ARTICLE

Microsoft Academic Graph: When experts are not enough

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia

Microsoft Research, Redmond, WA, 98052, USA

Keywords: citation networks, eigenvector centrality measure, knowledge graph, research assessments, saliency ranking, scholarly database

ABSTRACT

An ongoing project explores the extent to which artificial intelligence (AI), specifically in the areas of natural language processing and semantic reasoning, can be exploited to facilitate the studies of science by deploying software agents equipped with natural language understanding capabilities to read scholarly publications on the web. The knowledge extracted by these AI agents is organized into a heterogeneous graph, called Microsoft Academic Graph (MAG), where the nodes and the edges represent the entities engaging in scholarly communications and the relationships among them, respectively. The frequently updated data set and a few software tools central to the underlying AI components are distributed under an open data license for research and commercial applications. This paper describes the design, schema, and technical and business motivations behind MAG and elaborates how MAG can be used in analytics, search, and recommendation scenarios. How AI plays an important role in avoiding various biases and human induced errors in other data sets and how the technologies can be further improved in the future are also discussed.

1. INTRODUCTION

The field of science that studies the structure and the evolution of science has been established firmly on quantitative methodologies. It has seen tremendous growth since the pioneer work by Eugene Garfield at the Institute for Scientific Information (ISI), now part of Clarivate Analytics (Garfield, 1955, 1964, 1972). As a sign of growing interests beyond academic exercises, commercial entities that provide data sets to facilitate research and business exploitations have doubled in the last three years (Harzing, 2019). Despite these roots in the rigor of scientific pursuits, questionable results have also emerged in recent years. One of the most notorious incidents is the university rankings put forth by *U.S. News & World Report*, which received strong condemnation from the Computing Research Association (CRA). In their report (Berger et al., 2019), a CRA working group showed that incomplete data from Clarivate Analytics combined with surveys from human experts, who might no longer be up to date in their fields, may have led the news outlet to produce inaccurate rankings on research universities in the computer science area. More concerning, the same erroneous patterns that rank all the top U.S. computer science institutions as lagging behind a Middle Eastern university seem to be reproducible in the 2019 NTU Rankings[1] by National Taiwan University, where the rankings are derived using the same Clarivate data with a weighted sum of six indicators: the current year

[1] http://nturanking.lis.ntu.edu.tw/ranking/BySubject/ComputerSci.

an open access journal

Citation: Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

DOI: https://doi.org/10.1162/qss_a_00021

Received: 09 July 2019
Accepted: 10 December 2019

Corresponding Author:
Kuansan Wang
kuansanw@microsoft.com

Handling Editors:
Ludo Waltman and Vincent Larivière

and prior decade for each of publication count, citation count, and h-index of a university. These indicators are widely tracked by other databases for these universities, so NTU Rankings' disclosure allows insight into possible sources of the problem. There are numerous possible explanations, but they all seem to point to an inescapable conclusion that influential computer science publication venues must have been excluded; in other words, the cause of the erroneous results is highly likely to be inadequate selection biases in the data set rather than the analytic methodology conducted with the data. Given the leadership position Clarivate has in providing the community with data and how prominent its position is in facilitating data-driven decision-making processes in academic and research management institutions, the incompleteness of its data is of grave concern.

Our own experiences with Microsoft Academic Search and its subsequent relaunch in 2015 into Microsoft Academic Services (MAS) (Sinha et al., 2015) suggest that data collection methods considered adequate just a few years ago are actually in need of a major revamp to prevent quality degradation that might lead to seriously flawed results, as exemplified in the cases above. Many considerations that prompted the redesign remain valid today. First, we observe that the manners in which we conducted scholarly communications have changed dramatically. Although journal publications remain dominant and the volume continues to grow at an exponential rate, their roles in disseminating new discoveries and archiving known records are increasingly being supplemented with new breeds of online services, such as preprint services and collaboration websites, that are less expensive, more accessible, and able to offer fresher information (Tennant et al., 2019). Some important activities, such as journals dedicated to reproducing scientific results with replicated data and software (e.g., ReScience; see Rougier et al., 2017), are taking place exclusively on the web. Bolstered by the positive findings that open reviews do not compromise the peer review process (Chawla, 2019; Wang & Tahamtan, 2017), some societies have combined the online self-archiving and open review systems to conduct scholarly discourse, bypassing traditional publishers altogether (Soergel, Saunders, & McCallum, 2013). To faithfully capture the state of scholarly activities, we have to not only scale up to meet the voluminous demands in journal publications but also scale out to include the new channels of scholarly communications that are making good use of modern and popular information technologies.

Secondly, the technological landscape has also shifted significantly. Many technical standards, such as DOIs and ORCIDs, that serve as the bedrock for uniquely and persistently identifying publications and authors were proposed in the era of the semantic web, in which the proponents assumed that our societies would altruistically contribute manual efforts to annotate documents with standardized machine readable uniform resource identifiers (URIs) and resource definition frameworks (RDFs) so that the web could be smarter. Unfortunately, URIs are not designed with human factors in mind, such that their low usability and common typographical errors in DOIs and ORCIDs have frequently been a source of frustration and an adoption barrier. For instance, the top articles in the AI and machine learning areas are often published in *Journal of Machine Learning Research* (JMLR), Neurips, ICML, and ICLR, none of which use DOIs and ORCIDs. The DOIs for JMLR, one of the arguably most important journals in machine learning indexed in the Association for Computing Machinery (ACM) Digital Library, are now mostly reported as defunct by CrossRef, a DOI registration agency, contradictory to the promise that DOIs would be persistent. Scientists who would like to study the developments of these fields can therefore not rely on data sets with a strong dependency on DOIs. Most importantly, not having DOIs for these publications does not hamper their accessibility. In fact, our data shows these publications are often associated with multiple URLs on the web that can be easily served up by web search engines, making them more persistently accessible to readers than many with single DOIs as their sole

presence online. Similarly, the uptake of ORCIDs is also low, which warrants reflection after their heavy promotion for decades. Furthermore, in our own operations at MAS, we have observed that individuals would acquire multiple identifiers to partition their publications into separate profiles, defeating the design goal of ORCID being a unique identifier. On the other hand, AI technologies have advanced to the state that publications can be identified by their contents and authors disambiguated based on their fields of research, affiliations, coauthors, and other factors that are more natural to humans. Having to annotate ORCIDs is a mechanical and tedious data processing chore, better left to a machine. Given that scholarly communications are intended for human consumption, it appears more reasonable to challenge the machine to read more like humans rather than ask humans to work harder for the machine.

