The modified repeat rate described within a thermodynamic framework

The modified repeat rate described within a thermodynamic framework

----

[1] Gangan Prathap and [2,3] Ronald Rousseau

[1] APJ Abdul Kalam Technological University
Thiruvananthapuram, Kerala 695016, India
gangan_prathap@hotmail.com
ORCID: 0000-0003-1472-6358

[2] KU Leuven, MSI, Facultair Onderzoekscentrum ECOOM,
Naamsestraat 61, 3000 Leuven, Belgium
ronald.rousseau@kuleuven.be &
[3] University of Antwerp, Faculty of Social Sciences,
Middelheimlaan 1, 2020 Antwerp, Belgium
ronald.rousseau@uantwerpen.be
ORCID: 0000-0002-3252-2538

Abstract

Derek J. de Solla Price viewed science as a complex system and anticipated that the science of science can be developed via an analogy to thermodynamics. The main point of this article is to show a direct equivalence between a thermodynamic framework and the classical theory of evenness. It illustrates how thermodynamically inspired terms can lead to the measures used to quantify diversity (or lack thereof), balance, evenness, consistency, or concentration. A real-world example based on intersectional inequalities in science is used as an illustration.

Keywords: science of science, thermodynamic approach, measures of evenness, Simpson's index, weighted data

Introduction

The classical approach to developing a measure for evenness or balance is to use either a heuristic algorithm or apply statistical reasoning based on the Lorenz curve. This leads to measures such as the coefficient of variation or the Gini coefficient. In earlier publications, Prathap (2011a,b, 2014) introduced the idea that from a conceptual point of view, a thermodynamic approach based on first principles as in Gauss' least-squares error treatment of uneven distributions is to be preferred. We consider the thermodynamic approach as the more natural in the sense that fewer new concepts or arbitrary constructions are needed. This rich, natural process includes second-order terms which can be interpreted as energy, exergy, and entropy. Entropy is a measure of disorder and serves as a measure of unevenness or imbalance of a distribution.

Let X be a multiset of N non-negative numbers (not all zero), corresponding to data of nominal categories, such as numbers of publications by a given scientist in a set of N journals. We use the term multiset, and not set, because the same number may occur more than once. We rank these data in any way we like, leading to the array X $=(x_j)_{j=1, \ldots, N}$ of length N. Note that we use the term array and not vector because we consider a vector as an element of a vector space, which we do not have here because, in a vector space, one must be able to multiply with negative numbers.

For such an array Prathap (2011a,b, 2014) introduced a zero-order, a first-order, and a second-order indicator as follows:

$$\sum_{j=1}^{N}\left(x_j^0\right), \quad \sum_{j=1}^{N}\left(x_j^1\right), \quad \sum_{j=1}^{N}\left(x_j^2\right) \qquad (1)$$

The zeroth-order indicator is just the number of elements in the multiset X, namely N; the first-order indicator is the sum of the elements in X, and the second-order indicator is the sum of the squares of the elements in X. If X represents an array of citations received by the publications of an actor (author, research group, institute, country, etc.) over a given period, according to the database used for the study, the zeroth-order indicator is the number of publications of the actor, the first-order indicator is the total sum of received citations, denoted as C, and the second-order is called the energy of the system, here of the actor, denoted as E. Then the classical impact is nothing but C/N.

Prathap, see also (Rousseau et al., 2018, p. 242) further introduced the notions of exergy, denoted as X, entropy, denoted as S, and consistency, denoted as ν (the Greek letter nu). These three notions are defined as follows:

$$X = \frac{C^2}{N}, S = E - X, \nu = \frac{X}{E} \tag{2}$$

In the publication-citation context, consistency can be interpreted as the variation in the number of received citations in the publication portfolio.

Balance or evenness

It is well-known that the evenness of a given multiset of data is best represented by its Lorenz curve, see e.g. (Nijssen et al., 1998). As Lorenz curves of different multisets may intersect, they form a partial, not a total, order. Any function respecting this partial order is an acceptable evenness measure. It has been shown in (Nijssen et al., 1998) that, for variable N, the following measures are acceptable evenness measures.

