HITS Hits Readersourcing: Validating Peer Review Alternatives Using Network Analysis

Michael Soprano, Kevin Roitero, and Stefano Mizzaro

Dept. of Mathematics, Computer Science, and Physics. University of Udine, Italy.
michael.soprano@outlook.com, roitero.kevin@spes.uniud.it,
mizzaro@uniud.it

Abstract. Peer review is a well known mechanism exploited within the scholarly publishing process to ensure the quality of scientific literature. Such a mechanism, despite being well established and reasonable, is not free from problems, and alternative approaches to peer review have been developed. Such approaches exploit the readers of scientific publications and their opinions, and thus outsource the peer review activity to the scholar community; an example of this approach has been formalized in the *Readersourcing* model [5]. Our contribution is two-fold: (i) we propose a stochastic validation of the Readersourcing model, and (ii) we employ *network analysis* techniques to study the bias of the model, and in particular the interactions between readers and papers and their goodness and effectiveness scores. Our results show that by using network analysis interesting model properties can be derived.

1 Introduction

Peer review is an a priori mechanism exploited within the scholarly publishing process to ensure the quality of scientific literature; an article written by some authors undergoes peer review when it is judged and rated by colleagues of the same degree of competence. Such mechanism, despite being well established and reasonable, is not free from problems; indeed, it is characterized by various issues related to the process itself and the malicious behavior of some stakeholders [8].

In literature one can find alternative approaches to peer review, which exploit readers of scientific publications and their opinions as a "review force", thereby outsourcing the peer review activity itself to the community of readers. One of these approaches has been proposed by Mizzaro [5] and called *Readersourcing*, as a portmanteau for "crowdsourcing" and "readers", and it is based on a model proposed in a previous work [4]. Another similar model is *TrueReview* [2]. Soprano and Mizzaro [8] describe a general ecosystem called *Readersourcing 2.0* which provides an implementation for such models.

The aim of the Readersourcing model is to define a way to measure the overall quality of a published article as well as the reputation of a scholar as a reader / assessor; moreover, from these measures it is possible to derive the reputation of a scholar as an author. In other terms, the main issue to address is how the numerical judgments given to publications should be aggregated into indexes of quality and, from these indexes, how to compute indexes of reputation for the readers and, eventually, indexes of how much an author is able to publish

papers which are positively rated by their readers. Therefore, to each entity (i.e., publications, authors, and readers) is assigned one or more scores which measure how much good (skilled) it is.

Network analysis is a discipline which studies features and properties of (usually large) networks or graphs. Its algorithms can be can be quite general and, therefore, applicable to different domains. Mizzaro and Robertson [6] exploit link analysis techniques such as the HITS algorithm proposed by Kleinberg [3] to address a research question related to the effectiveness evaluation of Information Retrieval (IR) systems. The evaluation of IR systems is performed within different initiatives, such as TREC (Text REtrieval Conference). Before the actual conference, TREC provides a test collection made of documents and topics (i.e., representations of information needs); such a test collection is used as a benchmark to compare the performance of different IR systems. Participants use their systems to retrieve, and submit to TREC, a list of documents for each topic. System effectiveness is then measured by well established metrics like Mean Average Precision (MAP) and a final ranking is built. Mizzaro and Robertson [6] study the interactions between the difficulty of topics and the final rank of IR systems. In particular, they investigate the correlation between topic ease and the ability to predict system effectiveness and they find that to be effective, a system has to perform well on easy topics. Such finding is quite undesirable since difficult topic are more useful to allow IR to evolve. Roitero et al. [7] extend the work of Mizzaro and Robertson [6] by performing a more detailed analysis on three different datasets: they confirm that the original result is valid and general across datasets; they find that when only the most effective IR systems are considered there is no evidence that the ranking is affected only by easy topics; and they prove that such results are robust across different effectiveness metrics.

In this paper we take advantage of the methodology proposed by Mizzaro and Robertson [6] and extended by Roitero et al. [7] to address a similar research question related to the Readersourcing model. More in detail, we intend to study the interactions between the skill of a reader and the quality of a paper, where such quantities are computed by Readersourcing models. This paper is structured as follows. Section 2 details the related work; Section 4 describes the experiments performed; Section 5 discusses the results. Finally, Section 6 concludes the paper.

2 Background

In an attempt to make this paper self contained, in this section we summarize two major related work areas that we considered to do our analysis. Section 2.1 summarizes the Readersourcing model proposed by Mizzaro [5], while Section 2.2 summarizes the methodology proposed by Mizzaro and Robertson [6] to investigate the correlation between topic ease and the ability to predict system effectiveness within the effectiveness evaluation of IR systems activity.

