How Precise Does Document Scoring Need To Be?

Ziying Yang, Alistair Moffat, and Andrew Turpin

University of Melbourne

Background – Batch Evaluation Technique

Question

Is IR System A demonstrably better than IR System B?

Background – Batch Evaluation Technique

Question

Is IR System A demonstrably better than IR System B?

For each of system A and B

- For each of a set of topics
  - Compute the similarity score relative to the topic for each document
  - Generate a run in decreasing score order
  - Evaluate the run via relevance judgments and an effectiveness metric
  - Generate a run score for that system and topic

Background – Batch Evaluation Technique

Question

Is IR System A demonstrably better than IR System B?

For each of system A and B

- For each of a set of topics
  - Compute the similarity score relative to the topic for each document
  - Generate a run in decreasing score order
  - Evaluate the run via relevance judgments and an effectiveness metric
  - Generate a run score for that system and topic
- Aggregate run scores over the set of topics into a single system score

We then compare system A and B using their system scores.

How to Deal with Tied Similarity Scores?

Effectiveness metrics use ranks, not similarity scores.

Ties: when more than two items receive the same scores.

How should similarity score ties be handled?

Different orderings of tied documents may lead to different system scores, and might influence the outcome of the experiment.

How to Deal with Tied Similarity Scores?

Example

rank, k	1	2	3	4	5	6	7
document, d_k	D	H	A	C	M	S	W
gain, r_k	0	0	1	1	0	1	1
similarity score	9.8	9.3	9.3	9.3	8.4	8.4	8.2

Divide the ranking of a run into groups in which the documents have the same similarity score

How to Deal with Tied Similarity Scores?

Example

rank, k	1	2	3	4	5	6	7
document, d_k	D	H	A	C	M	S	W
gain, r_k	0	0	1	1	0	1	1
similarity score	9.8	9.3	9.3	9.3	8.4	8.4	8.2

Possible methods for dealing with ties in the run:

- Run Order : ignore the similarity score, use the presented ordering. (RBP0.5 = 0.211)

How to Deal with Tied Similarity Scores?

Example

rank, k	1	2	3	4	5	6	7
document, d_k	D	H	A	C	M	S	W
gain, r_k	0	0	1	1	0	1	1
similarity score	9.8	9.3	9.3	9.3	8.4	8.4	8.2

Possible methods for dealing with ties in the run:

- Run Order (RBP0.5 = 0.211)
- External Tie-Break Rule : re-order tied documents in each group using fixed ordering criterion, such as sorting by document ID (`trec_eval` program), filename, length and so on. (RBP0.5 = 0.227)

How to Deal with Tied Similarity Scores?

Example

rank, k	1	2	3	4	5	6	7
document, d_k	D	H	A	C	M	S	W
gain, r_k	0	0	1	1	0	1	1
similarity score	9.8	9.3	9.3	9.3	8.4	8.4	8.2

Possible methods for dealing with ties in the run:

- Run Order (RBP0.5 $= 0.211$)
- External Tie-Break Rule (RBP0.5 $= 0.227$)
- Limits : compute the best and worst run scores that may arise and present a score range instead of a single score value.
  (RBP0.5 $= [0.211, 0.414]$)

How to Deal with Tied Similarity Scores?

Example

rank, k	1	2	3	4	5	6	7
document, d_k	D	H	A	C	M	S	W
gain, r_k	0	0	1	1	0	1	1
similarity score	9.8	9.3	9.3	9.3	8.4	8.4	8.2

Possible methods for dealing with ties in the run:

- Run Order (RBP0.5 $= 0.211$)
- External Tie-Break Rule (RBP0.5 $= 0.227$)
- Limits (RBP0.5 $= [0.211, 0.414]$)
- Averaging Across Permutations : compute the average run score across all possible permutations of documents in each group. (RBP0.5 $= 0.323$)

Ties in TREC Experimentation

To explore the role of ties in TREC evaluation, we re-sorted the TREC7 submissions, using decreasing similarity score as a primary key, and increasing rank as a secondary key.

Table: The percentage of 103 systems, 50 $\times$ 103 runs and 4,900,042 documents affected by ties occurring in TREC7 Ad–Hoc runs after score-based re-sorting

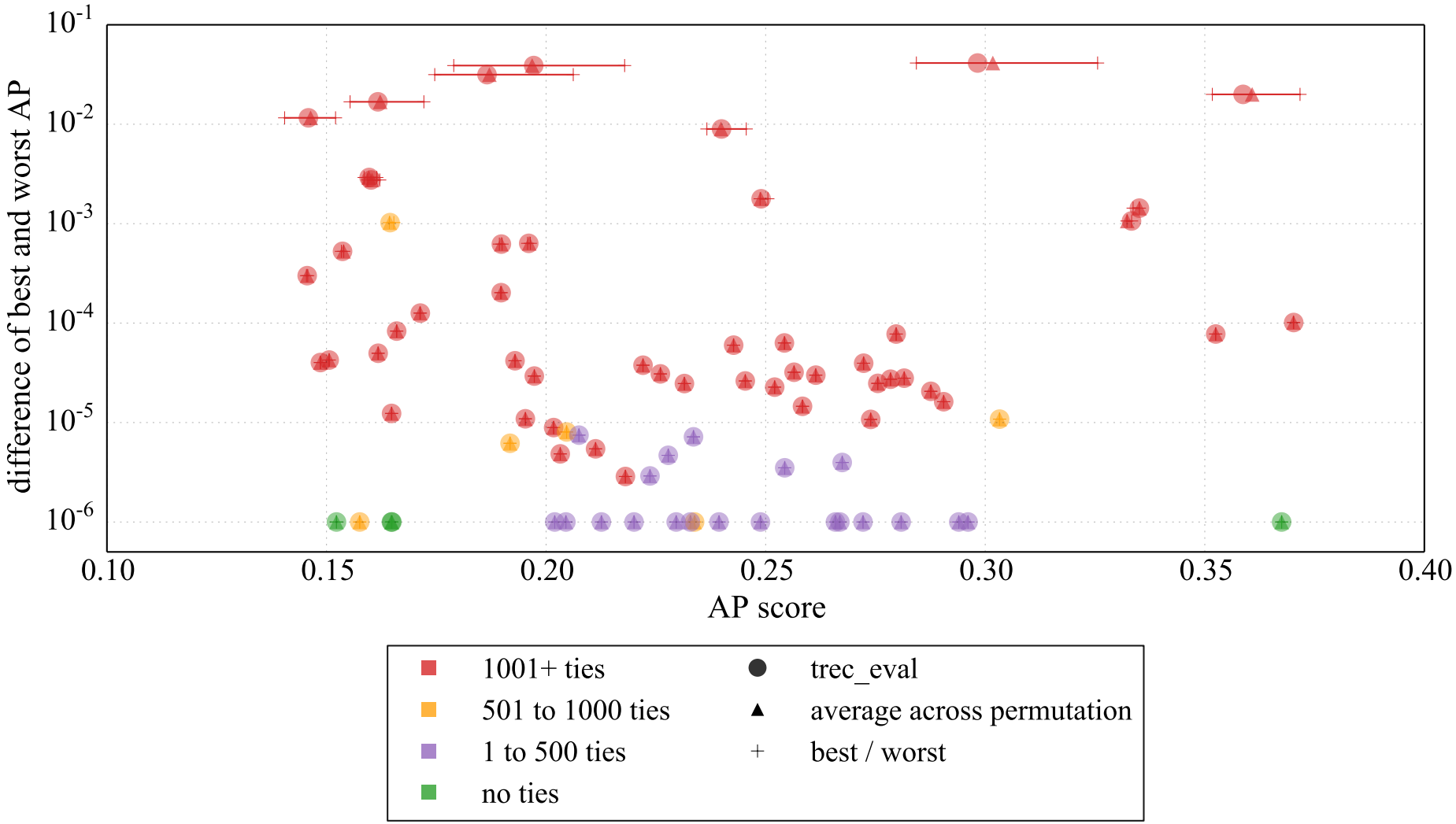
	systems	runs	documents
Tied similarity scores	95.2%	91.0%	14.0%
Rank/score contradictions	6.8%	4.2%	1.4%

Ties in TREC Experimentation

Examine the effect of ties on AP scores for systems

- For each system, calculate mean AP score over 50 topics using `trec_eval`.
- Select the top 80 systems, as ordered by mean AP score.
- For each selected system, compute
  - Run Order: mean (across topics) of the AP score (by `trec_eval`).
  - Limits: mean (across topics) of the best and worst AP score.
  - Permuations: mean (across topics) of the average (across permutations of the tied groups in the run) AP score.

