Technical Report 2023-08-01
Internetworking and Media Communications Research Laboratories
Department of Computer Science, Kent State University
http://medianet.kent.edu/technicalreports.html

A Data Lake for Correlated Visualization of Network Connectivity and Scientific Productivity for World's Research Organizations

Javed I. Khan, Amjad Hossain, Amal Babour, Adam Petric, and Adam Tishler
Network and Media Communications Lab
Department of Computer Science, Kent State University
javed@ kent.edu | mhossai2@kent.edu | apetric1@kent.edu | ababour@kent.edu | atischle@kent.edu

***Abstract*— Recently, collaborative research has gained considerable attention among research organizations worldwide. Collaboration can increasing science productivity. Over the last two decades, spearheaded by the emergence of national and regional Research and Education Networks (REN) there has been major new investment in science cyberinfrastructure that is reshaping scientific collaboration. The central component of this new development is capable cyber-connectivity between institutions- that enables far-beyond human-to-human communication, it also is enabling very large scale science data sharing, and advanced scientific equipment. An interesting question is how to gain a data centric methodology understanding how global interconnectivity is reshaping the extent, intensity, pattern, and fundamental nature of global collaboration. This paper presents a datalake framework to study the potential correlation between large scale data transfer/communication infrastructure to research creativity and research productivity. To that effects the prototype brings relevant data from the globally deployed federated network telemetric infrastructure perfSONAR ad well as the MAG data set that bears one of the most comprehensive signatures of scientific productivity worldwide with record from more than 250 million publications. The prototype also developed a set of visualization tools to jointly visualize these datasets at the level of almost all (+99,609) research institutions worldwide including all higher education institutions (HEIs).**

Keywords*— *collaboration, science cyberinfrastructure, research productivity.

I. INTRODUCTION

Scientific collaboration is an ancient concept. Classically, research collaboration has been defined as research work conducted by multiple scientists, research groups whether from the same or different institutions, country, and disciplines or sectors.

It is long understood that collaborative research plays an important role in exchanging ideas, learning new skills, producing new information and knowledge, resource sharing, improving the quality, and enhancing the depth and the impact of the research.

In ancient Greece, philosophers and scholars often engaged in collaborative discussions and debates. While not necessarily formalized as "scientific collaboration" in the modern sense, the intellectual exchange in places like Athens and Alexandria laid the groundwork for the development of various scientific disciplines. During the Islamic Golden Age, 8th - 14th centuries, scholars in the Islamic world made significant contributions to various scientific fields, including astronomy, chemistry, mathematics, medicine, and optics. Greek and Roman texts were translated into Arabic, and scholars from different regions collaborated on these translations and built the foundation of modern knowledge. Monasteries in medieval Europe collaborated in scriptoria, copying manuscripts, and producing illuminated texts and helped preservation, dissemination, and transmission of scientific and philosophical ideas. The 17th century witnessed the establishment of scientific societies, for example, the Royal Society of London for Improving Natural Knowledge, founded in 1660, provided a platform for scientists to present their work and engage in discussions. There were also scientific exploration and expeditions, such as those led by James Cook, involved collaboration among scientists, naturalists, and navigators. These collaborative efforts contributed to the understanding of geography, biology, and anthropology.

Whether the cooperation is local or international, it has a great value in helping each of the research groups to see the research problem from different perspectives resulting in raising awareness and creating innovative solutions. Such experimental research in the field of life sciences, earth sciences, health sciences require using expensive technologies for experimental work as well as

data collections, samples and technical skills which refer to the importance of collaboration in resource sharing. This type of collaboration can result in the formation of relatively stable networks of researchers who interact frequently over a longer period of time.

Over the last two decades, there has been major new development in science cyberinfrastructure that is reshaping scientific collaboration. The internet itself emerged from the concept of sharing super-computers for US defense research. There is now enormous global effort and investment to build enabler science cyber-infrastructure to encourage researchers to collaborate. This is also fundamentally reshaping the nature and pattern of scientific collaboration.

The emergence of Research and Education Networks (REN) around the countries education systems has enabled research universities to be gradually connected to each other by dedicated networks. The RENs connect the higher-education institutions (HEI)- particularly the research universities within a country with a high capacity network. The regional and federations such as Internet2, GEANT, APAN, further has interconnected these national networks into a global network. Besides the HEI's these networks are also connecting specialized and high value science resources and equipment located in various HEI's as well as laboratories remote sharable by the researchers. These include advanced super-computers and large data centers to provide computation and data storage needs of the researchers.

More and more specialized facilities are attached. The varied type of scientific instruments being connected includes Particle Aaccelerators (used in high-energy and nuclear physics research for accelerating charged particles), Mass Spectrometers (for identifying and quantifying molecules in a sample), Nuclear Magnetic Resonance (NMR) Spectrometers (for understanding the structure and dynamics of molecules), Electron Microscopes (for high-resolution imaging of small structures, such as cells and nanoparticles), X-ray Crystallography Equipment (for determines the atomic and molecular structure of a crystal), Flow Cytometers (it analyzes and sorts cells based on various properties, including size, complexity, and fluorescence), Next-Generation Sequencers (for high-throughput DNA and RNA sequencing, Chromatography Systems (separates and analyzes mixtures of chemicals), Cryogenic Electron Microscopes Cryo-EM (high-resolution images of biological macromolecules), and increasingly being connected to the global science/REN networks.

The significance of hyperconnectivity for science research in bringing major investment and innovation. A recent example is the National Science Foundation's initiative to build the Data Transfer Network (DTN) with nodes around the HEI's that will create 40-100 Gbps fiber links with minimum impendence. It will enable very large data transfer between research institutions in order of magnitude faster significantly improving data centric research workflow.

Our objective is to gain a data centric methodology to study and understanding how emergence of global cyberinfrastructure- a capable network at the core, is reshaping the extent, intensity, pattern, and fundamental nature of global collaboration in a comprehensive way.

A key question is how to characterize collaboration as well as the cyber-infrastructure in a large scale.

There are many manifestations of collaboration. But bibliometric measures is currently considered to be capable of providing most insight across more disciplines and over long time span. De Haan (1997) suggested six indicators to measure collaboration between researchers in the field of social sciences and humanities: co-authorship, shared editorship of publications, shared supervision in PhD projects, writing research proposals together, participation in formal research programs, and shared organization of scientific conferences. Many patterns of research collaborations do not result in co-authored publications (Katz et al. 1997; Melin et al. 1996; Laudel 2002). Half of scientific collaborations are invisible, either because they do not result in co-authored publications, nor do they receive formal acknowledgments in scientific texts (Laudel, 2002). Yet co-authorship publications are part of the visible institutionalized structure of science, whereas informal communications are not.

Formal co-authorship is still likely most active and prevalent form of collaboration between researchers (Price, 1963) for the scale of the issue. There has been tremendous growth in the number of co-authored publications in all fields of science. The first co-authored scientific paper was published in 1665 (Lukkonen et al. 1992). The number of co-authored publications has increased dramatically in the second part of the 20th century and at the beginning of the 21st century. Bibliometric analyses have shown a continuous increase in the number of co-authored publications in nearly all scientific disciplines. These collaborations have occurred both within and across countries and regions within countries (see Rodriguez et al. 2008; Glaenzel et al. 2004 and Wray 2002). Recently, many arguments have been advanced to support the claim that the most important value of collaboration lies in the enhancement of epistemological authority. Included are arguments supporting the thesis that co-authored publications, because of the scientific collaboration involved, have greater epistemic authority than research performed by single individuals (Beaver 2004; Wray 2002).

A similar problem exists for characterization of the collaboration cyberinfrastructure. One of the basic indicators is the characteristics of the network performance, which is measured via throughput, ping delay, hop count, round trip time, one way delay,

availability etc. These are dynamic quantities, and a very different approach is needed. Fortunately the deployment of perfSONAR ("Performance focused Service Oriented Network monitoring ARchitecture ")[AH1], provides a rich set of network capacity information in a Global scale. For this initial framework we thus use the PerfSONAR data.

The main contributions of this project and this report is the following:

- Calculating network measurements between each pair of research organizations in the world.
- Calculating the amount of research publications between each pair of research organizations in the world.
- Finding the correlation between the number of shared publications and the network measurements.
- Developing a web-hosted visualization platform presenting the network measurements and the amount of research publications among each set of research organizations.

This report paper, however, does not provide any correlated analysis of the information presented. Which we plan to do in a future series of report papers.