Perhaps the most important impetus for MAS redesign is our observation that intelligent technologies, when used inappropriately, could inadvertently incorporate and amplify the well-documented cognitive biases (Haselton, Nettle, & Andrews, 2015) in humans that lead to seriously negative impacts to society. We can no longer isolate ourselves within the confines of the technology community and must take a proactive role in considering the potential societal impacts from the technology we deploy and the data we distribute. As these views evolve rapidly from aspirational values to conventional wisdom in recent years, our professional society, ACM, and our employer, Microsoft, have both recently revised their codes of ethics to reflect the heightened responsibilities a modern technologist bears toward the society at large. Applied to data releases and related computational services, they imply a full disclosure of the protocols in the creation of data sets and the implementation of algorithms, including an explanation on the potential biases that might be introduced through the uses of particular techniques and methodologies. Most importantly, the developers should feel accountable for the misuses of their products and take initiatives to prevent them from being widely spread. Accordingly, the MAS redesign has a strong focus on the publishability of not only the data but also the underlying software components that contain the fewest sources of biases and codify the best practices in preempting inadvertent misuses. As noted above and further elaborated later, one of the most visible changes in MAS is the strategic move to gather raw data from the entire public web to minimize the selection and sampling biases that might have impaired the credibility of Clarivate's data in the two independent university ranking results described above. Furthermore, to reduce potential publication biases, MAS now includes work in all stages of scholarly communications, from preprints to conference or journal publications and reprints. As more scholarly work achieves impacts in patent publications, they are also included in MAS. To use the web data responsibly, one must address head-on the issues of web spam and other malicious content. To this end, an elaborated algorithm, called the *saliency measure* (Wang et al., 2019), is deeply ingrained into MAS to estimate the believability of web data and weight them accordingly. Perhaps the most challenging problem to mitigate biases is those that originated from human-generated exemplars for machine learning purposes, which inevitably include confirmation and other cognitive biases that are hard to detect, reconcile, and, most importantly, to explain. One recent example is the machine learning technique being used to categorize the scholarly articles in Dimensions, in which a single human expert is reported to use search results to guide the machine to learn the categories of publications (Hook, Porter, & Herzog, 2018). As a specialization of the framework known as *relevance feedback*, its efficacy in producing personalized recommendation and search results is widely reported in the literature (Manning, Raghavan, & Schütze, 2008). However, it is the very power of kowtowing to users on their subjective preferences that makes the technology a precarious tool for curating objective information. The Dimensions report claimed that search-engine-assisted human annotation was suitable but did not evaluate the agreement rate if

additional human experts were involved to review the labels for machine learning. Our experience, however, suggests that the cognitive biases among human experts are often significant and their annotation agreement rate tends to be low, leading to a questionable machine learning outcome due to incompleteness in training exemplars. Similar observations and potentially promising techniques to rectify biases are abundant in the literature (e.g., Joachims, Swaminathan, & Schnabel, 2017; Wang, Bendersky, Metzler, & Najork, 2016; White, 2013). This type of bias is difficult to explain, so we have elected to employ alternative approaches in the MAS redesign.

These considerations precipitated the new focus of MAS from an academic search engine relying mainly on publisher feeds to a web-based approach that actively exploits the cutting-edge cloud computing and AI technologies to tackle the scale challenges and cognitive biases. The publisher feeds have undergone quality checks before they are distributed, so they remain a valuable resource in MAS to validate the quality of data automatically curated with the technologies disclosed in Wang et al. (2019). The move nevertheless allows MAS to replicate the success of Google Scholar, which utilizes the massive document index from a web search engine to achieve comprehensive coverage of contemporary scholarly materials, many of which are not published and distributed through traditional channels and not assigned DOIs. In contrast to the index size of 40 million in 2014, the web crawl approach has enabled MAS to improve its coverage dramatically to include, by the end of November 2019, more than 225 million publications with more than 2 billion unique citations, growing at more than 1 million new publications a month in recent years. This improved coverage is a key in alleviating concerns about sampling biases in studies using incomplete data sets. To make the project useful for science of science studies in general, however, having a comprehensive coverage of scholarly materials is just a beginning: The data set needs to be able to robustly conflate references to the same entity, disambiguate mentions to distinct entities without heavily relying on human annotated URIs, and extract concepts being communicated in each article. Given that the explosive amount of scholarly materials being created has exceeded the human capacity to process, MAS aggressively utilizes the state-of-the-art machine learning techniques to achieve these objectives with little human intervention. Numerous studies seem to confirm that machine curated results in MAS achieve reasonable if not greater accuracy over commercial data sets with considerable amount of human effort (Harzing & Alakangas, 2017; Herrmannova & Knoth, 2016; Hug & Brändle, 2017; Hug, Ochsner, & Brändle, 2017; Kousha, Thelwall, & Abdoli, 2018; Thelwall, 2017, 2018a, 2018b, 2018c).

Within the scope of this article, we focus our discussion on MAG, a data component in MAS representing the knowledge acquired by the AI automatically. Readers interested in the technologies underpinning the creation of MAG and the applications of advanced knowledge reasoning and inferences are referred to a separate article (Wang et al., 2019). MAG is the knowledge representation modeling the scholarly communications on the web as observed and understood by the MAS AI agents. Scanning the entire web takes approximately two weeks, so MAG is updated biweekly and made publicly available under an open data license known as ODC-BY that explicitly protects MAG users from Microsoft's copyright and text and data mining claims for either research or commercial applications. To understand the potentials of MAG and why it is organized in its current form, however, a deeper understanding of the technologies behind its creation is warranted and provided below.

2. MAG SCHEMA

MAG uses a graph structure in which the nodes and the edges represent the entities involved in the scholarly communications and the interactions or relationships among them, respectively.

MAG is a heterogeneous graph, as it contains multiple types of nodes and the edges that are described in detail in this section. Similarly, MAG is a directed and attributed graph because the edges represent noncommunicative relationships and each node and edge can have its own properties, such as time stamps, probabilistic scores, and web addresses of the data sources. MAG is also a dynamic graph with a rapidly expanding topology because new authors, publications, and citations are constantly being discovered and added to the graph.

2.1. Publications

Publications are the main entity type in that all other entities are connected to a publication one way or another. This design choice is distinct from other choices that more greatly emphasize the social aspects of the scholarly world, such as Aminer (Tang, 2016; Zhang, Zhang, Yao, & Tang, 2018), where people are the focal type of entities in the graph. Accordingly, in MAG, authorship relations are modeled as "has authors" and represented by a directed edge from a publication to each of its authors, rather than the other way around, representing a "having authored" relation of an author. Each authorship relation is further annotated with the sequence number in which the author is found on the order of the publication record and the multiple affiliations each author has declared. Similarly, a publication has the "addressing the concept of" to a concept and "is published at" to a venue entity, respectively. In addition to the "citing" relation among publications, the AI agents in MAS further infer an "is related to" relation based on the semantic similarities of the contents, of which software utilities are also included in the MAG distribution, as described below.

2.1.1. Multiple points of presence

A unique characteristic of using the web crawl approach to source the data is that a publication can appear multiple times on the web, in contrast to the design of DOI that only recognizes one official web address for each publication. In addition to publishers' properties, publications can often be found on authors' personal websites and online archive facilities that are an inexpensive way to promote or fulfill the open access mandate. The discoveries of these additional versions of a publication are greatly aided by MAS having direct access to a web index: One can efficiently discover a large number of publications by searching the titles in a reference list of known publications. More publications can often be found once an address domain is discovered. Essentially, the web page discovery algorithm is a web crawler that can be adapted to amass an ample collection of publications online. All these web addresses, when marked safe by the crawler, are reported in MAG as a "paper URLs" attribute of a publication. On the average, each publication is associated with 5.5 web addresses based on the recent MAG releases. In many cases, one of the web addresses is the DOI. Particularly for articles in the medical fields, the corresponding PubMed IDs are provided.

