Analysis of Academic Diaspora Networks

Robert P. Trevino, Parikshit Deshpande, Cameron Tew
CS224W Project Final Report

Abstract

The migration patterns of research scientists often creates an academic diaspora with potentially significant, long-lasting societal impacts. By studying the underlying structural network of a large volume of researcher movements (as cataloged by their research publications), we can better understand the extent of academic diaspora over time and more accurately predict the likelihood of academic movements in future time periods.

Introduction

Scientists and scholars oftentimes transfer to different institutions that better accommodate their research agenda. Opportunities and incentives can frequently draw them away from their country of birth, and when this happens disproportionately relative to the amount of researchers coming *into* a country, a scenario of "academic diaspora" can arise. The common underlying assumption is that a small group of notable researchers is critical to the strength of a nation's establishment within a particular field of science (Laudel 2003; Capuano and Marfouk 2013; Dohlman et al. 2019). Countries often put in place policies in an attempt to retain this talent and, as such, the study of these patterns is of particular interest in evaluating the success of these policies.

Hunter et al. (Hunter, Oswald, and Charlton 2009) detail the extent of the academic "brain drain" on nations due to emigration of academic intellectuals. Their theory-based model demonstrated that scientists' "productivity" increased when emigrating from their country of origin (based on the lowering of associated "costs" to immigration and travel). The empirical analyses of ˜160 physicists extrapolated macro-level trends about international academic "brain-drain" and the potential effect(s) of migration on a researcher's productivity. Correlations uncovered, like the tendency for physicists to immigrate to high R&D-per-capita countries like the US and Switzerland, may have causal-links explained by the fact that the data is entirely focused on physicists (e.g. initiatives like the Large Hadron Collider on the Switzerland/France border might be a leading reason why many of these researchers emigrated; a discussion that the paper doesn't broach).

Laudel has also previously examined the academic "brain drain" of "elite" scientists (Laudel 2005). The study focuses on two scientific specialties. It also leans heavily on noteworthy publications (e.g. Science and Nature) to capture the "scientific elite". Similar to the work by Hunter et al. (Hunter, Oswald, and Charlton 2009), Laudel's theoretical framework assessed whether or not there is a "brain drain" in a given field, by characterizing elite scientists as possessing four basic features (Mulkay 1976):

1. Privileged with respect to awards and facilities, and highly cited
2. Social ties with each other are stronger than their ties with other scientists
3. Control or direct the activities of the others
4. They considerably influence recruitment

Interestingly enough, Laudel acknowledges the importance of these features while defining "elite" based on the first criteria outlined above. The study shed light on the importance on faceting by field of study, and on clearly delineating between normal movement/collaboration of the scientific elite, and on permanent moving/immigration.

More recently, it has been demonstrated that scientists across different fields tend to consistently migrate (John Bohannon 2017). Using the ORCID dataset (Bohannon and Doran 2018), the movements of scientists post-Phd were tracked with over 3 million CVs of research scientist as of 2017, providing a additional information on scientist movement. Bohannon cites the United Nations semiannual reports on global science state that as of 2015, the global head count comes to 8 million scientists with 20% residing in an EU country, 17% and 19% are in the United States and China, respectively. In addition, he underscores that it is much easier to keep track of students while they are matriculating at a University to receive a Ph.D. However, subsequent to obtaining a Ph.D. it becomes much more difficult to track large scale movement patterns of scientist across different fields.

This paper investigates utilizing a machine learning approach on the large-scale graph structure defined by millions of scientific publications to help identify and better understand trends and patterns of academic diaspora. Temporal analysis of the graph structure was accomplished for 1-year "snapshots" in time spanning from 1999 to 2015 across all

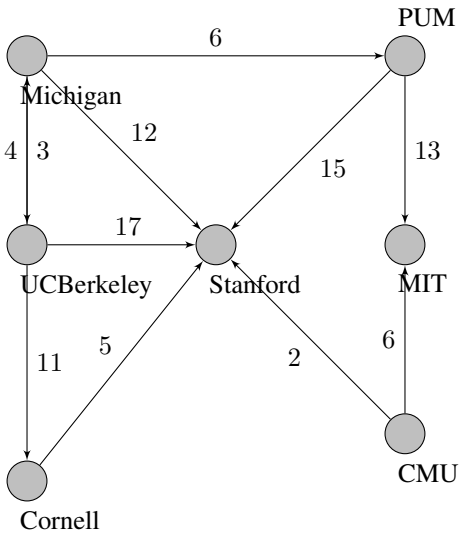


Figure 1: Example institution-to-institution graph

fields fields of study. In addition, unsupervised clustering techniques were leveraged to help form communities with similar graph structure that exposed underlying characteristics of the diaspora.

Problem Statement

Let $D = \{d_1, d_2, ..d_N\}$ be a corpus of N documents consisting of tracked scientific publications such that $N >> K$ for some large number K. Further, let each $d_i = \{o_i, n_i, a_i, y_i, t_i\}$ publication, where t_i is the text in the publication and $\{o_i, n_i, a_i, y_i\}$ is the meta data consisting of author organization, author name, author id, and year of publication respectively. The initial objective is defined as

$$\begin{aligned} D &\rightarrow G(V, E) \\ \text{S.T. } v_i &= o_i; e_{ij} = (a_i, o_i) \rightarrow (a_i, o_j); \\ V &= \cap\,(o_i), \forall (o_i ... o_N) \end{aligned}$$

where v_i is mapped to a single organization and e_{ij} represents migration. Thus, the corpus of D publications is transformed into $G(V, E)_t$, a directed graph at time t to $t + \delta$ using meta-data, where nodes $u, v \in V$ represent academic institutions and a weighted edge $e \in E$ represents the movement of research scientists from institution u to institution v from $t \rightarrow t + \delta$ as shown in Figure 1. The transformation from a large corpus of publications to a directed temporal graph structure is non-trivial compounded by the amount of noise inherent within the the corpus, i.e. duplication's, name homonyms, missing data, etc.