The rest of the paper is organized as follows. The related work is provided in Section 2. Section 3 provides the architecture of the overall datalake framework designed to continuously intake and interface to major global databases to obtain the required data. Section 4 describes the technical details how the network status is acquired from the globally deployed PerfSONAR framework. Section 5 provides the technical details how research collaboration data is extracted from major bibliometric database- in this case Microsoft Academic Graph (MAG). Finally, Visualization System's interface is presented in Section 6 with samples.

II. Related Work

A. Network measurements Data & visualization

Computer networks are monitored to identify and prevent unexpected behavior of the networks. The relevant network metrics such as latency, jitter, packet loss, throughput etc are also collected and stored for future analysis. The common tools or techniques for network data collection are iperf [AH5], traceroute[AH6], ping, one-way ping[AH7][AH8] etc. The stored data is analyzed to plan on improving overall performance of the network in terms of security, speed, reliability etc. There are many works that focus on security aspects based on network data analysis. For example, [AH2, AH3, AH4] address intrusion detection systems by synthesizing, analyzing and visualizing network data. However, most of these work are for network administration and usually performed for managing organizational networks or single administrative domains. There are few attempts to collect, analyze, and visualize data about inter-organizational networks (or nodes on the internet) such as PingER[AH9, AH10] and perfSONAR ("Performance focused Service Oriented Network monitoring ARchitecture ")[AH1]. PingER started with the objective of networking monitoring to understand present performance and allocate resources to optimize performance between laboratories, universities, and institutes collaborating on energy nuclear and particle physics experiments. However, it uses only the ping tool that measures only the round trip time(RTT).

On the other hand, perfSONAR forms a distributed network of monitoring nodes that enables a interoperable network measurements framework where data are gathered and exchanged in a multidomain, heterogeneous, federated manner. It collects different network metrics including RTT, one-way delay, throughput etc using different tools including traceroute, ping, OWAMP, iperf3 etc. It facilitates cross-domain troubleshooting based on the historical data achieved in a distributed database maintained within the network. Many researchers and network administrators use data from Perfsonar for designing new networks, developing tools, or improving performance of existing networks etc. TWAREN (Taiwan Advanced Research and Education Network) uses PerfSONAR developed network performance weathermap system for early detection and analysis of network quality degradations and failures[AH12]. [AH13] presents a network anomaly detection and diagnosis scheme for network wide visibility using perfSONAR data. They use principal component analysis(PCA) to transform data for accurate correlated and uncorrelated anomaly detection. A study on simulating network throughput by correlating perfSONAR measurements with link utilization is presented in [AH13]. They utilize delay and packet loss data from perfSONAR network and developed multiple machine learning models to predict link performance.

In this paper, we use perfSONAR measurement data (RTT, one-way delay, throughput etc) to report and visualize link performance between different research and educational organizations. We also aim to show correlation between network data and research collaboration among the organizations. The findings can be used for identifying lack of collaborations and optimal allocation of network resources between different organizations.

B. Collaborative Research Data & Visualization

The scientific research collaboration network is one of the most representative complex networks. Scientific research work which is done by two or more scientific researchers is called a collaborative relationship. The scientific research collaboration network is built by connecting many different scientific

research works in terms of authors, department, universities, countries, journals, ... etc. by collaborative relationship. Usually, such a network connected by scientific research work is called a scientific research collaboration network which can be used as a visualization tool to show the internal structure of scientific research collaboration work [AB1].

Chuang and Chen [AB2] applied social network analysis (SNA) to visualize international research collaboration patterns of the faculty members from all Management Information System departments in Taiwan (MIST) from 1982 to 2015. The authors first retrieved a dataset of the publications of 1,766 MIS professors in the study period from the Ministry of Science and Technology of Taiwan (MOST) website. Then, they merged the retrieved dataset with datasets obtained from the Web of Science (WoS), Google Scholar, IEEE Xplore, ScienceDirect, Airiti Library and SpringerLink databases and removed the redundant publications. The new merged dataset includes information about every MIS professor with the following fields: (1) journal keywords, (2) authors' Chinese and English names; (3) affiliation; (4) authors' home departments, universities and countries; (5) authors' titles; (6) publication journal titles and (7) coauthors. The authors applied D3.js to visualize the faculty members' international collaboration from all MIST, where every node indicates the following: (1) author; (2) university; (3) country; (4) keyword and (5) research field. The connecting lines present collaborations among faculty members, the darker the line, the more collaborations among these faculty members.

Hu and Zjang [AB3] visualized the patterns of collaboration networks among disciplines that are involved in publishing Big Data research. The data used in their study was retrieved from the WoS core collection and filtered using the term ''Big Data'' in both the title and author-provided keywords for the study period from1950 to 2015. They used the Science of Science software (SCI2) to generate the co-discipline network file from the retrieved data, where nodes representing disciplines and relations between those disciplines are presented as links. Using Pajek and VOSviewers softwares, the generated file visualizes the interdisciplinary network.

Huang and Wang [AB4] visualized the pattern of collaboration networks among regions in library science (LS) in China.

Table-1. Organizations Dataset

CCIID_of_institution	The institution's id given by the project team
grid_id	The institution's GRID identifier.
Affiliationid	The institution's MAG identifier.
Name	The institution name
wikipedia_url	A link to the English Wikipedia article describing this institution,
OfficialPage_link	A link to the institution's official homepage
email_address	The contact email address of the institution
established	The year the institute opened.
acronym	A list of short acronyms the institute is known as
latitude	The latitude of the affiliation.
longitude	The longitude of the affiliation.
city	The city of the affiliation.
state	The state of the affiliation.
state_code	The state code of the affiliation.
country	The country of the affiliation.
country_code	The ISO country code is a short for the ISO 3166 a standard published by the International Organization for Standardization (ISO).
Iso3166Code	A code published by the International Organization for Standardization (ISO) that defines two letters' codes for the names of countries.
external_id	Other IDs known to refer to the institute.
external_id_type	The type of the external ID.
iso639	A code published by the International Organization for Standardization that is concerned with representation of names for languages.
label	The name of the institute in different languages.
OrgType	The type of the organization/institution.
businessStatus	
aliases	A list of other names the institute is known as
status	The status of the institute if it is active/inactive.
ip_addresses	The IP addresses known to belong to the institution.
close	The year the institute closed.
continent	The continent of the affiliation.
PaperCount	The number of papers associated with the institution.
PaperFamilyCount	The number of primary family papers associated with the institution.
CitationCount	The number of citations of the institution.

They retrieved the publication data from the Journal of Library Science in China (JLS) from 2006 to 2015. By mapping authors' affiliations to regions (provinces and municipalities) and using SCI2, the co-region data file was generated where nodes represent regions, a link between nodes represents the relationship between two authors from different regions who have collaborated on at least one paper. The collaboration between the regions was visualized using the generated co-region data file, Pajek and VOSviewer software.

The Health Sciences Library (HSL) at the University of North Carolina at Chapel Hill (UNC-CH) conducted a collaboration visualization for the Cancer Cell Biology (CCB) research at the UNC Lineberger Comprehensive Cancer Center. They retrieved the CCB publication data

