The UMETRICS Initiative

Universities: Measuring the ImpacTs of Research on Innovation, Competitiveness, and Science

Universities have a central role in documenting the results of research – from the most fundamental science and humanities to the applied projects of professional schools, research institutes and affiliated hospitals. The UMETRICS initiative brings Science of Science Policy researchers together with university Vice Presidents for Research from CIC institutions to build a scientific framework that will inform research management, enable evidence-based decision-making and support credible advocacy.

In addition to generating frontier scientific analyses, the goal of the initiative is to present results that are readily accessible to the research policy community. Our projects embed research on how universities can influence the policy debate and advance institutional ability to effectively communicate the results of research to donors, policy makers and other key stakeholders. In addition, each project will develop a set of tools that will be directly useful for research administrators, including algorithms to quantify the impact of research and tools to foster and design high-performing teams.

The foundation for our efforts is a common, large-scale, automated data platform on the research enterprise made possible by the Federal STAR METRICS project and massive advances in the methods and tools to combine, mine, and analyze big data on research. The initiative focusses on answering 3 critical questions – the structure of the research *workforce*, the nature and evolution of *collaborations*, and the *results* of research. The work will build on and develop optimal ways of communicating STAR METRICS-relevant information, examining whether different communications (style and content) are differentially effective for different audiences.

The Research Workforce Although it is clear that the structure of the scientific workforce has changed dramatically in recent years, the data are flawed and little is known about how this change affects the production of research. We will establish a set of facts about these changes based on STAR METRICS data. We will study how the complex set of cross-subsidies that operate within institutions and are built into research grants shape the scientific workforce, including the use of graduate students and postdoctoral fellows to conduct research and their outcomes after completion. Our work will also assess how the increasing diversity of the workforce changes the potential for commercialization and analyze how the aging research workforce will respond to radical changes in the research landscape. Taken together, these analyses will both help institutions better manage their workforces and better account for their activities.

Research Collaborations Answering complex scientific questions increasingly requires that scientists collaborate in complex interdisciplinary teams. We will advance the scientific basis for creating high-performing research teams and develop tools to build high-performing teams. In doing so, we will unpack the socio-technical and cultural mechanisms that explain the assembly of teams and multi-team systems and quantify how these assembly mechanisms influence their effectiveness. This knowledge will enable us to develop recommendation algorithms for new teams and multi-team systems.

Research Results The results of research have historically been captured by mechanically documenting manual reports of grant activity. The UMETRICS approach is people centric and puts scientists and their collaborations at the center of the analytical framework. We will use this framework to examine the economic results generated by research in the form of (1) the benefits to and generated by students produced by universities, (2) spillovers to regional and national economies and (3) the public value

added to social wellbeing across the scientific spectrum, including innovations in health care, the environment, energy, and food system interventions, and even improvements in policies from social science research.

Data The foundation for our projects will be a dataset that includes STAR METRICS data from multiple universities. Such data provide much broader insights because they will inevitably exhibit greater institutional variation in all dimensions of the research enterprise, such as funding strategies, lab structures and collaborations. We will also leverage enabling technologies, notably a set of automated tools developed by members of our team, and will continue to pioneer approaches to systematically harvest and incorporate data across different sources. For example, data from the Survey of Earned Doctorates, data on jobs and employment from the Census Bureau's LEHD and Non Employer data, healthcare (Medicare), innovation (USPTO), finance (VentureXpert and CRISP, IPO databases), dissertation databases, industry announcements and information from curricula vitae could be extracted, structured, disambiguated, and linked to each other and to STAR METRICS data feeds from CIC institutions. The resulting large-scale, structured, linked, updatable dataset will permit new high-quality, large-scale analyses of the scientific enterprise at a variety of levels. It will permit building a common understanding of how to use and expand complex data. Security for sensitive university data will be provided by the state-of-the-art systems, like the NORC Data Enclave, an international standard bearer in secure microdata access. In order to make sure that the data are of maximum interest to key stakeholders, we will simultaneously engage them as the project develops these data.

