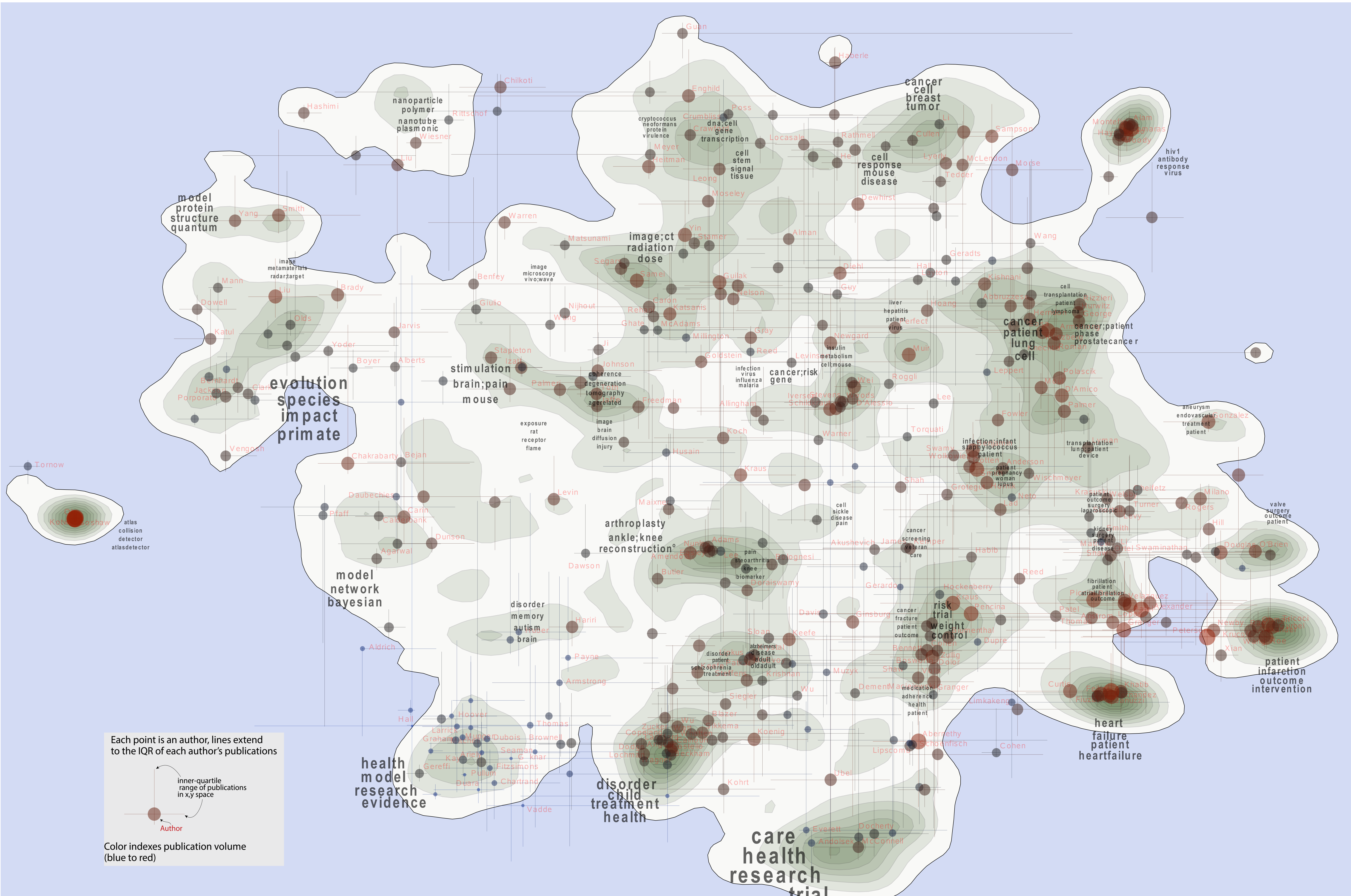
Towards an Intellectual Atlas of Scholars@Duke

DNAC Duke Network Analysis Center
DUKE UNIVERSITY
James Moody, Peter Mucha & The DNAC Lab



Each point is an author, lines extend to the IQR of each author's publications

Color indexes publication volume (blue to red)

