Bias and Fairness in Effectiveness Evaluation by Means of Network Analysis and Mixture Model

Michael Soprano, Kevin Roitero and Stefano Mizzaro

IIR 2019, Padova, 16 September 2019

Detect Bias in Information Retrieval

- Three existing works:
    - **HITS hits TREC**: detect bias in IR evaluation using network analysis [Mizzaro and Robertson 2007]
    - **HITS hits Readersourcing**: detect bias in the Readersourcing model using network analysis [Soprano et al. 2019]
    - **ANOVA to model IR effectiveness**: breaking down the effect caused by the components of a test collection in IR evaluation

We propose to extend results from these three works to define an engineered pipeline to find and correct the bias in the IR evaluation setting.

[Mizzaro and Robertson 2007] Mizzaro, Stefano and Robertson, Stephen. "HITS hits TREC: exploring IR evaluation results with network analysis." 30th ACM SIGIR, 2007.
[Soprano et. al 2019] Soprano, Michael and Roitero, Kevin and Mizzaro, Stefano. "HITS Hits Readersourcing: Validating Peer Review Alternatives Using Network Analysis." 4th BIRNDL Workshop @ 42th ACM SIGIR 2019.

HITS hits TREC

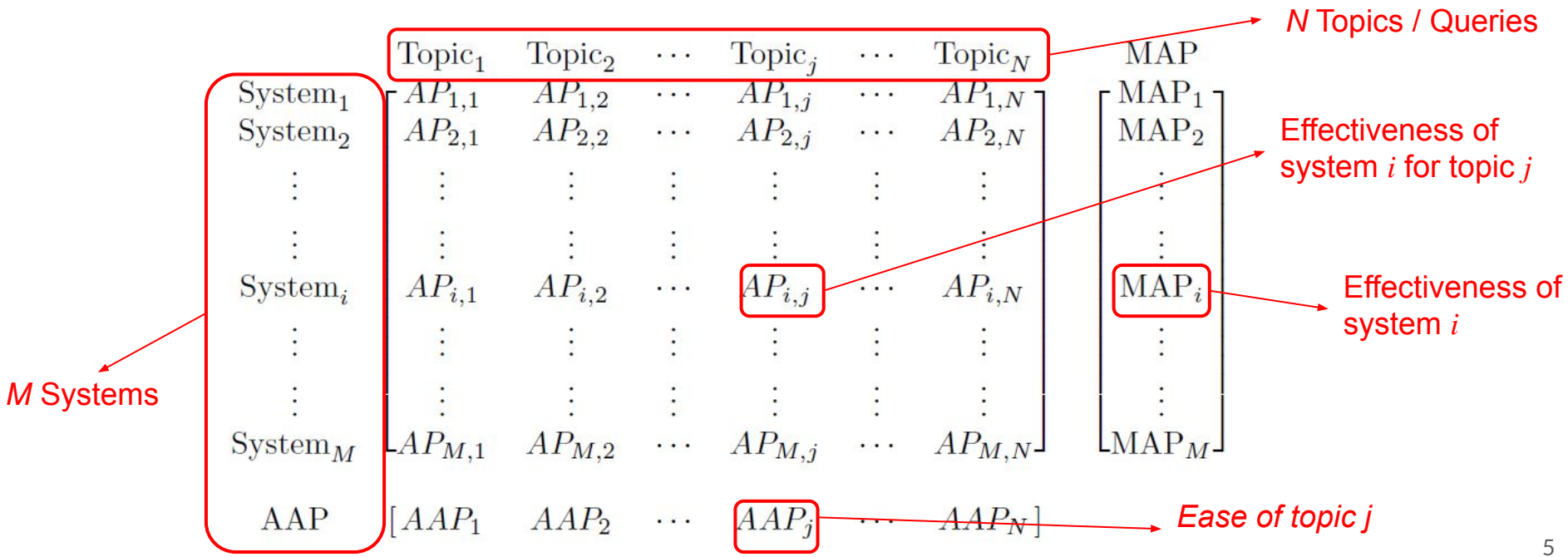
Information Retrieval Evaluation

- Test collection evaluation:
  - Document collection
  - Information Needs ≈ Queries (called topics)
  - Information Retrieval systems
- Each system retrieves a ranked list of documents for each topic
- Human made relevance judgments
- Metrics (such as Precision, Recall, NDCG, etc.) are computed
- Systems are then ranked according to the metrics

