Supporting Information for

Quantifying Behavior-based Gender Discrimination on Collaborative Platforms

Orsolya Vásárhelyi, Balázs Vedres

Orsolya Vásárhelyi.
E-mail: orsolya.vasarhelyi@uni-corvinus.hu

This PDF file includes:

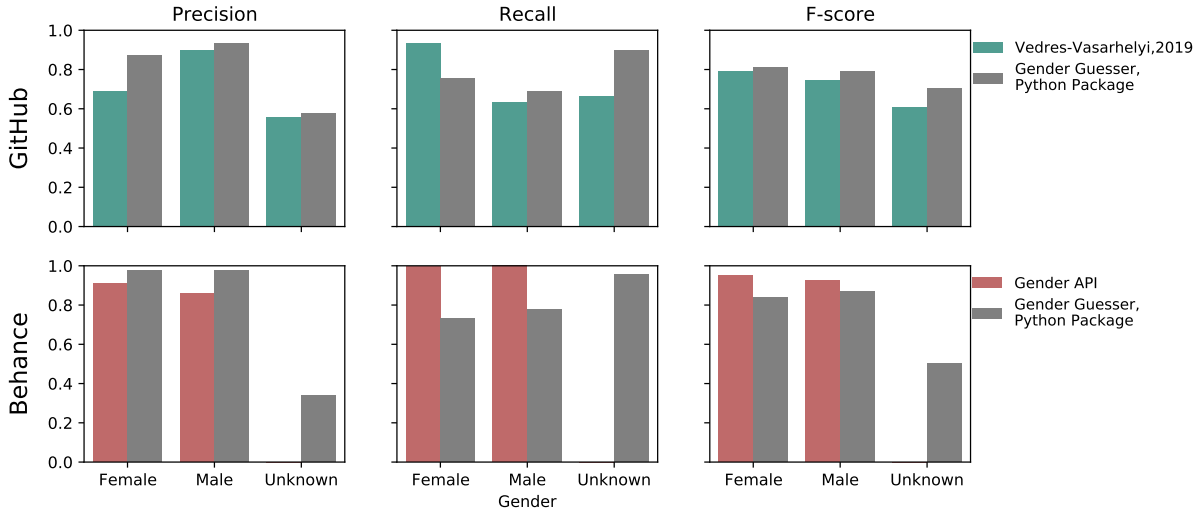

Fig. S1. **Gender Inferring Accuracy** Precision, Recall, and F-score of the GitHub (Vedres-Vasarhelyi, 2019) and Behance (Gender API) gender inferring methods against the manually inferred baseline method and a commonly used alternative method (Gender Guesser Python Package). Among GitHub users, our method and the default Python package yielded very similar results, optimized for high male precision. The used method's relative strength is female-recall, and it's weakness is unknown-recall. The commercial Gender API used to infer the gender of Behance users resulted in higher overall precision, recall, and f-score compared to the default python package. It is important to note that this dataset officially did not include unknown-gendered users, although we found 45 (11%) accounts which belong to companies, therefore, their gender could not be inferred.

Model Tables.

References

		MW	P	Female		Male	
				.25 quantile	.75 quantile	.25 quantile	.75 qunatile
Behance	Attention	21135593	0.000	25	401	41	927,5
	Success	20261832	0.000	47	980	66	2264
	Survival	17853000	0.128	1	1	1	1
GitHub	Attention	54302043	0.000	0	10	1	12
	Success	55770990	0.000	0	0	0	1
	Survival	52335000	0.000	1	1	1	1

Table S1. Mann-Whitney Statistics (MW),Significance (P) and Interquartile range of Attention, Success, Survival by gender

Fig. S2. Specializations on GitHub and Behance We used Scipy's PCA.decomposition package with Varimax Rotation to identify independent factors (?). Bar charts show the explained variance of each factor. The correlation matrices show the "importance" and the sign of the relationship of the language/design field in the component. On GitHub we identified 6 main specializations; 1) Frontend development, 2) Developers using Ruby for backend development, 3) Backend Development with high activity in Java, 4) Data Science, 5) iOS development, and 6) PHP enthusiastic with Frontend focus. On Behance our principal component analysis yield 8 main factors: 1) Photography, 2) Graphic Design, 3) Branding, 4) Art Direction, 5) Digital Art, 6) Fashion Photography 7) Fine Arts, and 8) Web design- UX. Each PC explains a certain percentage of the total variance in the data. We calculated the cumulative explained variance for each component, setting a threshold of at least 70% for GitHub and 80% for Behance. This difference arises from consulting two senior software engineers and two senior designers independently to label the fields based on the correlation matrices shown in Supplementary Materials Fig S2. They were able to label these fields and reached a consensus on the names without any collaboration.)

