Superstar Extinction and Brainstorming Ideation Notes

Sean Kelty

Started 19 February 2020

1 Superstar Extinction Status

Superstar academics have been shown to have an impact of collaborators work, signified by a drop in publications of collaborators after a Superstar's Sudden Death. We aim to take this idea of superstar influence further, looking at the content of the abstracts themselves and looking at how information content of papers influences future works in an academic network. We have a working PhraseLDA algorithm that embeds documents and phrases of abstracts into a latent vector representation on which to analyze content. Now we wish to utilize this infrastructure to uncover the way content and knowledge is developed through a network of academics.

Academia is a community in which the goal is to make novel, impactful, and creative contributions to the unniversal body of knowledge, so in theory, more creative work that combines a myriad of topics that were not yet connected would be highly influential. A citation network of these articles is highly complex, so how can we clarify driving sources in the network that induce creativity? We look at "superstars" in the network, defined by some accepted metric (h-index). The idea is that superstars are highly creative and their work is very influential. So following and being inspired by a superstar would result in higher acclaim as an academic.

We provide a list of hypotheses to investigate regarding the inspiration and creativity of an academic network and the influence of "superstars" in creative output.

1. **H1: More creative work is more influential than less creative work**

    - We use a metric for creativity called the "Creativity Quotient Q" which takes the max spanning tree of a fully connected network of

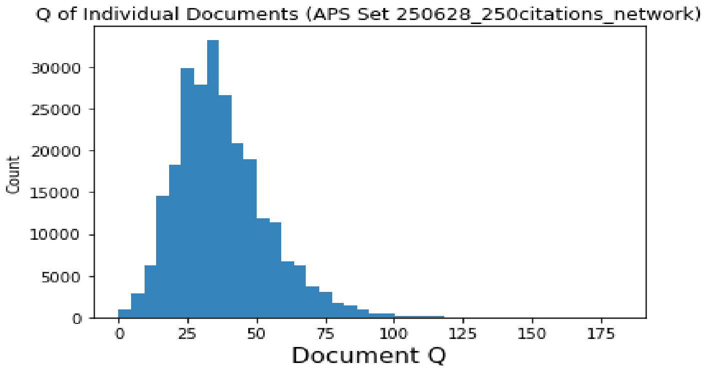


Figure 1: Histogram of Q for APS corpus of 250k Documents

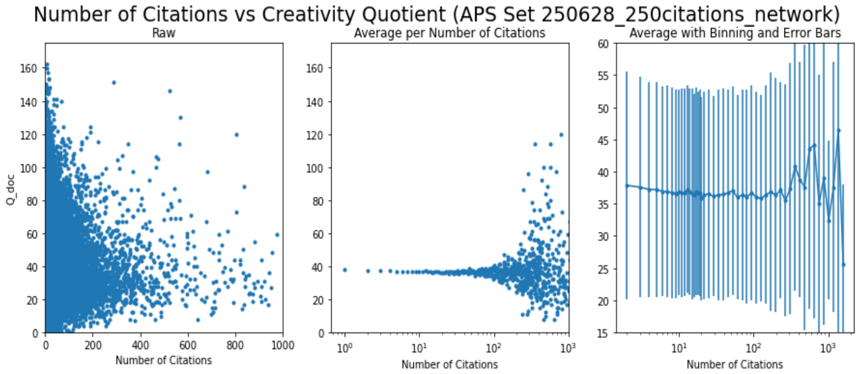


Figure 2: Correlation of Creativity Quotient Q to Number of Citations

all phrases in an abstract based on their vector embeddings. Fig 1 shows a distribution of Q in the corpus, resembling a poisson distribution in Q and centering around a mean of Q = 30.

- Fig 2 shows the correlation between the number of citations and the creativity quotient. We show that the Creativity Quotient is not positively correlated to the number of citations. In fact we see on average the creativity quotient stays fairly constant, slightly decreasing with number of citations, and becomes noisy in the realm of high citations.

2. **H2: Superstars produce more creative work**

   - According to the Creativity Quotient Q, Superstars are shown to have a near-identical distribution of creative an output as the

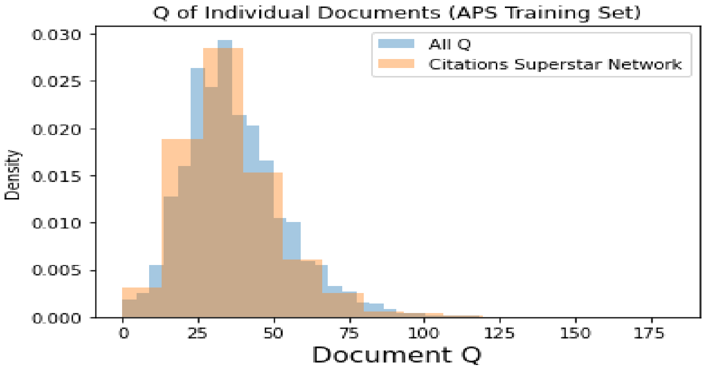


Figure 3: Correlation of Creativity Quotient Q to Number of Citations

entire corpus. Maybe looking at a different measure of creativity will yield a better signal.

