A Additional Tables and Figures

A.1 Tables

We provide seven additional tables showing results for recall@10 referenced in Sections 5.1 (Table A1), the full results for Table 3 in Section 5.2 and the corresponding results for recall@10 (Tables A2 and A3), the complete results for Table 4 for both nDCG@10 and recall@10 (Tables A4 and A5), the results for general recommenders with temporal LOO omitted from Section 5.3 (Table A6), and the results for general recommenders with split-by-timepoint LOO for recall@10 referenced in Section 5.3 (Table A7).

Table A1: Performance of sequential recommenders with temporal LOO and split-by-timestamp LOO, and the relative differences in Recall@10. Bold and underline indicate the best and second best performing models for each combination of evaluation strategy, evaluation metric and data set.

	Model	Popularity-sampled			Unsampled		
		T-LOO (Recall@10)	ST-LOO (Recall@10)	Perf. diff (%)	T-LOO (Recall@10)	ST-LOO (Recall@10)	Perf. diff (%)
ML-1m	FPMC	0.5439	0.2163	-60.23%	0.1934	0.0328	-83.04%
	GRU4Rec	0.6790	0.2525	-62.81%	0.2886	0.0320	-88.91%
	Caser	0.5899	0.2231	-62.18%	0.1935	0.0177	-90.85%
	SASRec	**0.6952**	0.2424	-65.13%	**0.3141**	0.0354	-88.73%
	BERT4Rec	0.6755	0.2096	-68.97%	0.2839	0.0227	-92.00%
	S^3-Rec	0.6874	**0.2780**	-59.56%	0.3121	**0.0469**	-84.97%
	LightSANs	0.6705	0.2197	-67.23%	0.2647	0.0311	-88.25%
	SINE	0.4755	0.1667	-64.94%	0.0990	0.0160	-83.84%
	FEARec	0.6671	0.2071	-68.96%	0.2469	0.0278	-88.74%
Yelp	FPMC	0.6305	0.4400	-30.21%	0.0354	0.0150	-57.63%
	GRU4Rec	0.6808	0.5055	-25.75%	0.0451	0.0201	-55.43%
	Caser	0.6549	0.4628	-29.33%	0.0334	0.0172	-48.50%
	SASRec	0.6985	0.5063	-27.52%	**0.0634**	**0.0293**	-53.79%
	BERT4Rec	0.6591	0.4848	-26.45%	0.0406	0.0192	-52.71%
	S^3-Rec	-	-	-	-	-	-
	LightSANs	**0.7132**	0.5228	-26.70%	0.0620	0.0290	-53.23%
	SINE	0.6786	0.5318	-21.63%	0.0528	0.0283	-46.40%
	FEARec	0.7073	**0.5388**	-23.82%	0.0593	0.0270	-54.47%
Steam	FPMC	0.1772	0.1466	-17.27%	0.1085	0.1007	-7.19%
	GRU4Rec	0.1987	**0.1623**	-18.32%	0.1211	0.0911	-24.77%
	Caser	0.1917	0.1526	-20.4%	0.1273	0.0983	-22.78%
	SASRec	0.2026	0.1603	-20.88%	0.1289	0.0954	-25.99%
	BERT4Rec	0.1861	0.1522	-18.22%	0.1251	0.0951	-23.98%
	S^3-Rec	0.2069	0.1555	-24.84%	0.1259	0.0915	-27.32%
	LightSANs	**0.2371**	0.1572	-33.7%	**0.1318**	0.0955	-27.54%
	SINE	0.1942	0.1496	-22.97%	0.1201	**0.0997**	-16.99%
	FEARec	0.2062	0.1601	-22.36%	0.1305	0.0983	-24.67%
Beauty	FPMC	0.3218	0.1141	-64.54%	0.0645	0.0139	-78.45%
	GRU4Rec	0.3271	0.1780	-45.58%	0.0574	0.0176	-69.34%
	Caser	0.3038	0.1462	-51.88%	0.0356	0.0132	-62.92%
	SASRec	0.3772	0.1903	-49.55%	0.0843	**0.0238**	-71.77%
	BERT4Rec	0.2908	0.1485	-48.93%	0.0360	0.0108	-70.00%
	S^3-Rec	0.3653	0.1793	-50.92%	0.0786	0.0177	-77.48%
	LightSANs	**0.4077**	**0.1941**	-52.39%	**0.0875**	0.0223	-74.51%
	SINE	0.3832	0.1089	-71.58%	0.0669	0.0043	-93.57%
	FEARec	0.4050	0.1309	-67.68%	0.0861	0.0183	-78.75%