1) The Gini evenness index, defined as one minus the Gini concentration index and formulated as:

$$G_E(X) = 1 - \frac{1}{2T}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i - x_j\right| \tag{3}$$

where $T = \sum_{j=1}^{N} x_j$. We deliberately use an expression for the Gini index which shows that its calculation doesn't require ranked data.

2) The modified Simpson index, defined as

$$\Lambda(X) = \frac{1}{N\sum_{j=1}^{N} p_j^2}, \text{ where } p_j = \frac{x_j}{\sum_{i=1}^{N} x_i} = \frac{x_j}{T} \tag{4}$$

The modified Simpson index is the reciprocal of the repeat rate (Rousseau, 2018) multiplied by N. For the reasons given in (Rousseau, 2018) we will from now on refer to Λ(X) as the modified repeat rate.

3) The reciprocal of the coefficient of variation, defined as

$$\frac{1}{V(X)} = \frac{\bar{x}}{\sqrt{\frac{1}{N}\left(\sum_{j=1}^{n}\left(x_j-\bar{x}\right)^2\right)}} \tag{5}$$

where $\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$ denotes the arithmetic average of the numbers in X.

4) The adapted entropy or Shannon-Wiener index, defined as

$$H_e(X) = \frac{1}{\ln(N) - \sum_{j=1}^{N}\frac{x_j}{T}ln\left(\frac{x_j}{T}\right)} \tag{6}$$

The Gini index is often used as a measure of evenness, but the other ones are rarely used, certainly in informetric studies.

The modified repeat rate expressed in terms of a thermodynamic framework

In this section, we will show that

$$\Lambda(X) = 1 - \frac{S}{E} = \frac{X}{E} \tag{7}$$

Indeed, $\Lambda(X) = \frac{1}{N\ \sum_{j=1}^{N} p_j^2} = \frac{1}{N\ \sum_{j=1}^{N}\left(\frac{x_j}{\sum_{k=1}^{N} x_k}\right)^2} = \frac{\frac{\left(\sum_{k=1}^{N} x_k\right)^2}{N}}{\sum_{j=1}^{N}\left(x_j\right)^2} = \frac{X}{E} = 1 - \frac{S}{E}$

In words: the modified repeat rate is equal to the ratio of the exergy over the energy, or one minus the ratio of the entropy over the energy. For later use we point out that

$$\Lambda(X) = \frac{\frac{\left(\sum_{j=1}^{N} x_j\right)^2}{N}}{\sum_{j=1}^{N}\left(x_j\right)^2} \tag{8}$$

Within a thermodynamic framework, X/E, being equal to the modified repeat rate is an acceptable evenness measure. Hence S/E is an acceptable inequality (or concentration) measure.

Different distributions with the same modified repeat rate

From equation (8) we see that multisets with the same number of elements, the same sum, and the same sum of squares have the same modified repeat rate. In the appendix, we show that such multisets exist, and even provide a method to construct them. We further show that there are infinitely many couples of multisets with the same modified repeat rate.

The reciprocal of the coefficient of variation too can be written using sums and sums of squares so that the sets constructed above also have the same reciprocals of the coefficient of variation (and of course also the same coefficient of variation).

Weighted data

Besides multisets X and corresponding sequences X = $(x_j)_j$, one may consider a multiset R of weights, associated with each number in the multiset X. We number the weights in R in such a way that the number x_j has weight r_j for j = 1,…, N.

It has been pointed out in (Rousseau, 2001) that such weights occur naturally when one wants to compare a data array with an internal or an external standard. A typical external standard occurs when comparing publications of countries with the countries' population or their number of scientists. An internal standard occurs when one has a two-way classification: then the row sums as well as the column sums can be considered as an internal standard, to be used for comparing evenness.

When the arrays X and R are normalized, this leads to the arrays P and W with coordinates:

$$p_j = \frac{x_j}{\sum_{i=1}^{N} x_i} \text{ and } w_j = \frac{r_j}{\sum_{i=1}^{N} r_i} \quad (9)$$

We re-arrange the arrays P and W such that for the new arrangement

$$\frac{p_1}{w_1} \geq \frac{p_2}{w_2} \geq \cdots \geq \frac{p_N}{w_N} \quad (10)$$

where we have assumed that all components of W are different from zero. Now we can construct the corresponding weighted Lorenz curve. This is the broken line connecting the origin (0,0) to the points with components

$$\left(\sum_{j=1}^{i} w_j \,, \sum_{j=1}^{i} p_j \right)_{i=1,\dots,N} \quad (11)$$

We easily see from its definition that if the normalized weight values are equal (to 1/N) then we obtain the standard Lorenz curve (Lorenz, 1905; Rousseau et al., 2018, p.88).