The multiple presences of a publication are a key factor for MAG to reconstruct the fullest contents of a publication even when some of the contents are missing or locked behind a paywall, and MAG reports all their web addresses as a good starting point to anyone that would like to recreate or replicate a sizeable data set, such as MAG.

2.1.2. Citation relations and citation contexts

The web crawling approach affords MAS multiple ways to access the text contents of a publication, giving rise to a more precise determination of the concepts described in the publication (see below) and, more importantly, a "citing" relation with richer attributes. Aside from acquiring knowledge about authors, venues, and publications from the bibliography section,

MAS also extracts individual paragraphs immediately preceding each citation from the full text and reports them as "citation contexts" in MAG. These citation contexts are analogous to anchor texts in the web search scenario, in which they play a key role in determining the strength of affiliation between the citing and cited documents and contributing to the identification of spurious references to defend against link spam (Gyöngyi & Garcia-Molina, 2005). In MAS, a similar idea is incorporated to differentiate publications that are referenced multiple times or just mentioned in passing in the text. A preliminary study (Wang et al., 2019) has shown that, for research assessments based on citation behavior, an eigencentrality measure taking citation contexts into account can avoid the drawbacks of citation counts that treat each reference equally. Given that institutions are observed to game their ranking positions by manipulating citations in a similar manner to web link spam, the citation context attribute may prove as useful as in the case of web search.

2.1.3. Patents as publications

The MAG data model for a publication was originally intended for scholarly articles in journals, conference proceedings, or books, but it appears equally applicable to patent disclosures, so MAG started to include it in the summer of 2018. There, we redefine the inventors as the authors of the invention and the patent assignees as the affiliations of the authors. Patents are always filed under particular jurisdictions, so the venues are defaulted to the patent offices with which the patent application is registered, even though the same disclosure can be discovered in many parts of the web, such as a scholarly article. Once the patent disclosures are treated in the same way as scholarly articles, citations among them are normalized. In other words, scholarly articles that are cited by patents are treated the same way as those cited by other scholarly articles. Through patent citation into scholarly articles, MAG can be used as a powerful tool to study how research has affected practices that are deemed worthwhile for intellectual property protection.

Note that each patent is tagged in accordance with a patent classification scheme that organizes the potential utilities of an invention into a hierarchy. The hierarchy is designed to describe the application areas of an invention independent of the fields of the invention. For example, a patent can be tagged with a code for transportation improvement regardless of the underlying invention being derived from discoveries in chemistry, physics, or engineering. In MAG, the same AI agents that recognize and extract concepts from a scholarly article are equally applied to patent disclosures, resulting in a mapping between scientific concepts and patent classification codes at a large scale. This mapping is potentially a powerful tool in understanding the impacts of various fields of study on technological advancements and applications.

2.1.4. Publication family

The multiple occurrence of a single publication, however, brings about the technical challenge of conflation, namely, how to robustly recognize documents that should be treated as a single entity. This turns out to be a nontrivial problem, and our current technique is described in detail in Wang et al. (2019). Suffice it to note that the evolution of publication is closely tracked in MAG in a special "family" attribute of the publication entity. For scholarly articles, a family includes versions of the articles at various stages describing the similar concepts and with identical authorships, ranging from preprints and conference papers to journal papers and reprints or book chapters. For patent disclosures, a family consists of the same invention filed under various jurisdictions. Although MAG reports the individual citation

relations for each member publication in a family, the family ID provided in MAG allows applications to aggregate citations from all the family members. For example, in computing the saliency, an eigencentrality measure that MAG uses to estimate the importance of each node on the graph is based on the citations of the family rather than an individual.

2.1.5. Data sets and software

Scientists share data and software tools to promote reproducibility, and historically, these additional resources are often accompanied by corresponding articles describing the design considerations, protocols, and intended uses. Their uptakes can therefore be tracked through the citations received by the articles describing them. For example, the papers describing protein databanks, cancer statistics, and various corpuses and utilities for benchmarking AI tasks are among those that receive the highest citations in their respective fields. For these cases, the data sets and software packages are currently included in MAG as supplementary materials to the publication. However, it is observed that more data sets and software packages are not associated with descriptive articles and, in many cases, are being cited directly. As in the case of patents, it appears that the MAG data model can be adapted to treat data sets and software packages as publications and, indeed, some of the data sets are already included as standalone publications in MAG. As the time of writing, the merits of the two modeling approaches are still being evaluated and are open to public comments and feedback.

2.1.6. Article-centric versus venue-centric principle

In contrast to other data sets that have taken a journal- or venue-centric approach to index publications, the MAS relaunch has switched from the journal-centric to the article-centric principle. As explicitly stated in the San Francisco Declaration on Research Assessment (DORA)[2], research communities are reaching a consensus that it is utterly inappropriate to judge the merit of a publication based on the venue it is published in. MAG embraces this spirit and includes all articles from the web that are deemed as scholarly by a machine-learning-based classifier. In fact, as described below, MAG does not recognize a publication venue as a node on the graph until the venue has exceeded a threshold based on the total saliency of its publications. This inclusion policy makes MAG distinct in many ways. First, articles published in obscure venues will be found in MAG, including journals that some considered "predatory" in nature. Given that the costs of publishing at highly regarded journals have grown to an exorbitant level, especially for scholars from poor regions of the world that arguably need to participate in scholarly communications most, it is understandable that many authors need to seek less expensive outlets. The problem, however, as well articulated in Tennant et al. (2019), rests with the publication industry and not the authors, who deserve to be heard. Furthermore, we believe that more publication venues can enrich scholarly discourse, and a good way to examine the quality of a new venue is to make its contents more openly and widely discoverable and observe the collective judgments from the entire research community on them. In fact, MAG's inclusion policy makes it a useful data set for objective research into the publication venues and, more broadly, the dynamics in scholarly communications in general.

Second, MAG also includes preprints that have yet to complete the peer review process and publications that are participating in an open peer review process. We note that these new developments in review processes, enabled by modern information technology, are not

[2] https://sfdora.org.

well-traveled paths and may indeed lead to unintended consequences. It is, however, encouraging to see that recent findings have shown that more openness does not seem to change the review quality, as many fear (Chawla, 2019). MAG can contribute to the further investigation of this issue further by providing the saliency measurement, which is expensive to compute (see below).

Finally, MAG's open inclusion policy currently extends only to the public web. Materials not found on the web or available electronically only within proprietary systems, such as student theses and printed manuscripts or books, are not included unless they are frequently cited. However, because the full text is typically unavailable for in-depth natural language processing, the concept recognition, citation context extraction, and other advanced intelligent inferences of these publications are typically of lower quality.