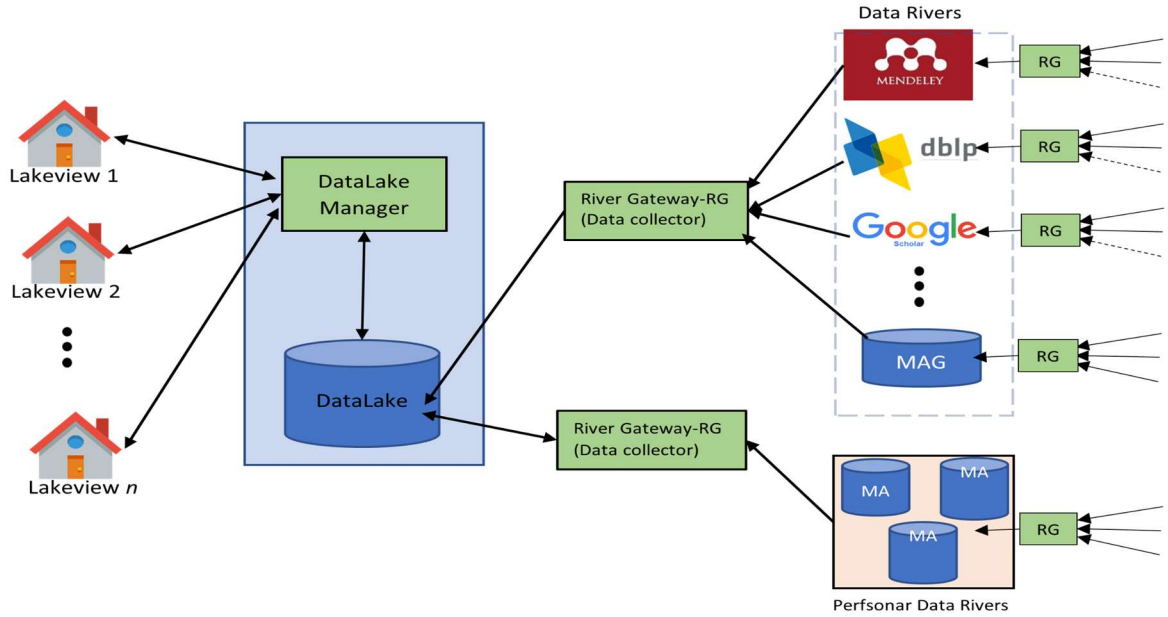

Fig-1 System Overview

from the Scopus database from 2010 to 2014. They utilized Tableau software to visualize the publication productivity and the average cite per paper per year. The same software was used to visualize the comparison citation among the Field-Weighted Citation Impact (FWCI), Citation Benchmarking (CB), Relative Citation Ratio (RCR), National Institutes of Health (NIH) of CCB publications. While VOSviewer software was used to visualize the co-author, country, and internal and external institutional collaboration networks and to produce the research topic network, and the topic density maps [AB5].

On the other hand, the authors in [AB6] developed a new visualizer called InterRing that shows in-depth sight of collaboration. Using the DBLP database, the tool extracts the co-authoring data and shows the weight of co-authorship of a researcher with other researchers in a given period and the knowledge domain of the researcher publications published in a past period. The tool presents the data in a series of connecting rings. Each ring represents a particular year and each sector in the ring represents a co-author identified by a color resulting in clarification of what researchers in which year have collaborative publications.

organization information

Information about the collaborative research organizations such as organization id, organization name, official page, latitude, longitude, etc are received from two merged sources which are Affiliation dataset belongs to MAG and Global Research Identifier Database (GRID) datasets.

1). Grid database

Global Research Identifier Database (GRID, https://www.grid.ac/) is an open free database that includes data about a collection of worldwide institutes associated with research organizations. The data is presented in ten datasets: acronyms, addresses, aliases, external_ids, geonames, institutes, labels, links, relationships, and types. For this paper, GRID (December 09, 2020) was used. The datasets were downloaded from the Grid website [*].

To receive information about the research organizations, eight datasets were utilized. The acronyms dataset lists short acronyms the institute is known as (e.g. KSU for Kent State University). The addresses dataset records the addresses associated with the institute such as city, state, country, latitude, and longitude. The aliases dataset lists other names the institute is known as (Kent State for Kent State University). The external_ids dataset lists ids known to refer to the institute other than grid id. The institute's dataset lists information about all institute records such as institution name, Wikipedia url, and established year. The link dataset lists the homepage for each institute. The types dataset lists types describing the institute (e,g, education, government, nonprofit, etc).

After merging the datasets, some MAG organizations were not having grid ids. So, a new identifier attribute is given by the project team to each organization resulting in an organization dataset covers information about research institutions in 31 attributes presented in Table-1.

III. The Framework OverView

The framework is designed based on the concept of lake park where the lake accumulates water from different creeks and rivers. The lake administrator can build different lakeviews for the park visitors. The

framework is designed around a Datalake as shown in Fig-1.

The Datalake stores the data about shared publications between different universities and the network measurement data between many pairs of locations around the globe. The data river gateway is a data socket for collecting data from different relevant data rivers(sources). It can accumulate data from one or more rivers such as google scholar, dblp, MAG etc and store the collected data in the DataLake (a central database). It also works as a data aggregator that collects data from the distributed data rivers (Measurement Archive) in the Perfsonar network.

The DataLake manager is a web based application that can generate one or more visualizations/views based on the data stored in the DataLake. The visitors to the DataLake can observe the different views (Lakeviews) of the DataLake through the web application. The Lakeviews help to understand the historical trend of networks and academic collaborations between different research entities and identify areas of potential improvement.

IV. Network Status

To understand the historical status of the network links between the college, universities and research organizations, the data about different network metrics such as link delays, throughput, packet loss etc are needed. The collected data can be filtered, analyzed and visualized.

1. Feeder Data stream: Network Measurements

1.1 Data of interest

We want to observe the historical status of network between different organizations. For that we collect delays such as Round Trip Time(RTT), One Way Delay(OWD) and the throughput between these organizations.

1.2 Description of data sources

In this project, we consider all the research organizations in the world(total *). We need network data between every possible pair of organizations which is not readily available. Fortunately, perfSONAR ("Performance focused Service Oriented Network monitoring ARchitecture ") is the closest framework that harvests pairwise network data and has the potential to match our requirements. It is an open source toolkit for running network measurements across multiple domains [AH1].

There are thousands of perfSONAR instances/nodes deployed worldwide, usually in the research organizations or universities. Many of these nodes are available for open testing of key measures of network performance. In the perfsonar network, any node can run tests to other nodes for measuring different network metrics including Round Trip Time (RTT), One Way Delay (OWD), Throughput, Packet Loss etc. The measured data values are stored in distributed databases deployed in PerfSONAR Measurement Archive (MA) servers. Each record in the database corresponds to a pair of IP addresses (Source IP and destination IP). Thus, the collected data from this global distributed infrastructure can be used to analyze current network status between different PerfSONAR measurement nodes. i.e. corresponding research organizations. So, we collect available network data from the many distributed data rivers (Measurement Archives of PerfSONAR), store them in the DataLake and visualize them upon user's filtered requests.

1.3 Data River Gateway/Socket

The data river gateway/socket for the network data is an independent module written in python. It can collect data from the distributed MAs of PerfSONAR. Using the gateway module, we collect data for throughput, Round Trip Time(RTT) and One Way Delay(OWD) and store them in the Datalake using IP-organization mapping.

1.3.1 Data Record Structure

The structure of the data record on PerfSONAR MAs is hierarchical as shown in Fig-2. The level 1 metadata contains the source and destination IP addresses, names of the tool used to collect this record, types of events/query, uri to the level 1 metadata etc.

The level 2 metadata contains the list of event names, base uri to the metadata, summary names (statistics, aggregations) etc. The level 2 metadata is obtained based on the tool-name and the related event generated by the tool. Because the tool-name tells what kind of data is stored in the record. The table-2 summarizes the available tool-names, related event names and corresponding data types. For example, if the tool_name in a record is 'powstream', then get_event_type ("histogram-owdelay") returns an **event type object** that can be used to collect metadata of the next level.

The event object has get_data() and get_all_summaries() functions that can be used to collect third level metadata. The third and final level of metadata contains the objects that store references to the series of data points. Each datapoint has one or more values associated with the timestamp of data collection. The values can be raw measures or the statistical measures such as mean variance, standard deviation etc.

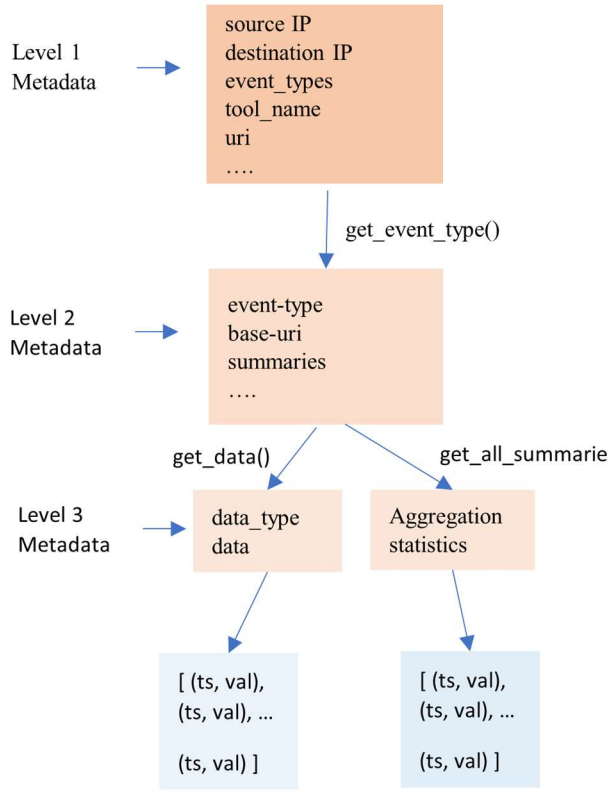

Fig-2 Data Record Structure

Table 2: List of Tools and corresponding events, and data types

Tool Name	Event Type	Data Type
Traceroute	packet-trace	RTT
powstream	histogram-owdelay	Oneway delay
ping	histogram-rtt	RTT
iperf3	throughput	throughput

1.2.2 Extraction Engine

The main component of the data river gateway is the extraction engine, Fig-3. It queries every MA one by one using their hostname or the IP address and collects all the records stored in the MA. The extraction engine follows the record structure to collect the data of our interest.