The ratios (p_j /w_j) are the slopes of the line segments of the weighted Lorenz curve. As these slopes decrease, see equation (10), the curve is concave. If we had ranked the values in (10) in an increasing way, we would have obtained an (equivalent) convex weighted Lorenz curve. Only if all (p_j /w_j)-values are equal to 1, the weighted Lorenz curve coincides with the diagonal of the unit square.

Evenness based on weighted data has been used in (Rousseau et al., 2022). Note that the term evenness in a weighted context refers to a comparison with a standard, whose distribution is followed or not. It can be calculated by the weighted Gini evenness index $G_w(X,R)$:

$$G_w(X,R) = 1 - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|w_i p_j - w_j p_i\right| \tag{12}$$

or by the weighted modified repeat rate

$$\Lambda_w(X,R) = \frac{1}{\sum_{j=1}^{N}\left(\frac{p_j^2}{w_j}\right)} \tag{13}$$

As $G_w(X,R)$ and $\Lambda_w(X,R)$ take values between zero and one, $1 - G_w(X,R) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|w_i p_j - w_j p_i\right|$ and $1 - \Lambda_w(X,R)$ are acceptable measures for weighted inequality.

Note that these measures do not depend on the shape, convex or concave, of the weighted Lorenz curve.

Weighted data in the thermodynamical framework

In this framework, the (weighted) exergy of the whole system (X,R), is equal to $\frac{\left(\sum_{j=1}^{N} x_j\right)^2}{\sum_{j=1}^{N} r_j}$ and the (weighted) energy E is equal to $\sum_{j=1}^{N}\left(\frac{x_j^2}{r_j}\right)$. Finally, the (weighted) entropy S is equal to E-X. We see that if all r_j are equal to 1 the definitions for the weighted case coincide with the ones for the unweighted case.

We now express the weighted modified repeat rate in terms of weighted thermodynamic quantities, showing that

$$\Lambda_w(X,R) = \frac{X}{E} = 1 - \frac{S}{E} \tag{14}$$

Indeed, $\Lambda_w(X,R) = \frac{1}{\sum_{j=1}^{N}\left(\frac{p_j^2}{w_j}\right)} = \frac{1}{\sum_{j=1}^{N}\left(\frac{x_j^2/\left(\sum_{k=1}^{N} x_k\right)^2}{r_j/\left(\sum_{k=1}^{N} r_k\right)}\right)} = \frac{1}{\sum_{j=1}^{N}\left(\frac{x_j^2}{r_j}\right).\frac{\sum_{k=1}^{N} r_k}{\left(\sum_{k=1}^{N} x_k\right)^2}} = \frac{1}{E.\frac{1}{X}} =$

$\frac{X}{E} = 1 - \frac{S}{E}$.

We observe that the weighted modified repeat rate is equal to the ratio of the weighted energy over the weighted energy. In a concentration context *S/E*, the weighted entropy over the weighted energy can be thought of as an alternative to the weighted Gini concentration index. In the next section, we shall take up a real-world example to illustrate the use of these two measures. As this example focuses on relative (with respect to the whole population) inequality in science we will use the concentration (inequality) measures:

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|w_i p_j - w_j p_i\right| \tag{15}$$

namely the weighted Gini concentration index, and

$$\Pi = \frac{S}{E} = 1 - \frac{1}{\sum_{j=1}^{N}\left(\frac{p_j^2}{w_j}\right)} \tag{16}$$

the thermodynamic weighted concentration measure.

A real-world example based on intersectional inequalities in science

We complete our investigation using a recent real-world example based on intersectional inequalities in science (Kozlowski et al., 2022). In this article, it was shown that the US scientific workforce is not representative of the population as a whole and is characterized by an unequal relative dispersion of sub-populations within the scientific community. The authors performed an analysis based on millions of scientific papers, offering real-world data to illustrate our indicators for evenness.

