Reconstructing Hidden Permutations Using the Average-Precision (AP) Correlation Statistic

Lorenzo De Stefani, Alessandro Epasto, Eli Upfal *(Brown Uuniversity)*, Fabio Vandin *(Univeristy of Padova)*
{lorenzo, epasto, eli} @cs.brown.edu, vadinfa@dei.unipd.it

MOTIVATION

Probabilistic models of rankings studied in:

- social sciences
- statistics
- machine learning
- computer science

Applications include:

- understanding user preferences
- ordering web search results
- aggregating crowd-sourcing data
- optimizing recommendation systems results

Most work in literature focused on the **Mallows model.**

THE MALLOWS MODEL

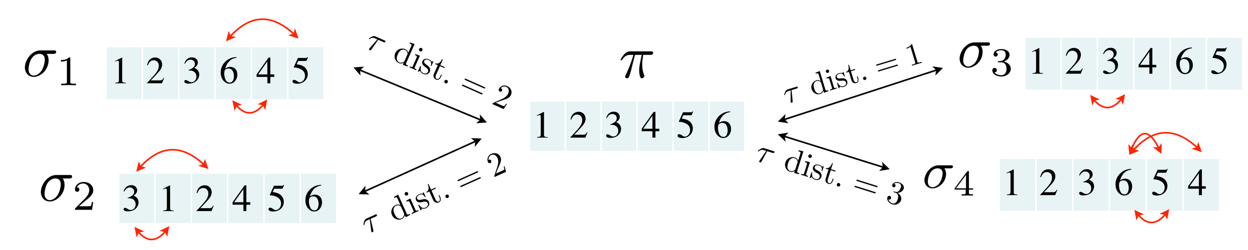
In the Mallows model $\mathcal{M}(\beta,\pi)$, the probability of observing a permutation σ is inversely proportional to its **tau distance** from a central (**"ground truth"**) permutation π

$$\boxed{\Pr_{\mathcal{M}(\beta,\pi)}(\sigma) = Z_\beta^{-1}\exp\left(-\beta d_K(\pi,\sigma)\right)}$$

normalization coefficient

dispersion coefficient

The **tau distance** $d_k(\pi,\sigma)$ between permutations counts the number of items whose order is inverted



All inversions are weighted in the same way!

In many cases, the **order** of items at the **top of the ranking** is **more significant** than the order of the items at the **bottom**!

THE AP - MODEL

We propose the new **AP-model** which uses the **AP-statistics** as measure of distance between permutations

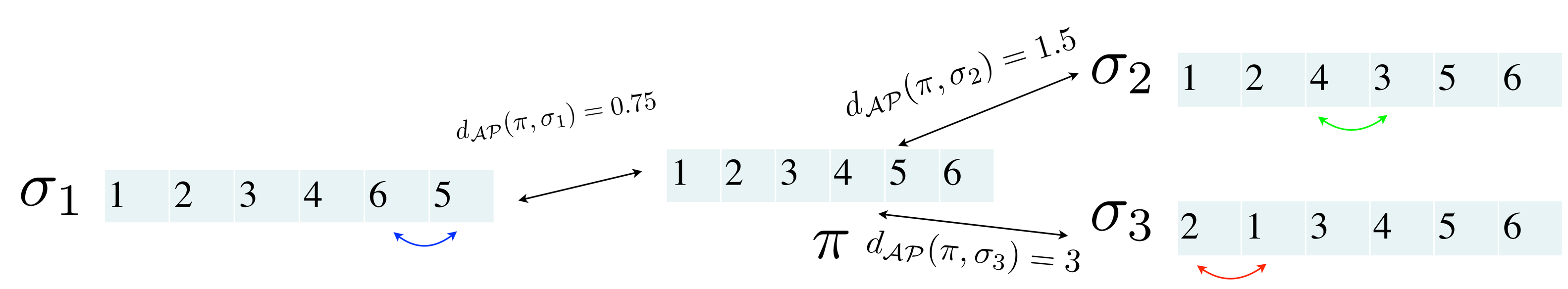
$$\boxed{\Pr_{\mathcal{M}_{AP}(\beta,\pi)}(\sigma) = Z_\beta^{-1}\exp\left(-\beta d_{AP}(\pi,\sigma)\right)}$$

$$d_{\mathcal{AP}}(\pi,\sigma) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_{ij}\frac{n}{2(j-1)}$$

$$d_{\mathcal{K}}(\pi,\sigma) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_{ij}$$

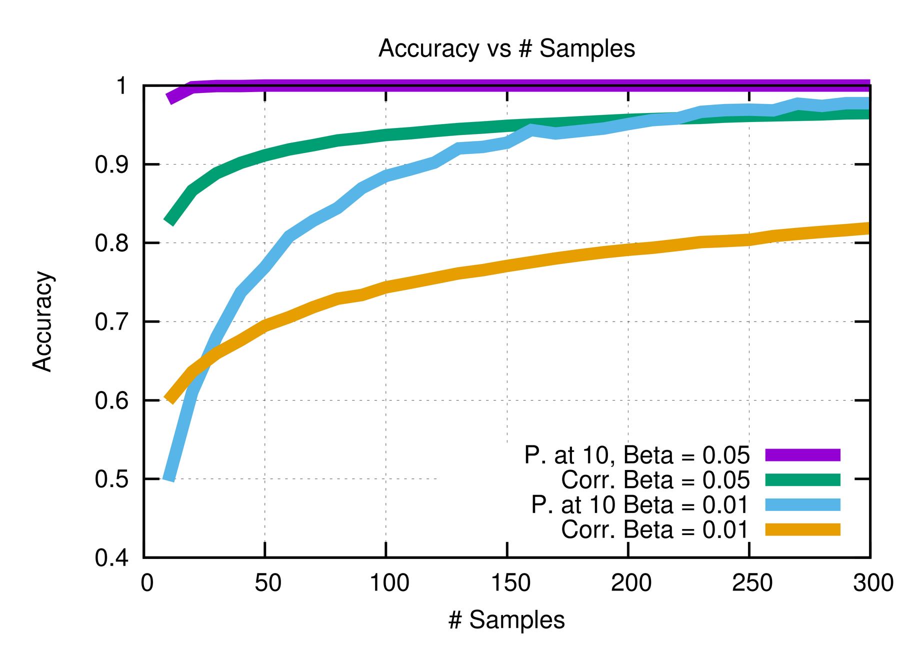
The **AP-statistics** counts inversions **weighting** them according to **the position of the swapped items** in the central permutation π

