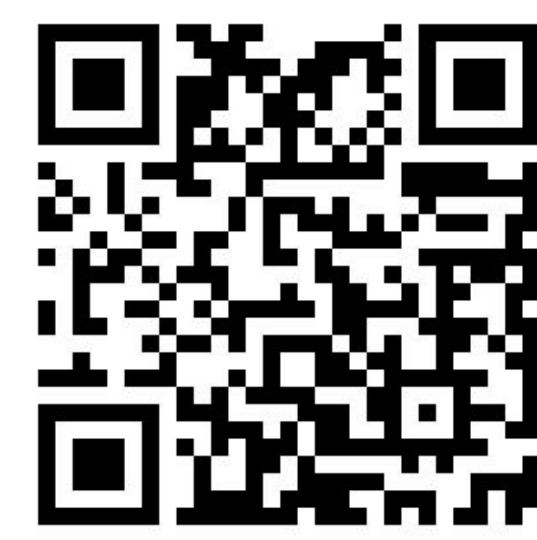


Identifying Fabricated Networks within Authorship-for-Sale Enterprises

Simon J. Porter
Leslie D. McIntosh, PhD



What is Authorship-for-sale?

Authorship-for-sale is a type of papermill product where an “author” buys a position on a paper. The research is often fake, and the authors generally do not know each other. Participants in Authorship-for-sale papers are typically young.

Are authorship-for-sale papers detectable?

Frequent participants in authorship-for-sale networks create research networks that are essentially random. This creates a rare network signal detectable by the clustering coefficient of their immediate coauthor network, and the number of people in it.

Distribution of Network Shapes: Each table column relates to a specific class of network shape. The column labelled ">n=0" are those networks seen fewer than 10 times; the column labelled ">n=1" are those network shapes that have appeared more than 10 times but fewer than 100 times. Percentages are relative to the size of the experience cohort. Thus, 0.10% of 'students' (<5 years publishing experience) are associated with co-authorship network shapes that occur fewer than 10 times in 2022.

cohort	> n=0	> n=1	> n=2	> n=3	> n=4	> n=5	> n=6
Student (I)	0.10%	0.61%	4.05%	12.26%	18.90%	23.71%	40.37%
Postdoc (II)	0.43%	2.08%	9.07%	18.02%	21.16%	17.69%	31.55%
Early Career (III)	1.14%	4.30%	13.19%	20.68%	21.08%	14.70%	24.91%
Established (IV)	2.24%	6.61%	16.00%	21.35%	19.83%	12.88%	21.08%
Career Building (V)	3.26%	7.90%	17.16%	20.89%	18.96%	12.16%	19.67%
Peak Production (VI)	3.82%	8.70%	17.46%	20.94%	18.22%	11.94%	18.91%
Advanced (VII)	4.33%	9.14%	17.87%	20.80%	17.71%	11.99%	18.16%
Senior Researcher (VIII)	3.83%	8.10%	16.17%	19.45%	17.64%	13.67%	21.14%