2.1 The Readersourcing Model

In the Readersourcing model three entities are identified: papers, readers, and authors. The score of an author is simply defined as a weighted average of his or her papers; we do not analyze it in detail in this paper, where we focus on

	p_1	$\cdots$	p_n	$\mathrm{S_r}$	σ_r
r_1	j_{r_1,p_1}	$\cdots$	j_{r_1,p_n}	$\mathrm{S_r}(r_1)$	$\sigma_\mathrm{r}(r_1)$
$\vdots$		$\ddots$		$\vdots$	$\vdots$
r_m	j_{r_m,p_1}	$\cdots$	j_{r_m,p_n}	$\mathrm{S_r}(r_m)$	$\sigma_\mathrm{r}(r_m)$
$\mathrm{S_p}$	$\mathrm{S_p}(p_1)$	$\cdots$	$\mathrm{S_p}(p_n)$		
σ_p	$\sigma_\mathrm{p}(p_1)$	$\cdots$	$\sigma_\mathrm{p}(p_n)$		

(a) RP matrix with judgments (RPJ)

	p_1	$\cdots$	p_n	$\mathrm{S_r}$	σ_r
r_1	$g(j_{r_1,p_1})$	$\cdots$	$g(j_{r_1,p_n})$	$\mathrm{S_r}(r_1)$	$\sigma_\mathrm{r}(r_1)$
$\vdots$		$\ddots$		$\vdots$	$\vdots$
r_m	$g(j_{r_m,p_1})$	$\cdots$	$g(j_{r_m,p_n})$	$\mathrm{S_r}(r_m)$	$\sigma_\mathrm{r}(r_m)$
$\mathrm{S_p}$	$\mathrm{S_p}(p_1)$	$\cdots$	$\mathrm{S_p}(p_n)$		
σ_p	$\sigma_\mathrm{p}(p_1)$	$\cdots$	$\sigma_\mathrm{p}(p_n)$		

(b) RP matrix with goodness values (RPG)

Fig. 1: Reader-paper matrices (RP) with judgments (goodness values), scores, and steadiness at a fixed timestamp.

more readers and papers. A generic reader is asked to give a numerical judgment to each paper he reads. Such judgments are used to compute a quality score for each paper. Likewise, each reader is characterized by a score which measures its skill/reputation. To each judgment is assigned a measure of its goodness with respect to other judgments given to the same paper. Moreover, to papers and readers is assigned a steadiness value which affects the update of the scores; a high (low) steadiness value leads to faster (slower) change of the score themselves.

Scores are dynamic and they change depending on user behaviour. For example, if an author with a low score publishes a paper positively rated by readers, his score increases; if a reader expresses a judgment which is judged as untruthful and/or biased because "distant" from other judgments (for a given paper) his score decreases, and so on. Therefore, there is a temporal dimension to consider, since the internal state of the model evolves as time passes. In the following, we hypothesize to "freeze" such state at a fixed timestamp, where no new judgments can be expressed and no new papers can be added. Figure 1 shows a representation of the model as a reader-paper matrix (RP) with m rows (i.e., readers) and n columns (i.e., papers) which can be represented in two ways. In the former (RPJ), each cell contains the numerical judgment given by reader r to paper p, while in the latter (RPG) each cell contains a measure of the goodness of the numerical judgment given by reader r to paper p. In both representations, each reader (paper) has a related score and steadiness pair, which are represented by the S_r and σ_r column (and S_p and σ_p row) vectors. These are computed according to the formulas defined by the Readersourcing model [4].

2.2 The HITS hits TREC methodology

The output of a TREC-like initiative can be represented as a system-topic matrix (ST) with m rows (i.e., systems) and n columns (i.e., topics). Each cell contains an effectiveness measure of each system with respect to each topic according to some metric such as Average Precision (AP). Each row is averaged to compute Mean Average Precision (MAP), which is a measure of system effectiveness with respect to all topics. Each column is averaged to compute Average Average Precision (AAP) which is a measure of topic ease.

The ST matrix is then normalized in two ways. Let us call AAP and MAP the AAP column and the MAP row of ST. In the former normalization, each $\mathrm{AP}(s_i, t_j)$ value is transformed into a $\overline{\mathrm{AP}}_\mathrm{A}(s_i, t_j)$ value (Normalized AP accord-

ing to AAP) by subtracting AAP to ST. In the latter, each $\mathrm{AP}(s_i, t_j)$ value is transformed into a $\overline{\mathrm{AP}_\mathrm{M}}(s_i, t_j)$ value (Normalized AP according to MAP) by subtracting MAP to ST. The normalized matrices ST_A and ST_M are exploited to study the interactions between topic ease and system effectiveness. More in detail, these two matrices can be merged into a single adjacency matrix which represents a complete weighted bipartite system-topic graph.

