A Scienceographic Comparison of Physics Papers from the arXiv and viXra Archives

David Kelk[1] and David Devine

[1]University of Ontario Institute of Technology
Oshawa, ON, Canada
`http://www.uoit.ca`
`{david.kelk@uoit.ca,davidthomas.devine@gmail.com}`

Abstract. arXiv is an e-print repository of papers in physics, computer science, and biology, amongst others. viXra is a newer repository of e-prints on similar topics. Scienceography is the study of the writing of science. In this work we perform a scienceographic comparison of a selection of papers from the physics section of each archive. We provide the first study of the viXra archive and describe key differences on how science is written by these communities.

Key words: arXiv, viXra, comparison, scienceography

1 Introduction

Bibliometrics [DB09,BF02] is the application of mathematical and statistical methods to books and other media of communication [HW01]. *Citation analysis* [DB09,Moe05] and *content analysis* [Kri03] are two bibliometric methods. *Scientometrics* [LM12,HW01] is bibliometrics applied to science and technology. Both bibliometrics and scientometrics are applied after publishing. *Scienceography* [CMY12] is the study of the writing of science. How papers are written and how they describe their results is an under-studied area. This paper performs a scienceographic analysis of 20 papers from the physics categories of the arXiv and viXra repositories. (40 papers total.) We examine whether or not there are differences in how these communities of authors write science. We were especially interested in learning if there are metrics indicating that a paper is from one archive or the other.

arXiv[1] is a repository and distribution server for research papers. It was started in August, 1991 and hosts approximately 771,000 documents[2]. Part of its *Goals and Mission* statement reads:

> arXiv is an openly accessible, moderated repository for scholarly articles in specific scientific disciplines. Material submitted to arXiv is expected to be of interest, relevance, and value to those disciplines. arXiv reserves the right to reject or reclassify any submission. Submissions are

[1] arxiv.org

[2] Retrieved on July 19, 2012

> reviewed by expert moderators to verify that they are topical and refereeable scientific contributions that follow accepted standards of scholarly communication (as exemplified by conventional journal articles)[3].

viXra[4] (arXiv spelled backwards) is a more recent e-print archive. It was started in July, 2009 and hosts approximately 3,140 papers[5]. viXra was created as a reaction to arXiv:

> ViXra.org is an e-print archive set up as an alternative to the popular arXiv.org service owned by Cornell University. It has been founded by scientists who find they are unable to submit their articles to arXiv.org because of Cornell University's policy of endorsements and moderation designed to filter out e-prints that they consider inappropriate.
>
> ViXra is an open repository for new scientific articles. It does not endorse e-prints accepted on its website, neither does it review them against criteria such as correctness or author's credentials.[6] [7]

In total, the 40 papers examined[8] had 486 pages, 60 authors, 1040 numbered equations and 751 references. We found there were differences between the papers in each archive. arXiv papers had on average 1.9 authors who were always affiliated with a university or an equivalent institution. viXra papers had an average of 1.1 authors who were university affiliated 35% of the time. 65% of arXiv papers were published in journals while 55% of viXra papers were published as web pages. arXiv papers averaged 14.3 pages in length, had a total of 481 numbered equations and 521 references. viXra papers averaged 9.9 pages, had a total of 559 numbered equations and 230 references. The complete data set is listed in Tables 2 through 5.

We found there are indicators identifying the source archive. If some or all of the authors had a university or equivalent institution affiliation, or the paper contained theorems or lemmas, it was likely to be from the arXiv repository. If a paper was published in web pages, had inline citations not in the references section, or had web hyperlinks outside the references section it was likely to be from the viXra archive. (See Table 6.)

The rest of this paper is organized as follows. Our survey methodology and research question is introduced in section 2. Results are described in section 3. Related work is discussed in 4 and is followed by threats to validity in 5. Conclusions and future work are discussed in section 6.

[3] arxiv.org/help/primer, Retrieved July 19, 2012

[4] vixra.org

[5] Retrieved on July 19, 2012

[6] vixra.org, retrieved July 19, 2012

[7] There is no connection or affiliation between the arXiv and viXra sites. We take no position on the archives, their policies, or the papers reviewed here.