[Mizzaro and Robertson 2007] proposed to identify possible biases in the model using network analysis

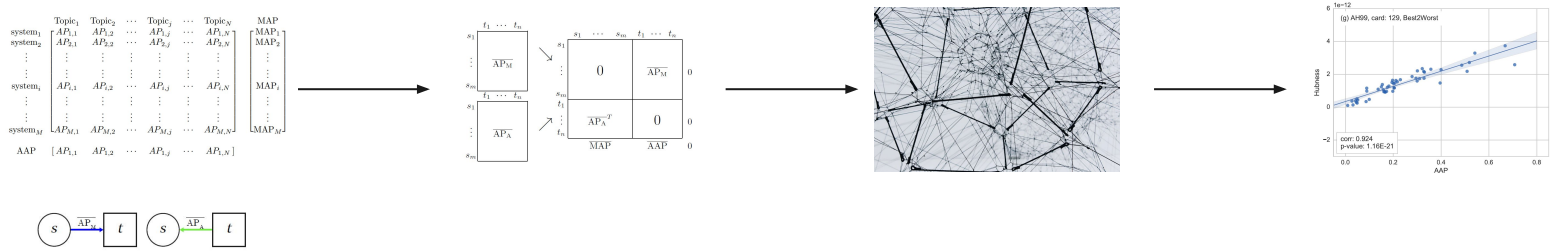
[Mizzaro and Robertson 2007] Mizzaro, Stefano and Robertson, Stephen. "HITS hits TREC: exploring IR evaluation results with network analysis." 30th ACM SIGIR, 2007.

Information Retrieval Evaluation -- Matrix



The Pipeline for Bias Detection

Matrix Normalization → Build the graph → Network Analysis (HITS) → Discover Bias

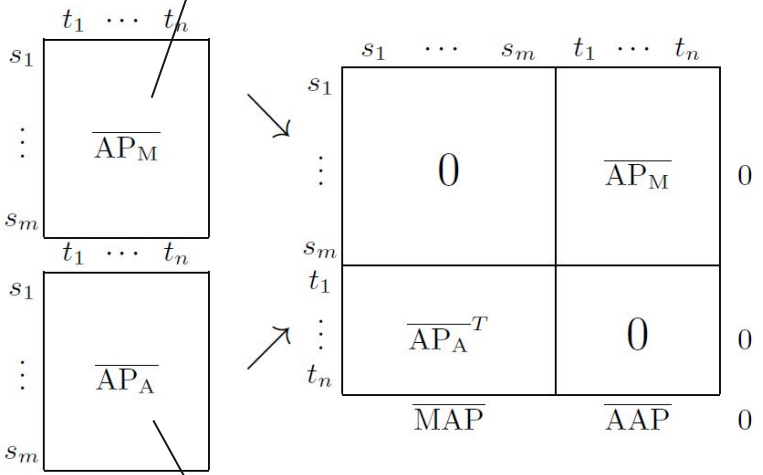


Building a Graph

How much a system "thinks" a topic is difficult

How much a topic "thinks" a system is effective

All systems are equally effective.
System effectiveness bias removed.



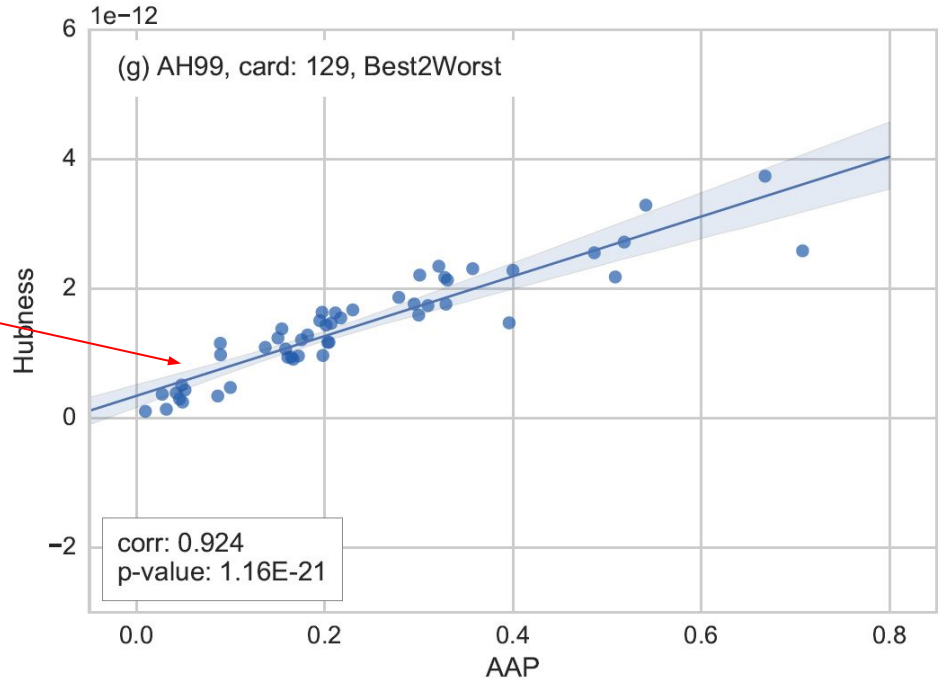
All topics have the same difficulty.
Topics difficulty bias removed.