To collect different network metrics, it needs to use tool names from the top level metadata of each record. So it collects all the possible tool names first and stores them in a file. Then the engine collects data from all the MAs using the following algorithm.

ExtractionEngine (IPsOfMA, toolNames)

*1. **for** each IP in the IPsOfMA*
2. connect to the server using IP
3. get the level1 metadata for all records
*4. **for** metadata of each record*
5. et = get_event_type(tool-name, event-name)
6. data-reference = getdata(et)
7. collect data using data-reference
8. calculate or collect aggregated measures
9. store data as (source IP, destIP, value, type)
*10 **return***

Fig-3 Extraction Engine

1.2.2 Upstream/ Downstream attribute convergence

The data collected from the PerfSONAR are stored as source IP and destination IP pair. These IP addresses are for different PerfSONAR nodes that are located in different universities or research organizations. So, we perform *organization to IP* mapping so that the SourceIP-destinationIP pair data can be used for SourceCCI_ID - destinationCCI_ID pair (Fig-4). However, many organizations don't have any PerfSONAR nodes installed. So, we collect and use location information (latitude and longitude) of PerfSONAR nodes(using IP) and the universities to find the closest PerfSONAR node of each of the organizations. Finally, the data pusher uses the CCI_ID, IP mapping to push the collected data to the DataLake based on SourceCCI_ID - destinationCCI_ID pair.

1.3.3 Timing and space complexity

For the PerfSONAR river gateway, there are two external input files, list of IPs of MA and the list of organizations. The time and space complexities of the gateway will depend on the size or number of entries in these files. Currently, there are around 2000 MA servers and 100K organizations in these files. As the MA servers store the historical data, the time and space complexities mainly depend on the time frame we want to collect data from. The current system collects data for the last 6 hours from the current time which takes around 7 hours to finish

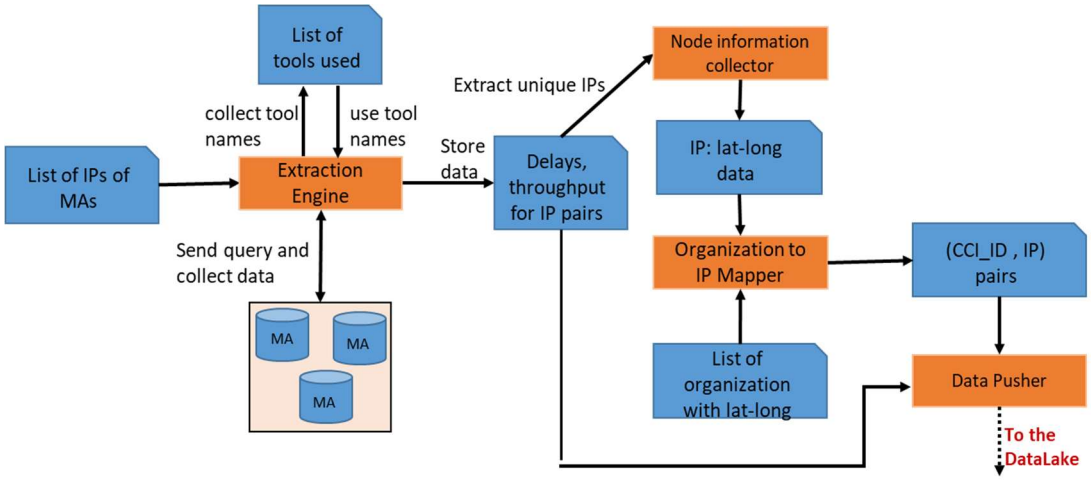

Fig.4. PerfSONAR to Datalake Data Flow Diagram.

data collection. The size of the collected data is around 20MB. The dataset provides around 7000 usable records for SourceCCI_ID - destinationCCI_ID pairs. If the time frame is increased, the dataset size and the collection time both will increase. The usable records will also increase with repeated records between organizations from different times/dates forming historical data.

The historical data is important to understand the change of network status. Considering the time complexity of data collection, we consider updating the network data of the DataLake every six months. This will also improve the historical data.

V. Research Collaboration

One of the most important pieces of knowledge referring to the collaboration research among the research affiliations is the number of shared publications among them. This type of knowledge is based on the availability of a dataset of scholarly publications containing detailed and structured information about publications affiliations such as affiliations' names and affiliations' countries to avoid ambiguity among affiliations holding the same name.

The scholarly publication has increased every day in terms of the volume of papers being published and collaboration of authors and affiliations. However, this causes a difficulty in quantifying the number of publications and collaborating of authors and affiliations. Although there are several quality scholarly knowledge databases which are free to use, they are often incomplete and lack essential information. For example, DBLP and Semantic Scholar do not have information about authors' affiliations. Crossref has a field for affiliations in its dataset but still the majority of publications lack this type of information. IEEE Xplore Digital Library has a field for affiliation in its dataset but exhibit several ambiguity issues in affiliation names such as (i) alternate names (e.g, "University of Akron" and "The university of Akron") (ii) different granularity and missing information (e.g., "Kent State University", "Department of Computer Science, Kent State University" and "Media Communications and Networking Research Lab Department of Computer Science Kent State University"), and (iii) Linguistic differences (e.g., "Polytechnic Institute of Setúbal") [AB7]. Google scholar (GS) has a field in its database about publication affiliations, but it does not make direct access to its database via an Application Program Interface (API) and the data is only available via the search portal which makes extracting the data a challenge [AB8]. On the other hand, several pay-walled databases offer information about publication affiliations such as: Web of Science (WOS) and Scopus that restrict access to their databases to paying subscribers via their API's [AB8].

2. Feeder Data stream: Academic Collaboration

2.1 Description of data source

2.1.1 Scopus

Scopus is a subscription-based abstract and indexing database that was produced by Elsevier. It covers scientific journals, books, and conference proceedings and receives daily updates covering more than 70 million publications [AB15, AB16, A17].

2.1.2 Google Scholar (GS)

Google Scholar is a free indexing scholarly database provided by Google. It provides information about journal articles, proceedings, theses, dissertations, books, book chapters, reports, manuscripts, newsletters, encyclopedia entries, government documents, and patents, including documents in many languages. It receives updates as soon as a new publication is released, providing more than 389 million documents in January 2018 [AB14]. The coverage of google scholar database is considered too broad and non-specific and the quality of its indexed data remains an issue. [AB12, AB13].

2.1.3 Microsoft Academic Graph (MAG)

Microsoft Academic Graph (MAG), is a downloadable, largest free to use scholarly database licensed under ODC-BY 1.0 published by Microsoft. It provides information about papers, authors, journals, conferences, affiliations, and citations. It receives regular updates every 1 or 2 weeks, providing more than 250 million scientific publications as of January 2021 [AB9, AB10].

In this study MAG database was chosen because it has the largest coverage of publications, including journals, conferences, books, and patents, when compared with Web of Science (WoS) and Scopus. In addition, it offers free and unrestricted access to the complete list of publications [AB11].

2.2 Data River Gateway

2.2.1 Extraction Engine

The following are the Azure platform set up that need to be performed to get access to MAG.

1. Create a Microsoft Azure account
2. Create an Azure data share service
3. In the created Azure data share service, Create an Azure storage account.
4. In the created Azure storage account, Create an Azure Blob Container.
5. Sign up for MAG provisioning

Microsoft Academic reviews the application. Then it sends an invitation through Azure Data Share for receiving MAG datasets. After accepting the invitation, MAG will be uploaded to the created Blob container located in the storage account.

6. In the Azure portal, create a databricks service.
7. Launch a workspace in the created Azure databricks.
8. Create a Spark Cluster in the created Azure databricks.
9. Create a notebook in the created workspace.

