SURVEYING USAGE OF ACADEMIC RESEARCH IN JOURNALISM

Logan Walls - Researcher | Isabelle Edwards - Researcher | Tin Ho - Project Manager

Information School UNIVERSITY of WASHINGTON

DataLab

PROBLEM STATEMENT

> How can we use **authorship** and **citations** to better understand **information diffusion** between **popular media** and **academic articles** for the purpose of **informing** the **general public** as well as the **academic community**?

PROCESS

RAW DATA COLLECTION

Obtained data from New York Times application programming interface (API)	Collect metadata about all of the articles which match our query.	Includes web URLs, headlines, keywords, publication dates, word counts, and more for each article
Use ontology query to retrieve web URLs of entities which have the “scholarly publication” attribute	Used regex to extract domain names domain names from URLs	Combine domain names to create a list of academic publication web domains
Parsed HTMLs recieved from API with BeautifulSoup for scholarly documents	Save links which appear to be citations (any URL in the list of academic publication domains)	URLs that contain ‘pubmed’, ‘.gov’, ‘.edu’, ‘doi’, ‘abstract’, or ‘pdf’ are also suspected to be citations
Collected digital object identifiers (DOI) from scholarly documents	If link leads to an HTML page, parsed the page for DOIs using BeautifulSoup and regex	If the link leads to a PDF, parse XML file generated by GROBID machine learning package for DOIs
Obtained metadata through DOI lookup service	For each DOI, send a request to http://dx.doi.org for a response in turtle format	Parse response using regex to retrieve DOI metadata (including article title, authors, publisher, etc.)
Analyzed data and created figures using GraphLab and Tableau	Requested Eigenfactor and open-access information from Dr. Jevin West for each DOI	Look at any anomalies or interesting trends in the data and figures