Furthermore, let $S \subseteq V$ be a subset of nodes that reflect some community structure with $G(V, E)$. The underlying structure is of interest, as it provides insight into the academic diaspora. Therefore, identifying these structures are critical. One way to measure the importance of the structure is through the use of conductance, which is subsequently defined. Let $cut(s)$ be defined where an edge connects two nodes in which one is a member of S and the other is a member of the complement $\bar{S} = S/V$. The conductance of a subset S is defined as:

$$\phi(S) = \frac{cut(S)}{min(vol(S), vol(\bar{S}))}$$

where $\sum_{u \in S} d_u$ represents the sum of degrees d_u of each node u in subset S (Yin et al. 2017). This type of of analysis can provide information on academic diaspora charateristics previously umknown. There are three primary objectives of this project:

- Transform a large corpus of published documents into a viable graph structure capable of informing on academic diaspora phenomenon.
- Investigate whether identifying a subset with low conductance yield cliques or communities that provide insight as to why the academic diaspora phenomenon originates in certain regions and countries and to what extent collaborative social influence may contribute to the phenomenon.
- Identify any significant graph structures such as high PageRank nodes that are in different communities that may explain the academic diaspora within and between different countries.

Background

The existing research landscape was surveyed in order to better understand several key topics that are relevant to academic diaspora network analysis:

- Existing techniques and datasets used in the modeling of academic collaboration networks.
- Measuring the "importance" or "significance" of a community within a network.
- Theoretical models around migration patterns in scientific fields of study.

Tang et al. (Tang et al. 2008) aimed at extracting and mining academic social networks using Open academic graph (OAG) data focusing on extracting researcher profiles, integrating the publication data, modeling the entire academic network, and providing search services over the data. The authors of the paper propose three generative probabilistic models for simultaneously modelling topical aspects of the papers, authors, and publication venues. They start by extending Friend-Of-A-Friend ontology as the profile schema and propose a unified approach based on conditional random fields. Using a unified probabilistic framework they deal with name ambiguity problem when integrating extracted researcher profiles and use a fixed schema structure for profile extraction. Starting with the raw data, the authors made a good attempt at extracting the profile of researchers. However, the authors do not account for the network structure associated with this data. Capitalizing on this approach on the OAG dataset, we identify and study patterns of academic diaspora using machine learning algorithms on the underlying graph structure.

Massucci and Doca (Massucci and Docampo 2019) applied the PageRank algorithm, a well known graph-traversing algorithm for defining web page popularity, to measure the academic reputation and prestige of a university. A citation network using five different Web of Science Subject Categories was used to test the proposed method, yielding a quantitative value that is less prone to bias when measuring the reputation of universities. The authors demonstrated that the citation network structure is well defined for the PageRank algorithm, since it uses measurements such as degree distribution and centrality to assist in measuring the reputation of an institution. The authors compared the proposed method to a well known academic standard, Academic Ranking of World Universities - Global Ranking of Academic Subjects (ARWU-GRAS), to determine the soundness of the algorithm. Principle Component Analysis was also used to plot the data points of the ranks of both the proposed PageRank as well as the ARWU-GRAS to show the how similar the datapoints were when projecting into the top 2 eigenvector feature space using 6 defined features pertaining to citations and rankings.

Foutohi and Rabbat (Fotouhi and Rabbat 2012) use a small world network topology to model migration trends. Each node represents an individual and the edges indicates a social connection to a neighbor that they interact with. Migration is predicted by analyzing economic gains, social influence, and patriotic bias. The strengths of this paper are the use of neighbors' behavior to assist in the modeling prediction of migration. It uses a simple network topology, infused with both economic and social influence including patriotic bias in determining the selection process of an individual for migration or not. The economic incentives were rather general, using the expected wage per capital for the home and potential destination countries. In addition, the social influence bias is limited to the proportion of neighbors that decide to migrate or stay at each time interval. The influence of each neighbor is measured as equal to each other. The method assumes a small-world based network, which may not be an appropriate assumption.

The proposed method differs from the previous methods in that no assumption will be made about the graph structure. Analysis will focus on aggregate movement of scientists from one institution to the next using graph-based machine learning models to identify community patterns that are not visibly apparent.

Methods

Conducting an analysis on academic diaspora required transforming massive quantities of publications into a graph structure that accurately reflected the diaspora phenomenon. Given the noise inherent within the corpus used, it was essential to commit considerable resources to cleansing and curating the data in order to subsequently perform statistical analysis and machine learning in an attempt to better understand the phenomenon.