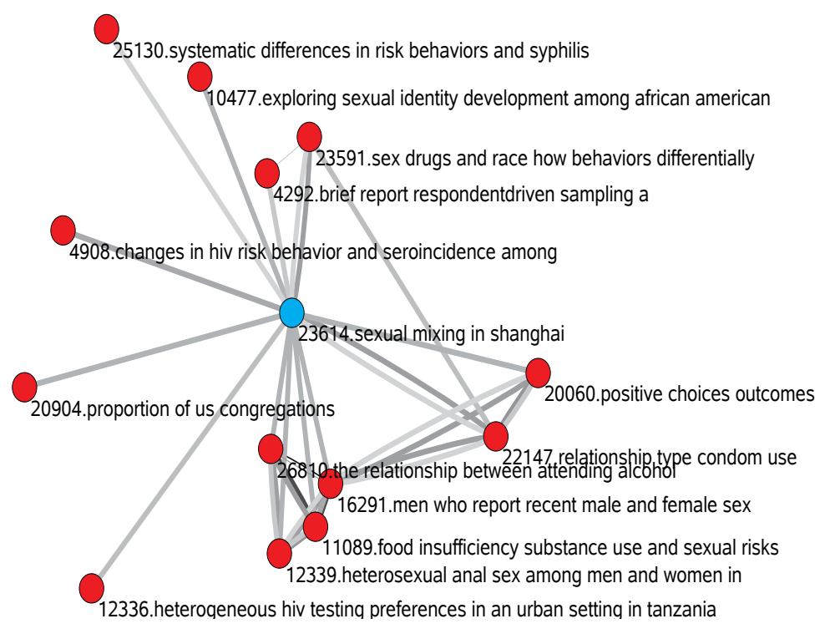
Interactive version provides detail on authors & publications: http://www.soc.duke.edu/~jmoody77/S@D/starfield_ia.html

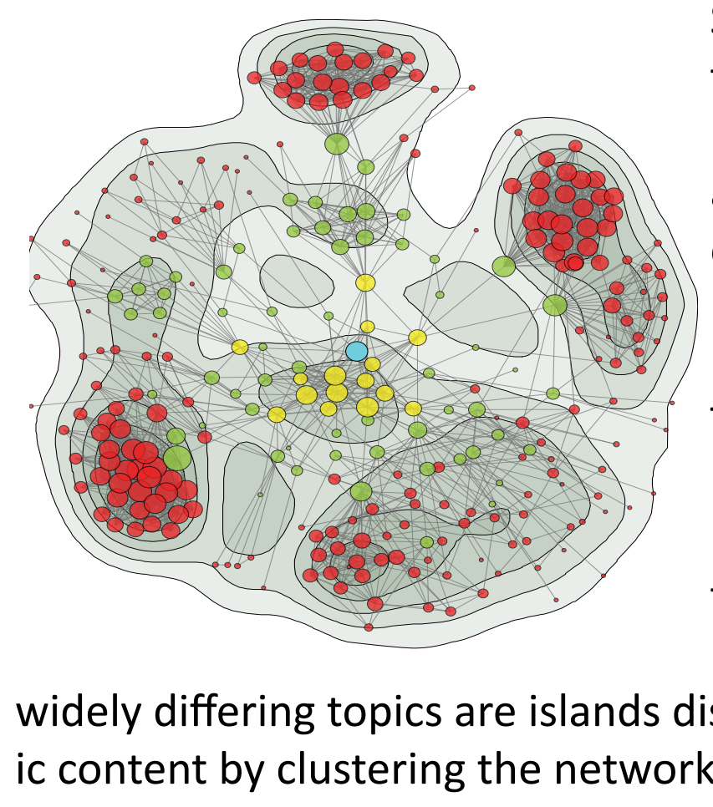
Goal

Our goal is to map scholarly production at Duke and use the distribution of scholarship as a frame for displaying relations amongst scholars.

Approach

Defining a landscape. Publications are the fundamental unit of scholarly production, so we start by building a publication network. We limit our sample to all unique papers with an abstract and construct links between papers based on how similar their abstracts are to each other. After some adjustments for density, we project the network in two dimensions using a layout algorithm that minimizes distance between connected nodes. This has the effect of placing papers with high similarity values close to each other in the display space, effectively grouping topically similar papers together.