Azure Storage is a cloud storage system optimized for storing massive amounts of unstructured data. It was used for the storage of MAG datasets and for the storage of the extracted datasets for the shared publications between each two affiliations. Azure Databricks and Azure Blob Storage. Azure Databricks is a cloud Apache Spark programming platform used to process massive amounts of data and supports code written in Python, R Scala, Spark, and SQL. On Databricks, codes are written in notebooks. The notebooks consist of a collection of cells where code can be written. Each cell can contain only one coding language while a notebook can contain cells of different languages. Codes written on notebook cells are run on clusters with custom settings and resources.

Microsoft Academic distributes the database for free through Microsoft's Azure Storage, but they charge for storage of the data and any computation done on the Azure platform [AB9, AB10]. When working with Databricks, the configuration of the Spark Cluster affects the performance of the algorithm and the executing time. For this paper, MAG (October 22, 2020) was used. The suggested configuration of the clusters created on Databricks is presented in Table 3.

Table 3. Configuration of the created cluster on databricks.

Databricks Runtime Version	9.0 (includes Apache Spark 3.1.2, Scala 2.12)
Worker Type	14 GB Memory, 4 Cores
Driver Type	14 GB Memory, 4 Cores
Number of minimum workers	2
Number of maximum workers	8

To extract the number of shared publications between any two research affiliations by year, three datasets from MAG are utilized: Affiliation, PaperAuthorAffilia-tions, and Papers. The Affiliation dataset records institution-related information such as affiliation id, name, grid id, official page and country. The PaperAuthorAffiliations dataset lists information about the authors and the affiliations of each paper such as paper id, author id, affiliation id and author sequence number. The Papers dataset consists of paper information such as title, digital object identifier (DOI), publication year, and publisher. The database schema among the utilized datasets from MAG are presented in Figure 5.

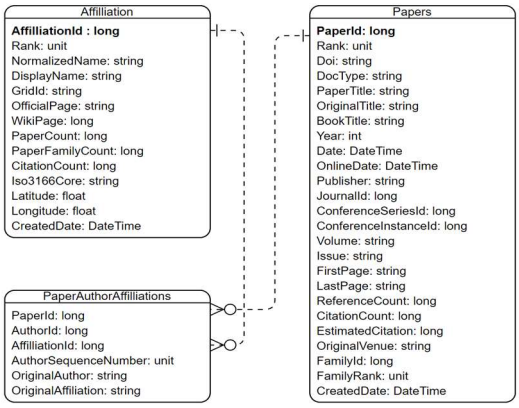

Figure 5. Subset of MAG database schema.

After completing the setting up tasks, execute the following codes in the created notebook as follows. In this research, the code for collecting the data was written in Python language.

1-Initialize the storage account and the container details created in Azure Storage. Replace the variables' values with the one created in your azure account.

```
AzureStorageAccount = '<AzureStorageAccount>'
AzureStorageAccessKey = '<AzureStorageAccessKey>'
MagContainer = '<MagContainer>'
```

```
OutputContainer = '<OutputContainer>'
```

2- Import samples/pyspark/MagClass.py uploaded in the created container. Then, define the imported class in the notebook.

```
%run "./MagClass"
```

3- Create a MicrosoftAcademicGraph instance to access MAG dataset located in the Azure storage.

```
MAG
=MicrosoftAcademicGraph(account=AzureStorageAccount,
key=AzureStorageAccessKey, container=MagContainer)
```

4- Create an AzureStorageUtil instance for saving the collected data.

```
ASU=AzureStorageUtil(container=OutputContainer,
account=AzureStorageAccount, key=AzureStorageAccessKey)
```

5- Import python libraries

```
from pyspark.sql import functions as F
from pyspark.sql.window import Window
```

6- Read Affiliations dataset stored in Azure Storage.

```
Affiliations = MAG.getDataframe('Affiliations')
Affiliations = Affiliations.select(Affiliations.AffiliationId)
```

7- Read PaperAuthorAffiliations dataset stored in Azure Storage.

```
PaperAuthorAffiliations =
MAG.getDataframe(PaperAuthorAffiliations)
PaperAuthorAffiliations =
PaperAuthorAffiliations.select(PaperAuthorAffiliations.PaperId,
PaperAuthorAffiliations.AffiliationId)
```

8- Read Papers dataset stored in Azure Storage.

```
Papers = MAG.getDataframe(Papers)
Papers = Papers.select(Papers.PaperId, Papers.Year)
```

9- Get all the paper id's published by x_i affiliationId.

```
paperId= PaperAuthorAffiliations
.where(PaperAuthorAffiliations.AffiliationId== xi) \
.select(PaperAuthorAffiliations.PaperId).distinct()
```

10- Get all the affiliationIds' for the paperIds found in step 9.

```
Paper_affiliations = PaperAuthorAffiliations
.where ((PaperAuthorAffiliations.PaperId.isin (paperId))) \
.select(PaperAuthorAffiliations.AffiliationId,
PaperAuthorAffiliations.PaperId)
```

11- Get the number of publications between x_affiliationId and the other affiliations by years.

```
shared_publication= Paper_affiliations \
.join(Papers, Paper_affiliations.PaperId==Papers.PaperId, 'inner')\
.select(Paper_affiliations.AffiliationId, Papers.Year)\
.groupBy(Paper_affiliations.AffiliationId,Papers.Year)\
.count()
```

12- Save the result in the Azure Storage.

```
ASU.save(shared_publication, filename, coalesce=True)
```

The previous steps get the data for one affiliation id. To find the data between each pair of affiliationId's in the Affiliations dataset, step 9 to 13 are executed in a for loop of size equals to the total number of affiliations in the Affiliation dataset as presented in Figure 6. For each row in Affiliations, the code assigns the AffiliationId in x_i. In line 4, the code gets all the paperIds published by x_i and saves them in paper_id. In line 6, it gets all affiliationIds that have shared publication with x_affiliationId and saves the result in Paper_affiliations. In line 8, it gets all AffiliationIds of the affiliations that have shared publications with x_i, the number of shared publications between x_i and the other affiliationIds by year and volume and saves them in the shared_publication dataset. In line 10, it adds a new column for x_i and names the attributes. Then, the shared_publication is saved in .csv file format in the Storage container as presented in line 11.

```
1.  for row in Affiliations.collect():
2.    xi=row['AffiliationId']

3.    # get all paper id's for xi
4.    paperId= PaperAuthorAffiliations
      .where(PaperAuthorAffiliations.AffiliationId== xi) \
      .select(PaperAuthorAffiliations.PaperId).distinct()

5.    // get all the affiliationIds' for the paperIds
6.    Paper_affiliations = PaperAuthorAffiliations
      .where ((PaperAuthorAffiliations.PaperId.isin
    (paperId))) \
      .select(PaperAuthorAffiliations.AffiliationId,
      PaperAuthorAffiliations.PaperId)

7.    // get the number of publications between xi and the
    other  universities  by years
8.    shared_publication= Paper_affiliations \
      .join(Papers,
    Paper_affiliations.PaperId==Papers.PaperId,  'inner')\
      .select(Paper_affiliations.AffiliationId, Papers.Year)\
      .groupBy(Paper_affiliations.AffiliationId,Papers.Year)\
      .count()

9.    shared_publication=shared_publication.
    withColumn('AffiliationId_1', lit(xi)) \
      .select('AffiliationId_1',  'AffiliationId_2', 'Year', 'count')

10.   # Save the result in the blob container
11.   ASU.save(shared_publication, filename,
    coalesce=True)
```

12. **End**

Fig.6. Data collection code.

Let us consider x_i = '149910238', the affiliation id for Kent State University. Some of the return output by the code is shown in Figure 7, where '58956616' is the affiliationId for Case Western Reserve University, '185163786' is the affiliationId for King Abdulaziz University, and '19820366' is the affiliationId for Chinese Academy of Sciences.

AffiliationId_1	AffiliationId_2	Pulication_Year	Publication_count
149910238	58956616	1960	1
	…….		
149910238	58956616	1988	5
149910238	58956616	1996	2
	…….		
149910238	58956616	2017	37
149910238	58956616	2018	16
149910238	58956616	2019	22
149910238	58956616	2020	18
	…….		
149910238	185163786	2015	3
149910238	185163786	2016	4
149910238	185163786	2017	1
149910238	185163786	2018	4
149910238	185163786	2019	3
	…….		
149910238	19820366	1987	2
149910238	19820366	2006	9
	……		
149910238	19820366	2018	20
149910238	19820366	2019	11
149910238	19820366	2020	14

Fig. 7. Example of the number of shared publications between Kent state university and other affiliations by year.