		Sample 1			Sample 2			Sample 3			Sample 4			Sample 5			5%				10%				25%			
	Variable	Coef.	SE.	P	Coef.	SE.	P	Coef.	SE.	P	Coef.	SE.	P	Coef.	SE.	P	Avg.	Min	Max	Sign.	Avg.	Min	Max	Sign.	Avg.	Min	Max	Sign.
Attention	Female	0.036	0.025	0.147	0.038	0.026	0.144	0.063	0.026	0.015	0.028	0.025	0.272	0.043	0.026	0.088	0.025	-0.026	0.095	20%	0.014	-0.028	0.081	9%	0.012	-0.024	0.060	6%
	Femaleness	-0.165	0.033	0.000	-0.135	0.032	0.000	-0.185	0.033	0.000	-0.174	0.032	0.000	-0.105	0.032	0.001	-0.078	-0.151	0.014	70%	-0.059	-0.113	0.014	55%	-0.022	-0.090	0.045	19%
	Female: Femaleness	0.074	0.045	0.101	0.053	0.045	0.236	0.045	0.045	0.323	0.075	0.044	0.093	0.022	0.045	0.618	0.024	-0.094	0.126	13%	0.025	-0.095	0.103	12%	0.004	-0.082	0.067	5%
	N. Own Repos (log)	-0.067	0.006	0.000	-0.066	0.006	0.000	-0.055	0.006	0.000	-0.067	0.006	0.000	-0.059	0.006	0.000	-0.055	-0.056	-0.054	90%	-0.056	-0.056	-0.055	90%	-0.056	-0.057	-0.056	90%
	N. Active Repos (log)	0.989	0.018	0.000	0.955	0.018	0.000	0.960	0.018	0.000	0.995	0.018	0.000	0.954	0.018	0.000	0.964	0.961	0.967	90%	0.965	0.962	0.967	90%	0.966	0.964	0.967	90%
	Tenure	0.167	0.004	0.000	0.169	0.004	0.000	0.169	0.004	0.000	0.171	0.004	0.000	0.170	0.004	0.000	0.171	0.170	0.172	90%	0.171	0.170	0.172	90%	0.172	0.171	0.172	90%
	Intercept	-0.878	0.024	0.000	-0.847	0.024	0.000	-0.872	0.024	0.000	-0.885	0.024	0.000	-0.877	0.024	0.000	-0.907	-0.937	-0.883	90%	-0.912	-0.938	-0.894	90%	-0.924	-0.951	-0.895	90%
	N			20000			20000			20000			20000			20000												
	R2			0.259			0.258			0.263			0.265			0.254												
Success	Female	-0.001	0.020	0.977	-0.021	0.021	0.307	-0.018	0.021	0.400	-0.009	0.021	0.664	-0.016	0.020	0.432	-0.029	-0.066	0.006	37%	-0.046	-0.085	-0.011	76%	-0.058	-0.080	-0.021	89%
	Femaleness	-0.142	0.026	0.000	-0.135	0.025	0.000	-0.165	0.026	0.000	-0.213	0.026	0.000	-0.175	0.026	0.000	-0.103	-0.146	-0.032	87%	-0.078	-0.116	-0.023	86%	-0.034	-0.082	0.014	49%
	Female: Femaleness	0.009	0.036	0.802	0.030	0.036	0.403	0.045	0.036	0.220	0.052	0.036	0.147	0.052	0.036	0.147	0.015	-0.054	0.080	7%	0.018	-0.056	0.084	6%	0.006	-0.072	0.054	4%
	N. Own Repos (log)	-0.022	0.005	0.000	-0.007	0.005	0.144	-0.015	0.005	0.001	-0.009	0.005	0.051	-0.009	0.005	0.063	-0.015	-0.016	-0.014	90%	-0.015	-0.016	-0.015	90%	-0.016	-0.016	-0.015	90%
	N. Active Repos (log)	0.585	0.014	0.000	0.534	0.014	0.000	0.557	0.014	0.000	0.568	0.015	0.000	0.540	0.014	0.000	0.559	0.556	0.563	90%	0.560	0.558	0.563	90%	0.562	0.559	0.564	90%
	Tenure	0.086	0.003	0.000	0.082	0.003	0.000	0.082	0.003	0.000	0.086	0.003	0.000	0.081	0.003	0.000	0.082	0.082	0.083	90%	0.083	0.082	0.084	90%	0.084	0.083	0.084	90%
	Intercept	-0.720	0.019	0.000	-0.683	0.019	0.000	-0.687	0.019	0.000	-0.700	0.019	0.000	-0.677	0.020	0.000	-0.705	-0.729	-0.691	90%	-0.713	-0.733	-0.700	90%	-0.726	-0.746	-0.705	90%
	N			20000			20000			20000			20000			20000												
	R2			0.170			0.163			0.164			0.173			0.159												
Survival	Female	-0.112	0.179	0.530	-0.340	0.178	0.056	-0.059	0.185	0.748	-0.175	0.178	0.326	-0.372	0.179	0.037	-0.029	-0.066	0.006	37%	-0.046	-0.085	-0.011	76	-0.058	-0.080	-0.021	89%
	Femaleness	-0.949	0.240	0.000	-0.340	0.252	0.177	-0.764	0.245	0.002	-0.661	0.248	0.008	-0.789	0.244	0.001	-0.103	-0.146	-0.032	87%	-0.078	-0.116	-0.023	86	-0.034	-0.082	0.014	49%
	Female: Femaleness	-0.104	0.319	0.744	-0.335	0.325	0.304	-0.361	0.326	0.267	-0.340	0.322	0.290	0.059	0.320	0.854	0.015	-0.054	0.080	7%	0.018	-0.056	0.084	6	0.006	-0.072	0.054	4%
	N. Own Repos (log)	-0.208	0.047	0.000	-0.271	0.049	0.000	-0.207	0.047	0.000	-0.163	0.046	0.000	-0.263	0.049	0.000	-0.015	-0.016	-0.014	90%	-0.015	-0.016	-0.015	90	-0.016	-0.016	-0.015	90%
	N. Active Repos (log)	2.168	0.157	0.000	2.530	0.161	0.000	2.157	0.155	0.000	2.059	0.155	0.000	2.540	0.164	0.000	0.559	0.556	0.563	90%	0.560	0.558	0.563	90	0.562	0.559	0.564	90%
	Tenure	-0.334	0.028	0.000	-0.335	0.028	0.000	-0.329	0.028	0.000	-0.309	0.028	0.000	-0.333	0.028	0.000	0.082	0.082	0.083	90%	0.083	0.082	0.084	90	0.084	0.083	0.084	90%
	Intercept	1.144	0.192	0.000	0.728	0.193	0.000	1.123	0.193	0.000	1.117	0.191	0.000	0.796	0.200	0.000	-0.705	-0.729	-0.691	90%	-0.713	-0.733	-0.700	90	-0.726	-0.746	-0.705	90%
	N			20000			20000			20000			20000			20000												
	AIC			20792			20519			20879			21651			20760												
	BIC			20847			20575			20934			21706			20815												

