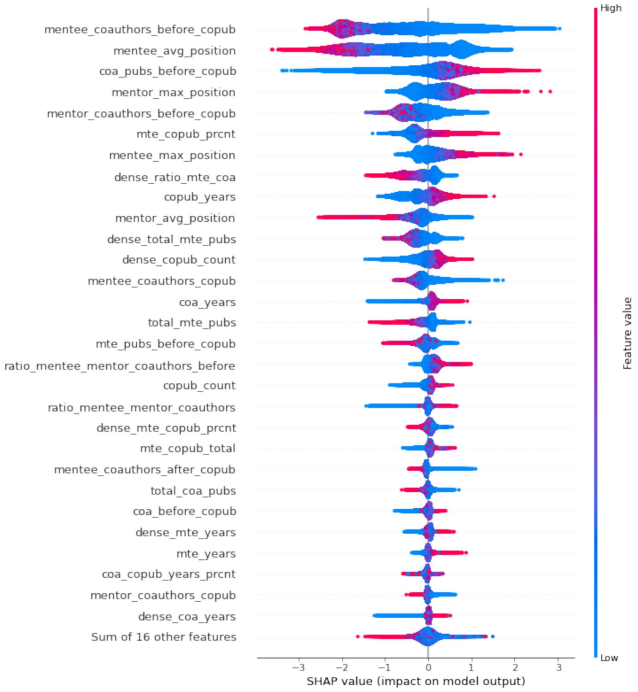

Figure 1: SHAP plot for first-stage pairwise features. The feature with the highest value is *mentee_coauthor_before_copub* interpreted as: higher value of this feature decreases the probability of mentorship.

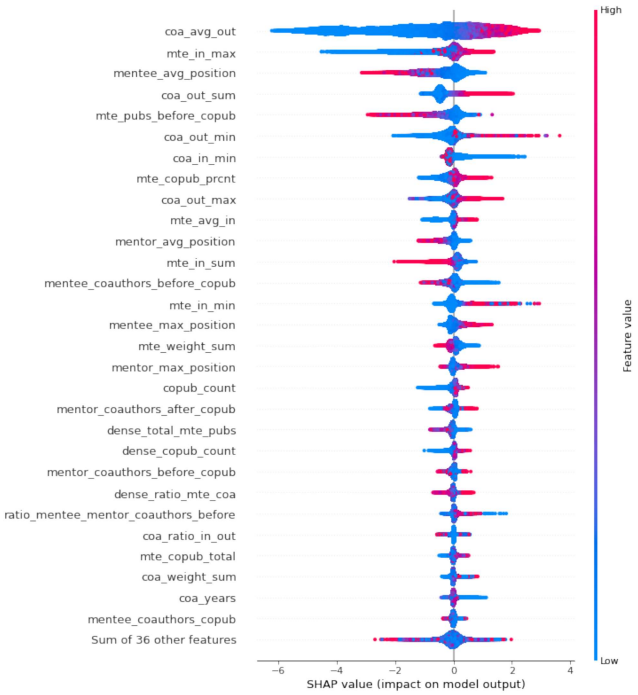

Figure 2: SHAP plot for second-stage graph features. The feature with the highest value is *coa_avg_out* (mean mentorship score) interpreted as: higher value of this feature increases the probability of mentorship.

A MENTOR AND MENTEE DISCOVERY

Here are some examples for scholars with high mentorship or menteeship sum score:

