A General Theory of IR Evaluation Measures

Marco Ferrante, Nicola Ferro and Silvia Pontarollo

Abstract—Interval scales are assumed by several basic descriptive statistics, such as mean and variance, and by many statistical significance tests which are daily used in IR to compare systems. Unfortunately, so far, there has not been any systematic and formal study to discover the actual scale properties of IR measures. Therefore, in this paper, we develop a theory of IR evaluation measures, based on the representational theory of measurements, to determine whether and when IR measures are interval scales.

We found that common set-based retrieval measures – namely Precision, Recall, and F-measure – always are interval scales in the case of binary relevance while this happens also in the case of multi-graded relevance only when the relevance degrees themselves are on a ratio scale and we define a specific partial order among systems.

In the case of rank-based retrieval measures – namely AP, gRBP, DCG, and ERR – only gRPB is an interval scale when we choose a specific value of the parameter p and define a specific total order among systems while all the other IR measures are not interval scales.

Besides the formal framework itself and the proof of the scale properties of several commonly used IR measures, the paper also defines some brand new set-based and rank-based IR evaluation measures which ensure to be interval scales.

Index Terms—representational theory of measurement, interval scale, IR evaluation measure, formal framework

1 Introduction

Information Retrieval (IR) is concerned with ranking document with respect to their degree of *relevance* to user information needs, typically expressed as vague and imprecise queries. IR systems are assessed by their *effectiveness*, i.e. retrieving and better ranking relevant documents while eliminating not relevant and noisy ones. *Experimental evaluation* is the way to measure the performance of IR systems from the effectiveness point of view [1] and most of our understanding of how IR systems work comes from research in the experimental evaluation area.

Experimental evaluation is based on the Cranfield paradigm [2] which makes use of experimental collections (D, T, GT) where: a corpus of documents D represents the domain of interest; a set of topics T represents the user information needs; and, human-made relevance judgements or ground-truth GT are the "correct" answers determining, for each topic, the relevant documents. The ranked result lists, i.e. the IR system outputs, are then scored with respect to the ground-truth using several *evaluation measures*, which aim at quantifying the effectiveness of a system.

Even if experimental evaluation is a main driver of progress in IR and evaluation measures lay at its foundations, our theoretical understanding of what evaluation measures are is still very limited, despite the several studies both in the past [3], [4], [5] and more recently [6], [7], [8].

In particular, *measurement scales* play a central role [9], [10] since they determine the operations that can be performed with the measured values and, as a consequence, the statistical analyses that can be applied. Stevens [10] identifies four major types of scales with increasing properties: (i) the *nominal scale* consists of discrete unordered values, i.e. categories; (ii) the *ordinal scale* introduces a natural order among the values; (iii) the *interval scale* preserves the equality of intervals or differences; and (iv) the *ratio scale* preserves the equality of ratios.

In experimental evaluation we daily perform operations, such as computing means and variances, which are also the basic "ingredients" of the more sophisticated statistical significance tests we use to compare IR systems and assess their differences [11]. However, all these operations can be performed only from interval scales onwards but, due to our limited knowledge of evaluation measures, we do not actually know which scales they rely on. For example, Robertson [12] claims that the assumption of *Average Precision (AP)* being an interval scale is somehow arbitrary.

This paper sets a theory of IR evaluation measures to formally investigate their properties and to study whether and when they use an interval scale. We frame our work within the *representational theory of measurement* [9], which is the measurement theory adopted in both physical and social sciences. In particular, we develop a fully comprehensive framework, comprising both *set-based* measures – namely Precision, Recall and F-measure – dealing with unordered result lists, and *rank-based* measures – namely, *Average Precision (AP)*, *Discounted Cumulated Gain (DCG)*, *Rank-Biased Precision (RBP)*, and *Expected Reciprocal Rank (ERR)* – dealing with ranked result lists. Moreover, we consider both *binary relevance*, i.e. when documents can be just relevant or not relevant, and *multi-graded relevance* [13], i.e. when documents can have different degrees of relevance, such as not relevant, partially relevant, and highly relevant.