Imprecision in AP Scores Caused by Ties



Similarity Score Rounding

Ties may have been caused by similarity score rounding.

Question

Do documents really need to assign similarity scores with high accuracy? How much similarity score rounding can be tolerated without greatly affecting system comparisons?

The finding offers the potential for approximate scoring regimes that provide faster search with little or no loss of effectiveness.

Deliberate Similarity Score Grouping

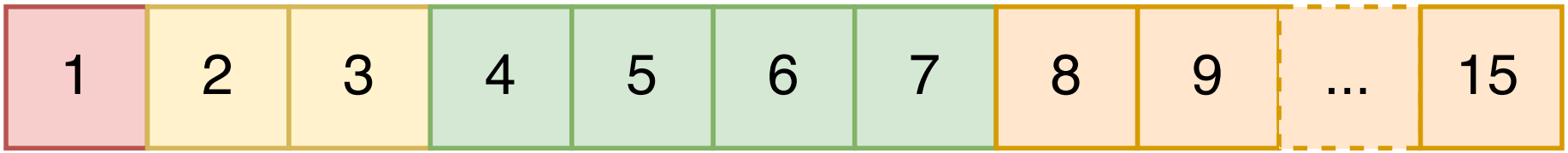
Will the deliberate use of ties affect retrieval quality?

Documents are scored and assigned in bands of the ranking. Bands are defined geometrically based on a parameter ρ ($\rho > 1$). More precisely:

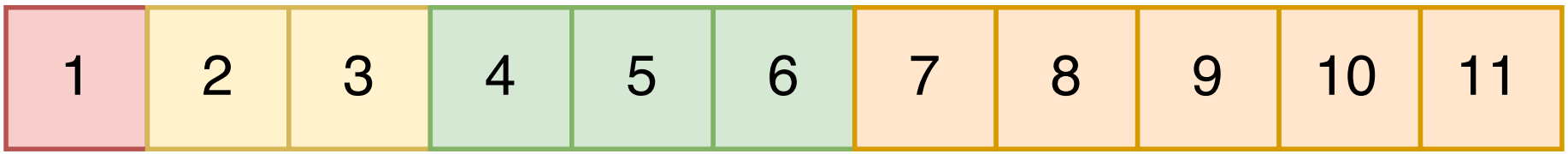
For the gth band:

- the begining rank: $b_g = \lceil \rho \cdot b_{g-1} \rceil$, $b_1 = 1$.
- the ending rank: $e_g = b_{g+1} - 1$

For example:
if $\rho = 2$



if $\rho = 1.62$ (the golden ratio)



