Propagation of Epidemics in Citation Networks

David Kartchner*
Rohit Mujumdar*
david.kartchner@gatech.edu
rohitmujumdar@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

Academic research *should* live or die by its own merits. However, human cognitive shortcuts have long believed to give undue advantage to particular institutions or researchers, sometimes blinding reviewers to errors of lack of rigor on their work. We investigate this imbalance by studying the spread of ideas across academic research networks using disease spread models adapted from epidemiology. We specifically focus on spread of ideas in the domain of Machine Learning (ML), a specialised area of research in Computer Science. We take some papers published in the ICLR 2018 as our set of base "pathogens" and assess the network growth dynamics of idea spread amongst major ML conferences. We choose publication citations as our medium of propagation and define data-centric measures for institutional prestige and idea quality and use both network-based simulation and statistical estimation to quantify how these factors affect idea propagation. To assess if idea spread is driven by connectivity amongst original authors or the explicit prestige of their institution, we use an epidemiological model to simulate the spread of an idea on our collaboration network. We discover that the quality of researcher connections seem to be the driving factor for idea spread instead of prestige. While we hope that these new measures and the subsequent network growth analysis can be instrumental in further research, we also hope that the work could be further helpful in correcting prestige-induced imbalance in academia.

KEYWORDS

epidemiology, heterogeneous information networks, citation networks, collaboration networks, science of science

ACM Reference Format:
David Kartchner and Rohit Mujumdar. 2020. Propagation of Epidemics in Citation Networks. In *Proceedings of CSE 8803: Data Science for Epidemiology (EPI '20).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

EPI '20, October 2020, Atlanta, GA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The acceptance and adoption of one's ideas is a key measure of influence and success in academia. Science is sometimes characterized as an ideal meritocracy, where ideas succeed or fail by their intrinsic merit and potential impact. However, the development of science is inseparably connected from scientists themselves, and recent research into human behavior has shown that humans are highly prone to biased judgements driven by cognitive shortcuts [6]. Since answering the question "Is this a high quality idea?" is difficult and mentally taxing, a natural response is to subconsciously substitute an easier question such as "Was this idea developed by at a prominent institution?" or "Is this researcher well-known in this field?" Such biasing, conscious or not, could lead to serious inequalities in publication and idea dissemination. This, in turn, could substantially shape the career trajectories of individuals and the development of science as a whole. Accordingly, we seek to combine epidemiological models for contagion (idea) spread in conjunction with large-scale empirical publication and affiliation data to answer the question "To what extent does institutional prestige lead to inequality in the spread and acceptance of ideas?"

2 PROBLEM DEFINITION

Assume that we are given a contact network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ of academic researchers, as well as a set of publication documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ with corresponding quality ratings $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$. Further, suppose that we can calculate the "infectivity" of each idea by calibrating an SIR model using empirical citation data over time. We investigate whether there is a significant difference in R_0 of ideas originating at prestigious universities after controlling idea quality for empirical differences in local contact network topology around the publishing researchers.

3 REVIEW OF LITERATURE

3.1 Idea spread on social networks

One of the earlier works in this area was [2], which tried to exploit the similarities between population dynamics in diffusion of Feynman diagrams and spread of infections. The key goal of the paper was applying several models such as SIR, SEI, SEIZ to empirical data in three, very different communities. While they concluded that suitably adapted models did a good job of fitting their data, they also emphasized that behavior of individuals may be very different for exposure to ideas and diseases. The strength of this paper lies determining how the events of idea diffusion are, or are not, analogous to population states commonly used in epidemiological models. We intend to borrow these analogies for our work.

However, this paper applies the model to a very specific dataset in theoretical physics, on 3 countries, across several years post World-War 2. Their results are heavily biased because of the socio-political factors influencing the information spread dynamics of that era. With the technology of today, which exists several decades later than the timeline of the information spread studied in this paper, and the speed, volume and diversity of research being conducted today, the features and parameters derived in the paper may not reflect the general characteristics of spread of ideas in today's times. Our work will assess characteristics of idea diffusion very relevant to the current ML research environment.

We also find relevant parallels in our proposed topic and the information diffusion modelling section of the comprehensive survey of information diffusion in online social networks, conducted in [4]. This paper considers nodes to be in only two states - activated or deactivated. The major idea that can be borrowed from here, to determine what constitutes as an infection in academic information spread, is that of *spreading cascade*, which is based on a simple assumption that people (nodes) in social networks are only influenced by actions taken by their connections. In academia too, the chances of someone getting influenced by an idea are high if their "connections", such as lab peers, publish work on that idea. This idea looks at information propagation as successive activation of nodes. The NETRATE [9] algorithms assume a static underlying network and formulate the exploration of correlation in nodes' infection times as maximum likelihood problems. The algorithm INFOPATH [10] instead, assumes a continuously changing network, and is noteworthy in its use of stochastic gradient descent to provide online estimates of the structure and temporal dynamics. Another idea that could be borrowed, is to be able to identify influential information spreaders, which in our case are the institutions we call prestigious or as having high calibre. We could borrow from the very popular HITS and PageRank algorithms for this.

Our work similarly draws from related work using epidemiological methods to model the spread of ideas in social networks. One recent work [8] shows that the way academic ideas spread at conferences can be modeled with SIR and SEIR models. Peri et al assume that each individual who posts with an official conference hashtag is "infected" and consider all twitter users who ever posted about the conference as "susceptible." However, their work fails to separate their proposed epidemiological models from the dynamics of conference attendance and event planning, which could potentially give rise to similar patterns.

Perhaps the most closely related work to our own is [7], which attempts to quantify whether university prestige leads to inequalities in the influence and perceived value of their research findings. To analyze this, the authors construct a faculty hiring network of the PhD-granting computer science departments in the U.S. and Canada by placing a directed edge (u, v) for each individual that received a doctorate at u and was in a tenured or tenure-track position at v in the 2011-2012 academic year. Each researcher was manually matched with his or her profile on DBLP, an online database spanning major CS journals and conferences. The authors model ideas as separate contagions that are spread across the network of institutions. Specifically, let each publication keyword be a particular contagion. They assume that institution i is "infected" at time t if i has at least 1 faculty member that publishes a paper with this keyword in the given time period.

The authors rank the "prestige" of each university by minimizing the number of rank violation in hiring decisions, i.e. the number of faculty that trained at less prestigious universities than those at which they were hired. In this network, institutional prestige was found to correlate with other institutional rankings (e.g. U.S. News and World Report), as well as other centrality measures such as degree, eigenvector, and closeness centrality.

After performing statistical analysis to confirm that faculty hiring does indeed lead to idea spread between departments, the authors simulate epidemics on the faculty hiring network using the SI model, allowing a particular edge to be used for transmission at most once. They assume that the transition probability p is a surrogate for the intrinsic merit of an idea and measure both the size of the infected population and the distance of the furthest infected node. Based on these simulations, the authors conclude that high-quality ideas will have far spread regardless of starting institution, but low quality ideas will spread significantly farther from prestigious institutions due entirely to the contact network structure. This conclusion still holds when relaxing the network structure to allow for random jumps, leading the authors to conclude that there is a structural advantage for researchers at prestigious institutions that leads to higher visibility and spread of ideas.

3.2 Models for citation network growth

Much of the current theory of idea spread and citation growth on networks has been has been heavily influenced by the Barabási-Albert model [1] of network growth known as preferential attachment. In this model, as new vertices (i.e. papers) are added to a network, they choose a fixed number of nodes to attach to, where the probability p_i of attaching to vertex i is $p_i = \frac{k_i}{\sum_i k_i}$[1]. This creates networks with a power law degree distribution, mirroring scale-free distribution observed in many real-world networks. Recent work has shown that when this model is combined with network growth, the result is a log-normal distribution of citation numbers, which more realistically describe real-world citation networks.

While these growth models may produce synthetic networks with similar statistical properties to real-world networks, they do not realistically model the transmission of ideas since researchers work in particular academic subfields. Accordingly, we seek to use heterogeneous networks composed of researchers, publication venues, institutions, and papers to more accurately capture how ideas spread between researchers and institutions, thus resulting in the generation of new papers. We hope that this will allow us to both more accurately characterize idea spread and accurately predict network growth dynamics.

A very recent work [3] has modeled the transfer of ideas from academic research to industry adoption ("translational research") via textual analysis key phrases that first appear in academic literature and proceed to "infect" ideas used in industry, measured by their appearance in subseqent patents and clinical trials. Specifically, the authors use AutoPhrase [11] to identify high quality phrases as a surrogate for actual research ideas, and then model different

[1]There are more complex formulations that allow for different forms of this attachment function to model nonlinear preferential attachment dynamics

features about the authors, publication patterns, and origin institutions that may influence whether an idea is adopted is translated into real-world use. They conclude that ideas are more likely to translate ideas spread in the research community and are repeated by the original author in multiple publications.

4 PROPOSED METHOD

4.1 Intuition

Our proposed method is a different from state-of-art in several ways, most prominent of them articulated in Table 1.

