Measuring Breakthrough Discovery and Invention with Algebraic Topology

Thomas Gebhart[1] and Russell Funk[2]

[1] Department of Computer Science, University of Minnesota

[2] Carlson School of Management, University of Minnesota

Motivation

Measuring progress in science and technology

- The post-war era has witnessed unprecedented advances in science and technology.
- But measuring the pace and extent of this progress is an open problem.
- Some observers suggest the pace of progress is slowing (Bloom et al., 2020; Gordon, 2017).
- We offer several explanations for why, which align with previous hypotheses, including
  - the most important/easiest discoveries have already been made (Cowen, 2011), and
  - the (re)-combinatorial growth of knowledge is a burden on scientists and technologists (Jones, 2009).
- From these observations, we propose a new framework for measuring innovation which
  - ties innovation to structural gaps in the landscape of scientific and technological knowledge and
  - contextualizes innovation according to the combinatorial complexity of its constituent knowledge.

Motivation

Measuring progress in science and technology

- Inspired by theories of recombination, many prior works view knowledge as a network.
  - Nodes are knowledge concepts, edges some relationship between these concepts (Ju et al., 2020).
  - New insights are generated by bringing together previously-disconnected concepts.
- While insightful, past work on the network structure of knowledge and innovation is limited.
  - Focus on low-level, dyadic relationships among knowledge lacks expressivity.
  - Measurement is typically one-dimensional, masking broader contours of the evolving structure of knowledge.
- Algebraic topology provides a natural framework for measuring progress in science and technology.
  - Simplicial complexes generalize networks, representing recombination in a formal and intuitive manner.
  - Persistent homology allows for characterization of the global, multi-dimensional structure of knowledge.
  - Formalizing the decomposition of science and technological networks provides additional insights into their structure.

Persistent Homology

A brief overview

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 46, Number 2, April 2009, Pages 255–308
S 0273-0979(09)01249-X
Article electronically published on January 29, 2009

TOPOLOGY AND DATA

GUNNAR CARLSSON

1. Introduction

An important feature of modern science and engineering is that data of various kinds is being produced at an unprecedented rate. This is so in part because of new experimental methods, and in part because of the increase in the availability of high powered computing technology. It is also clear that the *nature* of the data we are obtaining is significantly different. For example, it is now often the case that we are given data in the form of very long vectors, where all but a few of the coordinates turn out to be irrelevant to the questions of interest, and further that we don't necessarily know which coordinates are the interesting ones. A related fact is that the data is often very high-dimensional, which severely restricts our ability to visualize it. The data obtained is also often much noisier than in the past and has more missing information (missing data). This is particularly so in the case of biological data, particularly high throughput data from microarray or other sources. Our ability to analyze this data, both in terms of quantity and the nature of the data, is clearly not keeping pace with the data being produced. In this

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 45, Number 1, January 2008, Pages 61–75
S 0273-0979(07)01191-3
Article electronically published on October 26, 2007

BARCODES: THE PERSISTENT TOPOLOGY OF DATA

ROBERT GHRIST

Abstract. This article surveys recent work of Carlsson and collaborators on applications of computational algebraic topology to problems of feature detection and shape recognition in high-dimensional data. The primary mathematical tool considered is a homology theory for point-cloud data sets—**persistent homology**—and a novel representation of this algebraic characterization—**barcodes**. We sketch an application of these techniques to the classification of natural images.

1. The shape of data

When a topologist is asked, "How do you visualize a four-dimensional object?" the appropriate response is a Socratic rejoinder: "How do you visualize a three-dimensional object?" We do not see in three spatial dimensions directly, but rather

Aktas *et al. Applied Network Science* (2019) 4:61
https://doi.org/10.1007/s41109-019-0179-3

Applied Network Science

REVIEW Open Access

Persistence homology of networks: methods and applications

Mehmet E. Aktas[1*], Esra Akbas[2] and Ahmed El Fatmaoui[1]

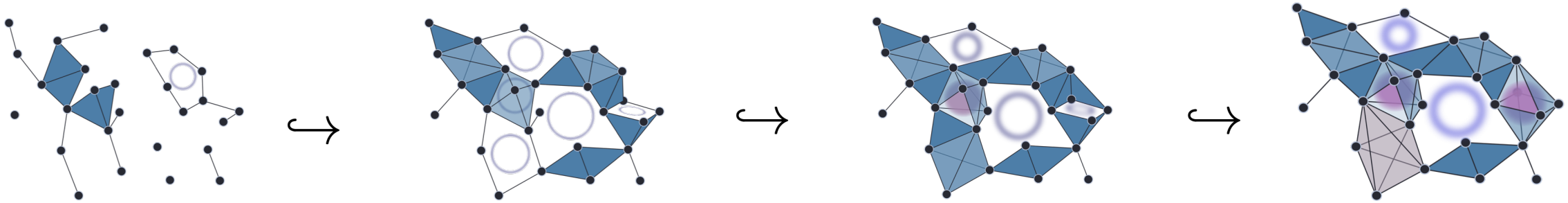
*Correspondence: maktas@uco.edu
[1]Department of Mathematics and Statistics, University of Central Oklahoma, Edmond, OK, USA
Full list of author information is available at the end of the article

Abstract
Information networks are becoming increasingly popular to capture complex relationships across various disciplines, such as social networks, citation networks, and biological networks. The primary challenge in this domain is measuring similarity or distance between networks based on topology. However, classical graph-theoretic measures are usually local and mainly based on differences between either node or edge measurements or correlations without considering the topology of networks such as the connected components or holes. In recent years, mathematical tools and deep learning based methods have become popular to extract the topological features of networks. Persistent homology (PH) is a mathematical tool in computational topology that measures the topological features of data that persist across multiple scales with applications ranging from biological networks to social networks.
In this paper, we provide a conceptual review of key advancements in this area of using PH on complex network science. We give a brief mathematical background on PH, review different methods (i.e. filtrations) to define PH on networks and highlight different algorithms and applications where PH is used in solving network mining problems. In doing so, we develop a unified framework to describe these recent approaches and emphasize major conceptual distinctions. We conclude with directions for future work. We focus our review on recent approaches that get significant attention in the mathematics and data mining communities working on network data. We believe our summary of the analysis of PH on networks will provide