Research Workforce

Motivation: We know little about the structure of the research workforce, particularly in the life sciences. Little is known about how workforce structure is linked to scientific productivity. There is minimal evidence about how federal subsidies are linked to students' time-to-degree, job placement, and career tracks, yet we know Ph.D. production in the U.S. is plagued by a mismatch between supply and demand; it is estimated that only one in five U.S. graduate students in science, technology, engineering, and math will land a tenure-track position within six years of completing a Ph.D. Our analysis would address several research questions:

Workforce Structure: How is the scientific workforce structured at the project and lab level? What are the factors that determine the use of postdoctoral researchers, graduate students and staff scientists? How does the structure of projects and labs, including faculty member experience and the role of graduate students and postdocs, affect research outcomes as well as completion and job placements of trainees?

University Subsidies: How are the ways that universities subsidize the research enterprise associated with Ph.D. student outcomes? Is the size and structure of these subsidies associated with a lab's productivity, in terms of Ph.D. completion and job placement?

Workforce Diversity: How do research organization strategies affect commercialization of underrepresented groups, including African American and female researchers? How can STAR METRICS data be used to identify potential new collaborations?

Technological change and workforce aging: How do new technologies affect the productivity of scientists? What are the characteristics of scientists and institutions who adapt and benefit? Can targeted funding affect career trajectories?

University Example: How do different funding structures affect the composition of Purdue University's research workforce, as identified by the University's STAR METRICS Level I data, and the nature of the science that its researchers do? How does this, in turn, affect the placement of Purdue's postdocs and graduate students?

Collaborations

Motivation: The most important and complex research in universities require collaborations among a network of researchers. To solve today's most critical social and intellectual problems, then, universities need teams with the best possible configuration of researchers. Yet, the same research program can be organized in many different ways. A single research lab or center can compete with a network of scholars spread across many institutions. The research activity of a tenured professor can, in part, be substituted by effort from non-tenure-track researchers, post-doctorate scholars, graduate students, technicians and machines. And yet, as recent research in the science of team science has demonstrated, assembling effective teams is a daunting task at which we only succeed sporadically.

A central challenge, catalyzed by developments in human centered computing, is that the size and complexity of research teams and how they operate has changed radically Even before the recent IT revolution, universities provided early historical examples of the emergence of ad hoc teams which brought together people with different skills from their latent networks for a specific project over a finite time period. Today, unfettered to their local confines and aided by a plethora of spigots providing information about potential collaborators, researchers exercise much greater autonomy in assembling teams. Furthermore, teams exercise much greater autonomy in linking with other teams to assemble multi-team systems in order to accomplish higher-order goals of larger, often interdisciplinary, research centers. While there is an incipient awareness of how research team collaborations can conduct crucial research that can spearhead socio-economic change, we still have sparse socio-technical knowledge of how teams and systems of teams are assembled, or how a given mode of assembly impacts effectiveness. Our lack of understanding also hinders our efforts at enabling the assembly of effective teams. Our analysis would address several research questions:

Research Organization: How can a university create a research environment that organizes research to most powerfully advance discovery, acclaim, intellectual property, education and alumni success? What are the implications of these organizing choices for the cost, success, credit, communication, application and broader impact that result from university research?

Team Assembly: What are the socio-technical and cultural mechanisms that explain the assembly of teams and multi-team systems? To what extent do these assembly mechanisms influence their effectiveness?

Tools: How do recommendation algorithms based on our knowledge of these assembly mechanisms, influence the assembly and effectiveness of new teams and multi-team systems?

University Example: What are the project linkages of Northwestern University's faculty, postdocs and graduate students? How dense are the social networks? What kinds of tools can be used by VPs for Research to help form successful multi-disciplinary teams?

