

Jun 3rd, 12:00 AM

Assessing the Impact of Publications Saved by Mendeley Users: Is There Any Different Pattern Among Users?

Zohreh Zahedi
Leiden University, z.zahedi.2@cwts.leidenuniv.nl

Rodrigo Costas
Leiden University, rcostas@cwts.leidenuniv.nl

Paul Wouters
Leiden University, p.f.wouters@cwts.leidenuniv.nl

Zohreh Zahedi, Rodrigo Costas, and Paul Wouters, "Assessing the Impact of Publications Saved by Mendeley Users: Is There Any Different Pattern Among Users?." *Proceedings of the IATUL Conferences*. Paper 4.
<https://docs.lib.purdue.edu/iatul/2014/altmetrics/4>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Assessing the impact of publications read (saved) by the different Mendeley users: is there any different pattern among users?

Zohreh Zahedi, Rodrigo Costas & Paul Wouters

Centre for Science and Technology Studies (CWTS), Faculty of Social and Behavioral Sciences, Leiden, University. PO Box 905, 2300 AX, Leiden (The Netherlands)
{ z.zahedi.2, rcostas, p.f.wouters}@cwts.leidenuniv.nl

Abstract:

The main focus of this paper is to investigate the impact of publications read (saved) by the different users in Mendeley in order to explore the extent to which their readership counts correlate with their citation indicators. The potential of filtering highly cited papers by Mendeley readerships and its different users have been also explored. For the analysis of the users, we have considered the information of the top three Mendeley 'users' reported by the Mendeley. Our results show that publications with Mendeley readerships tend to have higher citation and journal citation scores than publications without readerships. 'Biomedical & health sciences' and 'Mathematics and computer science' are the fields with respectively the most and the least readership activity in Mendeley. PhD students have the highest density of readerships per publication and Lecturers and Librarians have the lowest across all the different fields. Our precision-recall analysis indicates that in general, for publications with at least one reader in Mendeley, the capacity of readerships of filtering highly cited publications is better than (or at least as good as) Journal Citation Scores. We discuss the important limitation of Mendeley of only reporting the top three readers and not all of them in the potential development of indicators based on Mendeley and its users.

Keywords:

Mendeley; Career Stages; Altmetrics; Readership Impact; Citation Impact; Mendeley users

Introduction & Background:

Mendeley is a global reference management tool and a rich source of readerships. Mendeley collects a wide variety of different metadata per publication such as total number of readers, tags assigned by the users, statistics for discipline, country and academic status which are saved by the different types of users when creating their own libraries. Among these metadata, the 'career stage' of the users is provided. The previous analysis of the 'career stages' of the different Mendeley users (Zahedi, Costas & Wouters, 2013) showed that PhD students are the most common types of users per publication in Mendeley (similar results have been confirmed recently by Mohammadi et al., in press). Also, in terms of highly cited publications, apart from the unknown (i.e. non unidentified) users, Post Docs and PhD students tend to read papers with higher impact than other users in Mendeley. Other altmetrics studies on Mendeley investigated the relation between readership and citation counts in general (Li & Thelwall, 2012; Haustein et al., 2013a; Zahedi, Costas & Wouters, 2014) and between authors, departments, institutions and countries and readership and citation counts for different sets of publications (Sud & Thelwall, in press; Thelwall & Maflahi, in press) reporting different (moderate and weak) correlations. However, little is known about the readership activity of different types of users in Mendeley and the different patterns among them.

This study contributes to the idea of exploring the potential of Mendeley readerships as a new source of impact assessment compared to citation analysis. The focus is to explore the 'career stages' of Mendeley users to see their potential in detecting highly cited publications and to test which indicators (e.g. readerships scores compared to journal citation scores) can help better to

identify highly cited publications. In particular, we are addressing the following objectives and research questions:

- **General distribution of Mendeley readerships over publications across fields and by different career stages**
What is the distribution of Mendeley readerships across fields and by different career stages? Are there any differences among the different users and across fields?
- **Relationship (correlation) of Mendeley readerships with citations indicators**
To what extent do the readerships of the publications saved by the different users in Mendeley correlate with their citation indicators? What are the differences in correlation by career stages?
- **Potential of identifying highly cited papers across all fields of science and by different career stages in Mendeley**
To what extent can highly cited papers be identified by the different types of users in Mendeley ?

Data & Methodology:

For this study, we have collected a random sample of 100,000 publications from the Web of Science (WOS) database that were published in 2011, and for which a DOI was available¹. DOIs were used as the basis to extract readership from Mendeley by using the Mendeley REST API in May 2013. The data from Mendeley were matched back with CWTS in-house WoS database in order to calculate other bibliometric indicators². A variable citation window (until the end of 2012) was used and self-citations have been excluded. Taking into account only the publications which are classified as article or review, 88,905 publications were used for the final analysis, out of which, 39,804 (45.5%) publications received in total 295,964 Mendeley readerships and 96,975 WOS citations while 49,101 (55.5%) publications didn't get any readership in Mendeley. According to table 1, publications with readerships tend to have higher citation and journal citation scores than those without readerships. Also comparing citations and readerships per paper, on average, each publications received more readerships (3.32) than citations (2.17).