Persistent Homology

A brief overview

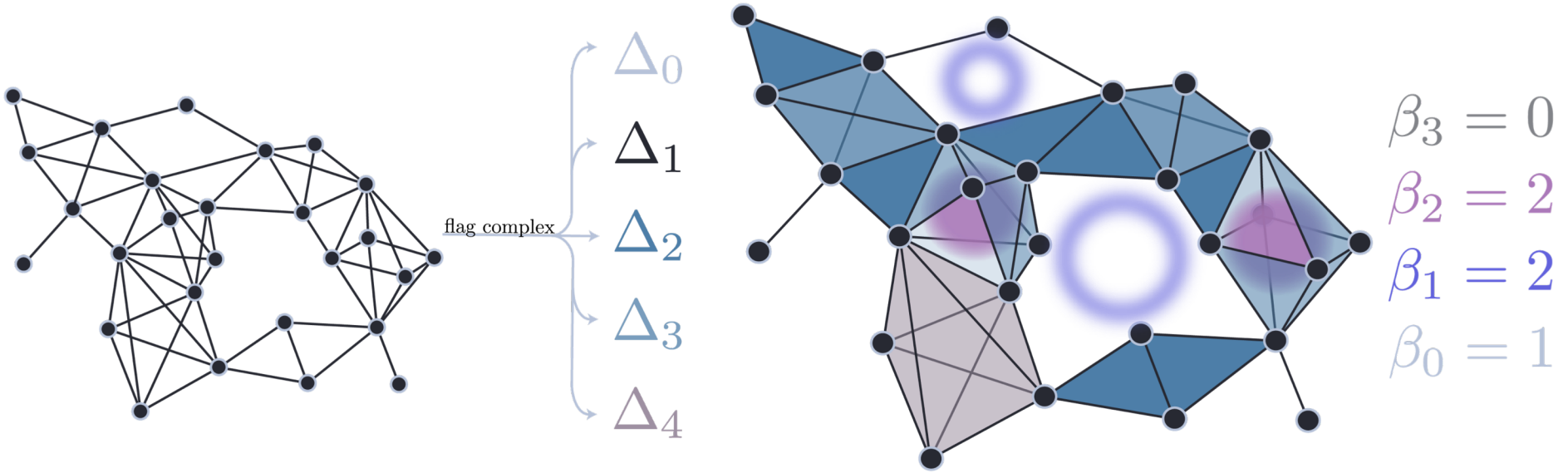
- The *homology* of a topological space is a set of invariants describing the structure of the space.
- PH tracks the birth and death of homological "holes" across a filtration of a simplicial complex.
- A *simplicial complex* K is a combinatorial set of *simplices* which themselves generalize triangles.
- A *filtration* of a simplicial complex is a nested family of simplicial complexes $\{K_i \mid i \in \mathbb{Z}\}$ related by inclusion such that $K_i \subseteq K_j$ for $i < j$.
- Persistent homology computes a common basis with which to track the homological features which emerge and disappear throughout the filtration.



Networks as Simplicial Complexes

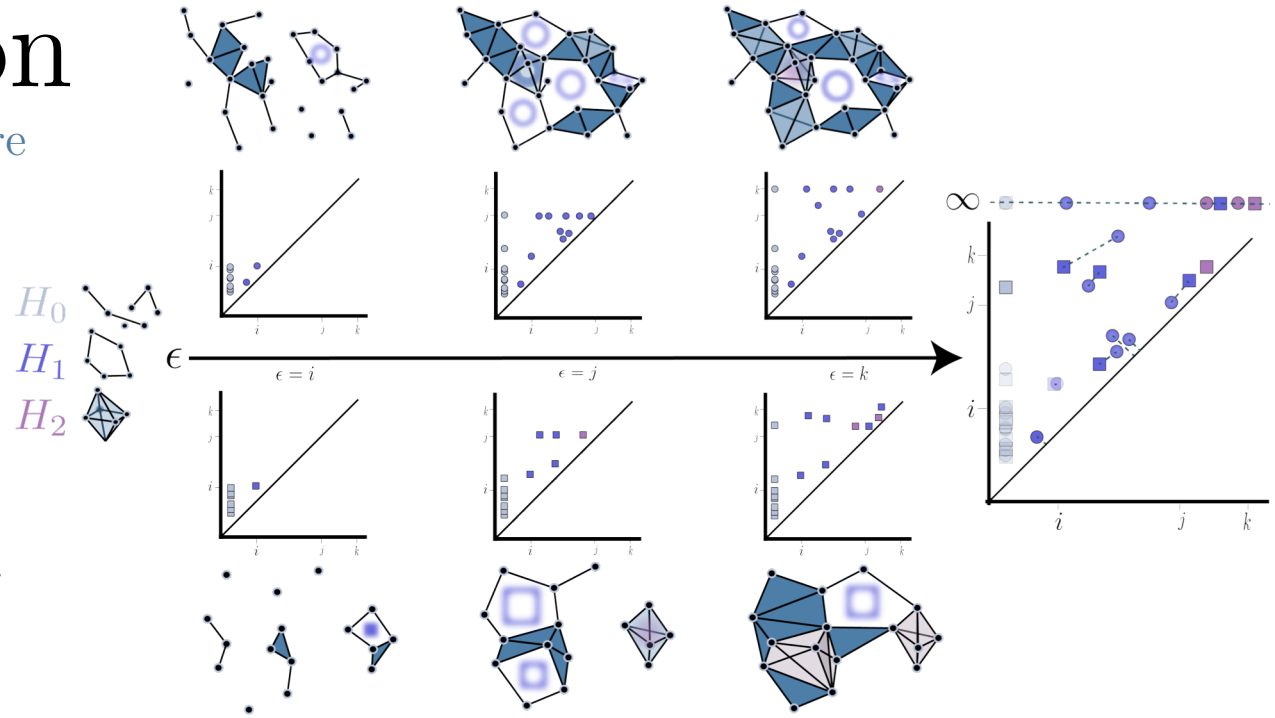
A brief overview

- We can encode networks as simplicial complexes through the *flag* or *clique complex*.
  - "Fill in" each k-clique of nodes and represent this clique as a (k-1)-dimensional simplex Δ_{k-1} .
  - The homology (Betti numbers β_{k-1}) of the resulting simplicial complex is an invariant of the underlying network.
- If the network is weighted, edge weights provide a natural filtration to decompose the network as subnetworks induced varying a threshold on the weights.



Benefits of Abstraction

Generalizing networks and characterizing structure



- Encoding cliques of concepts as simplices allows us to track higher-order knowledge from atomic concept to knowledge structure.
  - Structure in higher dimensions implies higher complexity.
- Betti numbers, simplex counts, and measures like the Wasserstein distance allow for comparison of scientific and technological network structures across time or field.
- Gaps in the structure of scientific and technological knowledge are formally defined.
  - Characterized by their topological representatives or their birth simplices and death simplices.

