USING MACHINE LEARNING AND AI IN SCIENCE OF SCIENCE

Daniele Pretolesi[1], Andrea Vian[2] and Annalisa Barla[3]

[1] Austrian Institute of Technology, Vienna, Austria (e-mail: daniele.pretolesi@aic.ac.at)

[2] Department of Architecture and Design, University of Genoa (e-mail: andrea.vian@unige.it)

[3] Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa and Machine Learning Genoa Center, University of Genoa (e-mail: annalisa.barla@unige.it)

ABSTRACT: Complexity has become deeply ingrained in every aspect of our society, and navigating this complexity has become a pressing challenge. Science, in particular, is evolving at an unprecedented rate, constantly pushing the boundaries of human knowledge and understanding. In order to make sense of this rapid scientific evolution, science itself must adapt, employing systems that facilitate the interaction among researchers and that allow to grasp the interconnectedness and evolution of various interdisciplinary fields. In this endeavor, technologies such as natural language processing, network analytics, and machine learning play a pivotal role. These tools provide the essential support needed to analyze vast amounts of scientific data, extract meaningful insights, and uncover hidden patterns.

KEYWORDS: complexity, machine learning, science of science, keyword attribution

1 Introduction

Complexity permeates every facet of our existence, spanning from personal connections to global issues like pandemics and climate change. It is undeniable that comprehending and effectively dealing with complexity has become the paramount challenge of our current era and will continue to be so in the times to come. In scientific research, the intricacy of the subject matter compels researchers to venture beyond their familiar territory and actively pursue collaboration and expertise from scholars in diverse domains. Over the past few years, there has been a rise in scientific collaboration among researchers, leading to intricate interactions involving individuals operating within the same discipline, as well as from different fields of study. This paper is set in the context of the *Science of Science (SciSci)* framework, where the main aim is to

leverage the ever-increasing digital information on scientific production and AI-driven approaches to gain insight into the progress of science, the amount of scientific collaboration between researchers, and the degree of openness (Fortunato *et al.*, 2018).

2 Materials and Methods

We analyze Academic Collaboration Networks (ACNs), which are complex graphs of researchers' scientific output. Each publication in ACNs has important attributes like title, abstract, and keywords, indicating the research topic. By preserving the semantics of collaboration graphs, we connect the academic community and recommend research topics, works, and people. We use advanced technologies like natural language processing (NLP), network analytics (Barabási, 2013), and machine learning (ML)(Hastie *et al.*, 2009; Goodfellow *et al.*, 2016) to attribute missing keywords to publications, which improves our understanding of researchers' scientific interests. This enhanced representation helps us comprehend their scientific endeavors and areas of expertise.

To display the effectiveness of statistical and AI-based methods in this context, we consider the MaLGa dataset, that represents the scientific production of a large interdisciplinary group of scientists in the field of machine learning research, namely the Machine Learning Genoa Center (MaLGa - `https://malga.unige.it`). We collected data of the papers published by the MaLGa faculty members ($n = 14$) during the period 1984–2022 Among a total number of 624 publications, 341 papers are equipped with previously assigned keywords by venues or authors, and 573 papers with abstracts.

With the avilable data, we build a heterogeneous graph considering: (**P**) the *papers* published by MaLGa members together with co-authors, (**Y**) the *year* of paper publication, (**V**) the *venue* in which the paper was published and (**A**) *authors*, i.e. MaLGa members and their co-authors. We consider edges of the type **A**-**P**, **P**-**V**, and **P**-**Y**. The resulting graph is depicted in Fig. 1. The obtained ACN comprises 2007 total nodes, of which 1023 authors, 624 publications, 322 venues, and 38 years. The total amount of edges is 4098, with 2854 **A**-**P** connections, 620 **P**-**V**, and 624 **P**-**Y**.

