
ECE 495N Lecture 15: Covalent Bonding	course lecture nanoelectronics quantum transport		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
ECE 495N Lecture 5: Quantitative Model for Nanodevices II	course lecture electrostatics nanoelectronics	7	3
ECE 495N Lecture 20: Bandstructures III	band structure course lecture nanoelectronics quantum transport		

Pair of Similar Resources	Tags	Total number of tags	Number of Common Tags
ECE 495N Lecture 15: Covalent Bonding	course lecture nanoelectronics quantum transport	7	3
ECE 495N Lecture 20: Bandstructures III	band structure course lecture nanoelectronics quantum transport		