2.1.7. *Principal component analysis*

Aside from the citation spam described earlier, another analogy to malicious web content for the scholarly domain is factitious documents that aim at tricking the content classifier into including low-quality materials as scholarly publications. The ease of this line of attack is convincingly demonstrated in the famous "Google Scholar Experiment" (López-Cózar, Robinson-García, & Torres-Salinas, 2014), which is still reproducible today, and, indeed, our practice attests that questionable content abounds on the web.

A multipronged approach is used in MAS to address this problem. First, the use of the saliency measure as a quality measure alleviates the weakness of simple aggregate statistics, such as the publication and citation counts, which are easily susceptible to the type of attack described in the Google Scholar Experiment. However, including the questionable content in processing and the MAG distribution is a waste of computational resource. Accordingly, in the very early stage of MAS processing, an additional step to conduct a principal component analysis (PCA) on the citation graph is taken and only the nodes corresponding to the largest component are selected for MAG. PCA is effective because questionable content typically manifests as local clusters in the citation graph and attacks with the means to infiltrate the entire scholarly communications globally have yet to be encountered. Unfortunately, PCA does have its drawbacks, the most regrettable of which is the bias against non-English publications. This is because the main principal component corresponds to publications oriented toward the international audience and collaboration. Although non-English publications cite English ones, the reverse is much less true. Similarly, self-publishing contents, such as student term papers and seminar reports, are also filtered out by the PCA even though the citations originating from these sources may be deemed legitimate and are indeed counted by Google Scholar. This differential motivates the feature in MAG to report two sets of "citation counts" for each publication, corresponding to citation sources that are included in or excluded from MAG, respectively.

2.2. Authors

Being able to identify the real-world person based on the name appearing on a publication is a highly desirable feature but also a technically challenging problem. Author name disambiguation has drawn intensive research interests (Cock et al., 2013; Kanani, McCallum, & Pal, 2007; Li et al., 2015; Liu, Lei, Liu, Wang, & Han, 2013; Roy et al., 2013; Wick, Kobren, & McCallum, 2013; Zhang, Xinhua, Huang, & Yang, 2019; Zhang, Zhang, Yao, & Tang, 2018; Zhong et al., 2013), yet the state-of-the-art techniques, using only information in the publication data such as coauthorship, affiliations, and topics, typically do not yield high enough

accuracy, especially for Asian or popular Western names. The reward of using these machine learning techniques is not high enough, so most systems have just used a simple name key (e.g., the author's last name prepended with first or middle initials, as in Google Scholar) to associate author names with publication clusters. The name keys are often not powerful enough to fully disambiguate author identities, making it more likely for each cluster to include publications from multiple authors. The ORCID standard could potentially address this problem were it not for its design problems and other adoption barriers, as previously discussed. Many systems, notably Google Scholar, employ additional crowdsourcing techniques, such as publication profile management websites that enlist the assistance of real-life people in author disambiguation. The URLs assigned to author profiles are typically more human friendly, making them strong alternatives to ORCID.

Both machine learning and crowdsourcing approaches are employed for MAG. In addition to machine learning techniques that only use publication records for author disambiguation, MAS further takes advantage of its unfettered access to a web search engine and includes public information, such as personal websites and public curricula vitae, to supplement the machine learning algorithm. In a way, the machine learning in MAS attempts to mimic the web search activities a human would typically undertake to infer the identity of an author. Observing that the name-key-based approaches, such as used in Google Scholar, tend to overconflate authors and thus assign more publication records to an author, MAS deliberately takes the opposite direction and decides to err on the conservative side, namely, publications bearing the same author name are not assigned to the same author node in MAG unless such assignments can exceed a 97% confidence threshold based on the machine learning algorithm. This artificially high threshold leads to author underconflation, where publications by the same author are split into multiple clusters if the variants in coauthors and topics are different enough to lower the confidence below the threshold. This design choice leads to the fact that the publication count of an author node in MAG can only be lower than the actual number of publications by the real-world author, complementing the upper bound estimates from systems that overconflate author publication records.

This design philosophy is extended to the treatment of crowdsourced data for author disambiguation that are collected from the profile management feature in the Microsoft Academic website. However, the collected data are treated only as, in machine learning terminology, supervision signals to recompute the decision boundaries but not to change the confidence threshold. The training algorithm is often not programmed to achieve perfect accuracy on the supervision data, so crowdsourced change requests can still be rejected by the learning algorithm and not reflected in MAG. The algorithm, nonetheless, cannot account for many real-life scenarios, especially in handling scholars who have undergone legal processes to change names or in cultures that have multiple family names. These scholars are encouraged to use the feedback mechanism on the Microsoft Academic website to request manual corrections. This remains one of the rare areas where the information in MAG is a result of human effort.

The strong design preference toward underconflating authors inevitably leads MAG to have more author nodes than there are real authors. In recent months, there have been constantly more than 220 million author nodes in MAG. It is, however, incorrect to estimate that the underconflating rate ranges between a factor of 4 and 5 because the underlying distribution, like many naturally occurring big data phenomena, features the "long tail" effect. Specifically, roughly three quarters of MAG author nodes are associated with a single publication and, when ignored, can bring the number of author nodes down considerably. Take the results from

the sample code in S.1 as an example. In the MAG snapshot taken on August 22, 2019, the total number of author nodes and the single publication nodes are 230,427,582 and 177,042,353, respectively, leaving the number of authors with two or more publications at 53,385,220, or 23.17% of the total. Many of these single-publication authors are indeed from patents that have only one inventor with no affiliation. Techniques powerful at dealing with long-tail data, such as the Good-Turing algorithm used in estimating the word frequencies for large linguistic corpuses (Gale & Sampson, 1995), may yet be proven useful again in smoothing the author counts from the raw observations as reported in MAG.

2.3. Institutions

Unlike other MAG entities, institutions are the only type of entities that do not have direct relations to publications, the primary type of entities in MAG. Rather, an institution entity occurs only as a secondary attribute in the "is authored by" relation to an author at the time of the publication. This design allows MAG to accommodate multiple affiliations that any author can declare on each publication, a feature that was not introduced to MAG until August 2018. Previous versions of MAG only allowed one dominant affiliation for each author, which causes minor but visible distortion in assessing the impacts of an institution based on its publications. This new modeling technique offers more precise control for analytic tasks, especially when authors move from one institution to another: Analyses that would like to consider all work by an institution's members even before they join it can simply select publications based on the author identifiers. On the other hand, when the use case needs to zoom in only on the work carried out at an institution, the identifier for the institution can be included in specifying the publications from MAG for analyses. This flexibility, however, can be tricky when using data analytic tools that treat attributed relations as distinct relations. For example, the sample code in S.1 demonstrates the special care needed in deduplicating the publication-author relations caused by multiple affiliations.