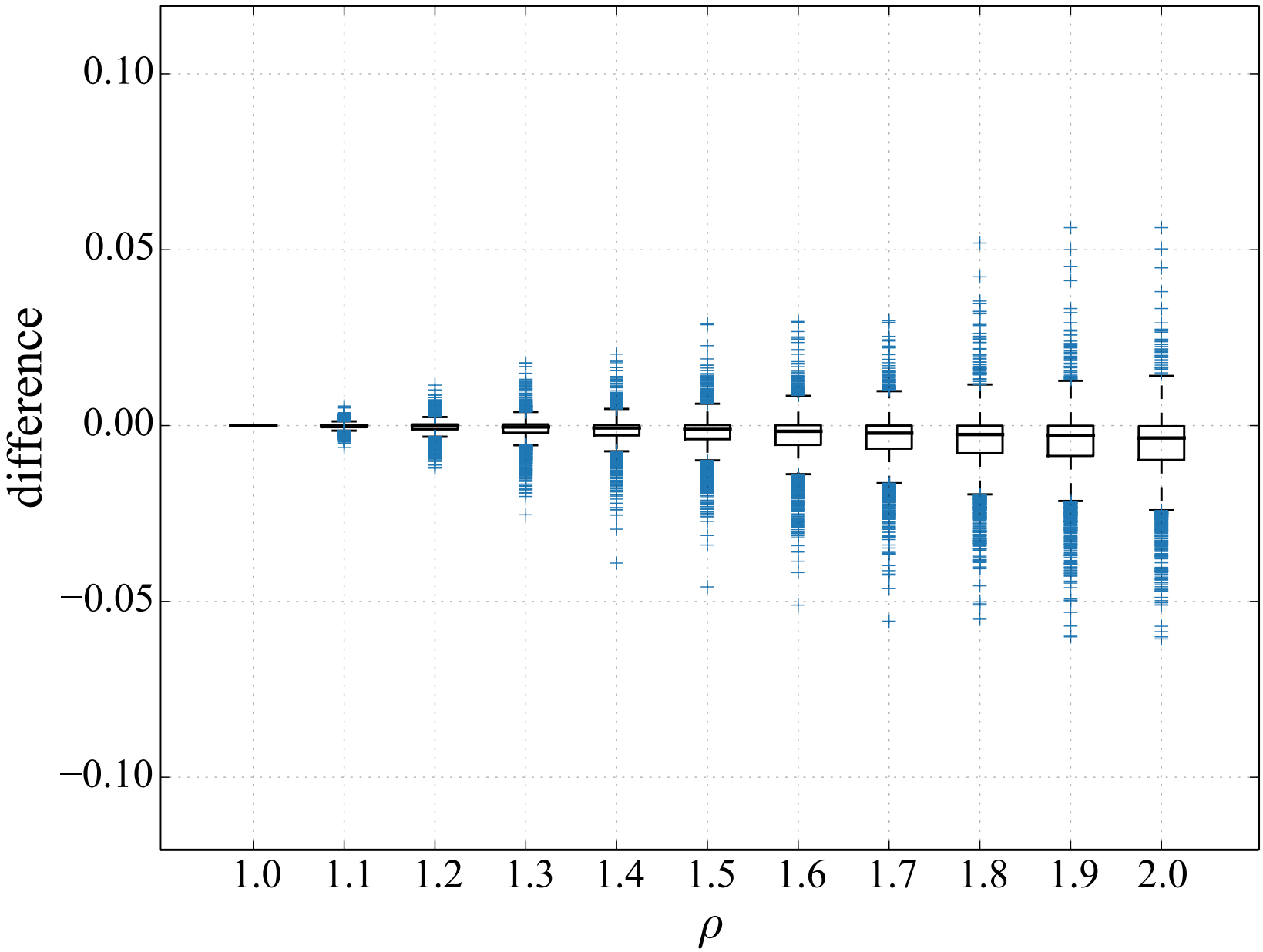
Run Score Differences in Practice

For each of the 80 $\times$ 50 system–topic runs

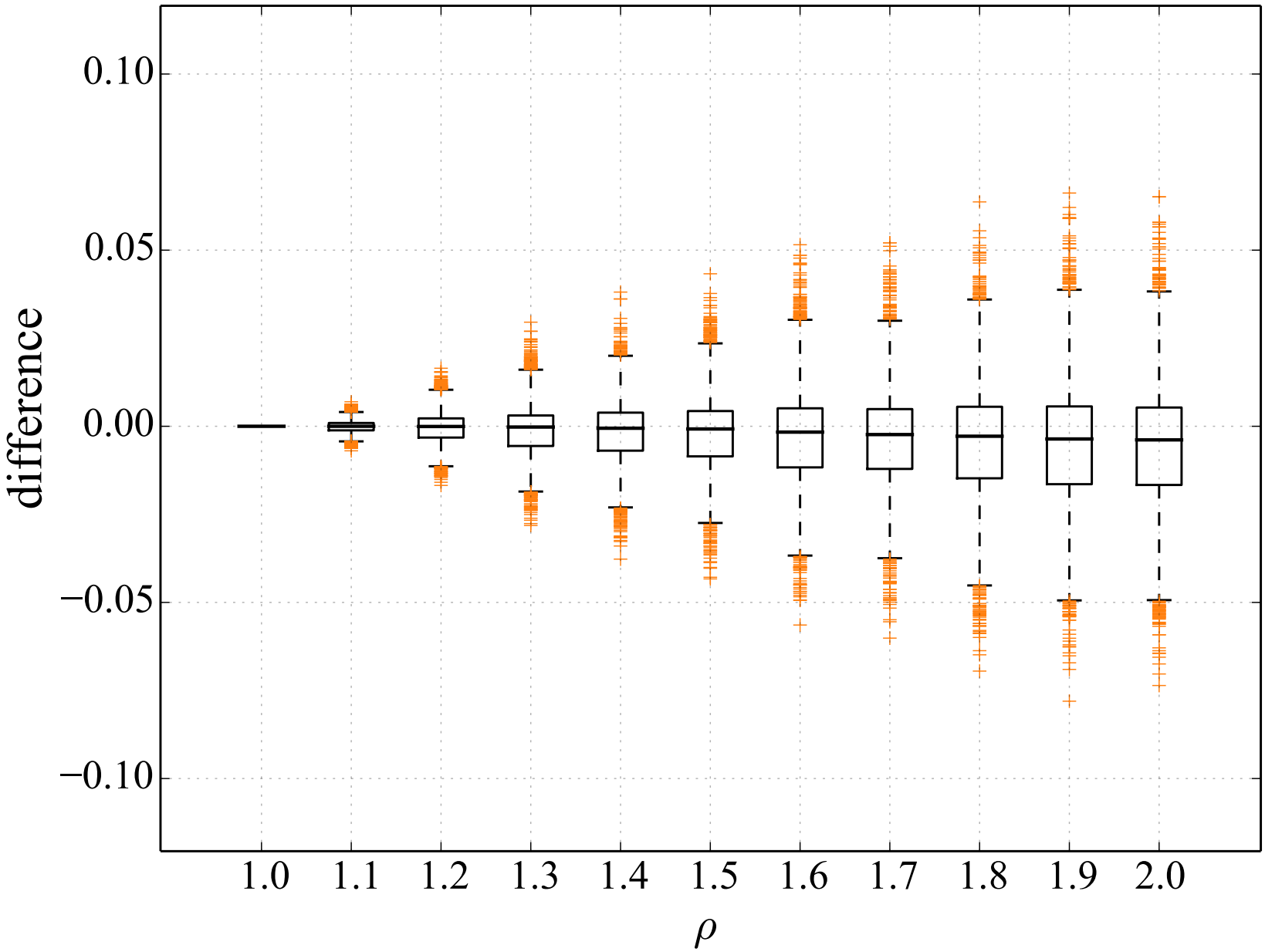
- compute effectiveness scores of metrics RR, RBP0.5, RBP0.85 and AP using the original run
- map original run to a banded ranking list using a ρ
- score the banded run using same metrics

Compute 80 $\times$ 50 run score differences.

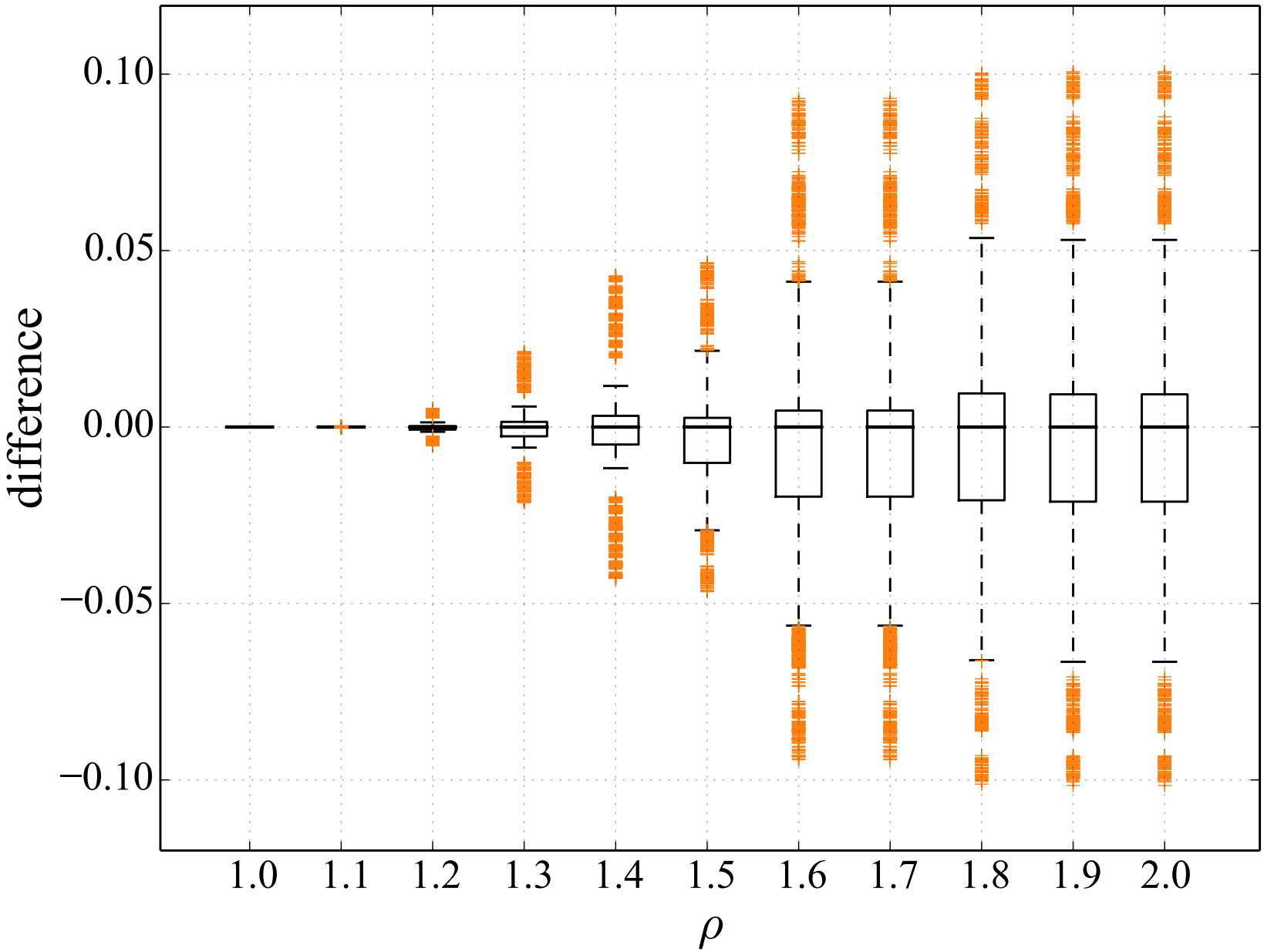
Variation in Run Score of AP



Variation in Run Score of RBP ($p = 0.85$)



Variation in Run Score of RBP ($p = 0.5$)



Are the Differences Significant?