We believe that these design choices would make our method better than the state-of-art. We choose citation networks and research collaboration networks as our ground-bed for infection spread. While faculty hiring can be instrumental in idea spread, we believe that one work citing another is the most robust and certain indication of an idea being spread. Relying only faculty hiring network can cause us to lose out on several cases of idea spread when researchers build on previous work or even derive inspiration of abstract or tangible ideas around modeling, solving or assessing a problem at hand. Thus, we choose a publication as our unit of contagion as opposed to a publication keyword. While one paper citing another is a necessary indication of infection, we further qualify this condition by assuming that an infection happens only when the said paper has an overlap with the abstract and keywords with the paper it is citing. This makes the infection condition sufficient and more rigorous than the current technique of using only keywords as a unit and condition of infection can be misleading. Two papers having only one keyword overlap may not be indicative of 'idea' spread since the keyword overlap may be indicative of them only sharing/working in the same domain or sub-domain - they may never cite each other or build on the each other's work.

We also bring in more nuance to our concepts of prestige and idea merit. We still use the prestige metric devised by the current method, but we add more nuance to our metric prestige by factoring in other rankings driven more by reputation and assign them higher weightage. Paper keywords on their own do not cannot have a merit value attached to them without a context, so using transition probabilities as surrogates for idea metric in the current technique makes sense. However, using a paper as a unit of infection allows us to actual paper review scores decided by field experts to have a more tangible and a standardised measure.

4.2 Description

4.2.1 ***Analogies with epidemiology.*** We base our work on analogies with concepts and models in epidemiology, modelling the spread of ideas on propagation of epidemics of infectious diseases. This involves drawing parallels between concepts and events in the spread of infectious diseases and spread of research in academia.

Following previous work [7], we let "contagions" be research ideas, denoted by papers with research topic keywords. We observe infections as a publication on a particular research topic within a time period t. We say that a paper A has infected paper B if B cites A and if the overlap of the domain and topics of A and B are very high, i.e. B has one of the same keywords as A. We further allow an individual to be in an "exposed" state, meaning that a researcher has seen and is incubating an idea but does not have an observed infection (publication) event yet. After a researcher has been dormant on a particular topic for a suitably long time, we assume that he or she returns to the susceptible pool. This is a realistic assumption since individuals tend publish repeatedly on topics throughout their careers. We therefore assume that retirement is the the only way a researcher is "removed" from the population of researchers.

4.2.2 ***Preparing the Citation and Infection Network.*** We query the Microsoft Academic Graph to extract data for papers published in 2018 and beyond for the major ML conferences ACL, ICCV, CVPR, ICLR, ICML, NAACL, NEURIPS. We create a network (a directed graph) of all papers citing one another. This graph has 409472 one-way edges. Here is a tiny subset of that graph.

Figure 1. A Tiny subset of the Citation Network

Using the citation network we built, from each pair of adjacent nodes, we select those edges, which satisfy the following conditions. If A is the source node (the paper/node which was cited) and B is the destination node (the node which cited A), then select the triplet *A-directed_edge-B* for our infection network if :

(1) The overlap of unique, normalised, stop-word filtered abstract words between A and B is more than 10%, **AND**
(2) The overlap of the field of study terms between A and B is more than 10%

After applying these conditions on our network, we get a subnetwork of infections, i.e. our **infection network** with 108326 directed edges.

4.2.3 ***Estimating Idea Quality.*** In most fields of study, idea quality is measured by the impact factor of the journal in which research is published, normalized to the subfield to which it belongs. More people tend to read top journals (e.g. Nature, New England Journal of Medicine), so these papers in turn gain citations both from quality and audience. Computer science conferences, however, offer an additional level of granularity in their accepted papers. While many top venues (e.g. NeurIPS, ICLR, ICCV) all have similar bars of quality, some also publish the reviewer ratings on their papers. This provides us with a quantitative metric of quality for papers within the same conference and allows us to see how much the quality of an idea drives it's acceptance and citation as opposed to other factors.

Model	Current Method	Proposed Method
Contact Network	Faculty hiring network only	Network of researchers (collaborators and colleagues) modeled on a network of citations
Unit of contagion	A publication keyword	A publication
Condition of Infection	Atleast one faculty publishes with the keyword	Paper A cites Paper B and they have a certain overlap of abstract and keywords
Prestige Metric	Determined by minimizing the number of rank violation in hiring decisions	Weighted average of rank violation prestige, US News Rankings Peer Assessment score, CS rankings Geometric Mean Count
Idea Calibre Metric	Uses transition probability as a surrogate for idea merit	Double-blind peer review scores of the contagion idea (publication)

Table 1. Differences between current and proposed method

In order to estimate paper quality, we queried openly available review data from OpenReview.net [12] for papers accepted to ICLR in 2018-2020. This data provides three independent estimates of paper strength, as well as a the final committee decision on the paper. We compute the average rating of each paper by taking the mean of these three independent estimates.

We attempted to retrieve this data for a larger number of conferences with a longer follow-up period, but were not able to retrieve individual reviews for any other conferences or for any years before 2018. Hence, we use data from the three available years of ICLR and leave expansion to other conferences as future work.

While ICLR papers are included in Microsoft Academic Graph, their unique identifiers are not linked to the ratings from open OpenReview, nor do their names provide an exact match. Thus, we link papers between corpus using fuzzy string matching on titles. Using a threshold of 90% similarity based on Levenstein distance, we are able to match 425 accepted and workshop papers between the two corpora.

In order to measure the affect of prestige in paper spread, we further filter papers to those where at least one author is affiliated university in the United States. This leaves us with a total of 123 ICLR papers papers linked to MAG with both prestige and quality data

Thus, we treat these papers from 2018 as our sources of infection, and call them **patient-zero papers**. We performed a fuzzy matching of the titles of 2018 papers from our infection network and this review data. For each of the patient-zero papers, we assign the average rating as its caliber metric. Since the average rating ranges from a scale of 1 to 9, we categorize the caliber of a paper as follows :

(1) STRONG, if rating is above 7
(2) GOOD, if rating is either 6 or 7
(3) AVERAGE, if rating is either 4 or 5
(4) WEAK, if rating is less than 4

4.2.4 ***Devising the prestige metric.*** We build on the work [7], to devise a prestige metric. The authors of this work rank the "prestige" of each university by minimizing the number of rank violation in hiring decisions, i.e. the number of faculty that trained at less prestigious universities than those at which they were hired. To access the prestige scores calculated by [7]. We worked with the original authors to understand how their metric was calculated and build on this metric by adding CSRankings and US News and World Report rankings.

We choose these two rankings to have a more balanced, nuanced and holistic metric of university prestige. CSRankings weighs departments by their presence at the most prestigious publication venues. Note that, including this as a metric is not in conflict with or contradictory to what we are trying to assess in our project. CS Ranking metrics are decidedly non-citation based, owing to alleged manipulation and logistical challenges in in incorporating citation impact on the rankings. Hence, for around 177 US universities we have collected the Geometric Mean Count (GMC) of papers published in the areas relevant to our task, viz. Artificial intelligence, Computer Vision, Machine learning and data mining, Natural language processing, The Web & information retrieval.

On the other hand, US News and World Report rankings, which are arguably the most popular and influential rankings, are heavily reputation-based and rely on surveys sent to department heads and directors of graduate studies. Apart from the rank of a university, the only metric visible to the general public is the Peer Assessment Score, which is a measure of how a school is regarded by administrators at peer institutions on a peer assessment survey. A school's **Peer Assessment Score (PAS)** is determined by surveying presidents, provosts and deans of admissions, or officials in equivalent positions, at institutions in the school's ranking category. Since this survey would be a direct reflection of the institution's prestige (in the eyes of its peers), it is analogous to the peer review process in assessing publications, which in turn is a reflection of the merit of the idea in the publications. Hence, we choose this metric as our third metric in determining an institution's prestige and have collected the PAS for around 188 US universities. Note that PAS is inversely proportional to the rank.

We take a weighted mean of all the three metrics for each university to compute the final prestige score. Let the prestige metrics based on faculty hiring, US News Rankings and CS rankings be denoted by pi, U, C respectively. Then the prestige score is given by :

$$prestige = \frac{(0.5 * pi) + (3 * U) + (1 * C)}{0.5 + 3 + 1}$$

Faculty hiring is given a lower weight because this prestige measures a department's ability to place its graduates as faculty at other institutions. However, this technique seems like it would bias things based on the proportion of faculty that go into industry vs academia, which will vary by university (e.g. Stanford has

a very strong startup culture). US News rankings has been given the highest weight owing to it being driven largely by reputation and peer-survey scores. We believe that calculating prestige as a function of the above three metrics would be a incorporate more 'aspects' of prestige in today's academic world.

Now that we have the prestige metric for each university, for each of our patient-zero papers, we take a mean of the affiliations of the authors of the paper and use it to decide the prestige of the origin of an idea i.e the infection source.

4.2.5 ***Preparing the Collaboration Network.*** Collaborations are an important way which ideas spread between individuals. The more often an individual is exposed to an idea, the more likely they are to reference it in their writings or research. Accordingly, we create a "contact network" between researchers over which ideas can spread via interaction and collaboration. We make an edge between two researchers if they have collaborated together on a publication i.e. if they are co-authors, or if they are colleagues in an institution.