Publication Dataset Acquisition

Publication meta-data used in defining a graph structure $G(V, E)$ was obtained via the Open Academic Graph (OAG), a large relational graph unifying two billion-scale academic published papers:

1. Microsoft Academic Graph (MAG)
2. AMiner

The OAG v2 dataset takes a snapshot of MAG from November 2018 and AMiner from January 2019. MAG contains information from ˜250 million authors and AMiner contains information from ˜113 million authors:

Dataset	Papers	Authors
Open Academic Society	91,137,597	1,717,680
AMiner	172,209,563	113,171,945
MAG	208,915,369	253,144,301

Figure 3: OAG, Aminer and MAG cardinality

The OAG v2 dataset matches linking pairs from both the MAG and AMiner data sources and filters out the authors with less than five publications. The dataset in its entirety contains 6 tables of relevant meta-data:

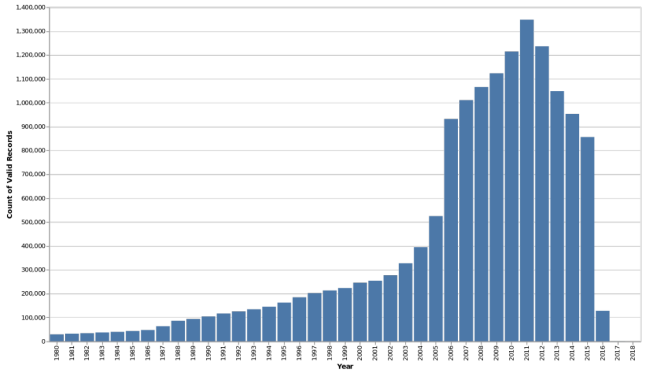
- Venue: information on locations and conferences where papers are published
- Authors: information on researchers/authors publishing papers
- Papers: the published paper along
- Papers Linking: a lookup table between AMiner and MAG papers
- Authors Linking: a lookup table between AMiner and MAG authors
- Venues Linking: a lookup table between AMiner and MAG venues

This provided distinct institutions from author and publication data using fields like "author affiliation". In cases where the data is missing, information from other fields was leveraged using various imputation methods. This provided a list of all the institutions in the dataset, that will be utilized to define the vertices $v \in V$ of the desired directed graph structure.

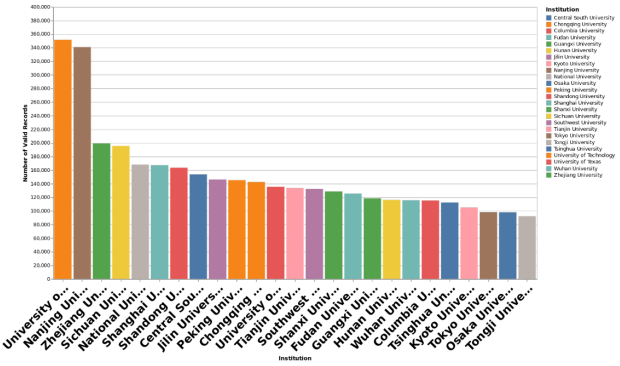
Migration patterns of authors were defined from one institution to another based on their previous publications. The aggregation of these movements over some time period, $t + \delta$, acts as a weighted edge in the resulting graph. For the cases where there is no movement seen (e.g. the paper is published at the university the researcher is currently "residing"), a self edge to the university was incremented. Given the inherent noise in the dataset, it was imperative to implement a thorough data curation process.

Data Curation

The OAG dataset follows a strict schema for papers, authors, and venues, well suited for the use of Google BigQuery - which proved necessary to handle the size of the OAG dataset and corresponding schema. This allowed for the use of the already-developed BigQuery-oag open-source



(a) University paper publication distribution by year



(b) University publication distribution

Figure 2: Paper publications distribution from raw data demonstrate inherent temporal bias peaking at around 2010.

project (Rada 2019) for uploading the OAG data into BigQuery.

From the raw OAG data, six tables of information were loaded to a BigQuery dataset containing the metat-data previously described as shown in 4:

Data Table	Rows	Attributes
authors	366,316,246	11
authors linking	1,717,680	2
papers	381,124,481	20
papers (linking)	91,137,597	2
venues	122,075	5
venues (linking)	29,841	2

Figure 4: OAG data overview

The scope of the project necessitates tracking the author's affiliation with an organization at time of paper publication. As such, the purview was limited to the authors, authors linking, and papers data , from the tables shown Figure 4. These tables act the building blocks of the eventual graph structure used to measure academic diaspora.

Data Extraction, Transformation, and Filtering At its core, OAG is a *union* of two datasets: Microsoft Academic Graph (MAG) and AMiner. Therefore, the first step is deduplicating $\sim 381M$ papers by removing those papers with both an AMiner ID **and** an MAG ID from the papers dataset. This reduces the dataset down by 22% to $\sim 297M$ papers.

To properly track author movement for edge $e \in E$ definition, each author is required be related to a paper and, subsequently, an organization and year of publication. Since author names may be similar, an already predefined author ID field was relied upon to distinguish between different authors with similar names. Curating this field would have required extensive resources prohibitive time and is, therefore, outside the scope of this paper. A "flattening" of the filtered paper data was performed to make author information easier to work with in future pipeline stages:

Pre-"Flattening" By Author

Row	ID	.....	authors.org	aurthors.name
1	100649178	.....	null	Tom J.
			Stanford	Susan H.

Row	ID	.....	authors.org	aurthors.name
1	100649178	.....	null	Tom J.
2	100649178	.....	Stanford	Susan H.

Post-"Flattening" By Author

Figure 5: "Flattening" process carried out on author name reduced table and query complexity.