3 Experimental results

We first use several keyword attribution techniques of increasing complexity to assign missing keywords to publications based on their title and abstract, then we exploit MetaPath2Vec Dong *et al.*, 2017 to represent the graph with

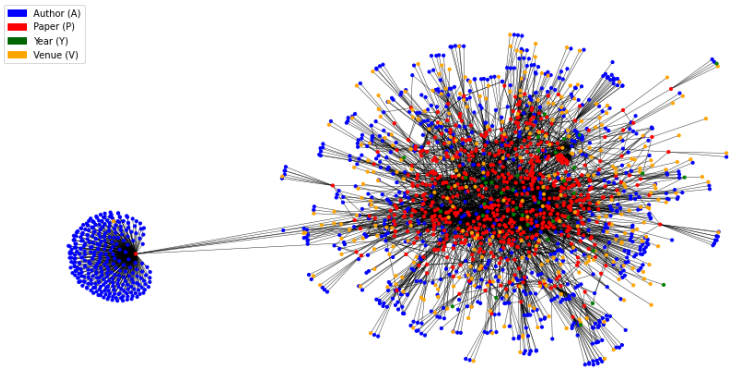

Figure 1. *MaLGa Collaboration as a heterogeneous graph. Blue nodes represent authors* ***A****, Red nodes represent publications* ***P****, green nodes are associated to years* ***Y****, and yellow nodes to venues* ***V****.*

and without keywords. We visualize such embedding on a two dimensional space and assess that this visualization is more informative than the one obtained from a keyword-less graph. We consider several keyword attribution methods: *n*-Grams, RAKE, BERT, SingleRank, TopicRank, MultipartiteRank, and YAKE. We evaluate their performance in terms of Recall, Precision, and F1-score against those ($n = 341$) papers that were associated with benchmark keywords, manually provided by authors or journals. Our results (data not shown) indicate that SingleRank keyword attribution method achieves the best performance for all three metrics considered. To visualise the impact of keywords as additional attributes of nodes, we consider the 14 faculty MaLGa members represented through the embedding with and without keywords and perform a PCA analysis, as shown in Fig. 2.

Adding the keywords into the graph representation improved the quality of the information stored in the network. Indeed, the visualisation clearly indicates that keywords improve the similarity among authors that share the same specific research interests. For example, if we consider M. Santacesaria[*] and G. S. Alberti[†], we note that using embedding that incorporates keywords significantly improve their similarity in terms of the distance measured in the projected space (right panel of Fig. 2, yellow and dark green dots).

[*]https://scholar.google.com/citations?user=iVlCw_gAAAAJ

[†]https://scholar.google.com/citations?user=boBf5cgAAAAJ

Figure 2. *PCA projections of MaLGa faculty members represented with embeddings with (right) and without (left) considering their keywords.*

4 Conclusion

In conclusion, our paper suggests how combining graph data structures, graph embeddings, NLP and ML techniques may provide valuable insights on complex topics. This integrated framework offers a holistic perspective by leveraging the strengths of each approach, enabling us to construct statistical models capable of accurately representing the intricate nature of the real world.

References

BARABÁSI, ALBERT-LÁSZLÓ. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**(1987), 20120375.

DONG, YUXIAO, CHAWLA, NITESH V, & SWAMI, ANANTHRAM. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. *Pages 135–144 of: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*.

FORTUNATO, SANTO, BERGSTROM, CARL T, BÖRNER, KATY, EVANS, JAMES A, HELBING, DIRK, MILOJEVIĆ, STAŠA, PETERSEN, ALEXANDER M, RADICCHI, FILIPPO, SINATRA, ROBERTA, UZZI, BRIAN, *et al.* 2018. Science of science. *Science*, **359**(6379), eaao0185.

GOODFELLOW, IAN, BENGIO, YOSHUA, & COURVILLE, AARON. 2016. *Deep learning*. MIT press.

HASTIE, TREVOR, TIBSHIRANI, ROBERT, FRIEDMAN, JEROME H, & FRIEDMAN, JEROME H. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.