Each link $s \rightarrow t$ with weight $\overline{\mathrm{AP}_\mathrm{M}}$ between a system s and a topic t of system-topic matrix (ST) represents how much s "thinks" that t is easy (or "un-easy", i.e., difficult, with $\overline{\mathrm{AP}_\mathrm{M}} < 0$). Each link $s \leftarrow t$ with weight $\overline{\mathrm{AP}_\mathrm{A}}$ represents how much t thinks that s is effective (or "un-effective", with $\overline{\mathrm{AP}_\mathrm{A}} < 0$).

Mizzaro and Robertson [6] exploit the complete weighted bipartite graph to compute *hubness* and *authority* values by using an extended version of the HITS algorithm proposed by Kleinberg [3] which allows to include negative values for links weights. As explained by Mizzaro and Robertson [6], the authority value of a topic t of the system-topic matrix (ST) represents its easiness; when considered for a system s, it represents its effectiveness. The hubness value of a topic t represents its ability to recognize effective systems; when considered for a system s, it represents its ability to recognize easy topics.

3 HITS Hits Readersourcing

We intend to study the interactions between reader skill and paper quality where such quantities are computed by the Readersourcing model proposed by Mizzaro [5] (Section 2.1) by taking advantage of the methodology proposed by Mizzaro and Robertson [6] (Section 2.2).

The starting point is a slightly different version of the RPJ matrix shown in Figure 1 (left), which is shown in Figure 2 (left). Let us consider the judgment matrix RPJ^*. The only difference with respect to RPJ is that RPJ^* has only one additional row and column. The former is called MJ_p and its values are used to normalize each column of RPJ^* (like AAP in the original methodology), while the latter is called MJ_r and its values are used to normalize each row of RPJ^* (like MAP). The goodness matrix RPG^* is built similarly. This formalization is useful since it allows to analyze different combinations of MJ_r and MJ_p (MG_r and MG_p) with judgment (goodness) matrices.

Once the set of MJ_r and MJ_p (MG_r and MG_p) have been computed, the RPJ^* matrix shown in Figure 2 (left) and the RPG^* one are normalized in two ways. In the former normalization, each $j_{r_i,p_j}/g(j_{r_i,p_j})$ value is transformed into a $ja_{r_i,p_j}/ga(j_{r_i,p_j})$ (Normalized Judgment/Goodness according to MJ_p) by subtracting MJ_p (MG_p) to RPJ^* (RPG^*). In the latter, each $j_{r_i,p_j}/g(j_{r_i,p_j})$ value is transformed into a $jm_{r_i,p_j}/gm(j_{r_i,p_j})$ (Normalized Judgment/Goodness according to MJ_r) value by subtracting MJ_r (MG_r) to RPJ^* (RPG^*).

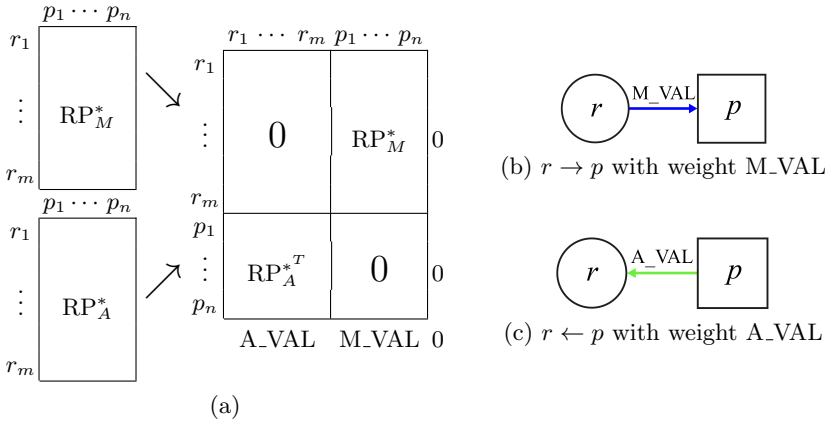
The normalized matrices $\mathrm{RPJ}^*_\mathrm{A}$ and $\mathrm{RPJ}^*_\mathrm{M}$ ($\mathrm{RPG}^*_\mathrm{A}$ and $\mathrm{RPG}^*_\mathrm{M}$) are then used to build a complete weighted bipartite reader-paper graph. Such a graph represents relationships between readers and papers which depend on the chosen set of MJ_p and MJ_r and it is used to compute hubness and authority values as done by Mizzaro and Robertson [6].