This flattening procedure produces $\sim 758M$ paper-author associations once complete, requiring significant "trimming down" of data that's relevant and usable for subsequent analysis. Entries are further filtered if the following conditions are not met:

- Author ID is present
- Author name is present
- Author institution is present
- Venue ID is present
- Venue name is present
- Publication year is $>= 1785$ and $<= 2019$

This filtration process resulted in a reduction of data to $\sim 44M$ paper-author records. It was discovered that in this filtered dataset, $\sim 1.5M$ (out of $21M$; $\sim 7\%$) of the authors had an author ID that corresponded to multiple distinct author names (referred to as an "ambiguous author"). For example, Figure 6 demonstrates a single author ID with multiple associated names:

Row	ID	aurthors.name
1	5434e4cbdabfaebba58797a7	F. A. Abreu
2	5434e4cbdabfaebba58797a7	a j gordon
3	5434e4cbdabfaebba58797a7	a j tomarkea
4	5434e4cbdabfaebba58797a7	abigail b sivan
5	5434e4cbdabfaebba58797a7	alice dybsky
6	5434e4cbdabfaebba58797a7	.....

Figure 6: Example ambiguous authors information capable of distorting edge definition in graph structure.

To eschew all paper-author records with an ambiguous author would result in losing $\sim 22M$ ($\sim 50\%$) of the $\sim 44M$ filtered records from the previous stages of transformation, much of which is still viable data capable of highlighting trends of academic diaspora. Instead, a process was implemented to filter out all paper-author records associated with an ambiguous author *except* for the author id-author name tuple that corresponded to having the highest publication count for that given author id. This operation dramatically reduced our noise (each author ID now maps to a single author name), and only reduced our paper-author pairs by $\sim 4M$ entries ($\sim 10\%$).

Golden List Analysis The resultant $\sim 40M$ paper-author entries provided $\sim 13M$ potential author institutions. It was essential to curate these institutions by attempting to identify the unique academic organizations, companies, independent researchers, etc. that live among the remaining $\sim 40M$ paper-author records. As a first step, an extensive open-source mapping of universities was leveraged to their corresponding two-digit IATA country codes (Gutiérrez 2013) and their public university web pages. This mapping, or "golden list", plays a critical role of identifying and defining a set of all possible academic institutions to consider as well as also filtering homonyms, misspellings, etc. This curating process leveraged an ensemble of different string-matching techniques.

Direct Substring Matching An initial first-pass at basic substring whole-word matching between all $\sim 40M$ paper-author records and the golden list results in $\sim 14.3M$ ($\sim 36\%$) one-to-one matches (e.g. a record's organization matched with only **one** entry in the golden list).

Levenshtein distance Matching A subsequent, less stringent, filtering process was conducted using Levenshtein distance as a metric to allow for "fuzzy" matching between a paper-author record's institution and a corresponding entry in the golden list. Fundamentally, the Levenshtein distance measures the number of insert, edit, or delete operations necessary to make two arbitrary strings equal. More formally, the Levenshtein distance between two strings is computed as:

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) \text{ if } min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1 \end{cases} \end{cases}$$

A variation of this technique was leveraged, whereby each entry is first split into text tokens on white space and sorted. Levenshtein distance is then run for the string, and a threshold is set by the operator to determine a threshold value for an "acceptable" match. Empirically, manual analysis found that a threshold of $>= 0.9$ (out of 1.0) was sufficient to reduce the number of false-positives that this kind of "fuzzy" matching can introduce, while still adding value.

This sorted-token Levenshtein technique was run for $\sim 9M$ unique author organizations that were unable to be mapped with naïve substring matching, and ultimately resulted in an additional $\sim 17K$ unique institutions that were mapped to entries in the golden list. This in turn translated to $\sim 1.1M$ additional paper-author records being mapped to golden list entries as shown in Figure 7. In total, this brought the number of records matched to $\sim 15.3M$ of the original $\sim 44M$ ($\sim 38.2\%$), used in academic diaspora analysis.

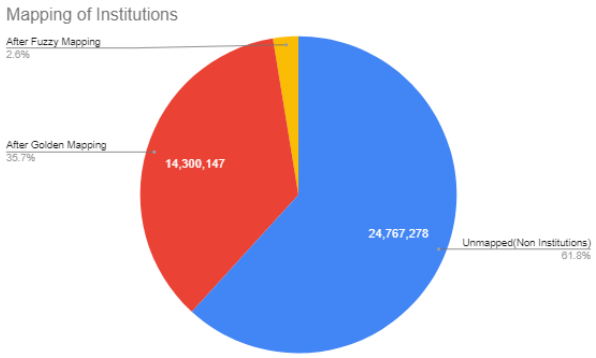


Figure 7: Distribution of data filtered after our mapping techniques

Phonetic matching based removal After generating the aggregate edge list for a given snapshot (e.g. Figure 1), additional pre-processing of the data found "near-duplicate" entries which were carried over from the golden list mapping that we used. Some of the institutions which represented the endpoints of an edge in actuality were the *same* institution but were mapping to satellite campuses, or using abbreviated/shortened names that were present in the original golden list.

In order to correct these errors, the aggregate edge list was examined for possible self loops using the ensemble of string-matching techniques previously mentioned in this paper, along with the Soundex phonetic matching algorithm. The end result of such an operation was a score depicting the belief that two edge endpoints were actually the same institution (even though they have non-equal strings). If the

score was above an empirically high threshold (> 0.95) these records were merged into a self-loop for the given institution and time period.

Figure 7 shows the effects of data size reduction using the various transformation and filtering techniques. Although significant, it was essential to ensure a higher quality analysis of academic diaspora.

