Web of Science™ as a Research Dataset

Katy Börner,[1] Valentin Pentchev,[2] Matthew Hutchinson,[2] James Pringle,[3] Jason Rollins,[3] Yadu N. Babuji,[4] & Eamon Duede[5]

[1]katy@indiana.edu CNS, SOIC & Network Science Institute, Indiana University, Bloomington, IN, US
[2]maahutch@iu.edu & vpentche@iu.edu IUNI, Indiana University, Bloomington, USA

[3]jason.rollins@clarivate.com & james.pringle@clarivate.com Clarivate Analytics, USA
[4]yadunand@uchicago.edu Computation Institute, Knowledge Lab, UofChicago, USA

[5]eduede@uchicago.edu Computation Institute, Knowledge Lab, Committee on the Conceptual and Historical Studies of Science, Uof Chicago, USA

Introduction

The Clarivate Analytics Web of Science (WoS) has served as a research dataset for more than 9,000 scholarly articles in the past 15 years alone—across a wide a range of fields and disciplines from toxicology to computer science to economics. Scientists and scholars have been particularly interested in the WoS citation network, a massive graph containing billions of links that can proxy the structure and dynamics of not only scholarly communication, but knowledge diffusion, the evolution of fields, and the career lifecycles of individuals and institutions. To power these investigations, scholars are increasingly employing a number of compute-intensive methodologies, sophisticated big data infrastructures, and so called collaborative "discovery science" tools and techniques. Suddenly, in addition to deep, domain specific expertise, world-class computational knowhow appears to be a new prerequisite for analysis of scholarly data at the scale represented by WoS. While cloud-based computing and tools are more prevalent and accessible than ever before, harnessing these technologies remains both a challenge and opportunity for researchers and data providers (i.e., Clarivate Analytics and similar commercial data vendors and non-commercial aggregators). While the opportunities made possible by scholarly data at the size and scope of WoS for discovery and innovation are limited only by imagination, two general prospects come readily to mind. First, access to these data coupled with the appropriate computational and analytical capabilities opens up a wide range of funding and subsequent publishing opportunities in high impact venues. Second, data providers can pursue new business opportunities, including novel data access models, new types of analytic products, and new kinds of academic/industry partnerships. In this poster paper, we briefly explore 1) the new computational infrastructures that are being developed to enable collaborative research that leverages scholarly datasets such as WoS that are both big and proprietary; 2) some recent findings that have been made possible by these infrastructures; and, 3) new commercial offerings that have been enabled and demanded in response to increasing reliance on the WoS as a research dataset.

New Computational Infrastructures

Research leveraging big, scholarly datasets like WoS presents researchers with challenges related to the data's size, inherently relational format, and sensitive (proprietary) nature. To overcome these challenges, researchers have developed a new generation of enclave supported, high performance, and cloud-based, collaborative research environments that are both elastic enough to provide substantial computational resources when needed while remaining secure enough to protect data providers' commercial interests. Moreover, these environments have evolved to allow geographically distributed researchers to collaborate on research projects in the fast paced, iterative style that has come to dominate research in the era of Big Data—namely, "discovery science".

IUNI WoS Data Enclave

The Indiana University Network Science Institute (IUNI) acquired the complete set of Clarivate Analytics' Web of Science XML raw data (Web of Knowledge version 5). The data was parsed and stored in a well-documented Postgresql database, see entity-relationship diagram, database schema, and data dictionary on http://iuni.iu.edu/resources/web-of-science. The code used to parse the WoS XML format and to save data in the Postgresql database was made available freely on GitHub, see **https://github.iu.edu/CNS/generic_parser**. All data can be accessed via the IUNI WoS Data Enclave, a secure repository that uses IU's Karst high-throughput computing cluster designed to deliver large amounts of processing capacity over long periods of time. Access to the XML data and the PostgreSQL database is granted to a user's Karst account. IU faculty, staff, and qualifying sponsored affiliates can request accounts on Karst to use the data for academic research and without any sharing of data. A simple web browser based query interface to the WoS dataset was implemented to support custom queries for specific terms, journals, or authors. Datasets can be downloaded in CSV data format compatible with data mining and visualization tools such as Gephi or the Sci2 Tool (http://sci2.cns.iu.edu) (Sci2 Team, 2009). More about the IUNI WoS Data Enclave can be found at **http://iuni.iu.edu/resources/web-of-science**.

