Private Companies and Scholarly Infrastructure – Google Scholar and Academic Autonomy

By Jake Goldenfein (Cornell Tech), Sebastian Benthall (NYU), Daniel Griffin (UC Berkeley) and Eran Toch (Tel Aviv University)

Private Companies and Scholarly Infrastructure – Google Scholar and Academic Autonomy

By Jake Goldenfein (Cornell Tech), Sebastian Benthall (NYU), Daniel Griffin (UC Berkeley) and Eran Toch (Tel Aviv University)

The academic funding scandals plaguing 2019 have highlighted some of the more problematic dynamics between tech industry money and academia (see e.g. Williams 2019, Orlowski 2017). But the tech industry's deeper impacts on academia and knowledge production actually stem from the entirely *non-scandalous* relationships between technology firms and academic institutions. Industry support heavily subsidizes academic work. That support comes in the form of direct funding for departments, centers, scholars, and events, but also through the provision of academic *infrastructures* like communications platforms, computational resources, and research tools. In light of the reality that infrastructures are themselves political, it is imperative to unpack the political dimensions of scholarly infrastructures provided by big technology firms, and question whether they might problematically impact knowledge production and the academic field more broadly.

Specifically, we have been studying Google Scholar and its growing centrality in academic search, evaluative bibliometrics, and scholar profiling. Leveraging Google's massive investment in user experience, cloud infrastructures, and search algorithms, Google Scholar 'search' was introduced in 2004, citation counting in 2006, and scholar profiles (as part of 'Google Scholar Citations') in 2011. These have since become critical scholarly tools for both conducting and evaluating research. While there is a great deal of research on how Google Scholar works, for instance, in terms of how it is different from other academic search and bibliometrics services (Jacsó 2005; Falagas et al., 2008; Adriaanse, L. S., & Rensleigh, C., 2013), the excellent work on, for instance, the politics of search engines (Itrona and Nissenbaum, 2006; Rogers, 2004) and search accountability (Grimmelman, 2010; Bracha and Pasquale, 2008) has not been extended into the scholarly domain. We're not suggesting that Google Scholar is actively manipulating search results or citation counts. Rather, we're suggesting that there are political and ethical consequences to Google Scholar's centrality that stem simply from how the technology works. Google Scholar shifts, disrupts, undermines, or otherwise alters the conventions and norms of academic work (following on work by e.g. Jamali & Asadi, 2010; Vaidhyanathan, 2012, Kemman et al., 2013), and those changes must be evaluated for their consequences on knowledge production. Unfortunately, understanding the political dimensions of how Google Scholar works is extremely difficult, because its automated systems are entirely opaque and non-accountable.

Citation indexing and evaluative bibliometrics

Although Google Scholar is relatively new, the history of the organization of scientific information offers a useful lens for analyzing its impacts on academia. The development of 'citation indexing' in the 1950s – the coordination of scientific knowledge around references – was a revolutionary development in information science. The massive increase of scientific output after World War II had overwhelmed the capacities of libraries to abstract and index work. 'Subject indexing' – indexing scholarly material according to its content – required human readers, and could take up to 12 months per article. Too much scientific output and not enough scientific organization thus prompted investigation into new mechanisms for *automating* the indexing of scholarly information.

Several punch-card based subject indexing systems were developed to address the problem, but these were fundamentally limited because they required pre-defined dictionaries for each discipline. The effect was to silo research into disciplines and sub-disciplines aware of the relevant subject dictionaries. However, in 1955, while working at John Hopkins medical library, information scientist Eugene Garfield came up with the idea of indexing articles by their citations rather than content. Not only was this approach far more amenable to automation by punch card and tabulation machine, but it also facilitated interdisciplinary research tools, premised on indexing an article's 'association-of-ideas' (Garfield, 1955).

It turned out, however, that Garfield's invention was useful for more than merely indexing scholarly information. Tracking how many times a particular work was referenced also generated a useful measure for the *quality* of work – a way to track its reception in the field. This was critical because rapidly growing research output had also overwhelmed the publishing capacities of scholarly societies, the traditional publishers of academic work, leading to a new publishing ecosystem to emerge that included commercial publishers and a vast proliferation of journals. The growing output of scholarly output meant new systems of fast evaluation were needed for researchers struggling to keep up with the literature, and research managers struggling to evaluate the growing number of scholars. The world of academic publishing was thus primed for the introduction of a unit of value that could rationalize and systematize the emerging competitive market for scholarly publishing.