Using the one-tail paired t–test, compare the banded run scores to the original run scores $\times$ 97%

Number of systems (max. 80) for which a t–test across 50 topics yields confidence at the $p \leq 0.05$ that the banded run score greater than or equal to 97% of the original run score.

ρ	Relative to 97% of original run score			
	RR	RBP0.5	RBP0.85	AP
1.1	80	80	80	80
1.2	80	80	80	80
1.4	80	80	80	80
1.7	80	67	80	77
2.0	80	61	71	20

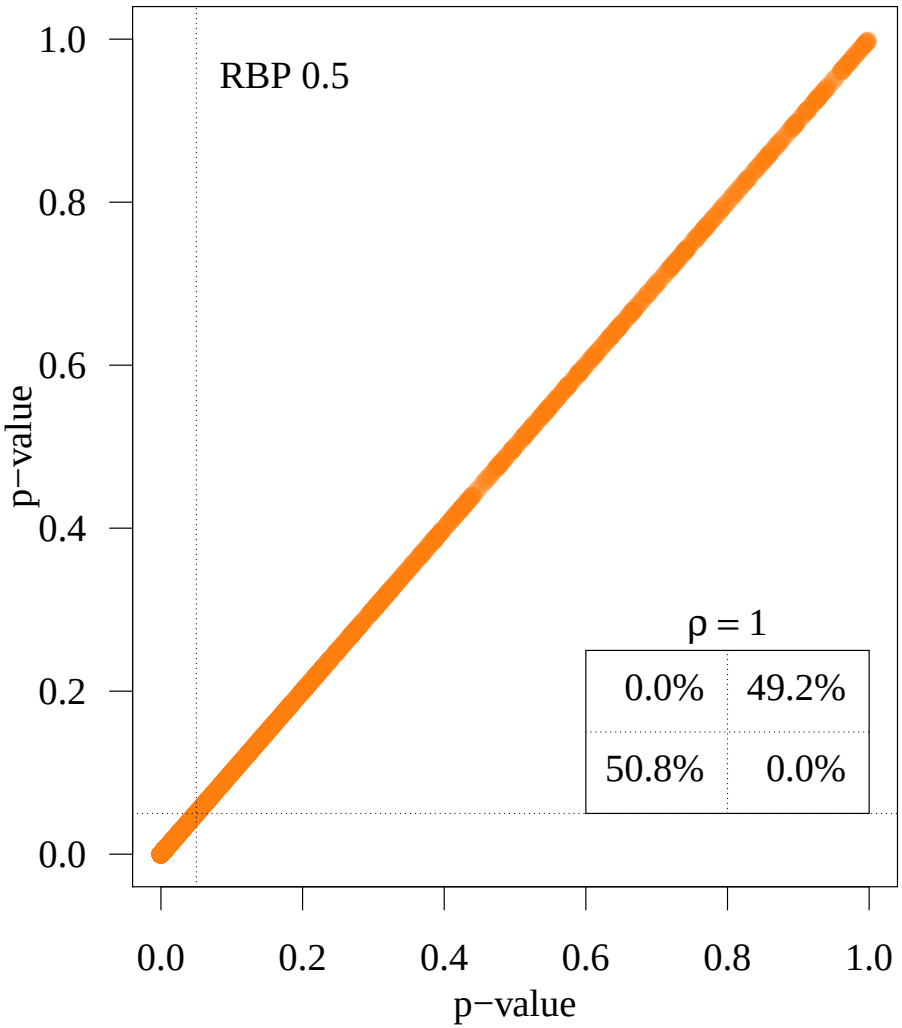
System Comparison Sensitivity

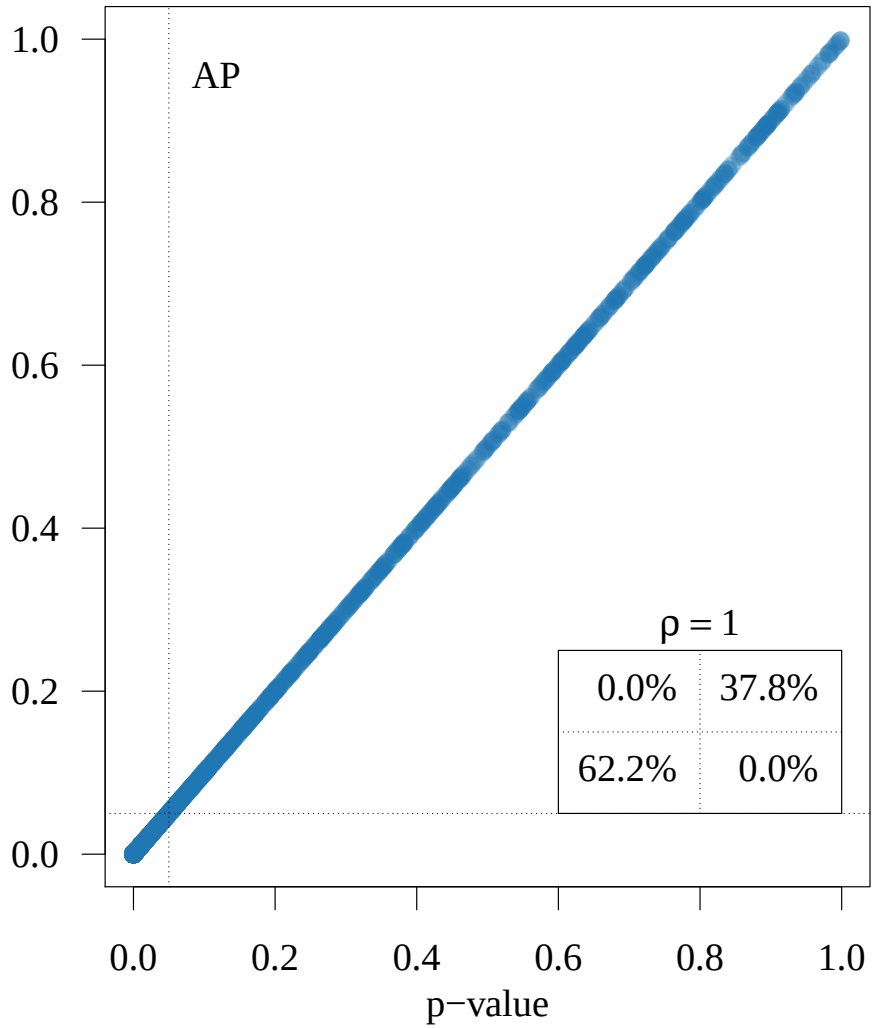
Compare systems in pairwise to explore the implications that the similarity score rounding has on ability the metrics to differentiate systems.