Discovering Bias

Ability to recognize most effective systems

Easy topics identify better the final rank of retrieval systems (positive correlation)

Ability to recognize least effective systems



difficult topics ⟷ easy topics

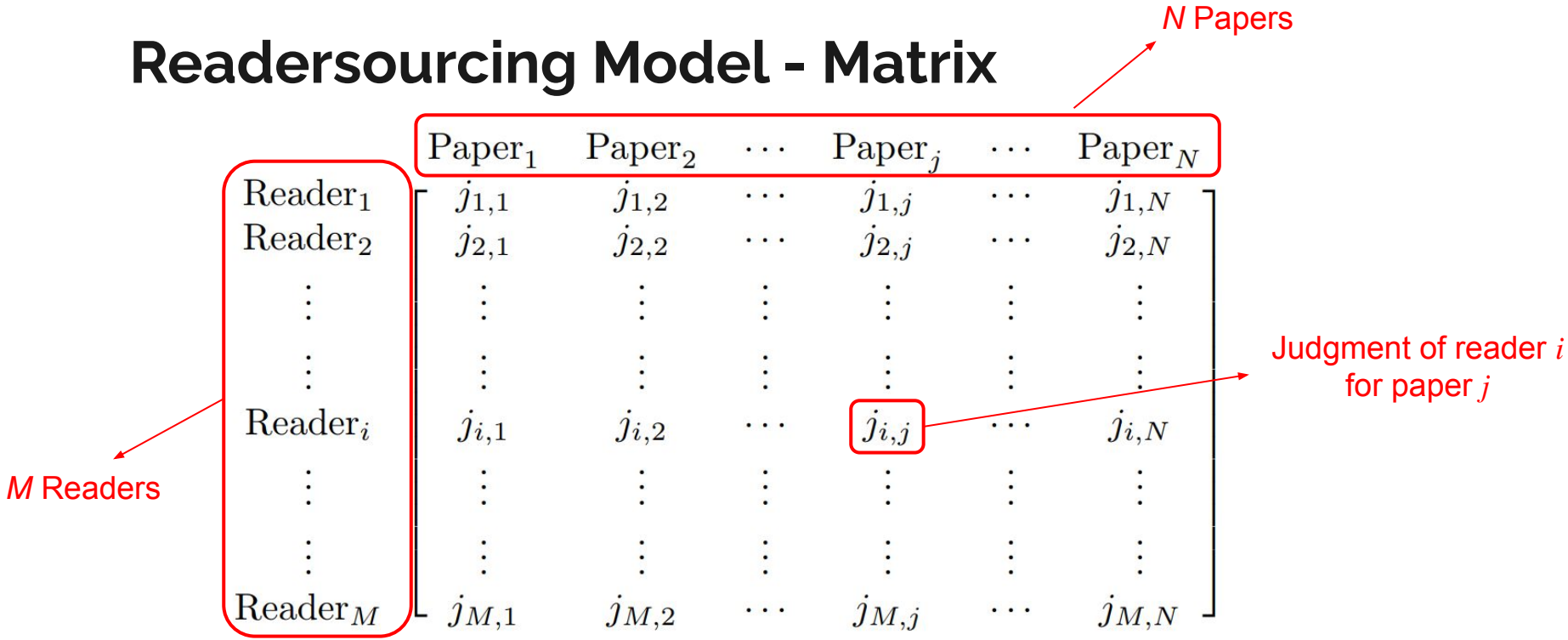
HITS hits Readersourcing

Readersourcing in a Nutshell

- A model to outsource the peer review activity of publications to their readers (Crowdsourcing)
- We have plenty of readers that read papers… a lot of readers in place of a few referees...
- Two models:
  - Readersourcing [Mizzaro 2012]
  - TrueReview [De Alfaro and Faella 2016]
- … we do not have enough time …
- Not only in theory! www.readersourcing.org [Soprano and Mizzaro 2019]
  - (Still beta-ish though it works!)

