Supplementary Text

Validation of gender prediction. We first validated the genderize.io algorithm using a set of 3265 names whose gender had been hand-coded based on appearance (Broderick & Casadevall cite). The names were supplied to the genderize algorithm both with and without the accompanying country data. The genderize algorithm returned gender predictions for 2899 queries when first names were given and 2167 when country data was also supplied (732 names were associated with countries unsupported by genderize).

Sensitivity and specificity, are measurements of the algorithm's tendency to return correct answers instead of false positives (e.g., a man incorrectly gendered as a woman) or false negatives (e.g., a woman incorrectly gendered as a man). The closer these values are to 1, the smaller the chance that the algorithm will return the correlating false response. Accuracy is a composite measure of the algorithm's ability to differentiate the genders correctly. These measurements were calculated from the data sets (with and without country data supplied) at three different probability threshold cutoffs: the default genderize (0.5), a probability threshold of 0.85 (0.85), and a modified probability of 0.85, which factors in the number of instances returned (pmod0.85)(citations).

At the 0.5 threshold, the data set returned a sensitivity of 0.8943 and specificity of 0.9339 for an accuracy of 0.911, compared to a marginally higher accuracy of 0.9146 for the data set where country data were included (Table S1). Generally speaking, the accuracy increases as the threshold increases, with slight trade offs between sensitivity and specificity. For the purposes of our analysis, we opted to use the pmod0.85 threshold moving forward (Table S1, in bold).

To understand the extent of geographic bias in our gender assignment against regions and languages with gender-less naming conventions, or that lack social media for incorporation into the genderize algorithm, we compared the number of names predicted without associated country data to when country data was also supplied. In our test data set, the top five countries associated with names were the United States, Germany, the United Kingdom, France, and

China. The countries with the highest proportion of un-predicted genders when country data were supplied are Cambodia, Iceland, Indonesia, Ireland, and Mexico, where the maximum number of names supplied ranged from 1 to 15. To determine the impact of each country towards the overall percentage of names whose genders were not predicted (27.14%), we found the difference between the percent of names un-predicted for each country and the overall percentage, multiplied by the proportion of observations from that country to the total observations and finally divided by the overall percentage of un-predicted names (Fig. S8A). The top five countries with the greatest impact on un-predicted names, and thus the countries receiving the most negative bias from genderize were Canada, China, Ireland, Belgium, and Sweden. These data suggest that there is likely some bias against countries with gender-neutral naming conventions (China), and indicates the stringency with which the algorithm applies gender to names that are accompanied by country data. For instance, strongly gendered names such as Peter and Pedro were not assigned gender when associated with Canada.

We next applied the genderize algorithm at the pmod0.85 threshold to our journals data set and tested its validity on a small portion. All first names collected from our data set were submitted to genderize both with and without country data and only those with a pmod equivalent to or greater than 0.85 were retained. Next, the predicted genders were assigned to individuals as described previously (Fig. S7). Given the relatively small number of editors and senior editors in our data set, the presenting gender (man/woman) of editors and senior editors in our data set was hand-validated using Google where possible. Of the 1072 editor names, 938 were predicted by our application of genderize for an accuracy of 0.9989, thus increasing our confidence in the gender predictions where made.

In our full data set, the five countries with the most individuals were the United States, China, Japan, France, and Germany. The countries with the highest proportion of un-predicted genders were Burundi, Chad, Kingman Reef, North Korea, and Maldives, where the maximum number of names supplied ranged from 1 to 4. Proportionally, fewer names in our full data set were assigned gender than in our validation data set (40.01% un-predicted versus 27.14% un-predicted, respectively). Since adjusting the workflow to predict the gender of names both with and without country data, the countries receiving the most negative bias from genderize were China, Japan,

South Korea, India, and Taiwan (Fig. S8B). These data indicate what we previously predicted, that the genderize algorithm has bias against countries with gender-neutral naming conventions.

Table S1. sensitivity/specificity/accuracy of genderize thresholds. Bolded text denotes the accuracy of the threshold used in all further analyses.

	First Names			Plus Country Data		
Measure	p0.5	p0.85	pmod0.85	p0.5	p0.85	pmod0.85
Sensitivity	0.8943	0.9516	0.971	0.9055	0.9471	0.9669
Specificity	0.9339	0.9593	0.972	0.9265	0.9553	0.9727
Accuracy	0.9110	0.9549	**0.9714**	0.9146	0.9507	**0.9695**