- **Author Name**: Yuen Kwok-yung
  h-index: 135
  Mentorship sum: 204.6
  Field of study: Medicine
  "He led a team identifying the SARS coronavirus that caused the SARS pandemic of 2003–4, and traced its genetic origins to wild bats. During the ongoing COVID-19 pandemic, he has acted as expert adviser to the Hong Kong government."[14]
- **Author Name**: Michael J. Black
  h-index: 96
  Menteeship sum: 4.8
  Field of study: Computer Science
  "[T]he only researcher in the field to have won all three major test-of-time prizes in computer vision: the 2010 Koenderink Prize at the European Conference on Computer Vision (ECCV), the 2013 Helmholtz Prize at the International Conference on Computer Vision (ICCV), and the 2020 Longuet-Higgins Prize at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)."[15]
- **Author Name**: Kaushik Roy
  h-index: 96
  Mentorship sum: 127.2
  Field of study: Electrical Engineering
  "He has supervised 91 Ph.D. dissertations and co-authored two books on Low Power CMOS VLSI Design (John Wiley & McGraw Hill)."[16]
- **Author Name**: Peter Nijkamp
  h-index: 96
  Mentorship sum: 169.8
  Field of study: Economics
  "He is ranked among the top 100 economists in the world according to IDEAS/RePEc, and is by far the most prolific economist."[17]

[14] https://en.wikipedia.org/wiki/Yuen_Kwok-yung
[15] https://en.wikipedia.org/wiki/Michael_J._Black
[16] https://en.wikipedia.org/wiki/Kaushik_Roy
[17] https://en.wikipedia.org/wiki/Peter_Nijkamp

```
HYPEROPT_SPACE = {
    "learning_rate": hp.choice("learning_rate", [0.1, 0.05, 0.01, 0.005, 0.001]),
    "num_leaves": scope.int(2 ** hp.quniform("num_leaves", 2, 7, 1)),
    "colsample_bytree": hp.quniform("colsample_bytree", 0.4, 1, 0.1),
    "subsample": hp.quniform("subsample", 0.4, 1, 0.1),
    "min_child_samples": scope.int(2 ** hp.quniform("min_child_samples", 0, 7, 1)),
    "min_child_weight": 10 ** hp.quniform("min_child_weight", -6, 0, 1),
    "reg_alpha": hp.choice("reg_alpha", [0, 10 ** hp.quniform("reg_alpha_pos", -6, 1, 1)]),
    "reg_lambda": hp.choice("reg_lambda", [0, 10 ** hp.quniform("reg_lambda_pos", -6, 1, 1)]),
    "max_depth": scope.int(hp.choice("max_depth",
        [-1, 2 ** hp.quniform("max_depth_pos", 1, 4, 1)]
    ))
}
```

Figure 3: We optimized the LightGBM models for both stages by running `hyperopt` for 50 iterations over the search space above.

Table 4: Description for features extracted from publication information retrieved from Semantic Scholar.