Analyses

The Emergence of Higher-order Structure

Topological Measures of Innovation

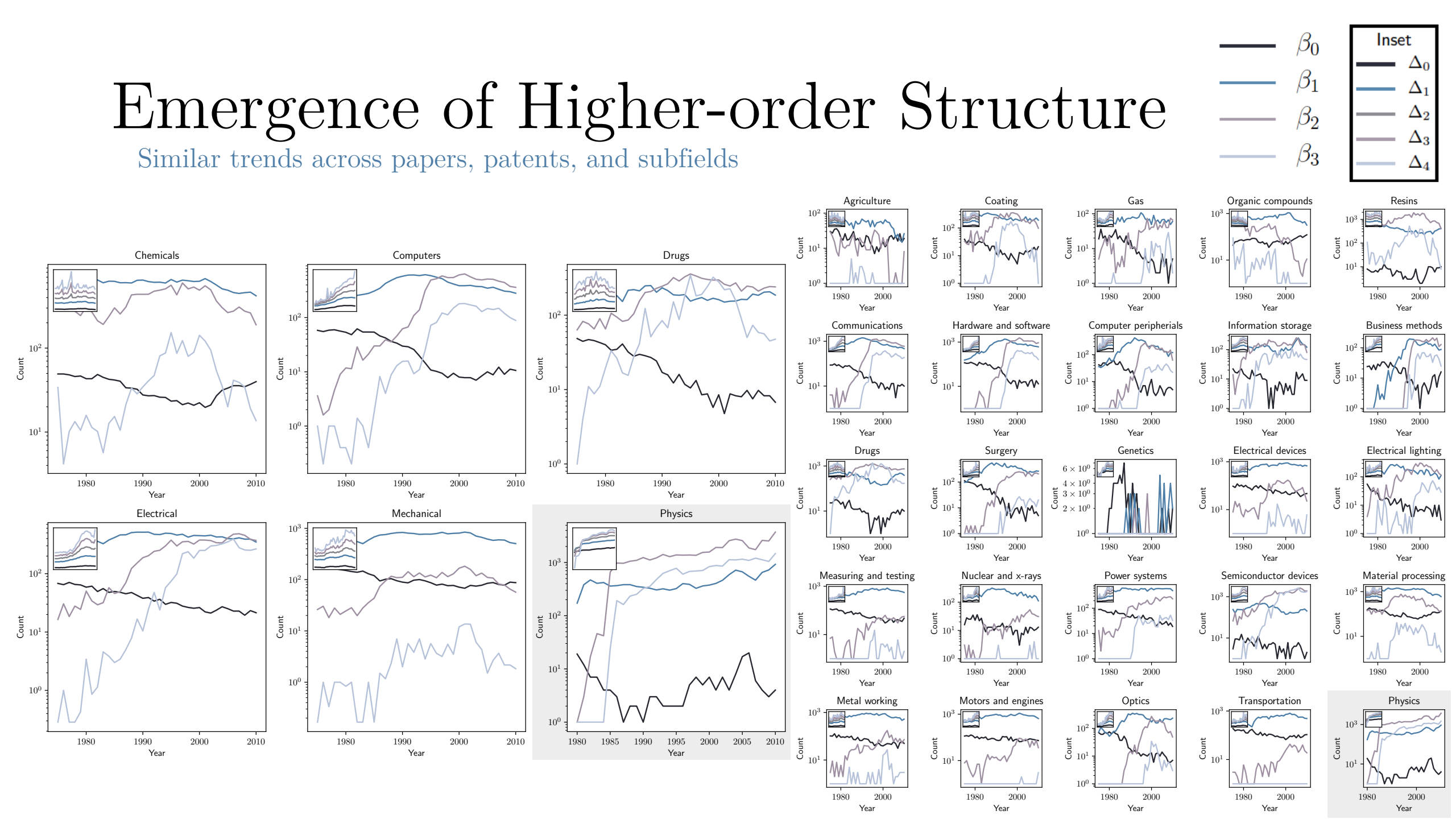
Emergence of Higher-order Structure

Data

- 630,000 American Physical Society (APS) papers between 1893 and 2018.
  - Nodes correspond to PACS codes (e.g., 04.30.-w “Gravitational waves”).
  - Edges indicate co-occurrence of two codes on the same paper.
  - **Edge weights correspond to the (inverse) count of co-occurrences.**
- 6.5 million utility patents granted between 1976 and 2017.
  - Patents assigned subfield according to their NBER technology categories.
  - Nodes correspond to USPC codes (e.g. 712/10+ “Array processors”).
  - Edges indicate co-occurrence of two codes on the same patent.
  - **Edge weights correspond to the (inverse) count of co-occurrences.**
- Also compute 3-year windowed collaboration networks for both datasets.
- We create these networks for each year and subfield.