2.2.2 Upstream/ Downstream attribute convergence

An important aspect is the convergence of the upstream and downstream data nomenclature. This is not discussed for the time being.

2.2.3 Timing and space complexity

The data collection process for the whole set of affiliation ids was accomplished in around two months, from December 18th, 2020 to February 17th, 2021. The size of the created shared_publication dataset is about 761 MB containing 17,300715 records about the number of shared publications between pairs of affiliations/institutions by year, where each record needs around 44 byte to be saved. Figure 4 shows the collected data volume over time.

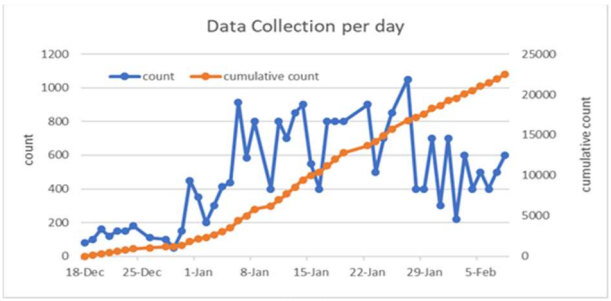

Fig.8 Data collection over time.

The executing time (in seconds) of the code presented in Figure 8 was calculated on a sample of 500 affiliation id collected on Feb 3, 2021. The minimum, maximum, average time to collect the data for an affiliation id as well as the total time needed to collect the data for the 500 ids are shown in Table-4.

Table 4. Execution Time

Minimum	90 s
Maximum	199.5 s
Average	134 s
Total Time	66847 s

The generated shared_publication dataset is supposed to be updated every year to include the count of organizations' shared publications that were published during the year.

VI. VISUALIZATION

All of this data is meaningless without a means to visualize it. Thus, the next step in the process is to create a web application that can visualize the data. We chose a geographical connection map, which is used to show network connections laid over a geographical map. In this style visual, individual affiliation can be represented as nodes on a map. Relational data with other affiliations on the map are represented as lines connecting two institutions with relevant relational data. In our visual, these lines vary in color depending on the quantity of the data in comparison with other connecting lines.

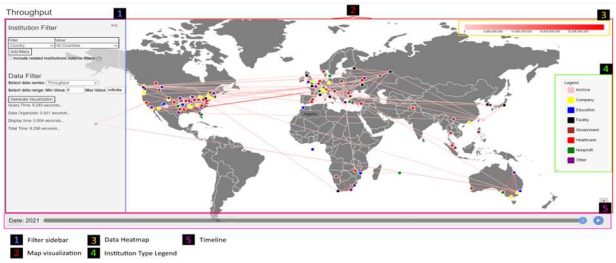

Fig. 9a connection map used in the visualization. the larger the quantity the darker the line.

Our visualization consists of five main parts (shown in figure 9a). First, the main geographical visual map takes up the majority of the space and is where the relational data (nodes and connecting lines) is displayed. Second, the filter side bar on the left of the visual is where

the user can set filters on the data and generate the visualization based on those filters. This sidebar can be collapsed to see more of the map. Third, the heatmap on the top right of the visual graphs out the changing color of the relational data lines. The numbers on this heatmap changed based on the smallest and largest value in the current filtered data set. Fourth, the institution legend, displayed on the right side of the visual, tells the user which colors correspond to each institution type for the nodes on the map. Finally, the fifth element of the visualization is the timeline located on the bottom. The timeline acts as a scrub bar that lets the user quickly scrub through the data based on year.

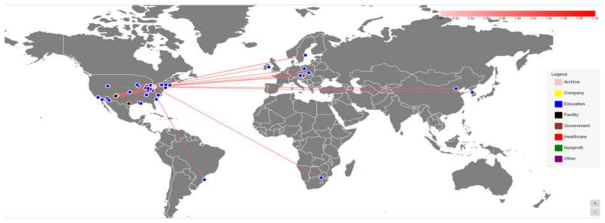

Fig.9b connection map used in the visualization- the larger the quantity the darker the line.

The application was built on the NET Core framework. Many elements of the visualization were built with the assistance of amCharts 4, a javascript library designed for various data visualization needs. The map projection, line series, heat legend, and connecting nodes were all realized with the help of this library.

1). Filters

The sheer scale of the data in our database presented a problem for our visualization. Displaying all the data at once makes the visual difficult to read and is resource intensive. The solution is to let the user filter what data they want to see.

All of the functionality for these filters is located within the collapsable filter sidebar on the left of the visualization. There are two main filter terminals within this sidebar: the "Institution Filter" and the "Data Filter".

The Institution Filter is an expandable list of filters that the user can use to specify which institutions the user wants to include in the visualization. These institution filters are "Country", "State (United States)", "Institution Name", and "Institution Type". The user can include as many of these institution filters as they want to further specify what institutions they want to see in the visualization. For instance, if you want to see shared data between "Kent State University" and "The University of Akron", you would add two "Institution Name" filters and set them to "Kent State University" and "The University of Akron" accordingly. Below the institution filter table is the "Include related institutions outside filters" button.

If this button is turned on, it will show institutions outside the institution filters from the institution filter that share data with institutions within the filters. For example, if you want to see every institution only Kent State University has shared publications with, then you would set an "Institution Name" filter to "Kent State University" and turn on the "Include related institutions outside filters" button. Even though only Kent State University is included in the filters, the "Included related institutions outside filters" button allows you to display any institution that shares data with institutions within the filters (In this case, the only institution in the filter is Kent State University).

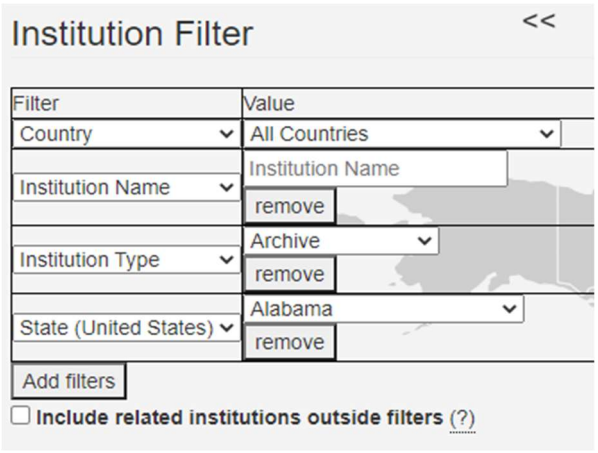

Fig.9c expandable institution filters

The data filter (figure 9d) located below the institution filter allows for three inputs from the user. The "Select data series" selection box allows the user to select which data series they would like to see. Currently, this includes shared publication data and various kinds of network data. Below this are two inputs for selecting the data range: min value and max value. These two inputs allow the user to set a minimum and maximum data range. For example, setting the data series to "Shared Publications", the minimum value to 100, and the maximum value to infinite, will show institutions who share over 100 publications with each other.

Fig. 9d Data Filter

Once the user has specified what filters they would like to apply, he/she can press the "Generate Visualization" button located below the filters to generate the visualization. Internally, the application takes many steps to query data from the database based on these user set filters. First, the application iterates through the selected filters in the institution and data filter and stores the values from them to a list. The application then runs an SQL stored procedure named "FilterLinks" using the

filter values gathered in the list as parameters. This stored procedure gathers the data from the database based on the user specified filters and sends them back to the application in JSON format. This JSON data is then separated into line data (for the connecting lines) and institution data (for the institution nodes) in a format recognizable by amCharts. This properly formatted data includes geographical coordinates, the data value of the data (e.g number of shared publications), and additional data not necessary for amCharts but helpful to the program. This now properly formatted line data and institution data is then put into the built in lineSeries and imageSeries classes of amCharts. "lineSeries" is a class used by amCharts to draw lines while "imageSeries" is a class used by amCharts to draw images. By putting properly formatted data into these classes, amCharts automatically draws the connecting lines and institution nodes on the map specified by the formatted data fed into it. Figure 10 shows a visual flow chart representation of this entire process.