[Mizzaro 2012] Mizzaro, Stefano. "Readersourcing - A Manifesto." JASIST 63(8), 1666-1672 (2012).
[De Alfaro and Faella 2016] De Alfaro, Luca and Faella, Marco. "TrueReview: A Platform for Post-Publication Peer Review". CoRR (2016).
[Soprano and Mizzaro 2019] Soprano, Michael and Mizzaro, Stefano. "Crowdsourcing Peer Review: As We May Do." Digital Libraries: Supporting Open Science. Springer. (2019)

Readersourcing Model - Matrix



$$
\begin{array}{c|cccccc}
 & \text{Paper}_1 & \text{Paper}_2 & \cdots & \text{Paper}_j & \cdots & \text{Paper}_N \\
\text{Reader}_1 & j_{1,1} & j_{1,2} & \cdots & j_{1,j} & \cdots & j_{1,N} \\
\text{Reader}_2 & j_{2,1} & j_{2,2} & \cdots & j_{2,j} & \cdots & j_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Reader}_i & j_{i,1} & j_{i,2} & \cdots & j_{i,j} & \cdots & j_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Reader}_M & j_{M,1} & j_{M,2} & \cdots & j_{M,j} & \cdots & j_{M,N}
\end{array}
$$

Stochastic Simulations to Populate the Matrix

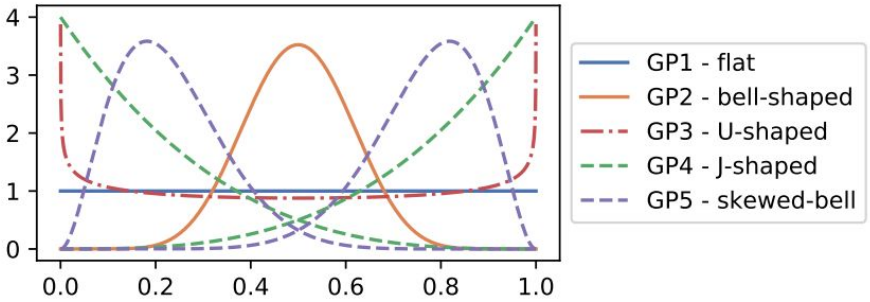
- Stochastic simulations
- **Readers** are divided into groups
- Each group of readers judges papers with a certain frequency

Group	Frequency	Amount
GR_1	1 x 2 Weeks	2
GR_2	1 x Week	4
GR_3	2 x Week	8
GR_4	1 x Day	30
GR_5	3 x Day	90

Stochastic Simulations to Populate the Matrix

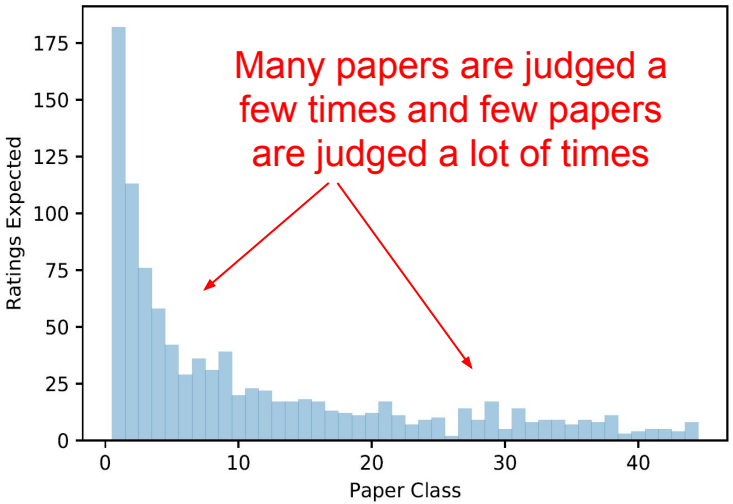
- **Papers** are divided into groups
- Each group of papers is rated along a curve of the beta distribution
- Different “rating behaviors”