DATA PROCESSING

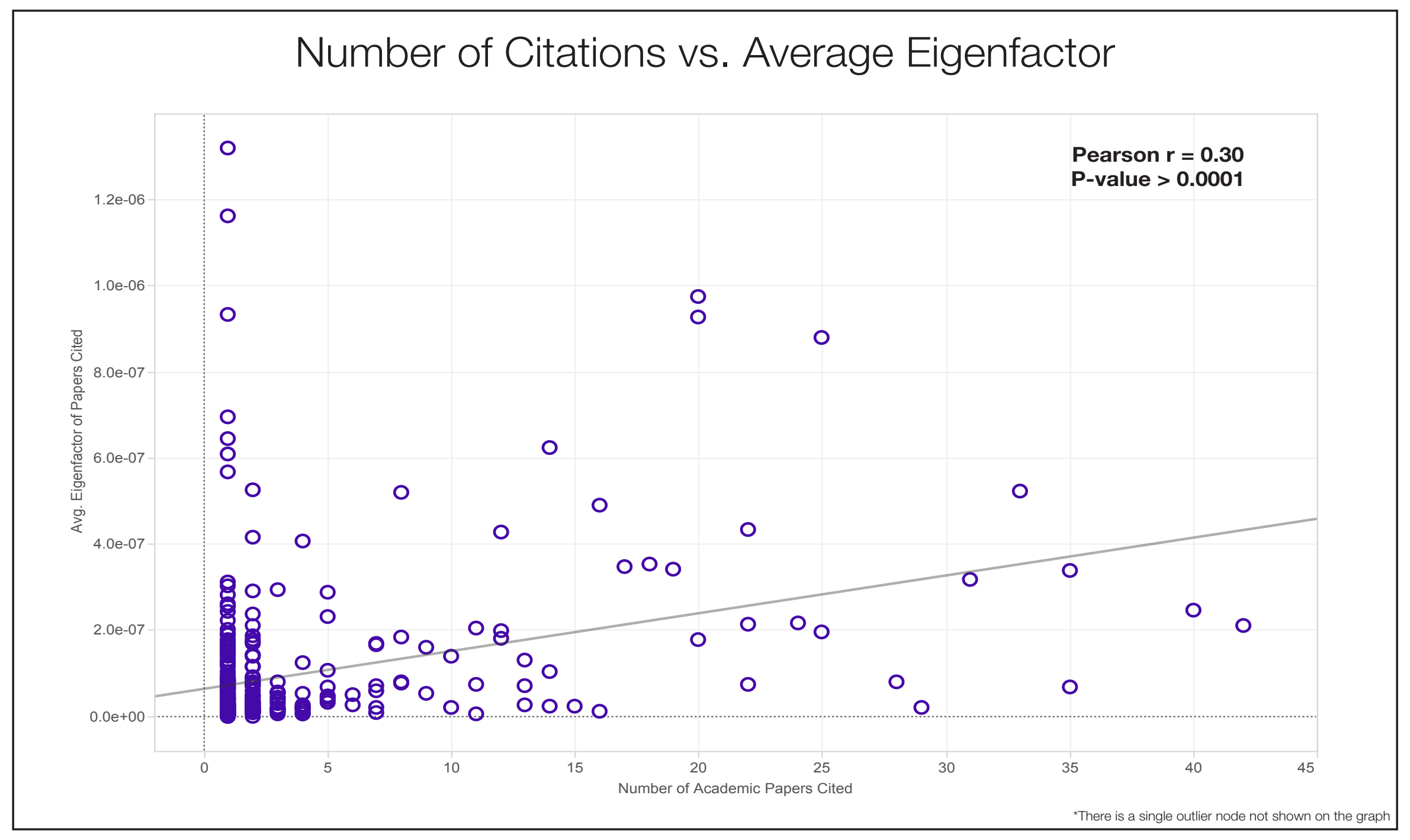
Combine the NYT and link data by concatenating all NYT metadata into one table using Graphlab	Sorted out nested data structures and extract relevant information	Join DOIs to each NYT article via the links from which the DOIs were retrieved
Merge all metadata retrieved from DOI lookup service into a single table	Reconcile inconsistent field-names and missing values	Join the metadata received from Dr. West into the table
Generate topic groupings for the NYT articles	Compute unigram set for each NYT article using body-text and filter by calculating Term-Frequency-Inverse-Document-Frequency score	Iteratively train topic models on the filtered unigrams using Graphlab, adjusting parameters as needed

RESULTS

Eigenfactor is a metric used to measure the influence of academic publications: by employing a similar algorithm to PageRank, the influence of each article is not merely determined by the number of citations it receives, but also by the influence of the papers which cite it. By plotting the number of academic citations in a New York Times article against the average eigenfactor of those citations we show a significant positive correlation (Pearson's r = 0.31, p < 0.0001).

