Collecting Diachronic Affiliation Data for Faculty at HBCUs Using Memento

Zarrillo, Deanna Drexel University, USA | dz364@drexel.edu
Kelly, Mat Drexel University, USA | mkelly@drexel.edu
Jackson, Christopher University of Tennessee - Knoxville, USA | cjacks75@vols.utk.edu
Yan, Erjia Drexel University, USA | ey86@drexel.edu

ABSTRACT
Academic mobility has accelerated in part due to recent civil rights movements and higher levels of social mobility. This trend increases the threat of brain drain from Historically Black Colleges and Universities (HBCUs), which already face significant logistical challenges despite broad success in the advancement of Black professionals. We aim to examine this threat from a Science of Science perspective by collecting diachronic data for a large-scale longitudinal analysis of HBCU faculty's academic mobility. Our study uses Memento, manual collection, and web scraping to aggregate historical identifiers (URI-Ms) of webpages from 35 HBCUs across multiple web archives. We are thus able to extend the use of "canonicalization" to associate past versions of webpages that resided at different URIs with their current URI allowing for a more accurate view of the pages over time. In this paper we define and execute a novel data collection method which is essential for our examination of HBCU human capital changes and supports a movement towards a more equitable academic workforce.

KEYWORDS
Memento; web archiving; Internet Archive; digital archive; academic mobility; HBCU

INTRODUCTION
Historically Black Colleges and Universities (HBCUs), established between 1837 and the passage of the Civil Rights Act in 1964, are the only institutions in the United States (U.S.) that were created for the purpose of educating Black citizens (Gasman, 2013). For most of the 19th century and the first half of the 20th century, HBCUs were the only institutions of higher education available to Black students and faculty. As of 2019, data from the National Center for Education Statistics reports 101 HBCUs in the U.S. and U.S. Virgin Islands with a combined enrollment of 298 thousand students (NCES, 2020). HBCUs are relatively small institutions and represent only a fraction (3%) of all higher education institutions in the U.S., but they have made substantial contributions in the preparation of Black professionals: they educated 80% of Black federal judges, 85% of Black doctors, 75% of Black Ph.D. graduates, 46% of all Black business professionals, and 50% of Black engineers (Seymore, 2005).

In the wake of *Brown v. Board of Education*, HBCUs found themselves competing with non-HBCUs to attract and retain outstanding faculty and students, which was often a losing battle (Seymore, 2005), and suffering a concomitant "brain drain." In HBCUs, brain drain has been a subject of constant concern because many Black professors and students transition from HBCUs to non-HBCUs, but few non-Black professors and students transition to Black colleges (Allen, 1991; Elmore & Blackburn, 1983; Miller, 1981; Mommsen, 1973; Morris, 1972). Brain drain in this context began as desegregation allowed Black students and faculty to enter non-HBCUs where academic quality was seen as higher (Morris, 1972); it is arguably exacerbated by elite universities in the U.S. prioritizing the hire of highly qualified Black faculty members to increase faculty diversity (Barrett & Smith, 2008; Mommsen, 1973). A 2010 U.S. Commission on Civil Rights report affirms that as resistance to Black attendance at non-HBCUs has faded and the need for segregated schooling has declined, it is reasonable to expect a brain drain from HBCUs (Rights, 2010).

While brain drain may pose a threat to any institution, it is especially damaging to HBCUs that face continual challenges like financial instability (Strauss, 2020; Walters, 2005), questionable governance structures (Minor, 2004), problems with student retention and progression (Brower & Ketterhagen, 2004; Lott & Davis, 2018), and declining enrollment (Strauss, 2020). HBCUs need to retain more capable faculty to address these issues, but the skilled employees that organizations need most to resolve crises are usually the first to leave (Levine, 1979). Meanwhile, studies of HBCU faculty mobility are scarce and, to date, have collected data using only interview or survey methods (Allen, 1991; Morris, 1972; Thompson, 1958). In the nearly 30 years following Allen's 1991 paper that explicitly discussed brain drain from HBCUs (Allen, 1991), there has been little quantitative reexamination of the issue. This may suggest that the matter is settled, but we argue that the reality is quite the opposite: social mobility and civil rights movements in the past few decades have accelerated academic mobility (Deville et al., 2014; Kato & Ando, 2017; Scellato et al., 2015; Siekierski et al., 2018; Sugimoto et al., 2016; Sugimoto et al., 2017; Van Noorden, 2012). Thus, the need to investigate academic mobility of faculty at HBCUs is increasing. Unlike previous research in this area, which has used small samples of data, this project will conduct extensive surveys and

