Understanding Knowledge Transfer from Academia to Social Media through Concept Level Analysis

Eric Shang Nan Chen

Grade 12
BASIS International School Guangzhou
Guangzhou, China
eric.ericsn@outlook.com

***Abstract*—Advances in social network analysis, natural language processing tools, and the availability of large online datasets have led to the proliferation of social science research on information dynamics and knowledge transfer. However, there has been little research on the cross-domain transfer of knowledge from scientific research venues to the public masses. Existing studies have analyzed the adoption of research in policy through citations and citations in new patents, but few have analyzed mass media or social media because citations are sparse. The present paper contributes the novel direction of understanding knowledge transfer at the concept level instead of the document level using computational phrase mining techniques. Specifically, I analyze the transfer of COVID-19 research concepts by correlating linguistic and social network structure features to the popularity of a given research concept. Using AutoPhrase, a text segmentation algorithm, more than 120,000 concepts were derived, and of a small sample of concepts, 67.5% were found to be transferred to Twitter. Furthermore, I propose several solutions to the current limitations of this study for ongoing and future work.**

***Index Terms*—Computational Social Science, Sociology, Information Science**

I. Introduction

The present paper approaches the topic of academic knowledge transfer by examining the cross domain knowledge transfer of academic research to social media through novel methods of concept-level analysis. Whereas in organizational theory, knowledge transfer is defined as "the process through which one unit (e.g., group, department, or division) is affected by the experience of another" [1], knowledge transfer in the context of the present paper refers to the dissemination of new research topics and ideas to the public instead of between organizations.

In this study, a number of social and linguistic features are correlated with the level of public exposure of a research concept, defined as a phrase corresponding to a novel topic or idea in research. A better understanding of the linguistic factors that lead to certain research concepts appearing more in public discourse can better inform researchers on how to frame their research for public dissemination, whereas insight into social and structural factors such as author fame, prestige of publication venue, and network metrics such as centrality can help scholars understand how best to collaborate in research. Specifically, I conduct analysis on COVID-19 related research in the Allen Institute of AI's CORD-19 Data-set [2] and its dissemination on Twitter. The application to COVID-19 research is significant because public health emergencies often require new research to disseminate quickly to the public, and the social adoption of novel ideas influences the effectiveness of response efforts.

II. Background and Related Work

Knowledge transfer and diffusion is a topic of high interest to management science, sociology, and communications researchers, as understanding knowledge transfer is crucial to informing decisions regarding how organizations foster innovation and disseminate the results effectively. With advances in sociological network models, data science tools, and the availability of large data sets, knowledge diffusion is a field that now sits at the intersection of sociology, information science, and data science.

Most existing studies of knowledge transfer mostly focus on one domain, relying on traditional statistical methods, data collection on a few case studies, and formal models to investigate the influence of sociological factors such as tie strength and structural holes [3]–[5]. Although there are studies on cross-domain knowledge diffusion, they mostly address the connection between academia and industry or government through tracing citations in public documents or patents [6]–[8].

Seldom have researchers been able to study the transfer of knowledge from academia to the masses. Although attempts have been made to detect scientific knowledge diffusion trends on social media through tracking mentions of individual papers or user surveys, they are limited to a few documents and do not capture a holistic view [9], [10].

To address the aforementioned limitations, I utilize a novel concept-level approach instead of a citation study, meaning instead of tracking individual research papers, I extract the key phrases and concepts presented in academic papers using a phrasal segmentation algorithm. Similar analysis has been applied to the detection of scientific concepts in patents [11], but the present study attempts to extend such efforts to the public domain, namely social media. A phrase-mining, concept level approach has three major advantages, especially when applied to knowledge diffusion to the public. First, public discourse on Twitter seldom includes explicit citations to academic papers,

so citation tracing is difficult to accomplish. Second, from the public's perspective, the diffusion of individual academic papers is not important; rather, it is the emergence of new concepts that spurs public discourse. Third, a phrase-mining approach is more fine-grained than existing methods of topic modeling [12] because it extracts the exact phrasing of the concepts rather than broad topic clusters.

III. Process

The academic research data-set used in this study is Allen Institute of AI's CORD-19 data-set containing more than 750,000 papers related to COVID-19 research [2]. After performing exploratory analysis on the time series, coauthor network, and entropy of publication values, I extracted 123,247 concepts from the paper abstracts using Autophrase, an automated text segmentation algorithm that extracts key phrases with reference to knowledge bases such as Wikipedia [13].

