Cross-Disciplinary Evolution of the Genomics Revolution

Alexander M. Petersen,[1] Dinesh Majeti,[2] Kyeongan Kwon,[2] Mohammed E. Ahmed,[2] Ioannis Pavlidis[2]

[1]Ernest and Julio Gallo Management Program, Department of Management of Complex Systems, School of Engineering, University of California Merced, California 95343

[2]Computational Physiology Laboratory, University of Houston, Houston, Texas 77204

Data Description

Summary

Merged Google Scholar, department affiliation, and NSF/NIH funding data for *Cross-disciplinary evolution of the genomics revolution* (Science Advances, 2018).

The assembly of the career data set. We selected 155 biology and computing departments in the United States following the 2014 U.S. News & World Report (see table S1 in ref. [1]). We confirmed that all the departments in the set had active PhD programs since the conception of Human Genome Project in the 1980s. We found no significant differences in the ranks of these 155 departments between the 2014 and 2018 U.S. News & World Report ranking ($P > 0.05$, Wilcoxon test).

We accessed the home pages of these departments and recorded the listed faculty as of spring 2017. In this master list (see **Faculty_GoogleScholar_Funding_Data_N4190.csv**), we identified the faculty, denoted by $\mathcal{F}_i$, that had Google Scholar (GS) pages. From this set we formed a database with their GS IDs, h-indices, departments, department rankings, and other aggregate bibliometric data. We also indexed their NSF and NIH grant data from the corresponding open-data repositories.

Data files included:

File 1: Faculty_GoogleScholar_Funding_Data_N4190.csv
File 2: GoogleScholar_paper_stats.csv
File 3: ComputerScience_citations_stats_CitationNormalizationData.csv
File 4: Biology_citations_stats_CitationNormalizationData.csv

Data in File 1 and File 3 can be merged via the Google Scholar unique identifier (google_id). File 3 and 4 are useful for converting the nominal citation counts $c_{i,p}$ provided in File 2 into the

normalized citation measure $z_{i,p}$ defined in Eq. (4) of [1]. This normalization method yields a stationary, normally distributed citation measure that is well suited for identifying longitudinal patterns of citation impact within and across research careers.

(i) Faculty data — merged Google Scholar and U.S. Funding data:

- **Faculty_GoogleScholar_Funding_Data_N4190.csv**

Each measure is specific to researcher $\mathcal{F}_i$:

google_id: [GS_i, string] Google Scholar unique identifier corresponding to $\mathcal{F}_i$, accessible at: https://scholar.google.com/citations?user=GS_i&hl=en

name: [string] Full name of $\mathcal{F}_i$

dept: [categorical] Primary departmental affiliation = CS or BIO

h_index: [h_i, integer] Hirsch h-index calculated by Google Scholar

i10index: [integer] i10-index calculated by Google Scholar

min_year: [y_i^0, integer] Year of first publication

max_year: [y_i^f, integer] Year of most recent publication

t_publication: [integer] Number of publications in Google Scholar profile

t_pubs_citations: [C_i, integer] Total number of citations calculated by Google Scholar

highest_citations: [integer] Total number of citations for the highest-cited publication

mean_of_IF: [$\overline{IF}_i$, float] Mean "impact factor" across the researcher's set of journal venues; impact factor values obtained from the 2015 Journal Citations Report (JCR).

mean_of_co_authors: [float] Mean number of coauthors per publication

num_nsf: [N_i^{NSF}, integer] Number of NSF grants

t_deflated_nsf: [$\$_i^{NSF}$, integer] Total amount of NSF funding deflated to 2010 U.S. dollars

num_nih: [N_i^{NIH}, integer] Number of NIH grants

t_deflated_nih: [$\$_i^{NIH}$, integer] Total amount of NIH funding deflated to 2010 U.S. dollars

Y05yr: [$y_{i,5}^0$, categorical] min_year grouped into 5-year periods

KTotal: [K_i, integer] Total number of coauthors, independent of coauthor type ($\mathcal{F}_i$ or pollinator)

KDirect: [K_i^D, integer] Total number of $\mathcal{F}_i$ coauthors

KMediated: [K_i^M, integer] Total number of pollinator coauthors

Chi: [χ_i, float] The degree of cross-disciplinarity, defined as the fraction of her/his KTotal collaborators who are cross-disciplinary.