25130.systematic differences in risk behaviors and syphilis
10477.exploring sexual identity development among african american
23591.sex drugs and race how behaviors differentially
4292.brief report respondentdriven sampling a
4908.changes in hiv risk behavior and seroincidence among
23614.sexual mixing in shanghai
20060.positive choices outcomes
20904.proportion of us congregations
22147.relationship type condom use
26810.the relationship between attending alcohol
16291.men who report recent male and female sex
11089.food insufficiency substance use and sexual risks
12339.heterosexual anal sex among men and women in
12336.heterogeneous hiv testing preferences in an urban setting in tanzania



Since the network is too large to display in traditional point and line format, we estimate a 2-dimensional kernel density function and use that as a smooth representation of the underlying distribution of papers in the network, which we represent with contours.

The distribution of papers across this space represents a landscape with varied topography—hills represent clusters of papers on similar topics, ridges link topics to each other, valleys represent gaps in the knhowledge space and widely differing topics are islands disconnected from the mainland. We identify topic content by clustering the network, and label the largest clusters with the terms found most frequently in each cluster's titles (label size proportional to number of papers).

Populating the Atlas

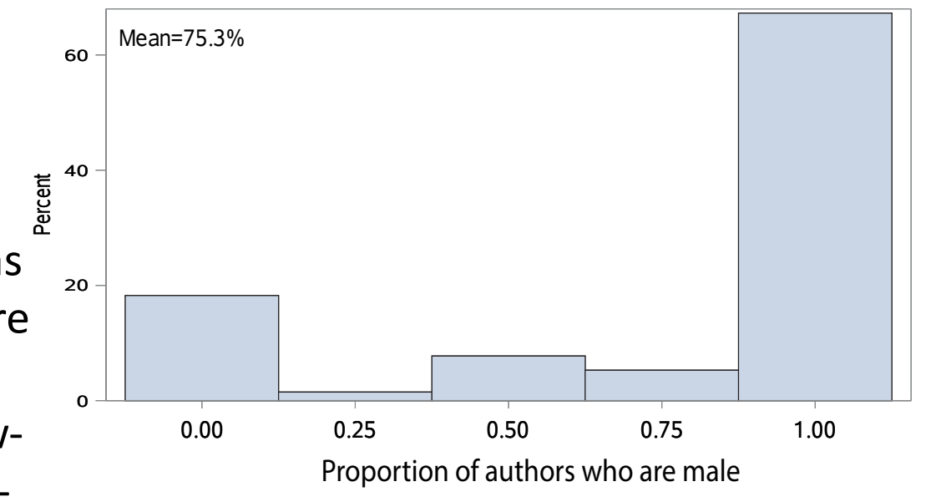
Like a geographic atlas, once we know the topography we can layer other information over this space. Here, we provide layers on high-volume producers, departments/units, gender, and collaboration.

Foxes and Hedgehogs. In the main figure, we have selected authors who are ranked in the top 15% by publication volume in their primary affiliation unit and place them at the centroid of their publications. Berlin (1953) famously contrasted thinkers working on a single area (hedgehogs) to those who draw widely from multiple intellectual sources (foxes). To capture this variability in topical range, we extend lines from each author to capture the inner-quartile range of their publication distribution across the space. Those who publish on a wide range of topics will have a longer reach across the space than those that focus narrowly in on a particular topic.