	pubs	%	tcs	%	mncs	js	mnjs	tcs /pub	total readerships	Read /pub
Pub with alt	39,804	45.5%	96,975	50.1%	2.44	2.33	1.20	1.09	295,964	3.32
Pub without alt	49,101	55.5%	96,507	49.9%	1.97	2.03	1.00	1.08		
Total	88,905	100%	193,482	100%	2.18	2.16	1.09	2.17		

Table 1. General distribution of publications with and without Mendeley readerships

In addition to the number of readerships, Mendeley also provides the percentage of users per category of readers (i.e. PhD students, Professors, PostDoc researchers, Students, Librarians, Lecturers, Other Professionals and Academic and non-Academic researchers). A strong limitation is that this information is limited to the top 3 most frequent users per publication

¹ We acknowledge the problem of not all publications having DOI information, but given the exploratory nature and the large amount of publications studied in this paper, we consider that this limitation does not pose a strong limitation for our results.

² Citation indicators (cf. Waltman et. al. 2011): Citation Score (CS): number of citations per publications; Normalized Citation Score (NCS): number of field-normalized citations per publications; Journal Citation Score (JCS): average number of citations per paper in the journal of publication; and Normalized Journal Score (NJS): field-normalized average number of citations received in the same journal; PP(top 1%) (proportion of top 1% publications):The proportion of the publications belong to the top 1% most frequently cited compared with other publications in the same field and in the same year.

(Gunn, 2013), therefore we don't know the other users beyond the top 3 (hereafter referred to as 'unknown users'). For every publication we have calculated the actual number of the different types of users and we also determined the total 'known' and 'unknown' users per publication. Table 2 shows that in terms of publications saved and readerships received by each user types, the most common types of users in Mendeley are PhD and Students (proportionally 34% and 17% of known users) and the least ones are lecturers (1%) and Librarians (.4%). Also, publications saved by PhDs have a higher readership density than citation density, while this is the contrary for the other known users in Mendeley (they received more citations than readerships).

Career Stages in Mendeley	No pubs read	% pubs Read	Readerships	% read	readership /total no pub	tcs	Citation/ total no pub
PhD	29,256	33%	101,772	34%	1.14	800,43	0.9
Unknown	18,173	20%	75,975	26%	0.8	664,10	0.7
Students	21,458	24%	49,182	17%	0.5	57,897	0.7
PostDoc	11,779	13%	27,419	9%	0.3	41,551	0.5
Researchers	10,347	12%	18,966	6%	0.2	27,045	0.3
Professors	7839	9%	12,080	4%	0.14	18,247	0.2
Other Professions	4519	5%	6955	2%	0.08	11,415	0.1
Lecturers	2288	3%	2572	1%	0.03	484	0.1
Librarians	879	1%	1043	0.4%	0.01	1511	0.0

Table 2. Distribution of Mendeley readerships by the different career stages

Findings & Results:

Distribution of Mendeley readerships across 5 major fields of science³

For all fields, publications with Mendeley readerships have higher citation scores than publications without readerships. Table 3 shows that Biomedical & health sciences (13.6%) have the highest share of publications with readerships while Mathematics and computer science (4.2%) have the lowest share. In terms of readerships per publication (i.e. RPP scores) the Life & earth science have the highest values (9.5) followed by the Social science & humanities (8.9). In contrast, Mathematics and computer science (6.2) and Natural sciences & engineering (6.0) exhibit the lowest readerships density. This is in line with both the Mendeley global report⁴ and our previous study which showed that the second best covered publications in Mendeley are publications from the Medical and life sciences (Zahedi, Costas & Wouters, 2014). Also, on average, all fields show higher RPP scores than CPP scores. This is probably caused by the fact that these are recent publications (from 2011) and still need some time to get their optimum levels of citations, while in terms of social media, the uptake tends to be much faster (Haustein et al, 2013b), although the proper analysis of 'readerships windows' in Mendeley is a topic that still needs further research.

	Main fields	Life & earth sc	Biomed & health sc	Natural sc & eng	Social sc & hum	Mat & com sc	Total
Pubs without Mendeley	no	7,836	22,416	16,772	4,275	5,175	56,474
	%	7.6%	21.60%	16.20%	4.1%	55%	54.50%
	tcs	16,387	59,671	30,275	3,807	3,467	11,3607
	tcs/pub	2.1	2.7	1.8	0.9	0.7	2.0
Pubs with	no	9,830	14,145	13,952	4,929	4,382	47,238
	%	9.5%	13.60%	13.50%	4.85	4.2%	45.5%

3 <http://www.leidenranking.com/ranking/2013>

4 http://www.mendeley.com/global-research-report/#.Ux3Tw_k2xv8

Mendeley	tcs	21,94	45,403	37,532	5,641	4,622	115,138
	tcs/pub	2.2	3.2	2.7	1.1	1.1	2.4
	tot read	93,165	119931	83093	43679	27099	366967
	%	25%	33%	23%	12%	7%	100%
	read/pub	9.5	8.5	6.0	8.9	6.2	7.8
Total pubs	no	17,666	36,561	30,724	9,204	9,557	103,712
	%	17%	35.30%	29.60%	8.90%	9.20%	100%
	tcs	38,327	105,074	67,807	9,448	8,089	228,745
	tcs/pub	2.2	2.9	2.2	1.0	0.8	2.2

Table 3. Distribution of Mendeley readerships across major fields of science

Distribution of Mendeley readerships by the different users across Leiden Ranking (LR) fields

Figure 1 shows the proportion of readerships by users across LR fields. Apart from the 'unknown' users, PhD and Students are the most common types of users while Lecturers and Librarian are the least ones across all LR fields.