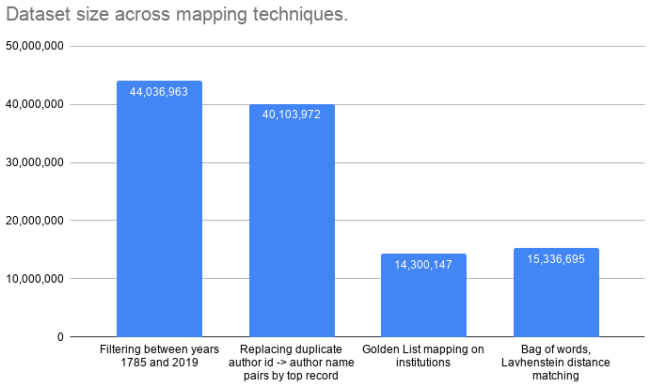


Figure 8: Data size after different mapping techniques

Approximately, 2 TB of data was acquired and processed through the data curation and cleansing tasks. The process required 2 days on a 24 core, 90 GB RAM, 200GB SSD machine to accomplish.

Graph Constructions

Subsequent to curating, a graph structure was created using Networkx (Hagberg, Schult, and Swart 2008) library in python, where the total number of nodes was upper bounded by the golden list enteries of $\sim 10,000$. Machine learning and statistical analysis was performed on the resultant graph structure.

The analysis of the academic diaspora graph took a top-down approach beginning with PageRank analysis for higher order graph graph properties and trends. Spectral clustering was subsequently performed for a more nuanced perspective on community structures within the graph. Finally, a hypergeometric test was applied granular analysis of individual institutions that contribute to the academic diaspora phenomenon.

PageRank Analysis on Graph Structure

PageRank (Page et al. 1998) is one of the most well known algorithms for graph analysis today. As previously noted, Massucci and Doca (Massucci and Docampo 2019) used PageRank to determine the prestige of an institution. Similarly, PageRank was implemented to determine whether any academic institution was considered more likely to be visited than others in a given year time span. The PageRank for any node in a graph defined as:

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + \frac{[1-\beta]}{N}$$

where d_i is the of neighbor i and β is the probability of following an outbound edge to another node. The algorithm was implemented and ran to determine which nodes are most probable given a random walk and provides useful information on an academic organization's contributions to the academic diaspora phenomenon.

Spectral Clustering Analysis

In addition, spectral clustering (Shi and Malik 2000) was implemented to find optimal "cuts" that identify interesting communities within the academic diaspora. The spectral clustering algorithm attempts to find the minimum number of normalized cuts that create groups of academic organizations; it is formally defined as

$$\min_{x \in R^n} \frac{x^T L x}{x^T D x}$$
$$S.T. x^T D e = 0, x^T D x = 2m$$

, where $2m$ is the sum of the degrees of a graph, e is vector of ones, x is the assignment vector and $L = D - A$ is the Laplacian matrix using degree matrix D and adjacency matrix A. It has been shown that the second smallest Eigenvector referred to as the Fiedler vector provides an optimal solution for the above optimization problem. The implementation relies upon a connected and undirected graph structure to perform properly. This requires transforming the academic diaspora graph into an undirected graph structure that is fully connected. Therefore, the academic diaspora graph was transformed into an undirected graph by transforming directed weighted edges into undirected edges and summing reciprocal edges between nodes. The largest weakly connected component of the undirected academic diaspora was then used in spectral clustering analysis.

Hypergeometric Distribution Analysis

In order to determine which academic institutions disproportionately contributed to the academic diaspora movement, the hypergeometric test was used on the outbound edges of the graph structure, which represents migrating researchers. The hypergeometric test uses the hypergeometric distribution to determine the statistical significance of observing k items of interest when taking a random sample; it is formally defined as:

$$P_{hg} = \frac{({}_K C_k)({}_{(N-K)} C_{(n-k)})}{({}_N C_n)}$$

, where N is the total population size to be selected from, n is the sample size selected, K is the total number of items of interest, and k is the number of success of selecting items of interest for the given sample size drawn. The hypergeometric test was used to determine which academic institutions provided out-going researchers in a statistically significant manner as compared to other organizations. This provides information on potential "brain drains" at the institution level.

Results

Using Colab to accomplish initial analysis of the raw data, several important factors were learned. First, as demonstrated in Figure 2a, there is an inherent temporal bias in the acquired data. The distribution patterns demonstrate that paper publications peak out around 2012. Although, the publications tend to range for the late 1700s to 2019, for the sake of contemporary analytics, analysis of the raw was limited beginning in 1999. The results further demonstrate that a significant number of those publications derive from organizations in east Asia as shown in Figure 2b. Although these biases may be as a result of the collection methods for published papers, determining collection veracity is outside the scope of this paper.

High Order Graph Structure Analysis

To better understand trends of academic diaspora, the heterogeneity of the underlying structural data was leveraged. The resultant graph structure was imported via the Newtorkx library (Hagberg, Schult, and Swart 2008) for subsequent analysis and visualization. A world map was generated and shown in Figure 9 to observe overall trends in academic diaspora. The results focused on 1-year snapshots over the time period ranging from 1999 to 2015.

Trends Captured Through PageRank Using the academic diaspora graph structure, analysis was conducted from 1999-2000 to 2014-2015 using PageRank to capture which organizations are more likely to be visited in the diaspora phenomenon.

