How to Cheat the Page Limit: the 2022 Update

Wouter Duivesteijn[1,2] (✉) and Sibylle Hess[1,2]

[1] Proceedings Chair of ECML PKDD 2022
[2] Data Mining Group, Technische Universiteit Eindhoven, the Netherlands,
{w.duivesteijn,s.c.hess}@tue.nl

Abstract. At ECMLPKDD 2019, the proceedings chairs released a report on the topic of page limit cheating: authors tweaking the parameters of the game such that they can squeeze more content into their paper. This paper provides the 2022 update: what has changed since the 2020 and 2019 editions, and what hasn't.

Keywords: Scientific integrity · Conference organization · Reviewing process · LaTeX· Program chairs

1 What Is Going On?

LaTeX offers far too many ways to change the appearance of a paper. As a consequence, it is far too easy for authors to squeeze in more material than the page limit would allow if one were to play the game fairly. How the game is supposed to be played fairly varies from venue to venue. At ECML PKDD, we publish our proceedings with Springer, in the Lecture Notes in Computer Science series. That means that authors should prepare their papers in accordance with Springer's Guidelines for Proceedings Authors [7].

As we have seen in recent ECML PKDD editions, authors tend to violate the guidelines to sneak more material into their allocated 16 pages. In our[3] role as Proceedings Chairs at ECML PKDD 2019, 2020, and 2022, we received the LaTeX sources of all papers accepted into the Research and Applied Data Science tracks. With those LaTeX sources, we did the following. First, we compiled the sources as the authors delivered them to us. This gives us the page length of the papers as the authors intended, on our system. Subsequently, we removed all commands that violate the guidelines (cf. [1] for some details). This gives us the page length of the papers, in the form compliant with Springer's guidelines [7]. We end up with the following data.

2 The Data

Raw results for ECMLPKDD 2019 can be found in Appendix A, for ECMLPKDD 2020 in Appendix B, and for ECML PKDD 2022 in Appendix C. Here,

[3] each ECML PKDD edition from the years 2019, 2020, and 2022 had as Proceedings Chairs a different subset from {Wouter Duivesteijn, Sibylle Hess, Xin Du}; the intersection across the three years is nonempty.

Table 1. Statistics across all papers accepted to the Research and Applied Data Science tracks at three editions of ECML PKDD. The first column contains the acronym in use on the ECML PKDD website of that year, the second column contains the year. Subsequent columns contain: the percentage of papers that were overlength; the number of pages used in the longest paper; the percentage of papers that used any space-cheating commands (even though their paper may still be under the page limit); the percentage of papers whose LaTeX sources wouldn't compile as the authors submitted them; the percentage of papers whose authors didn't manage to completely fill out Springer's Licence to Publish form. Except for the year, in all columns, lower is better.

Acronym	Year	%overlength	longest	%cheats	%TeX-errors	%form-errors
ECMLPKDD	2019	47.7	24	76.2	21.5	08.5
ECML-PKDD	2020	49.2	19	79.5	23.6	71.8
ECML PKDD	2022	59.3	22	79.7	18.3	32.4

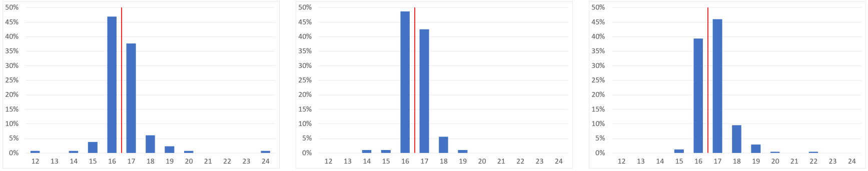
we only present aggregated results that can be compared across the years, normalizing by the number of submitted papers in each year.

From Table 1, the following main conclusions can be drawn:

1. The percentage of papers that are over the page limit keeps increasing year after year. Sometime between the 2020 and 2022 editions, the percentage increased drastically. Since we raised awareness of the problem, the problem only grew bigger, so clearly a soft hand is not working.
2. The length of the longest paper dipped in 2020, the year after a speech on this topic was held at the ECMLPKDD 2019 community meeting. This seemed to have scared off the worst excesses for one ECML PKDD edition, but the effect did not last.
3. Only one in five papers do not use any commands violating Springer's Guidelines. This number is slowly dropping from an already alarmingly low starting point.
4. On a lighter note, a disturbingly large percentage of the ECML PKDD community is incapable of submitting LaTeX source files that compile out of the box. However, the number for 2022 is making a move in the right direction.
5. Finally, in 2022 Springer introduced a new Licence to Publish form. This form no longer has very pronounced grey boxes indicating which fields to fill out on the first page. Instead, lighter grey text is used. Many authors do not spot these fields as something actionable for them. Hence, Proceedings Chairs need to chase an annoyingly large number of incompletely filled out forms. This problem quadrupled in size from 2019, when the old form was in use. The number for 2020 cannot be compared, because in that year we forgot to communicate in one track that authors had to fill out a Licence to Publish form at all; as the number shows, consequences were terrible.

