Monitoring women's scholarly production during the COVID-19 pandemic

Philippe Vincent-Lamarre[1], Cassidy R. Sugimoto[2], Vincent Larivière[1,3]

[1] École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec, Canada.
[2] School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA.
[3] Observatoire des sciences et des technologies, Université du Québec à Montréal, Montréal, Québec, Canada.

Methods

We used a web scrapper to directly collect the metadata of all submissions since 2019-01-01 on arXiv, bioRxiv, medRxiv, the National Bureau of Economic Research (NBER), Preprints.org, and F1000Research. We collected the metadata of preprints and preregistration on the Open Science Framework (OSF) and data from repositories with a significant number of submissions that are archived on the OSF database, including EarthArXiv, SocArXiv and PsyArXiv. We downloaded the preregistrations of the American Economic Association (AEA) and ClinicalTrials.gov directly from their website. We used custom python scripts (https://github.com/lamvin/manuscripts-tracker) that called an API in the case of arXiv, or that used the search function from the other repositories in order to pull the metadata, including the author names from the submissions.

We then matched the first names of authors to our gender disambiguation algorithm (see: https://www.nature.com/news/polopoly_fs/7.14227.1386700530!/suppinfoFile/504211a_s1.pdf). A gender was assigned to 92% of authors. This percentage will oscillate as preprints and registered reports are added to the database. Tables 1 and 2 present the number of submissions retrieved, authors, and authors matched to a gender by platform, while Tables 3 and 4 provide the same information by arXiv subrepository. The arXiv data in Tables 1 and 2 exclude the hep-ex and nucl-ex subrepositories due to the low % of gender match (caused by the widespread use of initials instead of first names). To identify the COVID-related manuscripts, we ran a regular expression that returned a positive value if the title contained the following keywords: covid, ncov, coronavirus or sars-cov-2.

Table 1. Total number of submissions collected from the repositories, and statistics on the gender inference.

All data (2019-2020)				
Platform	# Submissions	Total authors	Total authors matched	% Matched
EarthArXiv	801	3548	3379	95%
F1000Research	1219	3769	3506	93%
NBER	1651	5759	4590	80%
Preprints.org	5821	25720	24002	93%
PsyArXiv	5196	17715	17023	96%
SocArXiv	2198	4902	4624	94%
arXiv	210129	856624	673197	79%
bioRxiv	44703	335031	307209	92%
medRxiv	4203	35888	34876	97%
AEA	1224	3308	3181	96%
ClinicalTrials.gov	22823	27930	26174	94%
OSF Preregistration	7491	18691	17765	95%
Total	**307459**	**1338885**	**1119526**	**92%**

Table 2. Number of submissions collected from the repositories, and statistics on the gender inference for the first four months of 2019 and 2020.

Jan-Feb-Mar-Apr (2019-2020)				
Platform	# Submissions	Total authors	Total authors matched	% Matched
EarthArXiv	397	1773	1687	95%
F1000Research	589	1864	1722	92%
NBER	842	2942	2341	80%
Preprints.org	3012	13278	12399	93%
PsyArXiv	2701	9236	8837	96%
SocArXiv	1135	2719	2587	95%
arXiv	103156	426709	332406	78%
bioRxiv	23532	177854	163386	92%
medRxiv	3354	28485	27727	97%
AEA	593	1554	1499	96%
ClinicalTrials.gov	12278	15034	14089	94%
OSF Preregistration	3985	9963	9399	94%
Total	**155574**	**691411**	**578079**	**92%**

Table 3. Total number of submissions collected from the arXiv sub repositories, and statistics on the gender inference.

All data (2019-2020)				
arXiv subrepository	# Submissions	Total authors	Total authors matched	% Matched
astro-ph	18798	165352	91116	55%
cond-mat	20388	99448	72562	73%
cs	60201	232413	221588	95%
econ	931	2265	2146	95%
eess	8549	35100	33056	94%
gr-qc	4134	12531	9043	72%
hep-ex	611	28987	2018	7%
hep-lat	634	2883	2215	77%
hep-ph	6032	20179	14560	72%
hep-th	4938	12327	10355	84%
math	45266	98073	88132	90%
math-ph	1935	4126	3367	82%
nlin	1095	3123	2411	77%
nucl-ex	545	11791	1077	9%
nucl-th	1674	5946	3191	54%
physics	18159	99864	64501	65%
q-bio	2826	11762	10665	91%
q-fin	1098	2796	2568	92%
quant-ph	7584	29440	23828	81%
stat	5887	18996	17893	94%
Total	**211285**	**897402**	**676292**	**73%**

Table 4. Number of submissions collected from the arXiv sub repositories, and statistics on the gender inference for the first four months of 2019 and 2020.