Kozlowski et al. (2022) consider the publication patterns of US-affiliated first authors between 2008 and 2019. Their data are based on 5,431,451 articles indexed in Clarivate Analytics' Web of Science (WOS) and include 1,609,107 distinct US-affiliated first authors. The study is restricted to first authors, on the assumption that "they are generally those who have contributed the most to an article" and "represent the most visible name in bibliographic references" (Kozlowski et al., 2022).

The actual population share in the United States by cohorts based on ethnicity and gender serves as the external standard (see the first two columns of Table 1). The actual share of these cohorts within each scientific community based on discipline or area provides the data array for study (the remaining four columns of Table 1).

Table 1. Data array for five disciplines and eight intersectional cohorts from Kozlowski et al. (2022)

Cohorts	**Population share**	**Soc Sc share**	**Health share**	**Eng & Tech share**	**Nursing share**
Asian-M	2.3	9.4	5.3	33.2	2.1
Asian-W	2.6	5.5	6.8	9.1	7.2
Black-M	6	5.5	3.3	4.6	1.3
Black-W	6.6	3.5	6.4	1	8.7
Latin-M	8.5	3.5	2.1	4.1	1
Latin-W	8.3	2.4	3.9	0.9	4.4
White-M	32.3	43.3	25.1	39.1	9.5
White-W	33.4	26.9	47.2	7.9	65.9
TOTAL	100	100	100.1	99.9	100.1

Thus, in Table 1 we have data for four disciplines and eight cohorts. Due to rounding off errors the numbers do not add up exactly to 100.0 but do nearly so. The four ethnic/racial categories are:

1) Non-Hispanic White Alone (White)
2) Non-Hispanic Black or African American Alone (Black),
3) Non-Hispanic Asian, Native Hawaiian, and Other Pacific Islander Alone (Asian)
4) Hispanic or Latino origin (Latin).

In each case, the (binary) gender categories are men (M) and women (W).

The population share is based on the 2010 US Census. The four columns correspond to the share of the scientific workforce in the Social Sciences (Soc Sc), Health, Engineering & Technology (Eng & Tech), and Nursing as curated from the Web of Science. For precise information about the data used in their – and our - study, we refer to (Kozlowski et al., 2022).

Calculations are performed using formulae (15) and (16). Weighted Gini indices were also checked separately by constructing weighted Lorenz curves (this requires ranking the data). Fig. 2 summarizes the difference between these two main measures of inequality. As there is no such thing as a ground truth in these matters, it is not possible to say that one is better than the other. As in the unweighted case, the difference, and hence preference for one of these measures, is a matter of sensitivity for transfers (Allison, 1978; Egghe and Rousseau, 1990). Without investigating this deeper, we note that we have a preference for the Π-index as the Gini-index (at least in the unweighted case) is known to be dependent on ranks (Egghe and Rousseau, 1990). In Fig. 2 we also indicated the inequality in the cohort of the distinct US-affiliated first authors from the Web of Science (WOS US) and also those in the smaller cohort of US permanent residents from the same database (WoS US Res).