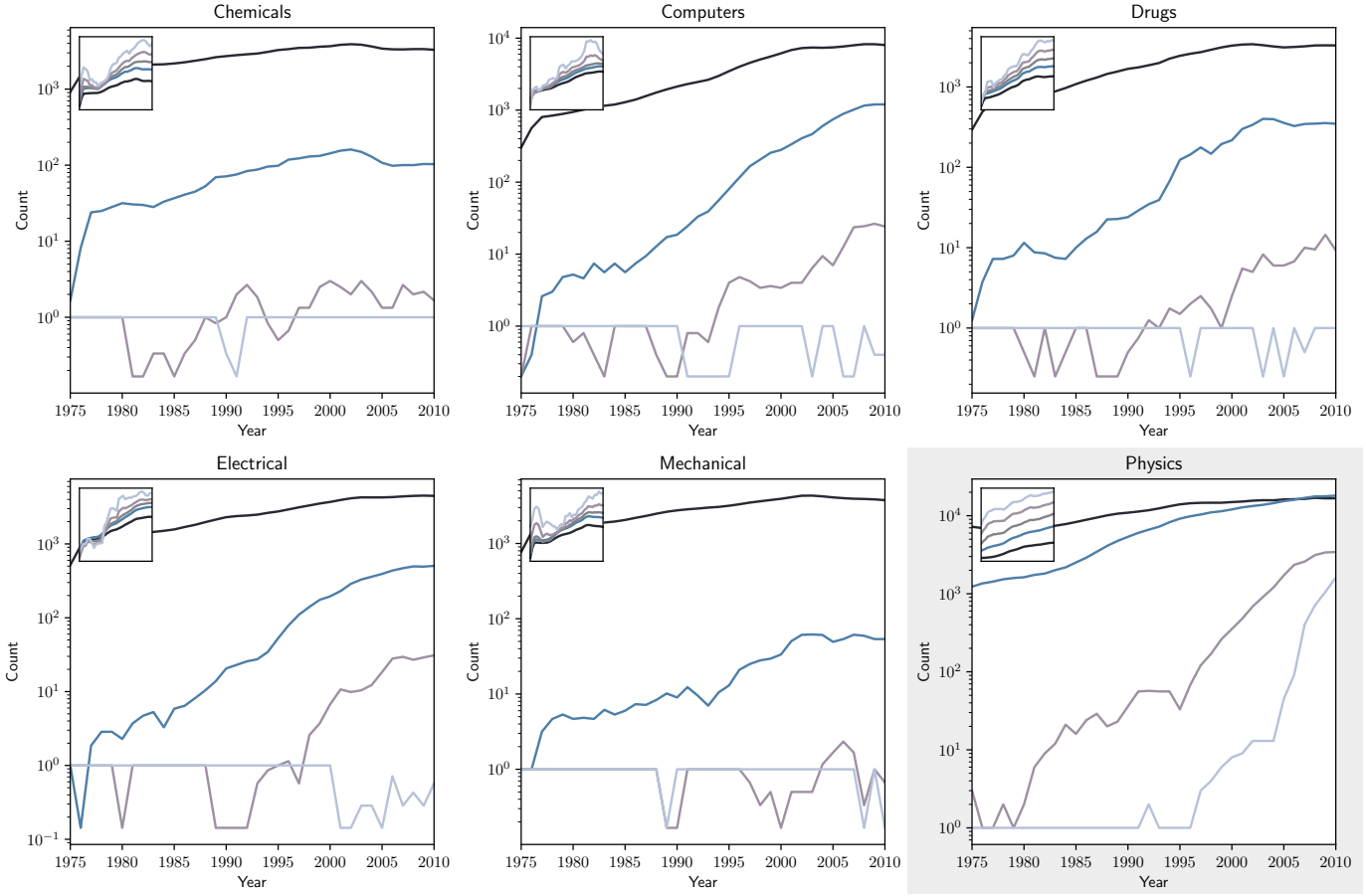
Emergence of Higher-order Structure

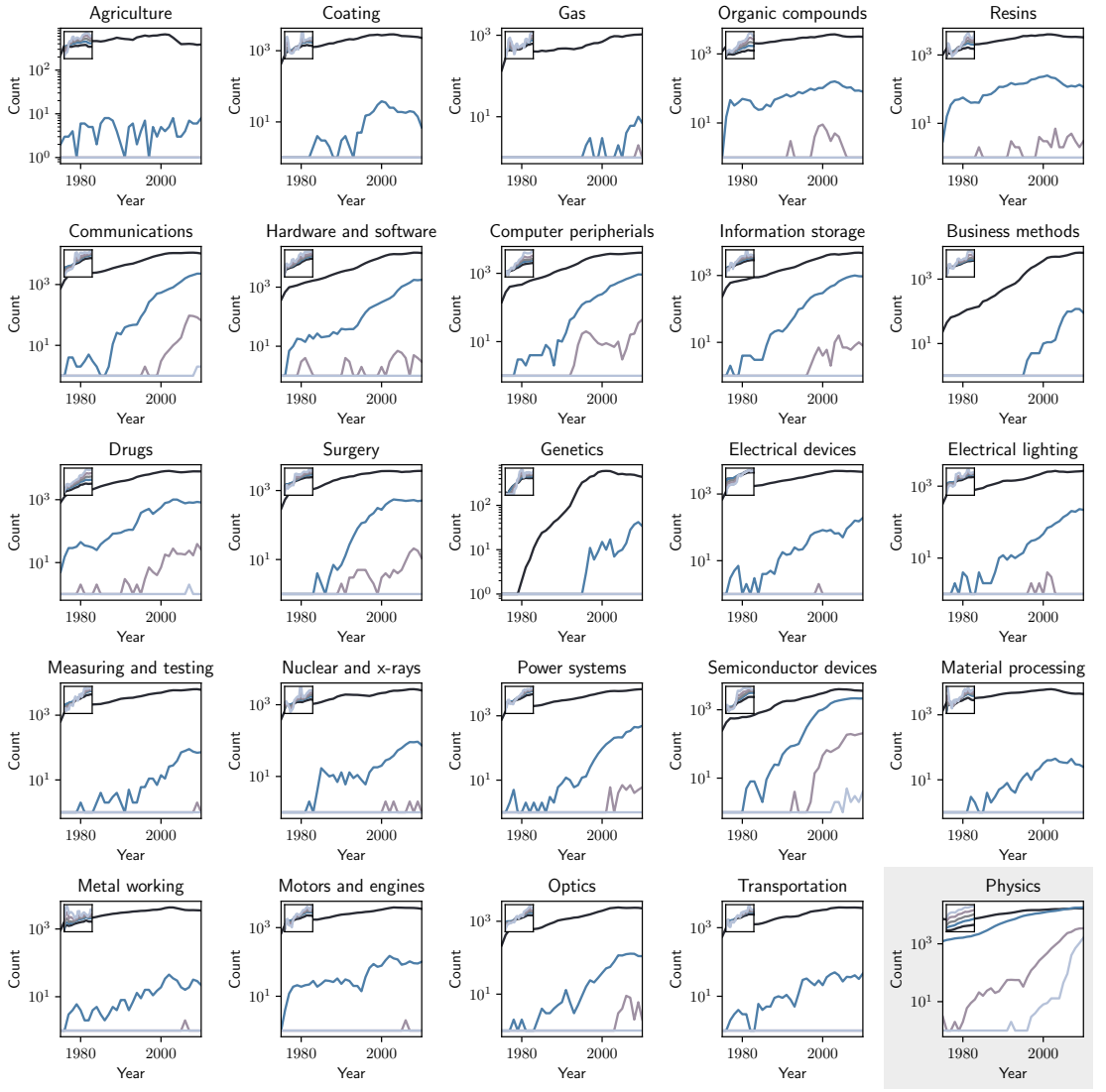
Similar trends across papers, patents, and subfields