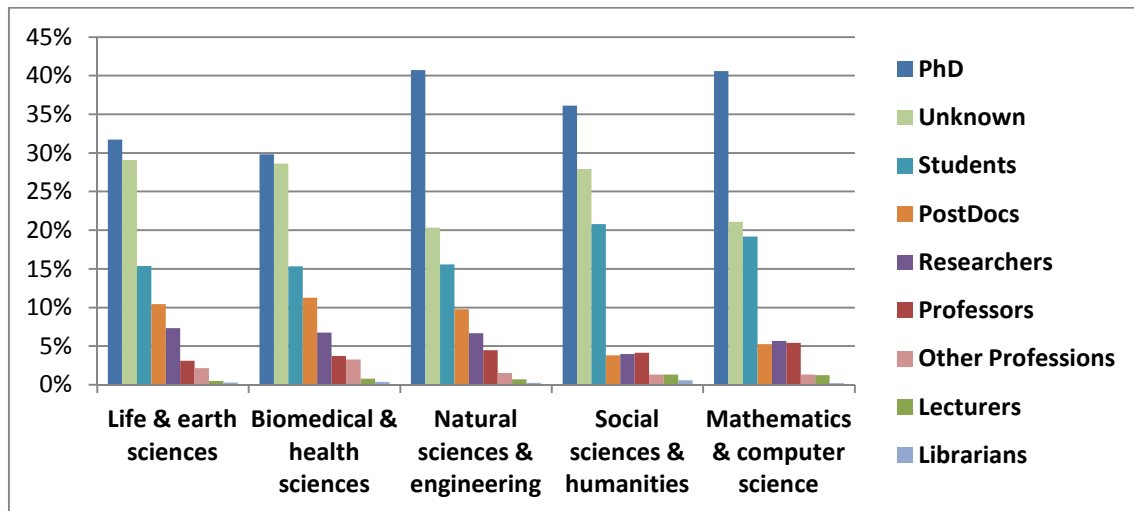


Figure 1. Distribution of Mendeley readerships by the different career stages across LR fields

Relationship between Mendeley readerships and bibliometrics indicators

The focus here is to explore to what extent the readerships for the publications saved by the different users in Mendeley relate to their citation and journal indicators. The results of factor analysis, correlation analysis (including correlations among readerships and bibliometric indicators by the different users) and the precision-recall analysis will be presented in the next sections.

Factor Analysis of Mendeley readerships and bibliometrics indicators

An exploratory factor analysis has been performed using SPSS version 21 in order to know more about the underlying structure and relationship among the variables. The factor analysis was performed between the readerships received by the publications saved by the different types of users in Mendeley and their bibliometrics indicators. The principal Component Analysis (PCA) revealed the presence of 4 main components (dimensions) with eigenvalues exceeding 1, explaining 61% of the total variance.

Table 4 shows that the first dimension is dominated by the bibliometric indicators in which both direct and normalized citation and journal indicators are loaded together. In the second dimension we find with the highest loadings Students, the Unknown users, the Total readership (and to a lower extent also PhD students), thus this is a dimension characterized by the most common users in Mendeley that correlate the most with absolute readerships values. The third dimension is related to the more academic/scientific users in Mendeley. The fourth dimension is

particularly related to other professional/educational users (Librarians, Other professionals or Lecturers).

	Component			
	1	2	3	4
CS	.849	.251		
NCS	.802	.272	-.114	
JCS	.780	-.112	.307	
MNJS	.771	-.117	.282	
Students		.924		
Unknown	.118	.857	.416	
Total readers	.127	.805	.572	
PhD	.133	.413	.737	
PostDocs	.137	.212	.736	-.137
Researchers			.550	.112
Professors			.317	
Other Professional			.117	.647
Librarian				.637
Lecturers				.423

Table 4. Factor analysis of types of users and bibliometric indicators

Correlation analysis of Mendeley readerships and bibliometrics indicators

Table 5 Presents the correlation between readerships of the publications saved by the different types of users in Mendeley, their citation and journal citation scores. It can be seen that correlations of the total account of readership are slightly better with journal indicators (JCS and NJCS) than with article level impact indicators (CS and NCS). Although, the overall correlation scores among total readerships and bibliometrics indicators are rather low (varies between .172 and .218), regarding the different “career stages”, apart from Unknown (which shows the highest correlations with citation indicators!), it is higher for readerships from PhDs, followed by PostDocs and Students; other types of users have very low correlation. Different patterns also observed in terms of correlations among the readerships and the other different career stages which may suggest different readership interests (e.g. PhD have higher correlations with students than with PostDocs).

