How to Cheat the Page Limit: the 2020 Update

Wouter Duivesteijn[1,2] (✉)

[1] Proceedings Chair of ECMLPKDD 2019
[2] Data Mining Group, Technische Universiteit Eindhoven, Eindhoven, the Netherlands, `w.duivesteijn@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 2020 update: what has changed since last year, and what hasn't.

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

1 What Happened Last Year? A Short Recap of the Proceedings Chairs Presentation at the ECMLPKDD 2019 Community Meeting

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 ECMLPKDD, both in the editions of 2019 [3] and 2020 [4], 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 [5].

As we have seen in 2019, authors tend to violate the guidelines to sneak more material into their allocated 16 pages. In our[3] role as Proceedings Chairs at ECMLPKDD 2019, we received the LATEX sources of all 130 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 details). This gives us the page length of the papers, in the form compliant with Springer's guidelines [5]. We end up with the following data.

Paper length histograms are given in Table 1, both in the form as originally submitted by the authors, as in the reformatted form. In Table 2, we summarize for both versions how many papers complied with the limit.

---

[3] Sibylle Hess, Xin Du, and I; there were three Proceedings Chairs last year

Table 1. 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 2. Condensed form of Table 1, 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

1.1 Main Conclusions and Recommendations

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).

Based on this data, we recommended that ECMLPKDD 2020 and onwards would include a TEXnical Desk Reject Phase just after paper submission, to eradicate the phenomenon of space cheating (cf. Section 3.2 of [1] for details).

2 What Happened Between Editions? Decisions and Events Between ECMLPKDD 2019 and 2020

Three main things have happened between the two editions. This report discusses them in chronological order.

2.1 Process Decisions for ECMLPKDD 2020

The General and Program Chairs for ECMLPKDD 2020 decided not to follow our recommendation to include a TEXnical Desk Reject Phase. While I think that this is the wrong decision, I would like to take a moment to point out that I strongly support the right of the General and Program Chairs to make that decision the way they made it. The General and Program Chairs have quite a bit of leeway in shaping their edition of ECMLPKDD in the way they see fit, and that's the way it should be; the Proceedings Chairs may have a mind of their own, but they serve the needs of the General and Program Chairs.

Part of the decision not to include a TEXnical Reject Phase, was infused by the desire by the General and Program Chairs to have an automated system that informs authors whether or not their paper sources conform to the guidelines. This automation, however, is nigh-on impossible to get right, as outlined in Section 3.1 of [1]. Particularly, it is a non-trivial problem to determine whether certain LATEX commands are used in a way that violates the guidelines: a command like `\scriptsize` severely reduces the font size, and Springer's guidelines determine that the font size should be 10pt. If `\scriptsize` occurs in the main text, this violates the guidelines. If, however, `\scriptsize` occurs in the labeling of a TikZ picture, this picture may become a figure that is blown up to reasonable proportions by enveloping commands, which does not violate the guidelines. Whether the occurrence of `\scriptsize` violates the guidelines depends on the context, and this context may not be easy to determine automatically.

2.2 Formal Publication of the 2019 Report

The report the proceedings chairs of ECMLPKDD 2019 wrote on the space cheating problem, has been published as a journal paper by WIREs Data Mining and Knowledge Discovery [1]. This means that we, as a community, can now refer to a publication containing a zoo of LATEX commands that should not be used: any command appearing in Section 1 of [1] should be avoided. Having this resource available should make it easier to make hard rules for future editions of ECMLPKDD.

2.3 Springer Automates

As discussed in Section 2.1 and outlined in detail in Section 3.1 of [1], it is impossible to make an automated *decision* whether or not a paper violates the guidelines. However, this does not mean that *helpful* tools cannot be made. Springer's editorial office took notice of our report, and incorporated its list of offending commands into their `texmlbus` software [2], available as open-source software.

The authors and I have agreed that I would get in touch with them as soon as the updated list of offending commands from the 2020 edition is available; these commands will subsequently be incorporated. This software flags up papers; I would not advise yet to use it for making hard rejection decision, because of the aforementioned context problems. I have not checked its FPR/TPR.

3 What Happened This Year? An Update of Space Cheating at ECMLPKDD 2020

ECMLPKDD is growing. The 2019 edition accepted a grand total of 130 papers to the Research and Applied Data Science tracks; the 2020 edition accepted 195. The updated versions of Tables 1 and 2 for 2020 can be found in Tables 3 and 4, respectively.

Table 3. Histogram of paper lengths in pages (partially) used, across the 195 papers accepted for the Research and Applied Data Science tracks at ECMLPKDD 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 [**?**] 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 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	191	4
reformatted version	99	96

3.1 Updated Main Conclusions

The three main conclusions, updated for 2020, 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.

3.2 New Cheats for 2020

In Section 1 of [1], we outlined all the commands used for space cheating in the 2019 batch of papers. The vast majority of space cheating commands occurring in the 2020 batch of papers, also occurs in that overview; there were very few creative cheaters this year. Still, there is a small extension to our 2019 collection.

In due time, I will list this extension in an updated version of this section.

4   Conclusions

The General and Program Chairs of ECMLPKDD 2020 have elected to treat the problem of space cheating with a soft hand. They are well within their rights to do so. By comparing the results from Sections 1.1 and 3.1, we can conclude that this soft hand does not result in reduced cheating.

Authors of papers at ECMLPKDD 2019 had plausible deniability; while Springer's Guidelines for Proceedings Authors have always been available to them, we as ECMLPKDD community never made explicit what is and isn't allowed. Authors of papers at ECMLPKDD 2020 have no such deniability; those who attended the ECMLPKDD 2019 community meeting heard our presentation, and others could access our report and subsequent publication. Despite the fact that we made explicit what isn't allowed, the problem did not go away or even reduce in size. While the longest paper reduced in size from 24 to 19 pages, the percentage of papers that cheated the page limit increased, both in terms of papers that violated the allowed number of pages (up from 47.7% to 49.2%) and papers that employed disallowed commands (up from 76.2% to 79.5%).

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]. The resources outlined in Sections 2.2 and 2.3 will be helpful in gaining wider acceptance of this phase in our ECMLPKDD community.

References

1. Wouter Duivesteijn, Sibylle Hess, Xin Du: *How to Cheat the Page Limit*. WIREs Data Mining and Knowledge Discovery 10(3):e1361, 2020.
2. Heinrich Stamerjohanns et al., texmlbus. Available online: `https://github.com/stamer/texmlbus`. Accesssed June 05, 2020.
3. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases: *ECMLPKDD 2019*. Available online: `http://www.ecmlpkdd2019.org/`. Accessed August 15, 2019.
4. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases: *ECMLPKDD 2020*. Available online: `http://www.ecmlpkdd2020.net/`. Accessed September 14, 2020.
5. Springer: *Guidelines for Proceedings Authors*. Available online: `ftp://ftp.springernature.com/cs-proceeding/svproc/guidelines/Springer_Guidelines_for_Authors_of_Proceedings_CS.pdf`. Accessed August 15, 2019.