The concept of predatory publishing is complex and often controversial. There are numerous – often similar – definitions of predatory publishing (e.g., Clark and Smith, 2015; Grudniewicz et al., 2019). However, these definitions are broad and rely on subjective, normative judgments of best practices and professional ethics in academic publishing. In part due to the subjectivity involved in many definitions of predatory publishing, debates about predatory publishing are often highly contentious. Different scholars and academic stakeholders may have different philosophies and beliefs regarding the appropriateness and legitimacy of varying publishing practices. Our research provides empirical data on non-indexed publishers to enable scholars and academic stakeholders to make informed decisions about the legitimacy – or lack thereof – of journals and publishers, regardless of professional preferences and philosophies.

The opacity of peer review in most academic journals makes it difficult to directly observe the legitimacy and quality of peer review. Although manuscript development and gatekeeping processes vary between different journals, it is often difficult to know what happens in the "black box" of peer review. This challenge is compounded with predatory or questionable academic journals, as such publishers often operate in covert and/or deceptive manners. Further, predatory journals and publishers are rarely indexed by institutions like the Web of Science, which makes surveilling and analyzing such journals difficult. However, document data and metadata can provide empirical evidence about the professionalism and operating procedures of questionable academic journals. Using a variety of web scraping techniques, we developed a database of non-indexed academic publishers, which includes numerous publishers on the Cabells Predatory Reports list. This database – named *The Lacuna Database* – is currently comprised of over 1,000,000 articles and chronicles the scholarly work published through in journals.

Recently, scholars have used computational analyses of large-scale data to reveal fraud, misinformation and questionable research practices in academic publishing (Bik, Casadevall and Fang, 2016; Petersen, 2019; Bishop, 2020). Our research follows this work. Using the Lacuna Database, we conducted empirical analyses of metadata of the most prominent publishers included on the Cabells Predatory Reports list. For additional comparison, we also included three prominent – but often controversial – Open Access-only publishers that are not on the Cabells list. However, the three publishers have also – fairly or not – historically faced charges of predation or illegitimacy: Frontiers, Hindawi and MDPI. Siler (2020) labelled those three publishers as *grey publishers* of uncertain and/or contested status.

The metadata included in the Lacuna Database can be used to empirically identify characteristics of publishers widely seen as predatory, or merely borderline.