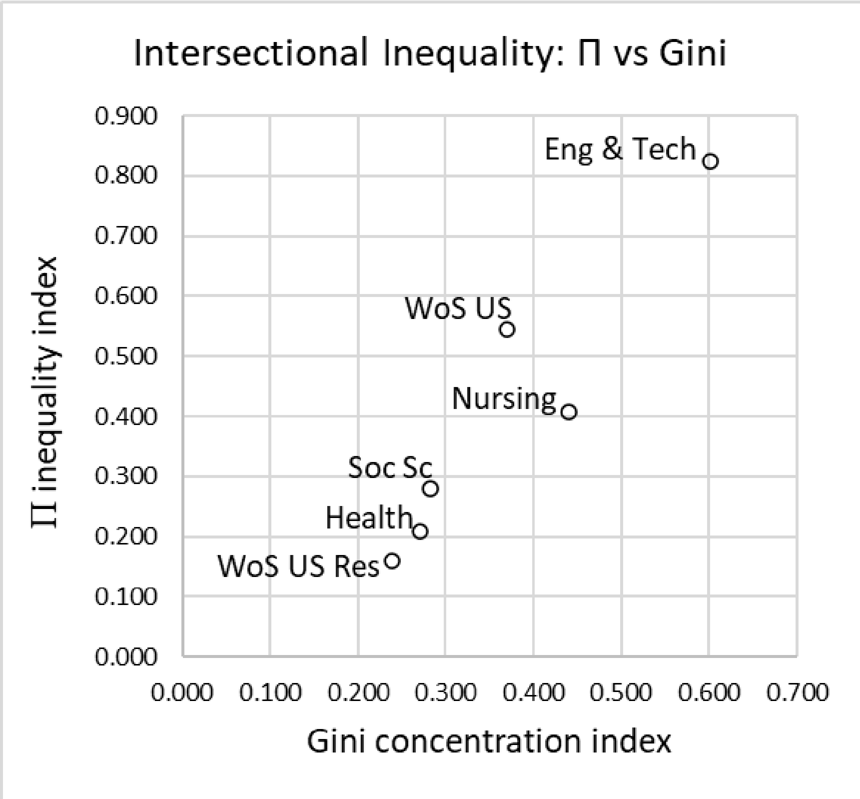

Fig. 2. Difference between the Π-index and the weighted Gini concentration index.

It is clear from Fig. 2 that the highest inequalities are seen in Engineering and Technology where the Asian cohorts have a disproportionate share of the scientific workforce. For this case, Fig. 3 shows the weighted Lorenz curve corresponding to a very high Π-index of 0.824 and a Gini concentration index of 0.620. We also illustrated that there is a considerable gender disparity. We see, for instance, that the intersection of Latin and Women is far more disadvantaged than the intersection of Asian and Men.

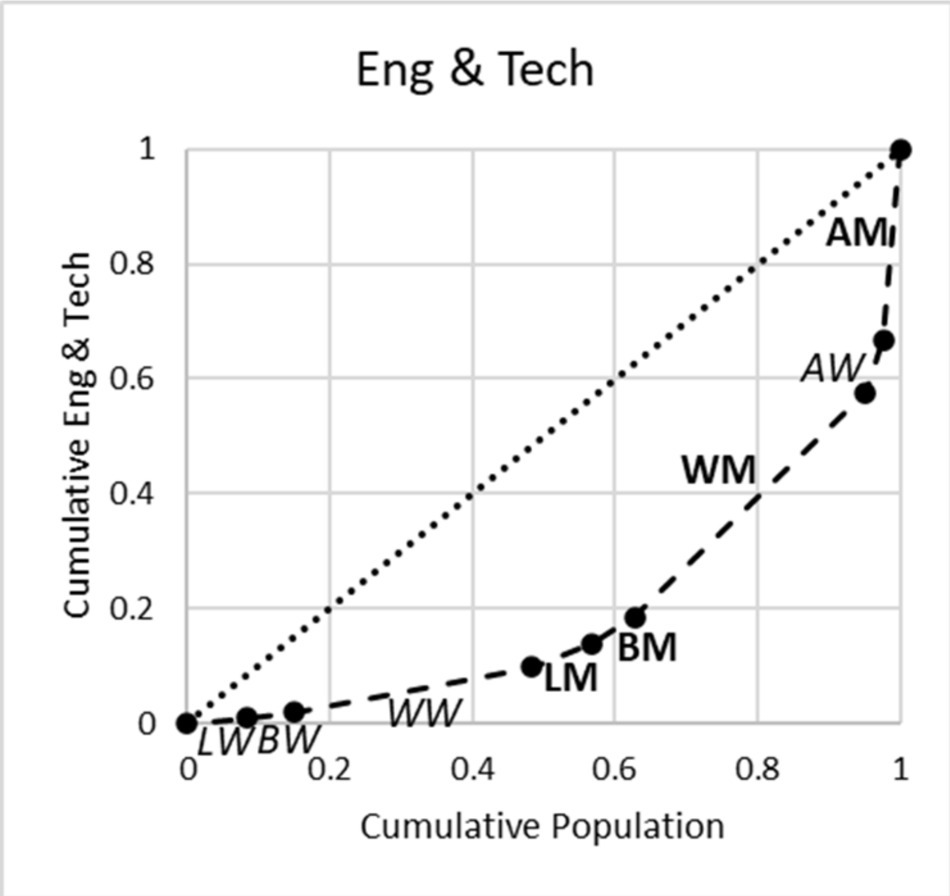

Fig. 3. The weighted Lorenz curve for Engineering &Technology shows the dominance of the Asian cohorts in the scientific workforce. This is reflected in a Π-index of 0.824 and a Gini concentration index of 0.620.