We create a collaboration network of machine learning researchers by taking all papers published in top conferences between 2016-2020 and creating links between every pair of authors who published a paper together during that time period. We consider the following conferences to obtain publications for constructing our collaboration network:

(1) **Computer Vision:** Internation al Conference on Computer Vision (ICCV), Computer Vision and Pattern Recognition (CVPR)
(2) **Natural Language Processing:** Association for Computational Linguistics (ACL), North American ACL (NAACL), European ACL (EACL), Empirical Methods in Natural Language Processing (EMNLP), Computational Linguistics (COLING)
(3) **Data Mining:** ACM Conference on Knowledge Discovery from Data (KDD), ACM World Wide Web Conference (WWW)
(4) **General Maching Learning**: International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), Neural Information Processing Systems (NeurIPS), Association for the Advancement of Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI)

With this, we get 66,452 vertices (researchers/authors) and 296,268 edges.

4.3 Simulating Epidemics

In order to evaluate if idea spread is tied to the connectivity and collaborations of the original authors rather than the explicit prestige of their institution, we simulate the spread of an idea on our collaboration network using the Susceptible - Infected - Recovered (SIR) model of epidemics. At each time step t, an infected node can spread its contagion to each of its neighbors with probability β. Let S, I, and R denote the sets of nodes that are respectively susceptible, infected, and recovered. While this progresses as a stochastic process, its expectation at any given timestep is modeled as:

$$\Delta S = -\sum_{s \in S} \left(1 - (1-\beta)^{k_s}\right)$$

where k_s is the number of infected neighbors of node s. We further assume that infected nodes recover with some probability γ at each time step, yielding an expected number of recoveries as:

$$\Delta R = \sum_{i \in I} \gamma$$

Combining these two equations and noting that $\Delta S + \Delta I + \Delta R = 0$ yields:

$$\begin{aligned}\Delta I &= -(\Delta S + \Delta R) \\ &= \sum_{s \in S} \left(1 - (1-\beta)^{k_s}\right) - \sum_{i \in I} \gamma\end{aligned}$$

To approximate these quantities in practice, we run multiple simulations and average results over all runs.

5 EXPERIMENTS AND RESULTS

5.1 Experimental Questions

Using our collaboration network, we intend to determine whether idea quality and institutional prestige affect idea spread. Specifically, we answer the following questions:

(1) Do collaboration networks accurately model the spread of ideas in academia?
(2) How much do institutional prestige and paper ratings affect how much ideas will spread?

5.2 Evaluation Measures

5.2.1 ***Collaboration Network Evaluation.*** We adopt a variety of measures to gauge the strength and spread of ideas in our network. We begin with an evaluation of our network to characterize differences in network topology that may contribute to the spread of ideas, including the following for authors and institutions:

- Degree distribution
- Author degree
- Closeness centrality

5.2.2 ***Tracing idea infections.*** To evaluate idea spread, we use publication data from ICLR 2018 from Open Review with corresponding paper rankings and author affiliation data. We compute a quality score for each paper as described in section 4.2.3. We further compute a prestige score for each paper by averaging the prestige of each author's institution, where prestige is calculated as described in section 4.2.4.

We use the networkx [5] package to represent the infected subnetwork of papers. Using relevant functions from this package, we trace (perform a Depth-First-Search) for each of the patient-zero papers (walk down the path of of all the papers that cited the source directly or indirectly) until we reach all of its last descendants. We note the authors for the patient-zero papers (we call them **founder authors**) and we also make a set of all the authors of the descendants of patient-zero. These would be all the authors who ever cited the patient-zero paper and we call them We use the networkx [5] package to represent this data **descendant authors**. We then determine the idea's infectivity based on the number of other authors subsequently infected

Once we have quantified the idea spread, we perform linear regression to analyze which factors influence idea propagation,

including prestige and reviewer rating. We additionally control for the position in his/her research collaboration network by adding a control for the average size of a simulated epidemic originating from the authors of the paper. Given that the papers come from the same year of the same venue, we assume that no other control variables are needed. For simulated epidemics, we repeat simulations over 20 runs and average results to get a more robust estimate of total number of infections. We fix $\beta = .06$ as the probability of an infection being spread across an edge with $\gamma = .2$ as the probability of an infected node recovering in any given timestep.

5.3 Results

5.3.1 ***Collaboration Network Connectivity.*** A plot of the degree distribution of our collaboration network is shown in Figure 2. From this, we observe that our graph follows a power law degree distribution. Additionally, scatter plots of author degree vs. university rank and author closeness centrality vs. university rank are shown in Figure 3. From these plots, it appears that highly ranked universities have more connected researchers, though this may be driven by the fact that they have more research faculty in general.

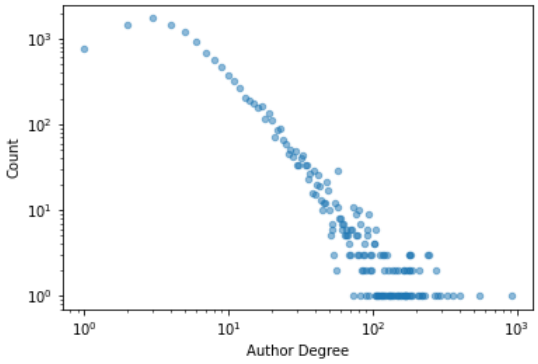

Figure 2. Degree distribution of collaboration network

It is worth noting that there are exceptions to the high degree and centrality of researchers in top institutions. York University, a school with relatively low prestige, had the highest average degree and closeness centrality. Thus, many researchers have highly connected faculty in the research community which could facilitate the spread of ideas through collaboration.

5.3.2 ***Infection Subnetwork.*** Once we have the idea calibre, prestige, and the founder and descendants author details for each patient-zero paper, we can visualize relationships between variables to identify meaningful patterns.

We see from Fig. 4 that there are only a few obvious, discernible patterns at the outset. We see that that the graphs of the number of researchers and papers infected by an idea show similar patterns with respect to the calibre of an idea and the prestige of its university origin. It is interesting to note that ideas of both, extremely high and low caliber infect fewer researchers. It follows that most researchers get infected with good or moderate ideas. We also see that on an average, extremely weak ideas seem to originate at low prestige

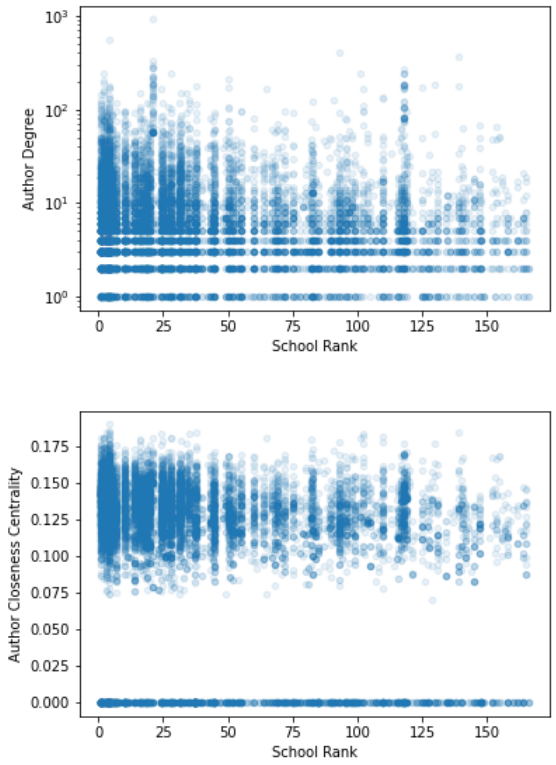

Figure 3. Scatter plots of university rank vs. author degree (top) and author closeness centrality (bottom)

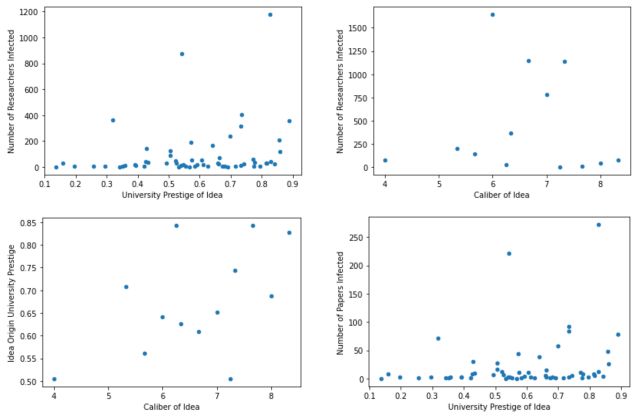

Figure 4. Visualizing patient-zero paper attributes

universities only. However, that is not necessarily true the other way round - we have a data point showing that low prestige universities also come up with high-calibre (STRONG category ideas). However, the over-arching observation is that, prestige and idea caliber seem to have a strong (positive) correlation.