The box plot in Figure 10 demonstrates a subtle decrease in PageRank for the first portion of years, which indicates the probability of randomly traversing a node is decreasing. Undoubtedly, this trend is influenced by the distribution patterns of the available published papers. In order to mitigate for this type of bias, the Wilcoxon Rank Sum Test was used to determine whether the change in PageRank of an academic organization from one year to the next was statistically significant and in which direction. The Figure 11 clearly shows that the change in PageRank cannot be completely explained by publication distribution bias alone. The academic organizations' rankings using PageRank tend to shift quite significantly over the course of 15 years. Assuming unbiased methods of collecting papers publislhed, if the shift in PageRank was due to a temporal bias of papers published as opposed to migration changes, then the proportion of papers published by an academic institution would increase in accordance with the increase in quantity of papers published and no notable changes would be observed in PageRank values from one year to the next. However, the Wilcoxon Rank Sum Test, shows that the academic diaspora have caused a statistically relevant shift of PageRank in different academic organizations from 1999 to 2015. Analysis indicates that two trends are evident: (1) Academic institutions that contribute to the academic diaspora phenomenon have significantly changed from 1999 to 2015. (2) The PageRank scores have decreased for all academic institution from 1999 to 2015.

It is notable that from 1999 until 2004, the universities with the highest PageRank scores originated from Japan (Osaka University, Tokyo University, Kyoto University) with Osaka topping the list with score of 0.011 in 1999. Starting around 2004 until 2015, the universities with the highest PageRanks were originating from China (Tsinghua University, China's University of Technology, Peking University), with Tsinghua University having the largest PageRank in 2005 of 0.006. This change is consistent with global changes as China has asserted its presence in technology.

Strength Of Community Using Spectral Clustering Communities were detected in the resultant academic diaspora graph using spectral clustering. Cluster size was varied from 2 to 30 to visualize how the average conductance across the different communities behaved. Figure 12 demonstrates that the conductance continues leveling out past 30 clusters.

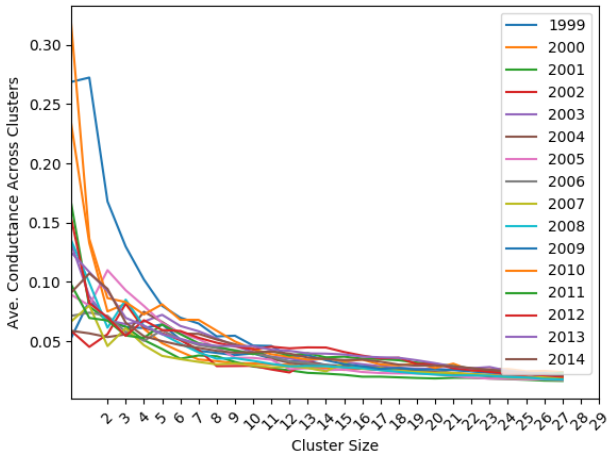


Figure 12: Conductance from Spectral Clustering using largest weakly connected component of the academic diaspora graph.

This behavior is most likely explainable by the number of node pairs that exist. Figure 13 demonstrates the weakly connected component with spectral cluster analysis ran using 20 clusters. The clustering demonstrates strong node interactions between many institutions primarily in Europe, North America, and East Asia. There is also an institution in Africa that tends to be well connected to the three previously mentioned regions. As will be discussed in subsequent analysis, this may be a key example of what is referred to as "brain drain". Regions that were disconnected after clustering were primarily located in South America and southwest Asia. This tends to indicate that the academic diaspora phenomenon is primarily withing the Europe-North America-East Asia region.

Academic Organization Statistically Significant Contributions The hypergeometric test was used on the outbound edges of the academic diaspora graph to determine

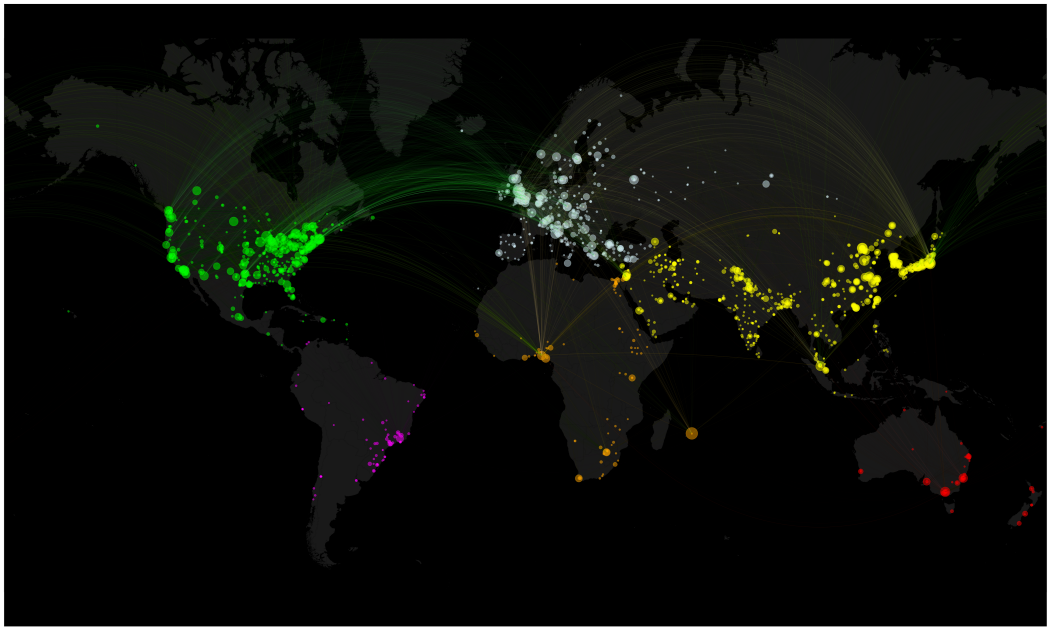


Figure 9: Institution ”importance” in the role of academic movement as calculated via PageRank (2014-2015)

which organizations contributed disproportionately to the academic diaspora phenomenon. A simple Bonferroni correction was applied to ensure statistical relevance across the different organizations tested using an alpha value of $\alpha = 0.05$ for p-value statistical significance. Table 1 provides a p-value score for each university given the number of outbound edges, as this is a direct indicator for the contribution of an organization to the academic diaspora phenomenon.