Jan-Feb-Mar-Apr (2019-2020)				
arXiv subrepository	# Submissions	Total authors	Total authors matched	% Matched
astro-ph	9309	83407	47105	56%
cond-mat	10108	49701	36003	72%
cs	29670	115387	109900	95%
econ	440	1116	1062	95%
eess	3903	15598	14668	94%
gr-qc	2045	6870	4484	65%
hep-ex	288	14050	1038	7%
hep-lat	302	1440	1083	75%
hep-ph	2813	9826	6894	70%
hep-th	2178	5333	4420	83%
math	22544	48624	43649	90%
math-ph	937	2012	1664	83%
nlin	503	1452	1092	75%
nucl-ex	280	5565	430	8%
nucl-th	846	3084	1626	53%
physics	8952	51555	31497	61%
q-bio	1561	6532	5892	90%
q-fin	533	1306	1214	93%
quant-ph	3680	14288	11538	81%
stat	2832	9178	8615	94%
Total	**103724**	**446324**	**333874**	**72%**

Figures

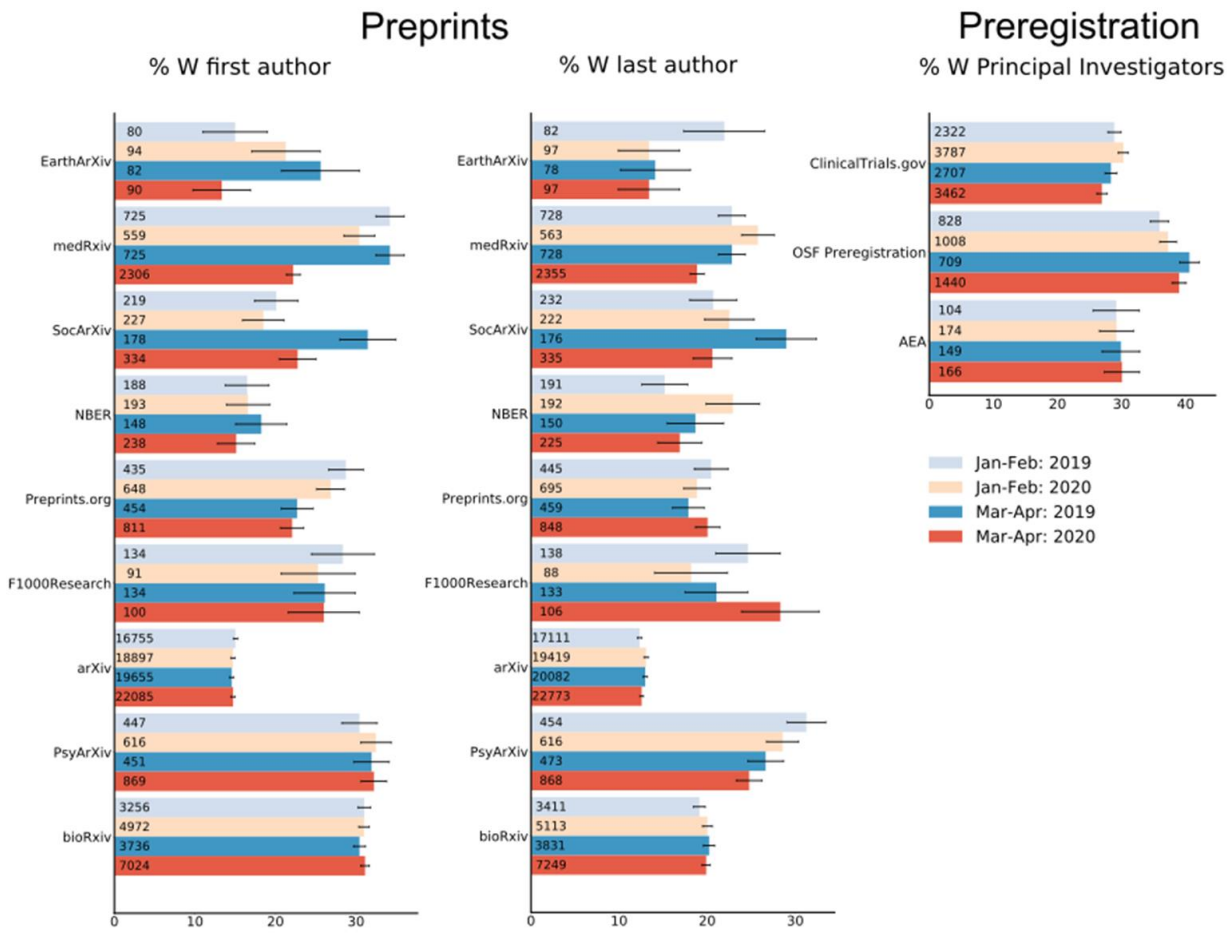

Figure 1. Percentage of women in first and last authorships, by preprint repository, and percentage of women principal investigators on the registration platform. Data are provided for two-month periods (January-February