Garfield took advantage of the new environment and commercialized his citation indexing system into a metric for journal evaluation called Journal Impact Factor (**JIF**). JIF estimated the quality of a journal by tracking the number of citations its publications received over a specific time frame. Alongside measuring the quality of a journal, it also became a way to measure the prestige of a scholar (whether they could publish in high-impact journals), a mechanism for establishing the price of a journal subscription (prices were determined by impact factor in the competitive market), and even a way to evaluate the productivity of an institution or research center (by measuring cost-per-citation) (Cameron, 2005). JIF thus become

the common unit of value linking scholarly prestige, the economic capital of publishers, and the funding imperatives of institutions (Fyfe et al., 2017).

The centrality of evaluative bibliometrics, and particularly JIF, in academic publishing is now recognized as a major problem for the academic field. It generates incentives contrary to academic ideals (Alberts 2013, European Commission, 2019), such as the generation of highly citable 'review' style literature instead of genuinely innovative work, as well as a problematic chasing and gaming of metrics. This had led some to argue for abandoning metrics for academic evaluation entirely, or at least moderating their centrality (Priem et al., 2010; Hicks, 2015). And while proper *human* evaluation is always going to be more meaningful than metrics related to JIF, or even newer metrics associated with social media and networked communication such as views, downloads, likes, retweets, or raw citation counts, there is also always more literature than scholars can keep up with and evaluate, and more scholars than research managers can keep up with and evaluate. That is especially so in the context of interdisciplinary work where the norms for evaluating the quality of people and their work are less clear.

Those realities mean metrics have become necessary tools in many academic contexts, and research managers will candidly tell you that sometimes the first thing they look at when considering a scholar for a position is their metrics (typically in the form of their Google Scholar profile). In fact, metrics, particularly bibliometrics based on citation count, are only growing in importance. Not only does citation count help evaluate the quality of a journal (and by *de facto* the researchers publishing in it) as with JIF, but enable the direct evaluation of scholars with, for instance Google Scholar Profiles, while also establishing what works are seen and read. For instance, Google Scholar (as well as other academic search engines) uses an article's citation count as the primary way to evaluate its 'relevance' to an academic search query (Rovira et al., 2019). The more citations (over a particular time period), the higher a work will rank in search results. Citations in this context are more than simply acknowledgements of intellectual lineage, or even mechanisms for evaluating research, researchers, journals, or institutions. They are recursive operators in a network that feeds back into the visibility of scholarly work.

Unfortunately, the precise mechanisms by which Google Scholar's citation counts are generated are unclear. The algorithms that build the Google Scholar index (i.e. determine what work is scholarly), parse documents for their references (i.e. extract bibliometric information), and rank academic works according to relevance, are all opaque, and Google offers effectively zero accountability for the operation of those systems. To that end, our work is not a critique of metrics, but rather an evaluation of the consequences to academia and knowledge production of evaluative bibliometrics being controlled primarily by Google rather than academic publishers.

Google Scholar has a remarkably usable interface for search, bibliometrics, and scholar profiling. It returns good results, quickly. It indexes more work, and returns higher citation counts than other services. Using Google Scholar is thus incredibly easy and powerful, as well as free, whereas access to Web of Science – the bibliometrics services associated with academic publishing and JIF – is not. It has accordingly become an absolutely central scholarly infrastructure. However, we argue that Google Scholar's non-transparent, non-accountable, but totally dominating system of bibliometrics introduces issues that the academic field should no longer simply accept as a trade-off for efficiency or usability.

Google Scholar's control of academic metrics

Academics, research managers at universities, and research policy makers rely on technology platforms offered by private companies routinely in academic work and the evaluation of scholarship. While these services have been provided by commercial actors for some time, and are implicated in the deeply problematic academic publishing industry, the role and motivation of commercial academic publishers in the scholarly ecosystem is relatively clear. Their metrics services, such a JIF, are quite transparent. Google Scholar's bibliometrics, on the other hand, are opaque. Traditional publishers actively police metrics for manipulation (typically self-citation), and introduce punishments like suspending journals from impact factor reporting. Google Scholar, on the other hand, neither monitors, nor facilitates contestation of its metrics. Responsibility is instead displaced onto publishers and repositories, with the suggestion that they ensure their publishing metadata is provided in a format that Google systems can parse. Google further absolves itself of responsibility by arguing that Google Scholar is free, with only a small number of employees, claiming that Google Scholar is not a Google 'product'. Nonetheless, we suggest there are several reasons academics should be wary of this lack of accountability.

