UMETRICS as a Tool for Quantifying the Value of Research and Assessing Underrepresentation

Barbara McFadden Allen (Committee on Institutional Cooperation)
Julia I. Lane (New York University)
Rebecca Rosen (New York University)
Jason Owen Smith (University of Michigan)
Bruce A. Weinberg (Ohio State University, IZA, and NBER)

Concern about the progress of women employed in STEM fields has been raised in the popular press, by science agencies and by the White House.[1] Yet little is known the environments in which women STEM researchers train, how those training environments compare to those of men, and how women's training environments affect their subsequent career outcomes. These questions are critical not just for assessing disparities in outcomes but also for improving the entire functioning of the STEM enterprise. This article introduces a new conceptual framework for understanding the value of the STEM enterprise and complementary new, linked, administrative "big data" capable of answering such fundamental questions comprehensively, economically, and at scale. These data are owned and managed by participating universities, known as UMETRICS, and housed under strict confidentiality protections at the newly-founded Institute for Research on Innovation and Science (IRIS, http://iris.isr.umich.edu)

At a conceptual level, the framework treats individual researchers as the primary actors in the STEM enterprise. People produce research. People transmit knowledge when they move from one lab to another or to industry. It is people that develop new technologies, new treatments, and products. In the words of J. Robert Oppenheimer, "The best way to send information is to wrap it up in a person".[2] Moreover, most scientific projects and most advanced research training occurs in collaborative teams that often span multiple related projects to bring together relatively large groups of people. Understanding the social environment in which training and discovery occur thus requires attention to the ways in which individuals are differently embedded in collaborative teams and the ways in which teams vary in their size, composition, and access to resources.

The heart of the new data infrastructure is a transaction-based dataset, known as UMETRICS, that captures all people paid and all purchases from vendors and sub-contracts for all federally funded grants to participating institutions. These data also provide information on the specific job titles that people hold as well as the source of funding for research projects. Such comprehensive and detailed data on the academic research workforce enables us to reconstruct, for the first time, the full teams working on research projects.

Figure 1 shows one faculty member (the large blue node), all of the people supported on grants with him, and all the people supported directly on grants with those people. In this image nodes represent individuals (red indicates female and blue indicates male) and shapes represent occupations. The figure shows that among the faculty (circles) and graduate students (squares), many of the women are connected to each other and less central to this portion of the network. While this figure represents just a small portion of the network at a single university, which may or may not be representative, such methods allow researchers to characterize the collaborative structure of science in very fine detail. We argue that these communities are the primary environment for research training and hence the appropriate level of analysis for research linking environmental differences to disparities in access or outcomes. This data can, in turn, be linked to a wide range of (i) naturally occurring data on research materials, including dissertations, publications, citations, funded grants, and patents as well as (ii) survey and administrative data held at the Census Bureau on job placements within the United States, as well as many business start-ups.

The types of analyses that can be done could quantify the value of research and shed light on underrepresentation well beyond gender differences in STEM fields. For instance, parallel methods can be applied to study underrepresentation on the basis of race, ethnicity and national origin. Moreover, the public nature of scientific works and funding and these unique data on STEM researchers allows for the quantitative analysis the informal processes of mentoring and network dynamics that determine outcomes across the economy (and in many non-economic domains) and constitute a first step toward a larger quantitative analysis of these processes.

Figure 2 illustrates the key stages and outcomes of a research career and critical variables for assessing the quality of experiences and outcomes. The following describes how the data infrastructure can provide powerful, low burden insights into research careers at each stage shown.

Quantifying STEM training environments

Training environments are formative and advanced research training is primarily based on active research experience

in laboratory or field research settings. Advanced research training is, at base, a long and often specialized apprenticeship in an established research team. The literature on the quality of K-12 education is voluminous and the literature on the quality of undergraduate education is growing, but unlike graduate training both of those types of education largely occur in relatively bounded and standardized settings such as classrooms. Perhaps as a result of the complex environment in which graduate training occurs, the literature on graduate education is underdeveloped relative to other types of education. The developing data infrastructure will provide a new window into the composition and quality of advanced training experiences.