Perform the paired t–test using the original runs and grouped runs with different value of ρ:

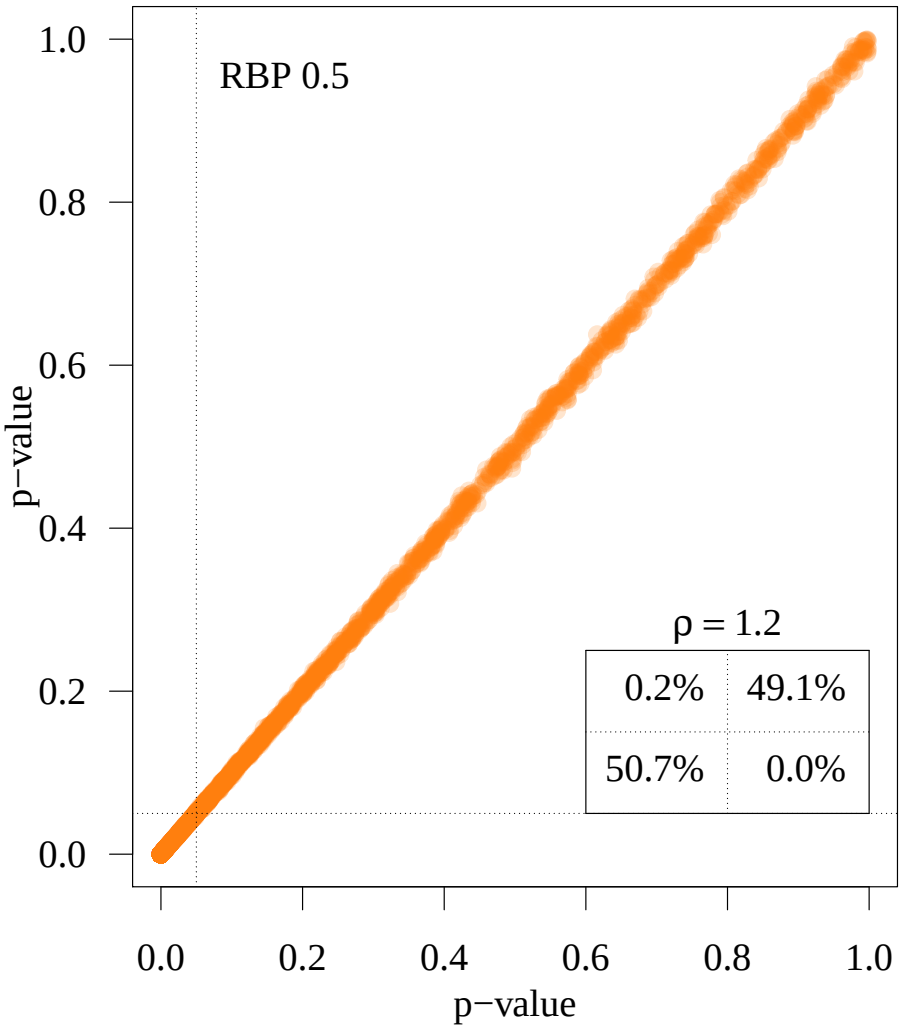
- Generate $80 \times 79/2$ system pairs
- Use their run scores given by metrics
- Explore the null hypothesis: two systems in a pair are same
- Check if the generated $p \leq 0.05$

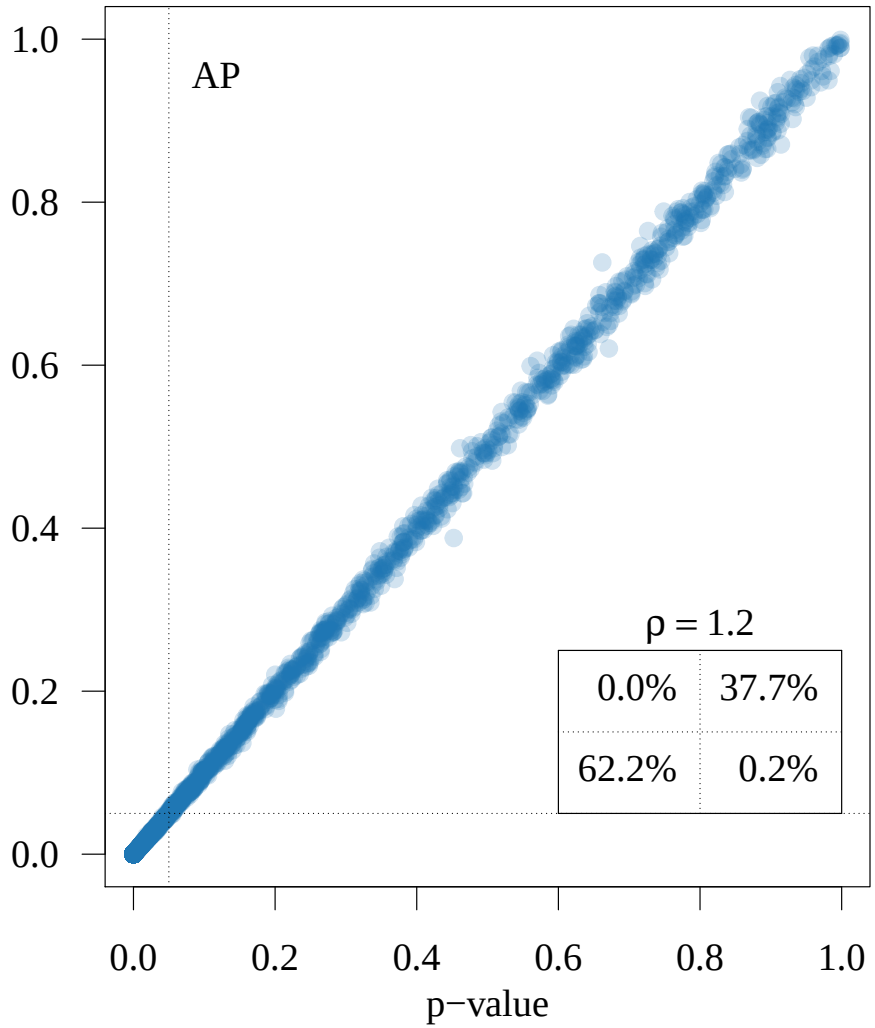
Correlation of p Values for All System Pairs