Category	Name	Description
publication	*copub_count*	total papers published together
	total_mte_pubs	total publications of the mentee till copub end date
	total_coa_pubs	total publications of the coauthor till copub end date
	mte_copub_total	# of papers published by mentee in copub period
	coa_copub_total	# of papers published by coauthor in copub period
	mte_copub_prcnt	ratio of copub_count to mte_copub_total
	coa_copub_prcnt	ratio of copub_count to coa_copub_total
	ratio_mte_coa	ratio of total_mte_pubs to total_coa_pubs
	copub_years	# of years of collaboration
	mte_years	mentee publication years till copub end date
	coa_years	coauthor publication years till copub end date
	mte_copub_years_prcnt	ratio of copub_years to mte_years
	coa_copub_years_prcnt	ratio of copub_years to coa_years
	dense_mte_copub_total	# of papers published by mentee in dense copub period
	dense_coa_copub_total	# of papers published by coauthor in dense copub period
	dense_total_coa_pubs	total publications of the coauthor till dense copub end date
	dense_total_mte_pubs	total publications of the mentee till dense copub end date
	dense_copub_count	total papers published together during the dense copub period
	dense_mte_copub_prcnt	ratio of dense copub_count to mte_copub_total
	dense_coa_copub_prcnt	ratio of dense copub_count to coa_copub_total
	dense_ratio_mte_coa	ratio of dense_total_mte_pubs to dense_total_coa_pubs
	dense_mte_years	mentee publication years till dense copub end date
	dense_coa_years	coauthor publication years till dense copub end date
	dense_mte_copub_years_prcnt	ratio of dense copub_years to mte_years
	dense_coa_copub_years_prcnt	ratio of dense copub_years to coa_years
	coa_pubs_before_copub	coauthor publication count before co-publication period
	mte_pubs_before_copub	mentee publication count before co-publication period
co-author	*mentee_coauthors_before_copub*	number of coauthors of mentee before copublication period
	mentor_coauthors_before_copub	number of coauthors of mentor before copublication period
	mentee_coauthors_after_copub	number of coauthors of mentee at the end of copublication period
	mentor_coauthors_after_copub	number of coauthors of mentor at the end of copublication period
	mentor_coauthors_copub	number of coauthors of mentee during copublication period
	ratio_mentee_mentor_coauthors	mentee_coauthors_copub divided by mentor_coauthors_copub
	ratio_mentee_mentor_coauthors_before	mentee_coauthors_before_copub divided by mentor_coauthors_before_copub
	ratio_mentee_mentor_coauthors_after	mentee_coauthors_after_copub divided by mentor_coauthors_after_copub
position	*mentee_min_position*	min position of authorship of mentee in copublications
	mentor_min_position	min position of authorship of mentor in copublications
	mentee_max_position	max position of authorship of mentee in copublications
	mentor_max_position	max position of authorship of mentor in copublications
	mentee_avg_position	avg position of authorship of mentee in copublications
	mentor_avg_position	avg position of authorship of mentor in copublications
graph	*coa_out_min*	coauthor out-edge min weight
	coa_in_min	coauthor in-edge min weight
	mte_out_min	mentee out-edge min weight
	mte_in_min	mentee in-edge min weight
	coa_out_max	coauthor out-edge max weight
	coa_in_max	coauthor in-edge max weight
	mte_out_max	mentee out-edge max weight
	mte_in_max	mentee in-edge max weight
	coa_out_sum	sum of out-edge weights for coauthor
	coa_in_sum	sum of in-edge weights for coauthor
	mte_out_sum	sum of out-edge weights for mentee
	mte_in_sum	sum of in-edge weights for mentee
	mte_weight_sum	mte_out_sum + mte_in_sum
	coa_weight_sum	coa_out_sum + coa_in_sum
	mte_avg_in	average of in-edge weights for mentee
	mte_avg_out	average of out-edge weights for mentee
	coa_avg_in	average of in-edge weights for coauthor
	coa_avg_out	average of out-edge weights for coauthor
	mte_ratio_in_out	ratio of mte_in_sum to mte_out_sum
	coa_ratio_in_out	ratio of coa_in_sum to coa_out_sum

Table 5: Complete results from the fitted `statsmodels` [15] negative binomial GLM. Note that for field of study, "unknown" was used as the reference 0 coefficient category. Note further that before fitting, 99 percentile outlier removal was performed on paper count, citation count and the two sum covariates.