	p_1	$\cdots$	p_n	$\mathrm{MJ_r}$
r_1	j_{r_1,p_1}	$\cdots$	j_{r_1,p_n}	$\mathrm{MJ_r}(r_1)$
$\vdots$		$\ddots$		$\vdots$
r_m	j_{r_m,p_1}	$\cdots$	j_{r_m,p_n}	$\mathrm{MJ_r}(r_m)$
$\mathrm{MJ_p}$	$\mathrm{MJ_p}(p_1)$	$\cdots$	$\mathrm{MJ_p}(p_n)$	

	p_1	$\cdots$	p_n	$\mathrm{MJ_r}$
r_1	ja_{r_1,p_1}	$\cdots$	ja_{r_1,p_n}	$\mathrm{MJ_r}(r_1)$
$\vdots$		$\ddots$		$\vdots$
r_m	ja_{r_m,p_1}	$\cdots$	ja_{r_m,p_n}	$\mathrm{MJ_r}(r_m)$
	0	$\cdots$	0	

	p_1	$\cdots$	p_n	
r_1	jm_{r_1,p_1}	$\cdots$	jm_{r_1,p_n}	0
$\vdots$		$\ddots$		$\vdots$
r_m	jm_{r_m,p_1}	$\cdots$	jm_{r_m,p_n}	0
$\mathrm{MJ_p}$	$\mathrm{MJ_p}(p_1)$	$\cdots$	$\mathrm{MJ_p}(p_n)$	

Fig. 2: Reader-paper matrix (RPJ^*) with judgments, $\mathrm{MJ_r}$, and $\mathrm{MJ_p}$ (left), Judgment-paper matrix normalized according to $\mathrm{MJ_p}$ ($\mathrm{RPJ}^*_{\mathrm{A}}$) (middle), and judgment-paper matrix normalized according to $\mathrm{MJ_r}$ ($\mathrm{RPJ}^*_{\mathrm{M}}$) (right).



(a)

(b) $r \rightarrow p$ with weight M_VAL

(c) $r \leftarrow p$ with weight A_VAL

Fig. 3: (a) Construction of the adjacency matrix. RP^{*T}_A is the transpose of RP^*_A. (b-c) Relationships between readers and papers of RP^* matrix with weight M_VAL and A_VAL.

4 Experiments

In our experiments we hypothesize a scenario in which there is a publishing system; authors submit their papers to such a system and readers are able to rate the papers. We run some stochastic simulation experiments, in which readers express stochastic judgments on papers according to some predefined setting, and we measure the outcome. There are 5,000 readers, 10,000 papers, and 134,000 judgments. We simulate one month of activity. Readers are partitioned into five groups GR_i of equal size. The members of each group rate a certain amount of papers, as shown in Table 1 (left). Papers are partitioned into five groups GP_i of different size to simulate the internal state of a publishing system.

For each reader a sample of papers is picked, whose size depends on his group. Every paper is simulated by a beta distribution defined by two parameters α and β; its support is the $[0, 1]$ interval. The beta distribution probability density function can assume five shapes which are represented in Figure 4, depending on the chosen set of α and β parameters. The beta distribution allows us to represent five different distributions of user behavior across papers, thus representing five different kinds of paper. Each of the distribution shapes (shown

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

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

Table 1: Amount of rated papers for each group of readers in one month (left), and amount of papers for each group with beta distribution parameters (right).

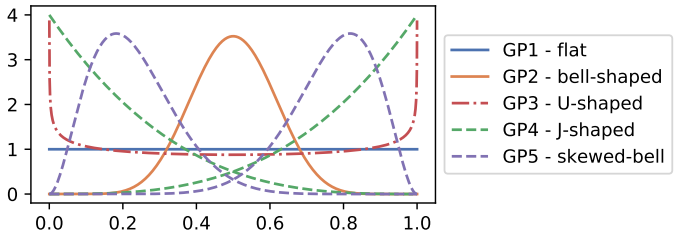


Fig. 4: Beta distributions used to generate the simulation data.

in Figure 4) originates a different simulation of the judgment agreement over the paper: the flat distribution (GP_1) simulates a completely random judging behavior; the bell shaped distribution (GP_2) simulates a judgment distribution centered around a data point in the centre of the judgment scale, simulating a case of high agreement; the U-shaped distribution (GP_3) simulates the case of maximum disagreement, where two-polarized behavior act in the opposite boundaries of the judgment scale; the J-shaped distribution (GP_4), and the skewed bell distribution (GP_5) simulate, as well as the bell-shaped distribution, the case of high agreement distributed near to the scale boundaries. The usage of the Beta distribution to capture and mimic different level of agreement, as well as the relationships between agreement and scale boundaries has been formally discussed in detail by Checco et al. [1]. The beta distributions for each paper are generated in the following way: the set of all papers is partitioned into five groups GP_i (one for each configuration) where each group contains a fixed percentage of papers. To each of these papers an instance of the beta distribution is assigned, whose α and β parameters depend on the paper group. Table 1 (right) shows such paper groups and parameters. Therefore, the stochastic judgment given to a paper by a reader is generated by sampling a value from the corresponding beta distribution.

