Estimating the Gini coefficient of M5S's Rousseau's commentaries

Enka Blanchard
CNRS, UPHF

January 3, 2023

1 Introduction

The goal of this preliminary report is to look at the proposals and comments on the Rousseau platform and to try to estimate the distribution of comments, using only data from Lorenzo Mosca's 2018 article[Mos20]. We know that one proposal (out of 326) gathered more than 10% of all comments, and are seeking a metric to compare the whole distribution with those of other platforms. We choose the Gini coefficient and show that it is at least equal to 0.124 (using conservative estimates and ignoring multiple elements which would increase it). Given the information available, we cannot compute this factor directly but must bound it (from both direction). This metric has multiple weaknesses, but one we should mention is that we are looking at temporally distributed data (as we are looking at 4 years of proposals and comments). The non-uniformity of the resulting distribution is hence not just due to the interest being captured by certain proposals (as well as random fluctuations) but also the decreasing interest in the platform itself (according to [Mos20]).

From the initial article, we know the following elements:

- The total number of proposals, by year, were 79, 81, 79 and 87 (a total of 326).
- Overall, a total of 71464 comments were submitted.
- The two main (initial) proposals gathered 4457 and 8153 comments.
- Only four proposals got more than 1000 comments during the first year.
- According to the author "during [years 2 and 3], only three proposals reached over 500 comments" and "in [the 4th year], only four laws collected between 150 and 200 comments".
- The average number of comments per law amounted to 446 in year 1, then 184, 144 and 63. The initial proposal with 4457 votes is not counted in these averages.
- The article also features a histogram of comments by month and proposals by month but this is sadly not directly usable as we do not know to which proposals correspond which comments, and more importantly, we do not know the exact duration of each consultation.

2 Lower bounds

From the definition of the Gini coefficient G, we get that

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2\overline{x}}$$

To facilitate computation, let's set P as the sum of all contributions and

$$G' = 2n^2\overline{x}G = 2nPG = \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$$

By setting p_i as the number of comments on proposal i and ranking the values in decreasing order, we can get the following:

$$G' = \sum_{i=1}^{n}(n-2i) \times p_i = \sum_{i=1}^{n}(n-i) \times p_i - \sum_{i=1}^{n} i \times p_i$$

Let's set $\hat{P}_i = \sum_{j=i}^{n} p_j$. We then get

$$G' = \sum_{i=1}^{n}(n-i) \times p_i - \sum_{i=1}^{n} \hat{P}_i = \sum_{i=1}^{n}(n-i) \times p_i - \hat{P}_i$$

We then have that for any i :

$$(n-i) \times p_i - \hat{P}_i = \sum_{j=i}^{n} p_i - p_j \geq 0$$

We can have one way of counting the direct contribution $G_i = \frac{(n-i) \times p_i - \hat{P}_i}{2nP}$ to the Gini provided by each p_i in this equation, with the Gini being the sum of all contributions. As each contribution is positive, a partial sum of these contributions yields a lower bound for the Gini. We can also notice that if $p_i = p_j$, their contributions are equal.

For p_1, we get

$$G_1 = \frac{(n-1) \times p_1 - \sum_{j=1}^{n} p_j}{2nP} = \frac{(n-1) \times p_1 - P}{2nP}$$

Replacing with the value for p_1 makes the contribution to G equal to

$$\frac{325 \times 8153 - 71464}{652 \times 71464} \approx 0.0553$$

The 2nd element's contribution is equal to

$$G_2 = \frac{(n-2) \times p_2 - \sum_{j=2}^{n} p_j}{2nP} = \frac{(n-2) \times p_2 + p_1 - P}{2nP}$$

$$\frac{324 \times 4457 + 8153 - 71464}{652 \times 71464} \approx 0.0298$$

We can do the same for the other proposals above 1000 and 500 comments. To give lower bounds, let's assume that they respectively equal 1000 and 500. For the ones above 1000, we then get at least :

$$\frac{323 \times 1000 + (8153 + 4457) - 71464}{652 \times 71464} \approx 0.0056$$

Using the histogram in [Mos20], it seems reasonable to assume that the four laws above 1000 comments do not include the top two. Let us then assume that there are 4 such laws. Similarly, we can infer from the histogram that the proposals above 500 comments were per year. This gives us a contribution for each of the ones above 500 at least :

$$\frac{319 \times 500 + (8153 + 4457 + 4 \times 1000) - 71464}{652 \times 71464} \approx 0.0023$$

Summing over those top proposals, we know that the contributions to the Gini coefficient is at least

$$0.0553 + 0.0298 + 4 \times 0.0056 + 6 \times 0.00225 \approx 0.121$$

Let us now consider the first year. As the average is 446 comments per law, considering potential approximations, that means that there were at least 35155 comments. We know that p_1 as well as the four proposals above 1000 votes all happened that year. We know that the Gini is only lowered when we redistribute more equally by giving from some above the average to others below the average. To continue working on a lower bound for the Gini, we can then assume that the pattern for the first year is that one proposal had 8153, four had 1000 and the rest split the remaining votes equally. That would give at least 310 votes for each of the 74 proposals. Making similar assumptions for the other years would make those come immediately after the contributions with 500 votes. We can then compute a lower bound for the total contributions of those 74 proposals :

$$74 \times \frac{313 \times 310 + (8153 + 4457 + 4 \times 1000 + 6 \times 500) - 71464}{652 \times 71464} \approx 0.0717$$

A similar analysis for the second year adds an additional 0.020.

This brings the lower bound for the Gini to 0.213

3 Upper bound

To compute the upper bound, let's consider the situation that maximises the Gini. That would mean as many proposals with 0 comments and as many comments as possible on the others (while keeping with the constraints). We could then compute a very exaggerated distribution where the first year has p_1, then four proposals with 1999 comments, then only proposals with 999 comments. Similarly, the second and third year would have three proposals with 999 comments and the rest at 499. The final year would have 3 proposals at 200 and the rest at 149. We then obtain a strict upper bound of 0.825.

If instead we assume more reasonably that the bounds also apply to the previous years (so year 1 is bounded by 500, and year 2 and 3 by 2000), we instead obtain an upper bound of 0.65.

References

[Mos20] Lorenzo Mosca. Democratic vision and online participatory spaces in the italian movimento 5 stelle. *Acta politica*, 55(1):1–18, 2020.