Covariate	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]
Intercept	-1.634325	0.003414	-478.698723	0.000000e+00	-1.641016	-1.627633
Agricultural And Food Sciences	0.140752	0.003383	41.609452	0.000000e+00	0.134122	0.147382
Art	-0.008203	0.007007	-1.170767	2.416926e-01	-0.021937	0.005530
Biology	0.140790	0.003033	46.423433	0.000000e+00	0.134846	0.146734
Business	0.089883	0.004022	22.348532	1.247803e-110	0.082000	0.097765
Chemistry	0.164536	0.003305	49.790898	0.000000e+00	0.158059	0.171013
Computer Science	0.130829	0.003319	39.413902	0.000000e+00	0.124324	0.137335
Economics	0.103034	0.004238	24.310960	1.501202e-130	0.094728	0.111341
Education	0.085535	0.004309	19.849071	1.122578e-87	0.077089	0.093981
Engineering	0.003292	0.003321	0.991141	3.216166e-01	-0.003218	0.009801
Environmental Science	0.108753	0.003325	32.708699	1.174727e-234	0.102236	0.115270
Geography	0.150294	0.018407	8.165211	3.208731e-16	0.114218	0.186371
Geology	0.129722	0.004449	29.154761	7.270255e-187	0.121001	0.138442
History	0.007579	0.007482	1.013009	3.110559e-01	-0.007085	0.022244
Law	0.054235	0.010239	5.296658	1.179416e-07	0.034166	0.074303
Linguistics	0.055035	0.010163	5.415055	6.126987e-08	0.035115	0.074954
Materials Science	0.131333	0.003322	39.528443	0.000000e+00	0.124821	0.137845
Mathematics	0.172346	0.004922	35.015816	1.292697e-268	0.162699	0.181993
Medicine	0.116418	0.002956	39.380783	0.000000e+00	0.110624	0.122212
Philosophy	0.028260	0.012606	2.241714	2.497989e-02	0.003552	0.052968
Physics	0.109517	0.003219	34.027028	8.877297e-254	0.103209	0.115825
Political Science	0.096860	0.006898	14.040819	8.769281e-45	0.083339	0.110381
Psychology	0.159464	0.003762	42.383743	0.000000e+00	0.152090	0.166838
Sociology	0.102675	0.009682	10.604561	2.837925e-26	0.083698	0.121651
Paper Count (2nd Quintile)	0.207493	0.001534	135.256638	0.000000e+00	0.204487	0.210500
Paper Count (3rd Quintile).	0.330930	0.001603	206.492021	0.000000e+00	0.327788	0.334071
Paper Count (4th Quintile).	0.440313	0.001742	252.711182	0.000000e+00	0.436898	0.443728
Paper Count (5th Quintile)	0.696573	0.002017	345.349712	0.000000e+00	0.692620	0.700526
Citation Count (2nd Quintile)	1.828905	0.001861	982.650323	0.000000e+00	1.825257	1.832553
Citation Count (3rd Quintile)	2.461976	0.001868	1317.670812	0.000000e+00	2.458314	2.465638
Citation Count (4th Quintile)	2.917600	0.001933	1509.490243	0.000000e+00	2.913812	2.921388
Citation Count (5th Quintile)	3.509647	0.002063	1700.923846	0.000000e+00	3.505602	3.513691
Menteeship Sum (2nd Quntile)	0.057742	0.001434	40.266133	0.000000e+00	0.054931	0.060552
Menteeship Sum (3rd Quntile)	0.080814	0.001419	56.963054	0.000000e+00	0.078034	0.083595
Menteeship Sum (4th Quntile)	0.105018	0.001429	73.485605	0.000000e+00	0.102217	0.107818
Menteeship Sum (5th Quntile)	0.138352	0.001477	93.700988	0.000000e+00	0.135458	0.141246
Menteeship Mean (2nd Quntile)	-0.020535	0.001287	-15.950493	2.826550e-57	-0.023058	-0.018012
Menteeship Mean (3rd Quntile)	-0.032761	0.001339	-24.473922	2.800337e-132	-0.035385	-0.030138
Menteeship Mean (4th Quntile)	-0.035933	0.001395	-25.765417	2.165674e-146	-0.038666	-0.033200
Menteeship Mean (5th Quntile)	-0.023244	0.001444	-16.100688	2.522834e-58	-0.026073	-0.020414
Mentorship Sum (2nd Quntile)	-0.005399	0.001882	-2.868472	4.124601e-03	-0.009088	-0.001710
Mentorship Sum (3rd Quntile)	-0.012412	0.002341	-5.301161	1.150686e-07	-0.017001	-0.007823
Mentorship Sum (4th Quntile)	-0.012612	0.002635	-4.785979	1.701561e-06	-0.017776	-0.007447
Mentorship Sum (5th Quntile)	0.098365	0.002937	33.497248	5.285543e-246	0.092609	0.104120
Mentorship Mean (2nd Quntile)	0.032508	0.001870	17.386650	1.041458e-67	0.028843	0.036172
Mentorship Mean (3rd Quntile)	0.046286	0.002297	20.147111	2.852842e-90	0.041783	0.050789
Mentorship Mean (4th Quntile)	0.048493	0.002584	18.766996	1.406156e-78	0.043428	0.053557
Mentorship Mean (5th Quntile)	0.025579	0.002768	9.240218	2.459948e-20	0.020153	0.031005