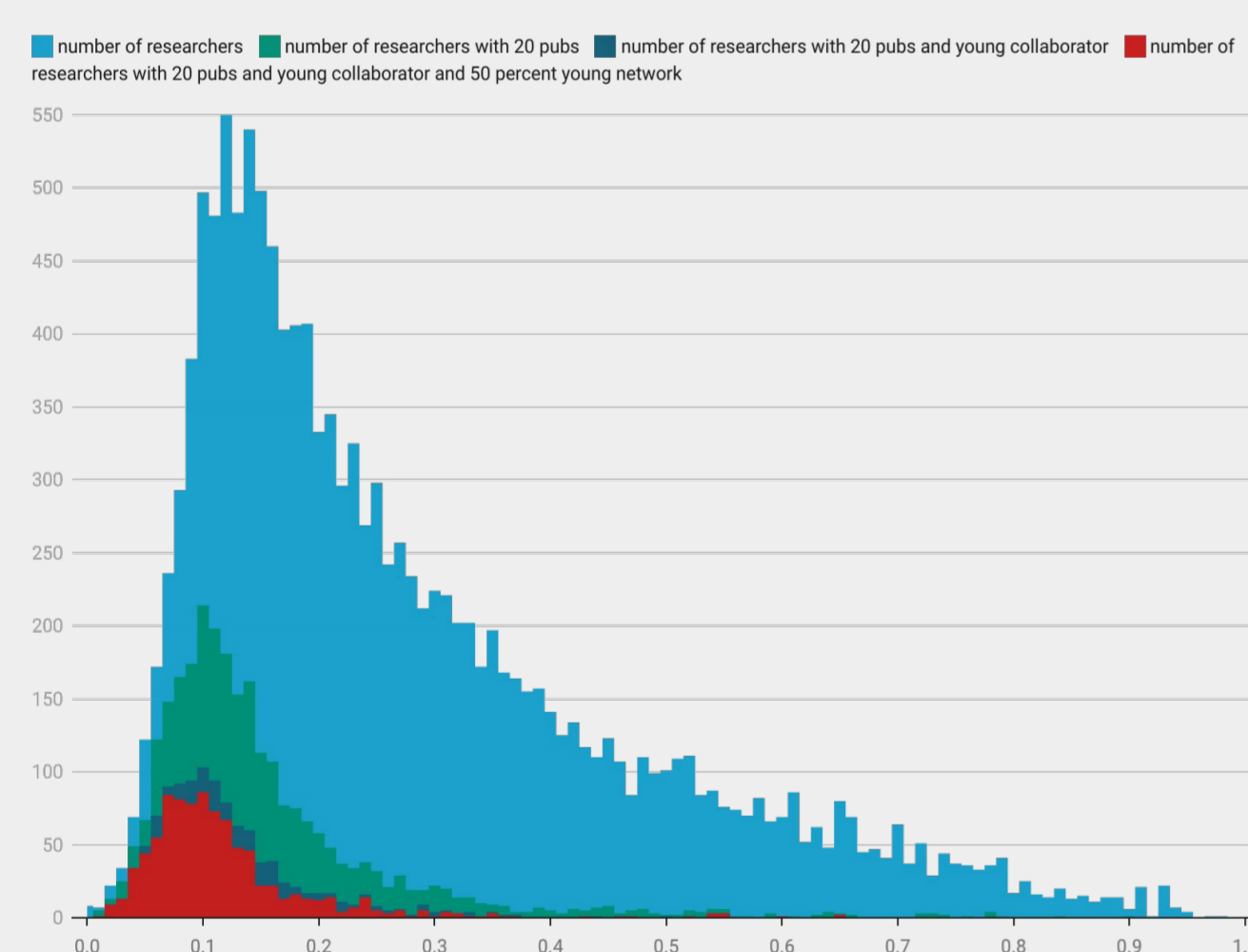
Finding Network shapes that match the suspicious author model... (Part 1)