5.3.3 ***Quantitative Idea Spread.*** For each paper in our OpenReview dataset, we performed 20 SIR simulations with the paper's authors set as the initial infected nodes. We additionally calculated

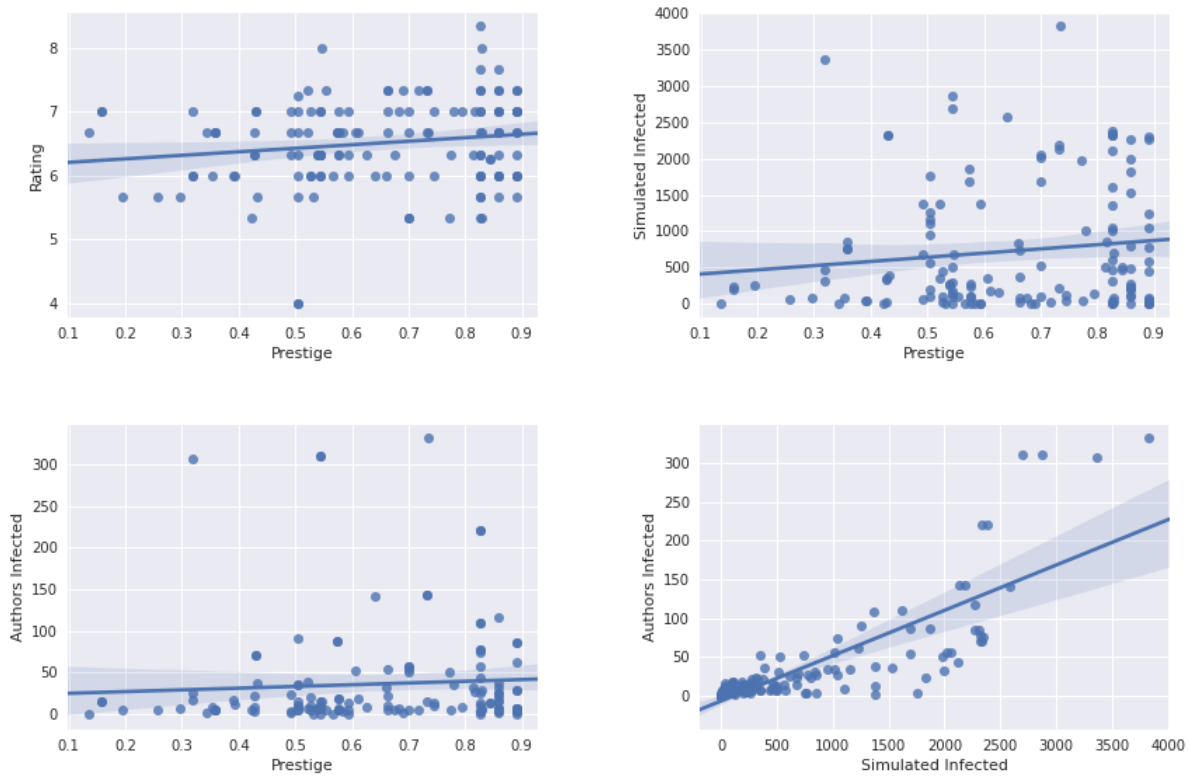

Figure 5. Scatter plots of prestige with paper rating (top left), simulated number of author infections (top right), and actual number of infections (bottom left). We also show the correlation between the simulated and actual number of author infections (bottom right).

the total authors infected by the paper as the set of all authors in its descendants, as described in the previous section. Since descendants are purely from publications, not collaborations, this avenue allows for authors to be infected by encountering a paper in a conference or online venue, rather than via direct collaboration.

We plot comparisons of prestige, rating, and idea spread in Figure 5. We see slight positive correlations between prestige and both paper rating and idea spread (measured as number of infected authors in two years after publication). We additionally see a slight positive correlation between paper prestige and the size of simulated epidemics. However, we note a strong positive correlation between simulated epidemic size and actual number of infected authors, suggesting that local collaboration network structure is an important vehicle for paper spread, especially in the first few years after publication.

We additionally performed a statistical analysis of the variables impacting empirical idea spread. In doing so, we find no significant relationship between either prestige or paper rating after controlling for simulated epidemic size. This suggests that the local structure of collaborations surrounding paper authors are the most important factors influencing their spread, at least in the earliest stages of adoption. This further suggests that idea spreading could be egocentric, with focus more on individual researchers than the institutions at which they work.

Variable	Coefficient	P-Value
Rating	1.312	0.777
Prestige	-13.924	0.392
Simulated epidemic size***	0.0587	<0.001

Table 2. Regression coefficients for actual number of infected authors. The only statistically significant variable is the simulated epidemic size, with p-value < 0.001.

5.4 Conclusion and Future Work

Idea spread and uptake is the lifeblood that sustains the career of an academic. It also determines what discoveries will go on to shape the future of the world. In this project, we investigated the mechanisms that drive idea propagation and discover that neither the prestige of an institution, nor the rating given by peer reviewers predicts whether an academic idea will spread. Rather, it appears to be driven by how well connected the researchers themselves are, which can lead to uptake by collaborators and thus better spread and name recognition.

There are multiple promising avenues to extend this work. First, there is great possibility to expand these ideas about idea propagation to other domains of science, such as biology and medicine. The main difficulty for such a project would be obtaining reviewer scores and normalizing them appropriately across journals and fields of science (in this work, we were only able to locate a single

ML conference that shared it's reviewer ratings). Such expansion would add robustness to the ideas developed here by adding data that is both more varied and larger in scope. Since MAG also contains data about patents and patents also cite one another in their prior art search, our methods can be used to in the intellectual property domain.

Second, there many other avenues through which we could encapsulate ideas and thus model their spread between papers. We spent considerable time exploring unsupervised phrase mining via AutoPhrase [11] to represent ideas, but ultimately decided to use keyword overlap for the sake of simplicity. However, high-quality phrases could potentially be used to replace user-defined keywords, making overlap between papers more objective.

Finally, our results provide preliminary evidence that academic ideas spread in an ego-centric manner, following highly relevant individuals rather than institutions. Future work could investigate how much citations skew towards individuals and the whether this is driven actual differences in content or rather differences in author notoriety and his/her marketing of ideas.

REFERENCES

[1] Albert László Barabási, Réka Albert, and Hawoong Jeong. 1999. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications* 272, 1 (10 1999), 173–187. https://doi.org/10.1016/S0378-4371(99)00291-5

[2] Luís M.A. Bettencourt, Ariel Cintrón-Arias, David I. Kaiser, and Carlos Castillo-Chávez. 2006. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications* 364 (5 2006), 513–536. https://doi.org/10.1016/j.physa.2005.08.083

[3] Hancheng Cao, Mengjie Cheng, Zhepeng Cen, Daniel A. McFarland, and Xiang Ren. 2020. Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora. (2020). http://arxiv.org/abs/2010.06657

[4] Adrien Guille and Hakim Hacid. [n.d.]. Information diffusion in online social networks: A Survey. ([n. d.]).

[5] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, Gaël Varoquaux, Travis Vaught, and Jarrod Millman (Eds.). Pasadena, CA USA, 11–15.

[6] Daniel Kahneman. 2011. *Thinking Fast and Slow.* Macmillan.

[7] Allison C. Morgan, Dimitrios J. Economou, Samuel F. Way, and Aaron Clauset. 2018. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* 7, 1 (2018). https://doi.org/10.1140/epjds/s13688-018-0166-4

[8] Sai Santosh Sasank Peri, Angela Liegey Dougall, Bodong Chen, and George Siemens. 2020. Towards understanding the lifespan and spread of ideas: Epidemiological modeling of participation on twitter. *ACM International Conference Proceeding Series* (2020), 197–202. https://doi.org/10.1145/3375462.3375515

[9] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the Temporal Dynamics of Diffusion Networks. (5 2011). http://arxiv.org/abs/1105.0697

[10] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. 2012. Structure and Dynamics of Information Pathways in Online Media. (12 2012). http://arxiv.org/abs/1212.1464

[11] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated Phrase Mining from Massive Text Corpora. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1825–1837. https://doi.org/10.1109/TKDE.2018.2812203

[12] David Soergel, A Saunders, and A McCallum. 2013. Open Scholarship and Peer Review: a Time for Experimentation. *Proceedings of the 30th International Conference on Machine Learning* 28 (2013). http://openreview.net/document/28cb8b58-d6f9-45c9-936f-c6c60e674381

Propagation of Epidemics in Citation Networks

David Kartchner*
Rohit Mujumdar*
david.kartchner@gatech.edu
rohitmujumdar@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

Academic research *should* live or die by its own merits. However, human cognitive shortcuts have long believed to give undue advantage to particular institutions or researchers, sometimes blinding reviewers to errors of lack of rigor on their work. We investigate this imbalance by studying the spread of ideas across academic research networks using disease spread models adapted from epidemiology. We specifically focus on spread of ideas in the domain of Machine Learning (ML), a specialised area of research in Computer Science. We take the papers published in the year 2014 as our set of base "pathogens" and assess the network growth dynamics of idea spread amongst major ML conferences. We choose publication citations as our medium of propagation. We assess if ideas of the same calibre spread faster when originating from more prestigious institutions. We define data-centric measures for institutional prestige and idea quality and use both network-based simulation and statistical estimation to quantify how these factors affect idea propagation. This also entails defining measures of and around idea propagation. We determine new ways of measuring these concepts and come up with novel metrics for the same. While we hope that these new measures and the subsequent network growth analysis can be instrumental in further research, we also hope that the work could be further helpful in correcting prestige-induced imbalance in academia.