1. *Google Scholar is changing research and research evaluation*

In our research, we are finding that Google (through Google Scholar) is playing an increasingly significant role in establishing what knowledge looks like today. Its technological configuration influences what gets seen, read, cited by scholars, and transmitted to students. There is also a growing body of empirical work describing how the use of Google Scholar is affecting both research practices, and how research is evaluated. However, with the Google Scholar system, many of the old gatekeepers of academic quality, such as peer-review, are being undermined. Google's scholar 'scholarliness' algorithm indexes a great deal of material from outside of traditional scholarly hierarchies. This might increase the visibility of certain types of work, but may also be detrimental to knowledge production generally. In particular, it changes the goalposts for how academic work is evaluated but without revealing how. Peer-review may be flawed, but at least we know how it works. And because of the growing centrality of Google Scholar bibliometrics in

evaluative processes like hiring and funding, universities seem to be accepting these changes *de facto* without paying enough attention to the true cost.

Other effects on academia also need close attention. For instance, the way Google Scholar's relevance algorithm works appears to be shifting the distribution of knowledge and 'impact' between disciplines. Because it does not pay attention to the specificities of different academic fields, it privileges physical sciences and computer science over social sciences and humanities in terms of search results and citation counts. This is a feature of how the technology functions in combination with the different citation and authorship conventions in STEM fields. Google Scholar therefore participates in a broader paradigm shift towards computer science appearing as the discipline around which all knowledge is organized. Citation count comparison across an ontologically flat scholarly network contributes intellectually to disciplinary collapse (Benthall, 2015), while establishing the primacy of computer engineering as a matter of sociotechnical fact.

2. *Platform economics*

Google scholar is free and does not serve advertising (although advertising does appear to have been one of the service's original intended purposes), presenting the illusion that it has no commercial dimension. However, if you want to edit your Google Scholar profile you need a Google account. That means agreeing to Google's terms and conditions, including those related to data collection and processing. Unlike other Google services in academia that are often governed by *sui generis* 'G-Suite' terms and conditions that universities can negotiate, Google Scholar users typically submit to relatively unmitigated consumer surveillance. That is, users of Google Scholar are having their data gathered and aggregated with ordinary Google profiles, to participate in the two-sided data market. Platform economics (Xu, J., & Zhang, X. J., 2006; Eisenmann et al., 2011) thus enters into the academic environment, and academic work (i.e. the practices of doing, finding, accessing, evaluating, and sharing research) becomes part of a system of surveillance, insights, and behavioral advertising.

Associated with this shifting of research tools onto technology platforms is the centralization of academic infrastructures in Google services. Increased centralization, for instance with universities using Google for communications and computational tools, as well as research and evaluation infrastructure, makes the academy more dependent on Google, more vulnerable to its dominance, and less capable of building its own tools.

3. *Academic autonomy*

Alongside dismantling quality review mechanisms and 'platformizing' academic work, we argue that because Google Scholar's tools are not transparent or testable, they further risk undermining academic autonomy. For some thinkers, organizing the academic field and its hierarchies through systems of evaluation that academics establish for themselves, that can be tweaked and tested, is the basis of scientific objectivity (Bourdieu, 2004). While many academics have not given up on building research and evaluation tools for academia, the centrality of Google Scholar is not abating, and there seems to be acceptance of whatever Google offers, simply because of usability. It is imperative however, to reflect on whether using the services that technology firms offer for free might simultaneously undermine the independence required of scholarship, because the systems used to evaluate scholarship and scholars are controlled by external entities, using mechanisms that cannot be properly examined. To that end, the academic field must ask itself whether it is acceptable that the currency for evaluating academic work and workers, that simultaneously determines the visibility of research, is entirely shaped and controlled by Google, in a non-accountable way.

4. *Accountability and political identity*

One solution to all this might be better accountability for Google Scholar. However, because Google Scholar is free and not a typical Google 'product', accountability provided by Google is more or less non-existent. However, we suggest greater accountability is critical, on one hand, because of Google Scholar's consequences for academic work, and on the other hand, because of Google's political identity. One thing that recent academic funding scandals have made very clear is that Google is a stakeholder in contemporary debates about the role of data science and platform economics in society. Google also uses its position and capital to influence technology policy. While Google may not fund to directly intervene in academic debates, it clearly supports academics whose work is sympathetic to its own interests.