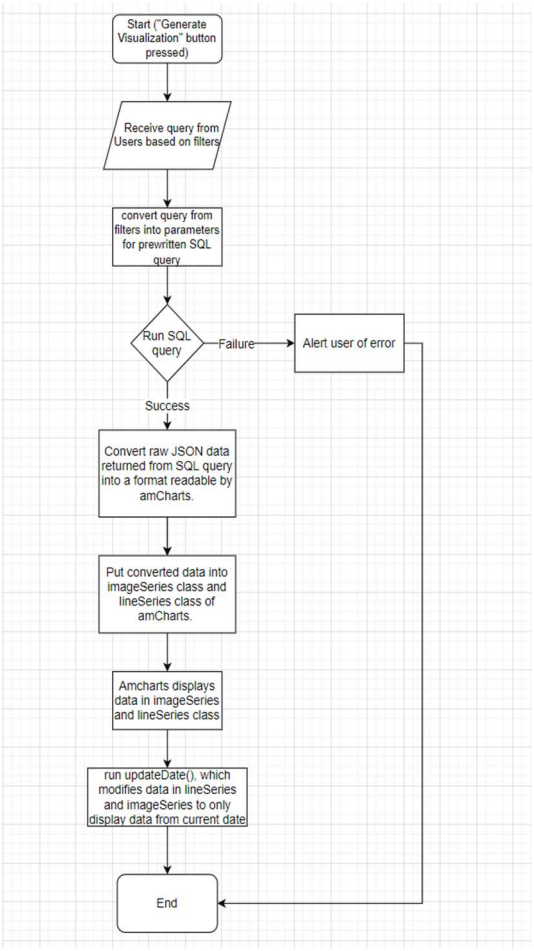

Fig.10 Flowchart of process of generating visualization

2). Timeline

An important part of the visual is being able to see how the data changes over time. Once the visual is generated, the user can use the timeline scrub bar (#5 in figure ?) at the bottom to quickly scrub through different years of the data. The earliest and last date on the timeline is the earliest and last date of the current queried data respectivley. The timeline is set to play by default and goes through each year sequentially. This can be paused at any time and can be manually scrubbed in a fashion similar to most common video players. Much of the timeline functionality is provided by amCharts. However, updating the actual data within the map is done manually by the application outside of amCharts.

3). Database Schema

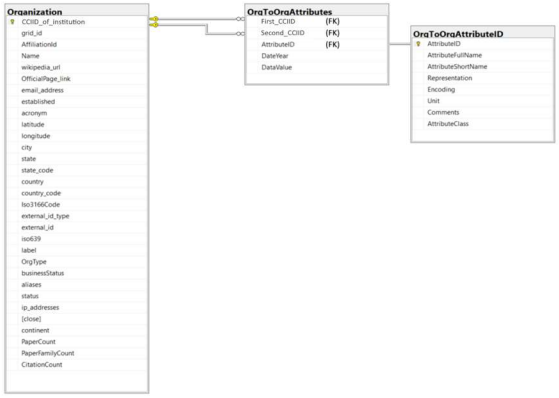

Fig. 11 Database schematics

Figure 11 shows the schematics of the database. The "Organization" table contains every institution in the database and their attributing static data. This static data contains data relating to the geographical position and various other information. This attributing data is what allows the visualization program to let the user filter through institutions. Each institution has a unique ID "CCI_ID", which is the primary key in the relationship.

The table "OrgToOrgAttributes" contains (Fig-12) the shared data between institutions, including collaborative research and shared network data. It contains fields for two IDs (First_CCIID, Second_CCIID) that reference the organizations in the linked Organization table, as well as the data (DataValue) and the related year the data is from. DataValue by itself is just a number. AttributeID is what gives context to what kind of data DataValue is by linking the data to an attribute in the "OrgToOrgAttributeID" table. OrgToOrgAttributeID contains information about the data, including what kind of data, the unit it is measured in, and more.

	AttributeID	AttributeFullName	AttributeShortName	Representation	Encoding	Unit	Comments	AttributeClass
1	1	Throughput_powstream	TPS	NULL	NULL	NULL	NULL	NULL
2	2	PingDelay	PDelay	NULL	NULL	NULL	NULL	NULL
3	3	CoauthoredPapers	CAP	NULL	NULL	NULL	NULL	NULL
4	4	Text_MbpsPerMonth	NULL	NULL	NULL	NULL	NULL	NULL
5	5	Audio_MbpsPerMonth	NULL	NULL	NULL	NULL	NULL	NULL
6	6	Video_MbpsPerMonth	NULL	NULL	NULL	NULL	NULL	NULL
7	7	ResearchCollaboration	NULL	NULL	NULL	NULL	NULL	NULL
8	8	GrantMoney	NULL	NULL	NULL	NULL	NULL	NULL
9	9	JointDoctoralCitations	NULL	NULL	NULL	NULL	NULL	NULL
10	10	HopCount	HC	int	NULL	NULL	NULL	NULL
11	11	RoundTripTime	RTT	float	NULL	ms	measure using multiple tools	NULL
12	12	OneWayDelay	OWD	float	NULL	NULL	NULL	NULL

Fig. 12 OrgToOrgAttributeID table.

4). Visualization Results

The final visualization web application allows the user to see many things. Here are some example queries that the user can generate:

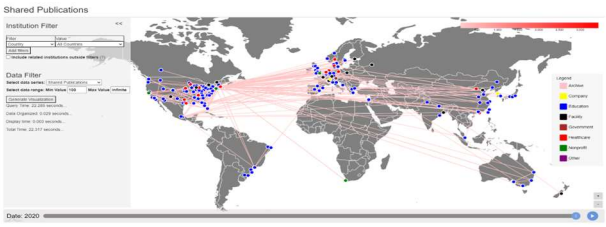

Fig. 13a Institutions with over 100 shared publications between each other in the year 2020

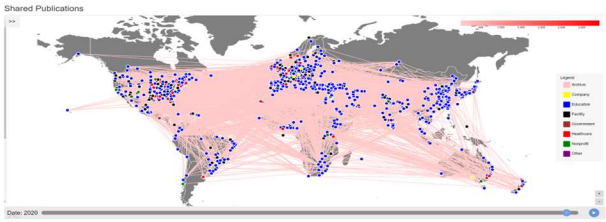

Fig. 13b Institutions with over 10 shared publications between each other in the year 2020

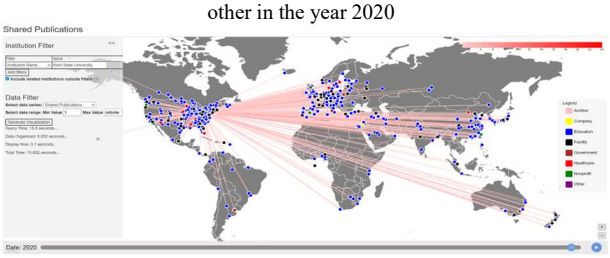

Fig. 13c All Kent State University shared publications in the year 2020

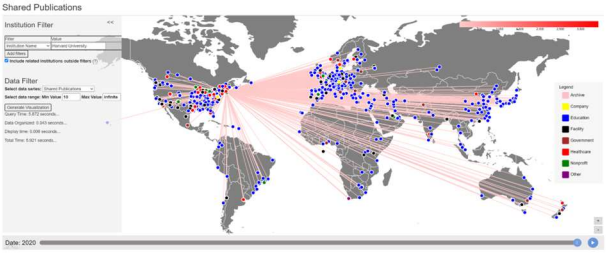

Fig. 13d Harvard University shared publications in the year 2020

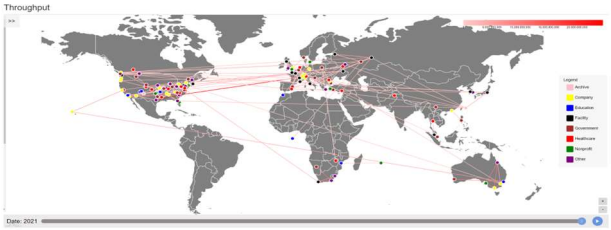

Fig. 13e All 2020 throughput data

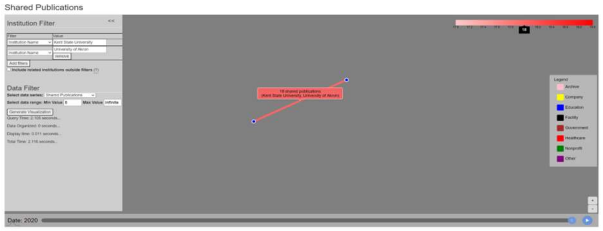

Fig. 13f Shared publications between Kent State University and the University of Akron. Queries can be quite specific.