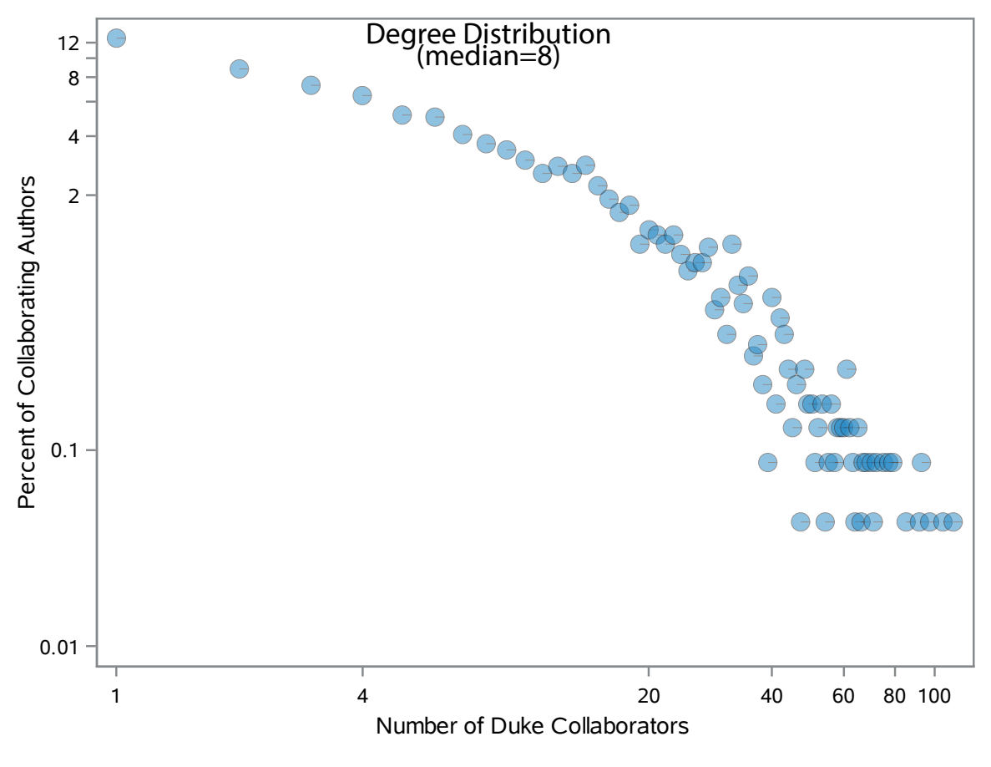
Departmental Production. Academic departments control hiring, tenure and promotion and thus generally shape the broad contours of scholarly production at the university. Layering departmental coverage over the intellectual landscape lets us see how similar departments are to each other, highlighting opportunities for collaboration. The dominant role of the Medical school is clear here as well (though some of this is due to coverage bias in the corpus, see limitations).

Gender Distribution. Concern over the involvement and retention of women in STEM fields is a top priority for the NSF (Cordova, 2016). By looking at the gender distribution of authors on each publication, we can estimate the gender representation across the intellectual landscape. Scholarly production at Duke is disproportionately male. While women are active in most areas of the space, they are somewhat more common in the community health and population research clusters (lower-right) and more rare in the computational areas (center-left).