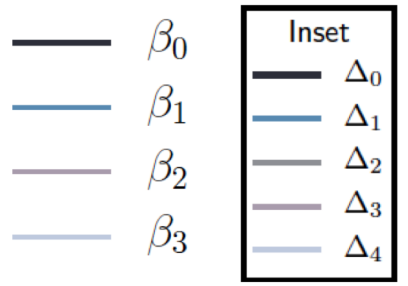


Emergence of Higher-order Structure

Collaboration networks show much less complex growth

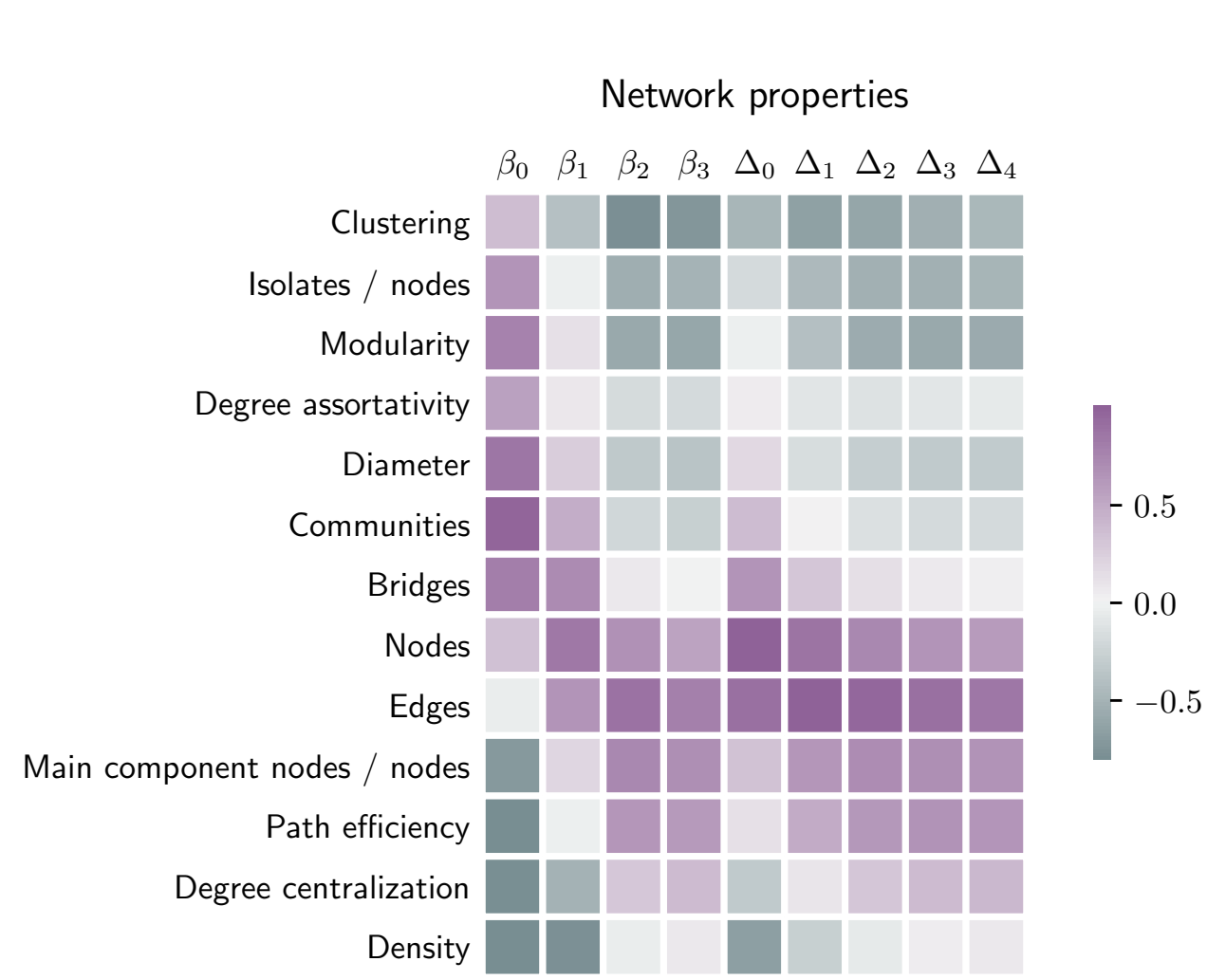


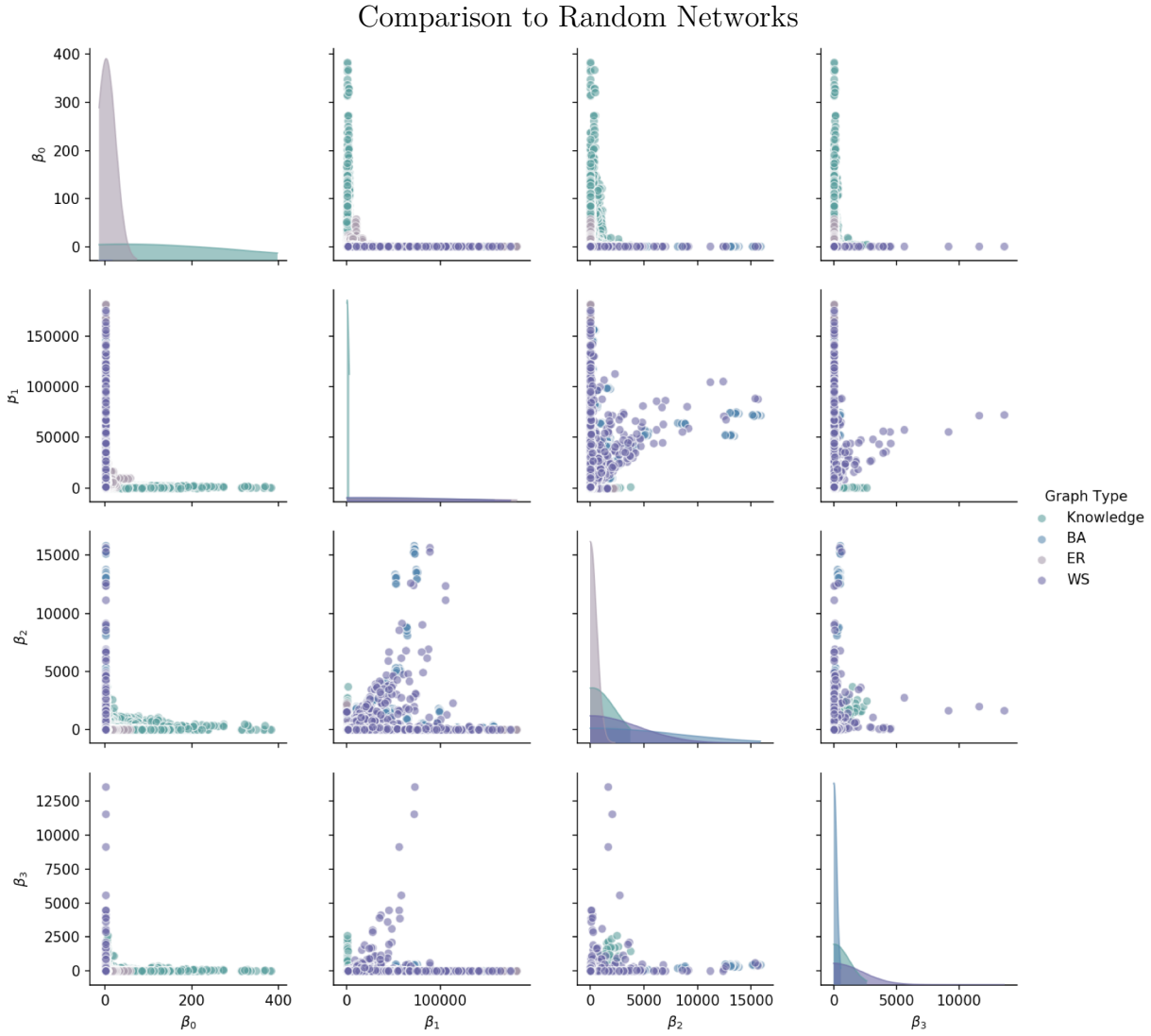




Emergence of Higher-order Structure

Structure is not explainable by low-order measures or as random processes





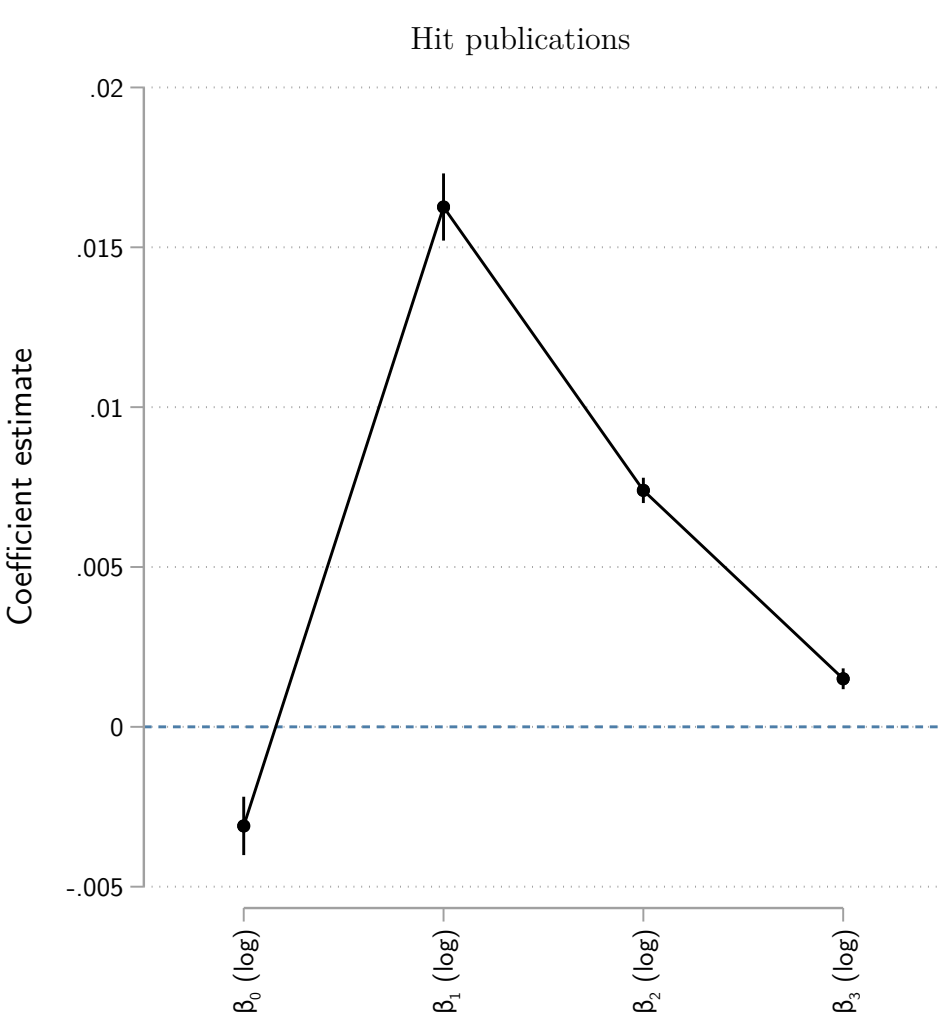
Mapping Structure to Innovation

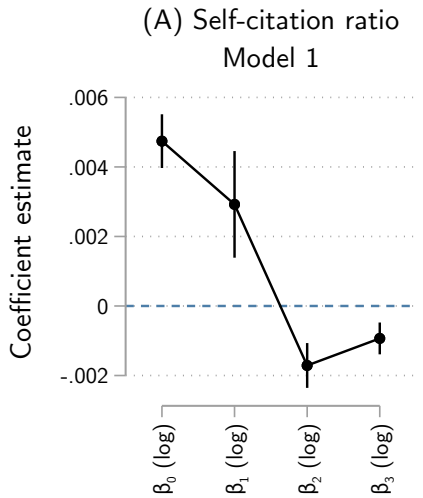
Regression on some proxies for innovation

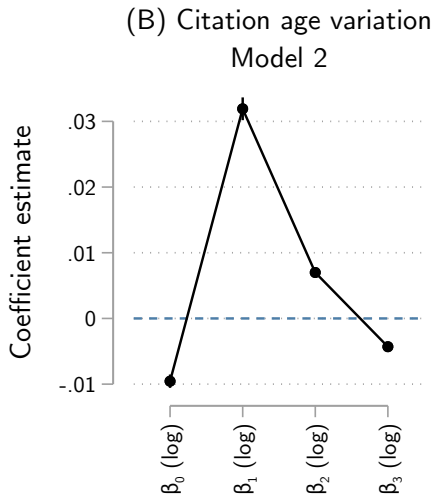
- Data
  - APS, $N = 390{,}823$ papers between 1980 and 2010
  - Patents, $N = 3{,}855{,}730$ patents granted between 1976 and 2010
- Outcomes
  - "Hit" publication/patent
  - Search depth
  - Lexical novelty
  - Lexical complexity
- Predictors
  - Betti numbers
  - Simplex counts
- Specifications
  - OLS regression with fixed effects for year and field
  - Robust standard errors