The impact of a small group of researchers as it pertains to tracking academic diaspora movement is evident for the top institutions. This surprising find demonstrated that researchers that are tracked based on unique author IDs associated with paper publications can significantly influence diaspora movement. This is evident in University of Aizu in 2014-2015 time frame. The number of out going edges is clearly distorted relative to the number of authors tracked in 2014. One explanation is that authors that depart an institution and subsequently migrate to multiple institutions the following year are going to have a disproportionate impact on graph structure versus those that move once. The next explanation, more probable, is that the unique author IDs may not be as unique as claimed and, therefore, require curation at the author ID meta-data level.

Nevertheless, the findings are still consistent with previous findings demonstrating that east Asia, namely Japan and China, are contributing disproportionately to the academic diaspora. There is an additional finding that suggests that a ”brain drain” is occurring within Africa as well. The University of Benin contributed to the academic diaspora for the last four out of five years. This would have been missed in the PageRank analysis, since there is no significant movement into that university.

Discussion

Several challenges existed in order to complete the project. The first challenge was the quality of the corpus data set. Pre-processing and curating of data required extensive effort to contend with missing data, author name homonyms, duplication, misspellings, etc. The discovery of the OAG data in Big query significantly reduced the time and effort needed in curating the data. In addition, defining node and edge values undoubtedly impacted the subsequent structure and, therefore, required considerable contemplation where re-definition was done in an initial iterative manner as limitations of the data were realized. The final graph structure of nodes as institutions and edges as an aggregate of researcher migration provided the most relevant diaspora information.

The sheer size of the data and resultant graph structure also posed a challenge requiring large-scale high performance cloud computing to carry out much of the analysis. The analysis included the mapping of organizations to nodes using a combination of exact substring match and Levinshtein distance on a pre-defined golden list. Although sig-

Table 1: TOP CONTRIBUTING ACADEMIC INSTITUTIONS TO ACADEMIC DIASPORA.

Year	Organization	P-value	Total Out Edges	Out Edges	Unique Author IDs
99-00	California Coast U.	3.340e-12	49	34	171
00-01	U. of Tulsa	1.902e-53	102	92	165
01-02	California Coast U.	1.4109e-19	57	44	253
02-03	California Coast U.	1.204e-31	85	66	280
03-04	U. of Wisconsin - Madison	6.159e-59	259	166	1309
04-05	Michigan State U.	1.311e-123	467	323	1240
05-06	Shanghai 2nd Medical U.	4.256e-167	669	541	792
06-07	Shanghai 2nd Medical U.	4.490e-180	620	585	877
07-08	Northwest A&F U.	5.668e-127	1288	989	807
08-09	Beijing U. of Aero & Astro	5.681e-88	1871	1255	1725
09-10	Beijing U. of Aero & Astro	3.267e-152	2138	1550	1749
10-11	Northwest A&F U.	2.846e-110	2545	1766	1999
11-12	U. of Benin	2.004e-202	1743	1532	464
12-13	U. of Benin	8.863e-319	2838	2540	563
13-14	U.of Benin	0.0	2879	2629	392
14-15	U. of Aizu	0.0	3329	3153	315

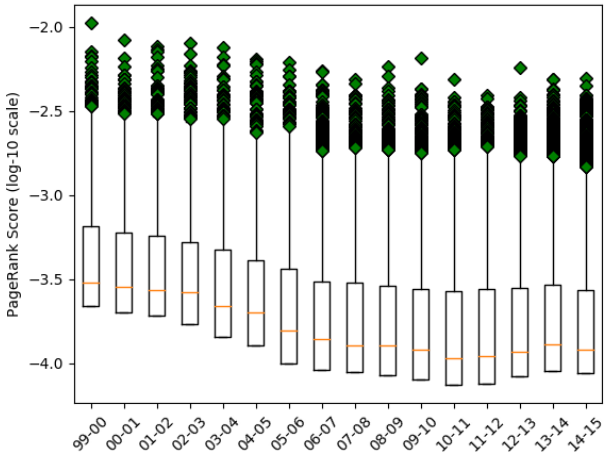


Figure 10: PageRank distribution analysis on academic organizations demonstrate a subtle but consistent decrease in value.

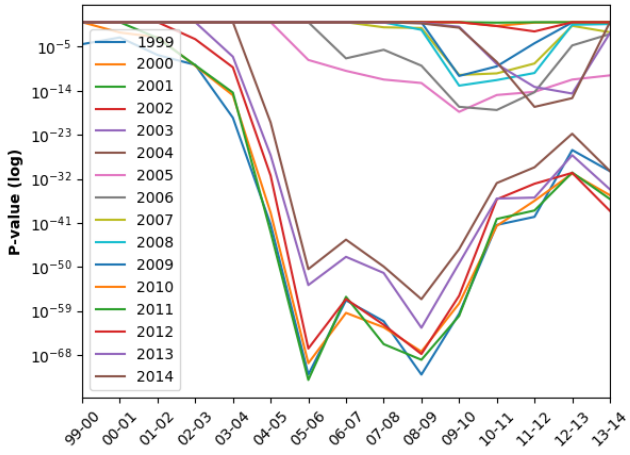


Figure 11: Wilcoxon Rank Sum Test comparing academic organizations' PageRank changes in consecutive years.

nificant amount of curating and cleaning on the data was accomplished, there were still certain biases that could not be removed. For instance, Figure 2 demonstrated a persistent temporal distribution bias pot-curating. This should be further investigated to determine why it exists.