[8] Raw data is available here (Google drive)

Item	**Description**
Area 1	Authors, their affiliation and collaboration
Area 2	Publishing and citation
Area 3	Writing metrics (# versions, # pages, # theorems and lemmas, . . .)
Area 4	Referencing
RQ 1	Do any of the metrics identify a paper as coming from arXiv or viXra?

Table 1. Research areas and research question.

2 Survey Methodology

Both the arXiv and viXra e-print servers are divided into a number of top-level categories: physics, mathematics, computer science and others. This paper studied only the physics category. Papers from the years 2007-2012 were considered from arXiv and 2009-2012 for viXra, due to the latter's more recent creation. Twenty papers were selected from each archive. Papers were chosen in pairs: a paper was chosen from the same sub-category from each site. As the sub-category names don't match up exactly we chose the closest matching one. For example, if we chose a paper from *High Energy Physics - Phenomenology* from arXiv, a paper from *High Energy Particle Physics* was chosen from viXra. Year and month of submission for each member of the pair were chosen randomly. To reduce bias, the paper at numerical position 10 was always selected. For viXra there were a number of months where the total number of submissions was under 10. In this case we randomly chose a year and month with 10 or more submissions. If no month had 10 or more submissions, we always selected the paper at numerical position 5. This latter case didn't occur in the survey.

Data collection was divided into 4 broad areas and guided by our research question, summarized in Table 1.

3 Survey Results

Each of the four survey areas are examined in subsections 3.1 to 3.4 and summarized in Tables 2 to 5. Our research question is answered in subsection 3.5 and Table 6.

3.1 Authors, Affiliation and Collaboration

More than half of the selected arXiv papers had two or more authors, all of whom were affiliated with a university or equivalent institution. There was a moderate amount of collaboration between institutions[9]. (See Table 2.) In contrast, the vast majority of viXra papers surveyed had one author who was not likely to be affiliated with a university. It also follows then the rate of collaboration was very low.

[9] Authors from the same institution do not count as collaboration for this metric.

Metric	arXiv	viXra
Avg. # authors	1.9	1.1
Papers with one author	35%	90%
University affiliation	100%	35%
Collaboration across institutions	35%	10%

Table 2. Authors, their affiliation and collaboration.

Metric	arXiv	viXra
Published (Journal or equivalent)	65%	15%
Published (Web page or equivalent)	0%	55%
Avg. # of citations received	1.3	0.11
Has one or more citations	35%	10%

Table 3. Publishing and citation (Google Scholar, May 1, 2012)

3.2 Publishing and Citations

Two-thirds of the arXiv papers have been published in journals. (See Table 3.) On average they have received 1.3 citations. In contrast, very few viXra papers have been published in journals and have garnered few citations as a result. Instead of journal publishing, half of the viXra authors chose to self-publish on web pages or in web journals[10]. None of the arXiv papers have done this[11].

3.3 Writing Metrics

Many of the writing metrics were very similar between the two archives: the average number of figures (per paper), numbered equations (per paper) and versions a paper had gone through. (Summarized in Table 4.) arXiv papers were about 45% longer than viXra and almost twice as likely to contain figures or tables. The largest difference was in the use of theorems and lemmas. Four arXiv papers contained a total of 30 while only one viXra paper contained one.

3.4 Referencing

arXiv papers had twice as many references, pointed[12] to them twice as often and self-referenced twice as often. (Table 5) When averaged per page, the arXiv reference numbers were about 50% larger than viXra's.

We found a number of citing behaviours in viXra papers not occurring in arXiv papers: all references were self references, the use of inline citations not in

[10] viXra papers appearing in web journals are classified as Published (Web page or equivalent).

[11] Making a PDF available is not considered web publishing for this metric.

[12] For the purposes of this paper, [HW01] is an example of a pointer to a reference.