The simulation produces a list of tuples $\langle t, r, p, a, s \rangle$: at timestamp t reader r judges paper p written by author a with a score equal to s. Such a list is provided as input data to an implementation of the Readersourcing model and from its output the final RPJ (RPG) matrix (Figure 1) is built. For each RPJ (RPG) matrix the corresponding RPJ* (RPG*) matrix is built, where each of them is a judgment (goodness) matrix characterized by a related set of $\mathrm{MJ_p}$ and $\mathrm{MJ_r}$ ($\mathrm{MG_p}$

	MJ_p	MG_p	S_p	σ_p
MJ_p		0.12	$0.94^{\dagger}$	−0.01
MG_p	0.25		$0.12^{\star}$	−0.04
S_p	$0.97^{\dagger}$	$0.25^{\star}$		−0.01
σ_p	−0.01	−0.046	−0.01	

	MJ_r	MG_r	S_r	σ_r
MJ_r		0.07	$0.06^{\times}$	−0.02
MG_r	0.11		$0.87^{\ddagger}$	0.02
S_r	$0.11^{\times}$	$0.98^{\ddagger}$		0.03
σ_r	−0.01	0.03	0.032	

Table 2: Paper (left) and reader (right). Pearson's ρ in the lower triangular part, and Kendall's τ in the upper triangular part

and MG_r) values. The RP^* matrices are then normalized to build adjacency matrices which are then used to compute hubness and authority values. In the following section we will discuss the meaning of the resulting relations (i.e., the links of the complete weighted bipartite graph) and hubness/authority values.

5 Results

We now detail the results of our experiments: Section 5.1 focuses on the measures defined in the Readersourcing model and analyzes the correlations between them; Section 5.2 discusses the outcome of HITS applied to our simulations.

5.1 Correlation Between Readersourcing Measures

The Readersourcing model produces both score and steadiness values for both readers and papers (i.e., S_r, σ_r, S_p, and σ_p). We also compute: the mean judgment received by a paper (MJ_p) the mean goodness of the judgments received by a paper (MG_p), the mean judgment expressed by a reader (MJ_r), and the mean goodness of the judgments expressed by a reader (MG_r). Table 2 shows the correlation values for the paper (left) and reader (right) scores, from which we can draw several remarks.

Let us focus on the mean judgment of a paper and the paper score (i.e., MJ_p and S_p of the left table, highlighted with †), and between the mean goodness of a reader and the reader score (i.e., MG_r and S_r of the right table, highlighted with ‡). The first correlation highlights some potential bias in how we generate the simulated data: there is lack of variance in the judgments of readers of a given paper. In other words, for each paper the vast majority of the readers that rated it present high agreement in their scores. If we look at Table 1 (right) we see that the beta distributions that induce high agreement between readers (GP_2, GP_4, and GP_5) represent the 75% of the total scores. We leave for future work the analysis of a different group distribution in the statistical simulation. The second correlation strengthens, and is a consequence of, the previous remark: there is a lack of variance in the quality of readers of a given paper; once a reader expressed a judgment on a paper, all the other readers of the same paper tend to express judgments of the similar quality.

Conversely, when looking at the dual scenario (i.e., MG_p and S_p of the left table, highlighted with *, and MJ_r and S_r of the right table, highlighted with $^{\times}$), we see that a sort of dual symmetry is present: neither mean judgment of a reader nor the mean goodness of a reader are correlated with respectively the paper and the reader scores. This suggests that: (i) the readers vote using the

whole judgment scale, and (ii) the papers receive judgments that span across all the judgment scale. This shows that the current simulation setting is able to cover all the judgment scale.

The correlations between all other measures are low and not interesting.

5.2 HITS Algorithm and Hubness

In this section we detail the results of the HITS algorithm when run on the normalized RP* matrices, both when considering the judgments (i.e., RPJ*) and the goodness (i.e., RPG*). As detailed in previous work [6, 7], the most interesting index we obtain from running HITS is the hubness of readers and papers; when we consider the judgment matrix, the hubness of a reader measures its capability to recognize papers that tend to obtain high judgments, while the hubness of a paper measures the paper capability to recognize reader that tend to give high judgments (or, in other words, readers that are biased towards giving high judgments). Symmetrically, when we consider the goodness matrix the hubness of a paper measures its capability to recognize readers that tend to give judgments that have a high quality (or, in other words, high quality readers), while the hubness of a reader measures the reader capability to recognize papers that tend to receive judgments of a high quality (i.e., papers that tend to be judged from high quality readers).

We start with the judgments, i.e., the RPJ* matrix. Figure 5 shows some scatterplots. All the y-axes report the hubness computed by HITS. In the plots on the left column, the x-axes report the model measures that refer to a paper, while in the plots on the right column, the x-axes report the model measures that refer to a reader; thus, in the scatterplots of the left each point is a paper, while in the scatterplots of the right each point is a reader. Each scatterplot also shows the respective Pearson's ρ and Kendall's τ correlations. The meaning of the correlations of each plot in the figure can be detailed as follows.