$E_{ij} = 1$ iff the the *i*-th element of π is ranked after the *j*-th element of π in σ



ALGORITHMS

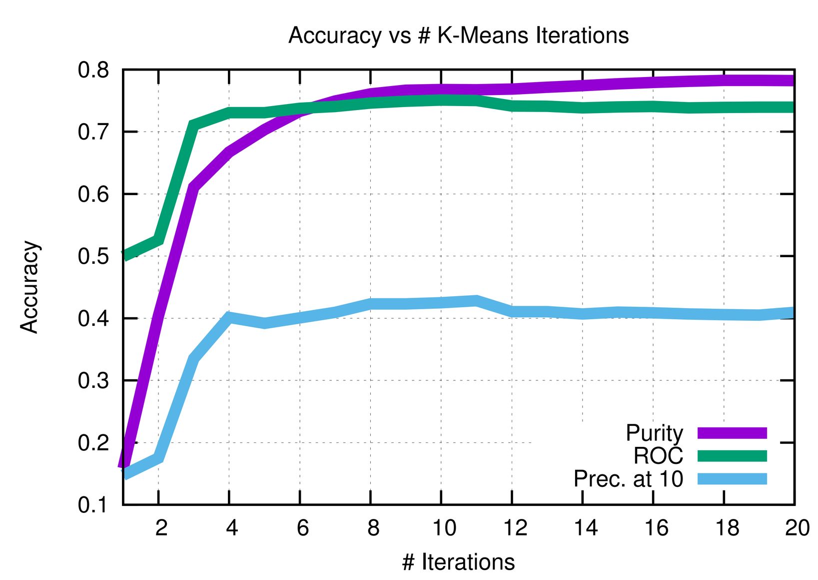
We provide efficient algorithms for reconstructing the central permutation π on *n* items from $\mathcal{O}(\log_2 n)$ observations

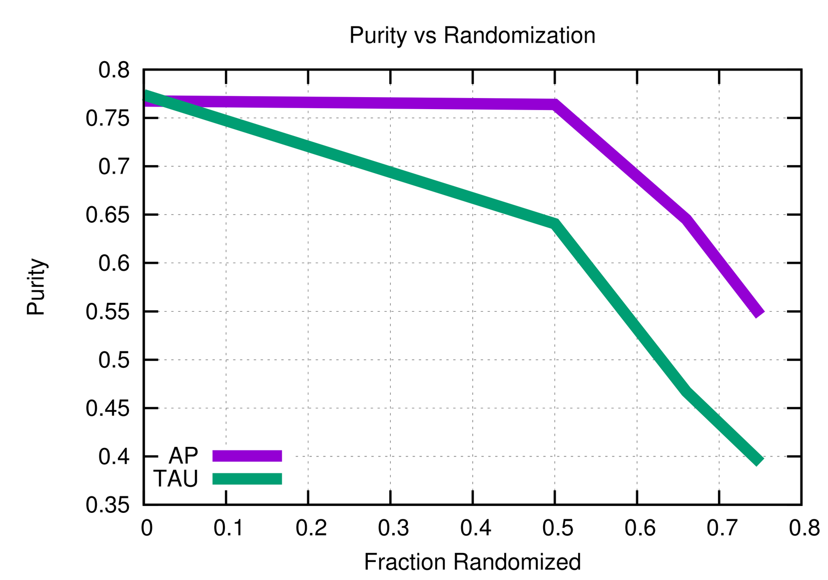


APPLICATION TO CLUSTERING

We use the AP-model estimators for an **unsupervised clustering algorithm** based on k-means.

- We apply the algorithm to cluster web pages
- **Purity** and **ROC** converge to over 80% after a few iterations
- Our approach is resilient to noise in the lower position of the rankings





APPLICATION TO CLASSIFICATION

It has been observed that in the context of high- dimensional gene expressions data ($> 10^4$ genes), **the relative order of the genes** is more important than their absolute magnitude.

Dataset	Prec. Tau	Prec. AP
BC1	0.662	**0.674**
BC2	**0.621**	0.601
CT	0.848	**0.868**
LA1	0.666	**0.685**
LC2	0.986	**0.993**
MB	0.613	**0.648**
OV	**0.836**	0.817
PC1	0.657	**0.667**
PC2	**0.499**	0.493
Average	0.709	**0.716**

- We use the AP-model to **classify gene expressions** into one of two binary classes (e.g., "Normal vs. Tumor")
- AP distance improves over tau distance-based methods for most datasets!