Table A2: Performance of sequential recommenders with temporal LOO split into active and inactive users in nDCG@10. Active users have interactions after the cut-off date used in split-by-timepoint LOO. All other users are inactive.

	Model	Popularity-sampled + Temporal LOO			Unsampled + Temporal LOO		
		Active users (nDCG@10)	Inactive users (nDCG@10)	Perf. diff (%)	Active users (nDCG@10)	Inactive users (nDCG@10)	Perf. diff (%)
ML-1m	FPMC	0.1818	0.3699	+103.47%	0.0486	0.1158	+138.27%
	GRU4Rec	0.2377	0.5068	+113.21%	0.0739	0.1842	+149.26%
	Caser	0.1785	0.4044	+126.55%	0.0429	0.1118	+160.61%
	SASRec	0.2600	0.5271	+102.73%	0.0789	0.1977	+150.57%
	BERT4Rec	0.2446	0.5018	+105.15%	0.0732	0.1754	+139.62%
	S^3-Rec	0.2645	0.5192	+96.29%	0.0756	0.1975	+161.24%
	LightSANs	0.2456	0.4942	+101.22%	0.0749	0.1630	+117.62%
	SINE	0.1296	0.2894	+123.30%	0.0178	0.0496	+178.65%
	FEARec	0.2376	0.4915	+106.86%	0.0567	0.1462	+157.85%
Yelp	FPMC	0.3444	0.3981	+15.59%	0.0164	0.0231	+40.85%
	GRU4Rec	0.3913	0.4567	+16.71%	0.0185	0.0292	+57.84%
	Caser	0.3624	0.4215	+16.31%	0.0129	0.0216	+67.44%
	SASRec	0.4167	0.4782	+14.76%	0.0342	0.0414	+21.05%
	BERT4Rec	0.3711	0.4360	+17.49%	0.0157	0.0271	+72.61%
	S^3-Rec	-	-	-	-	-	-
	LightSANs	0.4277	0.4896	+14.47%	0.0319	0.0401	+25.71%
	SINE	0.3926	0.4631	+17.96%	0.0251	0.0351	+39.84%
	FEARec	0.4165	0.4814	+15.58%	0.0316	0.0392	+24.05%
Steam	FPMC	0.0836	0.1000	+19.62%	0.0543	0.0742	+36.65%
	GRU4Rec	0.0986	0.1133	+14.91%	0.0619	0.0752	+21.49%
	Caser	0.0918	0.1171	+27.56%	0.0636	0.0815	+28.14%
	SASRec	0.1017	0.1153	+13.37%	0.0666	0.0772	+15.92%
	BERT4Rec	0.0913	0.1084	+18.73%	0.0642	0.0825	+28.50%
	S^3-Rec	0.1034	0.1108	+7.16%	0.0643	0.0740	+15.09%
	LightSANs	0.1233	0.1326	+7.54%	0.0741	0.0875	+18.08%
	SINE	0.0952	0.1159	+21.74%	0.0605	0.0769	+27.11%
	FEARec	0.1025	0.1177	+14.83%	0.0669	0.0784	+17.19%
Beauty	FPMC	0.1683	0.2117	+25.79%	0.0249	0.0478	+91.97%
	GRU4Rec	0.1735	0.2146	+23.69%	0.0205	0.0436	+112.68%
	Caser	0.1510	0.1708	+13.11%	0.0123	0.0252	+104.88%
	SASRec	0.2111	0.2525	+19.61%	0.0315	0.0530	+68.25%
	BERT4Rec	0.1420	0.1753	+23.45%	0.0129	0.0242	+87.60%
	S^3-Rec	0.2012	0.2426	+20.58%	0.0320	0.0559	+74.69%
	LightSANs	0.2292	0.2676	+16.75%	0.0341	0.0579	+69.79%
	SINE	0.1983	0.2410	+21.53%	0.0217	0.0421	+94.01%
	FEARec	0.2304	0.2643	+14.71%	0.0318	0.0542	+70.44%