Table S2. GitHub models. OLS Regression Model tables predicting Log(Attention), Log(Success), and logistic regression models predicting staying active on GitHub one year after data collection. All variables are normalized between 0 and 1. Samples indicate independent samples containing 10,000 female and male GitHub users. Robustness results are run on 100 5%, 10%, and 25% randomly gender-swapped datasets. Avg. shows the average coefficients of the models, Min and Max indicate the lowest and highest values of 100 runs, and Sign. shows the ratio of significant coefficients.

		Sample 1			Sample 2			Sample 3			Sample 4			Sample 5			5%				10%				25%			
	Variable	Coef.	SE.	P	Coef.	SE.	P	Coef.	SE.	P	Coef.	SE.	P	Coef.	SE.	P	Avg.	Min	Max	Sign.	Avg.	Min	Max	Sign.	Avg.	Min	Max	Sign.
Attention	Female	-0.040	0.045	0.374	-0.018	0.046	0.694	-0.063	0.046	0.170	0.011	0.046	0.814	-0.046	0.046	0.312	-0.059	-0.148	0.051	16%	-0.108	-0.201	0.049	57%	-0.179	-0.312	0.003	76%
	Femaleness	-1.023	0.061	0.000	-1.019	0.059	0.000	-1.050	0.061	0.000	-1.126	0.060	0.000	-1.113	0.061	0.000	-1.021	-1.171	-0.808	90%	-0.923	-1.149	-0.688	90%	-0.279	-0.964	0.211	76%
	Female: Femaleness	0.191	0.084	0.023	0.154	0.084	0.066	0.236	0.085	0.005	0.146	0.085	0.086	0.237	0.085	0.005	0.164	-0.075	0.289	39%	0.158	-0.091	0.322	32%	0.052	-0.328	0.327	13%
	Tenure	0.154	0.003	0.000	0.154	0.003	0.000	0.154	0.003	0.000	0.153	0.003	0.000	0.155	0.003	0.000	0.651	0.628	0.683	90%	0.634	0.603	0.658	90%	0.600	0.590	0.621	90%
	N. projects (log)	0.655	0.020	0.000	0.656	0.020	0.000	0.655	0.020	0.000	0.670	0.020	0.000	0.649	0.020	0.000	0.155	0.153	0.157	90%	0.157	0.153	0.159	90%	0.161	0.158	0.162	90%
	Total activity (log)	0.649	0.038	0.000	0.655	0.038	0.000	0.677	0.038	0.000	0.623	0.038	0.000	0.674	0.037	0.000	0.692	0.612	0.741	90%	0.738	0.679	0.789	90%	0.832	0.809	0.854	90%
	Intercept	0.122	0.060	0.043	0.112	0.060	0.060	0.090	0.059	0.131	0.195	0.060	0.001	0.123	0.059	0.038	0.071	-0.063	0.213	21%	-0.013	-0.165	0.113	2%	-0.401	-0.666	-0.038	89%
	N			6000			6000			6000						6000												
	R2			0.431			0.432			0.432			0.436			0.434												
Success	Female	-0.124	0.050	0.013	-0.117	0.051	0.020	-0.149	0.051	0.003	-0.074	0.051	0.147	-0.133	0.051	0.009	-0.123	-0.243	-0.007	61%	-0.151	-0.262	-0.008	74%	-0.151	-0.297	0.047	64%
	Femaleness	-0.812	0.067	0.000	-0.803	0.066	0.000	-0.839	0.067	0.000	-0.944	0.067	0.000	-0.931	0.068	0.000	-0.831	-1.021	-0.533	90%	-0.774	-1.049	-0.556	90%	-0.223	-0.926	0.315	59%