cohort	> n=0
Student (I)	0.10%
Postdoc (II)	0.43%

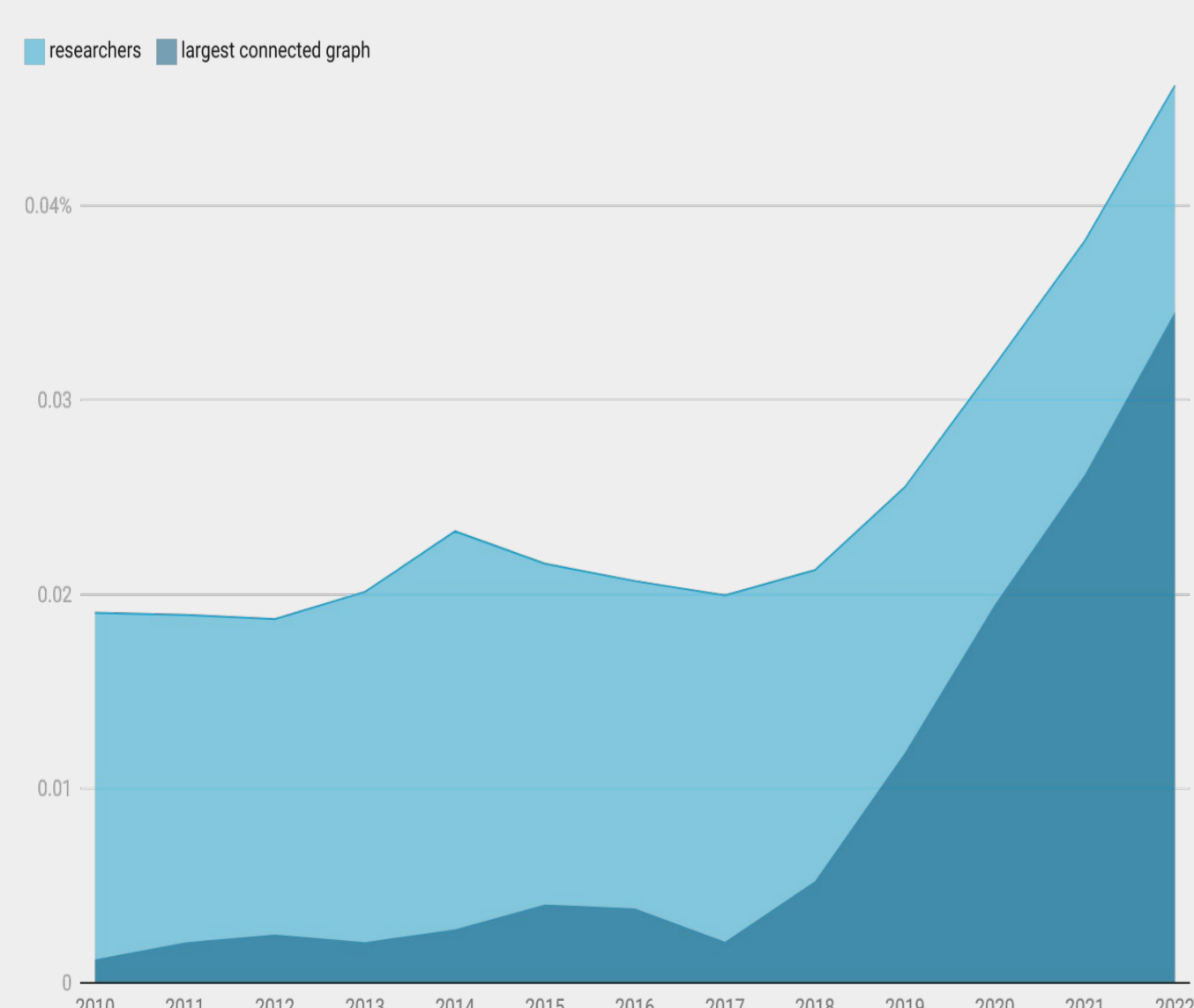
Histogram of Stage I and Stage II researchers with network shapes that have been repeated fewer than 10 times in 2022. Researchers are distributed across the x-axis by their clustering coefficient. The middle green histogram represents the subset of researchers that have produced greater than 20 publications in the year. In the bottom red subset, the most frequent collaborator is a Stage I-III researcher, and greater than 50% of the network is made up of Stage I-II researchers.

A property of large random graphs is they become connected once they reach a certain density

The area between the top of the light blue region and the x-axis represents all suspicious Stage I and II researchers as a percentage of all Stage I and Stage II researchers. The dark blue area shows just those that are part of the largest connected component of the suspicious co-author network defined by our model.