Table A3: Performance of sequential recommenders with temporal LOO split into active and inactive users in Recall@10. Active users have interactions after the cut-off date used in split-by-timepoint LOO. All other users are inactive.

	Model	Popularity-sampled + Temporal LOO			Unsampled + Temporal LOO		
		Active users (Recall@10)	Inactive users (Recall@10)	Perf. diff (%)	Active users (Recall@10)	Inactive users (Recall@10)	Perf. diff (%)
ML-1m	FPMC	0.3153	0.5782	+83.38%	0.0782	0.2117	+170.72%
	GRU4Rec	0.3791	0.7211	+90.21%	0.1395	0.3252	+133.12%
	Caser	0.3213	0.6320	+96.7%	0.0830	0.2112	+154.46%
	SASRec	0.4103	0.7312	+78.21%	0.1360	0.3425	+151.84%
	BERT4Rec	0.3959	0.7195	+81.74%	0.1300	0.3085	+137.31%
	S^3-Rec	0.4212	0.7278	+72.79%	0.1215	0.3425	+181.89%
	LightSANs	0.3875	0.7184	+85.39%	0.1479	0.2934	+98.38%
	SINE	0.2599	0.5087	+95.73%	0.0361	0.1090	+201.94%
	FEARec	0.3875	0.7095	+83.10%	0.1119	0.2684	+139.86%
Yelp	FPMC	0.5858	0.6694	+14.27%	0.0292	0.0433	+48.29%
	GRU4Rec	0.6333	0.7225	+14.08%	0.0355	0.0572	+61.13%
	Caser	0.6107	0.6925	+13.39%	0.0257	0.0431	+67.70%
	SASRec	0.6538	0.7371	+12.74%	0.0546	0.0745	+36.45%
	BERT4Rec	0.6107	0.7003	+14.67%	0.0309	0.0530	+71.52%
	S^3-Rec	-	-	-	-	-	-
	LightSANs	0.6697	0.7506	+12.08%	0.0534	0.0730	+36.70%
	SINE	0.6305	0.7226	+14.61%	0.0433	0.0649	+49.88%
	FEARec	0.6624	0.7472	+12.80%	0.0509	0.0701	+37.72%
Steam	FPMC	0.1756	0.1760	+0.23%	0.1083	0.1151	+6.28%
	GRU4Rec	0.2001	0.2122	+6.05%	0.1210	0.1266	+4.63%
	Caser	0.1911	0.2270	+18.80%	0.1268	0.1464	+15.46%
	SASRec	0.2029	0.2155	+6.21%	0.1288	0.1316	+2.17%
	BERT4Rec	0.1861	0.1990	+6.93%	0.1247	0.1431	+14.77%
	S^3-Rec	0.2066	0.2155	+4.31%	0.1257	0.1332	+5.976%
	LightSANs	0.2369	0.2434	+2.74%	0.1315	0.1447	+10.05%
	SINE	0.1928	0.2253	+16.86%	0.1196	0.1398	+16.89%
	FEARec	0.2063	0.2319	+12.41%	0.1301	0.1480	+13.76%
Beauty	FPMC	0.2864	0.3403	+18.82%	0.0461	0.0855	+85.47%
	GRU4Rec	0.2984	0.3467	+16.19 %	0.0404	0.0770	+90.59%
	Caser	0.2865	0.3010	+5.06%	0.0249	0.0480	+92.77%
	SASRec	0.3454	0.3941	+14.10%	0.0656	0.1058	+61.28%
	BERT4Rec	0.2596	0.3060	+17.87%	0.0259	0.0475	+83.40%
	S^3-Rec	0.3349	0.3819	+14.03%	0.0603	0.0996	+65.17%
	LightSANs	0.3754	0.4202	+11.93%	0.0686	0.1091	+59.04%
	SINE	0.3522	0.3972	+12.78%	0.0465	0.0904	+94.41%
	FEARec	0.3768	0.4097	+8.73%	0.0658	0.1094	+66.26%