and March-April) of both 2019 and 2020, to allow comparison of women's submission patterns of the current social isolation period (March-April 2020) with the two previous months (January-February 2020) as well as the same period last year. January-February 2019 is shown to provide an indication of growth in women's preprints and registration until COVID-19 reached the level of a pandemic and shelter-in-place orders were instituted in most countries. Because medRxiv started accepting submissions in June 2019, we used the June to December time period of 2019 instead of January to April period for 2019. The numbers inside the bars show the number of submissions for which we could infer gender to estimate each average.

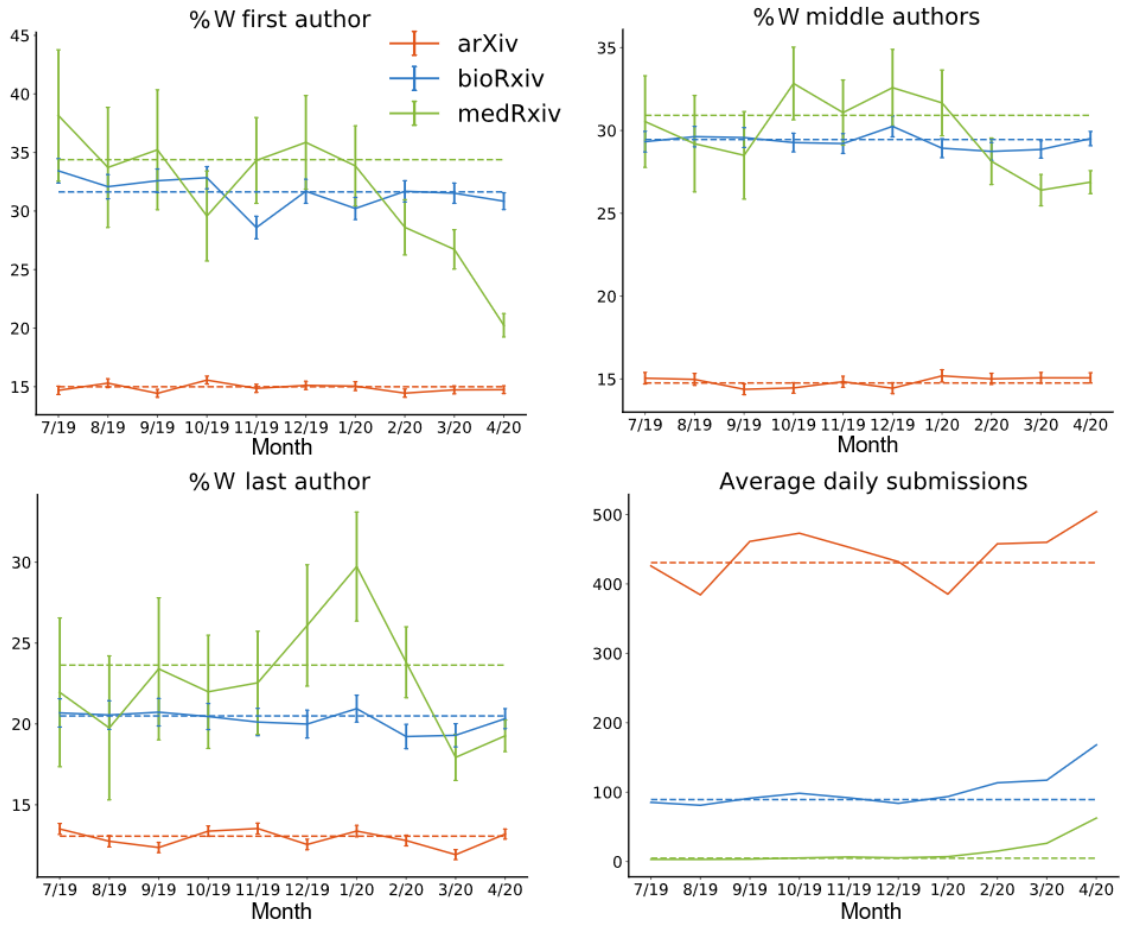

Figure 2. Percentage of women first, middle, last authorship, as well as average daily number of submissions, by preprint repository (arXiv, bioRxiv, medRxiv), July 2019 to April 2020.