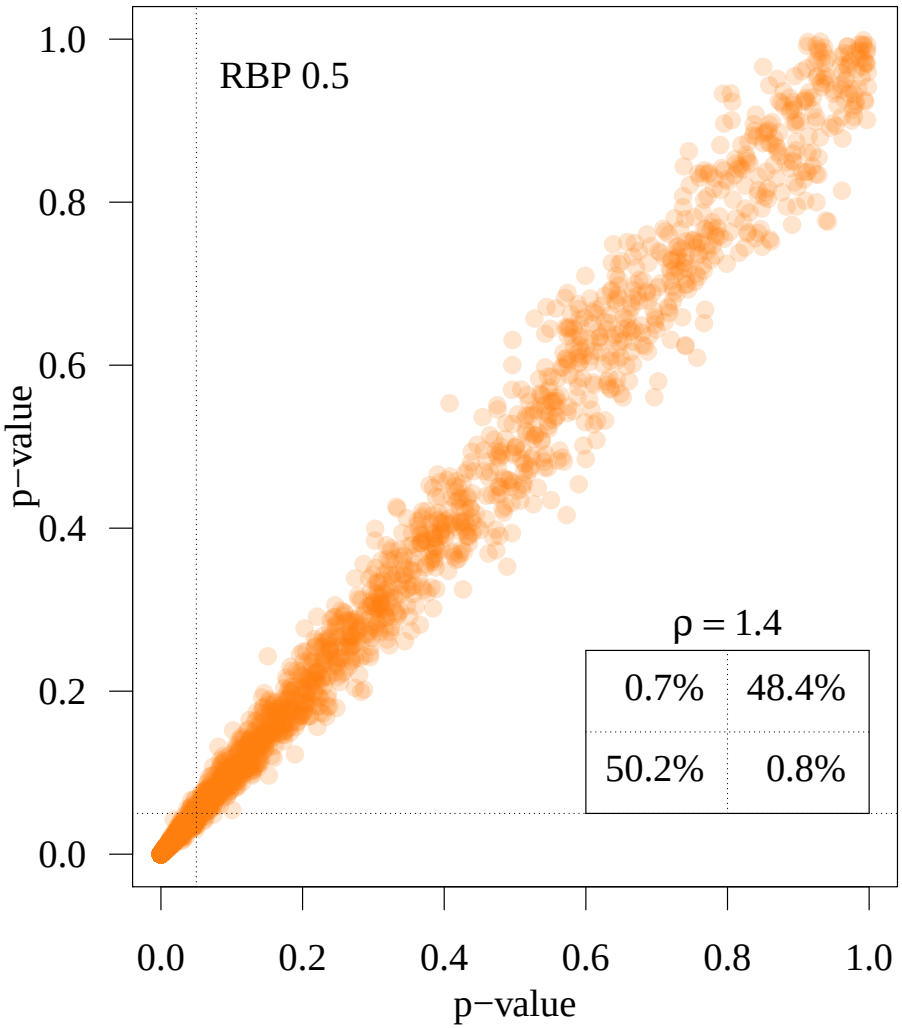


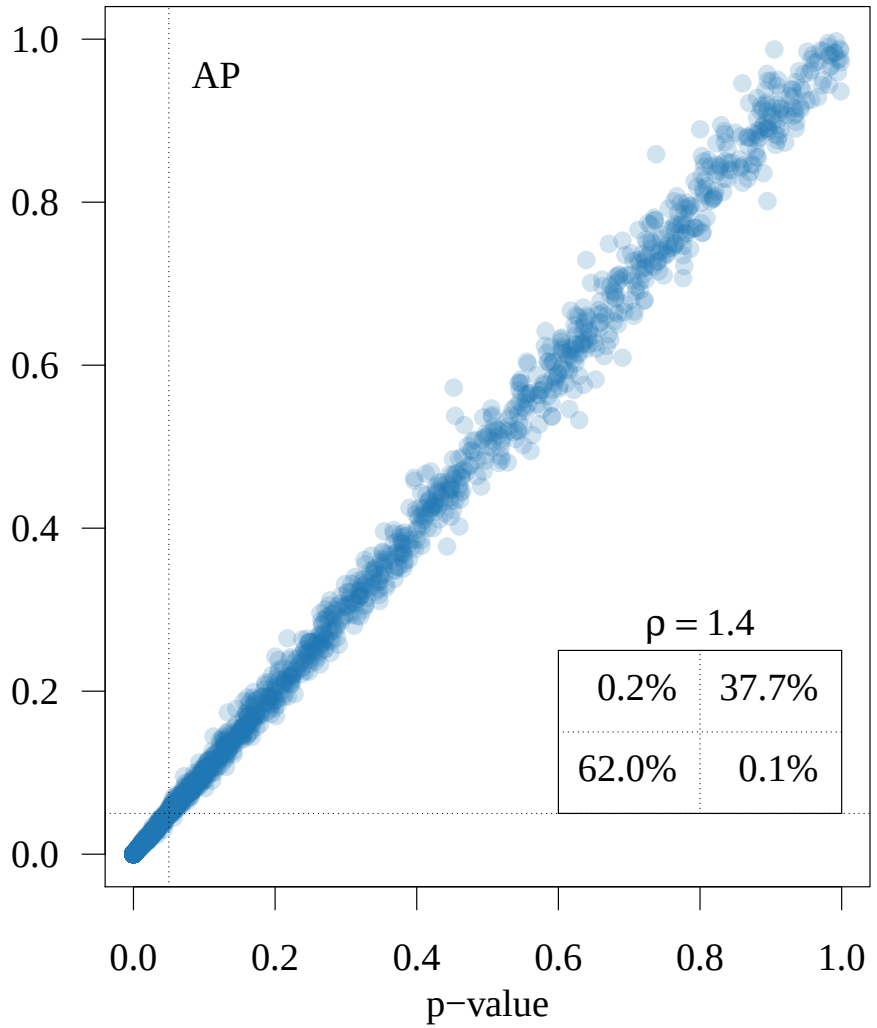
Correlation of p Values for All System Pairs