Table A4: Performance of sequential recommenders using temporal LOO with subsampled training data (indicated as T-LOO′) and split-by-timepoint LOO (indicated as ST-LOO), and the relative differences in nDCG@10. Bold and underline indicate the best and second best performing models for each combination of data splitting strategy, evaluation metric and data set. "-" indicates model could not be run due to data set size.

	Model	Popularity-sampled			Unsampled		
		T-LOO′ (nDCG@10)	ST-LOO (nDCG@10)	Perf. diff (%)	T-LOO′ (nDCG@10)	ST-LOO (nDCG@10)	Perf. diff (%)
ML-1m	FPMC	0.3412	0.1118	-67.23%	0.1078	0.0158	-85.34%
	GRU4Rec	0.4528	0.1251	-72.37%	0.1537	0.0138	-91.02%
	Caser	0.3885	0.1078	-72.25%	0.1030	0.0081	-92.14%
	SASRec	0.4387	0.1250	-71.51%	0.1418	0.0174	-87.73%
	BERT4Rec	0.4409	0.0968	-78.04%	0.1411	0.0110	-92.20%
	S^3-Rec	**0.4714**	**0.1410**	-70.09%	**0.1634**	**0.0231**	-85.86%
	LightSANs	0.4383	0.1100	-74.90%	0.1292	0.0137	-89.40%
	SINE	0.2353	0.0782	-66.77%	0.0363	0.0064	-82.37%
	FEARec	0.4447	0.1032	-76.79%	0.1288	0.0106	-91.77%
Yelp	FPMC	0.3020	0.2395	-20.70%	0.0089	0.0082	-7.87%
	GRU4Rec	0.4033	0.3024	-25.02%	0.0155	0.0101	-34.84%
	Caser	0.3697	0.2688	-27.29%	0.0118	0.0091	-22.88%
	SASRec	0.4262	0.3066	-28.06%	**0.0189**	**0.0175**	-7.41%
	BERT4Rec	0.3823	0.2827	-26.05%	0.0164	0.0094	-42.68%
	S^3-Rec	-	-	-	-	-	-
	LightSANs	**0.4358**	0.3191	-26.78%	0.0187	0.0164	-12.3%
	SINE	0.4219	0.3215	-23.80%	0.0166	0.0155	-6.63%
	FEARec	0.3559	**0.3243**	-8.88%	0.0128	0.0155	21.09%
Steam	FPMC	0.0708	0.0745	5.23%	0.0449	**0.0556**	23.83%
	GRU4Rec	0.0978	0.0791	-19.12%	0.0603	0.0501	-16.92%
	Caser	0.0871	0.0761	-12.63%	0.0587	0.0547	-6.81%
	SASRec	0.0974	0.0789	-18.99%	0.0481	0.0533	10.81%
	BERT4Rec	0.0854	0.0761	-10.89%	0.0553	0.0540	-2.35%
	S^3-Rec	0.0944	0.0777	-17.69%	0.0491	0.0510	3.87%
	LightSANs	**0.1081**	0.0782	-27.66%	**0.0676**	0.0537	-20.56%
	SINE	0.1047	0.0750	-28.37%	0.0537	0.0549	2.23%
	FEARec	0.0993	**0.0793**	-20.14%	0.0618	0.0535	-13.43%
Beauty	FPMC	0.1673	0.0598	-64.26%	0.0405	0.0063	-84.44%
	GRU4Rec	0.0995	0.0910	-8.54%	0.0128	0.0084	-34.38%
	Caser	0.0599	0.0714	19.20%	0.0101	0.0057	-43.56%
	SASRec	0.2323	0.1002	-56.87%	0.0439	**0.0102**	-76.77%
	BERT4Rec	0.1701	0.0775	-54.44%	0.0184	0.0050	-72.83%
	S^3-Rec	0.2286	0.0936	-59.06%	0.0450	0.0079	-82.44%
	LightSANs	**0.2474**	**0.1061**	-57.11%	**0.0455**	0.0092	-79.78%
	SINE	0.0690	0.0503	-27.10%	0.0111	0.0022	-80.18%
	FEARec	0.1827	0.0628	-65.63%	0.0357	0.0093	-73.95%

