Accuracy, performance, average distance of different methods measured using each of the 12 pairs of corpus dataset and query dataset are summarized in the following table.

- The accuracy consists of both *precision* and *recall*.
- Duration is the execution time.
- The *average distan*ce column is the average of each query's average Jaccard distance to its neighbours (neighbours are calculated by each method). As discussed in week10 lecture, Jaccard distance = 1 – Jaccard similarity. The larger the Jaccard distance is between two data points, the less similar these two data points are.
- The *difference of Average Distances* column calculates the difference between the exact and approximate solutions in terms of *Average Distance.*
- *And* refers to AndConstruction which uses intersections to combine results of two base constructions, one with the same seed (42) as the *base* construction tested.
- *Or* refers to OrConstruction which uses unions to combine results of two base constructions, one with the same seed (42) as the *base* construction tested.

	Precision	**Recall**	**Duration (second)**	**Average Distance**	**Diff of Avg Dist**
Corpus-1, query-1-2					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	2.080938809	0.0735210478724629	0
Base	1	0.881027342863626	0.107248236	0	-0.0735210478724629
Balanced	1	0.881027342863626	0.27252697	0	-0.0735210478724629
Broadcast	1	0.881027342863626	0.038395416	0	-0.0735210478724629
And	1	0.881027342863626	0.391303503	0	-0.0735210478724629
Or	0.736921618405704	0.916666666666667	0.341727171	0.247376213873628	0.173855166001165
Corpus-1, query-1-2-skew					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	2.382774296	0.0558035136160137	0
Base	1	0.960789703853918	0.104442822	0	-0.0558035136160137
Balanced	1	0.960789703853918	0.27232758	0	-0.0558035136160137
Broadcast	1	0.960789703853918	0.03637888	0	-0.0558035136160137
And	1	0.960789703853918	0.404554581	0	-0.0558035136160137
Or	0.559808741671483	0.971059930302449	0.376084809	0.286525319694991	0.230721806078977
Corpus-1, query-1-10					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	5.163136565	0.0752777427226649	0

Base	1	0.873356168920685	0.1124863	0	-0.0752777427226649
Balanced	1	0.873356168920685	0.30414845	0	-0.0752777427226649
Broadcast	1	0.873356168920685	0.049154642	0	-0.0752777427226649
And	1	0.873356168920685	0.438878298	0	-0.0752777427226649
Or	0.776954778569527	0.904399050770019	0.476832274	0.235304716340835	0.16002697361817
Corpus-1, query-1-10-skew					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	5.459142009	0.0605853169097216	0
Base	1	0.96120590050717	0.114394043	0	-0.0605853169097216
Balanced	1	0.96120590050717	0.339216363	0	-0.0605853169097216
Broadcast	1	0.96120590050717	0.072037462	0	-0.0605853169097216
And	1	0.96120590050717	0.502054865	0	-0.0605853169097216
Or	0.512468811477304	0.965633856499405	0.465042309	0.319335502580927	0.258750185671205
Corpus-10, query-10-2					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	71.94665724	0.214048948796594	0
Base	0.363375936535918	0.794013725260827	0.254069155	0.652906703881789	0.438857755085195
Balanced	0.363375936535918	0.794013725260827	0.6595105	0.652906703881789	0.438857755085195
Broadcast	0.363375936535918	0.794013725260827	0.137641335	0.652906703881789	0.438857755085195
And	0.828864893981686	0.731179830325612	1.228284951	0.210799759233506	-0.003249189563088
Or	0.300788669524545	0.847936309288371	1.24721138	0.706474862088189	0.492425913291595
Corpus-10, query-10-2-skew					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	61.803146445	0.173507868089636	0
Base	0.156469886702454	0.855934487631531	0.279870613	0.702456382888906	0.52894851479927
Balanced	0.156469886702454	0.855934487631531	0.582237099	0.702456382888906	0.52894851479927
Broadcast	0.156469886702454	0.855934487631531	0.152317152	0.702456382888906	0.52894851479927
And	0.8909609823854	0.827493376651555	1.112597978	0.207034903288245	0.033527035198609
Or	0.107219307012833	0.883429346294206	1.043954636	0.757369319111677	0.583861451022041
Corpus-10, query-10-10					

Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	331.503049177	0.208648356877452	0
Base	0.380288587098451	0.798168508372717	0.372364557	0.640275133807012	0.43162677692956
Balanced	0.380288587098451	0.798168508372717	1.200407839	0.640275133807012	0.43162677692956
Broadcast	0.380288587098451	0.798168508372717	0.310770669	0.640275133807012	0.43162677692956
And	0.818073977000917	0.737328323820174	2.156506091	0.217393027513977	0.008744670636525
Or	0.315583171396914	0.855026800075799	2.153676124	0.695057412083241	0.486409055205789
Corpus-10, query-10-10-skew					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	292.544396788	0.173107810405354	0
Base	0.137983335213956	0.876919464722672	0.332527107	0.707231525805447	0.534123715400093
Balanced	0.137983335213956	0.876919464722672	1.256443577	0.707231525805447	0.534123715400093
Broadcast	0.137983335213956	0.876919464722672	0.254587971	0.707231525805447	0.534123715400093
And	0.918541430686336	0.84946294681797	1.855668267	0.196724548233271	0.023616737827917
Or	0.0908345839095808	0.889055779903811	1.625957658	0.760544146732042	0.587436336326688
Corpus-20, query-20-2 (in results1.txt)					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	250.678566116	0.241398688930943	0
Base	0.339919264409063	0.766239437721151	0.361636521	0.67678597921143	0.435387290280487
Balanced	0.339919264409063	0.766239437721151	1.053975454	0.67678597921143	0.435387290280487
Broadcast	0.339919264409063	0.766239437721151	0.273235506	0.67678597921143	0.435387290280487
And	0.785422579123582	0.692039646353395	2.211551916	0.252958056832447	0.011559367901504
Or	0.290889311512216	0.834145118282226	2.175273245	0.718304312235118	0.476905623304175
Corpus-20, query-20-2-skew (in debug.txt)					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	220.760331421	0.204049138706951	0
Base	0.103742571994513	0.863534857501443	0.353750503	0.746423869579874	0.542374730872923
Balanced	0.103742571994513	0.863534857501443	0.979542168	0.746423869579874	0.542374730872923
Broadcast	0.103742571994513	0.863534857501443	0.201778429	0.746423869579874	0.542374730872923
And	0.864760741376244	0.829243229612255	1.853333238	0.238557510360552	0.034508371653601
Or	0.0797141316076181	0.909613140886388	1.687411873	0.783581921140153	0.579532782433202
Corpus-20, query-20-10 (in results20.txt)					

