National Institute of General Medical Sciences

Analysis and Modeling of Demographic Cohorts in the Population of PhD Recipients in Science and Engineering

PRB

Pacific Northwest NATIONAL LABORATORY

INDIANA UNIVERSITY

Antonio Sanfilippo – PNNL

Katy Börner, James P. Crutchfield, Dowman P. Varn et al. – IU, UCDavis

Marlene Lee – PRB

Progress report, SWAM Program, January 26, 2012

Overview

- Research Aims
- Progress report (by team)
  - Data Access and Analysis
  - Modeling
  - Future work

Research Aims

- Collect and analyze evidence relevant to demographic dynamics of the scientific workforce and related activities
- Use evidence collected to develop, calibrate and evaluate evidence-based dynamic models of population change in the scientific workforce
- Answer questions such as
  - What is the impact of the increasing proportion of foreign graduate students and postdoctoral scholars?
  - What are the proportions of female and minority students?
  - What are the effects of geolocation on a successful career?
  - How much should the rate of participation be increased to get to population proportions?

National Institute of General Medical Sciences

PNNL Progress Report

- Data Access and Analysis
- Initial data analysis insights
- Initial modeling efforts
- Future work

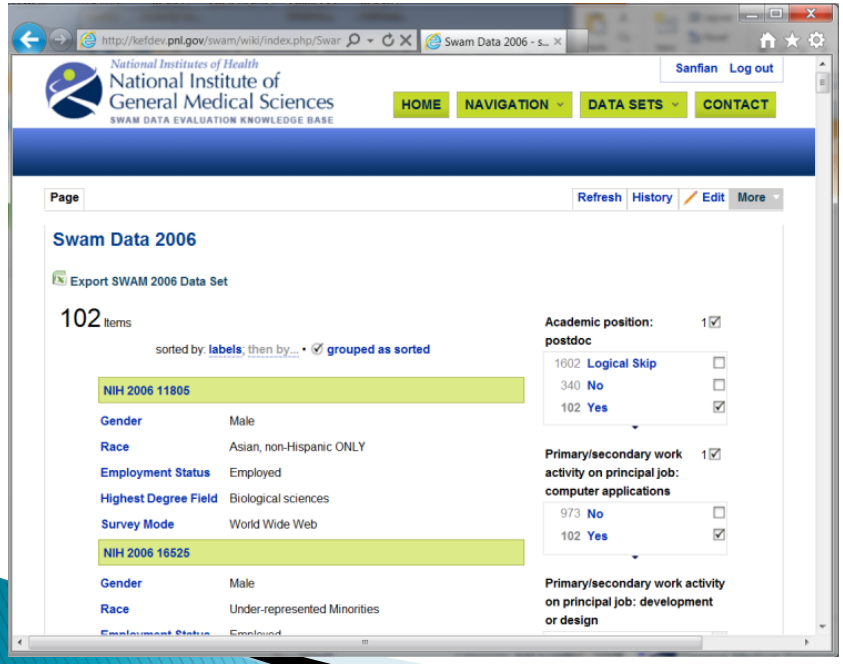
PRB

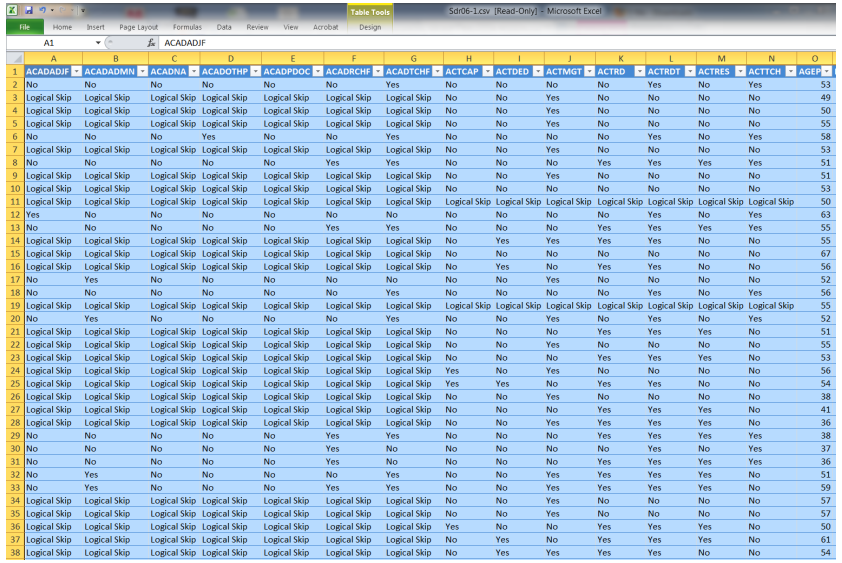
Pacific Northwest NATIONAL LABORATORY

INDIANA UNIVERSITY

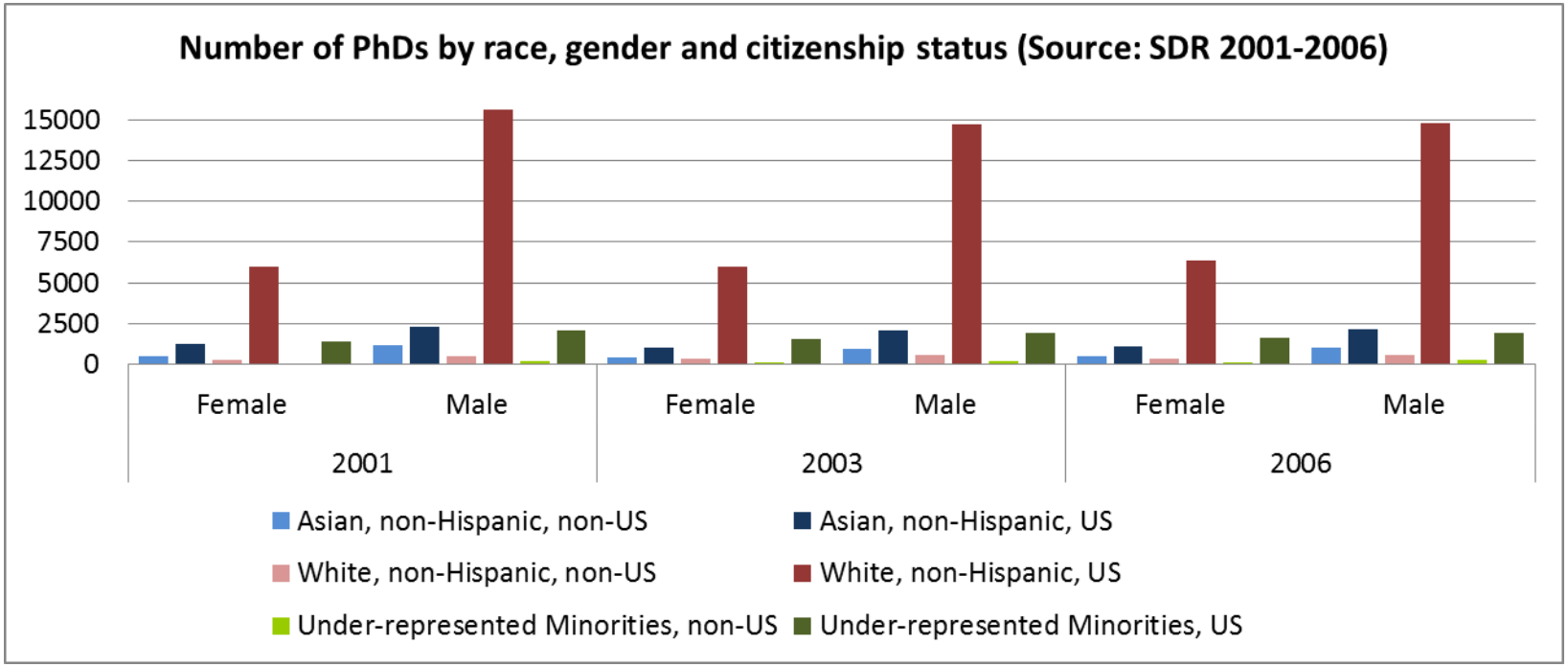
Data Access and Analysis

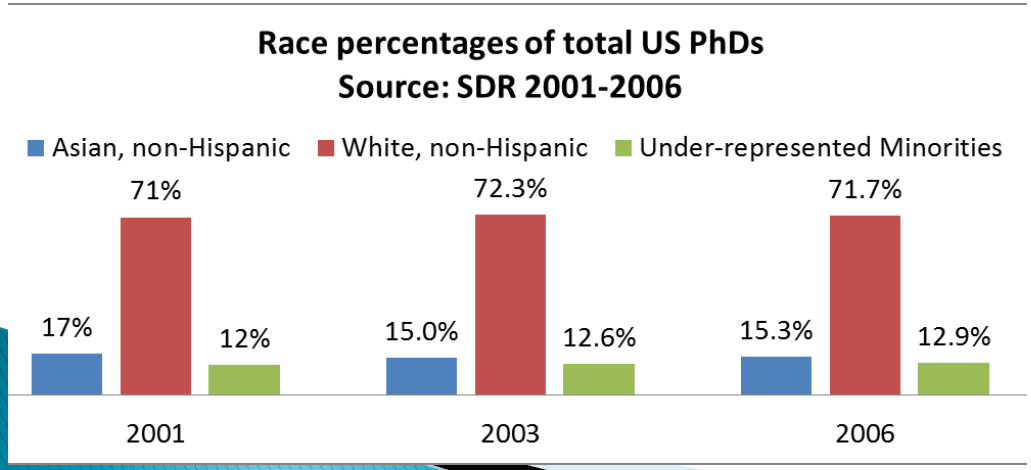
- Received NSF approval for licensing restricted SDR, NSRCG and integrated SESTAT Data 1993-2008
- Developed knowledge discovery wiki for SWAM with 2003 and 2006 SDR public data
  - Use **faceted search** to create data subset and export these as MS Excel CVS files suitable for analysis and modeling

