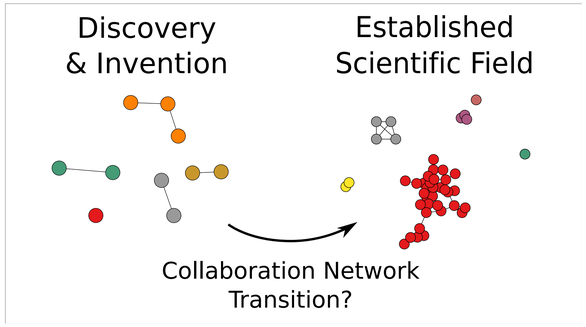
Measuring Topological Transitions in Scientific Collaboration Networks Using Topic Modeling for Subfield Detection

Daniel T. Citron[1], Samuel F. Way[2], Laurence Brandenberger[3]

[1] *Laboratory of Atomic and Solid State Physics, Cornell University* [2] *Department of Computer Science, University of Colorado* [3] *Institute of Political Science, University of Bern & Eawag*

Scientific Collaborations

How do scientific fields develop? How does a network of researchers coalesce around a scientific topic? How does a network of scientific collaborators assemble over time? **Do we observe social restructuring among researchers as their field develops?**



A global topological transition: Previous studies have observed that fields begin as disparate, disconnected groups. Over time, enough collaborators join the field such that it forms a single, densely-connected giant component.

Topic modeling: Previous studies were restricted to a small survey of fields due to reliance on human experts to curate their data sets. Our contribution is to bypass this limitation using document classifying algorithms on a large scientific corpus.

- Access to document corpora enables large-scale analyses
- Topic modeling enables rapid classification of increased number of subfields

ArXiv Data Set

The arXiv - a freely available online repository of scientific preprints, mostly related to Physics, Computer Science, and Mathematics. We focus on condensed matter physics ("**cond-mat**") articles:

- 189,000 articles total
- 680,000 unique authors
- Taken from April 1992 through June 2015
- Use titles, abstracts of articles, author names
- Month & year of submission encoded in arXiv ID

Contact: dtc65@cornell.edu

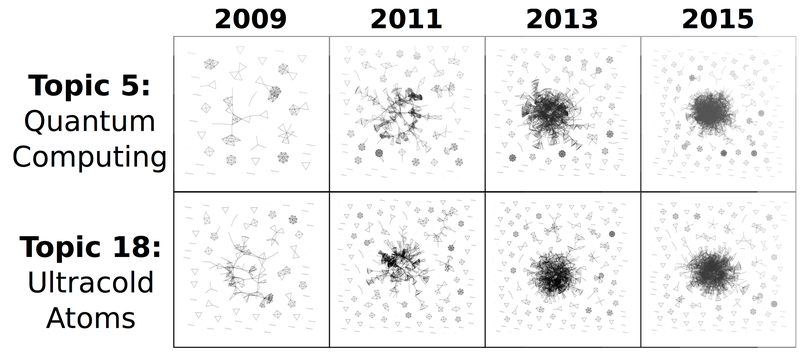
Topic Modeling

We implement **Latent Dirichlet Allocation** to classify the article titles and abstracts from **cond-mat**. LDA classifies documents by characterizing thematic content underlying the textual content. Given word co-occurrence in the document set, LDA returns probability distributions of words across topics and topics across documents. We use N**= 50 topics** and find that 45/50 are readily interpretable as scientific subfields.

Example: Interpreting Topic 5

- Keywords: *quantum state qubit coupling measurement qubits entanglement cavity coupled decoherence*
- Example paper title: "Controllable coupling between flux qubits"
- Interpretation: **Quantum Computing**

Network Assembly



We find groups of articles strongly associated ($p > 0.6$) with each topic, construct the corresponding co-authorship network, and observe how it changes over time. For topics that represent scientific subfields, we **consistently observe the formation of a dense giant cluster for 28/45 of the scientific subfields identified with LDA**.

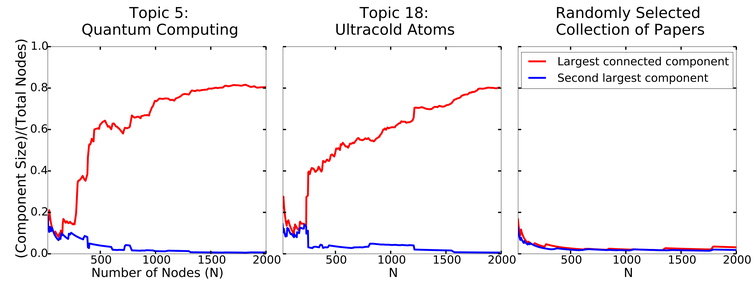
Citations and Acknowledgements

1. L. M. A. Bettencourt, *et al.* Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics* (2009).
2. D. M. Blei, *et al.* Latent dirichlet allocation. *The Journal of Machine Learning Research* (2003).

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144153. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

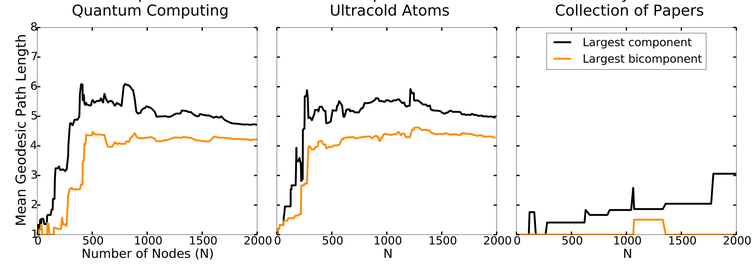
Network Assembly Measurements

Giant Component Formation



- Largest connected component dominates network
- **Global topological transition**

Mean Geodesic Paths



- Initial growth, rapid consolidation
- Establishment of **long ties**
- Distant research groups overlap at transition

Discussion

For many topics in condensed matter physics, we observe the same pattern: a **global topological transition** to a densely-connected community of collaborators following a period of scatterred initial growth. The formation of the giant cluster represents **social reorganization** as a subfield develops over time.

Future work:

- Microscopic model for network assembly to explain quantitative features, test hypotheses of contagion and cooperation
- Homophily - can we model how an author's past research activities influence future collaborations?