Table A5: Performance of sequential recommenders using temporal LOO with subsampled training data (indicated as T-LOO′) and split-by-timepoint LOO (indicated as ST-LOO), and the relative differences in Recall@10. Bold and underline indicate the best and second best performing models for each combination of data splitting strategy, evaluation metric and data set. "-" indicates model could not be run due to data set size.

	Model	Popularity-sampled			Unsampled		
		T-LOO′ (Recall@10)	ST-LOO (Recall@10)	Perf. diff (%)	T-LOO′ (Recall@10)	ST-LOO (Recall@10)	Perf. diff (%)
ML-1m	FPMC	0.5369	0.2163	-59.71%	0.1927	0.0328	-82.98%
	GRU4Rec	0.6639	0.2525	-61.97%	0.2747	0.0320	-88.35%
	Caser	0.6147	0.2231	-63.71%	0.1993	0.0177	-91.12%
	SASRec	0.6524	0.2424	-62.84%	0.2644	0.0354	-86.61%
	BERT4Rec	0.6598	0.2096	-68.23%	0.2583	0.0227	-91.21%
	S^3-Rec	**0.6811**	**0.2780**	-59.18%	**0.2880**	**0.0469**	-83.72%
	LightSANs	0.6563	0.2197	-66.52%	0.2392	0.0311	-87.00%
	SINE	0.4429	0.1667	-62.36%	0.0756	0.0160	-78.84%
	FEARec	0.6602	0.2071	-68.63%	0.2436	0.0278	-88.59%
Yelp	FPMC	0.5645	0.4400	-22.05%	0.0183	0.0150	-18.03%
	GRU4Rec	0.6496	0.5055	-22.18%	0.0324	0.0201	-37.96%
	Caser	0.6183	0.4628	-25.15%	0.0252	0.0172	-31.75%
	SASRec	0.6826	0.5063	-25.83%	**0.0394**	**0.0293**	-25.63%
	BERT4Rec	0.6312	0.4848	-23.19%	0.0338	0.0192	-43.20%
	S^3-Rec	-	-	-	-	-	-
	LightSANs	**0.6921**	0.5228	-24.46%	0.0394	0.0290	-26.40%
	SINE	0.6836	0.5318	-22.21%	0.0355	0.0283	-20.28%
	FEARec	0.6291	**0.5388**	-14.35%	0.0260	0.0270	3.85%
Steam	FPMC	0.1550	0.1466	-5.42%	0.0936	**0.1007**	7.59%
	GRU4Rec	0.1994	**0.1623**	-18.61%	0.1173	0.0911	-22.34%
	Caser	0.1800	0.1526	-15.22%	0.1197	0.0983	-17.88%
	SASRec	0.1990	0.1603	-19.45%	0.1020	0.0954	-6.47%
	BERT4Rec	0.1805	0.1522	-15.68%	0.1134	0.0951	-16.14%
	S^3-Rec	0.1926	0.1555	-19.26%	0.1004	0.0915	-8.86%
	LightSANs	0.2105	0.1572	-25.32%	**0.1264**	0.0955	-24.45%
	SINE	**0.2262**	0.1496	-33.86%	0.1085	0.0997	-8.11%
	FEARec	0.2037	0.1601	-21.40%	0.1237	0.0983	-20.53%
Beauty	FPMC	0.2824	0.1141	-59.60%	0.0717	0.0139	-80.61%
	GRU4Rec	0.2074	0.1780	-14.18%	0.0269	0.0176	-34.57%
	Caser	0.1326	0.1462	10.26%	0.0203	0.0132	-34.98%
	SASRec	0.3857	0.1903	-50.66%	0.0871	**0.0238**	-72.68%
	BERT4Rec	0.3176	0.1485	-53.24%	0.0378	0.0108	-71.43%
	S^3-Rec	0.3716	0.1793	-51.75%	0.0901	0.0177	-80.36%
	LightSANs	**0.4026**	**0.1941**	-51.79%	**0.0942**	0.0223	-76.33%
	SINE	0.1384	0.1089	-21.32%	0.0240	0.0043	-82.08%
	FEARec	0.3260	0.1309	-59.85%	0.0684	0.0183	-73.25%