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

take advantage of modern information technologies to collect a large, longitudinal data set of faculty appointments over time. In this pilot study, we use web archives to collect archival Internet data through a standards-based interoperability mechanism, Memento (Van de Sompel et al., 2013). Memento provides a common interface for access to one or more web archives and is supported by both institutional and national web archives inclusive of the Internet Archive Wayback Machine.

METHODOLOGY

Our two-stage data collection process is outlined below. First, we justify our sample selection of HBCUs and our reasoning for employing Memento as our primary data collection tool. We then discuss the first stage, which includes manual data collection and the generation of aggregated historical URIs via Memento. The second stage involves the development of a web crawler for the collection of supplemental data. Finally, we provide a high-level overview of the entire workflow. The goal of this process is to collect diachronic faculty affiliation data for mobile and non-mobile HBCU professors. It will provide us with the data necessary to examine the relationship between mobility and the productivity, impact, and career paths of professors at HBCUs. We will also study additional factors influencing this relationship such as gender, academic rank, and career length.

Selection of HBCUs

Of the current 101 HBCUs, 11 are doctoral-level institutions, 24 are master's-level institutions, and 66 are associate's-level or four-year institutions as designated by the Carnegie Classification of Institutions of Higher Education (CCIHE). We selected the 35 master's or doctoral degree granting HBCUs with potentially higher levels of research intensity. CCIHE publishes the number of full-time professors in ladder ranks (assistant, associate, and full professors) for the 11 doctoral-level institutions (master's-level institution data were not available). The average number of professors at HBCUs in 2020 is 320, with a maximum of 801 (Howard University) and a minimum of 156 (Clark Atlanta University). The total number of professors for the 35 HBCUs is estimated at ten thousand for a given year.

Using Memento to Collect Historical Faculty Affiliation Data

In a preliminary study, we collected professor mobility data from ORCID. We found that coverage of master's-level institutions was insufficient thus concluding that ORCID would be ineffective as the primary data source. A more reliable data source is the professor data published by institutional websites. Researchers have collected hiring and placement data from such websites (Clauset et al., 2015; Franzoni et al., 2018; Williamson & Cable, 2003; Zhu & Yan, 2017; Zhu et al., 2016). This project will significantly expand the scope of data collection by including those who have been affiliated with HBCUs from 2006 to 2020. Accessing web archives through their common Memento interface is essential in providing such historical data.

Memento is a standard mechanism that provides well-formed syntax and semantics for interfacing with web archives. All well-known web archives implement Memento in their archival replay system – the web-based interface where archival captures may be re-experienced. A replay system is critical to "re-assemble" the various resource representations (e.g., HTML, images, scripts) that make up a webpage. Because many web archives implement Memento, these common endpoints can be queried through an abstraction consisting of a single endpoint, known as an aggregator. An aggregator receives a single request, handles the asynchronous responses, and re-assemble the web archives' HTTP responses into a single listing of captures, known as a TimeMap in Memento parlance. We leverage the user-deployable open source MemGator (Alam & Nelson, 2016) software for aggregation in this project.

MemGator is initially configured with a set of well-known public web archive endpoints to query including Internet Archive (archive.org), Bibliotheca Alexandrina (bibalex.org), and the Archive-It service. This default set of web archives was sufficient for our initial data collection stage, as it queries international web archives for potentially valuable representation. In the following subsections, we detail the workflow we developed to associate HBCU faculty both past and present to their respective institution(s) and thus create a dataset for future analysis by leveraging web archives using Memento.

