**Characterization of Fair Experiments for Recommender System Evaluation – A Formal Analysis**

**Evaluation fairness condition**

**Elements of an experiment**

- $\mathcal{U} \times \mathcal{J}$ All user-item pairs
- $\mathcal{T}$ Target pairs
- $\mathcal{J}_{train}$ Training sample
- $\mathcal{J}_{test}$ Test sample
- $\mathcal{R}$ Relevant pairs
- $\mathcal{J}_{train} \cap \mathcal{J}_{test} = \emptyset$

**Metric definition**

- $P = \frac{|\mathcal{R} \cap \mathcal{J}|}{|\mathcal{R}|} = p(\mathcal{R}|\mathcal{J})$
- $Recall = \frac{|\mathcal{R} \cap \mathcal{J}|}{|\mathcal{R}|} = p(\mathcal{R}|\mathcal{J})$

**Fair estimates**

Preservation of system comparisons: $P(R_1) \leq P(R_2) \Rightarrow \hat{P}(R_1) \leq \hat{P}(R_2)$ – we say $\hat{P} \propto P$

**Analysis of common experimental protocols**

Null hypothesis recreation: taking e.g. MovieLens 1M, keep ratings (judgment set $\mathcal{J}$) but shuffle rating values ($\mathcal{R}$ set) over ratings
- User preferences become random, uniform and independently distributed over items (the sample $\mathcal{J}$ may not)
- Run an experiment over $\mathcal{J}, \mathcal{R}$ for a set of recommendation algorithms
- Some system is better than random recommendation $\Rightarrow$ Then your experiment is unfair (the data sampling/subsampling, the metric, etc.)

1. **Free user feedback**

   - Random rating split
   - Flat test [1]
   - Popularity strata [1]

2. **Randomized (forced) test judgments** [2,3]

   - Simulated random test sample of random preferences

**Conclusions**

- Only randomized test judgments or $\mathcal{J} \leftarrow \mathcal{J}_{test}$ ensure fairness
  - But $\mathcal{J} \leftarrow \mathcal{J}_{test}$ is not as realistic as $\mathcal{J} \leftarrow (\mathcal{U} \times \mathcal{J}) \setminus \mathcal{J}_{train}$ (plus coverage shortfalls)
  - Forced judgments to be handled with some care to be fully fair (see in paper)
- Other protocols are biased to non-random patterns in observations
  - Popularity, inter-user dependences, etc. (avg rating would not seem affected though)

- We also examine experimental protocols analytically
  - Empirical fairness test is consistent with analytical fairness condition
  - Temporal split can be usually expected to be still biased
  - Interleaved AB tests should be fair