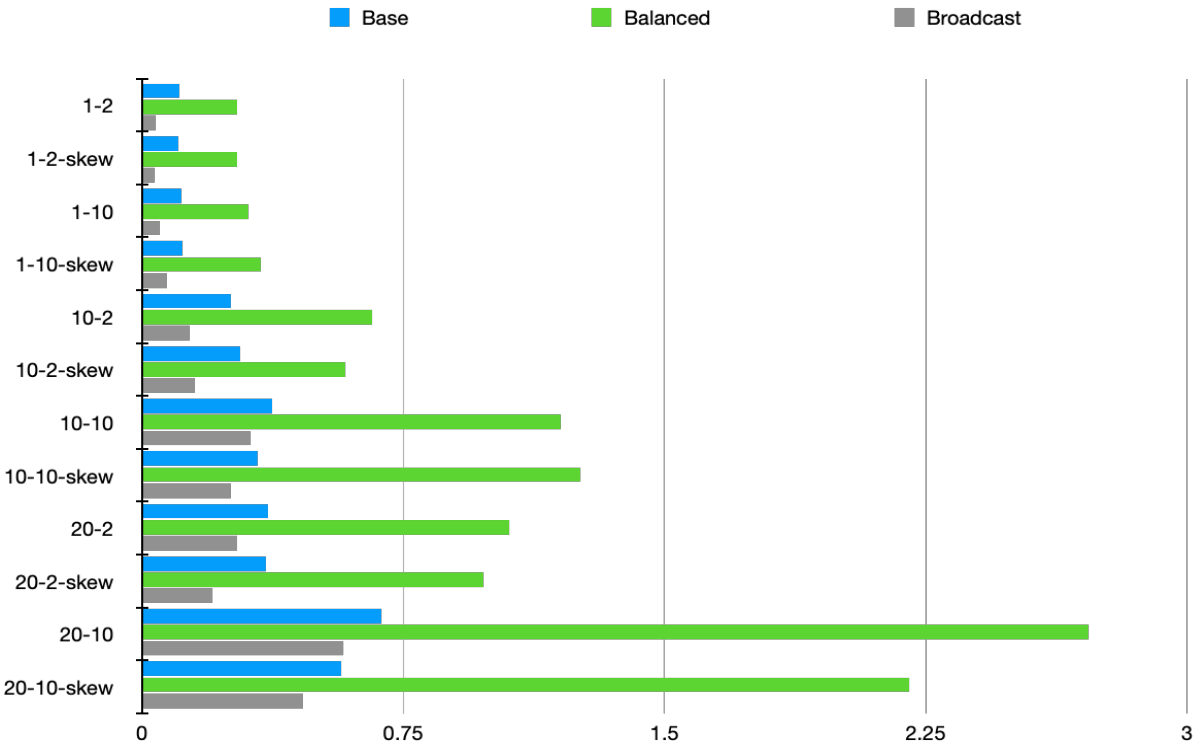
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	1175.619747309	0.243886186747839	0
Base	0.347200804006859	0.768676832063605	0.686600786	0.678924813866851	0.435038627119012
Balanced	0.347200804006859	0.768676832063605	2.717598215	0.678924813866851	0.435038627119012
Broadcast	0.347200804006859	0.768676832063605	0.57817444	0.678924813866851	0.435038627119012
And	0.785415149681149	0.689574120554173	4.067361963	0.256125376523021	0.012239189775182
Or	0.294579413524628	0.834885141402211	3.642179849	0.721831471749735	0.477945285001896
Corpus-20, query-20-10-skew (in results20.txt)					
Methods	Precision	Recall	Duration(sec)	Avg Dist	Diff of Avg Distance
ExactNN	1	1	947.173539916	0.202952042419327	0
Base	0.0889748128823646	0.860725984868002	0.572328804	0.745847091180629	0.542895048761302
Balanced	0.0889748128823646	0.860725984868002	2.203023374	0.745847091180629	0.542895048761302
Broadcast	0.0889748128823646	0.860725984868002	0.46111499	0.745847091180629	0.542895048761302
And	0.866833703069864	0.822821664143263	2.905662322	0.240670147411168	0.037718104991841
Or	0.067383092035333	0.90722245604674	2.83971073	0.783726539620908	0.580774497201581

Corpus-1.csv: ~2k data points
Corpus-10.csv: ~20k data points
Corpus-20.csv: ~40k data points

Execution time plot

Performance of different methods measured using each of the 12 pairs of corpus dataset and query dataset are summarized in the following chart. X-Y(-skew) refers to using query dataset query-X-Y(-skew).csv on corpus dataset corpus-X.csv.

- Different colours represent different methods.
- The vertical axis is the different pairs of datasets used.
- The horizontal axis is the duration in seconds.



(A execution time plot that also includes And/Or constructions is presented on the next page.)

When is each method preferable?

ExactNN:
Preferable only when the dataset is really small (e.g. corpus-1.csv, ~2k data points), and you really care about the accuracy but not so much about the performance.
By definition ExactNN can find exactly all the qualifying neighbours for each query but its performance doesn't scale well – even for corpus10.csv (20k data points) it will take more than 1min.