Locating Historical University Home URIs

We began the data collection process by manually gathering the current URI (Fielding et. al, 2005) of each HBCU homepage and its respective department-level pages into a spreadsheet. This data serves as input into a Python script that utilizes the MemGator aggregator to create and retain a TimeMap for each URI that passes through the program (Figure 1). The semantic and temporal context of these URI associations is represented in the TimeMap as relative relation attributes. For each HBCU homepage, we were able to collect a complete time range of historical URIs, which supported the later web scraping phase of data collection (Figure 2). Across all 35 homepages, we collected a total of 182,980 mementos with a mean of 5,228 mementos per original URI and a median of 5,322. In comparison, Drexel University and University of Tennessee at Knoxville's homepages each have over 10,000 mementos. However, while collecting and identifying the temporal extent of coverage, we found some department-level

TimeMaps did not contain a complete set of captures from our time range. This indicated to us that the URIs for these pages were different in the past and were consequently not captured. For example, Howard University's pharmacology department currently exists at "medicine.howard.edu/graduate-programs/pharmacology" whereas in 2008, this page was located at "med.howard.edu/pharmacology". Using the TimeMaps generated for each HBCU homepage, we began a process to programmatically collect these archival department URIs (URI-Ms) not represented in the original TimeMap.

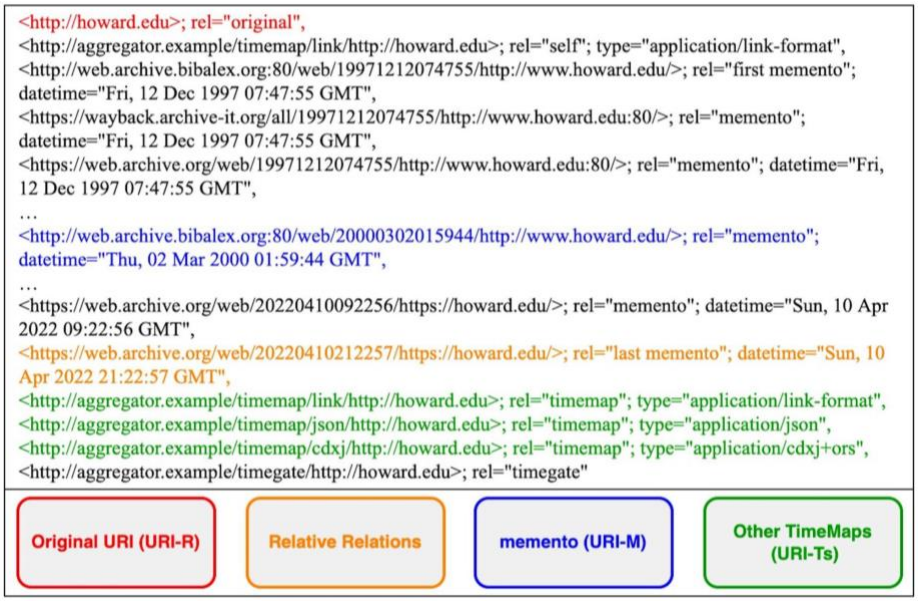

Figure 1. An abbreviated Memento TimeMap for the homepage of one HBCUs in our data set (Howard University at howard.edu) contains identifiers for historical captures (URI-Ms for mementos) as well as syntactic and semantic context per the Memento Framework (e.g., the relative relation "last memento").

Locating Historical Departmental URLs and Backfilling

Associating the past locations of a departmental homepage with its current live webpage and likewise historical captures of the past live page is an extended form of "canonicalization" (Kelly et al., 2017). Typically, canonicalization is meant to associate variations of a URI like "www.howard.edu" and simply "howard.edu". However, our approach would allow for the association of historical URI-Ms with their current live counterparts like the example of Howard University's pharmacology department previously mentioned. This association is critical for obtaining a more accurate picture of webpages over time. Our process to locate, collect, and canonicalize the missing historical department URIs has two stages. The first stage involved inspecting the evolution of sample HBCU websites by manually backfilling the missing URI-Ms in the TimeMap files. First, we identified institutions to manually canonicalize. Then, after identifying the oldest capture in each department's TimeMap, we utilized the Internet Archive to search for and collect the first instance of each archival URI-M in all missing years back to 2006 (Mabe et al., 2020; Kelly et al., 2014). This exploratory process helped us understand requirements for the second stage of backfilling, namely the programmatic implementation of a web scraper to avoid intense manual efforts. These requirements include the need for the crawler to only scrape data from within the university's network location, represented with the original URI in a memento's "Link" HTTP response header. We put this failsafe in-place to prevent the crawler from straying into out-of-scope URIs and for handling obsolete software and multiple encoding types in the URI-Ms. We also identified idiosyncrasies in the website structures such as the location of faculty lists that either existed as their own webpages, as lists on department or college level homepages, or within larger university level directories. When trying to identify the historical locations of the faculty details at scale by using the crawler, we will need to consider a broader scope in our search parameters.