Research Results

Motivation: Universities perform research that spans the most fundamental science and humanities work and the applied projects of professional schools, research institutes and affiliated hospitals. Yet policy-makers, the public, and other key audiences have only a limited vision of these varied research processes and outputs. It is thus very difficult to accurately characterize the varied social and economic effects of academic research. Current measures of university outputs do not go far enough to characterize the full range of academic activities and are thus likely to misrepresent or under-estimate the social impact of research. Too few clear, compelling, and rigorous empirical studies that identify the diverse channels through which the processes and products of academic research contribute to social and economic wellbeing exist.

Put simply, there is no part of contemporary economy or society that is not touched and influenced by work done on university campuses. The research we propose here will articulate and quantify the mechanisms by which research universities contribute to multiple fields at varied time scales.

Our analysis would address several research questions:

Knowledge Spillovers. We take the primary outputs of research to be new knowledge and skilled graduates and seek to trace the varied paths students and discoveries take from campus to society. Detailing those pathways will enable systematic efforts to measure or estimate the impact of academic R&D (and thus the research university) on social and economic wellbeing. We will emphasize three broad types of social and economic impact:

(1) **Student Flows**, serve as an important conduit for spillovers by carrying *skills* from research training into a variety of sectors and organizations. What are students trained in as a result of federal funding? Where do students get jobs, in which industries and with what starting salaries?

(2) **Regional and National Economic Spillovers** How large is the *economic resilience* created by federal grant spending on campus and to local and national vendors? How large is the *economic growth* driven by entrepreneurship and innovations that underpin new products, processes, technologies and industries?

(3) **Public Value** What is the role of universities in answering pressing social problems including *population health and wellbeing, food security, energy independence and environmental quality*? How can we quantify the public value of social science and professional school research in fields ranging from education, and law, to economic, energy, and science policy?

University Example: What are the results of federal funding to research in science and technology fields at the University of Minnesota, How does the transfer of university discoveries and skills spill over to influence such key areas food security, health , water purification, environmental remediation, and robotics related to advanced manufacturing?

Data

Data have become the lingua franca for assessing the impact of scientific research and education across academia, government, industry, and society and is vital for their success. However, not all data is easy to acquire nor is it in a format that makes it usable in a large-scale systematic analysis of science and scientists. For example, data reflecting the scientific knowledge and the activities of scientists is *scattered* across a variety of sources (e.g., bibliographic data sources vs. institutional administrative databases vs. personal web pages), in *unstructured, inconsistent, and noisy* format (e.g., text and html), and is often *ambiguous* (lack labels and identifiers) and *implicit* (desired information such as gender and ethnicity or topic may need to be inferred), and *changes rapidly* over time. Much of the relevant data is born digital, but without the associated labels that make it accessible by databases and data processing systems. Even labeled data can have inadequate labels such as author names that require disambiguation. Furthermore, to make data useful for discovery and evaluation, links and tags need to be established. Manual processing is not practical for large data sets that change over time; search engines all use automated methods. A solution is automated information extraction methods which have progressed to the point of being readily applicable to many problems.

We propose to develop automated tools that will help systematically harvest data across sources, extract and structure the data, disambiguate, link and infer data elements in order to create a *large-scale, structured, linked, updatable dataset* that will permit *new large-scale analyses of the scientific enterprise* at a variety of levels by researchers that are part of the UMETRICS initiative. We will engage key stakeholders in the development of the data and tools, in order to ensure that these analyses should not only help administrators make informed decisions (e.g., what to fund, connecting researchers) but could also help individual scientists in their daily activities (e.g., to identify promising new research areas or collaborators).

More specifically, we will harvest (using focused crawlers) and extract bibliographic data linked to scholars with special attention to individuals affiliated with CIC institutions. We will cover journal/conference publications, books, patents, grants, contracts, citations from freely available sources such repositories, PubMed, CiteSeer, DBLP, Microsoft Academic Search, HathiTrust, USPTO, NIH/NSF. The structured data will be disambiguated and assigned identifiers (e.g., author names, affiliation strings, reference strings) and linked to Star Metrics data from participating CIC institutions. The overall goal is to create rich descriptions of individual scholars over time, including inferred descriptions such as gender and ethnicity when possible to infer from a given name, or topics inferred from bag-of-words. Extraction, disambiguation, inference, and linking methods will be optimized and cross-validated in supervised and semi-supervised machine learning frameworks including standard techniques such as support vector machines, conditional random fields, and logistic regression.