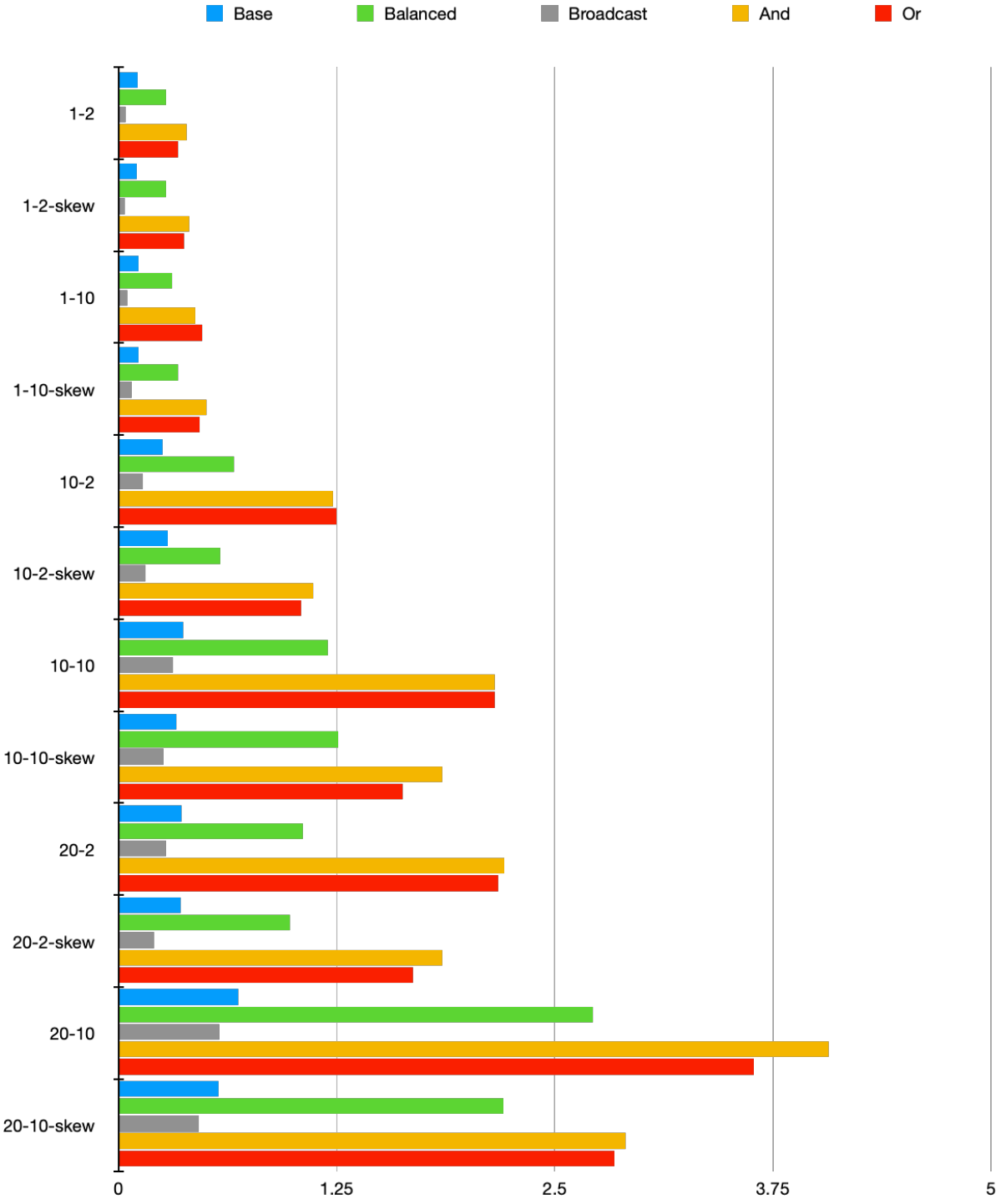
Base:
Preferable when corpus dataset is so large that it cannot fit in each node's memory (so that Broadcast is not suitable), and also not skewed (otherwise balanced might be preferable).