KEYWORDS

epidemiology, heterogeneous information networks

ACM Reference Format:
David Kartchner and Rohit Mujumdar. 2020. Propagation of Epidemics in Citation Networks. In *Proceedings of CSE 8803: Data Science for Epidemiology (EPI '20).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

EPI '20, October 2020, Atlanta, GA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The acceptance and adoption of one's ideas is a key measure of influence and success in academia. Science is sometimes characterized as an ideal meritocracy, where ideas succeed or fail by their intrinsic merit and potential impact. However, the development of science is inseparably connected from scientists themselves, and recent research into human behavior has shown that humans are highly prone to biased judgements driven by cognitive shortcuts [6]. Since answering the question "Is this a high quality idea?" is difficult and mentally taxing, a natural response is to subconsciously substitute an easier question such as "Was this idea developed by at a prominent institution?" or "Is this researcher well-known in this field?" Such biasing, conscious or not, could lead to serious inequalities in publication and idea dissemination. This, in turn, could substantially shape the career trajectories of individuals and the development of science as a whole. Accordingly, we seek to combine epidemiological models for contagion (idea) spread in conjunction with large-scale empirical publication and affiliation data to answer the question "To what extent does institutional prestige lead to inequality in the spread and acceptance of ideas?"

1.1 Addressing Proposal Feedback

One of our proposal feedbacks stated that that there should be some EPI link to our project. In this regard, we have decided to also run our experiments on Allen AI's CORD-19 dataset [7]. The CORD-19 dataset has a Microsoft Academic Graph Metadata ID Mapping, which we intend to use. The overall idea of our task remains the same - to assess if ideas of the same calibre spread faster when originating from more prestigious institutions. The other feedback was about taking down some default text from the template such as the default commands are for a PROCEEDINGS abstract or paper.

2 PROBLEM DEFINITION

Assume that we are given a contact network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ of academic researchers, as well as a set of publication documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ with corresponding quality ratings $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$. Further, suppose that we can calculate the "infectivity" of each idea by calibrating an SEIR model using empirical citation data over time. We investigate whether there is a significant difference in R_0 of ideas originating at prestigious universities after controlling idea quality for empirical differences in local contact network topology around the publishing researchers.

3 REVIEW OF LITERATURE

3.1 Idea spread on social networks

One of the earlier works in this area was [2], which tried to exploit the similarities between population dynamics in diffusion of Feynman diagrams and spread of infections. The key goal of the paper was applying several models such as SIR, SEI, SEIZ to empirical data in three, very different communities. While they concluded that suitably adapted models did a good job of fitting their data, they also emphasized that behavior of individuals may be very different for exposure to ideas and diseases. The strength of this paper lies determining how the events of idea diffusion are, or are not, analogous to population states commonly used in epidemiological models. We intend to borrow these analogies for our work. However, this paper applies the model to a very specific dataset in theoretical physics, on 3 countries, across several years post World-War 2. Their results are heavily biased because of the socio-political factors influencing the information spread dynamics of that era. With the technology of today, which exists several decades later than the timeline of the information spread studied in this paper, and the speed, volume and diversity of research being conducted today, the features and parameters derived in the paper may not reflect the general characteristics of spread of ideas in today's times. Our work will assess characteristics of idea diffusion very relevant to the current ML research environment.

We also find relevant parallels in our proposed topic and the information diffusion modelling section of the comprehensive survey of information diffusion in online social networks, conducted in [4]. This paper considers nodes to be in only two states - activated or deactivated. The major idea that can be borrowed from here, to determine what constitutes as an infection in academic information spread, is that of *spreading cascade*, which is based on a simple assumption that people (nodes) in social networks are only influenced by actions taken by their connections. In academia too, the chances of someone getting influenced by an idea are high if their "connections", such as lab peers, publish work on that idea. This idea looks at information propagation as successive activation of nodes. The NETRATE [10] algorithms assume a static underlying network and formulate the exploration of correlation in nodes' infection times as maximum likelihood problems. The algorithm INFOPATH [11] instead, assumes a continuously changing network, and is noteworthy in its use of stochastic gradient descent to provide online estimates of the structure and temporal dynamics. Another idea that could be borrowed, is to be able to identify influential information spreaders, which in our case are the institutions we call prestigious or as having high calibre. We could borrow from the very popular HITS and PageRank algorithms for this.

Our work similarly draws from related work using epidemiological methods to model the spread of ideas in social networks. One recent work [9] shows that the way academic ideas spread at conferences can be modeled with SIR and SEIR models. Peri et al assume that each individual who posts with an official conference hashtag is "infected" and consider all twitter users who ever posted about the conference as "susceptible." However, their work fails to separate their proposed epidemiological models from the dynamics of conference attendance and event planning, which could potentially give rise to similar patterns.

Perhaps the most closely related work to our own is [8], which attempts to quantify whether university prestige leads to inequalities in the influence and perceived value of their research findings. To analyze this, the authors construct a faculty hiring network of the PhD-granting computer science departments in the U.S. and Canada by placing a directed edge (u, v) for each individual that received a doctorate at u and was in a tenured or tenure-track position at v in the 2011-2012 academic year. Each researcher was manually matched with his or her profile on DBLP, an online database spanning major CS journals and conferences. The authors model ideas as separate contagions that are spread across the network of institutions. Specifically, let each publication keyword be a particular contagion. We denote that institution i is "infected" at time t if i has at least 1 faculty member that publishes a paper with this keyword in the given time period.

The authors rank the "prestige" of each university by minimizing the number of rank violation in hiring decisions, i.e. the number of faculty that trained at less prestigious universities than those at which they were hired. In this network, institutional prestige was found to correlate with other institutional rankings (e.g. U.S. News and World Report), as well as other centrality measures such as degree, eigenvector, and closeness centrality.

After performing statistical analysis to confirm that faculty hiring does indeed lead to idea spread between departments, the authors simulate epidemics on the faculty hiring network using the SI model, allowing a particular edge to be used for transmission at most once. They assume that the transition probability p is a surrogate for the intrinsic merit of an idea and measure both the size of the infected population and the distance of the furthest infected node. Based on these simulations, the authors conclude that high-quality ideas will have far spread regardless of starting institution, but low quality ideas will spread significantly farther from prestigious institutions due entirely to the contact network structure. This conclusion still holds when relaxing the network structure to allow for random jumps, leading the authors to conclude that there is a structural advantage for researchers at prestigious institutions that leads to higher visibility and spread of ideas.

3.2 Models for citation network growth

Much of the current theory of idea spread and citation growth on networks has been has been heavily influenced by the Barabási-Albert model [1] of network growth known as preferential attachment. In this model, as new vertices (i.e. papers) are added to a network, they choose a fixed number of nodes to attach to, where the probability p_i of attaching to vertex i is $p_i = \frac{k_i}{\sum_i k_i}$[1]. This creates networks with a power law degree distribution, mirroring scale-free distribution observed in many real-world networks. Recent work has shown that when this model is combined with network growth, the result is a log-normal distribution of citation numbers, which more realistically describe real-world citation networks.

While these growth models may produce synthetic networks with similar statistical properties to real-world networks, they do not realistically model the transmission of ideas since researchers work in particular academic subfields. Accordingly, we seek to

[1]There are more complex formulations that allow for different forms of this attachment function to model nonlinear preferential attachment dynamics

use heterogeneous networks composed of researchers, publication venues, institutions, and papers to more accurately capture how ideas spread between researchers and institutions, thus resulting in the generation of new papers. We hope that this will allow us to both more accurately characterize idea spread and accurately predict network growth dynamics.

3.3 New Literature

A very recent work [3] has modeled the transfer of ideas from academic research to industry adoption ("translational research") via textual analysis key phrases that first appear in academic literature and proceed to "infect" ideas used in industry, measured by their appearance in subseqent patents and clinical trials. Specifically, the authors use AutoPhrase [12] to identify high quality phrases as a surrogate for actual research ideas, and then model different features about the authors, publication patterns, and origin institutions that may influence whether an idea is adopted is translated into real-world use. They conclude that ideas are more likely to translate ideas spread in the research community and are repeated by the original author in multiple publications.

4 PROJECT GOALS

The major goal of the project is to test the hypothesis that ideas of the same calibre spread faster when originating from prestigious institutions. Working towards this creates several sub-goals in the process. To complete our project, we intend to do the following:

- Bootstrap a time-dynamic academic contact network of researchers from existing heterogeneous publication and citation networks.
- Collect and integrate data on paper quality based on double-blind peer review scores and committee selection at top conferences.
- Develop a rigorous methodology to analyze empirical idea spread between researchers using existing data on citation, collaboration, paper timing, and paper topics.
- Quantify the extent to which institutional prestige leads to the spread of ideas after controlling for intrinsic quality and network effects.

We first start off by building a collaboration network of researchers and institutions based on paper citations and significant computational overlap of paper topics. We aim to develop new metrics to determine institution prestige and paper calibre and create an index around the metrics of idea propagation.

5 METHODS

5.1 Models

We base our work on analogies with concepts and models in epidemiology, modelling the spread of ideas on propagation of epidemics of infectious diseases. This involves drawing parallels between concepts and events in the spread of infectious diseases and spread of research in academia.