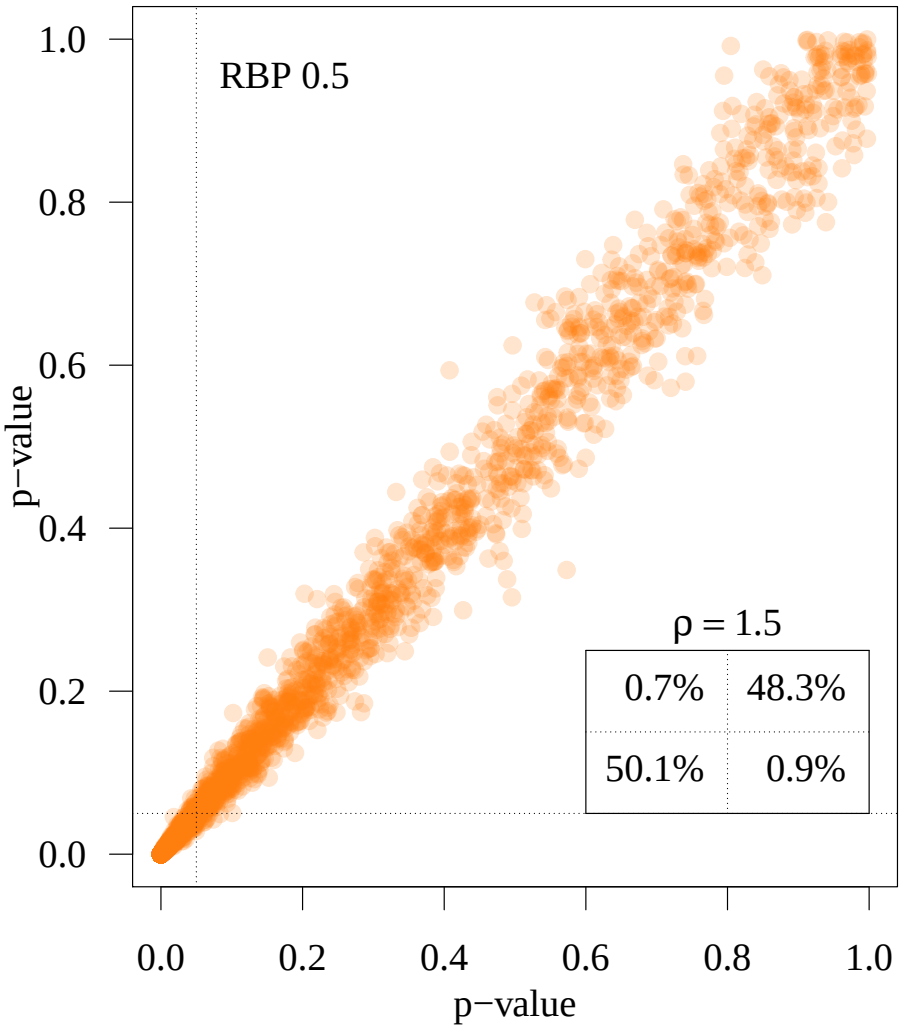


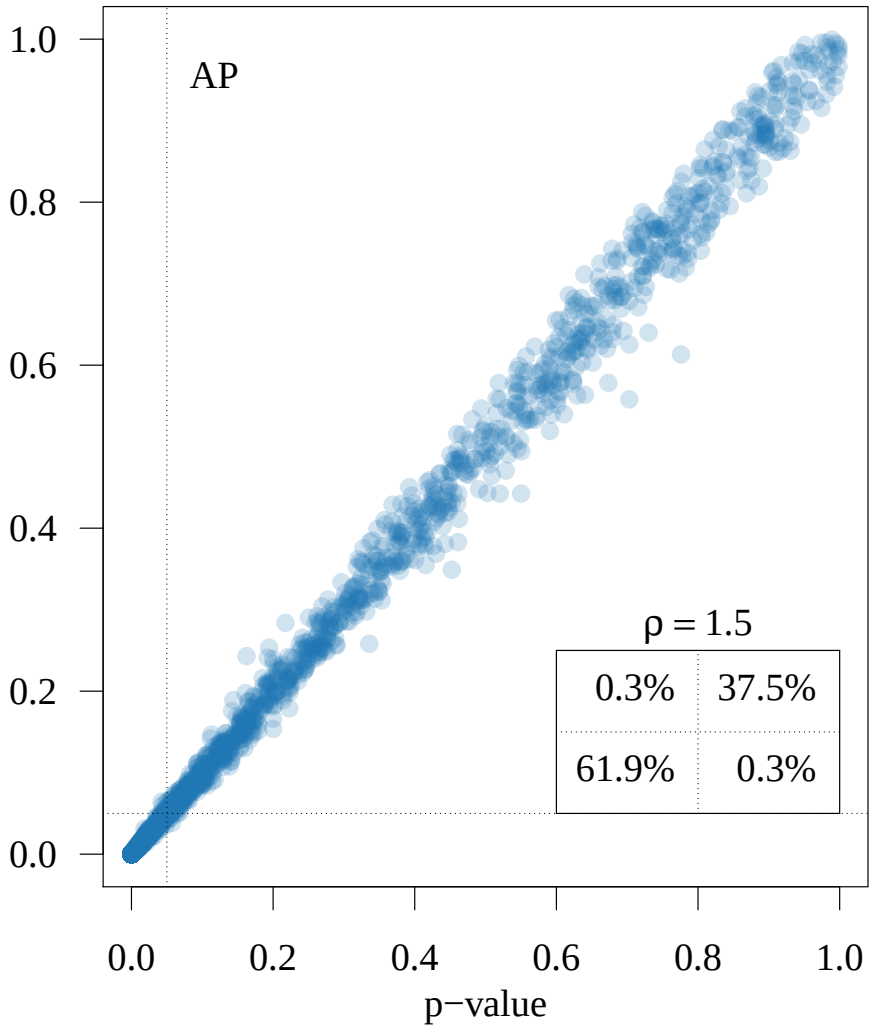
Correlation of p Values for All System Pairs



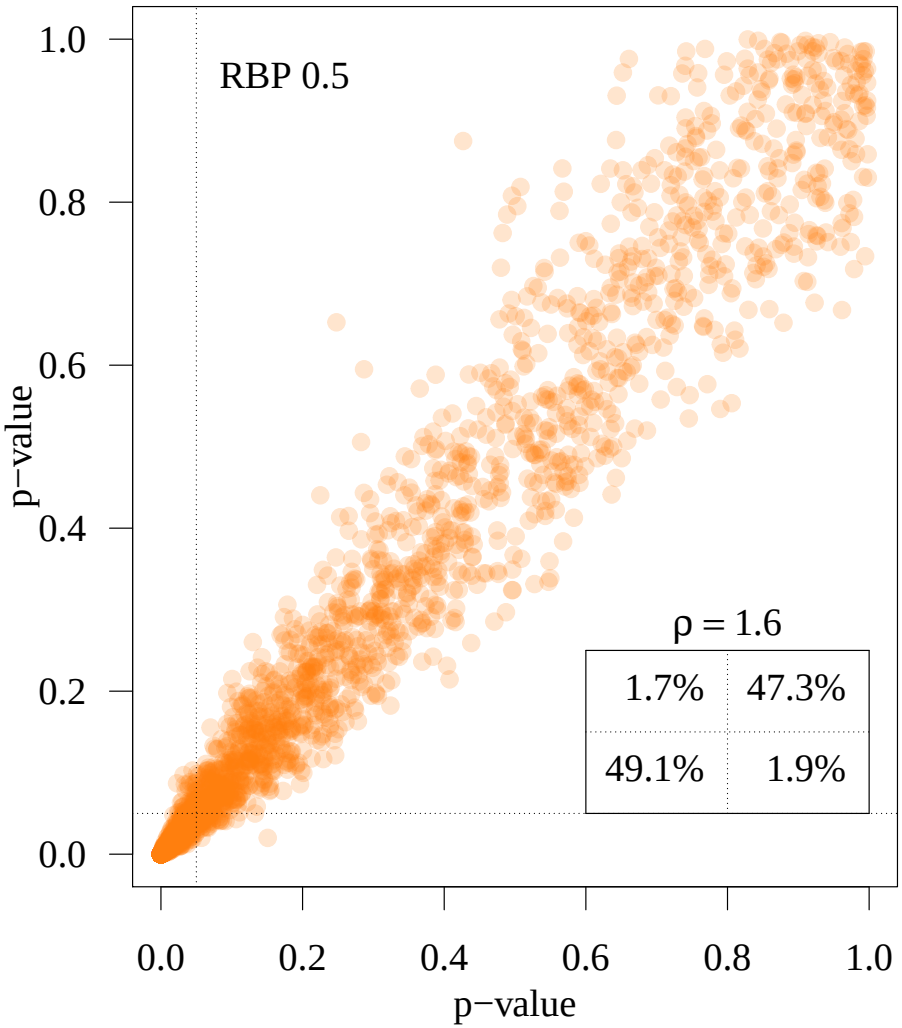


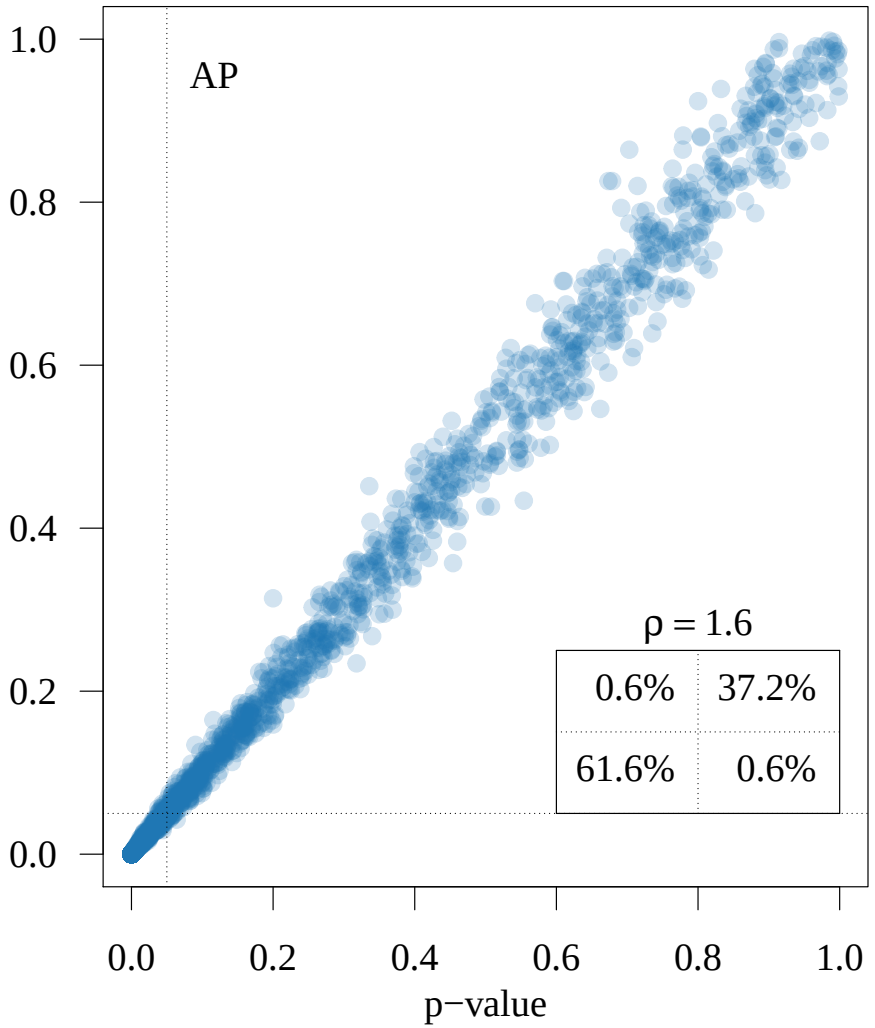
Correlation of p Values for All System Pairs