4). Potential Extensions

There are many improvements that can be made to this application in the future. The first potential improvement is performance. Figure 14 shows performance data gathered from the application under various user queries. For this performance data gathering, certain queries were generated then tested multiple times. The average time in seconds the queries took to generate were collected and are shown in the table. As shown, certain queries, especially related to shared publication data, can take 20 to 40 seconds to generate. This is because shared publication data is the largest dataset in the database. These SQL queries can be further optimized and potentially significantly improve generation times. In addition to optimizing SQL queries, performance can also potentially be optimized on the timeline feature. With larger datasets, the timeline can become slow and often freeze. These performance issues could potentially be alleviated by a better date parsing algorithm. However, the amount of data that needs to be drawn on screen by amCharts would still be a problem.

Filters	Minimum data range value	data series	Include related institutions outside filters	Average Time
Country: All Countries	100	Shared Publications	FALSE	17.9335
Country: All Countries	50	Shared Publications	FALSE	25.2825
Institution Name: Kent State University	1	Shared Publications	TRUE	20.4436
Country: All Countries	1	Throughput	FALSE	1.37275
Country: All Countries	1	Ping Delay	FALSE	1.3895
Country: All Countries	1	Hop Count	FALSE	1.37075
Country: All Countries	1	Round Trip Time	FALSE	4.7695
Country: All Countries	1	One Way Delay	FALSE	1.3525
State(United States): Ohio	1	Shared Publications	FALSE	3.26575
State(United States): Ohio	1	Shared Publications	TRUE	39.1465

Fig. 14 Performance data.

Another future extension could potentially be more filters and query options. For instance, cumulative data. As of now, the visualization displays data separated by years. In the future, there could be an option to display the data cumulatively. This means for instance that users will be able to see the total amount of shared publication data in any given year as opposed to only being able to see the amount of shared publications from that year alone. Additional filter options would also be ideal to allow the user to further specify the data they want to see. However, SQL query performance would need to be

improved first as additional filter parameters might further slow down query results.

Another future extension would be more charts to display the data. AmCharts has many data charting options beyond geographical connection maps. Bar charts, line charts, pie charts, and other kinds of charts can be used to visualize the data in ways that the geographical line map cannot.

CONCLUSION

The project implemented a prototype data lake that is connected to two large living data infrastructures. One is the Microsoft Academic Graph (MAG), which connects 200+ million publication records. It is also growing each year and is cataloguing about 250 million scientific papers [KH1, AB1, AB10]. It has over eight billion triples with information about scientific publications and related entities, such as authors, institutions, journals, and fields of study. The data set is based on the Microsoft Academic Graph and licensed under the Open Data Attributions license. It also finds the institutions of the Authors and attempts to network characteristics between the institutions using another mega network instrumentation currently in place perfSONAR.

The project is outcome from the NSF Office of Cyberinfrastructure Award# 1925678.

References

[AH1] Hanemann, A., Boote, J. W., Boyd, E. L., Durand, J., Kudarimoti, L.,Lapacz, R., Swany, D. M., Zurawski, J., Trocha, S., "PerfSONAR: AService Oriented Architecture for Multi–Domain Network Monitoring",In "Proceedings of the Third International Conference on ServiceOriented Computing", Springer Verlag, LNCS 3826, pp. 241–254, ACMSigsoft and Sigweb, Amsterdam, The Netherlands, December, 2005.

[AH2] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.

[AH3] A. H. Mirza and S. Cosan, "Computer network intrusion detection using sequential LSTM Neural Networks autoencoders," 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404689.

[AH4] Emilio Corchado, Álvaro Herrero, Neural visualization of network traffic data for intrusion detection, Applied Soft Computing, Volume 11, Issue 2, pp. 2042-2056, 2011.

[AH5] Iperf, https://iperf.fr/, accessed on February, 2022.

[AH6] Traceroute, https://www.hjp.at/doc/rfc/rfc1393.html, accessed on February, 2022

[AH7] L. De Vito, S. Rapuano and L. Tomaciello, "One-Way Delay Measurement: State of the Art," in IEEE Transactions on Instrumentation and Measurement, vol. 57, no. 12, pp. 2742-2750, Dec. 2008, doi: 10.1109/TIM.2008.926052.

[AH8]OWAMP, https://software.internet2.edu/owamp/owping.man.html , accessed in February 2022.

[AH9] Network Pinger. http://www.networkpinger.com/en/, accessed on March 2022.

[AH10] W. Matthews and L. Cottrell, "The PingER project: active Internet performance monitoring for the HENP community," in IEEE Communications Magazine, vol. 38, no. 5, pp. 130-136, May 2000, doi: 10.1109/35.841837.

[AH11] PerfSONAR, https://docs.perfsonar.net/intro_about.html , Accessed on March 2022.

[AH12] Yang, C., Ku, L., & Chen, J. (2021). A perfSONAR-Based Network Performance Weathermap System. International Journal of Grid and High Performance Computing (IJGHPC), 13(3), 43-55. http://doi.org/10.4018/IJGHPC.2021070104.

[*AB*1] Zhao, W., Luo, J., Fan, T., Ren, Y., & Xia, Y. (2021). Analyzing and visualizing scientific research collaboration network with core node evaluation and community detection based on network embedding. Pattern Recognition Letters, 144, 54-60.

[AB2] Chuang, Y. T., & Chen, Y. H. (2021). Social network analysis and data visualization of MIS international collaboration in Taiwan. Library Hi Tech.

[*AB*3] Hu, J., & Zhang, Y. (2017). Discovering the interdisciplinary nature of Big Data research through social network analysis and visualization. Scientometrics, 112(1), 91-109.

[*AB*4] Hu, J., Huang, R., & Wang, Y. (2018). Geographical visualization of research collaborations of library science in China. The Electronic Library.

[*AB*5] Yu, F., & Hayes, B. E. (2018). Applying data analytics and visualization to assessing the research impact of the Cancer Cell Biology (CCB) Program at the University of North Carolina at Chapel Hill. Journal of eScience Librarianship, 7(1), 4.

[*AB*6] Huang, T. H., & Huang, M. L. (2006, July). Analysis and visualization of co-authorship networks for

understanding academic collaboration and knowledge domain of individual researchers. In International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06) (pp. 18-23). IEEE.

[*AB7*] Mannocci, A., Osborne, F., & Motta, E. (2019). Geographical trends in academic conferences: An analysis of authors' affiliations. *Data Science*, *2*(1-2), 181-203.

[*AB8*] Paszcza, B. (2016). Comparison of Microsoft academic (graph) with web of science, scopus and google scholar (Doctoral dissertation, University of Southampton).

[*AB9*] A. Salatino, A. Mannocci, and F. Osborne, "Detection, Analysis, and Prediction of Research Topics with Scientific Knowledge Graphs," *arXiv preprint arXiv:2106.12875,* 2021.].

[*AB10*] T. Ropinski, "Combining Interactive Exploration and Search for Navigating Academic Citation Data," Ulm University, 2018.

[*AB11*] A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E. D. López-Cózar, "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations," *Scientometrics,* vol. 126, no. 1, pp. 871-906, 2021.].

[AB12] Lukman, L., Rianto, Y., Al Hakim, S., Nadhiroh, I., & Hidayat, D. (2018). Citation performance of Indonesian scholarly journals indexed in Scopus from Scopus and Google Scholar. *Science Editing*, *5*(1), 53-58.

[AB13] Sawangkul, S., Pinitpuwadol, W., Sakiyalak, D., & Choopong, P. (2020). Bibliometric Differences between Scopus and Google Scholar for Ophthalmology Academics in Thailand: A Comparative Study. *The THAI Journal of OPHTHALMOLOGY*, *34*(1), 30-38.

[AB14] Agrawal, R. (2021). Plant Based Natural Fibers: A Bibliometric Analysis. *RESEARCH JOURNEY*, 9.

[A15] Monzani, A., Tagliaferri, F., Bellone, S., Genoni, G., & Rabbone, I. (2021). A Global Overview of COVID-19 Research in the Pediatric Field: Bibliometric Review. *JMIR Pediatrics and Parenting*, *4*(3), e24791.

[AB16] Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113-5142.

[AB17]Wraith, J., Norman, P., & Pickering, C. (2020). Orchid conservation and research: An analysis of gaps and priorities for globally Red Listed species. *Ambio*, 1-11.

[KH1] Arnab Sinha , Zhihong Shen , Yang Song , Hao Ma , Darrin Eide , Bo-June Paul Hsu , Kuansan Wang An Overview of Microsoft Academic Service (MAS) and Applications, International World Wide Web Conferences | May 2015, Published by Microsoft