Following previous work [8], we let "contagions" be research ideas, denoted by papers with research topic keywords. We observe infections as a publication on a particular research topic within a time period t. We say that a paper A has infected paper B if B cites A and if the overlap of the domain and topics of A and B are very high, i.e. B has one of the same keywords as A. We further allow an individual to be in an "exposed" state, meaning that a researcher has seen and is incubating an idea but does not have an observed infection (publication) event yet. After a researcher has been dormant on a particular topic for a suitably long time, we assume that he or she returns to the susceptible pool. This is a realistic assumption since individuals tend publish repeatedly on topics throughout their careers. We therefore assume that retirement is the the only way a researcher is "removed" from the population of researchers.

Milestone Updates :

5.1.1 ***Preparing the Citation Network.*** For each paper in our data we have an attribute that lists the IDs of the papers the paper in concern referenced. Using this information we build a citation network, which is a directed graph with 742246 edges (one-way citations and 14218 nodes (papers). We use the networkx [5] package to represent this data in a n directed network/graph format. Here is a tiny subset of that mega-graph.

Figure 1. A Tiny subset of the Citation Network

5.1.2 ***Preparing an Infection Network.*** We built our first version of the infection network. Using the citation network we built, from each pair of adjacent nodes, we select those edges, which satisfy the following conditions. If A is the source node (the paper/node which was cited) and B is the destination node (the node which cited A), then select the triplet *A-directed_edge-B* for our infection network if :

(1) The overlap of unique, normalised, stop-word filtered abstract words between A and B is more than 20%, **AND**
(2) The overlap of the field of study terms between A and B is more than 30%

After applying these conditions on our network, we get a subnetwork of infections with 20438 edges. This is only the first version of our infection network, and we intend to brainstorm and iterate over our conditions and come up with better models.

5.1.3 ***Towards devising the prestige metric.*** We intend to build on the work [8], to devise a prestige metric. The authors of this work rank the "prestige" of each university by minimizing the number of rank violation in hiring decisions, i.e. the number of faculty

that trained at less prestigious universities than those at which they were hired. To access the prestige scores calculated by [8], we have contacted the authors and hope to receive the reply from them soon. We build on this metric by adding CSRankings and US News and World Report rankings.

We choose these two rankings to have a more balanced, nuanced and holistic metric of university prestige. CSRankings weighs departments by their presence at the most prestigious publication venues. Note that, including this as a metric is not in conflict with or contradictory to what we are trying to assess in our project. CS Ranking metrics are decidedly non-citation based, owing to alleged manipulation and logistical challenges in in incorporating citation impact on the rankings. Hence, for around 177 US universities we have collected the Geometric Mean Count (GMC) of papers published in the areas relevant to our task, viz. Artificial intelligence, Computer Vision, Machine learning and data mining, Natural language processing, The Web & information retrieval.

On the other hand, US News and World Report rankings, which are arguably the most popular and influential rankings, are heavily reputation-based and rely on surveys sent to department heads and directors of graduate studies. Apart from the rank of a university, the only metric visible to the general public is the Peer Assessment Score, which is a measure of how a school is regarded by administrators at peer institutions on a peer assessment survey. A school's Peer Assessment Score (PAS) is determined by surveying presidents, provosts and deans of admissions, or officials in equivalent positions, at institutions in the school's ranking category. Since this survey would be a direct reflection of the institution's prestige (in the eyes of its peers), it is analogous to the peer review process in assessing publications, which in turn is a reflection of the merit of the idea in the publications. Hence, we choose this metric as our third metric in determining an institution's prestige and have collected the PAS for around 188 US universities. Note that PAS is inversely proportional to the rank.

We intend to take the mean of all the three metrics for each university to compute the final prestige score. We believe that calculating prestige as a function of the above three metrics would be a incorporate more 'aspects' of prestige in today's academic world.

5.2 Evaluation Measures

We adopt a variety of measures to gauge the strength and spread of ideas in our network. We begin with an evaluation of our network to characterize differences in network topology that may contribute to the spread of ideas, including the following for authors and institutions:

- Degree distribution
- Closeness centrality
- Betweenness centrality

We then evaluate different measures of idea propagation, including:

- Spread distance (maximum distance of idea spread)
- Number infected
- Infectivity constant β for a given idea
- R_0 for each idea, calculated from calibration of SEIRS model
- Idea longevity, calculated as the number of consecutive years that an idea has $R_0 > 0$

Finally, we will perform a statistical analysis of idea spread controlling for researcher centrality, publication venue, sub-field, and idea quality to determine if institutional prestige affects total idea spread.

5.3 Data (Milestone Update)

5.3.1 *Data Collection and Processing.*

(1) **MAG Preprocessing** : The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study. We use this as our primary source of data and access it using REST-based Microsoft Academic Knowledge API. Out of the four REST methods, we use the 'evaluate' method which evaluates a query expression and returns Academic Knowledge entity results. We pass a query to return attributes of publication data dating 2014 and after for the following conferences :
   (a) ACL
   (b) CVPR
   (c) EMNLP
   (d) ICCV
   (e) ICML
   (f) ICLR
   (g) NAACL
   (h) NEURIPS.

   The data statistics could be summarised as follows:

Conference	Number of Papers	Size in KB
ACL	4474	47911
CVPR	8805	140026
EMNLP	2746	28743
ICCV	3457	54297
ICML	5485	15502
ICLR	2459	23547
NAACL	2496	23681
NEURIPS	5112	31475

   We also extract as many as 50 attributes for each paper, the most important ones, broadly being:
   (a) author information such as name and affiliation
   (b) year, date and venue of publication
   (c) citation count, IDs and citation contexts
   (d) field of study, normalised title and unique, stopword-filtered abstract words

(2) **Paper quality estimation** : In order to estimate paper quality, we queried openly available review data from OpenReview.net [13] for papers accepted to ICLR in 2018-2020. This data provides three independent estimates of paper strength, as well as a the final committee decision on the paper. We attempted to retrieve this data for a larger number of conferences with a longer follow-up period, but were not able to retrieve individual reviews. Accordingly, we have opted to

use two subsets of papers to test our hypotheses about idea propagation on academic networks

(3) **Linking Text Abstracts**
Note that since the MAG data did not provide us raw, unprocessed abstracts, we downloaded bulk data of around 6M papers from S2ORC [7], using its links to MAG ID. We will additionally

6 OTHER MILESTONE UPDATES

6.1 Outline of remaining tasks

So far we have completed the following parts of the project :

(1) Collection of citation network data for authors/papers
(2) Collection of relevant paper metadata, such as title and abstract, from Semantic Scholar. We have linked this data to MAG using a shared ID after collection.
(3) Collection of CORD-19 data for evaluation of idea spread for research during a pandemic
(4) Collection of university rankings from widely-used online sources
(5) Creation of a citation subnetwork of papers from top AI/ML conferences

Following tasks are to be done for the final version :

(1) Run epidemiological simulations on derived citation network
(2) Compute metrics of idea longevity, influence, and spread for various papers from top conferences
(3) Run statistical analyses to determine to what extent institutional prestige influences idea propagation and acceptance.

6.2 Problems faced and Deviations

(1) Changes in data quantity : To have a larger quantity of data, but to also stay relevant to the field of ML, whose growth gained momentum in the past 6 to 7 years, we have changed our base year from 2015 to 2014.
(2) Issues in data collection and aggregation : The abstracts for papers in the MAG dataset were lists of unique, normalised and stopword-filtered words. Since we intended to run keyphrase extraction on abstracts to extract 'infectious ideas', this data was not suitable for the task. We then tried using the Digital Object Identifier (DOI) attribute for each paper to scrape the web to extract abstracts. However, we realised that not all papers in the MAG have a DOI attribute (many of them simply have a Nan value). We then explored Semantic Scholar RESTful API, which provides convenient linking to Semantic Scholar pages and pulling information about individual records on demand. We were able to use our MAG IDs to link to the Semantic Scholar API. However, while the API is freely available, it enforces a rate limit and will respond with HTTP status 429 'Too Many Requests' if the limit is exceeded (100 requests per 5 minute window per IP address). Since this issue came up very recently, we resorted to writing and running a script to download the Semantic Scholar Open Research Corpus (S2ORC) [7] in bulk (approx 186M papers) and filter out the subset corresponding to CS (approx 6M papers).

REFERENCES

[1] Albert László Barabási, Réka Albert, and Hawoong Jeong. 1999. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications* 272, 1 (10 1999), 173–187. https://doi.org/10.1016/S0378-4371(99)00291-5

[2] Luís M.A. Bettencourt, Ariel Cintrón-Arias, David I. Kaiser, and Carlos Castillo-Chávez. 2006. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications* 364 (5 2006), 513–536. https://doi.org/10.1016/j.physa.2005.08.083

[3] Hancheng Cao, Mengjie Cheng, Zhepeng Cen, Daniel A. McFarland, and Xiang Ren. 2020. Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora. (2020). http://arxiv.org/abs/2010.06657

[4] Adrien Guille and Hakim Hacid. [n.d.]. Information diffusion in online social networks: A Survey. ([n. d.]).