The STAR METRICS Level I data

The STAR METRICS (Science and Technology for America's Reinvestment—Measuring the EffecTs of Research on Innovation, Competitiveness, and Science) was born amidst efforts to stimulate the U.S. economy with the American Recovery and Reinvestment Act of 2009.(Lane, 2009; Lane & Bertuzzi, 2011) STAR METRICS is a voluntary collaborative effort to begin to build an information platform that will provide more detailed information about federal science investments and their results. It is a bottom-up approach that draws on both individual and team contributions to the scientific endeavor. Almost 100 institutions and six agencies are involved in the program, which is supported by the Federal Demonstration Partnership, as well as AAU, APLU and AAMC.

The basic Level I data provide the following information by quarter for each institution for each federal project: the number of people on each project, the amount of time allocated to each project, the occupational status of each (faculty, postdoc, staff scientist, etc.), the purchase of equipment and supplies associated with project, and the subcontracts involving researchers at other institutions.

There are four basic ways in which existing data on Federal S&T expenditures can be repurposed and traced. Individuals can be directly employed on a Federal grant. Some Federal funds go to support researchers at collaborating institutions. Scientific supplies are purchased from vendors. Infrastructure support, including financial, IT, physical space and research services is also provided. Each of these activities creates a financial transaction that can be used to calculate the associated activities. Figure 1 provides a stylized description of flow of these financial transactions in a typical administrative system. The flow on the left hand side demonstrates how the Human Resources system in a research institution can be used to identify, on a quarterly basis, the universe of individuals (Principal Investigators (PIs), co-PIs, post-doctoral researchers graduate and undergraduate students, lab technicians, science administrators, etc.) supported by any funding mechanism. Just as the LEHD program used unemployment insurance wage records to capture the flows of workers across firms, this approach tracks the expenditure trail generated by financial reporting requirements to capture each transaction charged to the funding source. All payroll transactions, which include the occupational classifications of the payees, can thus be used to automatically generate reports on who is paid, and how much, from each source of funding. Additionally, disbursements to vendors and those receiving sub-awards can be traced in the administrative records of the reporting institutions.

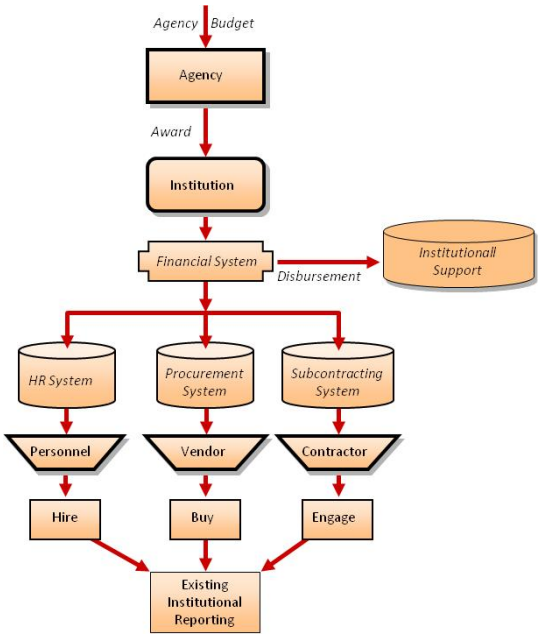

FIGURE 1: THE FLOW OF ADMINISTRATIVE TRANSACTIONS ASSOCIATED WITH FEDERAL FUNDING TO A RESEARCH INSTITUTION

The NORC Data Enclave

Launched in 2006, the NORC Data Enclave (http://www.dataenclave.org) provides a confidential, protected environment within which authorized researchers may access sensitive microdata remotely. NORC developed its secure, virtual data access platform and e-collaborative with original funding from the National Institute of Standards and Technology (NIST). Since then, the environment has grown to host sensitive microdata from a number of federal and state government agencies, including the US Department of Commerce, US Department of Agriculture, Centers for Medicare and Medicaid Services, National Science Foundation, State of Maine and as well as other data producing foundations, e.g., the Ewing Marion Kauffman Foundation, MacArthur Foundation, Annie E. Casey Foundation, and the Private Capital Research Institute. Currently, more than 350 researchers have active researcher accounts in the Enclave.