BetCentrality: [$\mathscr{C}_i^B$, float] Betweeness centrality calculated for the "direct network" that includes only $\mathcal{F}$

PRCentrality: [$\mathscr{C}_i^{PR}$, float] PageRank centrality calculated for the "direct network" that includes only $\mathcal{F}$, using "damping factor" $d = 0.85$

XDIndicator: [O(Φ_i), categorical] 0 = $\mathcal{F}_i$ only collaborated with other faculty with dept **=** BIO; 1 = $\mathcal{F}_i$ only collaborated with other faculty with dept **=** CS; 2 = $\mathcal{F}_i$ collaborated with both types of faculty, and is categorized as cross-disciplinary (O(Φ_i)$\equiv XD_\Phi$)

SchoolRank: [r_i, integer] Rank of department according to the 2014 U.S. News & World Report (see Table S1 in ref. [1])

(ii) Google Scholar — publication and citation data:

- **GoogleScholar_paper_stats.csv**

Panel data – each observation (data line) refers to a single Google Scholar item with publication year in the range [1970,2015], as implemeted in the panel regression analysis:

google_id: [GS_i, string] This publication belongs to the Google Scholar profile indicated by this unique identifier

year: [t_p, integer] Publication year

citations: [c_p, integer] Number of citations reported by Google Scholar in 2017

coauthor_codes: [string] Indicates the number of coauthors and types of coauthors belonging to the publication. Numbers indicate "pollinator" coauthors (individuals who do not belong to the group of 4,190 $\mathcal{F}$ set). These pollinators are coded as 0,1, or 2, whereas $\mathcal{F}_i$ are coded according to their google_id. For example, the first publication line is: "1,1,1,--nVNvIAAAAJ", indicating that this publication has 4 coauthors, and only one is a faculty member ($\mathcal{F}_i$), which in this case corresponds to the google_id corresponding to the Google Schoar profile researcher. Pollinator codes correspond to their disciplinary type determined by the union of that particular individual's set of coauthors: 0 = pollinator *j* only appeared in our dataset with other BIO $\mathcal{F}_i$; 1 = pollinator *j* only appeared with other CS $\mathcal{F}_i$; and 2 corresponds to a mixture of CS and BIO $\mathcal{F}_i$ - i.e. coauthor *j* is a cross-pollinator.

(iii) Citation distribution normalization data – each descriptive statistic is calculated from all publications observed from the indicated year:

- **ComputerScience_citations_stats_CitationNormalizationData.csv**
- **Biology_citations_stats_CitationNormalizationData.csv**

YEAR: [t_p, integer] Year of publication cohort for which we compute the statistics for each discipline

num_pub: [integer] Number of publications in that particular year and discipline

sum_citations: [integer] Total number of citations accrued by all publications published in that year and discipline

std_citations: [float] Standard deviation of citations accrued by all publications published in that year and discipline

sum_ln_citations: [float] Sum of ln $(1 + c_p)$ accrued by all publications published in that year and discipline

std_ln_citations: [float] Standard deviation of ln $(1 + c_p)$ accrued by all publications published in that year and discipline

These data are needed to normalize the nominal c_p values. This is achieved by removing the time-dependent trend in the location and scale of the underlying log-normal citation distribution P(c_p) by defining

$$z_{i,p} \equiv \left[\ln\left(1 + c_{i,p,s,t}\right) - \mu_t\right]/\sigma_t, \qquad \text{(see Eq. (4) in [1])}$$

where $\mu_t \equiv \overline{\ln\left(1 + c_{s,t}\right)}$ is the mean and $\sigma_t \equiv \sigma\left[\ln\left(1 + c_{s,t}\right)\right]$ is the standard deviation of the citation distribution, after adding 1 to each citation tally (to avoid the divergence of ln 0 associated with uncited publications) and applying the natural logarithm. As such, we calculated μ_t and σ_t within the subset of publications for a given year t_p and discipline: BIO or CS.

Publications:

[1] A. M. Petersen, D. Majeti, K. Kwon, M. E. Ahmed, I. Pavlidis, *Cross-disciplinary evolution of the genomics revolution*. Sci. Adv. 4, eaat4211 (2018).