Spearman Correlation	cs	ncs	mjcs	mnjs	readers	prof	Lect	PhD	Post Doc	stu	res	Lib	Other Prof	Un known
cs	1	.953	.558	.421	.172	.034	.007	.172	.146	.116	.069	.002	.043	.232
ncs		1	.434	.436	.183	.042	.019	.184	.124	.127	.070	.002	.040	.227
mjcs			1	.750	.177	.045	.002	.162	.191	.102	.085	.000	.052	.220
mnjs				1	.218	.073	.030	.208	.153	.140	.089	.011	.040	.228
readers					1	.322	.182	.842	.496	.695	.420	.107	.266	.758
professors						1	.011	.182	.060	.077	.042	.002	.024	.122
Lecturer							1	.093	.026	.078	.011	.021	.025	.154
PhD								1	.367	.509	.271	.033	.111	.645
PostDoc									1	.209	.124	.005	.055	.436
student										1	.131	.048	.104	.596
researchers											1	.017	.067	.292
Librarian												1	.047	.089
Other Professional													1	.229

Table 5. Correlation analysis of the rank values of citation and altmetrics variables (correlations > .150 have been highlighted)

Precision-recall analysis

In this section, we used the same precision-recall⁵ approach developed by Waltman & Costas (2014). The focus here is to explore the potential use of Mendeley users for filtering highly cited publications compared to journal citation scores. For this purpose, the top 1% of highly cited publications in the sample (based on the NCS indicator) have been detected. The precision-recall analysis has been performed for all publications in the sample ('extended approach') and publications with at least one readership in Mendeley ('tight approach'). We have explored the 5 LR fields and the different Mendeley users.

Figure 2 shows the general precision-recall analysis of total readership scores and Journal Citation Scores (JCS) both for extended vs tight approaches. This Figure shows that JCS performs overall better than readerships in identifying the top 1% most cited publications within our dataset, whereas, for the tight approach readerships slightly outperform the JCS score.

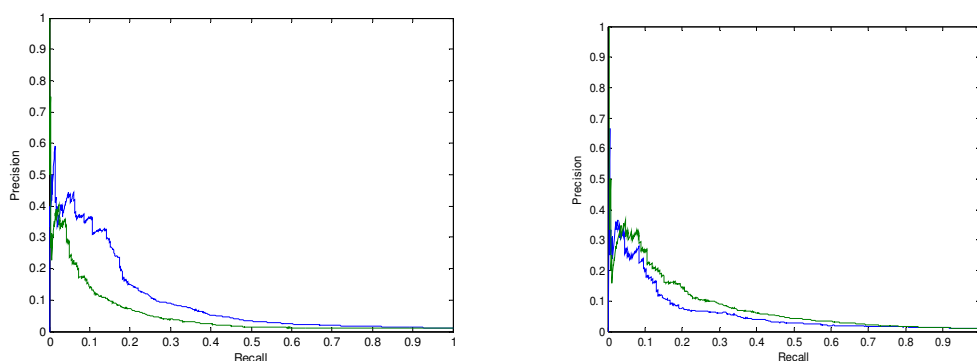


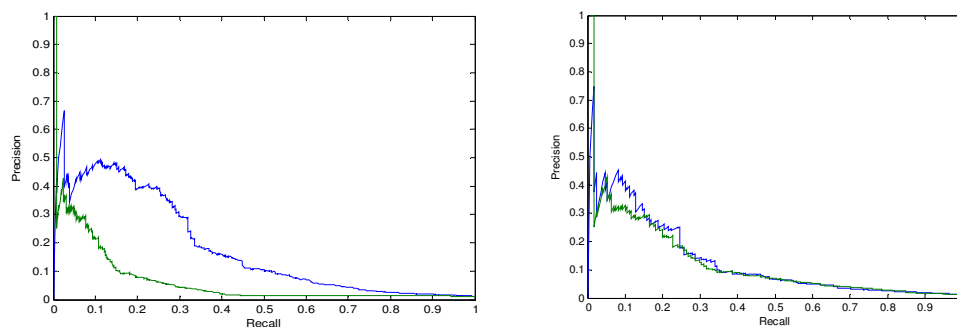
Figure 2. General Precision-recall curves for JCS (blue line) and total readerships (green line) for identifying 1% most highly cited publications (extended approach on the left and tight approach on the right)