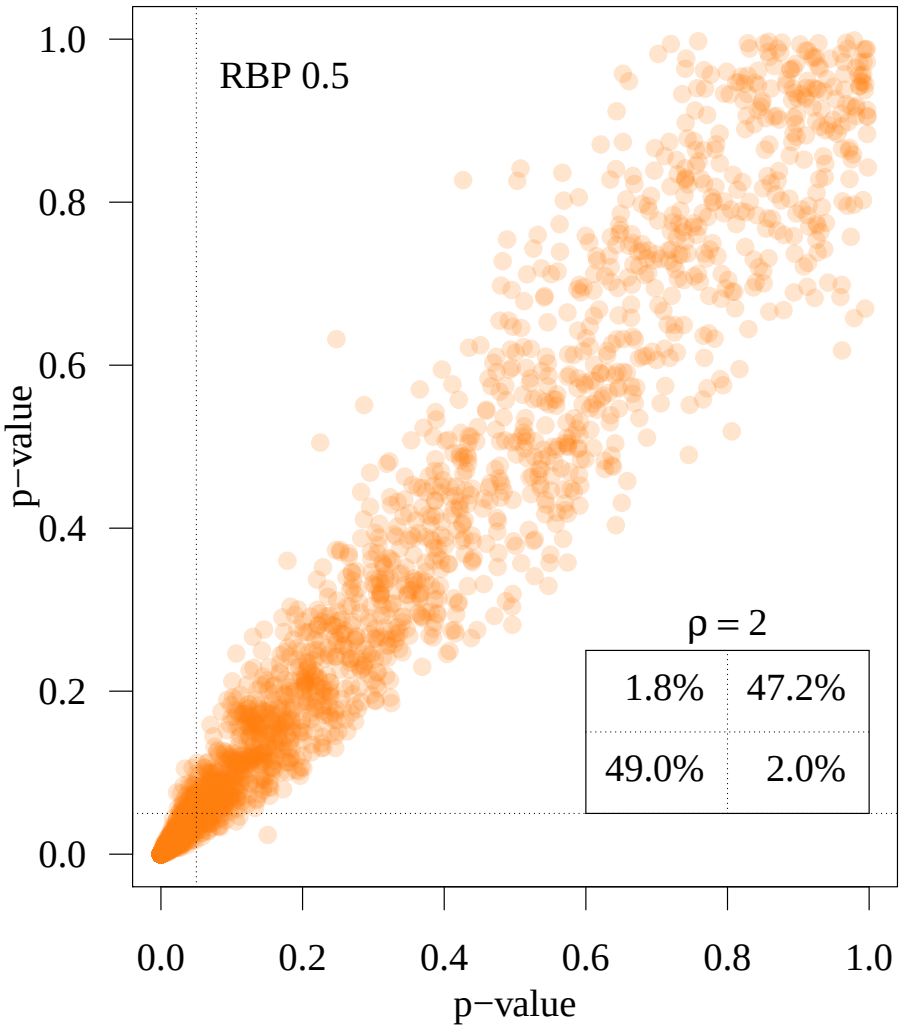


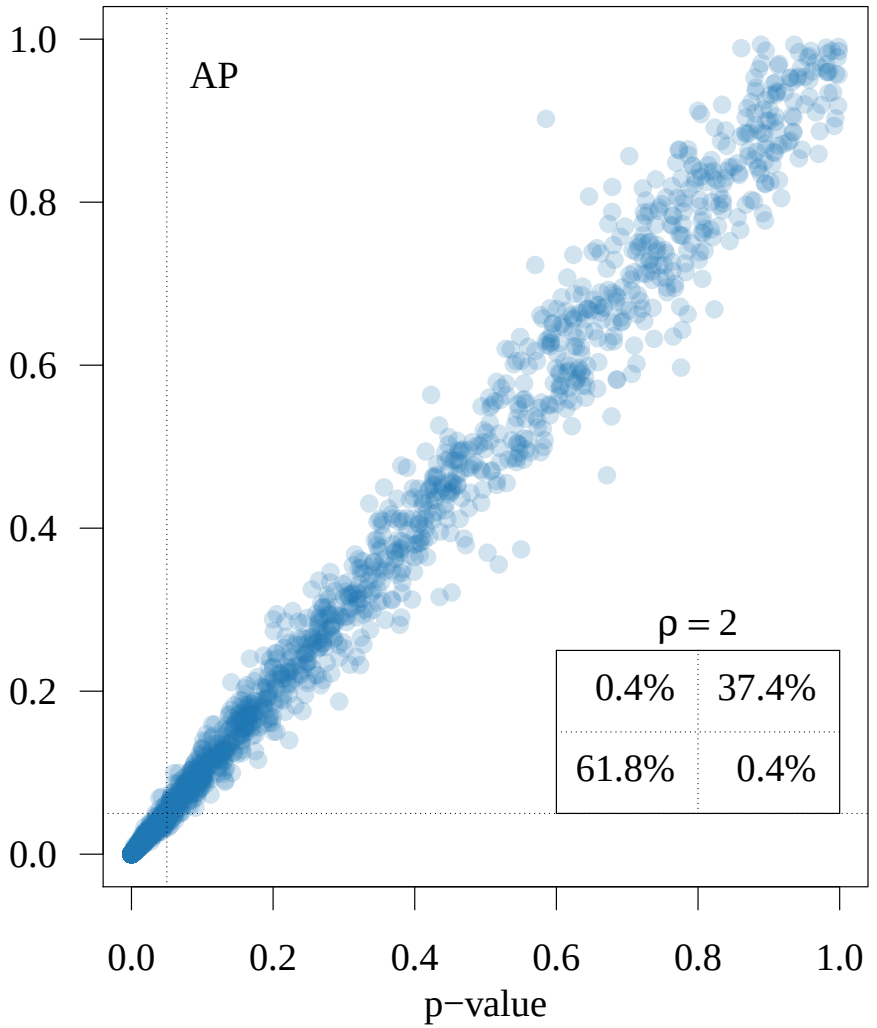
Correlation of p Values for All System Pairs





Correlation of p Values for All System Pairs





Summary

Similarity score ties do have potential to affect system comparisons. But fortunately, in practice, they did not.

Allowing deliberate introduction of ties in runs by grouping rules resulted only small changes in the ability to compare systems. Reducing the accuracy of similarity scores to improve search speed and reduce space used is feasible.