Table A6: Performance of general recommenders compared to sequential recommenders with temporal LOO. Bold indicates best-performing model overall for a given evaluation metric. Sequential recommenders consistently outperform general recommenders with temporal LOO (under the influence of data leakage, see Table 5).

Model	ML-1m		Yelp		Steam		Beauty	
	Pop. sampled (nDCG@10)	Unsampled (nDCG@10)	Pop. sampled (nDCG@10)	Unsampled (nDCG@10)	Pop. sampled (nDCG@10)	Unsampled (nDCG@10)	Pop. sampled (nDCG@10)	Unsampled (nDCG@10)
Pop	0.0670	0.0187	0.0434	0.0039	0.0761	0.0245	0.0268	0.0056
ItemKNN	0.1851	0.0458	0.2829	0.0180	0.0971	0.0676	0.1890	0.0239
BPR	0.2014	0.0384	0.3731	0.0200	0.0885	0.0526	0.1599	0.0227
SLIMElastic	0.1988	0.0529	-	-	0.0847	0.0549	0.0251	0.0002
NeuMF	0.2216	0.0341	0.3454	0.0136	0.0820	0.0313	0.1443	0.0136
NGCF	0.1962	0.0390	0.3429	0.0101	0.0808	0.0546	0.1814	0.0083
LightGCN	0.2055	0.0394	0.3573	0.0255	0.0825	0.0630	0.1978	0.0238
NCL	0.1310	0.0327	0.3997	0.0311	0.0776	0.0569	0.1489	0.0204
FPMC	0.3429	0.1065	0.3760	0.0194	0.0848	0.0547	0.1967	0.0356
GRU4Rec	0.4748	0.1624	0.4278	0.0232	0.0988	0.0622	0.1998	0.0313
Caser	0.3727	0.1023	0.3962	0.0168	0.0923	0.0640	0.1691	0.0184
SASRec	**0.4921**	**0.1814**	0.4515	**0.0374**	0.1017	0.0669	0.2406	0.0415
BERT4Rec	0.4654	0.1613	0.4081	0.0207	0.0916	0.0646	0.1642	0.0181
S^3-Rec	0.4875	0.1807	-	-	0.1036	0.0645	0.2295	0.0431
LightSANs	0.4592	0.1457	**0.4627**	0.0355	**0.1232**	**0.0745**	**0.2577**	**0.0452**
SINE	0.2656	0.0452	0.4313	0.0295	0.0959	0.0608	0.2286	0.0312
FEARec	0.4534	0.1339	0.4528	0.0349	0.1028	0.0672	0.2571	0.0422

Table A7: Performance of general recommenders compared to sequential recommenders with split-by-timepoint LOO in Recall@10. Bold indicates best-performing model overall for a given evaluation metric and cells highlighted in grey represent general recommenders that outperform best-performing sequential recommenders. "-" indicates that a model could not be run due to the size of the data set.