Precision-recall analysis of the different fields of science

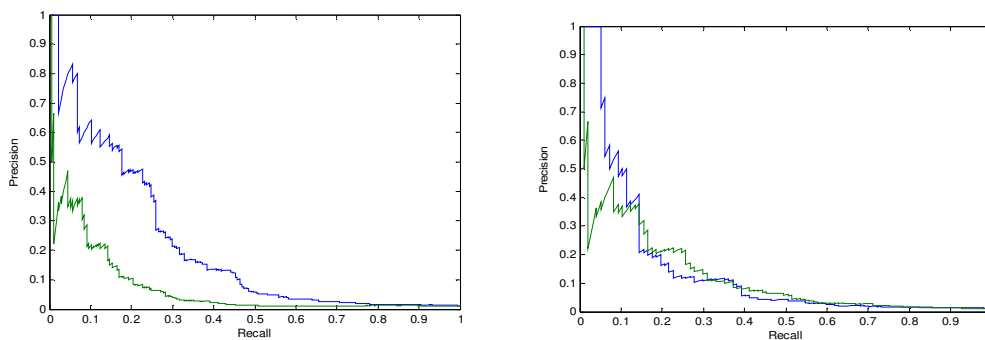
For this analysis, the top 1% most highly cited publications have been calculated individually for the 5 LR fields of science in LR. The results of the precision-recall analysis for all fields of science (extended approach including all publications in the sample), shows that JCS outperforms readership scores in filtering highly cited publications. However, an advantage of readership vs JCS within the lowest level of recall for fields such as Mathematics & computer science, Natural sciences & engineering and Social sciences & humanities has been observed (Figure 3). The same analysis for the publications with at least one reader in Mendeley (tight approach) shows no real improvement of readerships over JCS for Biomedical & health science, Life and earth science and Mathematics & computer science; for the Social sciences & humanities the two lines overlaps. Hence, for all of the fields JCS indicator have the same or slightly better filtering capabilities than the readerships. The only exception is the Natural sciences & engineering, where readership scores outperform JCS scores. This means that for this field the capacity of readership in detecting highly cited publications is higher than JCS.

⁵ Precision is defined as the number of highly cited publications in the selection divided by the total number of publications in the selection. Recall is defined as the number of highly cited publications in the selection divided by the total number of highly cited publications.