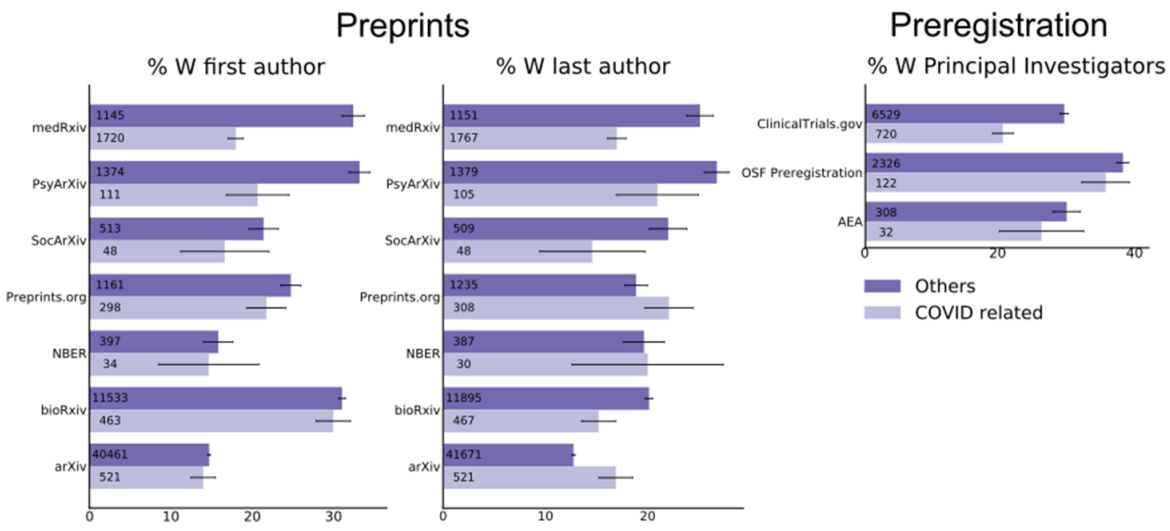

Figure 3. Percentage of women first and last authorships, by preprint repository, and percentage of women principal investigators in registration platform, for COVID-19 related research and other topics for 2020 submissions. The numbers inside the bars show the number of submissions for which we could infer gender to estimate each average.

