Recommendation distribution discussion of MIRRec

To prevent workload concentration on a limited number of core reviewers is an essential aspect that requires consideration beyond accuracy in reviewer recommendations. We employ Recommendation Distribution (RD) [22] as a metric to quantify the level of concentration in workload. A higher RD value signifies a more evenly distributed workload, leading to a more diverse set of recommended reviewers. Conversely, a lower RD value indicates poorer diversity in recommendations, with a higher likelihood of workload congestion among a few reviewers.

As shown in Table, we present the RD results for all recommenders on the dataset. Overall, MIRRec stands out as the top performer, followed by cHRev. They consistently rank first and second across all data, outperforming HGRec, while CN lags behind and RevFinder performs least, aligning with the findings in [22]. Specifically, in terms of the average RD for Top-1 recommendation results, cHRev performs the best, closely followed by MIRRec. However, MIRRec outperforms all the recommenders in Top-3 and Top-5 average RD. Although cHRev ranks next to MIRRec in RD performance, it consistently demonstrates poor performance in ACC and MRR compared to MIRRec. These results indicate that while maintaining excellent accuracy, MIRRec, in contrast to other baseline methods, is more adept at considering the balanced distribution of workload, thereby providing more diverse recommendation outcomes.

COMPARISON OF RD ACROSS DIFFERENT METHODS

Project	RevFinder			cHRev			CN			HGRec			MIRRec		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Bitcoin	0.147	0.309	0.388	**0.454**	**0.546**	**0.595**	0.309	0.451	0.523	0.237	0.399	0.477	0.224	0.437	0.523
Electron	0.164	0.385	0.450	0.264	0.428	0.495	0.312	0.432	0.486	0.272	0.415	0.467	**0.344**	**0.508**	**0.572**
Opencv	0.003	0.340	0.461	0.174	0.473	0.576	0.289	0.503	0.595	0.361	0.569	0.642	**0.388**	**0.661**	**0.750**
XBMC	0.145	0.376	0.489	**0.538**	0.638	0.676	0.422	0.588	0.661	0.372	0.564	0.632	0.443	**0.648**	**0.719**
React	0.103	0.285	0.383	0.257	0.408	0.491	0.310	0.396	0.472	0.358	0.432	0.489	**0.446**	**0.544**	**0.620**
Angular	0.285	0.423	0.492	0.502	0.579	0.622	0.453	0.530	0.578	0.495	0.548	0.582	**0.569**	**0.632**	**0.671**
Django	0.052	0.288	0.369	**0.301**	**0.476**	**0.567**	0.186	0.383	0.506	0.163	0.364	0.458	0.172	0.439	0.554
Symfony	0.013	0.222	0.317	**0.341**	**0.471**	**0.538**	0.201	0.369	0.449	0.223	0.368	0.445	0.243	0.419	0.514
Rails	0.089	0.319	0.391	0.381	0.517	0.575	0.277	0.431	0.508	0.373	0.503	0.560	**0.448**	**0.605**	**0.679**
Scala	0.124	0.388	0.497	**0.397**	**0.541**	**0.624**	0.311	0.503	0.602	0.252	0.456	0.557	0.271	0.506	**0.624**
Average	0.112	0.334	0.424	**0.361**	0.508	0.576	0.307	0.459	0.538	0.311	0.462	0.531	0.355	**0.540**	**0.623**

Note: 'Average', font and colors have the same meanings as in TableIV.

cHRev exhibits better performance in Top-1 RD than MIRRec, mainly because it scores candidate reviewers considering their workload and past review frequency. While MIRRec shows better performance in Top-3 and Top-5 RD, it may benefit from multiplex relation learning by adjusting the importance (α) of reviewers in review history, thereby increasing the diversity of reviewer candidates.

[22] G. Rong, Y. Zhang, L. Yang, F. Zhang, H. Kuang, and H. Zhang, "Modeling review history for reviewer recommendation:A Hypergraph Approach," in Proceedings of the 44th International Conference on Software Engineering (ICSE), pp. 1381-1392, 2022.