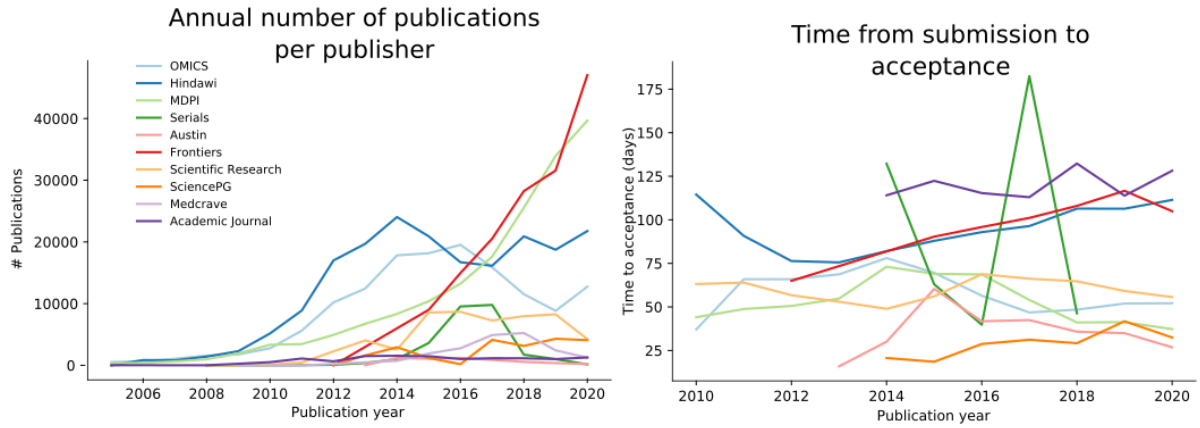

Figure 1 – Publications, Peer Review Durations of Various Open Access Publishers

Figure 1 shows annual growth levels of publishers of varying degrees of status and legitimacy. Notably, MDPI and Frontiers exhibited exponential growth in the 2010s, while Hindawi kept publishing values relatively constant since 2013. The left side of the figure shows publisher strategies regarding growth. The right side of the figure shows contrasts in average time from submission to publication across various publishers. There is considerable variation in the time publishers take to develop, gatekeep and publish articles. Three publishers (Hindawi, Frontiers, Academic journals) cluster around 110-130 days, while the remainder cluster around 30-60 days. As a comparison point to the publishers visualized in Figure 1, submission to publication times are 162-178 days at *PLOS ONE* (PLOS, 2021).

Publication speed is not necessarily deterministic of the quality of a journal or its peer review processes. However, both excessively fast or slow publication times are problematic, albeit for different reasons. For example, concerns have emerged regarding expedited peer review in COVID-19 research (Else, 2020). Many questionable or predatory publishers compete on speed; very rapid peer review attracts submissions for predatory publishers. Professional standards regarding optimal or acceptable peer review times are normative. Our research provides data to inform those normative decisions. For example, is MDPI's publication speed – which is similar to that of clearly predatory publishers such as OMICS – a well-earned competitive advantage achieved via technical and logistical innovations, or is it indicative of haphazard peer review? Is the gradual slowing of Frontiers manuscripts over time indicative of increasingly arduous peer review or a negative consequence of the publisher's exponential growth? The Lacuna Database will provide data to allow scholars and academic stakeholders make more-informed decisions about different academic publishers.

These analyses represent the tip of the proverbial iceberg with the Lacuna Database. In the future, we will analyze the demographics of publishing scholars (country, institutional status) in varying questionable journals. Textual analyses can also provide empirical evidence of publishing quality and legitimacy. Misspellings and typos in affiliations (e.g., countries, universities), as well as topical and textual properties of manuscripts can provide empirical indices of publishing quality and legitimacy. A philosophy underpinning The Lacuna Database is that "sunshine is the best disinfectant." Empirical analyses of questionable publishers can expose poor professional practices, but can also 'exonerate' publishers and journals unfairly under suspicion. It is expected that the first version of the Lacuna Database will be made public in late 2020, to enable researchers and academic stakeholders conduct large-scale research on questionable publishers.

WORKS CITED

Bik, Elisabeth M., Arturo Casadevall, Ferric C. Fang. 2016. "The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications." *mBio*, 7(3): e00809-16.

Bishop, Dorothy. 2020. "Percent by most prolific author score: a red flag for possible editorial bias." http://deevybee.blogspot.com/2020/07/percent-by-most-prolific-author-score.html

Clark, Jocalyn and Richard Smith. 2015. "Firm action needed on predatory journals." *The BMJ*, 350:h210.

Else, Holly. 2020. "How a torrent of COVID science changed research publishing – in seven charts." https://www.nature.com/articles/d41586-020-03564-y

Grudniewicz, Agnes et al. 2019. "Predatory journals: no definition, no defence." *Nature*, 576: 210-212.

Petersen, Alexander M. 2019. "Megajournal mismanagement: Manuscript decision bias and anomalous editor activity at PLOS ONE." *Journal of Informetrics*, 13(4): 1-22.

PLOS. 2021. "Journal Information." https://journals.plos.org/plosone/s/journal-information

Sciencemag. 2021. "Journal Metrics Overview" https://www.sciencemag.org/journal-metrics

Siler, Kyle. 2020. "Demarcating Spectrums of Predatory Publishing: Economic and Institutional Sources of Academic Legitimacy." *JASIST*, 71(11): 1386-1401.