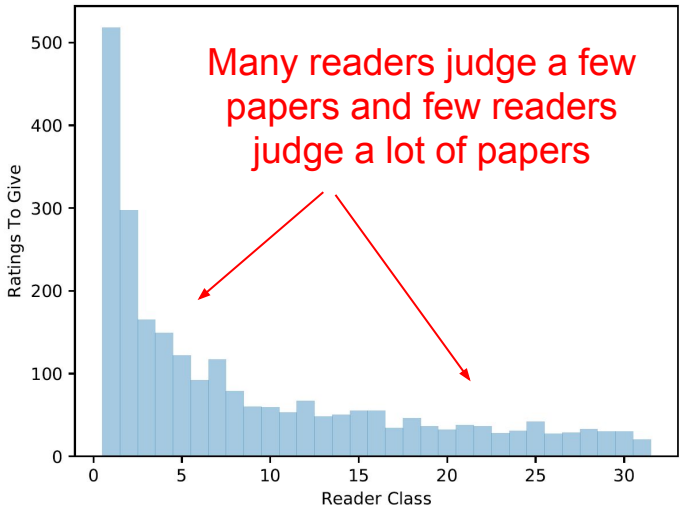
Group	% Parameters	Shape
GP_1	5% $(\alpha = 1) \wedge (\beta = 1)$	flat
GP_2	30% $(\alpha = \beta) \wedge (\alpha > 1) \wedge (\beta > 1)$	bell-shaped
GP_3	20% $(0 < \alpha < 1) \wedge (0 < \beta < 1)$	U-shaped
GP_4	30% $(\alpha > 1 \wedge \beta = 1) \vee (\alpha = 1 \wedge \beta > 1)$	J-shaped
GP_5	15% $(\alpha > 1 \wedge \beta > 1) \wedge (\alpha \neq \beta)$	skewed-bell



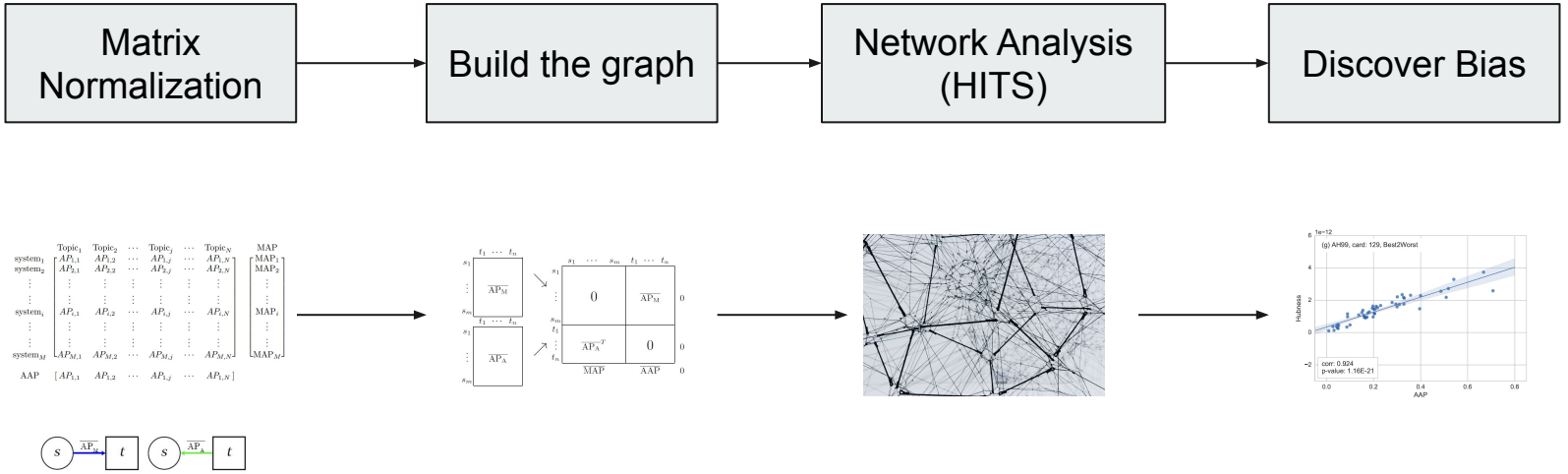
We Are Already Improving It...