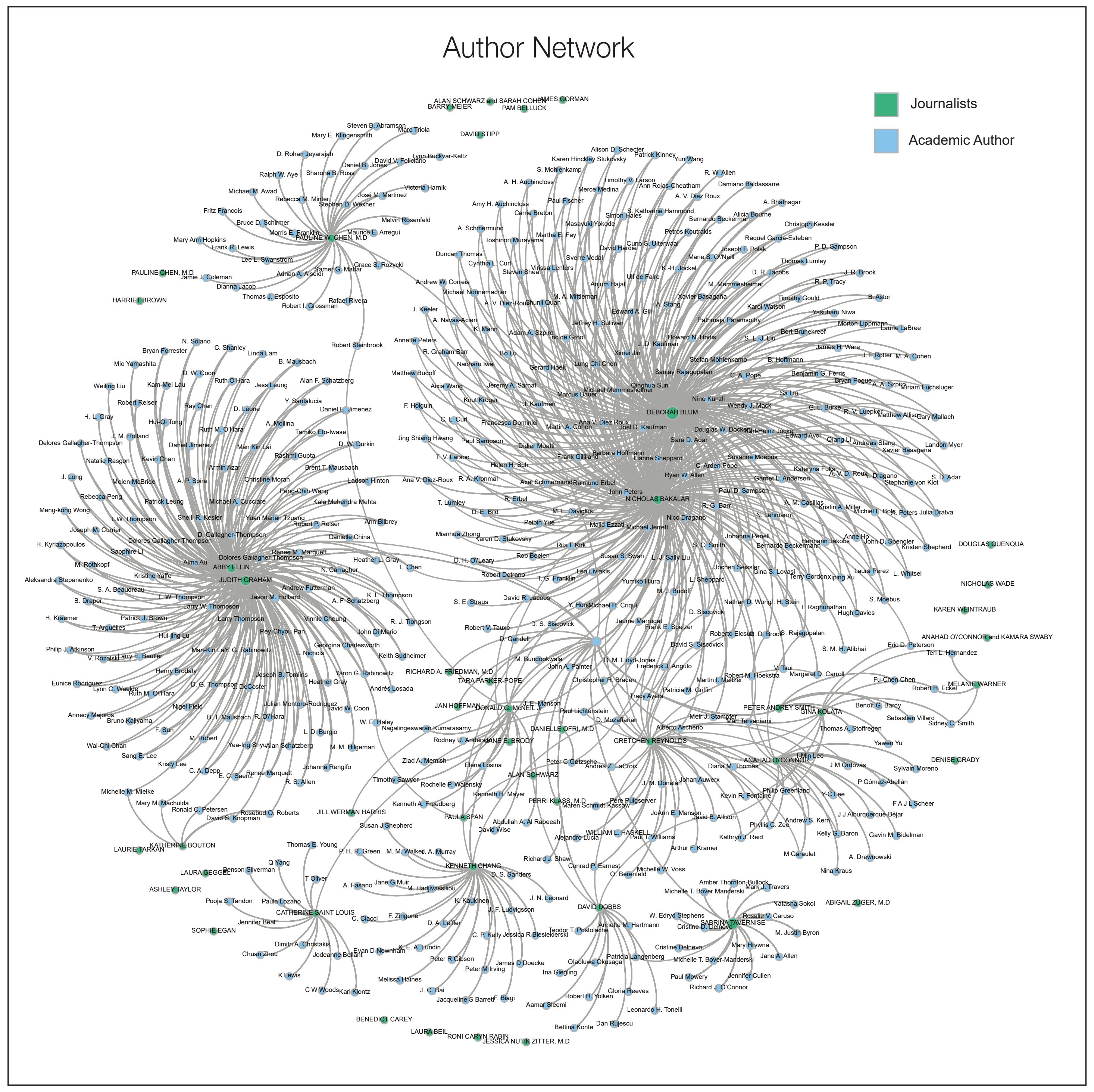
Initially we interpreted this as a difference in research styles between journalists, but upon examining the same variables aggregated by journalist the correlation was much weaker, suggesting that the pattern we observe between citation counts and eigenfactor is more content-driven than journalist-driven.

Number of Citations vs. Average Eigenfactor



*There is a single outlier node not shown on the graph

Author Network



The figure above is a network of all authors in our data set. The blue nodes are academic paper authors, and the green nodes are New York Times journalists. Each line coming from the journalist node is a citation to an academic article. This network only contains nodes that have received more than 5 citations (academic), and nodes that cite more than 5 articles (New York Times). The size of the nodes are determined by the amount of citations they have received or papers they have cited.

The journalists Deborah Blum and Nicholas Bakalar are shown to have cited many of the same academic articles (shown on the top left of the network). Similarly, the journalists Abby Ellin and Judith Graham are also shown to cite many of the same academic authors. Journalists in the center of the network have citations to authors all over the network, and do not seem to overlap too much with any of the other journalists. The journalists at the bottom of the network have a similar amount of citations as many of the other journalists, but are shown to have cited a fewer amount of academic authors. This could mean that they have cited a smaller sample authors on several occasions, or that they cited many different authors less than five times. Journalists on the edges (with no connections to academic authors) do cite more than five authors, but do not cite those authors more than five times.

Special thanks to Dr. Jevin West and Dr. Emma Spiro of the UW’s DataLab for guiding us along in our project