Cloud Kotta

One platform specifically developed with WoS in mind is Knowledge Lab's Cloud Kotta (CK). CK is a secure data enclave and analytics platform that serves the research needs of social sciences (Babuji 2016). By hosting CK in the Amazon Web Services cloud, the developers were able to take advantage of virtually limitless compute, cost-effective storage and the ability to implement a fine-grained security model ensuring the authorized collaborators could access both data and compute resources from any where in the world (Babuji 2016). Moreover, CK supports multiuser, rapid ideation and research iteration through a novel Python library that enables specific functions in an analysis code, written in a Jupyter Notebook to be seamlessly and securely submitted to the CK execution fabric (Babuji 2017). By allowing researchers to develop and share analysis code interactively over secure data like WoS, CK has removed the need for deep computational infrastructure expertise. The complete WoS XML dataset was ingested into a relational database housed in CK using a custom parser that has been made freely available on GitHub (see: **https://github.com/alexander-belikov/wos_parser**). The Cloud Kotta WoS database schema can be found on CK's documentation pages (see: **http://docs.cloudkotta.org/dataguide/wos.html**). More about Cloud Kotta can be found at **http://docs.cloudkotta.org**.

IU International Co-Affiliation Network, 2004-2013

CNS @ Indiana University
2016

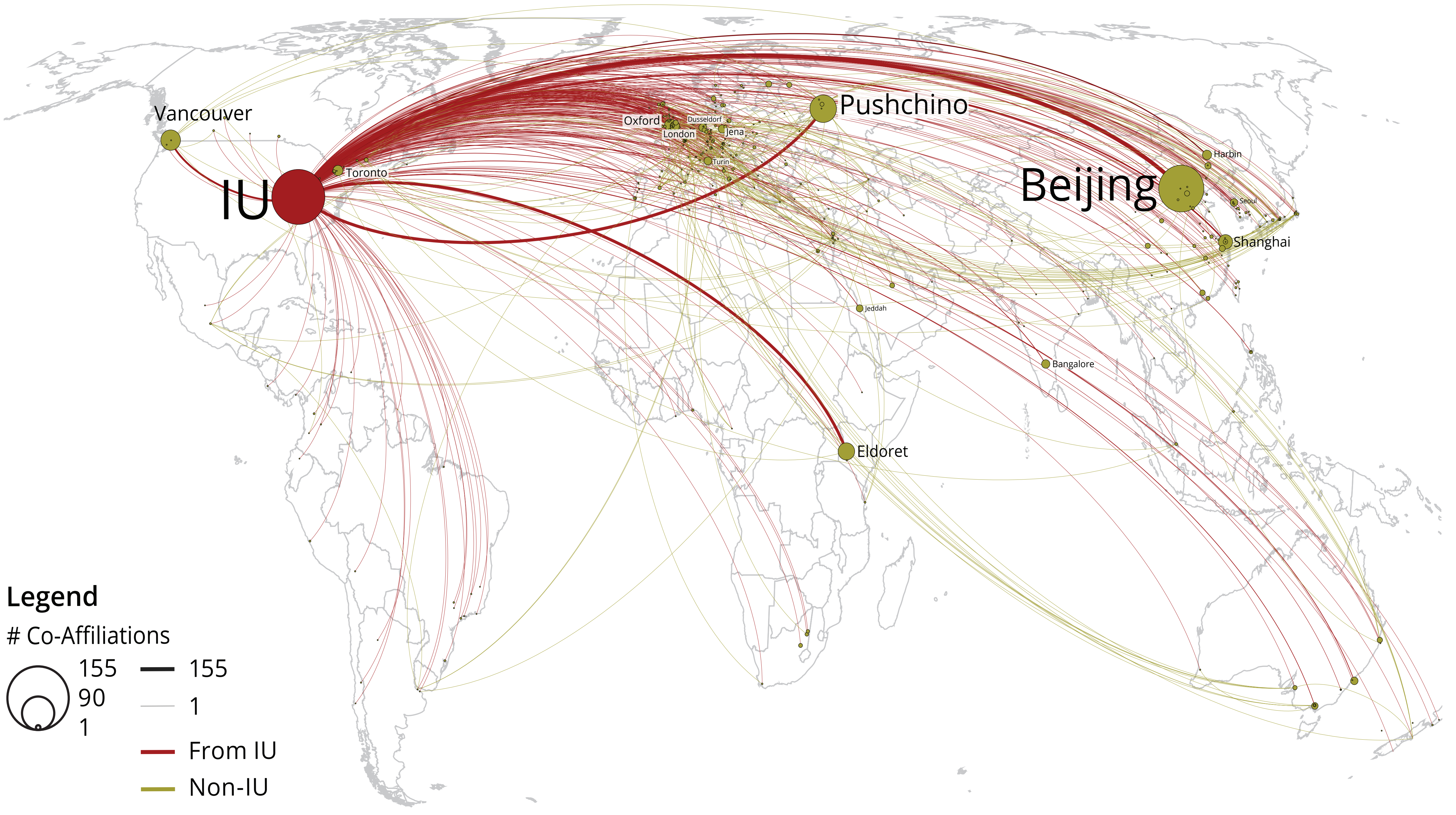

Figure 1. IU Co-Affiliation Network

New Computational Infrastructures

Fostering Global Collaboration