- Generation by means of two power-law distributions

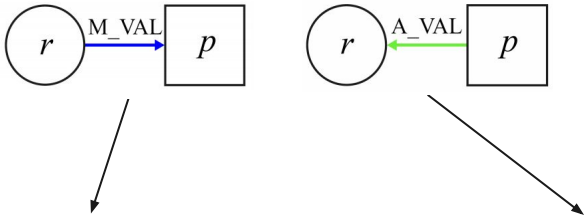




The Pipeline for Bias Detection



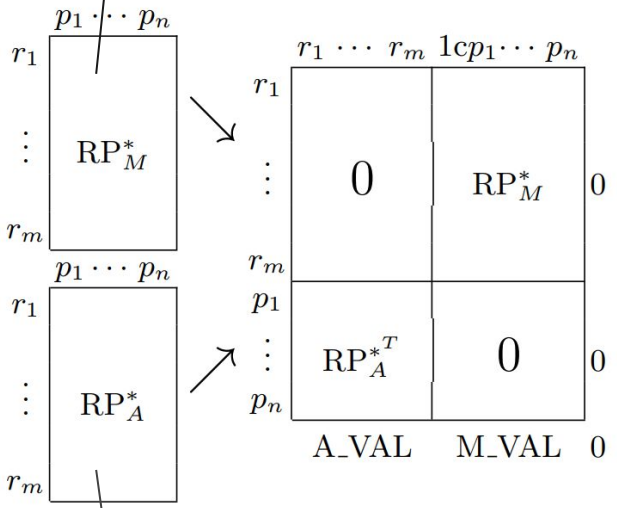
Building Multiple Graphs



how much a reader “thinks” a paper is good

how much a paper “thinks” a reader is of high quality

All readers are of the same quality.
All readers judge on the same scale.
...
Multiple Reader biases removed.



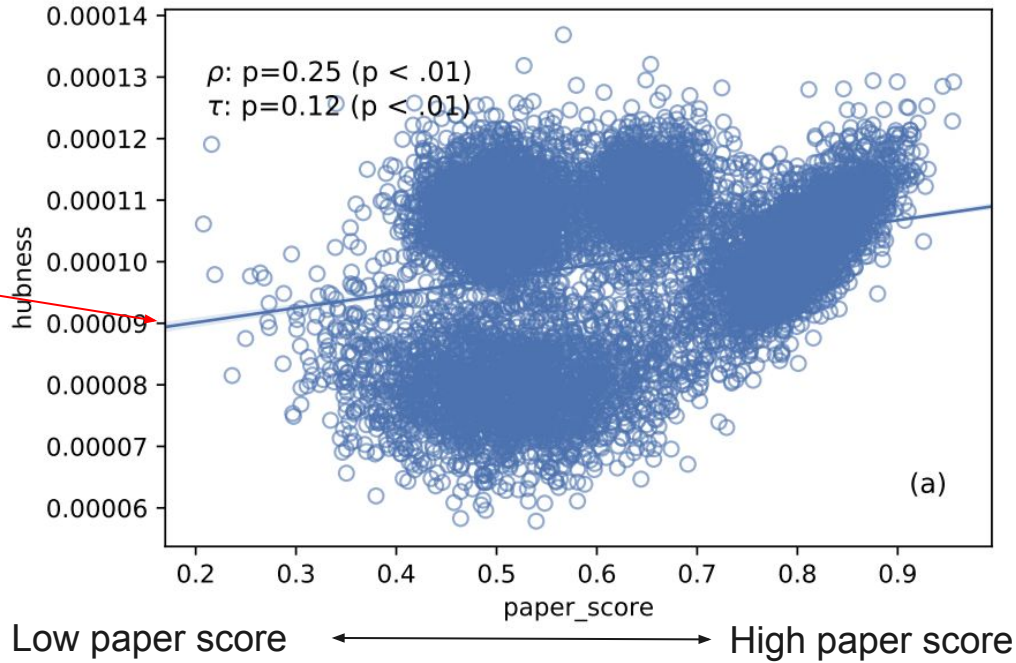
All papers have the same quality.
All papers have the same mean judgment.
...
Multiple Paper biases removed.