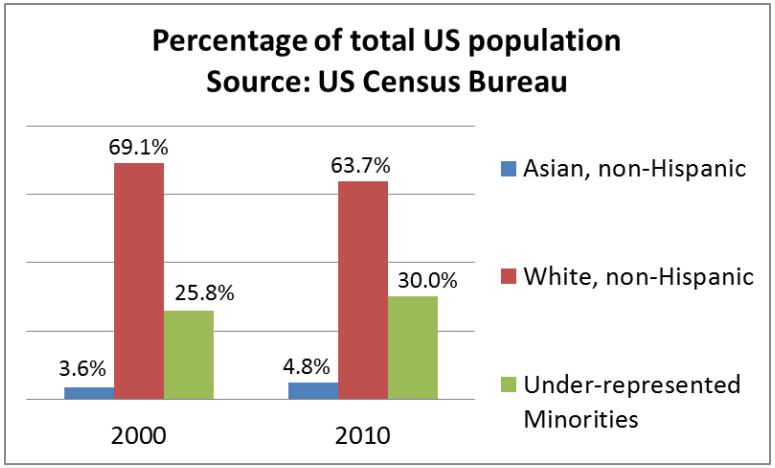




Initial data analysis insights

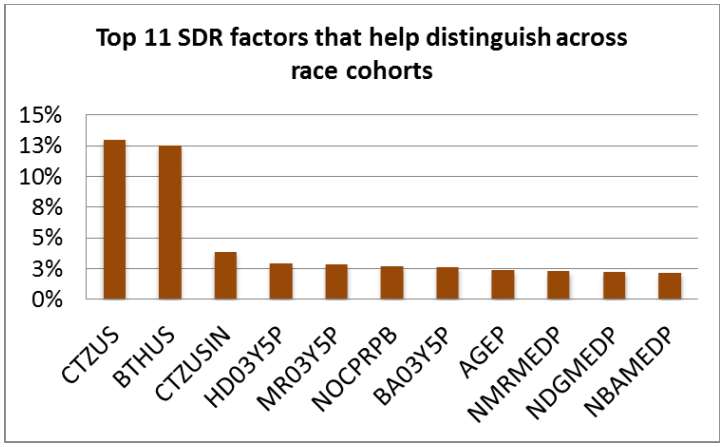


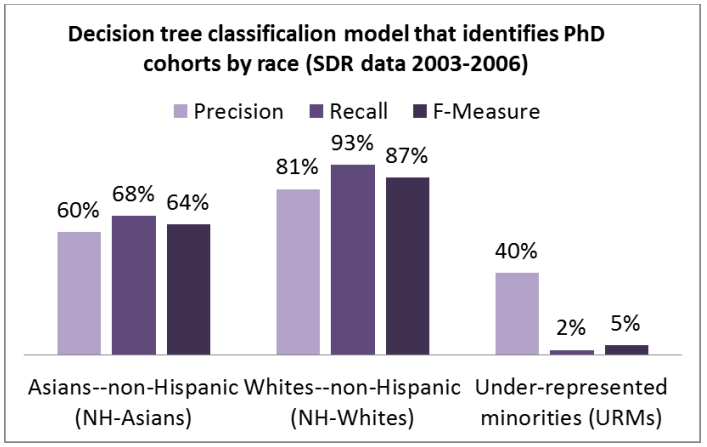




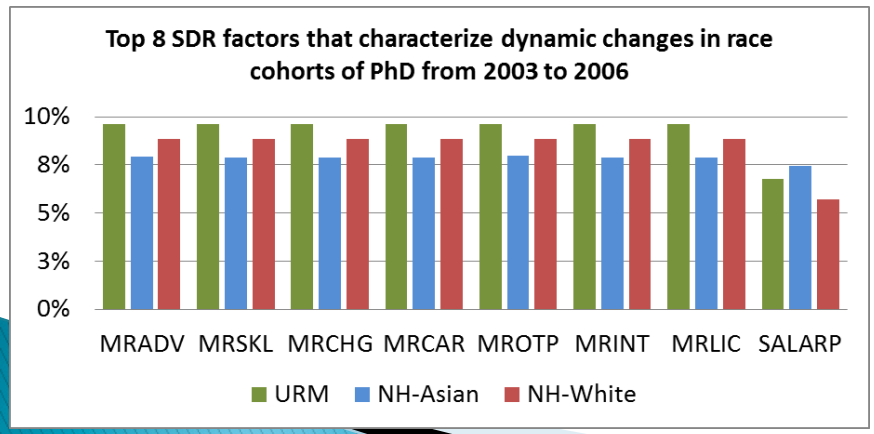