(a) The higher the score of a paper, the higher its capability of recognizing readers that tend to give high judgments. This correlation is expected to be high due to how the Readersourcing model is formalized; intuitively, if a score of a paper is high then the paper will be good in recognizing readers that tend to give high judgments.
(b) Since the correlation is really low, and close to zero, whatever the score of a reader (high/low, i.e, high-/low-quality reader), he has the same capability to recognize papers that tend to obtain high (and low) judgments. This is a good property of the Readersourcing model: a reader can be of a high (or low) quality independently of whether he expressed judgments on papers that have an average judgments that is either high or low. In other words, if a reader expresses a high quality judgment on a paper, his score as a reader will increase no matter what the judgment score is. Ideally, for a model to be completely fair, this correlation value should be exactly zero.
(c) The higher the mean judgment of a paper, the higher its capability of recognizing readers that tend to give high judgments. Also in this case, as for Figure 5(a), the high correlation value is expected and less interesting. Nevertheless, since the correlation value it exactly one, it can be also interpreted

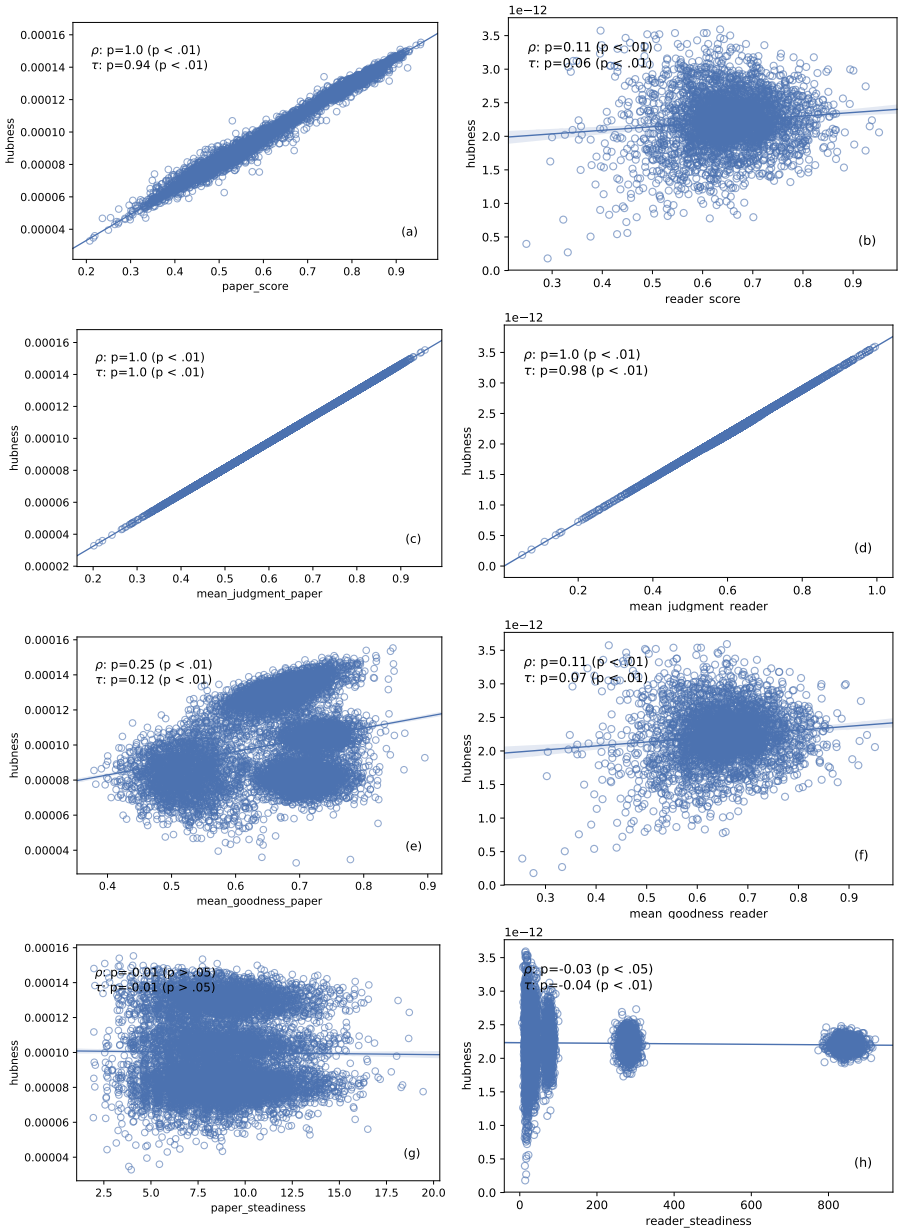


Fig. 5: Hubness vs. S_p, MJ_p, MG_p, and σ_p (left column) and vs. S_r, MJ_r MG_r, and σ_r (right column), computed on RPJ^* matrix

as a bias in how we generate the data: in fact, this plot shows that the variance of the judgments expressed by readers on each paper is on average very low, despite the beta distributions we use to generate the data. This is confirmed also by analyzing the next plot.