	Female: Femaleness	0.311	0.094	0.001	0.297	0.093	0.002	0.359	0.094	0.000	0.278	0.095	0.003	0.370	0.095	0.000	0.274	0.013	0.455	80%	0.260	0.050	0.427	61%	0.071	-0.345	0.345	14%
	Tenure	0.139	0.004	0.000	0.139	0.004	0.000	0.139	0.004	0.000	0.138	0.004	0.000	0.139	0.004	0.000	1.191	1.172	1.216	90%	1.180	1.156	1.205	90%	1.155	1.145	1.178	90%
	N. projects (log)	1.192	0.022	0.000	1.192	0.022	0.000	1.192	0.022	0.000	1.207	0.022	0.000	1.189	0.022	0.000	0.139	0.137	0.141	90%	0.140	0.137	0.142	90%	0.144	0.141	0.144	90%
	Total activity (log)	0.505	0.042	0.000	0.511	0.042	0.000	0.525	0.042	0.000	0.475	0.042	0.000	0.516	0.042	0.000	0.531	0.469	0.570	90%	0.562	0.518	0.598	90%	0.634	0.607	0.661	90%
	Intercept	-0.035	0.067	0.605	-0.045	0.067	0.496	-0.054	0.066	0.416	0.058	0.067	0.386	-0.002	0.066	0.979	-0.054	-0.212	0.074	8%	-0.105	-0.244	0.034	32%	-0.425	-0.721	-0.052	88%
	N			6000			6000			6000			6000			6000												
	R2			0.474			0.474			0.474			0.478			0.476												
Survival	Female	0.042	0.142	0.768	0.045	0.144	0.752	-0.012	0.144	0.932	0.005	0.147	0.975	-0.020	0.144	0.892	-0.307	-0.515	-0.052	52%	-0.307	-0.626	-0.040	45%	-0.243	-0.676	0.276	25%
	Femaleness	-0.637	0.197	0.001	-0.718	0.192	0.000	-0.769	0.196	0.000	-0.640	0.196	0.001	-0.690	0.196	0.000	-0.727	-1.051	-0.403	90%	-0.654	-1.068	-0.259	81%	-0.218	-0.787	0.449	18%
	Female: Femaleness	0.422	0.269	0.118	0.445	0.268	0.097	0.557	0.270	0.039	0.478	0.273	0.080	0.539	0.271	0.047	0.489	0.057	0.854	33%	0.436	-0.089	1.025	23%	0.175	-0.849	1.043	9%
	Tenure	-0.105	0.011	0.000	-0.105	0.011	0.000	-0.105	0.011	0.000	-0.104	0.011	0.000	-0.104	0.011	0.000	1.165	1.148	1.194	90%	1.156	1.140	1.188	90%	1.138	1.129	1.154	90%
	N. projects (log)	1.129	0.072	0.000	1.133	0.072	0.000	1.130	0.072	0.000	1.128	0.073	0.000	1.123	0.072	0.000	-0.110	-0.112	-0.108	90%	-0.109	-0.112	-0.107	90%	-0.106	-0.108	-0.106	90%
	Total activity (log)	1.283	0.131	0.000	1.271	0.131	0.000	1.285	0.130	0.000	1.288	0.131	0.000	1.299	0.130	0.000	1.430	1.381	1.464	90%	1.453	1.370	1.492	90%	1.503	1.472	1.526	90%
	Intercept	-3.406	0.201	0.000	-3.355	0.200	0.000	-3.356	0.199	0.000	-3.410	0.202	0.000	-3.405	0.199	0.000	-3.310	-3.469	-3.127	90%	-3.358	-3.563	-3.086	90%	-3.597	-3.935	-3.314	90%
	N			6000			6000			6000			6000			6000												
	AIC			14583			14579			14578			14584			14582												
	BIC			14635			14631			14630			14635			14634												