Table 2. Histograms (normalized by total number of papers accepted in that year, numbers given in percentages) with heatmaps of paper lengths in pages (partially) used, across the papers accepted for the Research and Applied Data Science tracks at ECML PKDD 2019, 2020, and 2022. Notice that the page limit is 16 pages. These rows contain the histograms of the papers recompiled after all space cheating commands from Section 1 of [1] were removed; data on intially submitted versions can be found in the Appendices.

number of pages	12	13	14	15	16	17	18	19	20	21	22	23	24
ECMLPKDD 2019	0.77	0.00	0.77	3.85	46.92	37.69	6.15	2.31	0.77	0.00	0.00	0.00	0.77
ECML-PKDD 2020	0.00	0.00	1.03	1.03	48.72	42.56	5.64	1.03	0.00	0.00	0.00	0.00	0.00
ECML PKDD 2022	0.00	0.00	0.00	1.24	39.42	46.06	9.54	2.90	0.41	0.00	0.41	0.00	0.00



(a) 2019 (b) 2020 (c) 2022

Fig. 1. The data from Table 2 in chart form.

Table 2 displays the histograms per year, of the distribution of paper length in pages (partially) used. Per year, the histogram is normalized by the total number of papers accepted that year, and displayed in percentages. The same data is also represented in Figure 1. We can conclude:

1. The histogram for 2020 is markedly narrower than for both other years: much fewer outliers are present, particularly on the right-hand side. In 2020, only 1.03% of all papers used 19 pages or over; this number is 3.85% for 2019 and 3.72% for 2022. Considering papers of 20 or more papers, there were none in 2020, compared to 1.54% in 2019 and 0.82% in 2022.
2. Particularly notable for 2022 is the sharp reduction in the percentage of papers that use precisely 16 pages: this is fewer than 2 in 5, where it used to be closer to half.
3. There seems to be a linear yearly increase, by about 4 percentage points per year, in the percentage of papers that use 17 or 18 pages. This percentage went from 43.84% in 2019, to 48.20% in 2020, to 55.60% in 2022.

3 Conclusions

Page limit cheating isn't going anywhere, and extent of the problem is getting bigger rather than smaller. We think the ECML PKDD community must fix this problem; raising awareness clearly hasn't been enough.

We reiterate our proposal to include a TEXnical Desk Reject Phase into the paper submission process, directly after the paper submission deadline. The details of this Phase are outlined in Section 3.2 of [1]. We could consider making use of the texmlbus software [2] created by Heinrich Stamerjohanns, downloadable from [3]. This software can highlight all commands outlined in [1].

However, a list of new offending commands must be compiled from the observations made in 2020 and 2022, and these must be included in the software before deployment; lots of new innovative ways to cheat the page limit have been developed in the last three years. We must also take care that a software solution does not lull authors into a false sense of security. Since LATEX space cheating is dependent on the context in which a command is used (see [1, Section 3.1] for details), and since new LATEX packages are released all the time, a software solution to identify space cheating will necessarily result in both false positives and false negatives. A human touch is still necessary.

References

1. Wouter Duivesteijn, Sibylle Hess, Xin Du: *How to Cheat the Page Limit*. WIREs Data Mining and Knowledge Discovery 10(3):e1361, 2020. `https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1361`
2. Heinrich Stamerjohanns: *Texmlbus, a build system to convert documents to XML and other formats*. TUGboat 42(2):132–134, 2021. `https://tug.org/TUGboat/tb42-2/tb131stamerjohanns-texmlbus.pdf`
3. Heinrich Stamerjohanns et al., texmlbus. Available online: `https://github.com/stamer/texmlbus`. Accesssed June 05, 2020.
4. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases: *ECMLPKDD 2019*. Available online: `http://www.ecmlpkdd2019.org/`. Accessed September 07, 2022.
5. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases: *ECML-PKDD 2020*. Available online: `http://www.ecmlpkdd2020.net/`. Accessed September 07, 2022.
6. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases: *ECML PKDD 2022*. Available online: `https://2022.ecmlpkdd.org/`. Accessed September 07, 2022.
7. Springer: *Instructions for Proceedings Authors*. Available online: `https://resource-cms.springernature.com/springer-cms/rest/v1/content/19242230/data/v9`. Accessed September 07, 2022.