Web scraping

We built a web crawler that utilizes a frontier rather than reusing an existing product because our primary purpose is to collect data that supplements and extends the TimeMaps, rather than as a primary source of data collection. A custom implementation allows us to directly manage the requirements of web archive domain specificity (Mohr et al., 2004) such as those found in the manual backfilling process. Our program will take the list of mementos that span from 2006 to 2020 in a TimeMap generated for an HBCU homepage (Figure 1), as a seed frontier. As the program visits each URI-M in the frontier "queue," it reads the content of the page then scrapes and collects all links that exist on that page. Once a page is crawled, it is added to a "crawled" list. Any URI collected that has not yet

been crawled will be appended to the frontier list and the program will iterate, subsequently reading from the queue. The program first checks if the crawled URIs are already in either the "queue" or "crawled" lists to prevent duplicative work before adding them to the frontier. We intend to address certain crawler traps to maximize the program's efficiency. Potential approaches include giving priority to certain URIs in the frontier and introducing parameters that limit how many pages deep the crawler can go from the homepage. We will then develop and implement a model on the final "crawled" list to classify faculty vs. non-faculty pages. Finally, we can then implement an additional program to scrape the content from these pages and collect the faculty data needed for our longitudinal study's data set.

WORKFLOW SUMMARY

Figure 2 depicts the high-level workflow of our data collection process. It shows manually collected homepage and department-level URIs as the input for the process, the overarching steps taken to create TimeMaps (Figure 1), the iterative process the crawler takes to supplement the TimeMaps with missing data, and the final output of extracted faculty and affiliation data into a compiled data set.

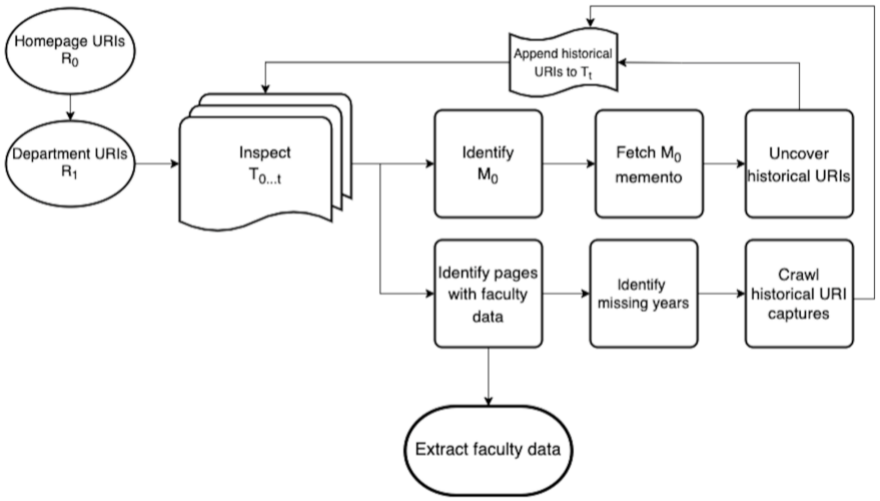

Figure 2: An illustration of the high-level workflow for the historical faculty affiliation data collection process

CONCLUSIONS AND FUTURE WORK

Increasing levels of academic mobility make it imperative to understand the impact of brain drain from HBCUs. We leverage contemporary, domain-specific approaches to collect large quantities of historic data for interpretation and analysis. We describe a novel canonicalization method for collecting webpages and their associated historical URI-Ms. Basing queries to web archives on the live web alone is insufficient for the task at hand as URIs are often not static. We leverage Memento to create TimeMaps by collecting historical URI-Ms across multiple web archives. We built a custom web crawler and scraper to capture URI-Ms that may not be present in the TimeMaps. Our novel contribution is the recursive procedure of using that which we observe in the archive for subsequent queries and discovery. Future data collection work includes completing the implementation of the specialized crawler and conducting test runs on a small portion of institutions to further refine the program. We will address questions such as: what patterns do we see in the evolution of URI-Ms, how can we ensure correct identification of departments across years, and how can we distinguish between faculty and non-faculty pages? Simultaneously, we will develop a program to identify faculty pages, scrape, and process faculty level data across all years. This process can potentially be scaled to fit more general or automated usage and can also serve as guidance for others wanting to analyze other types of web content over time.