Initial modeling efforts

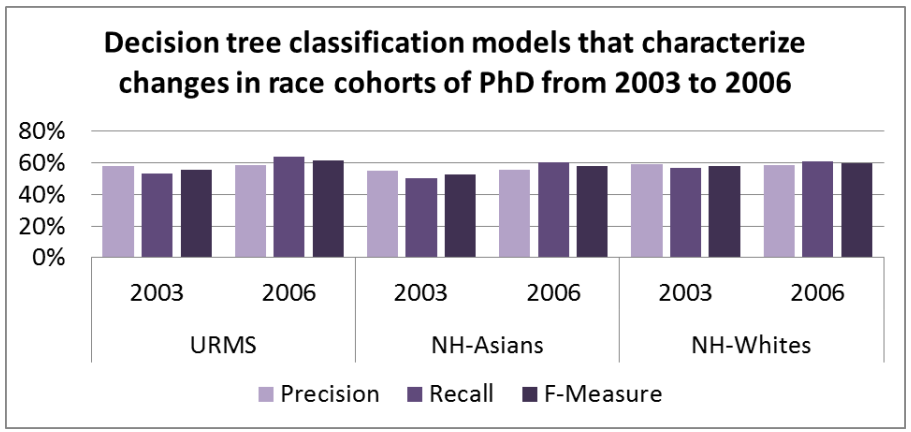
- Developed and evaluated classification models that
  - Identify URM, NH-White and NH-Asian PhD cohorts