To accurately assign a unique identifier to an institution based solely on author's disclosure, however, is far from a trivial matter, because the affiliation is often in an organizational hierarchy that changes over time and often disclosed by authors with varying and inconsistent strategies. One notable attempt to tackle this problem is the Global Research Identifier Database (GRID)[3] spearheaded by Digital Science, the creator of Dimensions database (Hook, Porter, & Herzog, 2018) mentioned earlier. Currently, GRID has, through manual effort, gathered more than 15,000 relations describing the hierarchies and physical and online addresses among the 90,000+ organizations spanning over 200 countries in the database, including research units from educational, commercial, governmental, and public service institutions. Although not following the semantic web's URI specification rigorously, GRID identifiers are prominently featured in MAG because, like DOIs, they serve the purpose of disambiguating lexical names when used correctly. However, like other manual efforts, GRID encounters scalability and recency problems frequently seen in applying humans' limited cognitive capacities to big data processing. Take our organization, MSR, a division of Microsoft Corporation, as an example. GRID correctly recognizes that MSR has labs in Cambridge, UK, Beijing, China, and Bengaluru, India, but our labs in New York, NY, Cambridge, MA, and Montreal areas and the earliest labs in Redmond, San Francisco, and Silicon Valley are missing and without their individual identifiers. A once-affiliated team in Munich Germany that has been reorganized out of MSR is still identified as part of MSR,

[3] https://grid.ac/.

and its sister teams in Aachen, Germany, Cairo, Egypt, and Herzliya, Israel are not treated in the same manner. Many of the web links in GRID that could provide supporting evidence to the validity of the data are broken or point to pages with stale content. Understandably, these organizational details and their evolution over time can easily escape the grasp of human experts outside of Microsoft or even just MSR, especially when the notion of data accuracy is often only meaningful and highly contemporaneous with the time of publication.

Further confounding the inconsistencies from author disclosure are the editorial interventions by the publishers. Particularly with the crawling method used to curate MAG, the metadata of one publication often appears multiple times on the web, with many instances from various publishers that can impose their own unique stylistic conventions. The results are a spread of lexical varieties in describing a single affiliation that are too tedious for manual rectifications but are nevertheless invaluable assets for machine learning algorithms to automatically acquire synonymous phrases that can enrich coverage in processing affiliation mentions. The specific approach described in Wang et al. (2019) is an example of such an unsupervised machine learning algorithm that can reason among the various affiliation mentions of the same author, through either multiple publications or multiple versions of a single publication, to arrive at the most commonly shared segment of expression for the affiliation. This process is then extended by aggregating expressions from authors likely to be in the same organization to yield the canonical expression for the affiliation, based on which a MAG affiliation entity is hypothesized. With the access to a web index, a hypothesis of a distinct entity can be verified and disambiguated, mostly based on the domain name of the web address, and then a unique identifier assigned. Without a human in the loop, the method is inexpensive, highly scalable, and, most importantly, shown to be mathematically optimal in the *maximum a posteriori* (MAP) sense (Wang et al., 2019). The approach is also efficient and can be executed within the time frame of the regular MAG update cycle.

The data-driven approach leads to notable characteristics in MAG's treatment of affiliations. First, the information in MAG is dynamic and time stamp sensitive. When two institutions are merged into one, their identities can be inferred as soon as enough evidence can be observed from the web. Each MAG snapshot will be consistently tagged, so all publications will be attributed to the new institution's identifier. This design is convenient for conducting analyses on the merged organization with a more up-to-date version of MAG yet preserving the possibility of uncovering separate identities of the constituents with an older version of MAG. Secondly, the consistency-oriented approach seems to smooth out the complicated and fluid nature in organizational boundaries within an institution, as MAG uses the same affiliation identifiers on both scholarly and patent publications. Accordingly, MAG affiliations are composed largely of tertiary schools, with their individual departments, colleges, and research centers lumped under them. However, professional law, business, and medical schools or affiliated hospitals retain their separate identities, as do individual campuses of a university system or alliance. Industrial research teams tend to be aggregated at the corporate level (e.g., Microsoft, Google), as are units inside large governmental facilities (e.g., Chinese Academic of Sciences, CNRS, and Max Planck Society).

2.4. Venues

Scholarly communications have been undergoing dramatic changes for decades: With the lowering cost of web publishing, everyone is becoming a publisher, and with the popularity of the web search technology, self-published content is just as discoverable as that from incumbent publishers. These observations lead us to broaden the definition for publications and,

accordingly, the publication venues in our relaunch in which, as previously mentioned, a significant change is to migrate from a publisher-centric to a publication approach in sourcing the MAG contents. In other words, publication venues are no longer just those neatly defined by the publishers but wherever the provenance of a publication resides. A notable side effect of this process is MAG may appear not to respect ISBNs because the nuances that publishers adopt in assigning multiple ISBNs to a single journal are hard to process consistently. For example, it is not uncommon for a publisher to report a paper as published at a journal with one ISBN, only to report it again with a different ISBN months or even years later. Using the publication-centric approach, MAG will recognize only the first publication event and treat the subsequent one as a reprint. We recognize that this outcome is a departure from the tradition in some research communities that exclusively use the hard-copy publication dates as official.

The same data-driven approach used in recognizing the affiliation entities described above is applied here for publication venues where, aside from the publication records on the web, the machine learning algorithm is greatly aided by the diverse expressions authors use to refer to venues in their citations. As in the case for affiliations, the data-driven approach is effective in taking advantage of the commonality inherently encoded into the big data to identify venues of publication, even for web-only proceedings that have not gone through traditional publishers. In contrast to ISSNs or ISBNs, which are not unique, the web address domains are used to disambiguate venues that have the similar or same names.

However, unlike in the case for affiliation entities, MAG does not immediately report a newly discovered publication venue as an entity as soon as it is recognized by the automatic algorithm, until the saliencies of its publications have jointly exceeded a manually chosen threshold. This is another area in MAG where we exercise human intervention on machine curated results. The policy is motivated by the observation that the ever-lowering cost of publishing has unfortunately lured nefarious players into the world of scholarly communications that exploit the publication pressure on scholars for their commercial gains. Derided as "predatory," these publishers behave like web spammers and create cheap publication venues with little concern for their quality to lure authors into publishing their research. As technologists making data available to the public, we find it our ethical duty to carefully distinguish the genuine newcomers from the repugnant ones and try our best to avoid inadvertently promoting predatory venues. Our approach in delaying the recognition of publication venues is complementary to the MAG policy in reporting publication entities, described earlier, and the consistent use of the saliency measure reinforces the notion that publishers and other members participating in the scholarly communications should focus on producing work that will be recognized positively by the scientific communities rather than gaming the system by manipulating simple measurements, such as publication or citation counts.

2.5. Concepts

Concepts are the entities MAG uses to categorize the fields of study of publications. They are abstract, and their recognition is therefore not as simple as in the case of other types of MAG entities, where the task amounts to accurately identifying the MAG entities with their the real-world counterparts. Furthermore, concepts are intuitively hierarchical, yet human experts disagree significantly about how to organize them, as library systems around the world use very different categorization schemes.