(d) The higher the mean judgment by a reader, the higher his ability to recognize papers that get high scores. As for previous plot, also in this case the correlation value is exactly one. While the high correlation of the previous plot is expected, this one is not. On the contrary, if the mean judgment by a reader is high, then it should not necessarily mean that the papers that s/he judged should get on average high scores, since the other readers judging the same papers could give lower judgments. This is an indication of a possible bias in how we generate the data. We leave for future work to use more sophisticated statistical methods to generate the data, such as for example vine copulas, that would allow to consider both the paper and reader distributions at the same time.

(e) Since the correlation is really low, whatever the goodness of the judgments received by a paper (i.e., high or low mean goodness), it has the same capability to recognize readers that tend to give high judgments. This is a good property of the Readersourcing model: a paper can be either good or bad (i.e., have a high or low mean goodness) independently from having been judged by readers biased towards high or low scores. In other words, the model formalization of the goodness measure of a paper is robust to the possible reader bias on the judgment scale.

(f) Since the correlation is really low, whatever the mean goodness of the judgments expressed by a reader (i.e., high or low mean goodness), he has the same capability to recognize papers that tend to get high scores. As for the previous plot, also this correlation is an indication of a good model property: a reader can be either good or bad (i.e., have a high or low mean goodness) independently from the fact that he has judged papers that get on average high or low scores. In other words, the model formalization of the goodness measure of a reader is robust to the possible behavior of other readers that express judgments on the same paper.

(g) Since the correlation is zero, whatever the steadiness of a paper, it has the same capability to recognize readers that tend to give high judgments. This reflects a good property of the model: the formalization of the steadiness measure of a paper is robust to the fact that the paper will get scores in the upper or lower part of the judgment scale.

(h) As for the previous plot, since the correlation is zero, whatever the reader steadiness, he has the same capability to recognize papers that tend to get high scores. Symmetrically from what derived from the previous plot, this hints that the steadiness measure of a reader is robust to the judgment behavior of the other readers that express judgments on the same paper.

We now turn to discuss Figure 6 which shows the same plots as in Figure 5 but when running the HITS algorithm on the goodness matrix RPG^*.

(a) Due to the low correlation, whether a paper has a high or low score, it has the same capability to recognize readers that tend to express high quality

judgments (judgments with high goodness). This highlights a good property of the Readersourcing model: the ability of a paper of recognize good readers is independent from the quality of the paper itself. In an ideal model, the correlation value of this plot should be zero.

(b) The higher the score of a reader, the higher its capability to recognize papers that tend to get high quality judgments. This high correlation highlights a possible bias in how we generate the simulations: in fact, if a high quality reader judges a paper, all the other readers that judge the same paper will tend to be of high quality. As for Figure 5(d), we leave for future work the use of more sophisticated models for the statistical generation of judgments.

(c) This plot is the same as Figure 6(a); this has a double meaning: the paper score and mean judgment of a paper are almost perfectly correlated (see the 0.97 value in Table 2) and, as for Figure 6(a), the ability of a paper of recognize good readers is independent from the mean judgment of the paper.

(d) Due to the very low correlation, 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. As for the previous plot, this highlights a good property of the model: the ability of a reader to recognize papers with high quality judgments is independent from the judgment location of the reader (i.e., it is independent from the judgment scale).

(e) The higher the mean goodness of the judgments received by a paper, the higher its capability of recognizing readers that tend to express high quality judgments. In this case the correlation is exactly one, and this is expected and derived from how the Readersourcing model is defined.

(f) The higher the mean goodness of the judgments expressed by a reader, the higher his capability to recognize papers that tend to get judgments having high quality. This, as the previous plot, is a natural consequence of how the Readersourcing model is defined.

(g) Due to the low correlation, whether a paper has a high or low steadiness, it has the same capability of recognizing readers that tend to express high quality judgments.

(h) Due to the correlation close to zero, whether a reader has a high or low steadiness, he has the same capability of recognizing papers that tend to get high quality judgments. Also in this case this highlights a good property of the model: the ability of a reader to recognize papers that receives high quality judgments is independent from its steadiness value.

6 Conclusions and Future Work

We have provided a two-fold contribution: (i) we proposed an experimental validation of the Readersourcing model carried out through a stochastic simulation, and (ii) we explored model properties using network analysis techniques.

This paper leaves plenty of space for future work like, for example, the usage of other stochastic models, and the analysis of other models that propose alternatives to peer review [2].