A Raw Results for ECMLPKDD 2019

Table 3. Histogram of paper lengths in pages (partially) used, across the 130 papers accepted for the Research and Applied Data Science tracks at ECMLPKDD 2019. Notice that the page limit is 16 pages. The second row contains the histogram of the papers compiled as they were originally submitted; the third row contains the histogram of the papers recompiled after all space cheating commands from Section 1 of [1] were removed.

number of pages	12	13	14	15	16	17	18	19	20	21	22	23	24
submitted version	1	0	2	5	112	10	0	0	0	0	0	0	0
reformatted version	1	0	1	5	61	49	8	3	1	0	0	0	1

Table 4. Condensed form of Table 3, retaining only the information whether papers were over or under the limit.

complied with page limit?	yes	no
submitted version	120	10
reformatted version	68	62

Paper length histograms of all 130 papers accepted to the Research and ADS Tracks at ECMLPKDD 2019 [4] are given in Table 3, both in the form as originally submitted by the authors, as in the reformatted form. In Table 4, we summarize for both versions how many papers complied with the limit. The main conclusions were trifold:

1. 62 out of 130 papers (47.7%) were over the page limit;
2. the longest paper was 24 pages long;
3. only 31 of the 130 papers (23.8%) did not contain any space cheats (not shown in tables).

B Raw Results for ECML-PKDD 2020

Table 5. Histogram of paper lengths in pages (partially) used, across the 195 papers accepted for the Research and Applied Data Science tracks at ECML-PKDD 2020. Notice that the page limit is 16 pages. The second row contains the histogram of the papers compiled as they were originally submitted; the third row contains the histogram of the papers recompiled after all space cheating commands from Section 1 of [1] were removed.

number of pages	14	15	16	17	18	19
submitted version	2	3	186	4	0	0
reformatted version	2	2	95	83	11	2

Table 6. Condensed form of Table 5, retaining only the information whether papers were over or under the limit.

complied with page limit?	yes	no
submitted version	191	4
reformatted version	99	96

Paper length histograms of all 195 papers accepted to the Research and ADS Tracks at ECML-PKDD 2020 [5] are given in Tables 5 and 6, respectively. The three main conclusions are:

1. 96 out of 195 papers (49.2%) were over the page limit;
2. the longest paper was 19 pages long;
3. only 40 of the 195 papers (20.5%) did not contain any space cheats (not shown in tables).

These conclusions compare to the previous year as follows. The good news is that the most egregious cases of space cheating are no longer as egregious; the longest paper is substantially shorter than last year. The bad news is that less blatant space cheating happens more often: the percentage of papers that were over the page limit went up, and the percentage of papers not containing any space cheats went down.

C Raw Results for ECML PKDD 2022

Table 7. Histogram of paper lengths in pages (partially) used, across the 241 papers accepted for the Research and Applied Data Science tracks at ECML PKDD 2022. Notice that the page limit is 16 pages. The second row contains the histogram of the papers compiled as they were originally submitted; the third row contains the histogram of the papers recompiled after all space cheating commands from Section 1 of [1] were removed.

number of pages	15	16	17	18	19	20	21	22
submitted version	5	224	12	0	0	0	0	0
reformatted version	3	95	111	23	7	1	0	1

Table 8. Condensed form of Table 7, retaining only the information whether papers were over or under the limit.

complied with page limit?	yes	no
submitted version	229	12
reformatted version	98	143

Paper length histograms of all 241 papers accepted to the Research and ADS Tracks at ECML PKDD 2022 [6] are given in Tables 7 and 8, respectively. The three main conclusions are:

1. 143 out of 241 papers (59.3%) were over the page limit;
2. the longest paper was 22 pages long;
3. only 49 of the 241 papers (20.3%) did not contain any space cheats (not shown in tables).

These conclusions compare to the previous years as follows. The percentage of overlength papers is significantly (*ten percentage points!*) higher than it was in 2019 and 2020. The number of completely clean papers is comparable to 2020 but down from 2019. The worst offender in 2022 is substantially worse than the one in 2020, but didn't quite achieve the extreme lengths of the worst offender in 2019.