Consistent with the data-driven design for MAG, a state-of-the-art semantic understanding algorithm (Wang et al., 2019) is employed to recognize and organize concepts in each publication. The algorithm relies on a core capability to quantify the semantic distance between

two textual paragraphs. With this semantic distance measure, passages that are addressing a similar topic are clustered together and their joint semantics forms the basis of a concept. A subsumption principle is introduced (Sinha et al., 2015) to organize the concepts into the hierarchical structure as currently in MAG. Simply put, if a concept is always present in the context of another, MAG treats the former as a child concept of the latter. As there is no mathematical guarantee that the hierarchy thus formed is unique, we manually define the top two levels of concepts where various categorization systems seem to agree most. The rest of the six levels are left to the algorithm to decide organically. Although new concepts are detected dynamically as they are observed in the documents, the hierarchy is only recomputed every six months to minimize the instabilities of the data sets.

Internally, each concept is a collection of semantic representations of textual passages, and these abstract representations, though not readable to humans, are adequate for machine reasoning, such as finding the "is related to" relations among concepts or publications in MAG (Kanakia, Shen, Eide, & Wang, 2019). To improve usability and facilitate our own debugging tasks, however, an attempt is made to find, for each concept, a tutorial article on the web that is closed in semantics and uses the title of the article as the human-readable label for the concept. Most of the tutorial articles are found in Wikipedia, although, as disclosed in MAG in the form of URLs, other sources are also considered. These human-readable labels are naturally not perfect, and users should be mindful that MAG concepts are determined and represented by a more sophisticated machine reading mechanism that itself is an actively studied topic.

2.6. Semantic Similarity Functions

For transparency, we include in the regular MAG updates two software tools that are at the center of determining the semantic similarity for the tasks of concept recognition and hierarchy formation. The first, called *Language Similarity*, contains a software object with a function to quantify the semantic distance for any two given text strings, using the same machine learning models as in the MAG production. This function is used to identify tutorial documents to generate human-readable labels for concepts. Based on this basic operation, the language similarity object also features an additional function that computes the distance between an input text to a concept, and, in turn, a third function that retrieves the top concepts closest to any given text. These functions are used in MAG to extract concepts from scholarly articles but are surprisingly effective for patents as well. The algorithm underlying Language Similarity is an application of an insightful observation proposed in Harris (1954), known as *distributional similarity*, that semantically similar words or phrases in natural language tend to exhibit similar occurrence patterns in linguistic contexts. Many recent advancements in natural language processing, starting with the word embedding technique (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), can all trace their theoretical origins to this principle.

The notion of distributional similarity has been extended from its linguistic root to network topology to assess the similarity of two nodes based on their connections to neighboring nodes. The second tool, called the *Network Similarity*, implements this extension to compute the similarity measure based on the MAG topology. The heterogeneous nature of MAG, though, makes the notion of contextual similarity highly dependent on the type of relation connecting the nodes. For example, two authors may be similar because they research in the same fields or just attend similar conferences. The network similarity functions therefore have an additional parameter that specifies under which entity relation the similarity is to be assessed.

Empirically, these two sets of similarity functions are complementary. As demonstrated in the task of publication recommendation (Kanakia, Shen, Eide, & Wang, 2019), the algorithm

combining the language and the network similarity features achieves the best outcome. In addition, the two functionalities are combined in determining the concepts for each publication, disambiguating authors, and identifying related and advanced entity relations in MAG.

3. SALIENCY MEASURE IN MAG

In addition to information gathering, this project also explores the extent to which the power of machines can be leveraged to facilitate studies into the science of science. Among the many topics in this field, how to access the importance of academic entities is of particular interest because not only are the problems scientifically challenging but also, practically, the web is full of malicious content that requires effective countermeasures for MAG to maintain acceptable data quality.

Eigenvector centrality is a well-known measure for characterizing the importance of nodes in a network with applications in many fields, the most notable one being the PageRank algorithm behind Google (Franceschet, 2011). PageRank is notable for its recursive property: The algorithm assigns a higher PageRank score to a document cited by higher PageRank documents. In other words, citations from higher PageRank sources weigh more. However, issues in adapting PageRank for assessing academic importance remain unsolved, either because the expedient factor, called the *teleportation probability*, is implausibly high for citation graphs or, more generally, there is no solid grounding in determining the right strategy to estimate the latent variables intrinsic to the framework (Maslov & Redner, 2008; Waltman & Yan, 2014).

We have found it effective to take a two-pronged approach to tackle these problems (Wang et al., 2019) and arrive at an eigencentrality-based measure we call *saliency*. First, in addition to weighing citations proportional to the saliencies of its sources, we adopt a heterogeneous generalization that further considers the saliencies of the authors, their affiliations, the venue, and the recency of the last citations for each publication. The natural richness in which two publications can be related to each other through heterogeneous relations makes it unnecessary to introduce the teleportation mechanism, which is hard to justify mathematically. Similarly, the recursive property notable for an eigencentrality measure also applies to saliency: A MAG entity will have high saliency if it contributes to a publication of high saliency.

Secondly, instead of manually deciding the heterogeneous weights with heuristics, we employ a reinforcement learning algorithm with a delayed reward function aiming at best predicting the future citation behaviors of the research community at large. Reinforcement learning (RL) in recent years has been shown to be a powerful technique for computers to dynamically adjust the latent variables in the internal scoring function to achieve long-term gains. Often, in an artificial environment or a game setting, where the objective function and the rules of the game can be precisely described, abundant computational power today allows straightforward RL programs to explore automatically synthesized scenarios and gain experience at a speed several orders of magnitude faster than real life. Accordingly, RL-trained machines can often outsmart top human players of the game by having centuries more of experience. The environment of scholarly communications is far from artificial, and citation behaviors cannot be automatically synthesized, yet MAG has more than two centuries of citation behaviors that an RL algorithm can use to imaginarily travel back and forth in time to gain experience in predicting the future. Empirically, as shown in Wang et al. (2019), five years' worth of citation behaviors are enough as a target for the RL algorithm to reliably learn the heterogeneous weights. The computational power needed for RL is nevertheless considerable, such that the saliencies of all MAG entities are precomputed and distributed with each update.

The properties of an eigencentrality measure imply that saliency is a tougher measure to game than a simple citation count and its derivatives, such as the h-index. This is indeed empirically shown in Wang et al. (2019). By analogy, if PageRank can thwart the attack of link spam because of the design of eigencentrality, saliency can also render "citation cartels" less effective and may even alleviate a worsening problem known as *coercive citations* (Wilhite & Fong, 2012), if they are indeed less relevant or expressed in less enthusiastic tones in the citation contexts. Although these conjectures have yet to be rigorously proved, MAG does provide the necessary data for such studies.

