Towards an Intellectual Atlas of Scholars@Duke

Goal
Our goal is to map scholarly production at Duke and use the distribution of scholars as a frame for displaying relations among 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 closer to each other in the display space, effectively grouping similarly sized papers together.

Since the network is too large to display in standard point and line format, we extract 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—highs represent clusters of papers on narrow topics, ridge-like topography connects to each other, valleys represent gaps in the knowledge space, and widely differing topics are islands disconnected from the mainland. We identify topographic features by clustering the network, and label the clusters with terms found most frequently in each cluster's titles (at least one word proportional to number of papers).

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

Fuses and Hedging. In the main figure, we have selected authors who are ranked in the top 15% by publication volume in their primary affiliation unit and placed them at the centroid of their publication clusters. Berge (1955) famously contrasted thinkers working in a single area (fused thinking) to those who draw widely from multiple intellectual sources (hedgehogs). 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 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 universities. Laying 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 (Corbett, 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).

Who Collaborates? Modern scholarly production is team based, particularly in the health and natural sciences (Manghi et al. 2017). We constructed a collaboration network from the publications. We then clustered using a modularity maximization routine (Blucho, ongoing) and label the clusters over the space. Forty percent of papers in the corpus have multiple Duke authors on the median author has 8 Duke collaborators. By definition, most collaborations 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, 15% coverage is uneven, under representing social science and humanities, and their relatively small footprint reflects the coverage. Second, any two-dimensional representation of an intellectual landscape is limited by the fit of the data 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.

Technical Tables: The original publication for includes 37798 abstracts, which map to 13132 unique authors and 3260 unique abstracts. The 1997-2012 corpus includes 97404 unique papers, authors were assigned to departments by aligning the NCSU-UCI match up with EDUCAUSE and the UC Berkeley match up with EDUCAUSE. The correspondence was then used to assign faculty and students to the appropriate department. Because the match was between US institutions, we were able to make use of the entire network of the abstracts to label the clusters. We used a number of edge-based algorithms (e.g., modularity) to assign authors to the top clusters, applying to the largest connected component individually to determine the clusters. The authors were then assigned to the cluster with the most common term in their abstracts that were more than 15% of the total number of terms.

Collaboration Clusters

Gender Distribution

Collaboration Clusters

Figure provided by the Duke Network Analysis Center, Duke University.