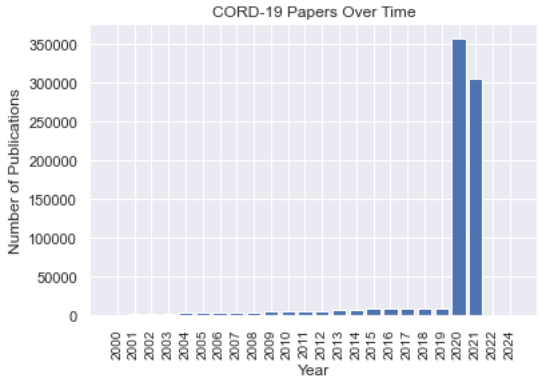

Fig. 1. Time series of CORD-19 papers

Of the returned concepts, most are background phrases such as "abdominal pain" and "atmospheric pollutant". To filter out these phrases, only phrases that first appeared after 2019, have a AutoPhrase quality score of over 0.85 out of 1.00, and have a frequency of more than five in the data-set were kept, resulting in 1,700 remaining concepts. Some examples of good research concepts in this list are "d614g mutation" and "3 chymotrypsin-like cysteine protease". Linguistic features such as the mean Dale Chall Readability Score [14] and VaderSentiment score [15] of all abstracts associated with a phrase, as well as social features such as information, betweenness, and degree centrality of the authors associated with a concept were calculated.

On the Twitter side, I utilize the Twitter Premium Full Archive Counts API to retrieve the number of Tweets associated with a concept in the year of 2020. Due to limited access to the API, only counts for 40 randomly selected concepts were returned. Finally, the counts are correlated with the previously calculated metrics for each concept using least squares regression.

IV. Results

Analysis of CORD-19's publication venues demonstrated that no journal ever represented more than 5% of the data-set and the normalized entropy of venues is consistently above 0.78, meaning CORD-19 is diverse and suitable for this project.

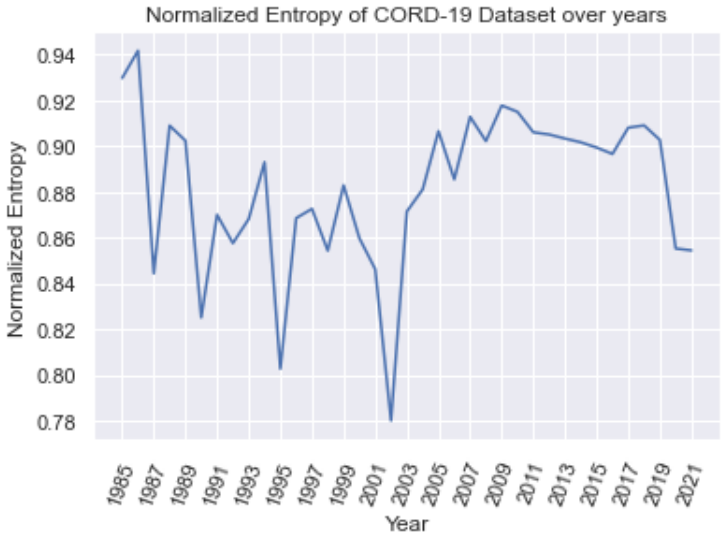

Fig. 2. Normalized entropy of venues over time

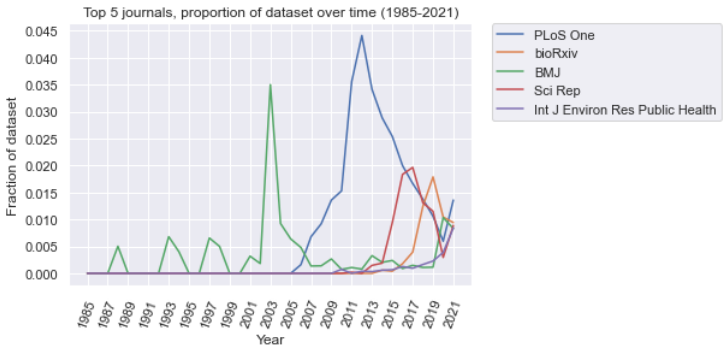

Fig. 3. Top 10 journals and their relative frequency over time

The coauthor graph has a clustering coefficient of 0.626, meaning the graph is moderately clustered and should be conducive to information flow.

Of the 40 randomly sampled concepts, 27 were transferred (more than 100 mentions on Twitter), amounting to 67.5%. So far, no statistically significant correlations (p less than 0.05) have been found between the independent features (linguistic and social) and frequency of appearance on Twitter. The lack of correlation is most likely due to the extremely limited sample size of the concepts used for the final analysis.