4. DISCUSSION

Much has changed since Clarivate's ISI was founded in 1960, when less than half a million scholarly publications were produced in that era. Today, MAG is discovering more than twice as many in a month. In fact, MAG shows the growth rate in annual publication output has been on an exponential pace for almost two centuries and shows no sign of abating. Suffice it to say that no commercial investments can sustain the corresponding human labor needed to even browse the literature, let alone to manually conduct the analyses with adequate depth. The industry is actively collaborating and innovating, but the floodgates have been opened, and the cracks, ranging from aberrant publication and citation behaviors to abhorrent analytics, such as university rankings that mislead the public in school choices, have emerged. Like fake news and many unscrupulous social media activities that have been negatively impacting the civil and political discourse in our societies, technologies play a central role in instigating many of these problems. Fortunately, our experience in producing MAG suggests technologies can also provide an effective answer. MAG has shown that the machine curated data set is no less accurate if not better than human efforts, yet it is more scalable and less expensive: The cloud computing cost needed for updating the MAG content is roughly equal to that of a single experienced data scientist. We cannot even fathom how many data scientists we would have to hire had all the functions in MAG had to be produced by manual effort. This is consistent with the case study, for instance, that shows that a credible research institution ranking in multiple metrics for all of more than half a million subject areas and all institutions in MAG can be done in minutes, not in months, and with a cost less than a cup of coffee (Microsoft, 2019). As part of the CRA effort, we are genuinely at a loss why entities such as *U.S. News & World Report* would continue to mislead. Encouragingly, though, we are constantly surprised by the diverse use cases of MAG, which can be tracked by inspecting the number of publications citing the WWW-2015 article that first describes MaAG: By the end of November 2019, there have been more than 261 articles documenting how MAG is being used in the research community at large.

Many compare our website with Google Scholar and portray the two efforts as competitors. Although we can understand the perception, as both are providing free services to the academic communities and both are sourcing the data from the web, the analogy is off in many respects. First and foremost, unlike Google, the main business of Microsoft does not rely on building the most detailed profile for each person for targeted services, a business model that has caused many privacy concerns and governmental interventions, most notably the European GDPR legislation. Accordingly, the Microsoft Academic website is not just an outlet to collect users' online behavioral data. In fact, the website does not use the browse and click information at all in its machine learning components. The only behavioral signal the website uses, as described earlier, is the author disambiguation hints, but even that is not contingent upon user identities. The most important behavioral signal driving the website is the citation

behaviors in MAG. In a way, the website is a demonstration of how a search engine may be constructed without using online behavioral data. Instead, Microsoft Academic follows the business model of being a platform that strives to "empower everyone and every organization to do more and achieve more." In other words, MAG is viewed as a data platform that promotes and adds value to Microsoft's cloud business, and the website is a demonstration of the unique values we would like to showcase on the platform. This fundamental business model difference manifests itself in many subtle ways. Take the discovery business in the libraries as an example. Unlike Google Scholar, we do not offer the "library links" feature with which the libraries allow Google to index their catalogues and their users to search library collections on Google Scholar. As the service is offered for free, many public libraries can no longer justify their technology budgets, partially because, in many jurisdictions, it is prohibited to spend public funds services that can be otherwise obtained for free. Furthermore, the arrangement implicitly assumes users understand that their uses of library services are no longer confined within the boundaries of the libraries. This model, although apparently mutually beneficial, leaves a lot of challenging privacy questions unanswered. On the other hand, under the platform model we follow, libraries can receive frequent MAG updates to merge with their private data and integrate with their existing discovery infrastructure, all without exposing the usage patterns of their services and users. The model ostensibly requires more technology work, which Microsoft has cultivated a healthy ecosystem to assist with, and is the only palatable option for many research institutions that require confidentiality in their work. With the platform model, it is therefore not surprising to see we are actively assisting other academic search engines with MAG. For example, Semantic Scholar[4] and Lens.org[5] are among the partners that have granted explicit permission to Microsoft to disclose our active collaborations.

Perhaps the most distinctive feature of our work is driven by a research objective to explore the nature of AI and how to use AI to better our lives with our research colleagues all around the world. Despite the achievements in MAG, we are mindfully aware that we are probably just scratching a tiny surface, and there is a long way to go before the cognitive capabilities of more computers can be fully unleashed. For example, the concept recognition in MAG is far from perfect, and the recommendations made by MAG are still too overwhelming and can be more contextually sensitive. The distributional similarity functions distributed in MAG are themselves actively researched topics that we would like to continue our contributions, especially with MAG, where semantics is encoded in the form of both unstructured natural language passages and highly organized topological structure. It is within this scientific pursuit and our belief in open science that we are and will continue distributing our data and software tools so that our work can be reproducible under the most rigorous scientific standards. It is also our hope that by effectively harnessing machine power to lower the cost of producing and distributing MAG, we can remove the access barriers to scholarly knowledge and open science, especially those from incumbent commercial entities that have made data access prohibitively expensive except for well-funded organizations. Aside from the monetary cost, many commercial interests further restrain scientists with contractual nondisclosure agreements and exploit legal arguments based on antihacking or copyright laws that prevent scientists from accessing necessary data and investigating apparent errors merely by threats of litigation. If only they could as actively pursue the misuses of their products (such as those documented in DORA) even when the misuses serve their business interests! For various historical reasons, these practices are largely tolerated by the scientific communities even though the cost of not knowing how many tenure applications are denied and how many research proposals are

[4] https://www.semanticscholar.org/.
[5] https://www.lens.org/.

rejected because of faulty yet opaque data hits very close to home. We hope that, in the very near future, these business-trumping-science practices will be widely regarded as anomalies or even unethical behaviors resulting from outdated business models that can be completely abandoned by effectively embracing new and cost-efficient technologies.

ACKNOWLEDGMENTS

Dr. Hao Ma led the efforts in creating many advanced features in MAG, and Drs. Bojune "Paul" Hsu and Rong Xiao participated in the design and implementation of the saliency framework. Dr. Junjie Qian contributed to the experiments of many large-scale algorithms described in this article. The work would not be possible without the strong support from Microsoft Bing engineering teams and colleagues in Microsoft Research labs around the globe.

COMPETING INTERESTS

All authors are employees of Microsoft Corporation, which provides the financial support for the work described in this article.

REFERENCES

Berger, E., Blackburn, S. M., Brodley, C., Jagadish, H. V., McKinley, K. S., Nascimento, M. A., … Xie, L. (2019). GOTO rankings considered helpful. *Communications of the ACM*, *62*(7), 29–30.

Chawla, D. (2019). Rare trial of open peer review allays common concerns. *Nature*. Retrieved from https://academic.microsoft.com/paper/2913755906

Cock, M., Roy, S., Savvana, S., Mandava, V., Dalessandro, B., Perlich, C., … Hamner, B. (2013). The Microsoft Academic Search challenges at KDD Cup 2013. *2013 IEEE International Conference on Big Data*, pp. 1–4. Retrieved from https://academic.microsoft.com/paper/2038101504

Franceschet, M. (2011). PageRank: Standing on the shoulders of giants. *Communications of the ACM*, *54*(6), 92–101.

Gale, W., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*(3), 217–237. Retrieved from https://academic.microsoft.com/paper/2063918473

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*(3159), 108–111.

Garfield, E. (1964). "Science Citation Index"—A new dimension in indexing. *Science*, *144*(3619), 649–654.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation journals can be ranked by frequency and impact of citations for science policy studies. *Science*, *178*(4060), 471–479.