Concluding remarks

Regarding the thermodynamic approach, we recall that Derek J. de Solla Price viewed science as a complex system and developed the science of science via an analogy to thermodynamics (Price, 1963). More concretely, he compared science to a gas with individual molecules (scientists) possessing velocities and interactions, exhibiting general properties. In his own words: *“One does not fix one's gaze on a specific molecule called George, traveling at a specific velocity and being in a specific place at some given instant; one considers only an average of the total assemblage in which some molecules are faster than others,* *and in which they are spaced out randomly and moving in different directions."* This averaging approach using Gauss' least squares error principle leads to a natural definition of the second-order terms X, E, and S.

In this article, we showed a direct relationship between the thermodynamic framework and the classical theory of evenness as measured using the modified repeat rate and its weighted form. This does not seem to be possible with the Gini evenness index.

As an extra, we recalled (see appendix) Sridhar Ramesh' construction of arrays with the same sum and the same sum of squares.

We completed our study with a recent real-world example based on intersectional inequalities in science (Kozlowski et al. 2022). The highest inequalities are seen in Engineering and Technology where the Asian cohorts have a disproportionate share of the scientific workforce. We also noted that there is a considerable gender disparity.

Conflict of interest. The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Ronald Rousseau is a member of the Distinguished Reviewer Board of *Scientometrics*.

References

Allison, P. (1978). Measures of inequality. *American Sociological Review*, *43*(6), 865-880.

Egghe, L., and Rousseau, R. (1990). Elements of concentration theory. In: *Informetrics 89/90* (L. Egghe & R. Rousseau, eds.) (pp. 97-137). Amsterdam: Elsevier.

Kozlowski, D., Larivière, V., Sugimoto, C. R. and Monroe-White, T. (2022). Intersectional inequalities in science. *Proceedings of the National Academy of Sciences, 119*(2), e2113067119. https://doi.org/10.1073/pnas.2113067119

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209-219.

Nijssen, D., Rousseau, R. and Van Hecke, P. (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses*, *13*(1), 33-38.

Prathap, G. (2011a). A thermodynamic explanation for the Glänzel-Schubert model for the h-index. *Journal of the American Society for Information Science and Technology*, *62*(5), 992-994.

Prathap, G. (2011b). The energy-exergy-entropy (or EEE) sequences in bibliometric assessment. *Scientometrics*, *87*(3), 515-524.

Prathap, G. (2014). Quantity, quality, and consistency as bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(1), 214-214.

Price, D. J. dS. (1963). *Little Science, Big Science*. New York: Columbia University Press. ISBN 0-231-08562-1.

Ramesh, S. (n.d.) What are two lists of integers having the same sum, same sum of squares, and same sum of cubes? (https://www.quora.com/What-are-two-lists-of-integers-having-the-same-sum-same-sum-of-squares-and-same-sum-of-cubes)

Rousseau, R. (2001). Concentration and evenness measures as macro-level scientometric indicators (in Chinese). In: *Research and university evaluation* (Keyan pingjia yu daxue pingjia), Jiang Guo-hua (ed.), 2001, Red Flag Publishing House, Beijing, p.72-89. An English version is available from the author.

Rousseau, R. (2018). The repeat rate: from Hirschman to Stirling. *Scientometrics*, *116*(1), 645–653.

Rousseau, R., Egghe, L., & Guns, R. (2018). *Becoming metric-wise. A bibliometric guide for researchers.* Kidlington (Chandos-Elsevier).

Rousseau, R., Zhang, L., & Sivertsen G. (2023). Using the weighted Lorenz curve to represent balance in collaborations: the BIC indicator. *Scientometrics*, *128*(1), 609-622. https://doi.org/10.1007/s11192-022-04533-0

Appendix.