[5] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, Gaël Varoquaux, Travis Vaught, and Jarrod Millman (Eds.). Pasadena, CA USA, 11–15.

[6] Daniel Kahneman. 2011. *Thinking Fast and Slow.* Macmillan.

[7] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. (2020), 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[8] Allison C. Morgan, Dimitrios J. Economou, Samuel F. Way, and Aaron Clauset. 2018. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* 7, 1 (2018). https://doi.org/10.1140/epjds/s13688-018-0166-4

[9] Sai Santosh Sasank Peri, Angela Liegey Dougall, Bodong Chen, and George Siemens. 2020. Towards understanding the lifespan and spread of ideas: Epidemiological modeling of participation on twitter. *ACM International Conference Proceeding Series* (2020), 197–202. https://doi.org/10.1145/3375462.3375515

[10] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the Temporal Dynamics of Diffusion Networks. (5 2011). http://arxiv.org/abs/1105.0697

[11] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. 2012. Structure and Dynamics of Information Pathways in Online Media. (12 2012). http://arxiv.org/abs/1212.1464

[12] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated Phrase Mining from Massive Text Corpora. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1825–1837. https://doi.org/10.1109/TKDE.2018.2812203

[13] David Soergel, A Saunders, and A McCallum. 2013. Open Scholarship and Peer Review: a Time for Experimentation. *Proceedings of the 30th International Conference on Machine Learning* 28 (2013). http://openreview.net/document/28cb8b58-d6f9-45c9-936f-c6c60e674381

Milestone Feedback :

I agree, use of Allen AI is good! Makes it also more relevant. - for prestige, you do not need to wait for Clauset et al. [8]. You can also simply use a well known ranking as a proxy to start. You can then see if your primary results change if you use a couple of other rankings and methods. e.g. for CORD-19 I would expect having a strong med school is also a useful indicator. - Sec 5.3.1 point (3) Linking Text Abstracts, ends with an incomplete sentence? - good progress, seems to be on track. looking forward to final results. some interesting case studies would be helpful! (so you should keep some time to interpret and understand the results!)

Propagation of Epidemics in Citation Networks

David Kartchner*
Rohit Mujumdar*
david.kartchner@gatech.edu
rohitmujumdar@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

Academic research *should* live or die by its own merits. However, human cognitive shortcuts have long believed to give undue advantage to particular institutions or researchers, sometimes blinding reviewers to errors of lack of rigor on their work. We investigate this imbalance by studying the spread of ideas across academic research networks using disease spread models adapted from epidemiology. We specifically focus on spread of ideas in the domain of Machine Learning (ML), a specialised area of research in Computer Science. We take the papers published in the year 2015 as our set of base "pathogens" and assess the network growth dynamics of idea spread amongst major ML conferences. We choose publication citations as our medium of propagation. We assess if ideas of the same calibre spread faster when originating from more prestigious institutions. We define data-centric measures for institutional prestige and idea quality and use both network-based simulation and statistical estimation to quantify how these factors affect idea propagation.ASIt also entails defining measures of and around idea propaZgation. We determine new ways of measuring these concepts and come up with novel metrics for the same. While we hope that these new measures and the subsequent network growth analysis can be instrumental in further research, we also hope that the work could be further helpful in correcting prestige-induced imbalance in academia.

KEYWORDS

epidemiology, heterogeneous information networks

ACM Reference Format:
David Kartchner and Rohit Mujumdar. 2020. Propagation of Epidemics in Citation Networks. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The acceptance and adoption of one's ideas is a key measure of influence and success in academia. Science is sometimes characterized as an ideal meritocracy, where ideas succeed or fail by their intrinsic merit and potential impact. However, the development of science is inseparably connected from scientists themselves, and recent research into human behavior has shown that humans are highly prone to biased judgements driven by cognitive shortcuts [4]. Since answering the question "Is this a high quality idea?" is difficult and mentally taxing, a natural response is to subconsciously substitute an easier question such as "Was this idea developed by at a prominent institution?" or "Is this researcher well-known in this field?" Such biasing, conscious or not, could lead to serious inequalities in publication and idea dissemination. This, in turn, could substantially shape the career trajectories of individuals and the development of science as a whole. Accordingly, we seek to combine epidemiological models for contagion (idea) spread in conjunction with large-scale empirical publication and affiliation data to answer the question "To what extent does institutional prestige lead to inequality in the spread and acceptance of ideas?"

*Both authors contributed equally to this research.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 PROBLEM DEFINITION

Assume that we are given a contact network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ of academic researchers, as well as a set of publication documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ with corresponding quality ratings $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$. Further suppose that we can calculate the "infectivity" of each idea by calibrating an SEIR model using empirical citation data over time. We investigate whether there is a significant difference in R_0 of ideas originating at prestigious universities after controlling idea quality for empirical differences in local contact network topology around the publishing researchers.

3 REVIEW OF LITERATURE

3.1 Idea spread on social networks

One of the earlier works in this area was [2], which tried to exploit the similarities between population dynamics in diffusion of Feynman diagrams and spread of infections. The key goal of the paper was applying several models such as SIR, SEI, SEIZ to empirical data in three, very different communities. While they concluded that suitably adapted models did a good job of fitting their data, they also emphasized that behavior of individuals may be very different for exposure to ideas and diseases. The strength of this paper lies determining how the events of idea diffusion are, or are not, analogous to population states commonly used in epidemiological models. We intend to borrow these analogies for our work. However, this paper applies the model to a very specific dataset in theoretical physics, on 3 countries, across several years post World-War 2. Their results are heavily biased because of the socio-political factors influencing the information spread dynamics of that era. With the technology of today, which exists several decades later

than the timeline of the information spread studied in this paper, and the speed, volume and diversity of research being conducted today, the features and parameters derived in the paper may not reflect the general characteristics of spread of ideas in today's times. Our work will assess characteristics of idea diffusion very relevant to the current ML research environment.

We also find relevant parallels in our proposed topic and the information diffusion modelling section of the comprehensive survey of information diffusion in online social networks, conducted in [3]. This paper considers nodes to be in only two states - activated or deactivated. The major idea that can be borrowed from here, to determine what constitutes as an infection in academic information spread, is that of *spreading cascade*, which is based on a simple assumption that people (nodes) in social networks are only influenced by actions taken by their connections. In academia too, the chances of someone getting influenced by an idea are high if their "connections", such as lab peers, publish work on that idea. This idea looks at information propagation as successive activation of nodes. The NETRATE [8] algorithms assume a static underlying network and formulate the exploration of correlation in nodes' infection times as maximum likelihood problems. The algorithm INFOPATH [9] instead, assumes a continuously changing network, and is noteworthy in its use of stochastic gradient descent to provide on-line estimates of the structure and temporal dynamics. Another idea that could be borrowed, is to be able to identify influential information spreaders, which in our case are the institutions we call prestigious or as having high calibre. We could borrow from the very popular HITS and PageRank algorithms for this.

Our work similarly draws from related work using epidemiological methods to model the spread of ideas in social networks. One recent work [7] shows that the way academic ideas spread at conferences can be modeled with SIR and SEIR models. Peri et al assume that each individual who posts with an official conference hashtag is "infected" and consider all twitter users who ever posted about the conference as "susceptible." However, their work fails to separate their proposed epidemiological models from the dynamics of conference attendance and event planning, which could potentially give rise to similar patterns.

Perhaps the most closely related work to our own is [6], which attempts to quantify whether university prestige leads to inequalities in the influence and perceived value of their research findings. To analyze this, the authors construct a faculty hiring network of the PhD-granting computer science departments in the U.S. and Canada by placing a directed edge (u, v) for each individual that received a doctorate at u and was in a tenured or tenure-track position at v in the 2011-2012 academic year. Each researcher was manually matched with his or her profile on DBLP, an online database spanning major CS journals and conferences. The authors model ideas as separate contagions that are spread across the network of institutions. Specifically, let each publication keyword be a particular contagion. We denote that institution i is "infected" at time t if i has at least 1 faculty member that publishes a paper with this keyword in the given time period.

The authors rank the "prestige" of each university by minimizing the number of rank violation in hiring decisions, i.e. the number of faculty that trained at less prestigious universities than those at which they were hired. In this network, institutional prestige was found to correlate with other institutional rankings (e.g. U.S. News and World Report), as well as other centrality measures such as degree, eigenvector, and closeness centrality.

After performing statistical analysis to confirm that faculty hiring does indeed lead to idea spread between departments, the authors simulate epidemics on the faculty hiring network using the SI model, allowing a particular edge to be used for transmission at most once. They assume that the transition probability p is a surrogate for the intrinsic merit of an idea and measure both the size of the infected population and the distance of the furthest infected node. Based on these simulations, the authors conclude that high-quality ideas will have far spread regardless of starting institution, but low quality ideas will spread significantly farther from prestigious institutions due entirely to the contact network structure. This conclusion still holds when relaxing the network structure to allow for random jumps, leading the authors to conclude that there is a structural advantage for researchers at prestigious institutions that leads to higher visibility and spread of ideas.