For those of us working on issues of technology and society, this reality generates urgent questions about the centrality of Google Scholar. When we know that Google actively influences policy, why do we feel so comfortable using its infrastructure, especially given its opacity and lack of accountability? For the reasons discussed above, including the reality that Google's is not a disinterested actor in the academic world, this lack of accountability is a serious concern. If we are not, however, able to make Google available for accounting commensurate with the far-reaching effects of Google Scholar services, perhaps the answer is to emphasize the importance and necessity of scholars being accountable to themselves for how they use them. In other words, scholars and research managers may need to reflect on their reliance on Google services, and consider what they are giving up in exchange for user experience.

References

Adriaanse, L. S., & Rensleigh, C. (2013). Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, *31*(6), 727-744.

Alberts, B. (2013) Impact Factor Distorions. *Science, 340*(6134), 787.

Benthall, S. (2015). Designing networked publics for communicative action. *Interface*, *1*(1), 3.

Bourdieu, P. (2004). *Science of science and reflexivity*. Polity.

Bracha, O. Pasquale, F. (2008). Federal Search Commission? Access, Fairness, and Accountability in the Law of Search. *Cornell Law Review*, 93, 1149-1210.

Cameron, B. D. (2005). Trends in the usage of ISI bibliometric data: Uses, abuses, and implications. *portal: Libraries and the Academy*, *5*(1), 105-125.

Eisenmann, T., Parker, G., & Van Alstyne, M. (2011). Platform envelopment. *Strategic Management Journal*, *32*(12), 1270-1285.

European Commission. (2019). Future of Scholarly Publishing and Scholarly Communication: Report of the Expert Group to the European Commission.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB journal*, *22*(2), 338-342.

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, 15 July 1955: 108-111.

Grimmelman, J. (2010). Some Skepticism About Search Neutrality. Szoka and Markus (eds) *The Next Digital Decade: Essays on the Future of the Internet*. Tech Freedom.

Hicks, D. et al. (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature, 520,* 23 April 2015: 429-431.

Itrona, L. Nissenbaum, H. (2006), Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society,* 16(3) 169-195.

Kemman, M., Kleppe, M., & Scagliola, S. (2013). Just Google It - Digital Research Practices of Humanities Scholars. In: Clare Mills, Michael Pidd and Esther Ward. Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities. Sheffield: The Digital Humanities Institute, 2014.

Fyfe, Al. et al. (2017). Untangling Academic Publishing: A History of the Relationship between commercial interests, academic prestige, and the circulation of research (https://doi.org/10.5281/zenodo.5461)

Jacsó, P. (2005). Google Scholar: the pros and the cons. *Online information review*, *29*(2), 208-214.

Jamali, H. R., & Asadi, S. (2010). Google and the scholar: the role of Google in scientists' information-seeking behaviour. *Online information review*, *34*(2), 282-294.

Orlowski, A. (2017). Academics “funded” by Google tend not to mention it in their work. *The Register*, 13 July 2017 <https://www.theregister.co.uk/2017/07/13/google_transparency_report_academics/>.

J. Priem, D. Taraborelli, P. Groth, C. Neylon (2010). Altmetrics: A manifesto, 26 October 2010. http://altmetrics.org/manifesto.

Rogers, R. (2004). *Information Politics on the Web*. MIT Press.

Rovira, C. Codinea, L. Guerrero-Solé, F. Lopezosa, C. (2019). Ranking by Relevance and Citation Counts, a Comparative Study: Google Scholar, Microsoft Academic, WoS and Scopus. *Future Internet*, *11(9)*, 202.

Vaidhyanathan, S. (2012). *The Googlization of everything:(and why we should worry)*. Univ of California Press.

Williams, O. (2019). How Big Tech funds the debate on AI Ethics. *New Statesman America*, 6 June 2019 < https://www.newstatesman.com/science-tech/technology/2019/06/how-big-tech-funds-debate-ai-ethics>.

Xu, J., & Zhang, X. J. (2006). Survey on Platform Economics [J]. *China Industrial Economy*, *5*.

✉ **Jake Goldenfein**
Cornell Tech
jg2323@cornell.edu

✉ **Sebastian Benthall**
New York University
spb413@nyu.edu

✉ **Daniel Griffin**
UC Berkeley
daniel.griffin@ischool.berkeley.edu

✉ **Eran Toch**
Tel Aviv University
erant@post.tau.ac.il