While public-use data may be disseminated in a variety of ways, fewer options exist for sharing sensitive microdata that have not been fully de-identified or anonymized. Whereas some data producers have sufficient economies of scale to develop advanced in-house solutions to serve the needs of external researchers, most lack the resources to archive, curate, and disseminate the datasets they collect. The Data Enclave provides our partner organizations with a secure platform where they can both host and sustain a thriving knowledge infrastructure around each dataset through its virtual, collaborative workspace, which enables geographically dispersed researchers to share information, replicate results, and provide feedback to fellow researchers and data producers.

As the datasets used by researchers have grown in size and complexity, the NORC Data Enclave has expanded its infrastructure to make these new datasets amenable to analysis, enabling research projects that would otherwise have been impossible. The latest version of the Enclave employs multiple high performance computing clusters that allow researchers to leverage large administrative data files, repositories of unstructured data and other emerging data sources in a rapid and efficient manner.

The NORC Data Enclave is a FISMA- and HIPAA-compliant system. It leverages multifactor authentication, thin client terminals and encryption to ensure that intruders cannot access system resources. Users have access to a virtual desktop and on-demand applications within the environment and are unable to import or remove data without prior authorization. This allows data owning organizations to know exactly where their data is at all times and to control and monitor the flow of analytic products out of the environment. Within the environment, all sponsor organizations are strictly separated from each other, and users in a particular sponsor area can only share data amongst themselves as authorized by the sponsor.

It is also worth noting that the NORC Data Enclave – or the "NORC Model" – is increasingly referred to as the international standard bearer in secure remote microdata access solutions. Among others, the "NORC Model" has been implemented for the UK Data Archive's new Secure Data Service (SDS) and is being considered by the Council of European Social Science Data Archives (CESSDA) as they plan to provide access to sensitive data across its 20-country consortium. In addition, NORC provided technical assistance to the University of Pennsylvania's and Columbia University's Population and Aging Research Centers in developing secure enclaves mimicking the NORC model.

Committee on Institutional Cooperation (CIC)

Headquartered in the Midwest, the Committee on Institutional Cooperation (CIC) is a consortium of the Big Ten member universities plus the University of Chicago. For more than half a century, these world-class research institutions have advanced their academic missions, generated unique opportunities for students and faculty, and served the common good by sharing expertise, leveraging campus resources, and collaborating on innovative programs. Governed and funded by the Provosts of the member universities, CIC mandates are coordinated by a staff from its Champaign, Illinois headquarters.

CIC Member Universities:

- University of Chicago
- University of Illinois
- Indiana University
- University of Iowa
- University of Michigan
- Michigan State University
- University of Minnesota
- University of Nebraska-Lincoln
- Northwestern University
- Ohio State University
- Pennsylvania State University
- Purdue University
- University of Wisconsin-Madison

CIC Science of Science Policy Researchers

Nosh Contractor, Northwestern University
Lisa Cook, Michigan State University
James Evans, University of Chicago
Ian Foster, University of Chicago
Lee Giles, Pennsylvania State University
Kaye Husbands Fealing, University of Minnesota
Mark Largent, Michigan State University
Maggie Levenstein, University of Michigan
Danielle Li, Northwestern University
Chris Morphew, University of Iowa
Lisa PytlikZillig, University of Nebraska
Jason Owen Smith, University of Michigan
Alan Tomkins, University of Nebraska
Vetle Torvik, University of Illinois
Bruce Weinberg, Ohio State University

And Julia Lane, American Institutes of Research

More information http://econ.ohio-state.edu/CIC/. Username=starmetrics Password: iC3mc26#