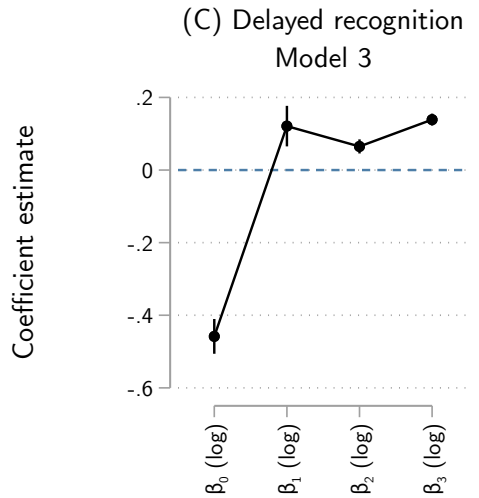
Mapping Structure to Innovation

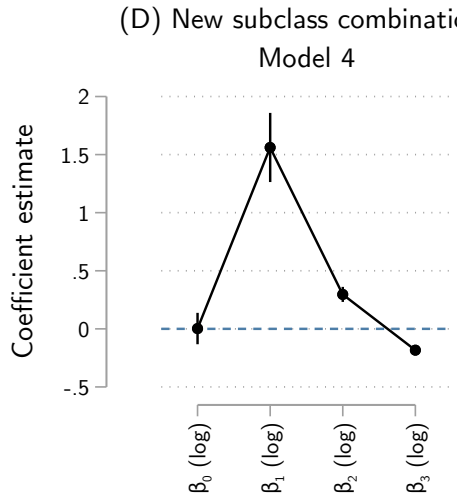
Regression on some proxies for innovation

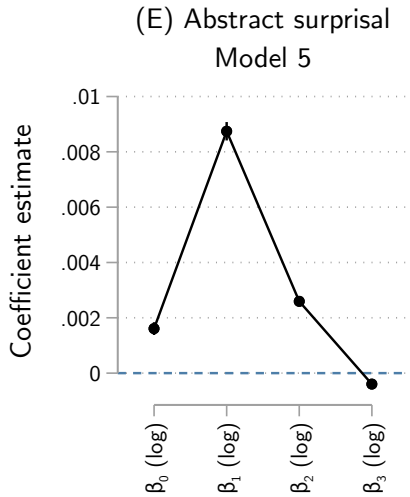


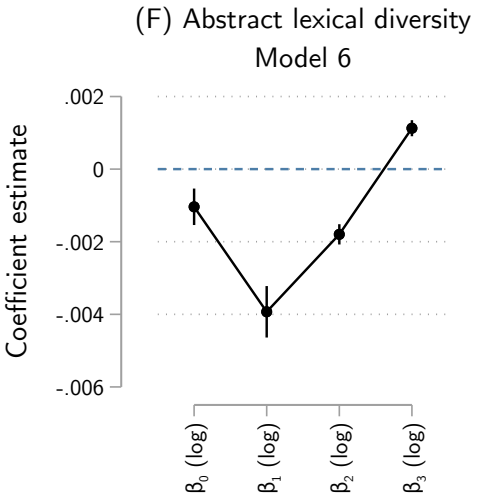












Emergence of Higher-order Structure

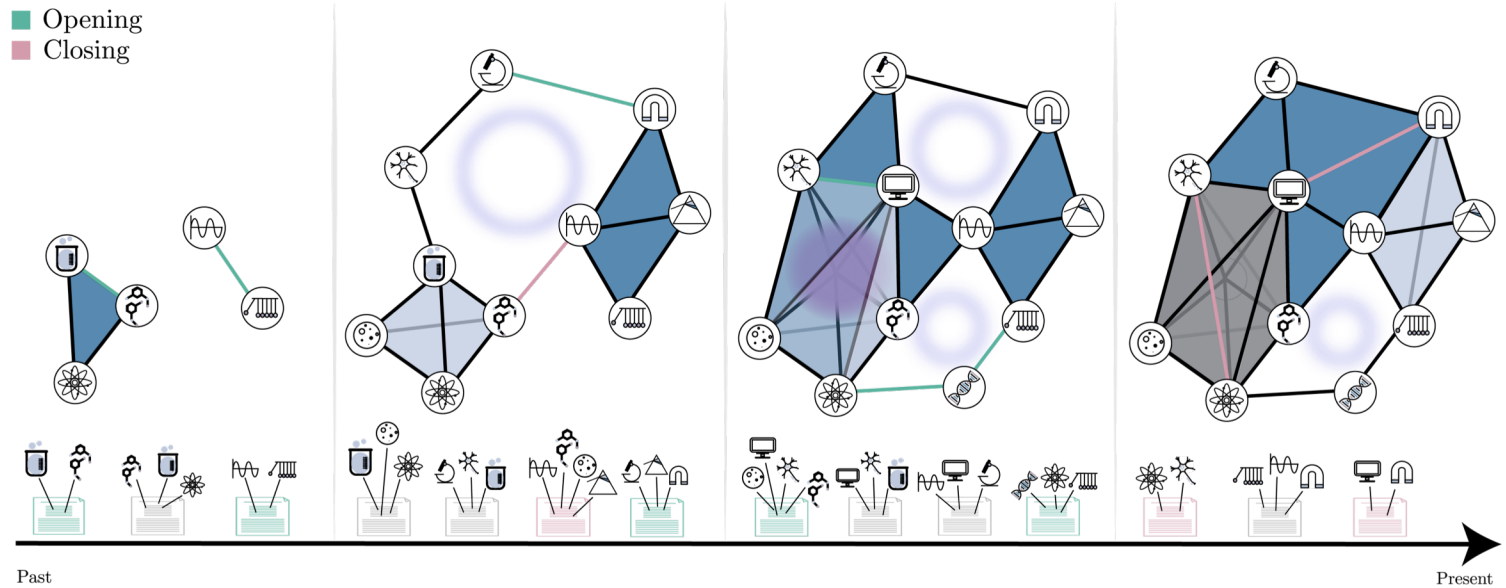
Summary of observations

- Recent shifts towards higher-order structure in scientific and technological knowledge networks.
  - These shifts are robust across datasets and appear in many subfields.
- Until very recently, the shift to higher-order structure is missing in the collaboration networks that produce this knowledge.
- This knowledge network structure is distinct from a number of popular random network models.
- Scientific and technological innovation may be difficult to measure with traditional measures.
- Innovation appears to be taking place at higher levels of knowledge abstraction.