Table S3. Behance models. OLS Regression Model tables predicting Log(Attention), Log(Success), and logistic regression models predicting staying active on Behance one year after data collection. All variables are normalized between 0 and 1. Samples indicate independent samples containing 6,000 female and male Behance users. Robustness results are run on 100 5%, 10%, and 25% randomly gender-swapped datasets. Avg. shows the average coefficients of the models, Min and Max indicate the lowest and highest values of 100 runs, and Sign. shows the ratio of significant coefficients.

	GitHub			Behance		
	Attention	Success	Survival	Attention	Success	Survival
Males' prediction at male's femaleness median	39.24	1.67	0.94	180.69	313.44	0.44
Males' prediction at female's femaleness median	22.70	1.32	0.92	87.02	171.09	0.39
Females' prediction at male's femaleness median	52.16	1.63	0.92	210.91	333.61	0.42
Females' prediction at female's femaleness median	37.62	1.24	0.88	113.81	221.01	0.39
Total female disadvantage	1.62	0.43	0.06	66.88	92.43	0.05
Female's behavior-based discrimination	14.54	0.39	0.04	97.10	112.60	0.03
Male's behavior-based discrimination	16.54	0.35	0.02	93.67	142.35	0.05
Direct discrimination	-12.92	0.04	0.01	-30.22	-20.17	0.02
% behavior-based discrimination out of total disadvantage of women	898%	90%	74%	145%	122%	60%
% direct discrimination out of total disadvantage of women	-798%	10%	26%	-45%	-22%	40%

Table S4. Medians of Attention, Success and Survival by males' and female's femaleness median, and quantified direct and behavior-based discrimination on GitHub and Behance. Total female disadvantage is calculated as the difference between predicted outcome values of men at males' femaleness median, and women's at females' femaleness median. Behavior-based gender discrimination is the slope of women's predicted outcome along the values of femaleness, specifically the difference between the the predicted outcome value of women at males' and female's femaleness median. Direct discrimination is the difference between the predicted outcome values of men and women at male's femaleness median.

		Attention		Success		Survival	
		Model 0	Model 1	Model 0	Model 1	Model 0	Model 1
GitHub	Gender	1.05	3.25	1.05	3.25	1.03	3.12
	Femaleness		3.30		3.30		3.17
	Number of repositories	1.61	1.61	1.61	1.61	1.75	1.74
	Repositories where active	1.64	1.64	1.64	1.64	1.77	1.76
	Tenure	1.07	1.09	1.07	1.09	1.07	1.08
Behance	Gender	1.02	1.73	1.02	1.73	1.03	1.76
	Femaleness		1.83		1.83		1.85
	Total Activity	1.74	1.82	1.74	1.82	1.78	1.85
	Number of projects	1.72	1.77	1.72	1.77	1.74	1.78
	Tenure	1.08	1.09	1.08	1.09	1.10	1.12

Table S5. Variance Inflation Factor (VIF) results to test for multicollinearity. If VIF equal to 1 variables are not correlated; VIF between 1 and 5: variables are moderately correlated, VIF greater than 5: variables are highly correlated.

