The Development of Facticity: from Preliminary Findings to Accepted Implicit Knowledge

Tianyu Du Yuze Sui Jingruo Sun Daniel A. McFarland
{tianyudu, yuzesui, jingruo, mcfarland}@stanford.edu

Stanford University

Stanford GRADUATE SCHOOL OF EDUCATION | Mimir Knowledge Creation Lab at Stanford University

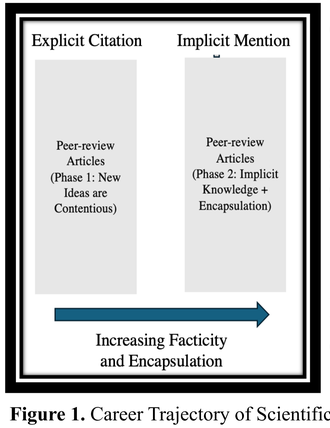
Introduction

Most research in the science of science literature focuses on the citation patterns in published journal articles or conference proceedings to study the **emergence and survival of new scientific ideas and concepts**. Are new concepts and their associations accepted when published? When do new ideas become accepted facts and extend the body of scientific knowledge? In this research, leveraging on NLP methods, we trace the "career trajectories" of scientific ideas and concepts and reveal their transition from **initial knowledge claims and loci of dispute** to more implicitly held and **accepted knowledge**.

Theory

We contend the process of usage and change of scientific ideas and concepts is core to knowledge accumulation and enables increased knowledge capacity and storage in science overall. This is similar to the concept of "attention allocation" in the organizational sociology literature. The cognitive resources of an organization are limited, and it has to convert routines into implicit knowledge and allocate attention to contingencies. Similarly, by actively converting consensus into implicit knowledge, science saves its cognitive bandwidth by focusing on contentious new ideas/concepts in published papers.

We argue newly published ideas and concepts get used in ways which reveal their transition from initial knowledge claims and loci of dispute to more implicitly held and accepted facts (Figure 1). When new ideas and concepts are first proposed, only a small subset of those published get used in ensuing papers and garner explicit attention and citation (phase 1). This recognition may be from competing camps that lack consensus.



Over time, the idea develops more stable associations and loses salience in its citation network, reflecting increasing consensus. At some point (phase 2), the concept becomes a category like "recurrent neural network (RNN)" in the machine learning literature. After becoming a category, it receives a diminished citation count but increased usage.

Figure 1. Career Trajectory of Scientific Concepts and Ideas

Data & Methods

Concepts (c): We use the keywords generated from topic modeling by the OpenAlex team as our unit of analysis. Based on titles and abstracts, up to five keywords are assigned to each paper. We study keywords whose first appearance was between 1950 and 2014.

Papers. 65 million publications from the *OpenAlex* with abstracts available **and** have type "article", "pre-print", "dissertation", or "book"; **and** being cited at least once.

Seed Papers ($P(c)$): For each keyword, we identify the first *ten* papers that mention the keyword as seed papers that propose the concept.

Papers Mentioned the Concept:

We also track the number of mentions each concept received in each year by counting the number of papers published in year t that were also associated with the concept.

Citation Count of the Concept:

We locate all subsequent papers that cite the seed paper proposing the concept using the reference section of each paper. The number of citations a concept received in year t is the sum of citation counts of all its seed papers.

Concepts' Career Trajectories: We have now constructed two time series capturing a concept's career trajectory since its first appearance. We also include a third time series measuring the ratio (# mentions + 1) / (#citations + 1).

Time Series Feature Extraction: Different concepts have trajectories of different lengths, we first construct fixed-length summary statistics of these time series data for each of these concepts (≈900 features in total).

Concept Clustering: We perform a dimension reduction and then cluster concepts into three major clusters based on their trajectory pattern.

Results

Figure 2 and 3 display the clustering results of scientific concepts and ideas based on three metrics: the number of mentions, the number of citations, and the ratio of mentions to citations over time. The results are divided into three clusters, each with distinct average values, indicating different patterns of prominence and decline in scientific influence over time. Figure 2 includes all detected scientific concepts and ideas on OpenAlex whereas Figure 3 only includes scientific concepts and ideas in Computer Science and Math. We see similar patterns between two figures.

Cluster 0 (blue) likely represents scientific ideas and concepts that lose credibility or relevance overtime, characterized by high contentiousness at the beginning (high citation) but flat mention/citation ratio overtime. Cluster 2 (light blue) likely represents those influential scientific ideas and concepts that keep have relevance in their fields. Cluster 1 (brown) likely represents those highly influential scientific concepts and ideas that eventually become categories and subfields, such as RNN in computer science.

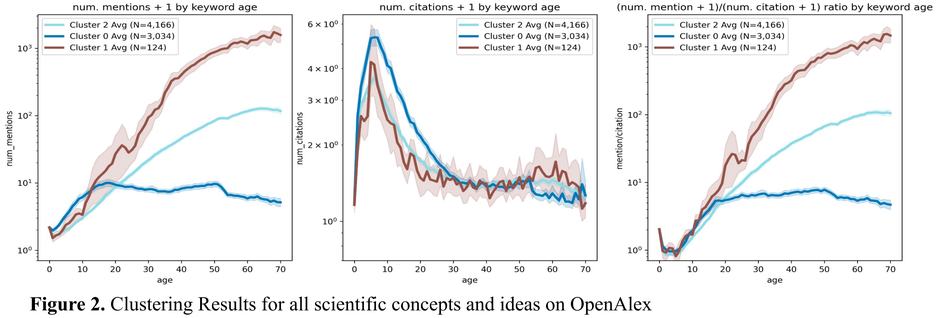


Figure 2. Clustering Results for all scientific concepts and ideas on OpenAlex

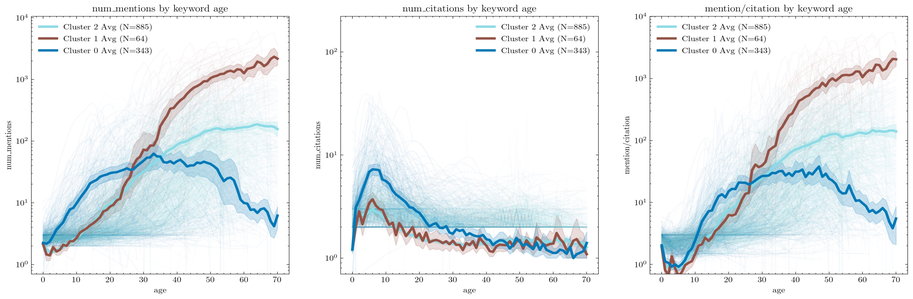


Figure 3. Clustering Results for scientific concepts and ideas in Computer Science and Mathematics on OpenAlex

References

[1] Johan S. G. Chu and James A. Evans. 2021. "Slowed canonical progress in large fields of science." Proceedings of the National Academy of Science.

[2] Meng, Xiangyi, Onur Varol, and Albert-László Barabási. 2024. "Hidden Citations Obscure True Impact in Science." PNAS Nexus 3(5): pgae155.

[3] Mengjie Cheng, Daniel Scott Smith, Xiang Ren, Hancheng Cao, Sanne Smith, and Daniel A McFarland. "How new ideas diffuse in science." American Sociological Review, 88(3):522–561, 2023.

[4] Michael Park, Erin Leahey, and Russell J Funk. Papers and patents are becoming less disruptive over time. Nature, 613(7942):138–144, 2023.