Discovering Biases

Ability to recognize high quality readers

Whether a paper has a high or low score, it has the same capability to recognize readers that tend to express high quality judgments

Ability to recognize low quality readers

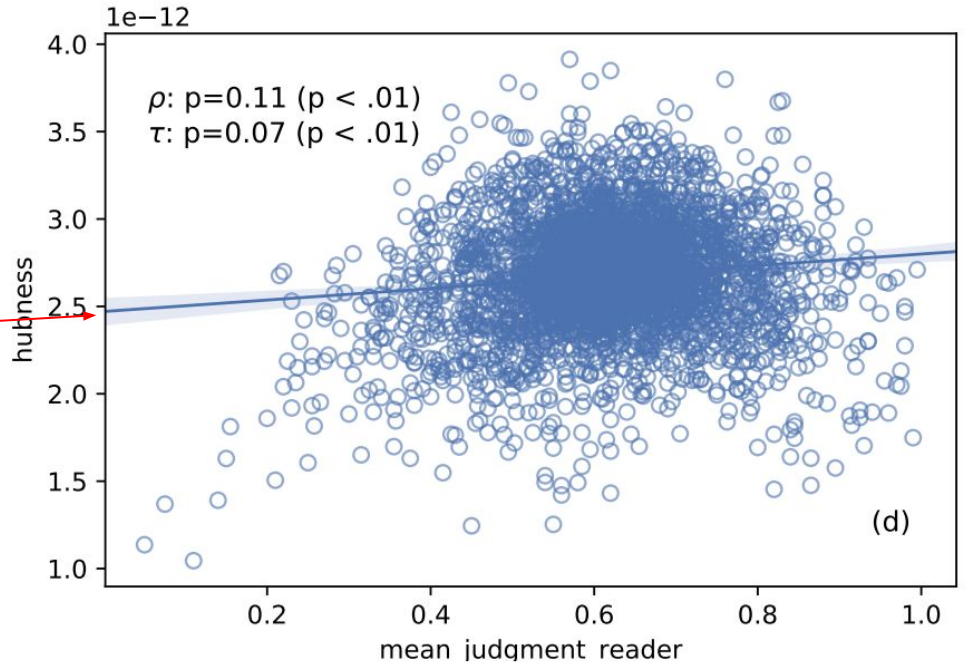


Discovering Biases

Ability to recognize papers which get high quality judgments

Whether a reader has a high or low mean judgment, it has the same capability of recognize papers that tend to get high quality judgments

Ability to recognize paper that get low quality judgments



Discovering Biases

- We build multiple graphs
  - i. e., we try multiple normalization
    - i.e., we try to remove different biases
- See [Soprano et. al 2019] for more results!

[Soprano et. al 2019] Soprano, Michael and Roitero, Kevin and Mizzaro, Stefano. "HITS Hits Readersourcing: Validating Peer Review Alternatives Using Network Analysis." 4th BIRNDL Workshop @ 42th ACM SIGIR 2019.

ANOVA to Model IR Effectiveness

Breaking Down Things

Breaking down the effect of a dimension on a complex system is a problem much studied in IR evaluation

- [Ferro and Silvello 2018] break down the system effectiveness score into its components by means of GLMM + ANOVA analysis
- [Ferro and Sanderson 2017] take the sub-corpora effect into consideration
- [Zamperi et al. 2019] provide a complete analysis on topic ease and the effect of system configurations, corpora, and interactions between components

[Ferro and Silvello 2018] Ferro, Nicola and Silvello, Gianmaria. "Toward an Anatomy of IR System Component Performances" JASIST 69(2), 187-200 (2018).
[Ferro and Sanderson 2017] Ferro, Nicola and Sanderson, Mark. "Sub-corpora Impact on System Effectiveness" JASIST 69(2), 187-200 (2018). 40th ACM SIGIR, 2017.
[Zamperi et al. 2019] Zamperi, Fabio and Roitero, Kevin and Culpepper, Shane and Kurland, Oren and Mizzaro, Stefano. "On Topic Difficulty in IR Evaluation: The Effect of Corpora, Systems, and System Components" 42th ACM SIGIR, 2019.

General Linear Mixed Model

- It is a mixed effect model, useful with data with more than one source of random variability
- It is an extension of the linear model to include fixed and random effects (mixed model)
- Allows to compute a size of effect index and quantify the magnitude of such effects
- Such techniques can be combined together along with the work of [Zamperi et. al 2019]
- [Ferro and Silvello 2018] and [Ferro and Sanderson 2017] use GLMM + ANOVA to study the effects of test collection components on system effectiveness in IR evaluation

[Ferro and Silvello 2018] Ferro, Nicola and Silvello, Gianmaria. "Toward an Anatomy of IR System Component Performances" JASIST 69(2), 187-200 (2018).
[Ferro and Sanderson 2017] Ferro, Nicola and Sanderson, Mark. "Sub-corpora Impact on System Effectiveness" JASIST 69(2), 187-200 (2018). 40th ACM SIGIR, 2017.
[Zamperi et al. 2019] Zamperi, Fabio and Roitero, Kevin and Culpepper, Shane and Kurland, Oren and Mizzaro, Stefano. "On Topic Difficulty in IR Evaluation: The Effect of Corpora, Systems, and System Components" 42th ACM SIGIR, 2019.