Among others, IU started to use the IUNI WoS data to understand existing and foster global research collaborations. The world map in Figure 1 shows the co-affiliations of authors that listed "Indiana Univ" and at least one other non-U.S. institution as affiliation on 1,590 scholarly papers published in 2004-2013. There are 344 affiliation locations (not counting IU) and 641 co-affiliation links. Nodes denote author locations and are area size coded by degree with the exception of IU, which has 1,592 co-affiliation links. Links denote co-affiliations, e.g., an author with three affiliations IU, X, Y will add three links; the two links that connect IU with X and Y are drawn in red while the link between X and Y is given in green. Links are size coded by the number of co-affiliations with the top-three being Beijing, China (155), Eldoret, Kenya (115), and Pushchino, Russia (90).

Impact vs. Disruptiveness

Researchers at the University of Chicago's Knowledge Lab and Northwestern University's NICO have used WoS data going back to 1900 to study the relationship between team size and impact and the relationship between team size and disruptiveness. This work, currently under review, finds striking differences between the scientific output of large and small teams. Looking across all fields represented in WoS, small teams are shown to disrupt science, patents, and software with new ideas and opportunities, while large teams contribute to existing ones. Figure 2 shows the relationship between impact and disruptiveness of articles (left panel) indexed by WoS, patents (middle), and software (right). In all three spaces, there is a strong, inverse relationship between citations and disruptiveness as team size increases.