3. **H3: Superstars produce more novel work**

   - This analysis has not yet been done. I think using the measure in the diversity-innovation paradox paper would be good: with novelty defined as the first time a term appears in the corpus, how much influence does this new term have over different topics in the corpus?

4. **H4: Individuals highly inspired by a superstar have more creative works**

   - We define an "inspirator" as an academic that cites a superstar paper. Highly inspired academics are in the top percentile of inspirators (those that cite the superstar most), while lowly inspired academics are in the lower percentile.
   - According to the creativity quotient Q, we see that (Fig **??**) highly and lowly inspired inspirators both show the same average creativity quotient based around $t = 0$, the year in which a superstar first publishes a paper. It is worth re-examining this plot with a different measure of creativity for documents. Perhaps document entropy?

5. **H5: Individuals highly inspired by a superstar have higher redundancy of topics among themselves than lowly inspired academics**

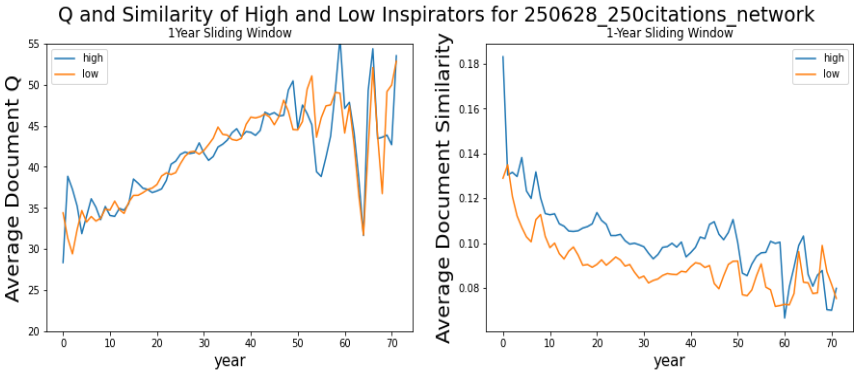


Figure 4: Correlation of Creativity Quotient Q to Number of Citations

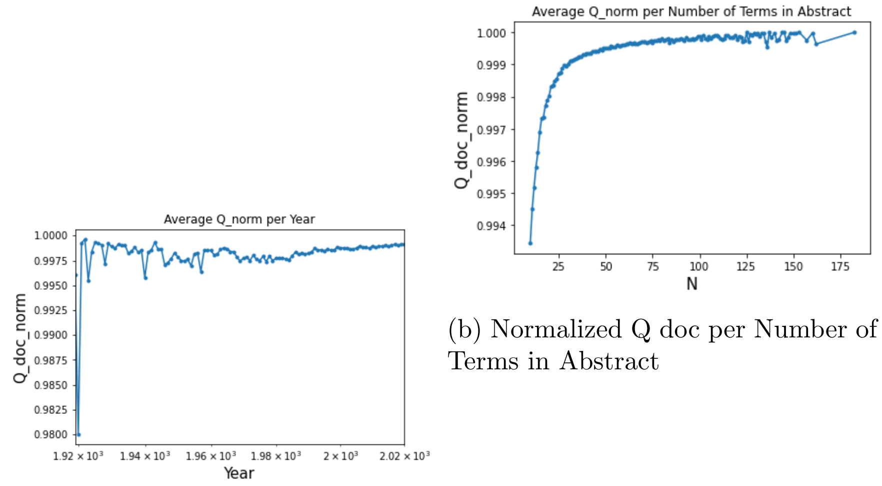
- We calculate similarity with the cosine similarity, where each document is a vector representation over topics. We see that highly inspired academics have higher similarity among their works than lowly inspired academics (Fig **??**).

6. **Q: Do superstars have a highly diverse body of work ranging from many different topics, or are they highly influential in their own field that induce much creative and spinoff work within their own academic communities?**

2 Issues with Analysis

2.1 Creativity Quotient Q as a measure for creativity

Our PLDA algorithm has developed vector representations for the phrases in the corpus. The algorithm uses document frequencies and corpus frequencies to determine a topic assignment for a phrase in a document. Many of the phrases are coerced towards one topic, many have few varying topic assignments. For the most part, phrases have only a select few topic representations over the whole corpus. So Q, which is calculated by a max spanning tree by-way-of a min spanning tree, is very high for most documents because word-topic vectors contain many zeroes, which makes traversing a min spanning tree fairly trivial. When scaled with the unique number of phrases N in a document (Fig **??**), we see that Q and N are very similar due to the dissimilarity of phrases based on the topic represenation. It could be worth revising or revisiting our measure for Creativity.



(a) Normalized Q doc per Year

(b) Normalized Q doc per Number of Terms in Abstract

Figure 5

2.2 Author-Name Disambiguation

Originally, I designed a loose author-name disambiguation where the author's name was compared with their affiliation and if the name and affiliation matched they were determined to be the same person. Some issues with this

- Some data have author names written in full, and some are italicized.
- Some affiliations have slight differences that represent the same affiliation but are not written the same. For instance "THE University of Rochester" versus "University of Rochester". Some postal codes are written differently or are gramatically different.

This is in my opinion the most annoying part of the analysis, so I'd like to discuss a systematic way for this to be done.