	Dependent variable:	
	Still Active GitHub	Still Active Behance
	1	2
Female	-0.030***	-0.005
	(0.007)	(0.012)
Femaleness	-0.056***	0.306**
	(0.016)	(0.150)
Number of followers (log)	0.039***	0.076
	(0.007)	(0.070)
Number of followers (log)^2	-0.005***	0.007
	(0.002)	(0.014)
Femaleness:Number of followers (log)	-0.004	-0.328***
	(0.012)	(0.127)
(Femaleness: Number of followers (log))^2	0.003	0.059**
	(0.003)	(0.026)
Tenure	-0.034***	-0.039***
	(0.002)	(0.003)
Number of Repositories or Projects	-0.010***	0.139***
	(0.003)	(0.018)
Number of Repos where active	0.044***	
	(0.005)	
Total Activity		0.339***
		(0.029)
Constant	0.857***	-0.283***
	(0.015)	(0.093)
N	20,000	10,857
AIC	7069	14407
BIC	7156	14488
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table S6. Quadratic Models testing whether survival and the Attention(number of followers) have a U shape relationship.

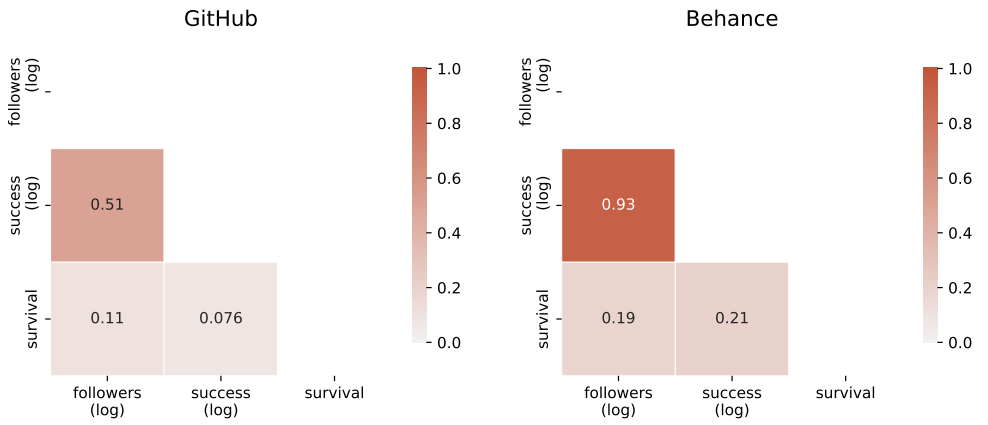

Fig. S3. Pearson Correlation of outcome variables of Behance and GitHub

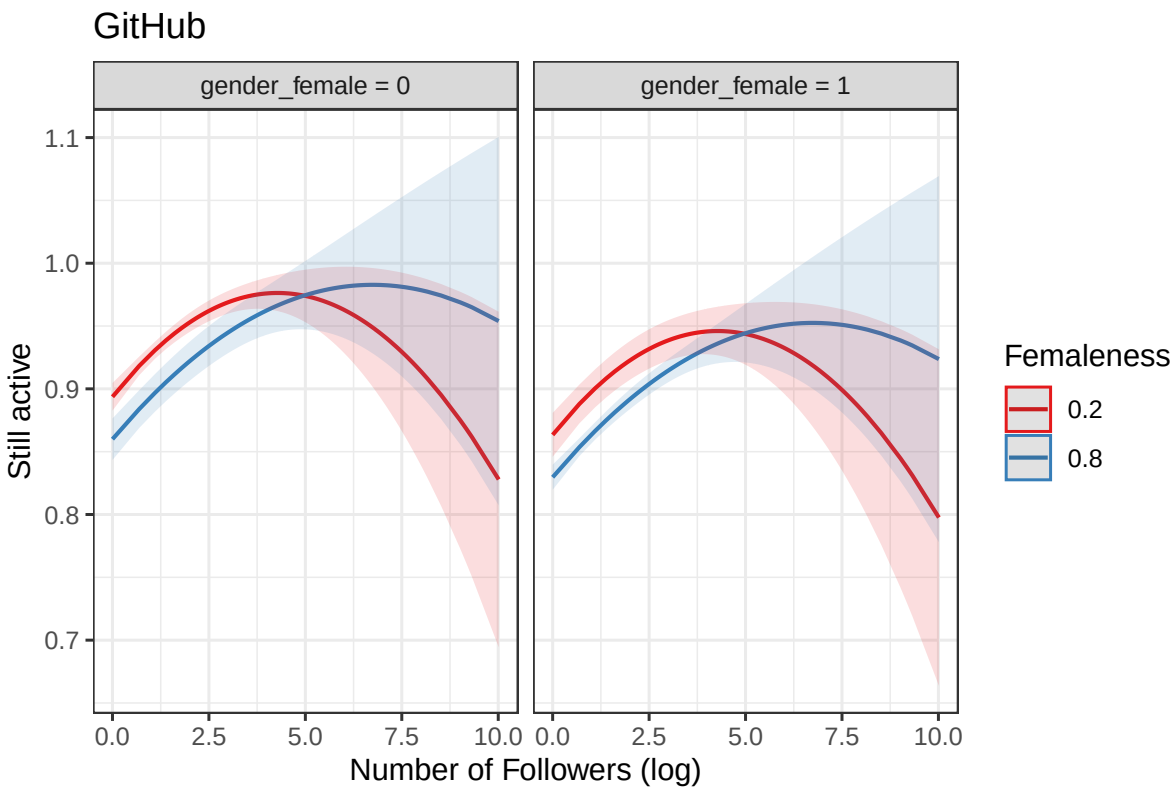

Fig. S4. Marginal Predictions of Survival on GitHub by the Number of Followers (log), Femaleness (red: low, blue: high) and gender (left: male, right: female) Based on Quadratic Models at Table S6

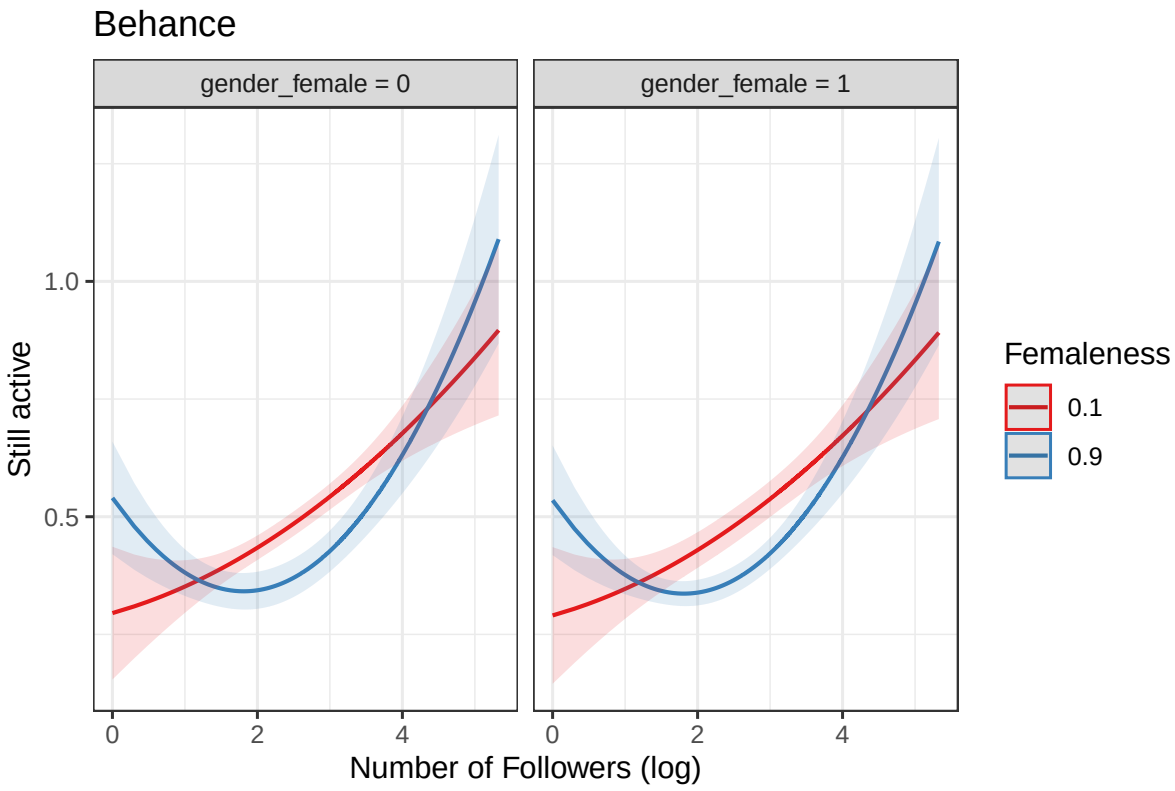

Fig. S5. Marginal Predictions of Survival on Behance by the Number of Followers (log), Femaleness (red: low, blue: high) and gender (left: male, right: female) Based on Quadratic Models at Table S6