As we proceed, results from this large-scale analysis will provide important evidence regarding factors which influence mobility of professors at HBCUs. This project will provide meaningful insights with data-backed evidence to support a diverse, inclusive, and equitable scientific workforce by analyzing diachronic institutional human capital changes within HBCUs. By doing so, we reveal potential brain drain and identify factors that attract and retain faculty and factors that make faculty leave HBCUs. Results of this project will illustrate key aspects of academic mobility at HBCUs along the dimensions of institutional and professor profiles. The triangulation of hand-collected archival data, data from the Web of Science, and survey and interview data will ensure the accuracy and value of the data.

REFERENCES

Alam, S. & Nelson, M. L. (2016). MemGator — A portable concurrent memento aggregator: Cross-platform CLI and server binaries in Go," In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, June 2016, 243-244.

Allen, H. L. (1991). The Mobility of Black Collegiate Faculty Revisited: Whatever Happened to the" Brain Drain"? *The Journal of Negro Education, 60*(1), 97-109.

Barrett, T. G., & Smith, T. (2008). Southern coup: Recruiting African American faculty members at an elite private Southern research university. *American Educational Research Journal, 45*(4), 946-973.

Berners-Lee, T., Fielding, R. T., & Masinter, L. (2005). Uniform Resource Identifier (URI): Generic Syntax, RFC 3986.

Brower, A. M., & Ketterhagen, A. (2004). Is there an inherent mismatch between how Black and White students expect to succeed in college and what their colleges expect from them? *Journal of Social Issues, 60*(1), 95-116.

Cheung, A. C. K., & Xu, L. (2015). To return or not to return: examining the return intentions of mainland Chinese students studying at elite universities in the United States. *Studies in Higher Education, 40*, 1605-1624. doi:10.1080/03075079.2014.899337

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science advances, 1*(1), e1400005.

Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V. D., & Barabási, A. L. (2014). Career on the move: Geography, stratification, and scientific impact. *Scientific reports, 4*, 1-7. doi:10.1038/srep04770

Dulam, T., & Franses, P. H. (2015). Emigration, wage differentials and brain drain: the case of Suriname. *Applied economics, 47*(23), 2339-2347.

Elmore, C. J., & Blackburn, R. T. (1983). Black and white faculty in white research universities. *The Journal of Higher Education, 54*(1), 1-15.

Foster, G. A. (2001). *Is There a Conspiracy to Keep Black Colleges Open?* : Kendall/Hunt Publishing Company Dubuque, IA.

Franzoni, C., Scellato, G., & Stephan, P. (2018). Context Factors and the Performance of Mobile Individuals in Research Teams. *Journal of Management Studies, 55*, 27-59. doi:10.1111/joms.12279

Gasman, M. (2013). The changing face of historically Black colleges and universities.

Hussain, S. M. (2015). Reversing the brain drain: Is it beneficial? *World Development, 67*, 310-322.

Kato, M., & Ando, A. (2017). National ties of international scientific collaboration and researcher mobility found in Nature and Science. *Scientometrics, 110*, 673-694. doi:10.1007/s11192-016-2183-z

Kelly, M., Alkwai, L. M., Alam, S., Nelson, M. L., Weigle, M. C., & Van de Sompel, H. (2017). Impact of URI Canonicalization on Memento Count, In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 303-304, arXiv:1703.03302.

Kelly, M., Nelson M. L., & Weigle M. C. (2014). Efficient Thumbnail Summarization for Web Archives, Digital Preservation 2014.

Le, T., & Bodman, P. M. (2011). Remittances or technological diffusion: which drives domestic gains from brain drain? *Applied economics, 43*(18), 2277-2285.

Levine, C. H. (1979). More on cutback management: Hard questions for hard times. *Public administration review, 39*(2), 179-183.