3.2 Models for citation network growth

Much of the current theory of idea spread and citation growth on networks has been has been heavily influenced by the Barabási-Albert model [1] of network growth known as preferential attachment. In this model, as new vertices (i.e. papers) are added to a network, they choose a fixed number of nodes to attach to, where the probability p_i of attaching to vertex i is $p_i = \frac{k_i}{\sum_i k_i}$[1]. This creates networks with a power law degree distribution, mirroring scale-free distribution observed in many real-world networks. Recent work has shown that when this model is combined with network growth, the result is a log-normal distribution of citation numbers, which more realistically describe real-world citation networks.

While these growth models may produce synthetic networks with similar statistical properties to real-world networks, they do not realistically model the transmission of ideas since researchers work in particular academic subfields. Accordingly, we seek to use heterogeneous networks composed of researchers, publication venues, institutions, and papers to more accurately capture how ideas spread between researchers and institutions, thus resulting in the generation of new papers. We hope that this will allow us to both more accurately characterize idea spread and accurately predict network growth dynamics.

4 PROJECT GOALS

The major goal of the project is to test the hypothesis that ideas of the same calibre spread faster when originating from prestigious institutions. Working towards this creates several sub-goals in the process. To complete our project, we intend to do the following:

- Bootstrap a time-dynamic academic contact network of researchers from existing heterogeneous publication and citation networks.
- Collect and integrate data on paper quality based on double-blind peer review scores and committee selection at top conferences.

[1] There are more complex formulations that allow for different forms of this attachment function to model nonlinear preferential attachment dynamics

- Develop a rigorous methodology to analyze empirical idea spread between researchers using existing data on citation, collaboration, paper timing, and paper topics.
- Quantify the extent to which institutional prestige leads to the spread of ideas after controlling for intrinsic quality and network effects.

We first start off by building a collaboration network of researchers and institutions based on paper citations and significant computational overlap of paper topics. We aim to develop new metrics to determine institution prestige and paper calibre and create an index around the metrics of idea propagation.

5 METHODS

5.1 Models

We base our work on analogies with concepts and models in epidemiology, modelling the spread of ideas on propagation of epidemics of infectious diseases. This involves drawing parallels between concepts and events in the spread of infectious diseases and spread of research in academia.

Following previous work [6], we let "contagions" be research ideas, denoted by papers with research topic keywords. We observe infections as a publication on a particular research topic within a time period t. We say that a paper A has infected paper B if B cites A and if the overlap of the domain and topics of A and B are very high, i.e. B has one of the same keywords as A. We further allow an individual to be in an "exposed" state, meaning that a researcher has seen and is incubating an idea but does not have an observed infection (publication) event yet. After a researcher has been dormant on a particular topic for a suitably long time, we assume that he or she returns to the susceptible pool. This is a realistic assumption since individuals tend publish repeatedly on topics throughout their careers. We therefore assume that retirement is the the only way a researcher is "removed" from the population of researchers.

5.2 Evaluation Measures

We adopt a variety of measures to gauge the strength and spread of ideas in our network. We begin with an evaluation of our network to characterize differences in network topology that may contribute to the spread of ideas, including the following for authors and institutions:

- Degree distribution
- Closeness centrality
- Betweenness centrality

We then evaluate different measures of idea propagation, including:

- Spread distance (maximum distance of idea spread)
- Number infected
- Infectivity constant β for a given idea
- R_0 for each idea, calculated from calibration of SEIRS model
- Idea longevity, calculated as the number of consecutive years that an idea has $R_0 > 0$

Finally, we will perform a statistical analysis of idea spread controlling for researcher centrality, publication venue, sub-field, and idea quality to determine if institutional prestige affects total idea spread.

5.3 Data

In order view the spread of ideas over time, we use the Open Academic Graph (OAG), which is a large knowledge graph unifying two, billion-scale academic graphs: Microsoft Academic Graph (MAG) [11], [10] and AMiner [5].

We specifically use the OAG v2 paper data for our work, which has the MAG November 2018 snapshot and AMiner January 2019 snapshot, and which amounts to 64GB of datasize. We intend to use papers published from the year 2015 onwards at the following major ML conferences - ICML, ICLR, NeurIPS and KDD. Thus, we need only a subset of the OAG data with a few fields for our task. Note that this will have to be done only once at the beginning of the project. We could use Microsoft Research's free REST-based Academic Knowledge API to run queries and collect the data we need. We could also collect the data by accessing the version uploaded to Google's BigQuery (https://github.com/ESHackathon/bigquery-oag). Another resort is to prepare an SQL Database of the specific data subset we require. The schema of this data of papers contains 23 fields, out of which the following subset of fields will be useful to us.

Table 1. Paper Data Schema (Selected Fields Only)

Field Name	Field Type	Description
id	string	paper ID
title	string	paper title
authors.name	string	author name
author.org	string	author affiliation
author.id	string	author ID
venue.raw	string	paper venue name
year	int	published year
keywords	list of strings	keywords
n_citation	int	citation number
abstract	string	abstract

Since our data is a combination of data from MAG and AMiner, there are several data entries which are common to both and are called as Linking Relations. Accordingly, the statistics of our data are delineated as follows:

Table 2. Paper Data Statistics

Data set	#Pairs/Papers	Date
Linking relations	91,137,597	2018.12
AMiner papers	172,209,563	2019.01
MAG papers	208,915,369	2018.11

The data is in JSON format. A truncated sample of the data looks like this:

```
id:          "53e9ab9eb7602d970354a97e"
title:       "Data mining: concepts and techniques"
authors:
  0:
    name:    "jiawei han"
    org:     "department of computer science university of illinois at urbana champaign"
year:        2000
keywords:
  0:         "data mining"
  1:         "structured data"
fos:
  0:         "relational database"
  1:         "data model"
n_citation:  29790
references:
  0:         "53e99ef4b7602d97027c2346"
  1:         "53e9aa23b7602d970338fb5e"
abstract:    "Our ability to generate .... and multi-relational data."
```


Figure 1. Truncated Dataset Sample

6 TIMELINE AND DIVISION OF LABOR

We have the following potential timeline :

(1) **By October 19** : Accessing OAG and building the data subset specific to our task
(2) **By November 2** :
    - Build collaboration network of researchers and institutions
    - Design and apply paper (idea) calibre metrics to data
    - Design and apply institution prestige/quality metrics to data
    - Define idea propagation index and assess its variance by data slices
(3) **By November 9** : Calibrate epidemiological models for papers from year in question
(4) **By November 16** : Statistically analyze the transmission probability as a function of institutional prestige
(5) **By November 23** :
    - Visualizations of methods and results
    - Prepare final report and presentation

We intend to contribute equally to the project work and have delineated the division as follows:

(1) Building data subset - Rohit
(2) Build collaboration network - Rohit & David
(3) Applying paper and institute quality metrics to data - Rohit
(4) Assessing idea propagation variance - David
(5) Calibrate EPI models - David & Rohit
(6) Transmission probability analysis - Rohit
(7) Visualization - David
(8) Milestone Report - Rohit & David
(9) Final Report and Presentation - David & Rohit

REFERENCES

[1] Albert László Barabási, Réka Albert, and Hawoong Jeong. 1999. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications* 272, 1 (10 1999), 173–187. https://doi.org/10.1016/S0378-4371(99)00291-5

[2] Luís M.A. Bettencourt, Ariel Cintrón-Arias, David I. Kaiser, and Carlos Castillo-Chávez. 2006. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications* 364 (5 2006), 513–536. https://doi.org/10.1016/j.physa.2005.08.083

[3] Adrien Guille and Hakim Hacid. [n.d.]. Information diffusion in online social networks: A Survey. ([n. d.]).

[4] Daniel Kahneman. 2011. *Thinking Fast and Slow.* Macmillan.

[5] Ying. Li, ACM Digital Library., Association for Computing Machinery. Special Interest Group on Knowledge Discovery & Data Mining., and Association for Computing Machinery. Special Interest Group on Management of Data. 2008. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM. 1102 pages.

[6] Allison C. Morgan, Dimitrios J. Economou, Samuel F. Way, and Aaron Clauset. 2018. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* 7, 1 (2018). https://doi.org/10.1140/epjds/s13688-018-0166-4

[7] Sai Santosh Sasank Peri, Angela Liegey Dougall, Bodong Chen, and George Siemens. 2020. Towards understanding the lifespan and spread of ideas: Epidemiological modeling of participation on twitter. *ACM International Conference Proceeding Series* (2020), 197–202. https://doi.org/10.1145/3375462.3375515

[8] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the Temporal Dynamics of Diffusion Networks. (5 2011). http://arxiv.org/abs/1105.0697

[9] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. 2012. Structure and Dynamics of Information Pathways in Online Media. (12 2012). http://arxiv.org/abs/1212.1464

[10] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (MAS) and applications. In *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web.* Association for Computing Machinery, Inc, 243–246. https://doi.org/10.1145/2740908.2742839

[11] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413. https://doi.org/10.1162/qss{_}a{_}00021

Proposal Feedback:

Let's remove the watermark, makes it harder to read.. - interesting proposal. I think you should try to make it also relevant to epi. How about you also try this out on a COVID related dataset? I recall OpenAI had released a large dataset of covid related papers. Some of these ideas can be tried on that (or atleast a portion of that). Feel free to focus on a few specific papers (like only those on forecasting etc).