These data make it possible to measure the size and composition of research teams in a fashion that reaches beyond

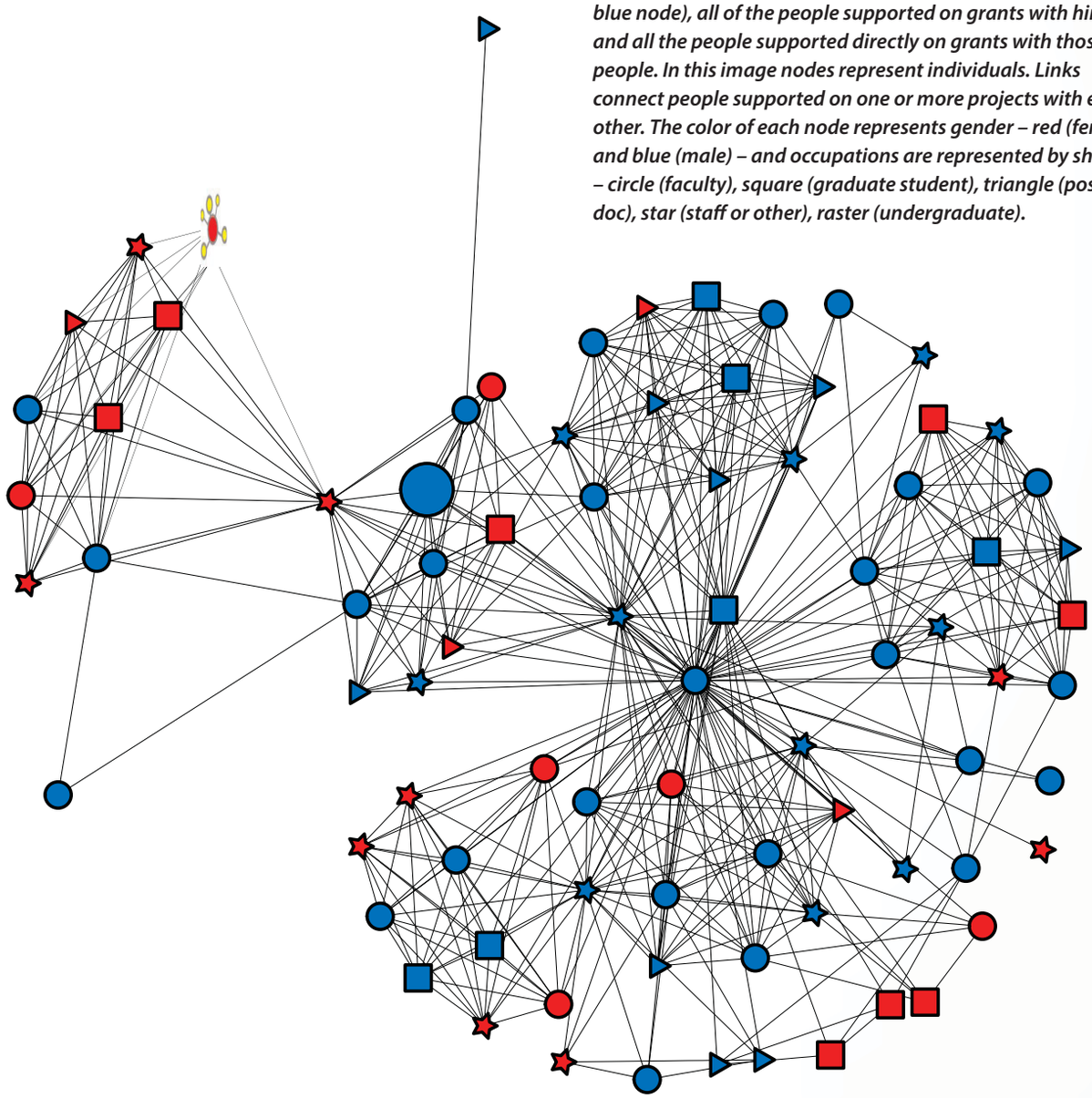

Figure 1. ***The figure shows a focal faculty member (the large blue node), all of the people supported on grants with him, and all the people supported directly on grants with those people. In this image nodes represent individuals. Links connect people supported on one or more projects with each other. The color of each node represents gender – red (female) and blue (male) – and occupations are represented by shapes – circle (faculty), square (graduate student), triangle (post-doc), star (staff or other), raster (undergraduate).***