Construction of different distributions with the same modified repeat rate

There exists an interesting way of constructing multisets with the same modified repeat rate. We illustrate this in the case of four numbers (N=4). The method explained here is due to Sridhar Ramesh (no date available).

Consider a three-dimensional cube with vertices a, b, c, and d on the upper facet and e, f, g, and h on the lower facet. Vertices are named counterclockwise and vertex e is situated below vertex a (see Fig.1).

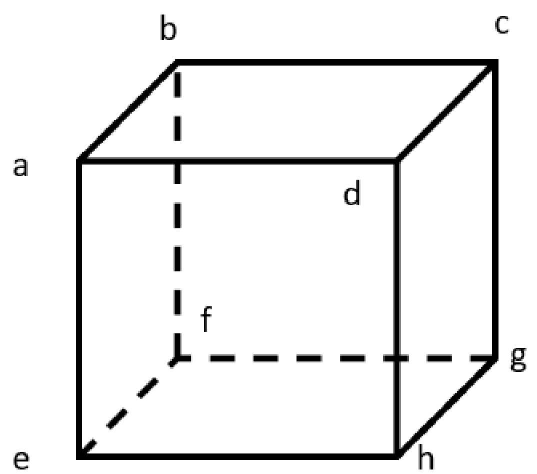

Fig. 1 Cube used to perform Ramesh's algorithm

In this theory vertices a, c, f, and h are called the white vertices, while vertices b, d, e, and g are called the black vertices. A cube has six facets, here bounded by a,b,c,d; a,b,f,e; b,c,g,f; c,d,h,b; a,d,h,e and finally by e,f,g, and h. Now assign to each facet a non-negative number and finally, assign to each vertex the sum of the numbers of the three facets adjacent to it. This number is called a V-number. The multiset of the four V-numbers of the white vertices has the same sum and the same sum of squares as the multiset of the four V-numbers of the black vertices, and hence they have the same modified repeat rate. A simple example is given in Table 2.

Table 2. An example of Ramesh's algorithm,

2a: facets

facet	number	number
abcd	3	u
abfe	4	v
bcgf	6	w
cdhg	2	x
adhe	5	y
efgh	1	z

2b: vertices

vertex	V-number	V-number
a	3+4+5=12	u+v+y
b	3+4+6=13	u+v+w
c	3+6+2=11	u+w+x
d	3+2+5=10	u+x+y
e	4+5+1=10	v+y+z
f	4+6+1=11	v+w+z
g	6+2+1=9	w+x+z
h	2+5+1=8	x+y+z

The multisets {12,11,11,8} and {13,10,10,9} have the same sum, namely 42, and the same sum of squares, namely 450. Proof of this property can be given based on symmetry properties of the assigned numbers, but we provide a very elementary proof by just doing the calculations for an abstract assignment of numbers to facets, as shown in the third column of Table 2a and of Table 2b.

We see that the sum of the white V-numbers is: (u+v+y)+(u+w+x) + (v+w+z) + (x+y+z) = 2(u+v+w+x+y+z), which is equal to the sum of the black numbers: (u+v+w)+(u+x+y)+(v+y+z)+(w+x+z). Similarly, one can easily check that $(u+v+y)^2+(u+w+x)^2 + (v+w+z)^2 + (x+y+z)^2 = (u+v+w)^2 +(u+x+y)^2 +(v+y+z)^2 +(w+x+z)^2$.

The previous construction deals with the case N = 4. Performing this construction two, three, or more times yields the cases N=8,12, etc. (take the union of the obtained multisets). Taking two sets with 4 numbers and appending the same number to both yields an example with N=5; appending two or three numbers to both multisets (the same numbers) yields examples for the cases N=6 and N=7. The construction performed on a cube and also be done on a square yielding the case N=2, such as {3,7} and {5,5}.

As, for N=4, the sum of the white V-numbers is equal to the sum of the black V-numbers, and is twice the sum of the freely chosen numbers u,v,w,x,y,z one can form as many couples of multisets with the same modified repeat rate as there are numbers with the same sum. This reasoning also applies to other values of N.