Gender Distribution of Authorship

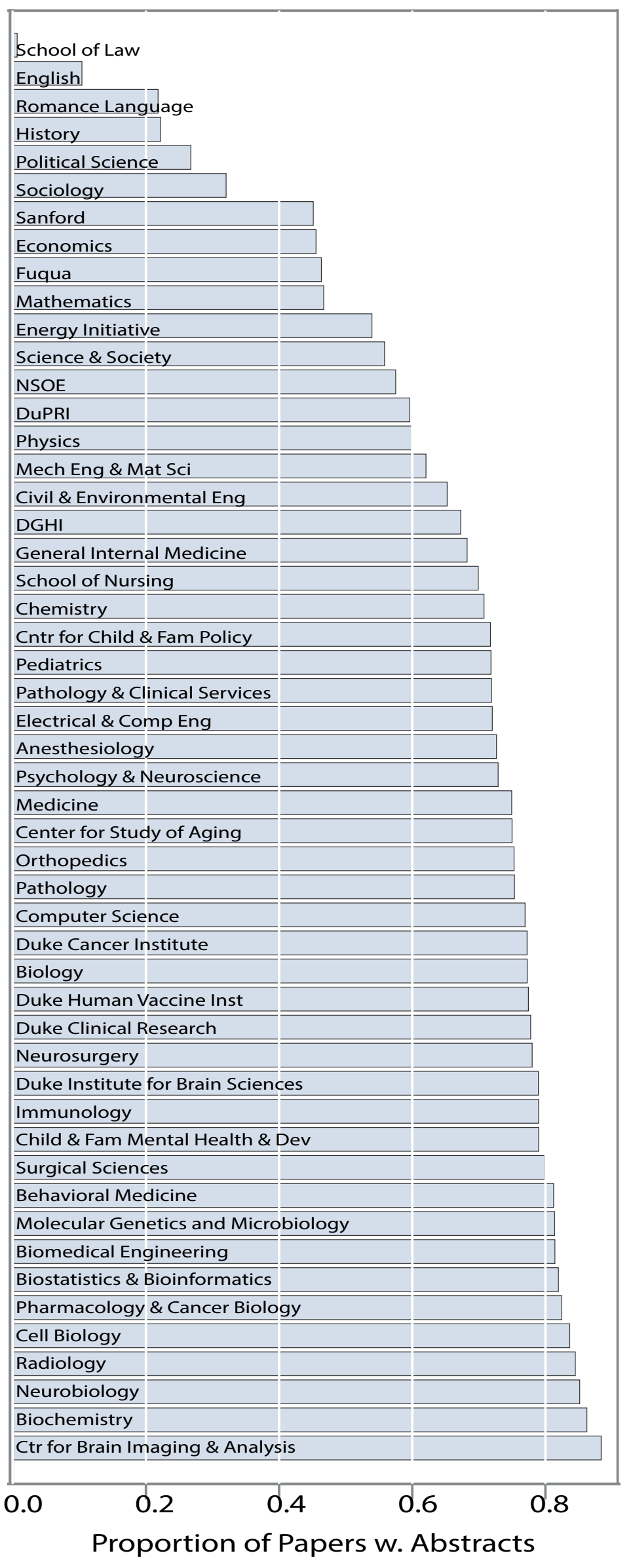


Who Collaborates? Modern scholarly production is team based, particularly in health and natural sciences (Wuchty et al 2007). We constructed a collaboration network from the publications file then clustered that using a modularity maximization routine (Mucha, ongoing) and layer the clusters over the space. Forty percent of papers in the corpus have multiple Duke authors an the median author has 8 Duke collaborators. By definition, most collaboration falls within these clusters, but there is significant cross-cluster interaction. Here we highlight the most common collaborations across clusters.



Limitations. There are at least three limitations to this work. First, S@D coverage is uneven, under representing social science and humanities, and their relatively small footprint reflects this coverage. Second, any two dimensional representation of an intellectual landscape is limited by the fit of the papers to the space, and there may be better projections. Finally, we took a naïve approach to the content of abstracts. While our routine parses for entities, common stop-words and parts of speech, one could improve on the similarity scoring by expanding a dictionary of technical synonyms, author-provided keywords or, cited references.

Scholars @ Duke Coverage



Technical Tidbits. The original publications file includes 77306 observations, which drops to 54133 when we exclude theses/dissertations and limit to papers with abstracts. This represents 29424 unique papers. Authors were assigned to departments based on the primary affiliation listed in the scholars file, authors with multiple primary affiliations were assigned alphabetically to the first one. We used the SAS HPMiner tool to parse papers and score similarity. We included common noun phrases and entities, excluded numbers, punctuation, pronouns, conjunctions, prepositions, articles and a common English stoplist. We stemmed terms to common parents and measured similarity as the cosine of the term-frequency, inverse document frequency (tf-idf) weighted term vectors for each pair of papers, retaining the strongest 1% in the initial network construction. We used a Fruchterman Reingold layout algorithm (implemented in Pajek)to define the topic space, applied to the largest connected component (29408/29424) of the network. The surface contour is smoothed using a 2d kernel density estimate. We used the Louvain method to cluster the paper network and label clusters with the most common terms appearing in the titles of papers. We constructed the collaboration network from the set of 29424 papers (i.e. not the collaboration set posted) and we identified collaboration clusters using a variant of the Louvain method optimizing over multiple resolution parameters. We produced the figures initially in SAS then edited them in Illustrator and compiled the poster in InDesign.

References

Berlin, Isaiah. 1953. *The Hedgehog and the Fox*. London: Weidenfeld & Nicolson.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008.