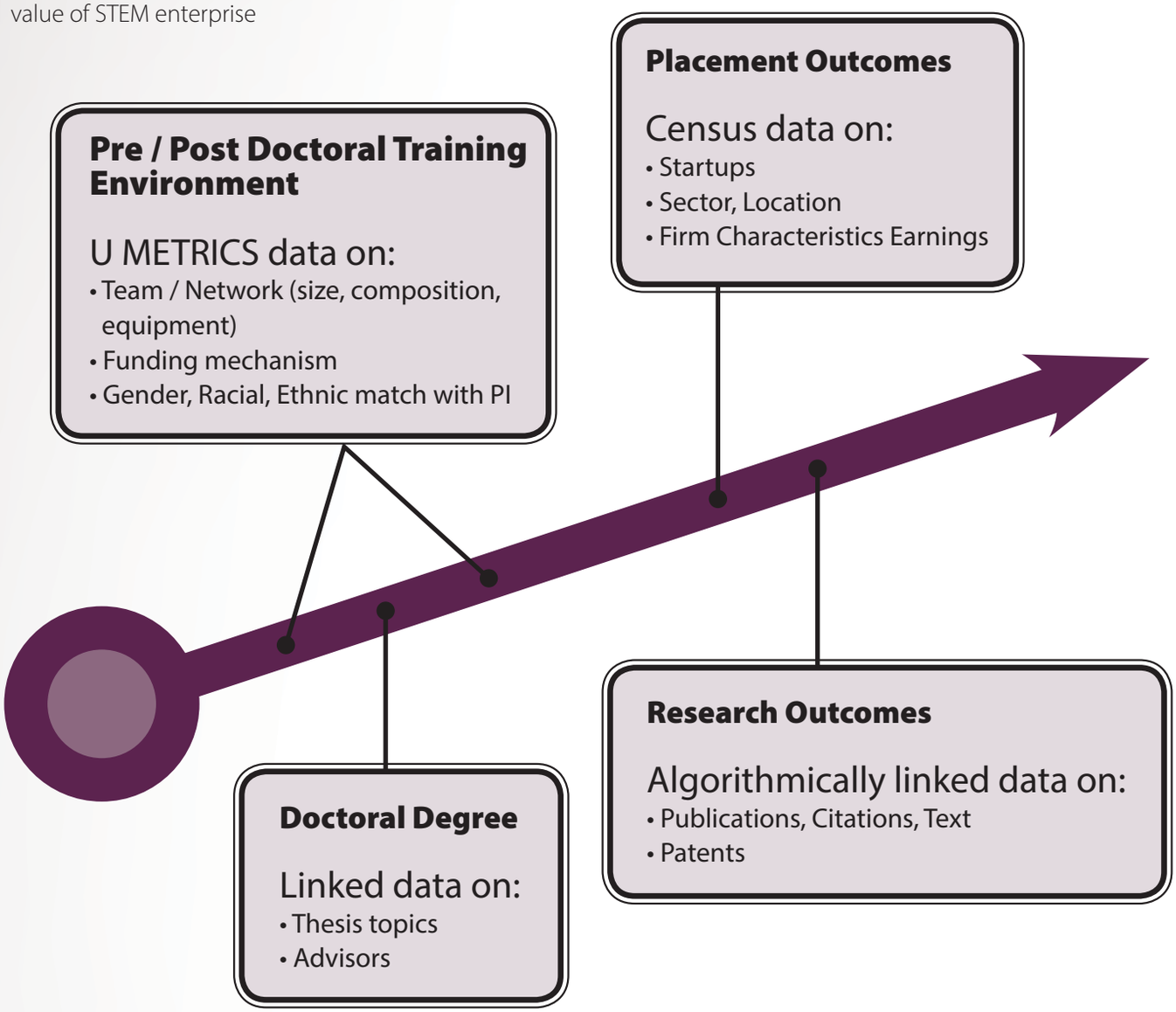

Figure 2. Career Trajectory

simply associating a principal investigator with a particular trainee. It is possible to quantify the number of faculty, staff, and postdoctoral researchers working with male and female undergraduate and graduate students on one or multiple projects throughout the students' training experience. It is also feasible to measure the gender, race, ethnicity, and national origin of trainees and whether they match advisors and/or principal investigator). As illustrated in Figure 1, there is substantial variation in undergraduate and graduate student locations within networks of research projects. Substantial research in Sociology and kindred fields suggests that these kinds of network positions are exceptionally important to success in difficult, innovative work.