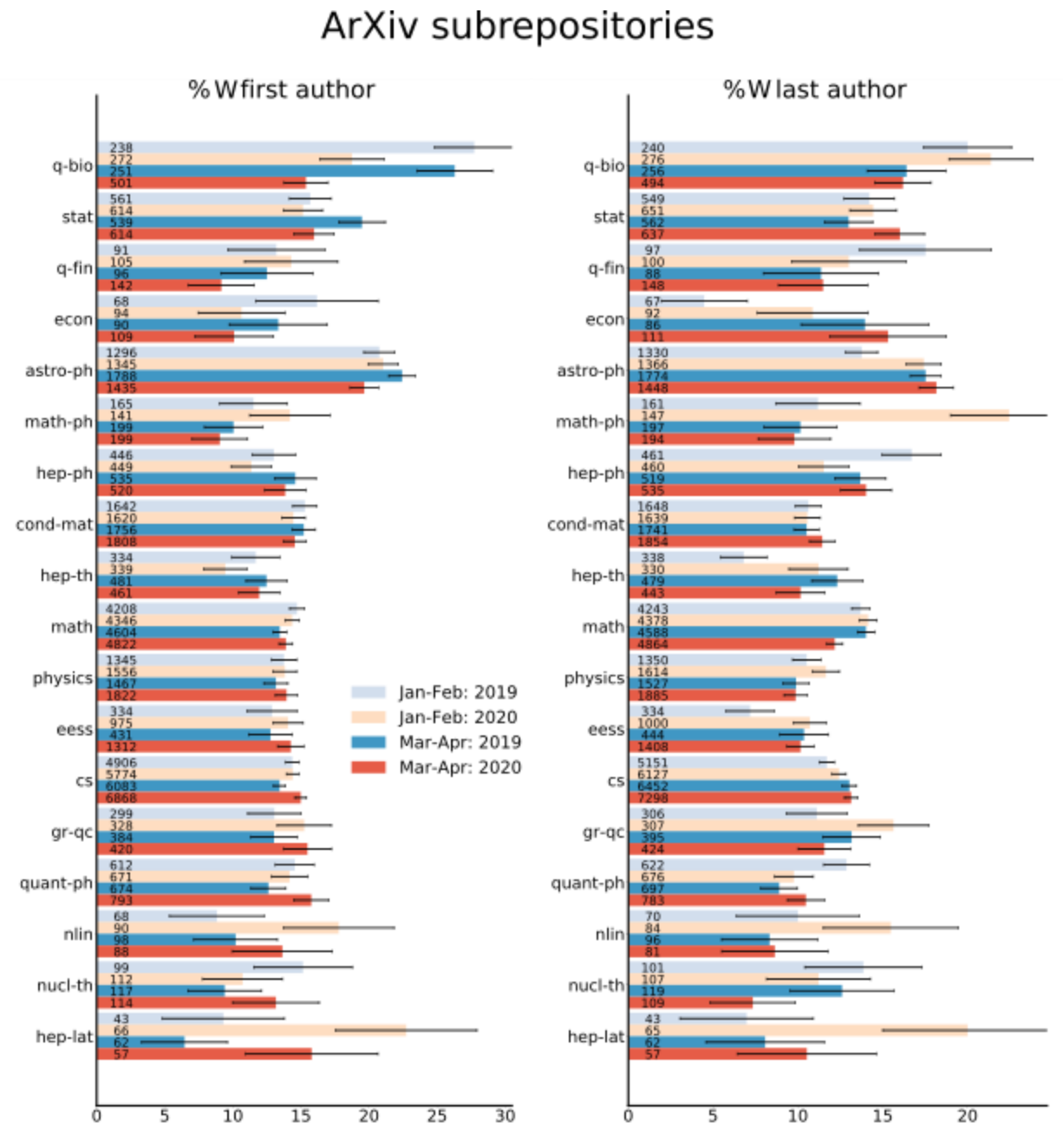

Figure S1: Figure 1.

Percentage of women in first and last authorships, by arXiv sub-repositories, preprint repository, and percentage of women principal investigators on the registration platform. The numbers inside the bars show the number of submissions for which we could infer gender to estimate each average.

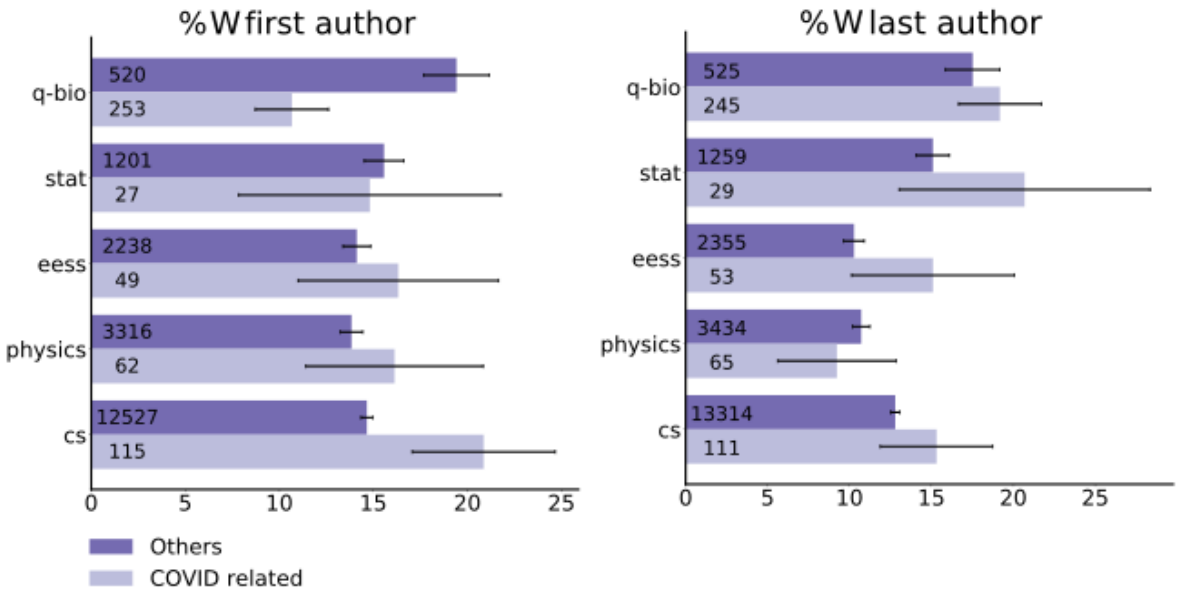

Figure S2. Percentage of women first and last authorships, by arXiv sub-repositories, and percentage of women principal investigators in registration platform, for COVID-19 related research and other topics for 2020 submissions. The numbers inside the bars show the number of submissions for which we could infer gender to estimate each average.