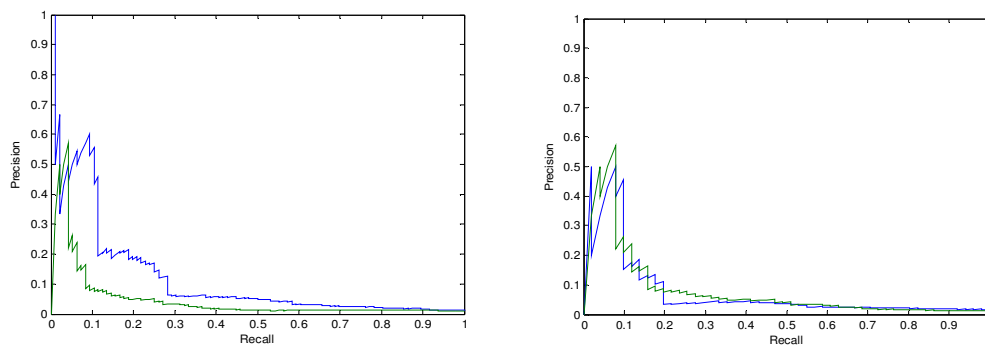
Biomedical and health sciences



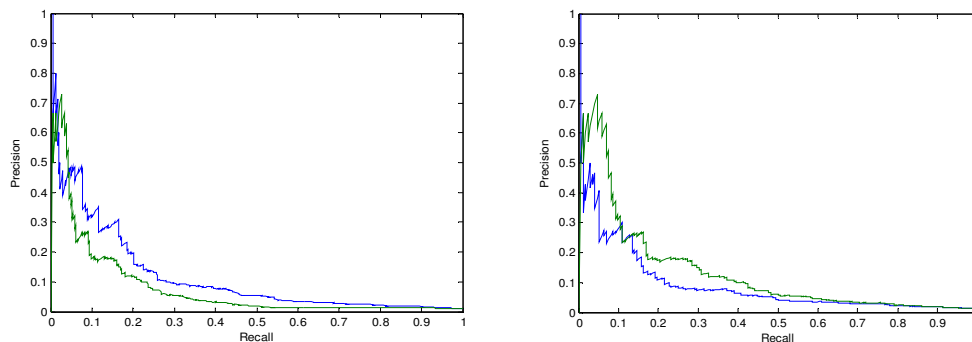
Life and earth sciences



Mathematics and computer science



Natural Sciences and engineering



Social sciences and humanities

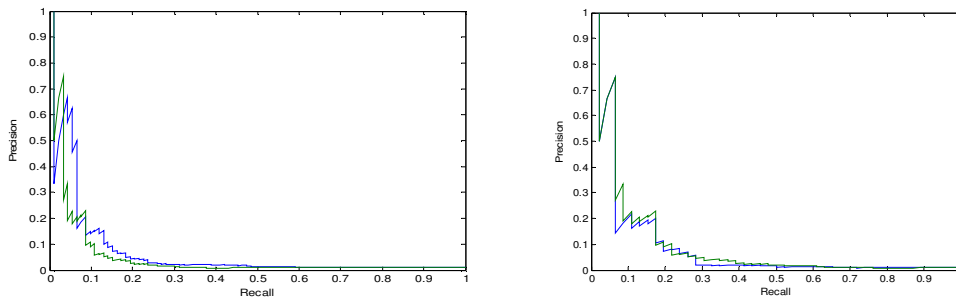


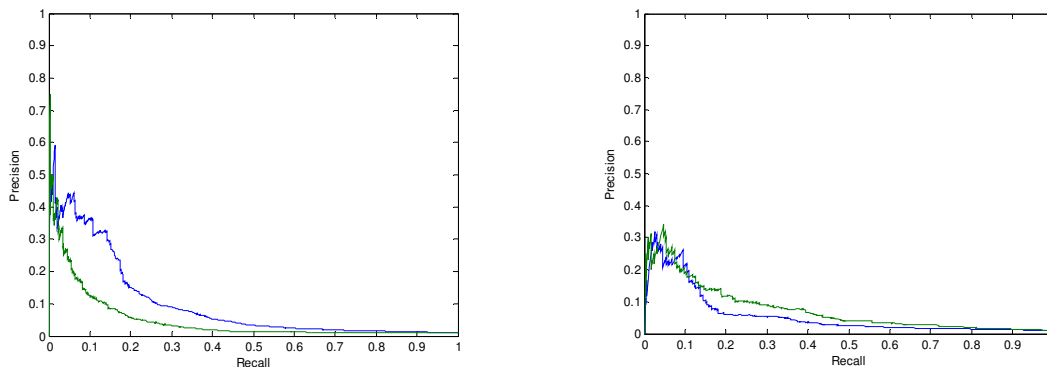
Figure 3. Precision-recall curves for JCS (blue line) and LR Fields (green line) for identifying 1% most highly cited publications (extended approach on the left and tight approach on the right)

Precision-recall analysis of types of different career stages in Mendeley

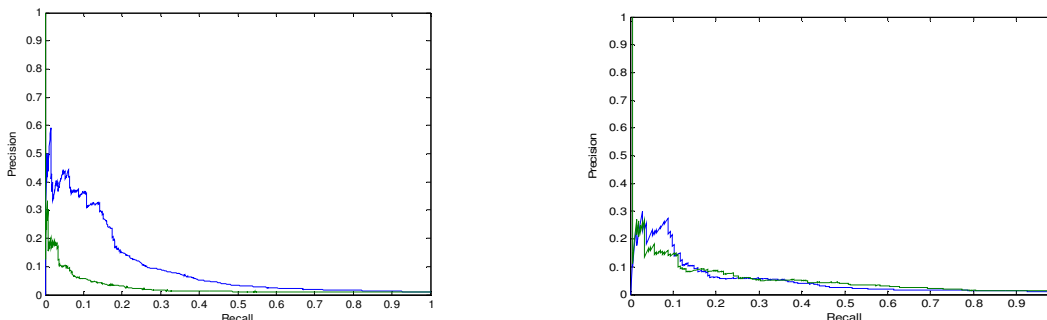
The same approach has been done based on different known Mendeley users. Here, the top 1% most highly cited publications is calculated individually for every type of user in Mendeley. Figure 4 shows the results of the precision-recall analysis of readerships scores by the different types of users in Mendeley and Journal Citation Score (JCS) both for extended and tight approaches. For the extended one, again, JCS performs better than readerships for all types of users in identifying the top 1% most highly cited publications within our dataset.

For the tight analysis, according to the figure for the PhD students, readerships slightly outperforms the JCS (particularly for the higher levels of recall, thus resembling the general pattern in Figure 2). In other words, there is a slight advantage of PhD readerships over JCS in identifying the top 1% most highly cited publications for the tight approach. The figures for the other users show doesn't show any distinguished pattern or any advantage of these indicators over each other. The curves for both JCS and readerships are similar to each other or there is an occasional advantage for JCS over readerships. The figures for Lecturers and librarians do not show any clear pattern maybe due to the low presence of these types of users per publications in this sample. In general, most of the readerships by different types of users do not show any improvement of the general pattern except for PhD, with the strongest similarity with the general pattern.

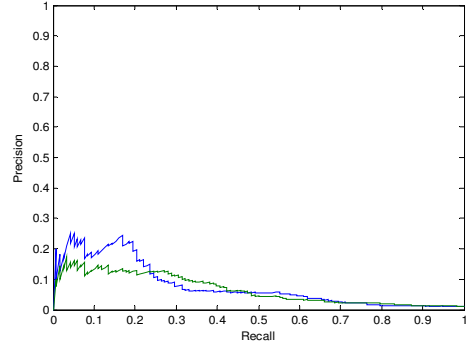
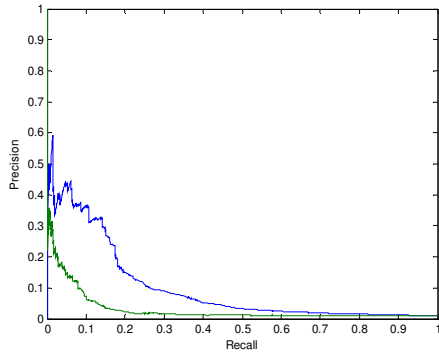
PhD



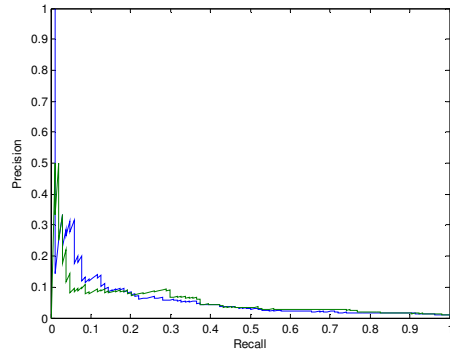
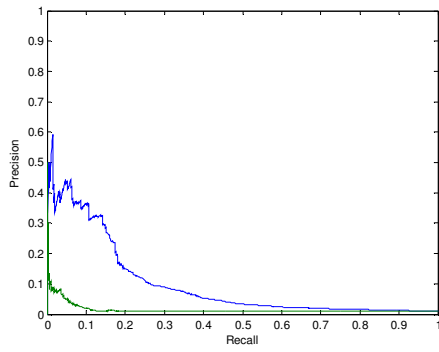
Students