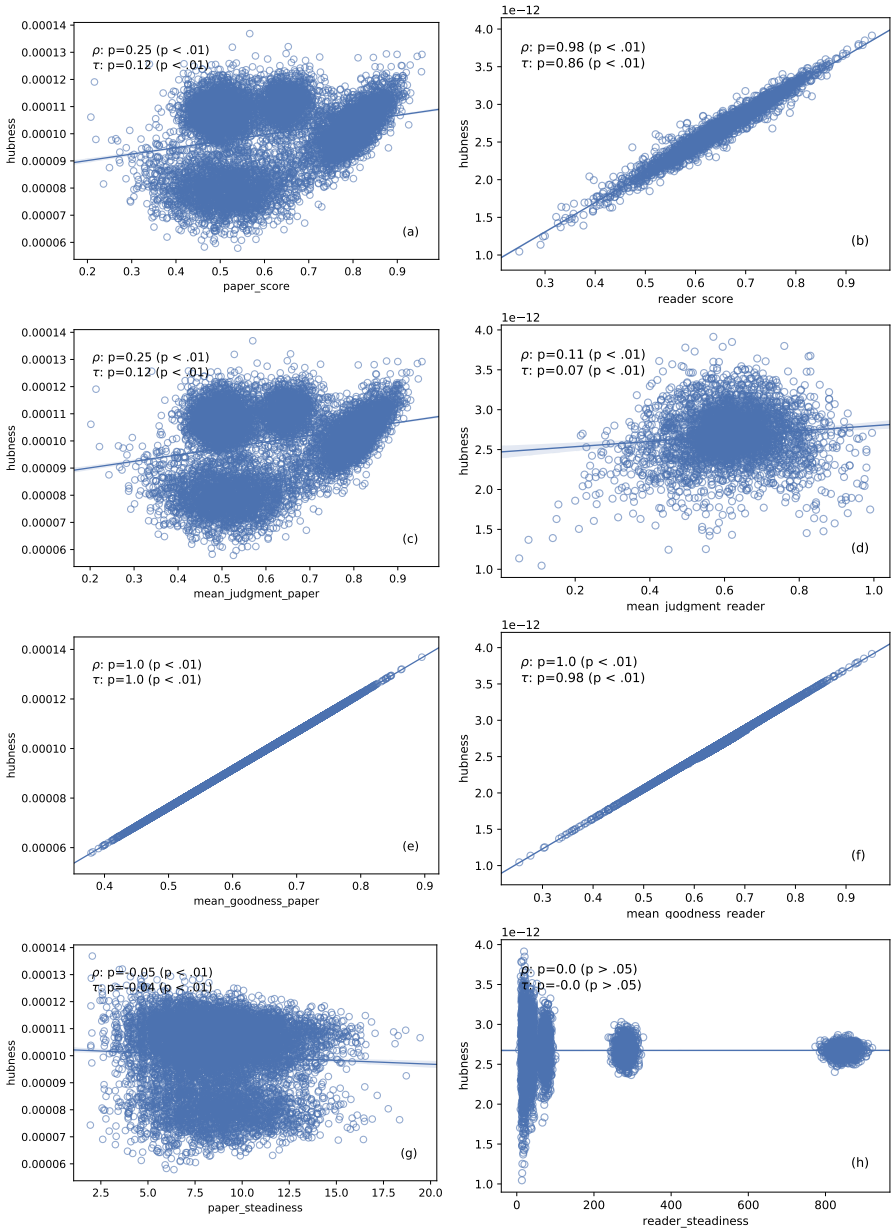


Fig. 6: Hubness vs. S_p, MJ_r, MG_r, and σ_p (left column) and vs. S_r, MJ_r MG_r, and σ_r (right column), computed on RPG^* matrix

References

[1] Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., Demartini, G.: Let's agree to disagree: Fixing agreement measures for crowdsourcing. In: 5th HCOMP (2017)

[2] De Alfaro, L., Faella, M.: TrueReview: A Platform for Post-Publication Peer Review. CoRR (2016), http://arxiv.org/abs/1608.07878

[3] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (Sep 1999), http://doi.acm.org/10.1145/324133.324140

[4] Mizzaro, S.: Quality control in scholarly publishing: A new proposal. JASIST 54(11), 989–1005 (2003), https://doi.org/10.1002/asi.22668

[5] Mizzaro, S.: Readersourcing - A Manifesto. JASIST 63(8), 1666–1672 (2012), https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22668

[6] Mizzaro, S., Robertson, S.: HITS Hits TREC: Exploring IR Evaluation Results with Network Analysis. In: Proceedings of 30th ACM SIGIR. pp. 479–486 (2007)

[7] Roitero, K., Maddalena, E., Mizzaro, S.: Do easy topics predict effectiveness better than difficult topics? In: ECIR. pp. 605–611. Springer (2017)

[8] Soprano, M., Mizzaro, S.: Crowdsourcing peer review: As we may do. In: Digital Libraries: Supporting Open Science. pp. 259–273. Springer International Publishing (2019)