Cordova, France A. 2016. "Dear Colleague Letter: NSF INCLUDES (Inclusion across the Nation of Communities of Learners of Underrepresented Discoverers in Engineering and Science)." https://www.nsf.gov/pubs/2016/nsf16048/nsf16048.pdf

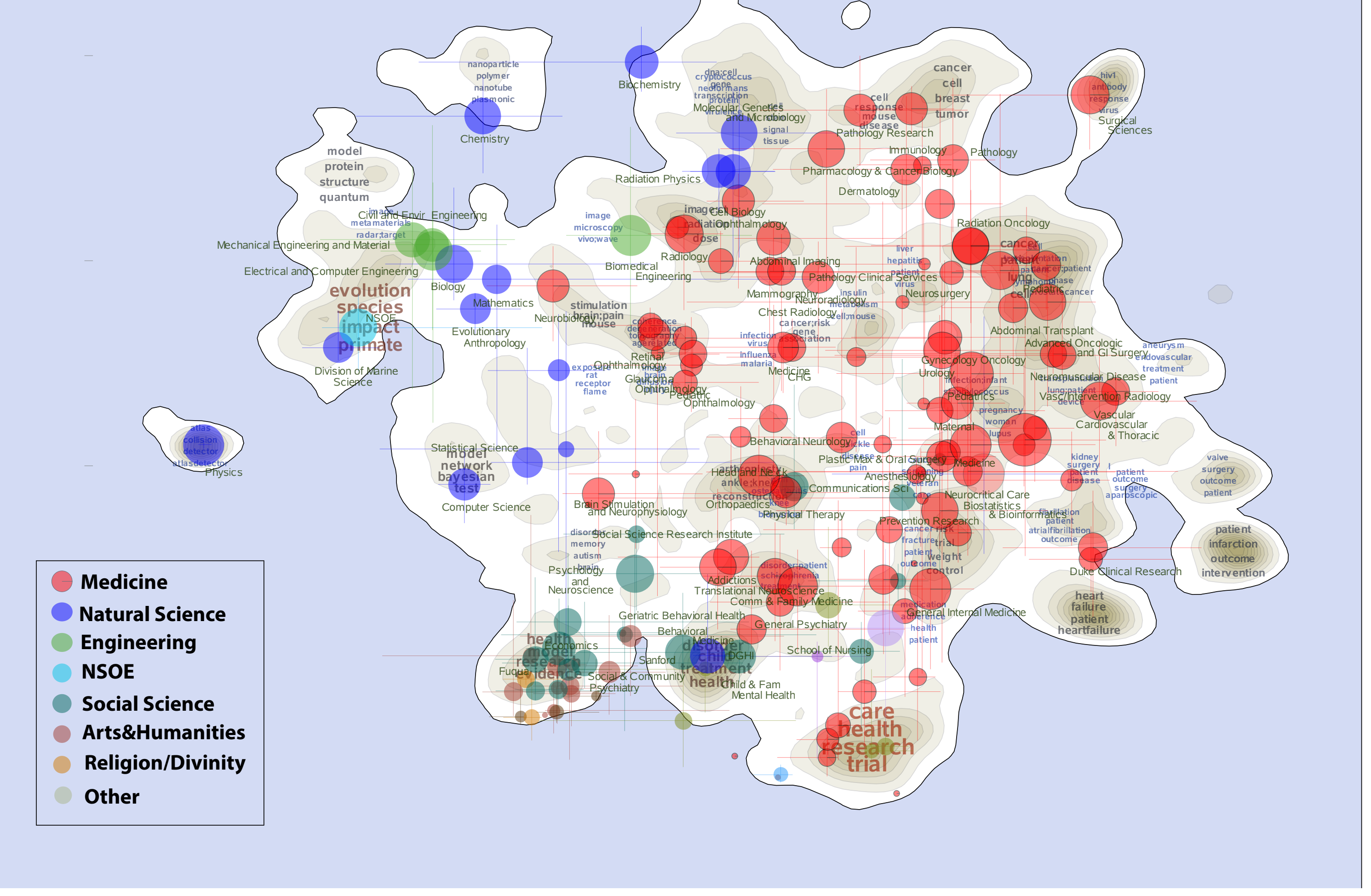
Lucas G. S. Jeub, Marya Bazzi, Inderjit S. Jutla, and Peter J. Mucha, "A generalized Louvain method for community detection implemented in MATLAB" http://netwiki.amath.unc.edu/GenLouvain/GenLouvain (2011-2016).