Balanced:
only preferable when the query dataset is really skewed and extremely large. As we can see even using query-20-10-skew.csv on corpus-20.csv was not enough to demonstrate the theoretical benefit load balancing should have.

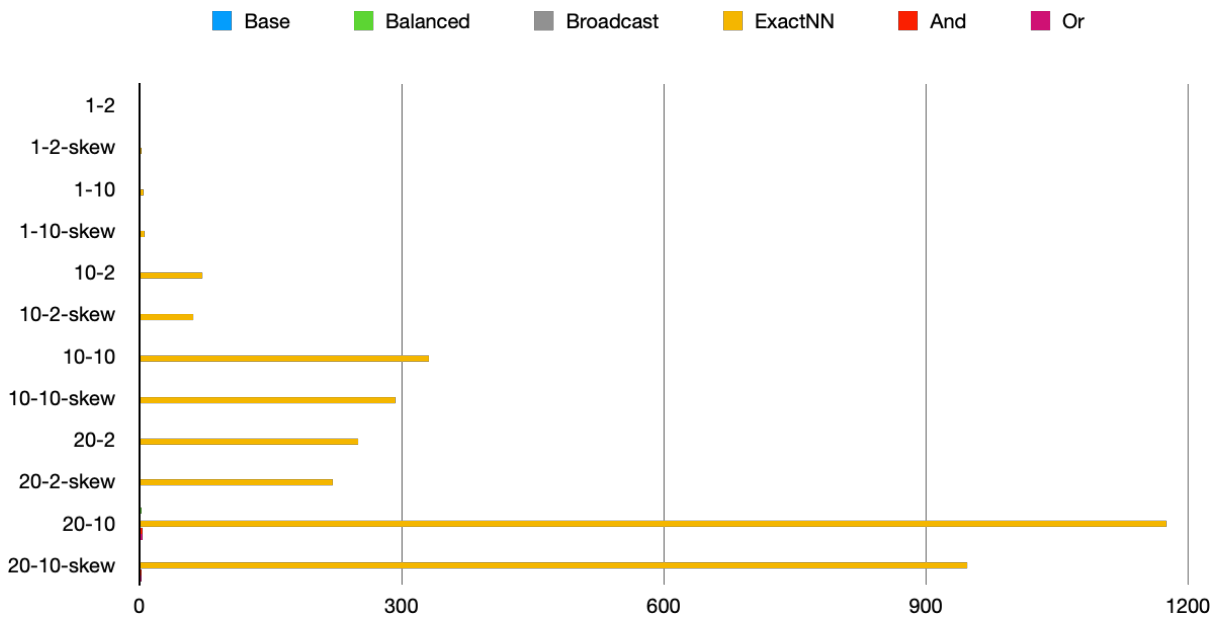
Broadcast:
Preferable when corpus dataset is not that large (e.g. <= 40k data points, as in corpus-20.csv) and can fit in the memory of each node. For our datasets broadcast is always the fastest method.

And/Or:
For all LSH methods, in the case where the precision or recall is not ideal, we can use And Construction to improve precision and use Or-construction to improve recall (And by improving precision/recall we can improve the average distance). Both And/Or constructions have longer execution time than single construction though.

Plot that also includes And/Or constructions, which have longer execution time than single atomic LSH methods.



Additional plot to show that ExactNN really doesn't scale well. LSH methods' time costs are so little compared to ExactNN that they are almost invisible.


Base
Balanced
Broadcast
ExactNN
And
Or
1-2
1-2-skew
1-10
1-10-skew
10-2
10-2-skew
10-10
10-10-skew
20-2
20-2-skew
20-10
20-10-skew
0
300
600
900
1200