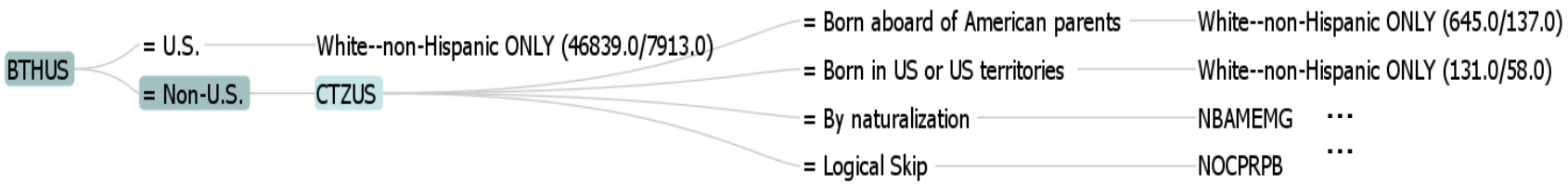
  - Characterize dynamic changes across each race cohort





Next Steps

- Modeling
  - Improve performance, visualization and analysis of models



  - Develop gender-based models
  - Normalize analyses and models using census data (w/PRB)
  - Start bringing science policy factors to bear on modeling task
  - Develop dynamic models of population change in the scientific workforce
  - Start working with restricted data
- SWAM Wiki
  - Make available to SWAM members at ***swam-us.org***
  - Load remaining public SESTAT data
  - Improve usability and add visual analytic functionality
  - Create standalone SWAM Wiki for restricted data

National Institute of General Medical Sciences

PRB

Pacific Northwest NATIONAL LABORATORY

INDIANA UNIVERSITY

Indiana University Progress Report

- Data Acquisition & Analysis
- Theory & Definitions
- Modeling Scholarly Dynamics
- Future work

Dowman P Varn,[1,2] Katy Börner,[1] Robert P Light,[1] Scott B Weingart,[1] & James P Crutchfield[2]

[1]Cyberinfrastructure for Network Science Center
School of Library & Information Science
Indiana University

[2]Department of Physics & Complexity Sciences Center
University of California, Davis

Data Acquisition & Analysis

- The National Science Foundation Survey of Doctoral Recipients (NSF-SDR) data contains information about recent PhDs in the sciences for the years 1993, 1995, 1997, 1999, 2001, 2003, 2006, & 2008. This longitudinal data give a rich, detailed picture of the scholars' career evolution. We are in the process of acquiring the complete surveys from NSF.
- In collaboration with Vincent Larivière of Universitè du Quèbec à Montrèal, we now have bibliometric data from the Web of Science (WoS) for the 13,513 most moved physicists between 1980 and 1987, each of whom has published at least six papers. This sample contains 258,021 publications with a total of 4,120,342 citations and is expected to be rich in postdoctoral researchers. We can track their geolocation vs. time as well as any topical changes in their research interests.

Theory & Definitions

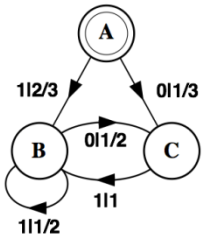
- Under what theoretical framework do we couch our questions? A necessary first task is to introduce mathematical definitions that parameterize our problem.
- **Scholar**: Let a scholar, $\alpha^{(i,j,k)}$, be an autonomous agent that can perform certain actions, *i,* (e.g., publish papers, work at institutions), has certain attributes, *j,* (e.g., gender, ethnicity), and can store and process information, *k*.
- **Event**: Let an event, ω_m, be something that happens to an agent or its environment, or that the agent does to itself or its environment. Examples are the publication of a paper or movement to another location.
- **Career**: Let the career of a scholar, Ψ_α, be the time ordered sequence of events that spans the active life of the scholar, i.e., $\Psi_\alpha = (\omega_1, \omega_2, \ldots, \omega_M)$.