Topological Measures of Innovation

Extending these to time-varying knowledge networks

- Can we measure the innovative nature of a scientific work via its impact on network topology?
- Create knowledge networks using concepts extracted from abstracts of scientific works.
  - Using millions of physics paper abstracts from APS and social science papers from Web of Science.
  - Each node is a concept (“labor market”).
  - Edges are given by co-occurrence of concepts within the abstracts.
  - **Edge weights correspond to the time of publication.**



Topological Measures of Innovation

Extending to time-varying knowledge networks

Physical sciences

- Key measures of innovation are positively associated to papers which close knowledge gaps.
- Closing recently-created gaps in knowledge is associated with higher future citation rates.

	All papers				Hole-closing papers only			
	(1) Pr("hit")	(2) Pr("hit")	(3) Pr(Nobel Prize)	(4) Pr(Nobel Prize)	(5) Pr("hit")	(6) Pr("hit")	(7) Pr("flop")	(8) Pr("flop")
Closes 1-dimensional hole (1=Yes)	0.0139***	0.0111**	0.0008	0.0008				
	(0.0048)	(0.0048)	(0.0007)	(0.0007)				
Closes 2-dimensional hole (1=Yes)	0.0110***	0.0077***	0.0007***	0.0007***				
	(0.0019)	(0.0019)	(0.0002)	(0.0002)				
Introduces novel concept pair (1=Yes)		0.0097***		-0.0000				
		(0.0008)		(0.0000)				
Lifetime of 1-dimensional hole(s) closed (days, log)					-0.0092**		0.0119**	
					(0.0040)		(0.0060)	
Lifetime of 2-dimensional hole(s) closed (days, log)						-0.0035**		0.0074***
						(0.0016)		(0.0024)
Field fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	373749	373749	373749	373749	2312	16617	2312	16617
Adjusted R2	0.01	0.01	0.00	0.00	0.03	0.01	0.01	0.01
Wald tests for topology predictors								
F	22.13	11.25	3.83	3.81	5.37	4.80	3.97	9.06
d.f.	2.00	2.00	2.00	2.00	1.00	1.00	1.00	1.00
p-value	0.00	0.00	0.02	0.02	0.02	0.03	0.05	0.00

Notes: Estimates are from ordinary-least-squares regressions (linear probability models). Models 1-4 evaluate the probability of a "hit" and Nobel Prize winning paper as a function of whether the paper closes a 1- or 2-dimensional hole or introduces a novel pairing of concepts in the knowledge network of its field. Hit papers are defined (using a 0/1 indicator variable) as those that are cited more than (or equal to) 95 percent of all other papers across fields and years in the first five years post publication. To allow time for the accumulation of citations, we limit the sample to papers published before 2010. Models 5-8 evaluate the probability of a "hit" and "flop" paper as a function of the lifetime of the hole closed. Hole lifetime is only defined for papers that close holes of a given dimension, and therefore we estimate separate models for the lifetime of 1- and 2-dimensional holes (only a small number of papers close holes of both dimensions). Flop papers are defined (using a 0/1 indicator variable) as those that are cited less than (or equal to) 95 percent of all other papers across fields and years in the first five years post publication. For more details on variables, see Table **??**. Wald tests reported below each model evaluate whether the included topological predictors significantly improve model fit. Robust standard errors are shown in parentheses; p-values correspond to two-tailed tests.
*p¡0.1; **p¡0.05; ***p¡0.01

Topological Measures of Innovation

Extending to time-varying knowledge networks

Social sciences

- Key measures of innovation are positively associated to papers which close knowledge gaps.
- Lifetime of closed knowledge gaps is not significantly associated with future citation rate.

	All papers		Hole-closing papers only			
	(1) Pr("hit")	(2) Pr("hit")	(3) Pr("hit")	(4) Pr("hit")	(5) Pr("flop")	(6) Pr("flop")
Closes 1-dimensional hole (1=Yes)	0.0103***	0.0034***				
	(0.0013)	(0.0013)				
Closes 2-dimensional hole (1=Yes)	0.0141***	0.0076***				
	(0.0008)	(0.0009)				
Introduces novel concept pair (1=Yes)		0.0311***				
		(0.0005)				
Lifetime of 1-dimensional hole(s) closed (years)			-0.0002		0.0017	
			(0.0007)		(0.0010)	
Lifetime of 2-dimensional hole(s) closed (years)				-0.0006		-0.0005
				(0.0005)		(0.0006)
Field fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	894449	894449	37905	156906	37905	156906
Adjusted R2	0.02	0.02	0.02	0.02	0.04	0.03
Wald tests for topology predictors						
F	186.19	44.22	0.07	1.09	2.67	0.54
d.f.	2.00	2.00	1.00	1.00	1.00	1.00
p-value	0.00	0.00	0.80	0.30	0.10	0.46

Notes: Estimates are from ordinary-least-squares regressions (linear probability models). Models 1-2 evaluate the probability of a "hit" paper as a function of whether the paper closes a 1- or 2-dimensional hole or introduces a novel pairing of concepts in the knowledge network of its field. Hit papers are defined (using a 0/1 indicator variable) as those that are cited more than (or equal to) 95 percent of all other papers across fields and years in the first five years post publication. To allow time for the accumulation of citations, we limit the sample to papers published before 2010. Models 3-6 evaluate the probability of a "hit" and "flop" paper as a function of the lifetime of the hole closed. Hole lifetime is only defined for papers that close holes of a given dimension, and therefore we estimate separate models for the lifetime of 1- and 2-dimensional holes (only a small number of papers close holes of both dimensions). Flop papers are defined (using a 0/1 indicator variable) as those that are cited less than (or equal to) 95 percent of all other papers across fields and years in the first five years post publication. For more details on variables, see Table **??**. Wald tests reported below each model evaluate whether the included topological predictors significantly improve model fit. Robust standard errors are shown in parentheses; p-values correspond to two-tailed tests.
*p¡0.1; **p¡0.05; ***p¡0.01

Open Questions and Upcoming Work

Measuring the topological characteristics of innovation

- Investigating features of topologically-disruptive works in science and technology
  - Are gap-closing papers associated to a particular knowledge concepts or particular styles of paper?
  - Are there lexical distinctions between the language of gap-closing papers compared to other works?
- Characterizing the topological structures which imply innovation
  - Do any particular topological structures imply the opportunity for innovation?
  - Are particular knowledge concepts more topologically significant than others?
- Measuring innovation at the paper level
  - Can we summarize a paper's innovative potential through a topologically-motivated measure?
  - Is there a measure of innovative potential at the field level based on the underlying concept network?
  - What knowledge gaps exist in the current network structure of science and technology?