The analysis of the resultant graph structure provided, interesting results that tend to enforce the notion that East Asia is significantly contributing to the academic diaspora phenomenon as both a source and sink node. PageRank analysis showed at around 2004-2005, China overtook Japan with respect to PageRank score in statistically significant manner. This was reinforced with spectral clustering for community detection where several institutions in East Asia were tightly connected to North America and Europe. In addition, the hypergeometric test on outbound edges demonstrated similar results with East Asian institutions. There was also a surprising find of University of Benin in Africa contributing significantly to the diaspora movement with outbound edges.

The resultant graph structure from published papers has provided an interesting glimpse into the academic diaspora phenomenon. However,caution must be taken as the data has been shown to be noisy and difficult to clean. Analysis of the nodes with the high degrees was revealing to the extent that it demonstrates many issues that must be resolved to move forward. Primarily, properly defining nodes is critical in order to have a viable graph structure. There are several instances of homonyms, misspellings, and concatenations including information such as department, country, etc.

Figure 13: Spectral clustering performed on the largest weakly connected component on the academic diaspora graph demonstrates strong connection between several institutions within Europe, North America, and East Asia.

that add a significant amount of noise, making it difficult to identify unique institutions. Although, this was mitigated by mapping of nodes onto a Golden List of Institutions, followed by Levenshtein's distance matching to remove some of the near duplicate entries further curating needs to be done to improve subsequent analysis. For instance, a small number of author IDs were able to significantly contribute to the hypergeometric test. It worth further investigating to determine whether this over-representation accurately reflects migration or whether there are shortcomings in the author ID assignment process.

Conclusion and Future Direction

Academic diaspora has several potential benefits as well as significant drawbacks. The sharing of scientific knowledge is critical in continuing to move forward with discoveries and inventions that revolutionize the way we live. However, academic diasporas have also been associated with "brain drains", a terminology used to describe the emigration flows of talented scientists from their country of origin. Therefore, better understanding the current trends provides a glimpse into the future of scientific movements. Analysis was focused on data curating and graph structure design with global community detection.

This paper provided an interesting, top-level perspective of the academic diaspora phenomenon. It also underscores the need for much more data curating and cleansing to create more accurate representations of movement patterns. Moreover, analysis that identify trends in movement such as PageRank and spectral clustering provide insight into the current state of academic diaspora. Indicators currently point to significant movement within and between east Asian countries. However, more research is needed to predict what the implication of these current trends mean for future movement patterns. Future work will focus on attempting to predict which organization will become more relevant in in contributing to future academic diaspora. Also since this analysis was focused on only the Academic Diaspora, a natural next step would be to include this analysis to non-academic research organizations, companies, and government organizations.

Contributions

All members contributed equally to the project.

References

[Bohannon and Doran 2018] Bohannon, J., and Doran, K. 2018. Data from: Introducing ORCID. *Dryad*.

[Capuano and Marfouk 2013] Capuano, S., and Marfouk, A. 2013. African brain drain and its impact on source countries: What do we know and what do we need to know? *Journal of Comparative Policy Analysis: Research and Practice* 15(4):297–314.

[Dohlman et al. 2019] Dohlman, L.; Dimeglio, M.; Hajj, J.; and Laudanski, K. 2019. Global brain drain: How can the maslow theory of motivation improve our understanding of physician migration? *International Journal of Environmental Research and Public Health* 16(7).

[Fotouhi and Rabbat 2012] Fotouhi, B., and Rabbat, M. G. 2012. Migration in a small world: A network approach to modeling immigration processes. *2012 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2012* 136–143.

[Gutiérrez 2013] Gutiérrez, E. 2013. world-universities-csv.

[Hagberg, Schult, and Swart 2008] Hagberg, A. A.; Schult, D. A.; and Swart, P. J. 2008. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)* 11–15.

[Hunter, Oswald, and Charlton 2009] Hunter, R. S.; Oswald, A. J.; and Charlton, B. G. 2009. The elite brain drain. *Economic Journal* 119(538).

[John Bohannon 2017] John Bohannon. 2017. Vast set of public CVs reveals the world's most migratory scientists. *Science*.

[Laudel 2003] Laudel, G. 2003. Studying the brain drain: Can bibliometric methods help? *SpringerLink* 57(2):215–237.

[Laudel 2005] Laudel, G. 2005. Migration currents among the scientific elite. *Minerva* 43(4):377–395.

[Massucci and Docampo 2019] Massucci, F. A., and Docampo, D. 2019. Measuring the academic reputation through citation networks via PageRank. *Journal of Informetrics* 13(1):185–201.

[Mulkay 1976] Mulkay, M. 1976. The Mediating Role of the Scientific Elite. *Social Studies of Science* 6:445–70.

[Page et al. 1998] Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1998. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, 161–172.

[Rada 2019] Rada, D. P. 2019. BigQuery-OAG.

[Shi and Malik 2000] Shi, J., and Malik, J. 2000. Normalized Cuts and Image Segmentation. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

[Tang et al. 2008] Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, 990–998. New York, NY, USA: ACM.

[Yin et al. 2017] Yin, H.; Benson, A. R.; Leskovec, J.; and Gleich, D. F. 2017. Local higher-order graph clustering. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Part F1296:555–564.