Lott, S., & Davis, B. L. (2018). Accelerated Baccalaureate Nursing Students' Perception of and the Variables Influencing Their Retention: An HBCU Perspective. *ABNF Journal, 29*(3), 76-85.

Mabe, A., Patel, D., Gunnam, M., Shankar, S., Kelly, M., Alam, S., Nelson, M. L., & Weigle, M.C. (2020). Visualizing webpage changes over time. *arXiv:2006.02487*.

Miller, C. L. (1981). Higher education for black Americans: Problems and issues. *The Journal of Negro Education, 50*(3), 208-223.

Minor, J. T. (2004). Introduction: Decision making in historically Black colleges and universities: Defining the governance context. *Journal of Negro Education*, 40-52.

Mohr, G., Stack, M., Ranitovic, I., Avery, D., & Kimpton, M. (2004). An Introduction to Heritrix, An open source archival quality web crawler. In *Proceedings of the 4th International Web Archiving Workshop (IWAW`04),* 109-115.

Mommsen, K. G. (1973). Professionalism and the racial context of career patterns among black American doctorates: A note on the" brain drain" hypothesis. *The Journal of Negro Education, 42*(2), 191-204.

Morris, E. W. (1972). The contemporary Negro college and the brain drain. *The Journal of Negro Education, 41*(4), 309-319.

NCES. (2020). College navigator. Retrieved from https://nces.ed.gov/COLLEGENAVIGATOR/?s=all&sp=4&pg=1

Rights, U. C. o. C. (2010). The educational effectiveness of historically Black colleges and universities: Author Washington, DC.

Rosenblatt, Z., & Sheaffer, Z. (2001). Brain drain in declining organizations: Toward a research agenda. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 22*(4), 409-424.

Scellato, G., Franzoni, C., & Stephan, P. (2015). Migrant scientists and international networks. *Research policy, 44*, 108-120. doi:10.1016/j.respol.2014.07.014

Seymore, S. B. (2005). I'm confused: How can the federal government promote diversity in higher education yet continue to strengthen historically Black colleges. *Wash. & Lee J. Civil Rts. & Soc. Just., 12*, 287.

Siekierski, P., Lima, M. C., & Borini, F. M. (2018). International Mobility of Academics: Brain Drain and Brain Gain. *European Management Review, 15*, 329-339. doi:10.1111/emre.12170

Solimano, A. (2008). *The international mobility of talent: Types, causes, and development impact*: Oxford University Press on Demand.

Strauss, V. (2020). Why the coronavirus crisis could hit historically black colleges and universities especially hard. Retrieved from https://www.washingtonpost.com/education/2020/04/07/why-coronavirus-crisis-could-hit-historically-black-colleges-universities-especially-hard/

Sugimoto, C. R., Robinson-Garcia, N., & Costas, R. (2016). Towards a global scientific brain: Indicators of researcher mobility using co-affiliation data. *arXiv:1609.06499*.

Sugimoto, C. R., Robinson-Garcia, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature, 550*, 29-31. doi:10.1038/550029a

Thompson, D. C. (1958). Career patterns of teachers in Negro colleges. *Social Forces*, 270-276.

Van de Sompel, H., Nelson, M., and R. Sanderson, (2013). HTTP Framework for Time-Based Access to Resource States -- Memento, RFC 7089, doi: 10.17487/RFC7089.

Van Noorden, R. (2012). Global mobility: Science on the move. *Nature News, 490*, 326.

Walters, A. (2005). Predominantly Black and historically Black colleges spar over federal funds. *The Chronicle of Higher Education, 52*(6), A28.

Ware, L. (1993). The Most Viable Vestige: Black Colleges after Fordice. *BCL Rev., 35*, 633.

Williamson, I. O., & Cable, D. M. (2003). Predicting early career research productivity: The case of management faculty. *Journal of Organizational Behavior, 24*, 25-44. doi:10.1002/job.178

Zhu, Y., & Yan, E. (2017). Examining academic ranking and inequality in library and information science through faculty hiring networks. *Journal of Informetrics, 11*(2), 641-654.

Zhu, Y., Yan, E., & Song, M. (2016). Understanding the evolving academic landscape of library and information science through faculty hiring data. *Scientometrics, 108*(3), 1461-1478.