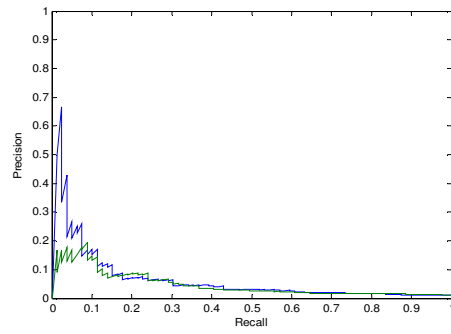
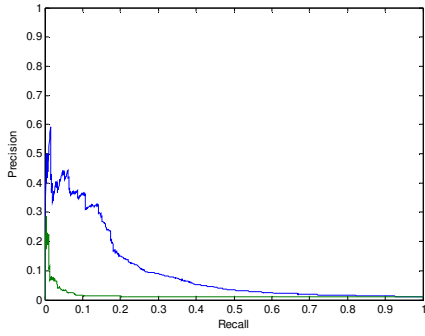
Post Docs



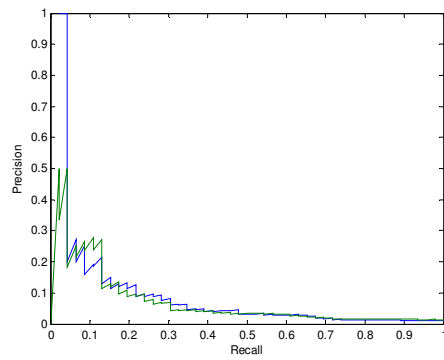
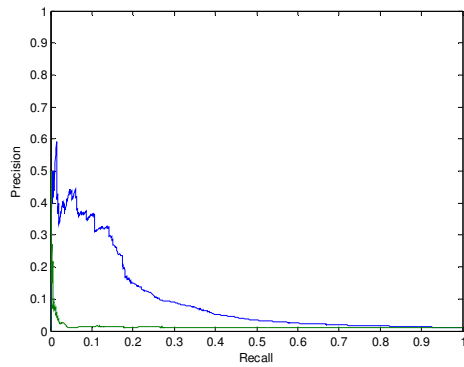
Researchers



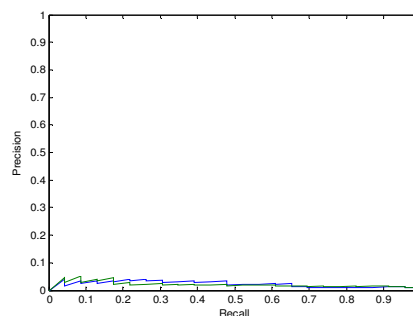
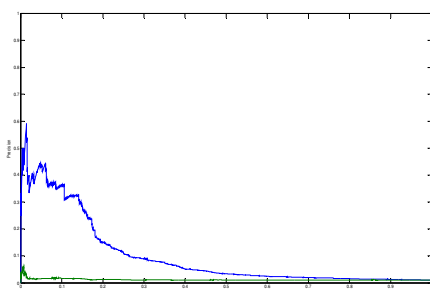
Professors



Other Professionals



Lecturers



Librarians

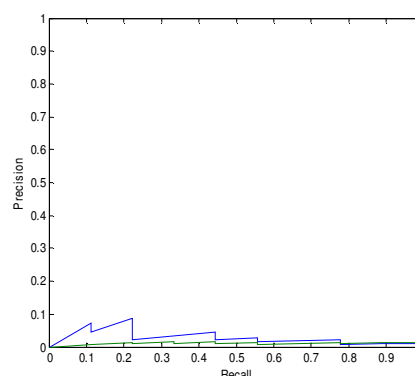
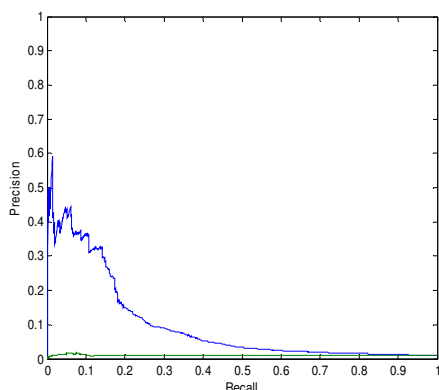


Figure 3. Precision-recall curves for JCS (blue line) and types of users readerships (green line) for identifying 1% most highly cited publications (extend approach on the left and tight approach on the right)

Conclusion & Discussions:

Mendeley is a global data source of readerships data for scholarly outputs; it collects a wide variety of metadata per publication saved by different users. Preliminary studies (Zahedi, Costas, & Wouters, 2013; Mohammadi, et al., in press) have suggested that the information on the 'User categories' or 'Career Stage' of users in Mendeley could represent a valuable source to inform different types of impact (e.g. scientific, educational, etc.). However, as suggested before (Zahedi, Costas, & Wouters, 2013), the fact that Mendeley only reports the three most frequent users of the publications introduces an important limitation and potential bias for this type of analysis. For this reason, for the time being, this study must be considered as an exploration of these relationships only for the 'known' users in Mendeley and certain caution is necessary as the potential uncertainty introduced by the 'unknown' users limits the strong interpretation of our results (as well as it hinders the introduction of Mendeley users as a reliable source of impact information). For example, the finding that citations correlate the best with the 'unknown' typology reinforces the idea that the disclosure of this typology is absolutely relevant for a much better and transparent understanding of the different users in Mendeley. In addition, more information on how Mendeley users are defined is very important, as well as on how the typologies are chosen and updated by the users. For example, the relatively strong correlation between PhDs and Students could suggest that (some) students that become PhD do not update their profiles and therefore they 'read' like PhD students but without updating their 'Academic status' in Mendeley.