Metric	arXiv	viXra
Avg. # of versions of paper	1.35	1.55
Avg. # of pages	14.3	9.9
Avg. # of figures and tables	3.5	3.9
Has figures and tables	80%	45%
Avg. # of theorems and lemmas	1.5	0.05
Has theorems or lemmas	20%	5%
Avg. # of numbered equations	24	28
Has numbered equations	90%	75%

Table 4. Writing metrics.

Metric	arXiv	viXra
Avg # of references	26	11.5
Has references	100%	90%
Avg # of inline pointers to references	37.7	20.8
Has inline pointers to references	100%	85%
Has unused references	0%	10%
Avg # of self-references	3.6	1.9
Has self-references	80%	55%
All references are self-references	0%	15%
Total # of inline citations not in references	0	21
Papers with inline citations not in references	0%	25%
Total # of hyperlinks outside references section	0	71
Papers with hyperlinks outside references section	0%	30%

Table 5. Referencing.

the references section[13] and the use of hyperlinks outside the references section. These occurred in a minority of papers, between one-sixth to one-third.

3.5 Metrics Identifying Source Archive

Metrics with large differences were extracted from Tables 2 to 5 to create the list of indicators for the source archive. (Summarized in Table 6.) The strong indicators for arXiv are unsurprising: university or equivalent affiliation and publication in a journal. That viXra has low university affiliation, low journal

[13] (Stephen Hawking, A Brief History of Time, 1988) is an example of an inline citation not in the references section.

Metric	**arXiv**	**viXra**
University affiliation	100%	35%
Published (Journal or equivalent)	65%	15%
Has theorems or lemmas	20%	5%
Collaboration across institutions	35%	10%
Has one or more citations	35%	10%
Papers with one author	35%	90%
Published (Web page or equivalent)	0%	55%
Papers with inline citations not in references	0%	25%
Papers with hyperlinks outside references section	0%	30%

Table 6. Metrics most likely to identify which archive a paper comes from.

publication rates and high self-publication rates, along with other stylistic differences from the other viXra-only indicators, indicate it has a more diverse pool of authors.

4 Related Work

arXiv has appeared in numerous studies. In [GBMB09] the High Energy Physics (HEP) sub-categories were studied to determine the relative advantages of publishing in repositories versus open access journals. They concluded there was an *"immense citation advantage"* for HEP papers in repositories.

One years worth of HEP papers from arXiv were studied in [MDVY06] to determine the share of HEP production by country and institution.

arXiv sends out daily email announcements of new papers. The effect on long term citation count based on position within the announcement list was studied in [HG09,HG10]. They found there was an enhancement to the citation rate for papers at the beginning and end of the list.

We searched for, but could not find any papers studying viXra.

A scienceographic study of arXiv was carried out in [CMY12]. Latex files from 65,000 papers from the mathematics and computer science disciplines were analysed. Items such as comments, authors and diagrams were quantified and compared between the two disciplines. Other metrics such as the number of pages and number of Latex packages used were tracked over a 15 year period.

This work is very complimentary to [CMY12]. A subset of metrics from each paper are the same (Avg. # of authors/pages/theorems, . . .). Where [CMY12] emphasizes Latex analysis (Comments, word counts, packages, . . .) this work considers PDF-level analysis (References, citations, equations, . . .). Where the metrics are the same, Table 7 synthesizes (with caveats) the results of the two papers.

Metric	**arXiv Physics**	**viXra**	**arXiv Mathematics**	**arXiv Computer Science**
Avg. # authors	1.9	1.1	1.2	1.7
Percentage of papers with a single author	35	90	More than half	38
Avg. # of pages	14.3	9.9	15	9
Avg. # of theorems and lemmas in papers containing theorems and lemmas	7.5	1	5.5	4.9
Percentage of papers with theorems or lemmas	20	5	71	48

Table 7. Synthesis of compatible data. Columns *arXiv: Mathematics* and *arXiv: Computer Science* are from [CMY12].Columns 2 and 3 were calculated from 20 papers each. Columns 4 and 5 were calculated from approx. 39,000 and 26,000 papers respectively.