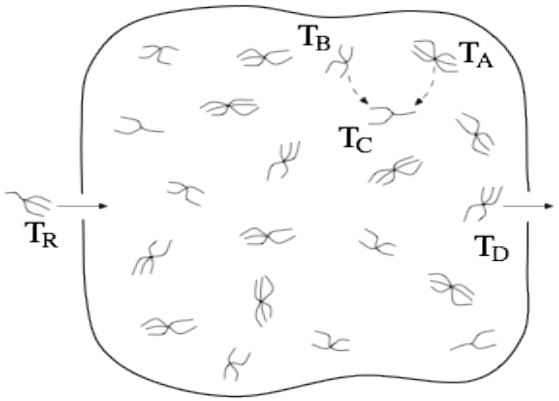
Modeling Scholarly Dynamics

- What parts of the dynamics of scholarly communities and of individual careers are endogenous and what parts are exogenous? That is, what is due to internal, spontaneously formed patterns inherent in the social population dynamics, and what is due to external funding, scholarly fashion, or other environmental factors? More pointedly, what in the observed dynamics of scholarly activity is amenable to change from external influences? What can policy makers really do?

Modeling Scholarly Dynamics II

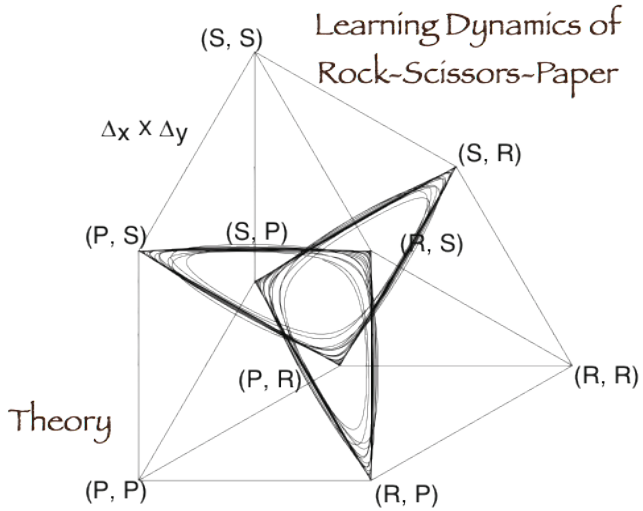
- Recent work in game-theoretical dynamical systems (Sato & Crutchfield, 2002; Sato et. al. 2004) and population dynamics (Gornerup & Crutchfield, 2008) shows that a rich variety of endogenous behaviors, including collective adaption and spontaneous hierarchical organization, can form spontaneously in multiagent systems.
- We are extending these models to scholarly dynamics.





To the left, an interacting "soup" of scholars.

To the right, the complex dynamics of the rock-scissors-paper game.



PRB INDIANA UNIVERSITY Pacific Northwest NATIONAL LABORATORY National Institute of General Medical Sciences

Next Steps

- Complete the theoretical (conceptual) framework necessary to parameterize the problem.
- Continue our analysis of the WoS data, in particular looking for patterns of changing geolocation and research interests and how these relate to productivity and success.
- Continue our preparations for the NSF-SDR survey data, and once acquired, begin extracting information about gender, ethnicity, movement, and topical interests of the participants.
- Develop models of active, interacting scholars. How well do these models predict publication patterns (both over topic space and geolocation space) and career trajectories? Do the parameters of this model differ for different genders and ethnic groups? What predictions do these models give when we introduce exogenous shocks to the population?

National Institute of General Medical Sciences

PRB Progress Report

- Data Access
- Analysis and Modeling
- Future work

PRB

Pacific Northwest NATIONAL LABORATORY

INDIANA UNIVERSITY

Status Update – Data Access

- Application for secure data in progress
- Computer and software acquisitions completed
- Testing new version of SPACE software for life tables with simulations

Status Update – Analysis & Modeling

- Definitions of population groups
- Definition of employment states
- Functional form of hazard models for transitions

Status Update – Next Steps

- Transition Matrices
- Life Table Simulations
- Involvement of collaborators (MPI)
- Research Assistant