A DATASET SOURCE AND PREPROCESSING STEPS

Table A1: Source and processing steps taken for the 51 datasets used in Sections 3.2 and 4

Dataset(s)	Source	Preprocessing Steps [1]
24 Amazon Datasets, e.g. Amazon (Books), Amazon (Electronics), etc.	jmcauley.ucsd.edu/data/amazon	We directly use the 5-core [2] version whereby all users and items have ≥ 5 interactions each.
Amazon Fine Food	kaggle.com/snap/amazon-fine-food-reviews	We preprocess these datasets to get its 5-core.
BookCrossing	www2.informatik.uni-freiburg.de/~cziegler/BX/	
CiteULike-a	github.com/js05212/citeulike-a	
CiteULike-t	github.com/js05212/citeulike-t	
Epinions	trustlet.org/downloaded_epinions.html	
FilmTrust	guoguibing.github.io/librec/datasets.html	
Flixster	sites.google.com/view/mohsenjamali/flixter-data-set	
GoodReads (Comics)	sites.google.com/eng.ucsd.edu/ucsdbookgraph/home	
HetRec2011-LastFM-2K	grouplens.org/datasets/hetrec-2011/	
Last.fm 1K	ocelma.net/MusicRecommendationDataset/lastfm-1K.html	
Last.fm 360K	ocelma.net/MusicRecommendationDataset/lastfm-360K.html	
Meetup (NYC)	personal.ntu.edu.sg/gaocong/datacode.htm	
Netflix	kaggle.com/netflix-inc/netflix-prize-data	
Yahoo! R1	webscope.sandbox.yahoo.com/catalog.php?datatype=r	
Yahoo! R4	webscope.sandbox.yahoo.com/catalog.php?datatype=r	
Yelp [3]	yelp.com/dataset	
Gowalla	github.com/dawenl/expo-mf	We use the dataset from [28] where venues have ≥ 20 check-ins, and we remove users with < 5 check-ins.
Million Song Dataset	millionsongdataset.com/	We follow [29] and retain users with ≥ 20 interactions and songs with ≥ 200 interactions.
HetRec2011-ML-2K	grouplens.org/datasets/hetrec-2011/	The users in these datasets have ≥ 20 interactions, and we remove items with < 5 interactions.
ML-100K	grouplens.org/datasets/movielens/100k/	
ML-10M	grouplens.org/datasets/movielens/10m/	
ML-1M	grouplens.org/datasets/movielens/1m/	
ML-20M	grouplens.org/datasets/movielens/20m/	
Pinterest	github.com/hexiangnan/neural_collaborative_filtering	
Goodbooks-10k	github.com/zygmuntz/goodbooks-10k	We directly use these datasets as they are.
Million Song Dataset (Taste Profile Subset)	millionsongdataset.com/tasteprofile/	
Yahoo! R2 [4]	webscope.sandbox.yahoo.com/catalog.php?datatype=r	

[1] As the final preprocessing step, for datasets with explicit feedback (e.g. ratings), we convert all the observed entries into positive interactions.

[2] The K-core (e.g. used in [50, 55]) is derived by using a *recursive filter* such that all remaining users and items have $\geq K$ interactions each.

[3] We use the latest version of the Yelp dataset (Retrieved on 20th January 2021).

[4] The original dataset consists of more than $1.8M$ users and $717M$ ratings, and it has been partitioned into 10 subsets. In this paper, we only utilise the first subset of $200K$ users.