	Dependent variable:
	Number of Merged Pull Requests GitHub
Female	0.047
	(0.054)
Femaleness	-0.211***
	(0.070)
Gender:Femaleness	0.106
	(0.096)
Tenure	-0.004
	(0.008)
Number of Repositories	-0.151***
	(0.013)
Number of Opened Pull Requests	-0.010***
	(0.001)
Number of Closed Pull Requests	0.166***
	(0.001)
Number of Repos where active	0.359***
	(0.021)
Constant	-0.766***
	(0.057)
Observations	20,000
θ	2.016*** (0.036)
Akaike Inf. Crit.	67,854.580
Note:	*p<0.1; **p<0.05; ***p<0.01

Table S7. Negative Binomial Regression Model predicting the Number if Merged Pull Requests.

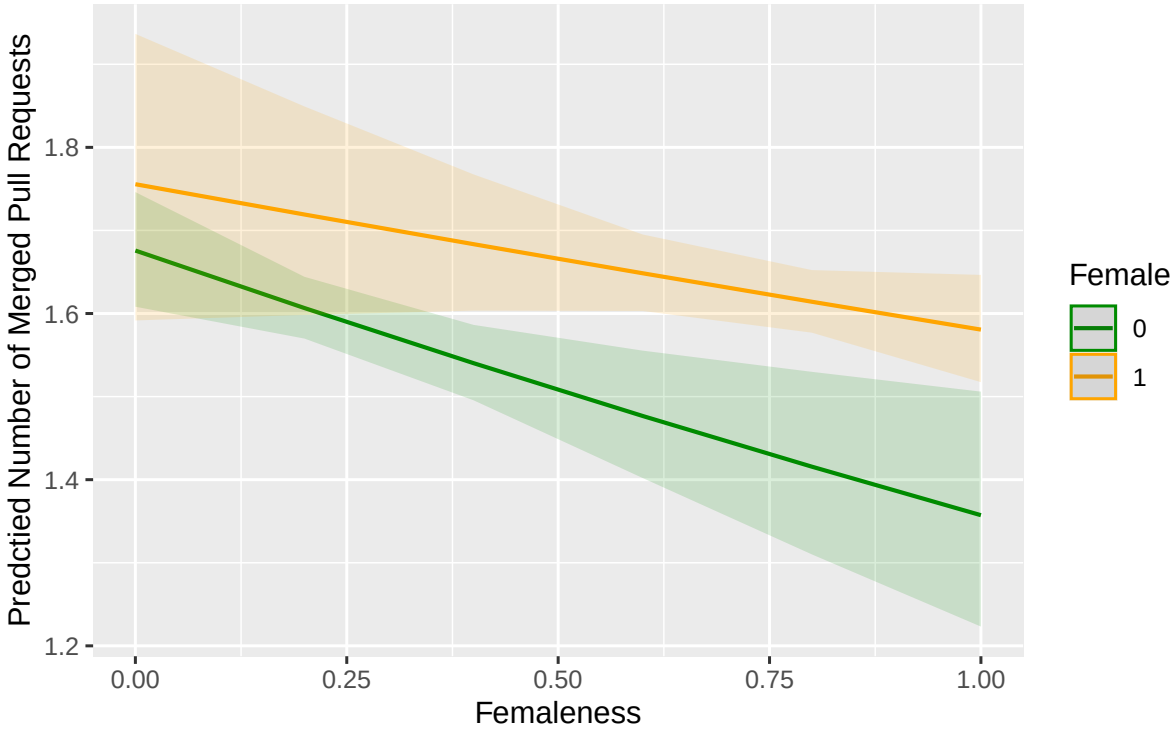

Fig. S6. Marginal Predictions of the Predicted Number of Merged Pull Requests on GitHub by Femaleness and gender (green: male, orange: female) Based on Negative Binomial Regression Model at Table S7