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall

Wuchty,S, Benjamin F. Jones and Brian Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge" *Science* 316:1036-39.

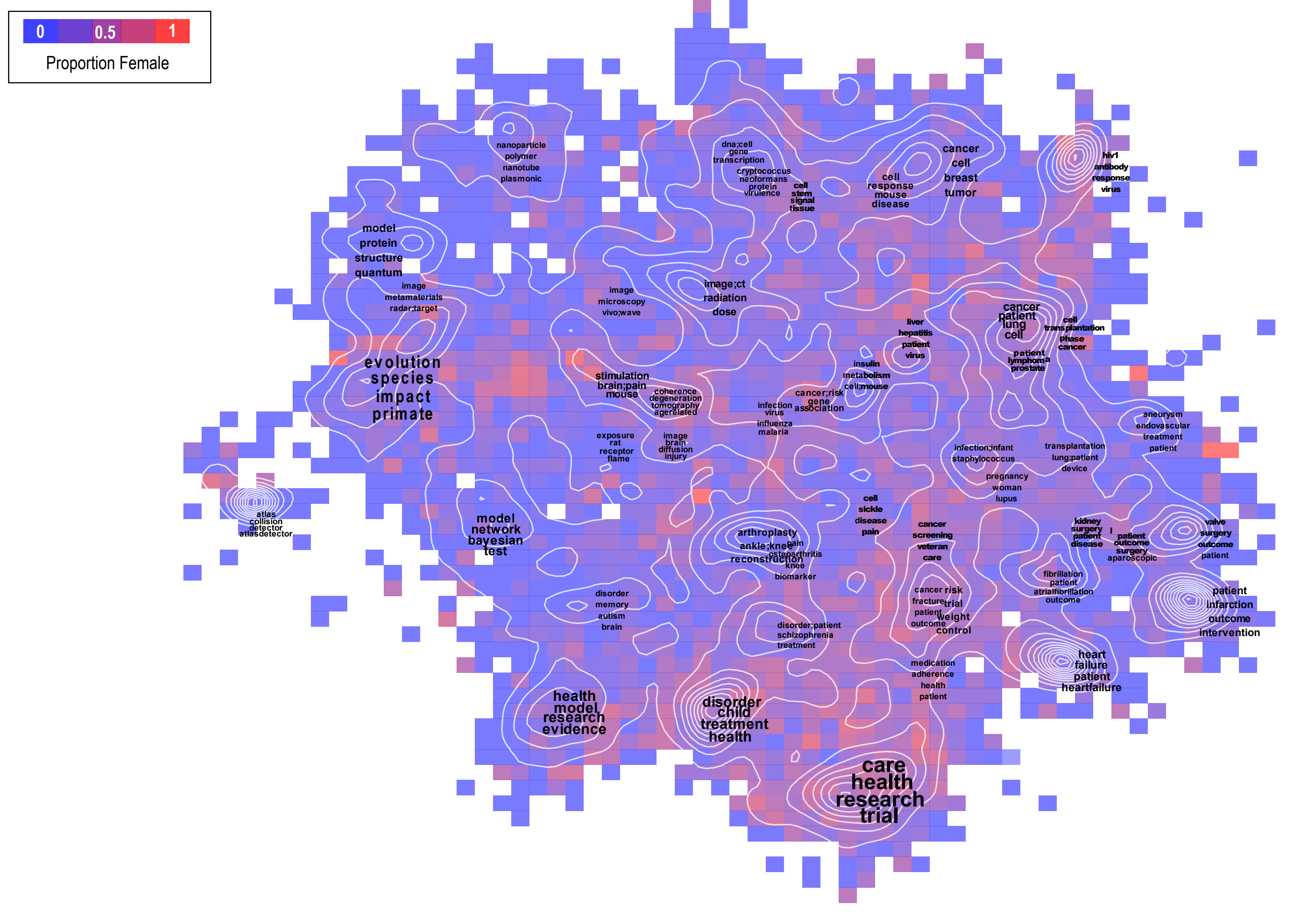
Departments/Units

by division



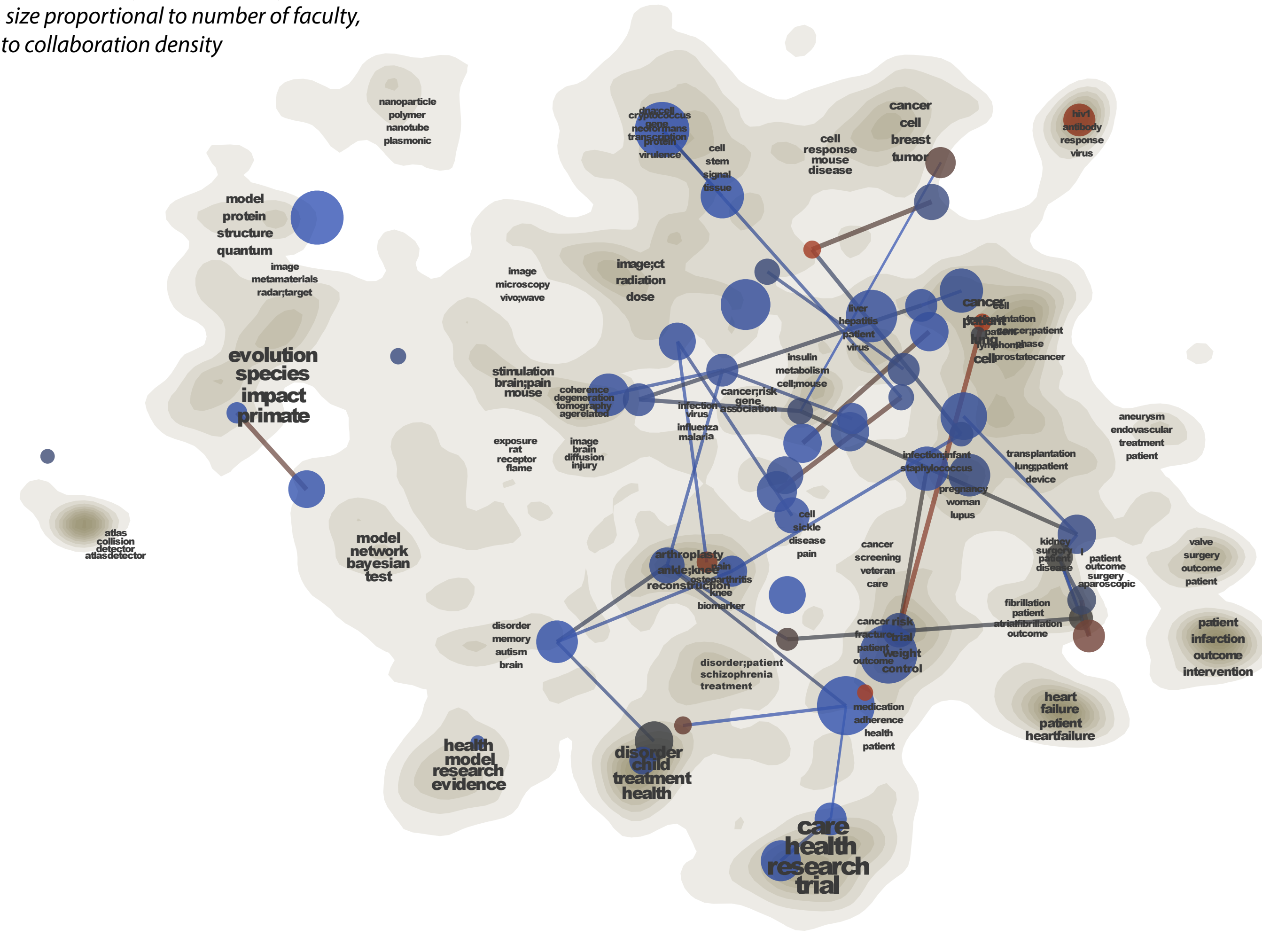
http://www.soc.duke.edu/~jmoody77/S@D/starfield_dept.html

Gender Distribution



Collaboration Clusters

Nodes are groups of faculty who frequently collaborate, size proportional to number of faculty, color scales to collaboration density



http://www.soc.duke.edu/~jmoody77/S@D/contour_collabclusts.html