5 Threats to Validity

After posing and answering the research question we were then responsible for considering threats to the validity of our results:

Internal threats: Any bias in experimental design could be an internal threat. One source of bias is selecting metrics favouring one archive over the other. This can be seen in extremely high or low numbers appearing consistently in results for arXiv or viXra. arXiv scored higher on 9 out of 14 criteria suggesting a bias in its favour. Given this it is interesting to observe the four scores of zero, {Publilshed on the Web or equivalent, number of inline references not in end references, all references being self-references and number of web links outside the references section} are for arXiv. These categories favoured viXra.

Papers selected may not be a representative sample of the physics papers in the arXiv and viXra archives. To mitigate this we used a consistent selection process (matching sub-categories across archives, selecting year and month randomly from the last six (four) years and selecting the paper at numerical position 10) to minimize human bias.

External threats: Any bias hurting the generalizability of the results is an external threat. arXiv scored 0 in 4 metrics. (See internal threats.) These metrics only appear in papers from the viXra archive so we cannot expect them to generalize to all archives.

6 Conclusions

This paper performed a scienceographic analysis of 20 papers from the physics category of the arXiv and viXra e-print archives. Differences were found between the writing styles and contents of the papers of each. These differences are captured as a series of indicators. Many of the indicators are weak, appearing in one-third or less of papers.

This paper makes two contributions: a first study of papers from the viXra archive and a scienceographic comparison of the papers from viXra and arXiv.

6.1 Future Work

viXra's open access policies have attracted a population of non-academically trained authors [Wer11]. It would be interesting to verify this and determine if there are metrics indicating which group a paper falls into[14] [15]. For example, are the 4 metrics where arXiv[16] papers scored 0 indicative of non-academically trained authors? Creating a classifier using machine learning or evolutionary techniques could help answer this question.

Expanding the study to include more sub-disciplines from arXiv, viXra and other repositories could give further interesting insights into how science is written by different communities and sub-communities.

7 Acknowledgements

We thank Margaret Wertheim for providing the inspiration [Wer11] to write this paper.

References

BF02. C.L. Borgman and J Furner. Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36(1):3–72, 2002.

CMY12. Graham Cormode, S Muthukrishnan, and Jinyun Yun. Scienceography: the study of how science is written. In *FUN'12*, pages 379–391, 2012.

DB09. Nicola De Bellis. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*, volume 31. Scarecrow Press, 2009.

GBMB09. Anne Gentil-Beccot, Salvatore Mele, and Travis Brooks. Citing and reading behaviours in high-energy physics. how a community stopped worrying about journals and learned to love repositories. *Journal of High Energy Physics*, pages 1–13, 2009.

[14] Simply looking to see if the author is academically affiliated may not be enough.

[15] arXiv papers with their arXiv:NNNN.NNNNvN stamp in them have also been posted on viXra.

[16] Based on this work we are implicitly assuming all arXiv authors are academically trained. Perhaps this isn't true.

HG09. Asif-ul Haque and Paul Ginsparg. Positional effects on citation and readership in arxiv. *Journal of the American Society for Information Science and Technology*, 60(11):2203–2218, 2009.

HG10. Asif-ul Haque and Paul Ginsparg. Last but not least: Additional positional effects on citation and readership in arxiv. *Journal of the American Society for Information Science and Technology*, 61(12):2381–2388, 2010.

HW01. William Hood and Concepcin Wilson. The literature of bibliometrics , scientometrics , and informetrics. *Scientometrics*, 52(2):291–314, 2001.

Kri03. Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*, volume 79. Sage Publications Inc., 2003.

LM12. L. Leydesdorff and S. Milojević. Scientometrics. *ArXiv e-prints*, August 2012.

MDVY06. Salvatore Mele, David Dallman, Jens Vigen, and Joanne Yeomans. Quantitative analysis of the publishing landscape in high-energy physics. *Journal of High Energy Physics*, 2006(12), 2006.

Moe05. H.F. Moed. *Citation Analysis in Research Evaluation*, volume 13 of *Information Science and Knowledge Management*. Springer, 2005.

Wer11. Margaret Wertheim. *Physics on the Fringe: Smoke Rings, Circlons, and Alternative Theories of Everything*. Walker Publishing Company, New York, 2011.