Filtering the suspicious author graph by the largest connected component

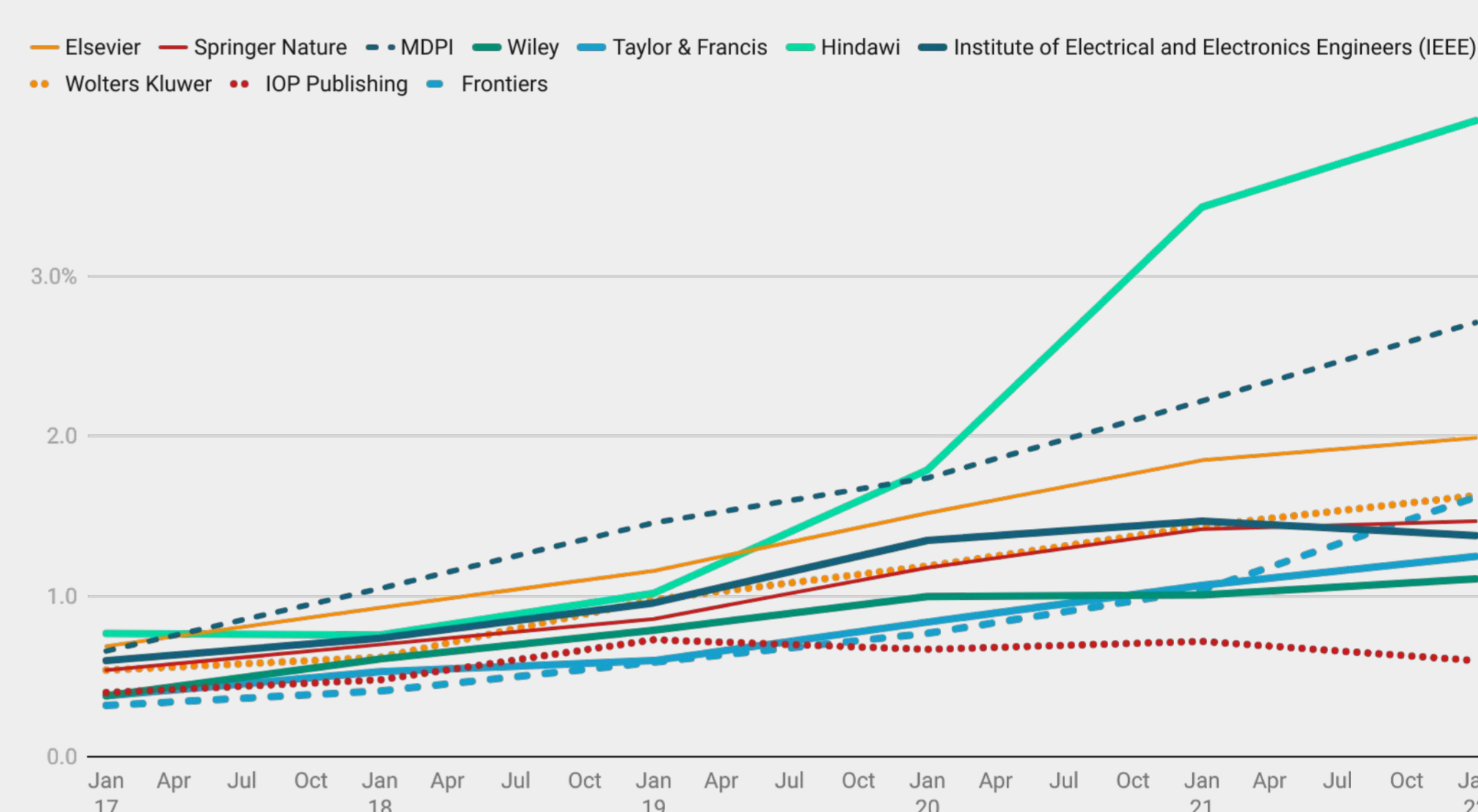


How can this methodology be used to detect paper mill papers?

Publishers

Publishers can use the suspicious author methodology to detect trends, and identify authors at the point of submission that require careful consideration

Percentage of papers with suspicious authors by publisher



Institutions

Institutions can identify young researchers that look to be connected to the authorship-for-sale network, providing the opportunity to help young researchers who may have lost their way.

Does an authorship-for-sale algorithm overlap with other paper mill detection methods?

Yes, we compared the detection methodology against the papers identified by the Problematic Paper Screener.

country	all publications	matched publications	percentage
Pakistan	266	251	94%
Saudi Arabia	345	317	92%
Malaysia	122	99	81%
Iran	88	65	74%
Turkey	96	66	69%
Iraq	234	119	51%
Egypt	209	106	51%
India	1,666	773	46%
United States	126	55	44%
China	1,059	263	25%

Absolute number of suspicious papers detected in the years 2020, 2021 and 2022 by country. The "All publications" column is the total number of publications identified via the curated Problematic Paper Screener method, and the "Matched publications" column reports the number of publications matched with the model proposed in the current paper. The Percentage column gives the proportion of matched publications. Different countries had different overlap rates suggesting different paper mill cultures (authorship-for-sale vs paper-for-sale) for instance.