**Understanding the Fast Changing Academic Eco System using Microsoft Academic**

Michael Fire and Carlos Guestrin

University of Washington

**Abstract:** *The academic world changed dramatically over the last decades, with ever growing number of publication each year, and changing publication patterns. In this study, we suggest utilizing recent advantages in data science tools and algorithms, combined with extensive cloud computing resources to analyze the massive Microsoft Academic Graph (#academicgraph). Our main research goal is to understand how the large-scale co-authors network and citation network evolve over the last century, and especially in the last two decades. Additionally, we plan to develop algorithms to identify unwelcome scientific behaviors, such as H-index inflating and ghost authorship. Our premillennial results indicate that our research can assist to more precisely understand how the academic eco system advanced. Moreover, we hope that using machine learning algorithms can help to reveal characteristic of high impact papers.*

In the last few decades, the academic world change rapidly. Today, it is easier than ever before to *share* publications using preprints repositories, like the [arXiv](https://arxiv.org/) and [bioRxiv](http://biorxiv.org/). Furthermore, mega journals, such as [PlosOne](http://journals.plos.org/plosone/) and [Nature Scientific Reports](https://www.nature.com/srep/), which publish tens of thousands of papers each year, make it *easier and faster to publish* tens of thousands of papers each year. In addition, website such as Microsoft Academic [1] and [Google Scholar](http://scholar.google.com/), make it easier *to search and find* articles. Moreover, both the numbers of new researchers and the number of new journals increased sharply in the last two decades.

Recent studies observed that unquestionably the academic world is dramatically changing, Nowadays, researchers have the tendency to publish more [2] , to cite other research more [3], and to internationally collaborate more [4]. However, the changing academic world created several undesirable phenomena. Due to the increasing pressure upon researchers, especially early career researchers [5], to publish more high impact papers in top ranked journals, (i.e., “publish or perish”) [5], there is increase in number of unwelcome research behaviors, such as scientific salami slicing [6], p-value hacking [7], H-Index inflating [8], ghost authorship [9] , and more.

In our research, we plan to utilize the many recent advantage in data science tools [10]–[12] combined with extensive cloud computing resources to analyze the massive Microsoft Academic datasets. Our research goals are fivefold: First, we intend to utilize state-of-the-art data science tools to precisely analyze papers publication patterns over time. Second, we plan to analyze the evolution of the co-authorship network, which consists of more than 117 million researchers, and more 6.7 billion edges, and to analyze the citations complex network consisted of 126 million papers and over half a billion links. Our premillennialanalysis results indicate that using new set of algorithms for analyzing large graphs can reveal new insights on how various novel insights on the way various science domains evolve over long period of time. Third, we hope to identify how prevalent is the described above unwelcomed behavior become in recent years. Similar to our previous studies [13]–[15] who hope to develop new algorithms to identify unwelcome behaviors. Fourth, we will utilize various machine learning tools to identify interesting patterns in the dataset that are results from the recent changes in publication trends. Lastly, we hope to uncover what are the characteristics that make high impact paper.

By analyzing the Microsoft Academic Graphthat was published as part of [2016 KDD Challenge](https://kddcup2016.azurewebsites.net/Data), we already succeeded to demonstrate that with the right data science tools, it possible to better understand how co-authorship networks of various research domains evolved over time. We hope that with the assistance Microsoft Azure for Research Award, we will have the resources to analyze large-scale networks, and data in to better understand how science evolve over the last centuries. This in turn can help to uncover how to produce a better science.

**References:**

[1] A. Sinha *et al.*, “An Overview of Microsoft Academic Service (MAS) and Applications,” in *Proceedings of the 24th International Conference on World Wide Web - WWW ’15 Companion*, 2015, pp. 243–246.

[2] R. Van Noorden, “Global scientific output doubles every nine years,” *Nature*, 2014. [Online]. Available: http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html. [Accessed: 12-Jun-2017].

[3] L. Bornmann and R. Mutz, “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references.”

[4] J. Adams, “Collaborations: The rise of research networks.,” *Nature*, vol. 490, no. 7420, pp. 335–6, Oct. 2012.

[5] M. A. Edwards and S. Roy, “Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition.”

[6] “The cost of salami slicing,” *Nat. Mater.*, vol. 4, no. 1, pp. 1–1, Jan. 2005.

[7] D. Chavalarias, J. D. Wallach, A. H. T. Li, J. P. A. Ioannidis, and J. MD, “Evolution of Reporting *P* Values in the Biomedical Literature, 1990-2015,” *JAMA*, vol. 315, no. 11, p. 1141, Mar. 2016.

[8] C. Bartneck and S. Kokkelmans, “Detecting h-index manipulation through self-citation analysis,” *Scientometrics*, 2010.

[9] G. Mowatt *et al.*, “Prevalence of Honorary and Ghost Authorship in Cochrane Reviews,” *JAMA*, vol. 287, no. 21, p. 2769, Jun. 2002.

[10] Y. Low, J. E. Gonzalez, A. Kyrola, D. Bickson, C. E. Guestrin, and J. Hellerstein, “GraphLab: A New Framework For Parallel Machine Learning,” Aug. 2014.

[11] X. Meng *et al.*, “MLlib: Machine Learning in Apache Spark,” *J. Mach. Learn. Res.*, vol. 17, pp. 1–7, 2016.

[12] M. Bostock, V. Ogievetsky, and J. Heer, “D3 Data-Driven Documents,” *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

[13] D. Kagan, Y. Elovici, and M. Fire, “Generic Anomalous Vertices Detection Utilizing a Link Prediction Algorithm,” Oct. 2016.

[14] M. Fire, G. Katz, and Y. Elovici, “Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies,” *homes.cs.washington.edu*.

[15] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, “Friend or foe? Fake profile identification in online social networks,” *Soc. Netw. Anal. Min.*, vol. 4, no. 1, p. 194, Dec. 2014.