Given all the previous limitations, the current study has explored the idea of analyzing and comparing the readership and citation impact of the scholarly publications used by the different Mendeley users and across different fields. Our results show that in general, the publications with Mendeley readerships received higher readership impact vs. citation impact per publication than those without readerships. Also, in terms of readership density across the 5 major LR fields, on average, all fields show higher RPP scores than CPP scores and some disciplinary differences among fields have been observed. These lower values of citations compared to readership can be explained by the relatively short citation window considered in this study

(only 1 year after the publication year) and also explain the lower correlation values between the two indicators. This also suggests a faster reception of Mendeley readerships as compared to citations and encourages the need to study the temporality and pace of readership counts. Regarding the career stages, the most common types of users in Mendeley are PhDs and Students (besides the 'unknown' users), and similar proportions are observed for all the LR fields. The correlation analysis shows relatively low relationships among the users with different career stages, thus introducing the idea that potentially the different types of Mendeley users could be reading different publications and therefore these "career stages" could help to detect different typologies of impact. However, here again we should warn about the limitation of the lack of information on the 'unknown' users. For example, by knowing the other types of users, it could be the case that the predictability for some of the less common users (e.g. Researchers, Professors, etc.) would change. In any case, our current results suggest that aside from PhD students, as the most common type of user in Mendeley, consideration of the other users does not really improve the filtering capacity of detecting highly cited publications by readerships over JCS.

All in all, one of the most promising results is that the overall (as well as for some of the fields) filtering capacity for detecting highly cited publications by Mendeley readerships (from a 'tight' perspective) tends to outperform (or at least is quite similar to) that of the JCS indicator, something that has not been observed for other altmetric sources (cf. Costas et al, 2014). Therefore, reinforcing the potential role that readerships could play for informing scientific and alternative impacts, provided that the coverage and data limitations in Mendeley are improved, and that the limitations highlighted in this paper on the number of users reported are solved.

References:

Costas, R., Zahedi, Z. & Wouters, P. (2014). Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the American Society for Information Science and Technology*, pp:1-31.

Gunn, W. (2013). Social Signals Reflect Academic Impact: What it Means When a Scholar Adds a Paper to Mendeley. *Information Standards Quarterly*. 25(2): 33-39.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2013a). Coverage and adoption of altmetrics sources in the bibliometric community. 14th International Society of Scientometrics and Informetrics Conference (pp. 1–12). *Digital Libraries*. Retrieved from: <http://arxiv.org/abs/1304.7300>

Haustein, S., Peters, I., Sugimoto, C.R., Thelwall, M., Larivière, V., & (2013b). Tweeting biomedicine: an analysis of tweets and citations in the biomedical literature. *Journal of the American Society for Information Science and Technology*. DOI: 10.1002/asi.23101

Li, X., & Thelwall, M. (2012). F1000, Mendeley and Traditional Bibliometric Indicators. In: *Proceedings of the 17th International Conference on Science and Technology Indicators*. Montréal, Canada. pp. 451–551

Mohammadi, E., Thelwall, M, Larivière, V., & Haustein, S. (in press). Who Reads Research Articles? An Altmetrics Analysis of Mendeley User Categories. *Journal of the Association for Information Sciences and Technology*.

Sud, P. & Thelwall, M. (in press). Not all international collaboration is beneficial: The Mendeley readership and citation impact of biochemical research collaboration. *Journal of the Association for Information Science and Technology*.

Thelwall, M. & Maflahi, N. (in press). Are scholarly articles disproportionately read in their own country? An analysis of Mendeley readers. *Journal of the Association for Information Science and Technology*.

Waltman, L., Van Eck, N.J., Van Leeuwen, T.N., Visser, M.S. & Van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*. 5(1). 37–47.

Waltman, L. & Costas, R. (2014). F1000 Recommendations as a Potential New Data Source for Research Evaluation: A Comparison With Citations. *Journal of the Association for Information Science and Technology*. 65: 433–445. DOI: 10.1002/asi.23040

Zahedi, Z., Costas, R. & Wouters, P. (2013). What is the impact of the publications read by the different Mendeley users? Could they help to identify alternative types of impact? In *PLoS ALM Workshop*, 7-9 October, 2013, San Francisco, US. <http://article-level-metrics.plos.org/alm-workshop-2013/>

Zahedi, Z., Costas, R. & Wouters, P. (2014). How well developed are Altmetrics? Cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications". *Scientometrics*, DOI:10.1007/s11192-014-1264-0