Gyöngyi, Z., & Garcia-Molina, H. (2005). Web spam taxonomy. *AIRWeb*, 39–47.

Harris, Z. (1954). Distributional structure. *WORD*, *10*(2), 146–162.

Harzing, A.-W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, *120*(1), 341–249.

Harzing, A.-W., & Alakangas, S. (2017). Microsoft Academic: Is the phoenix getting wings? *Scientometrics*, *110*(1), 371–383.

Haselton, M., Nettle, D., & Andrews, P. (2015). The evolution of cognitive bias. In *The Handbook of Evolutionary Psychology*, 1–20. Retrieved from https://academic.microsoft.com/paper/2269742369

Herrmannova, D., & Knoth, P. (2016). An analysis of the Microsoft Academic graph. *D-Lib Magazine*, *22*(7), 6.

Hook, D., Porter, S., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, *3*. Retrieved from https://academic.microsoft.com/paper/2888592790

Hug, S., & Brändle, M. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551–1571.

Hug, S., Ochsner, M., & Brändle, M. (2017). Citation analysis with Microsoft Academic. *Scientometrics*, *111*(1), 371–378.

Joachims, T., Swaminathan, A., & Schnabel, T. (2017). Unbiased learning-to-rank with biased feedback. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 781–789. Retrieved from https://academic.microsoft.com/paper/2507134384

Kanakia, A., Shen, Z., Eide, D., & Wang, K. (2019). A scalable hybrid research paper recommender system for Microsoft Academic. *Proceedings of WWW-2019*, pp. 2893–2899.

Kanani, P., McCallum, A., & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the web. *IJCAI'07 Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pp. 429–434. Retrieved from https://academic.microsoft.com/paper/1547705211

Kousha, K., Thelwall, M., & Abdoli, M. (2018). Can Microsoft Academic assess the early citation impact of in-press articles? A multi-discipline exploratory analysis. *Journal of Informetrics*, *12*(1), 287–298.

Li, C.-L., Su, Y.-C., Lin, T.-W., Tsai, C.-H., Chang, W.-C., Huang, K.-H., … Lin, C.-J. (2015). Combination of feature engineering and ranking models for paper-author identification in KDD Cup 2013. *Journal of Machine Learning Research*, *16*(1), 2921–2947. Retrieved from https://academic.microsoft.com/paper/2613826676

Liu, J., Lei, K., Liu, J., Wang, C., & Han, J. (2013). Ranking-based name matching for author disambiguation in bibliographic data. *Proceedings of ACM SIGKDD Annual Conference on Knowledge Discovery and Data Mining*, (p. 8). Retrieved from https://academic.microsoft.com/paper/2133494723

López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and

manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454.
Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Retrieved from https://academic.microsoft.com/paper/1532325895
Maslov, S., & Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *The Journal of Neuroscience*, *28*(44), 11103–11105.
Microsoft. (2019). GOTO ranking made affordable with MAG. Retrieved from https://www.microsoft.com/en-us/research/project/academic/articles/goto-ranking-made-affordable-with-mag/
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of Advances in Neural Information Processing Systems*, 3111–3119.
Rougier, N., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L., Benureau, F., … Zito, T. (2017). Sustainable computational science: The ReScience initiative. *PeerJ*, *3*(12), 1–8.
Roy, S., Cock, M., Mandava, V., Savanna, S., Dalessandro, B., Perlich, C., … Hamner, B. (2013). The Microsoft Academic search dataset and KDD Cup 2013. *Proceedings of ACM SIGKDD Annual Conference on Knowledge Discovery and Data Mining*, p. 1–6. Retrieved from https://academic.microsoft.com/paper/1975258157
Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. *Proceedings of the 24th International Conference on World Wide Web*, pp. 243–246.
Soergel, D., Saunders, A., & McCallum, A. (2013). Open scholarship and peer review: A time for experimentation. Retrieved from https://academic.microsoft.com/paper/2121395634
Tang, J. (2016). AMiner: Mining deep knowledge from Big Scholar Data. *WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 373–373. Retrieved from https://academic.microsoft.com/paper/2337399039
Tennant, J., Crane, H., Crick, T., Davila, J., Enkhbayar, A., Havemann, J., … Vanholsbeeck, M. (2019). Ten hot topics around scholarly publishing. *Publications*, *7*(2), 34. Retrieved from https://academic.microsoft.com/paper/2943934175
Thelwall, M. (2017). Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals. *Journal of Informetrics*, *11*(4), 1201–1212.
Thelwall, M. (2018a). Can Microsoft Academic be used for citation analysis of preprint archives? The case of the Social Science Research Network. *Scientometrics*, *115*(2), 913–928.
Thelwall, M. (2018b). Does Microsoft Academic find early citations? *Scientometrics*, *114*(1), 325–334.
Thelwall, M. (2018c). Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics*, *12*(1), 1–9.
Waltman, L., & Yan, E. (2014). PageRank-Related Methods for Analyzing Citation Networks. In Y. Ding, R. Rousseau, & D. Wolfram, *Measuring Scholarly Impact* (pp. 83–100). Switzerland: Springer.
Wang, K., Shen, Z., Huang, C.-Y., Wu, C.-H., Eide, D., Dong, Y., … Rogahn, R. (2019). A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data*, *2*, 45. https://doi.org/10.3389/fdata.2019.00045
Wang, P., & Tahamtan, I. (2017). The state-of-the-art of open peer review: Early adopters. *Proceedings of the Association for Information Science and Technology*, *54*(1), 819–820. Retrieved from https://academic.microsoft.com/paper/2765327763
Wang, X., Bendersky, M., Metzler, D., & Najork, M. (2016). Learning to Rank with selection bias in personal search. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 115–124. Retrieved from https://academic.microsoft.com/paper/2340526403
White, R. (2013). Beliefs and biases in web search. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12. Retrieved from https://academic.microsoft.com/paper/2159205954
Wick, M., Kobren, A., & McCallum, A. (2013). Large-scale author coreference via hierarchical entity representations. Retrieved from https://academic.microsoft.com/paper/2188509728
Wilhite, A., & Fong, E. (2012). Coercive citation in academic publishing. *Science*, *335*(6068), 542–543.
Zhang, S., Xinhua, E., Huang, T., & Yang, F. (2019). ANDMC: An algorithm for author name disambiguation based on molecular cross clustering. *International Conference on Database Systems for Advanced Applications*, pp. 173–185. Retrieved from https://academic.microsoft.com/paper/2942498978
Zhang, Y., Zhang, F., Yao, P., & Tang, J. (2018). Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1002–1011. Retrieved from https://academic.microsoft.com/paper/2809279178
Zhong, E., Li, L., Wang, N., Tan, B., Zhu, Y., Zhao, L., & Yang, Q. (2013). Contextual rule-based feature engineering for author-paper identification. *Proceedings of ACM SIGKDD Annual Conference on Knowledge Discovery and Data Mining*, p. 6. Retrieved from https://academic.microsoft.com/paper/2147658782