Model	ML-1m		Yelp		Steam		Beauty	
	Pop. sampled (Recall@10)	Unsampled (Recall@10)	Pop. sampled (Recall@10)	Unsampled (Recall@10)	Pop. sampled (Recall@10)	Unsampled (Recall@10)	Pop. sampled (Recall@10)	Unsampled (Recall@10)
Pop	0.1983	0.0504	0.0524	0.0032	0.1277	0.0869	0.0332	0.0113
ItemKNN	0.2059	0.0479	0.3002	0.0168	0.1355	0.0792	0.1661	0.0102
BPR	0.2429	0.0555	0.4266	0.0142	0.1392	0.0982	0.1142	0.0146
SLIMElastic	0.2202	**0.0597**	-	-	**0.1835**	0.0439	0.1146	0.0018
NeuMF	0.2437	0.0471	0.4097	0.0153	0.1354	0.0994	0.1162	0.0110
NGCF	0.2580	0.0244	0.4590	0.0037	0.1279	0.0670	0.1745	0.0079
LightGCN	0.2664	0.0563	0.4333	0.0231	0.1319	0.0985	0.1800	**0.0242**
NCL	0.2387	0.0437	0.4423	0.0214	0.1443	0.0983	0.1348	0.0217
GRU4Rec	0.2525	0.0320	0.5055	0.0201	0.1623	0.0911	0.1780	0.0176
SASRec	0.2424	0.0354	0.5063	**0.0293**	0.1603	0.0954	0.1903	0.0238
S^3-Rec	**0.2780**	0.0469	-	-	0.1555	0.0915	0.1793	0.0177
LightSANs	0.2197	0.0311	0.5228	0.0290	0.1572	0.0955	**0.1941**	0.0223
SINE	0.1667	0.0160	0.5318	0.0283	0.1496	**0.0997**	0.1089	0.0043
FEARec	0.2071	0.0278	**0.5388**	0.0270	0.1601	0.0983	0.1309	0.0183

A.2 Figures

We present one additional figure showing how the model rankings for sequential recommenders changes between temporal LOO and split-by-timepoint LOO for each data set using recall@10.

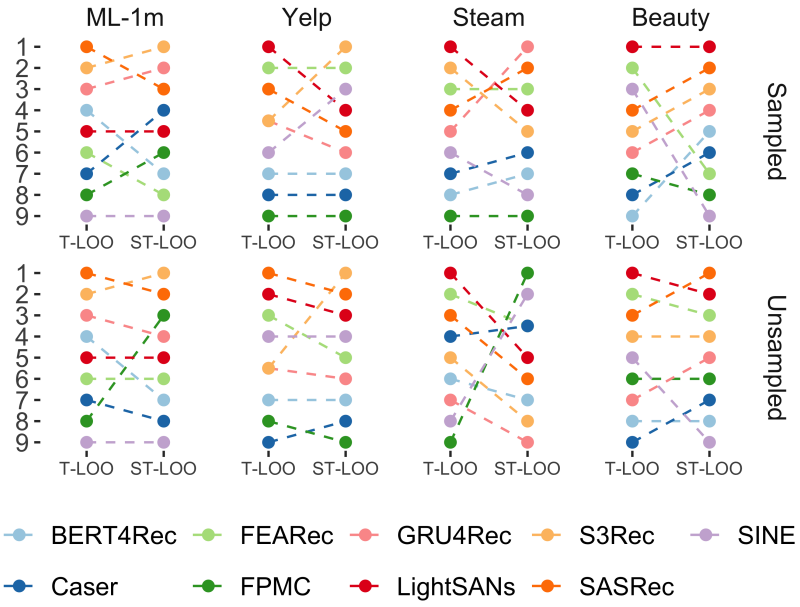

Figure A1: Inconsistent rankings of 9 sequential recommenders obtained with temporal LOO (T-LOO) and split-by-timepoint LOO (ST-LOO) across four datasets. Rankings are obtained based on both popularity-sampled (top) and unsampled (bottom) Recall@10.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009