Experiments

Information Retrieval Evaluation

- Test collection evaluation:
  - Document collection
  - Information Needs ≈ Queries (called topics)
  - Information Retrieval systems
- Each system retrieves a ranked list of documents for each topic
- Human made relevance judgments
- Metrics (such as Precision, Recall, NDCG, etc.) are computed
- Systems are then ranked according to the metrics

We propose to extend results from the previous work to find and correct bias in this setting...

… How?

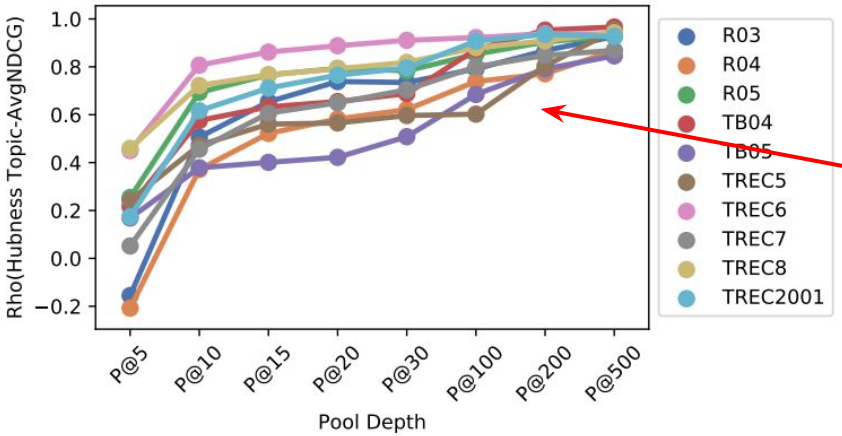
We plan to investigate how a specific bias (effective systems being recognized by easy topics) varies when varying the components of a test collection:

- Different pool depths effect
- Different collections and corpora effect
- Different effectiveness evaluation metric effect
- Different system and topic population effect

We propose to use a General Linear Mixed Model to compute the magnitude of the effect on model bias obtained by varying such components

Pool Depth Effect

We are planning to compute, for each effectiveness metric, its value at difference cut-offs.



Each point is a measure of the model bias at a given pool depth

In this preliminary result there is a trend which suggests that bias grows along with pool depth.

Effectiveness Metric Effect

We are planning to investigate the effect of different evaluation metric in the model bias

- Various approaches in the normalisation step can be used
  - Remove the average of system effectiveness and topic ease
  - More complex approaches
    - Remove the top-heaviness
    - Enhance precision-oriented systems
    - Enhance recall-oriented systems

Collection and Corpora Effect

We are planning to investigate the effect of different collections in the model bias and to break down the sub-corpora effect by considering the different corpora of the collections

- TREC Collections
  - Terabyte 2004 and 2005
  - Robust 2003, 2004 and 2005
  - TREC5, TREC6, TREC7, TREC8 and TREC2001

System and Topic Population Effect

We are planning to investigate the effect of different systems and topic populations in model bias

- Various approaches can be used
  - Systems ordered by effectiveness
  - Topics ordered by difficulty
  - ...

General Linear Mixed Model

To wrap-up the effects that different components of a test collection have on model bias the following GLMM can be defined:

$$\begin{aligned}\text{Bias}_{ijklm} = \text{Pool}_i \; + \; \text{Collection}_j \; + \; \text{Corpora}_k \; + \; \text{System-subset}_l \\ + \; \text{Topic-subset}_m \; + \; (\text{interactions}) + \; \text{Error}.\end{aligned}$$

This equation allows to compute the size of effect index and measure the magnitude given by the different components of a test collection on model bias.

Conclusions and Future Work

- We propose a tentative engineered pipeline base on network analysis and mixture model
  - Detect bias and its causes in IR evaluation and other domains (Readersourcing)
- We outline some preliminary results
- We are doing our experiments on more model biases

Thank You!