

# Automatically assembling a full census of an academic field

Allison C. Morgan<sup>1\*</sup>, Samuel F. Way<sup>1</sup>, Aaron Clauset<sup>1,2,3</sup>

**1** Department of Computer Science, University of Colorado, Boulder, CO, USA

**2** BioFrontiers Institute, University of Colorado, Boulder, CO, USA

**3** Santa Fe Institute, Santa Fe, NM, USA

\* allison.morgan@colorado.edu

## Supporting information

**S2 Appendix. Features of a faculty directory.** In addition to the stemmed words from directory and non-directory web pages, derived features of our random forest classifier included names, email addresses, phone numbers and titles. The complete feature set for our directory classifier can be found online at [https://github.com/allisonmorgan/academic\\_census](https://github.com/allisonmorgan/academic_census).