V. Future Work

The largest limitation of this study at the moment is the limited sample size of the Twitter frequency counts. However, beyond analyzing a larger sample of concepts, a necessary next step would be to curate a data-set of full Tweets and extract phrases from the Tweets via AutoPhrase to equate

the method of deriving concepts on both the academia and public side, thereby removing uncertainties introduced by Twitter's API. A larger sample size may reveal statistically significant correlations with linguistic and social features. Another inherent limitation of the phrase mining, concept level approach it the vagueness of the derived concepts. Without the context of the paper, many of the phrases do not sufficiently represent a complex research idea. Instead, the phrases derived seem more like topics. Future works should either complement phrase mining with other other methods to extract concepts or use a phrasal segmentation algorithm capable of detecting longer and more specific ideas.

Moreover, a temporal analysis of a select set of concepts and the users mentioning it on Twitter could reveal valuable insights on the underlying mechanisms and causes of knowledge transfer to the public domain. To go beyond measuring the exposure concepts, future studies may examine the attitudes and sentiment towards concepts.

Finally, as an extension of the current descriptive study would be to construct a formal sociological model for cross-domain knowledge transfer. The network model may be similar to that of information diffusion between two clustered communities, academia and Twitter, with connections being research venues that Tweet about findings, authors who self-promote their work on social media, and other Media coverage of scientific concepts on Twitter. Such a formal model would help researchers understand how to best present their findings to spread their ideas to the public.

REFERENCES

[1] L. Argote and P. Ingram, "Knowledge transfer: A basis for competitive advantage in firms," *Organizational behavior and human decision processes*, vol. 82, no. 1, pp. 150–169, 2000, publisher: Elsevier.

[2] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. M. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. A. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "CORD-19: The COVID-19 Open Research Dataset," *ArXiv*, 2020.

[3] Y.-A. de Montjoye, A. Stopczynski, E. Shmueli, A. Pentland, and S. Lehmann, "The Strength of the Strongest Ties in Collaborative Problem Solving," *Scientific Reports*, vol. 4, no. 1, p. 5277, May 2015. [Online]. Available: http://www.nature.com/articles/srep05277

[4] B. Hofstra, V. V. Kulkarni, S. Munoz-Najar Galvez, B. He, D. Jurafsky, and D. A. McFarland, "The Diversity–Innovation Paradox in Science," *Proceedings of the National Academy of Sciences*, vol. 117, no. 17, pp. 9284–9291, Apr. 2020. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1915378117

[5] R. S. Burt and others, "The social capital of structural holes," *The new economic sociology: Developments in an emerging field*, vol. 148, no. 90, p. 122, 2002.

[6] T. Hallett, O. Stapleton, and M. Sauder, "Public Ideas: Their Varieties and Careers," *American Sociological Review*, vol. 84, no. 3, pp. 545–576, Jun. 2019. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0003122419846628

[7] Y. Yin, J. Gao, B. F. Jones, and D. Wang, "Coevolution of policy and science during the pandemic," *Science*, vol. 371, no. 6525, pp. 128–130, Jan. 2021. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.abe3084

[8] M. Ahmadpoor and B. F. Jones, "The dual frontier: Patented inventions and prior scientific advance," *Science*, vol. 357, no. 6351, pp. 583–587, Aug. 2017. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aam9527

[9] J. P. Alperin, C. J. Gomez, and S. Haustein, "Identifying diffusion patterns of research articles on Twitter: A case study of online engagement with open access articles," *Public Understanding of Science*, vol. 28, no. 1, pp. 2–18, 2019, publisher: SAGE Publications Sage UK: London, England.

[10] E. Mohammadi, M. Thelwall, M. Kwasny, and K. L. Holmes, "Academic information on Twitter: A user survey," *PLOS ONE*, vol. 13, no. 5, p. e0197265, May 2018. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0197265

[11] H. Cao, M. Cheng, Z. Cen, D. A. McFarland, and X. Ren, "Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora," *arXiv:2010.06657 [cs]*, Oct. 2020, arXiv: 2010.06657. [Online]. Available: http://arxiv.org/abs/2010.06657

[12] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman, and C. Espina, "Classifying Twitter Topic-Networks Using Social Network Analysis," *Social Media + Society*, vol. 3, no. 1, p. 205630511769154, Jan. 2017. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2056305117691545

[13] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han, "Automated Phrase Mining from Massive Text Corpora," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1825–1837, Oct. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8306825/

[14] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948, publisher: JSTOR.

[15] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, 2014, issue: 1.