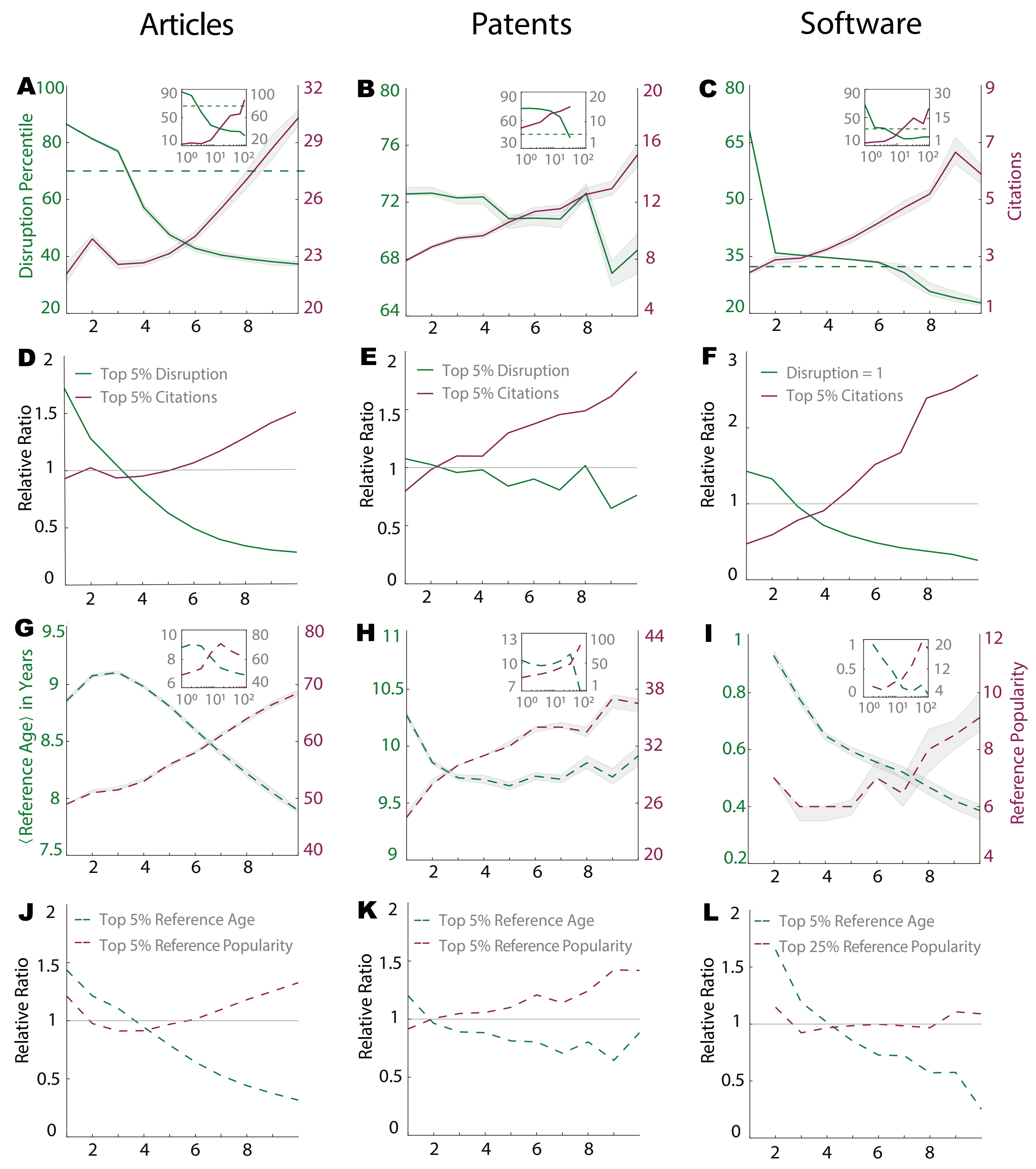

Figure 2. Tracing Inventive Teams

New Commercial Offerings

The value of the Web of Science as a search and discovery tool is well established at thousands of research institutions worldwide. But the commercial opportunities for the use of its high-quality metadata outside of the platform for big data studies are still emerging. When researchers need to study broad-scale trends in science, technology, and innovation, they very often turn to the Web of Science as the most comprehensive citation source to provide over 100 years of consistent, global publication data. Increasingly, user requests for this data take the form of custom reports, curated data sub-sets, and large-scale raw XML delivery. Clarivate Analytics is actively looking at compelling ways to meet these customer demands with new commercial products and data delivery choices. These options must balance scale and ease of use, with security and control over access to the proprietary WoS data. The lessons learned in the development of Cloud Kotta and IUNI WoS Data Enclave will very likely be instructive here, as they have proven their utility and leverage a mix of custom code built on proven commercial cloud services. Both self-service data access and secure use of analytical tools in a cloud "sandbox" seem like attractive features of these environments that could make commercial sense to meet the evolving expectation of Web of Science customers.

Acknowledgements

This work as partially supported by and contributes to research for IBM, Facebook, Jump Trading, AWS, Clarivate Analytics, the National Institutes of Health under awards P01 AG039347 and U01CA198934 and the National Science Foundation under awards NCSE-1538763, EAGER 1566393, and NCN CP Supplement 1553044. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work uses Web of Science data by Clarivate Analytics.

References

Babuji, Y. N., Chard, K., Gerow, A., & Duede, E. (2016). Cloud Kotta: Enabling Secure and Scalable Data Analytics in the Cloud. *IEEE Big Data 2016*.
Babuji, Y. N., Chard, K., Gerow, A., & Duede, E. (2016). A Secure Data Enclave and Analytics Platform for Social Scientists. *IEEE eScience 2016*.
Babuji, Y. N., Chard, K., & Duede, E. (2017). Enabling Interactive Analytics of Secure Data using Cloud Kotta. *Science Cloud Workshop: ACM International Symposium on High-Performance Parallel and Distributed Computing 2017* (Forthcoming)
Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, **http://sci2.cns.iu.edu**.