An information-theoretic approach to author order

Ari Trachtenberg Carl Friedrich Gauss

July 9, 2010

Abstract

We consider several ambiguities and inconsistencies that arise from interpretations of author order in published documents, and propose information-theoretic solutions for consideration as a broad-based convention for the academic community.

The order in which authors are presented in a publication traditionally encodes information about the relative contributions and values of the authors. The particular encoding depends on its academic community, with notable examples such as:

- **Alaphabetic.** Such an ordering may indicate that each author has contributed equally to the paper. In some communities, this style is the *de facto* standard for all publications.
- **Contribution.** The first author has provided the greatest contribution to the paper, followed (typically with exponential decay) by the second and subsequent authors. Some of the middle authors may be included simply for show.
- **Delineation.** Each aspect of the publication process is explicitly identified: who wrote the article, who did the research, who performed the statistical analysis, who secured the funding, who was mentioned because they are too important not to mention, who fudged the data, etc. This is valuable not just for assigning credit, but, more importantly, for assigning blame should the work subsequently be discredited.
- **Lottery.** Authors with regular collaborations take turns being first on their joint papers. If a paper wins a major award or prize, the author who was first on that paper wins the jackpot.
- **Altruistic.** The youngest member of the author team, or maybe the one closest to tenure or the one with the largest mortgage, is placed first, with the older, more secure authors last.
- **Political.** The leader of the lab/group/division is placed last. Everyone else does not matter.

1 The problem

Unfortunately, some communities operate under more than one author ordering convention. This is especially exacerbated by inter-disciplinary teams, where it is possible that different authors are using different conventions within the same paper.

For example, if Einstein, Florey, and Pascal were to collaborate on a paper, there could be several interpretations:

1. Each have contributed equally to the piece, as unlikely as that may be.
2. Einstein contributed the most to this work, followed by a much smaller contribution by Florey, and an imperceptible contribution by Pascal, who was legally dead at the time.
3. Einstein and Florey wrote the paper and put Pascal on at the end because he was:

(a) a person of note who had nothing to do with the paper but would increase its visibility.

(b) someone with valuable connections to industry/funding/organized crime.

(c) it looked strange with only two authors.

2 A proposed solution

We propose a novel, efficient, scalable, and optimal approach for completely disambiguating author ordering. Though we do not prove it, our approach is also elegant, general, and distributed, and it features some unspecified connection to the Grassmannian manifold.

To better understand our proposed solution, consider a set of permtutations π_i on n elements (in this case, authors). Define the distance between two permtuations as in [2]:

$$d(\pi_1, \pi_2) = \min\{d : \pi_1 g_1 g_2 \dots g_d = \pi_2 \text{ for some } g_i \in G\},$$

where G is a set of generators for the symmetric group S_n. Following the exposition in [2], define the *Ulam distance* $U(\pi_1, \pi_2)$ to be the distance associated with the generators $\{c_{ij}^{\pm 1}\}$ for c_{ij} being the cycle $(i, i+1, \dots j)$; it is thus obvious that $U(\iota, \pi) = n - l(\pi)$, where ι is the identity permutation and $l(\pi)$ remains undefined.

It is thus natural to establish a convention on the meaning of author order as a function of ordering deviation from the canonical (alphabetical) ordering. More precisely, we can ascribe a specific meaning to each possible author ordering, based on its Ulam distance from the alphabetic order of the authors' last names.

Theorem 1. *The following convention on author ordering provides unambiguous meaning to each possible ordering of authors names in a novel, efficient, scalable, distributed, optimal and general manner:*

Dist	***Meaning***
0	*No meaning may be infered from author ordering.*
1	*The first author did all the work. The others are for show.*
2	*The first author is just for show. The others did all the work.*
3	*The first and second author could not agree on an order, but they agree that all other authors are not important.*
4	*All authors contributed equally.*
$\vdots$	$\vdots$

where ***Dist*** *refers to the Ulam distance from the alphabetic permutation of the authors' names.*

Proof: It is clear that the convention is elegant as it occupies very little space on the page. Its efficiency follows from is utilization of a lookup table, which is scalable to any (finite) Ulam distance. The approach is optimal because it extracts the information theoretic maximum amount of information from the author ordering. □

3 Complications

Practical applications of the proposed convention may find a number of non-trivial issues.

Name repetitions Authors whose last names and initials are identical cannot be disambiguated with their author ordering. For example, `A Trachtenberg` could refer to Ari, Alan, or Alexander Trachtenberg, and if all of these wrote an article together, the list of their names would be fixed under permutation. Such issues can be solved using classic techniques, such as using full names in citations or making up additional middle initials for identical authors.

Ambiguity In some cases, ambiguity is not only helpful, it is necessary to permit a paper to be published. For example, if two authors disagree on who contributed the most to a result, they can still publish the paper without unnecessarily bruising either ego if each author assumes a different convention for the author ordering. This approach can be easily generalized to such cases of necessary ambiguity by agreeing on several tables (similar to what is in Theorem 1), any one of which could be inferred by a given reader.

We leave the joint optimization of these various tables to a future work.

More authors The amount of information that may be conveyed by a given author ordering naturally grows exponentially in the number of authors, as per Stirling's approximation [3]:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

In other words, it is advisable for as many authors as possible to be included on a submission so that the author contributions may be more finely conveyed.

Disagreement For papers in which authors cannot agree on the proper apportionment of credit, we propose the canonical author designation *Ari Trachtenberg and Carl Friedrich Gauss*, which will serve to indicate such disagreement.

4 Conclusion

This is (arguably) where we stopped thinking.

Acknowledgements

The authors would like to thank Prakash Ishwar, who refused to have his name associated with this work, Felicia Trachtenberg, who edited it, and Benni, Yoni, and Manni Trachtenberg, who had nothing to do with it whatsoever.

References

[1] S.M. Ulam. *Some ideas and prospects in biomathematics*. Ann. Rev. Biophys. Bioeng., 1972.

[2] David Aldous and Persi Diaconis, *Longest Increasing Subsequences From Patience Sorting to the Baik-Deift-Johansson Theorem*, Bull. Amer. Math. Soc, 1999.

[3] Miyabe, H. ; Hamaguchi, K. ; Takahashi, K., *An approach to the design of Stirling engine regenerator matrix using packs of wire gauzes*, Journal Volume: 4; Conference: 17. Intersociety Energy Conversion Engineering conference, Los Angeles, CA, USA, 8 Aug 1982.