6th International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2020, 75 Amherst St. Cambridge, MA USA

Are author, affiliation, and citation networks predictive of a journal getting blacklisted?

Lizhen Liang and Daniel Acuna
Syracuse University

Keywords: Predatory Journals, Citation Networks, Open Access Journals

Extended Abstract

Open access journals are becoming increasingly viable publication venues for scientists, educational organizations, and government funders. Unfortunately, unscrupulous publishers have taken advantage of this trend by creating journals that are open access but predatory or of low quality [1]. Some services attempt to remedy this situation by providing a white list and black list of journals, manually vetted by experts. Two examples of these expertly curated lists are the Directory of Open Access Journals (DOAJ) and the Cabbell's journal blacklist and whitelist. However, how these organizations choose journals is poorly understood. It would be beneficial to understand these decisions and also it would be important to improve on the detection accuracy of these services. In this preliminary work, we codify the rules that the DOAJ purports to use for journal auditing and examine their effectiveness in telling apart blacklisted vs whitelisted journals [2]. We compare these rules to features derived from the author, organization, and citation networks. We show that by using a combination of the DOAJ rules and network features, we can achieve significantly higher accuracy in our predictions. Finally, we examine the features that are most predictive and discuss our next steps.

Method. In this work, we estimate the effect of several features related to citation networks and features that the DOAJ considered best practices of reputable journals. In particular, we focus on three sets of features and models: 1) a model with features from the citation network of articles, authors, organizations, and journals. 2) a model based on what the Directory of Open Access Journals considers the publishing best practice. 3) A model using combined features.

Features of the citation network: Based on the publication-level citation network we obtain from Microsoft Academic Graph [3], we compute the following features about a journal: Summation, average, standard deviation (SASD) of citation for a journal, SADS of author self-citation for a journal, SADS of affiliation self-citation for a journal, and SADS of journal self-citation for a journal

Features from the Directory of Open Access Journals (DOAJ): Based on the Directory of Open Access Journals, there is a definition of publishing best practice [2]. Based on the guideline provided by DOAJ, we consider the following aspects as predictors of good journals: country code of affiliation frequently published in the journal, Website availability (HTTP status code), journal availability in Pubmed, the entropy of fields of study covered by the journal, and average academic age of authors publishing papers in a journal (measured by the year of the first publication for each author)

Results. We tested the ability of the three models described above to predict whether a journal is whitelisted in DOAJ

*Data***.** Based on Microsoft Academic Graph dataset [2] and Pubmed Open Access Subset dataset (https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/).

With the features mentioned above, we take the journals listed in the DOAJ white list and unwhite list and use the whitelisted status as labels. We train three logistic regression classifiers. The performance of the models is measured by cross-validated area under the ROC curve (AUC). This measures varies from 0.5 (random) to 1.0 (perfect prediction). Only the significant features of the regression are shown in Table 1. The model with only network features performs poorly with an AUC of 0.5339. The features suggested by the DOAJ do significantly better with an AUC of 0.7285. Combining both sets of features gives, however, the highest performance with AUC of 0.7306.

The regression Table 1 shows that for journals, the location of the affiliation of the authors submitting to a journal matters. Out-sized concentration of authors might be indicative of predatory behavior from the journal's point of view. We found that journals with a disperse set of topics tend to be blacklisted as well. This is surprising because most high-profile journals tend to have a wide topic base, but perhaps blacklisted journals are significantly beyond this multidisciplinary threshold. The other two features are somewhat expected: if the website of the journal is down or the journal age is short, it is more likely to be blacklisted.

Conclusion. In this work, we examined factors related to a journal being unwhitelisted from the Directory of Open Access Journals (DOAJ), including the own rules of the DOAJ and a set of features derived from the network. Our results shed light on the relative importance of DOAJ rules, and that network features are not particularly predictive.

Significant features	**Features sets for predicting whether a journal is whitelisted**		
	Model 1: Network features	**Model 2:** DOAJ features	**Model 3:** Network + DOAJ
Author affiliation location (Peru)		-0.418**	-0.451**
		(0.189)	(0.189)
		t = -2.215	t = -2.382
		p = 0.027	p = 0.018
Author affiliation location (Pakistan)		-0.218*	-0.247**
		(0.119)	(0.120)
		t = -1.822	t = -2.057
		p = 0.069	p = 0.040
Website unavailable (HTTP code 403)		-0.181**	-0.176**
		(0.071)	(0.072)
		t = -2.540	t = -2.430
		p = 0.012	p = 0.016
Entropy of fields of studies		-0.072***	-0.073***
		(0.007)	(0.007)
		t = -10.876	t = -10.722
		p = 0.000	p = 0.000
Average academic age		-0.016***	-0.016***
		(0.002)	(0.002)
		t = -10.294	t = -9.962
		p = 0.000	p = 0.000
Observations	2,928	3,038	2,928
Area Under the Curve	0.5339	0.7285	**0.7306**

1. Robert E Bartholome, 2014, "Science for sale: the rise of predatory journals", J R Soc Med. 107(10): 384–385.
2. DOAJ. 2020. Directory of Open Access Journals. Retrieved from https://doaj.org/publishers#applying

3. Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. WWW 2015