How do the environments in which women train compare to those of men?

The new ability to identify the entire teams of researchers employed conducting research provides a unique opportunity to identify how the environments in which women train compare to those of men. Preliminary analysis shows that women graduate students are more likely to be employed on teams with other women and on grants with women as principal investigators, suggesting the potential for sizeable differences in training environments.

The data also make it possible to compare training environments along a wide range of dimensions. For instance, researchers can examine whether women graduate students are employed on grants with a larger share of fellow graduate students or more staff or more faculty. Researchers can compare the number of grants on which women are supported, the type and source of these grants, the length of time, and the share of time charged.

Documenting career pathways

The data allow new fundamental questions to be answered about how these various aspects of training environments relate to career outcomes, both in terms of career pathways and research production.

University data is being linked to strictly protected Census Bureau data on people and businesses to allow for the characterization of establishments (in academia, industry, or government) that hire people after they leave research, particularly the industries in which they operate, their geographic location (e.g. in the state in which the person trained or elsewhere), their size, age, growth, and wages.

The Census data also contain information on entrepreneurship. The Integrated Longitudinal Business Database (iLBD) combines administrative records and survey-based data for all nonfarm employer and nonemployer business units in the United States and hence provides information about the dynamics of firm growth and transitions from nonemployer to employer status

The people-centered approach emphasizes that where people go is critical for diffusing knowledge throughout the economy; the integration with Census data permit documentation of the extent to which research doctorates (and others employed on research projects) enter the broader economy and determine which aspects of the training environment matter for placement.

In particular, researchers can estimate how the careers paths of men and women compare to each other holding constant the rich characteristics of the training environment already discussed and also permit the identification of the long-term ramifications of any differences in training environments. Are women more or less likely to obtain academic jobs versus go into industry? Are women who go into industry more or less likely to work at smaller firms or higher wage firms or more quickly growing firms or firms that are in different industries than observationally equivalent men? And, how much of any differences can be explained by training?

Blume-Kohout finds that women supported on industry funded postdocs are more likely to participate in entrepreneurship.[3] The new ability to identify the mechanisms on which people were supported as graduate students and postdocs and then trace them through to subsequent activity can shed additional light on the decision to enter entrepreneurship and on success probabilities for a large number of researchers.

Research production

There are few economically important activities where the output of people are as readily available and as measurable as the journal articles that researchers publish. (Athletics might be another example.) The public nature of journal publications (and patents and public research funding) provide a rare opportunity to obtain fundamental estimates of how training environments relate to actual productivity. And the data are ideally suited to quantifying the research achievements of women and men, how they differ, and how any differences close or widen over the career.

The sample frame based on people employed on grants, as opposed to people listed as coauthors on publications, is unique and particularly powerful way of studying the determinants of authorship. Specifically, researchers can examine the publication patterns of all the people who were employed on a project as well as their jobs and time charged to it. In this way, researchers can quantify the extent to which women are less (or more) likely to appear on coauthors on articles and assess how the ordering of authors differs controlling for a wide range of measures of involvement.

Conclusion

The new data infrastructure constitutes an important opportunity for breakthrough research on science and innovation that can inform many aspects of science policy. In addition to issues related to underrepresentation of women and other groups, they will support a wide range of analyses of the creation, transmission, and utilization of ideas and at an unprecedented level. They will rely on algorithmic, "big data" methods to combine and mine data from a wide range of sources at low burden. And, the resulting, confidentiality protected, data will be made available to the research community through the newly founded Institute for Research on Innovation and Science (IRIS) (http://iris.isr.umich.edu/).

References

[1] http://www.cnet.com/news/women-in-tech-the-numbers-dont-add-up/#ftag=CADf328eec

[2] J. Robert Oppenheimer as quoted in Anonymous. "The Eternal Apprentice." Time 52 (November 8, 1948): pp. 81.

[3] Margaret E. Blume-Kohout. 2014. "Understanding the Gender Gap in STEM Fields Entrepreneurship." MBK Analytics, LLC.