We found that in the case of set-based IR measures, Precision, Recall, and F-measure are always interval scales when using binary relevance while this holds in the multi-graded case only when the relevance degrees are a ratio

- *M. Ferrante and S. Pontarollo are with Department of Mathematics, University of Padua, Italy E-mail: {ferrante, spontaro}@math.unipd.it*
- *N. Ferro is with Department of Information Engineering, University of Padua, Italy E-mail: ferro@dei.unipd.it*

Manuscript received January 16, 2018; revised April 10, 2018; accepted May 10, 2018

scale. In the case of rank-based IR measures RBP is an interval scale for specific values of its parameter while all the other measures – namely AP, DCG, and ERR – are not.

Moreover, as an additional significant contribution, we derived from our theory new evaluation measures which guarantee to be interval scales in all the examined cases – see eq. (3), (5), (7), and (9).

Finally, the impact of this paper goes well beyond IR itself and reaches out to neighboring fields, such as databases and data mining, where many of the studied measures are widely adopted in benchmarking and it is equally crucial to know which scales they use as well as which descriptive statistics and which statistical analyses are permissible.

The paper is organized as follows. Section 2 briefly recalls some definitions and properties of posets and Hasse diagrams, which are the basic algebraic tools we rely on to build our theory. Section 3 introduces the basic concepts about the representational theory of measurement and how to determine if a measure is an interval scale. Section 4 introduces the basic formalism adopted in our framework to model the Cranfield paradigm and the IR evaluation measures. Section 5 analyses set-based evaluation measures while Section 6 deals with rank-based evaluation measures, in both cases considering binary and multi-graded relevance judgements. Section 7 reports some related works. Finally, Section 8 wraps up the discussion and outlooks some future work. The Electronic Appendix reports explanations, examples, and proofs, which do not fit here for space reasons.

2 Poset, Lattice, and Hasse Diagram

A partially ordered set P, **poset** for short, is a set with a partial order $\preceq$ defined on it [14]. A **partial order** $\preceq$ is a binary relation over P which is reflexive, antisymmetric and transitive. Given $s, t \in P$, we say that s and t are *comparable* if $s \preceq t$ or $t \preceq s$, otherwise they are *incomparable*.

A closed **interval** is a subset of P defined as $[s,t] := \{u \in P : s \preceq u \preceq t\}$, where $s, t \in P$ and $s \preceq t$. Moreover, we say that t **covers** s if $s \preceq t$ and $[s,t] = \{s,t\}$, that is there does not exist $u \in P$ such that $s \prec u \prec t$.

We can represent a finite poset P by using the **Hasse diagram** which is a graph where vertices are elements of P, edges represent *covers* relations, and if $s \prec t$ then s is below t in the diagram. Note that if $s, t \in P$ lie on the same horizontal level of the diagram, then they are incomparable by construction.

A subset C of a poset P is a **chain** if any two elements of C are comparable: a chain is a totally ordered subset of a poset. If C is a finite chain, the **length** of C, $\ell(C)$, is defined by $\ell(C) = |C| - 1$. A **maximal chain** of P is a chain that is not a proper subset of any other chain of P.

If every maximal chain of P has the same length n, we say that P is **graded of rank n**; in particular there exists a unique function $\rho : P \to \{0, 1, \ldots, n\}$, called the **rank function**, such that $\rho(s) = 0$, if s is a minimal element of P, and $\rho(t) = \rho(s) + 1$, if t covers s.

Finally, since any interval on a graded poset is graded, the **length of an interval** $[s,t]$ is given by $\ell(s,t) := \ell([s,t]) = \rho(t) - \rho(s)$.

Foldes [15] proves that in a graded poset P the length $\ell(\cdot,\cdot)$ of any interval, also called the **natural distance**, equals the length of the shortest path connecting the two endpoints of the interval in its Hasse diagram.

Please, see the Electronic Appendix for a more detailed description of posets, Hasse diagrams, and their properties.

3 Measurement Theory

3.1 Representational Theory of Measurement

The *representational theory of measurement* [9] sees measurement as the process of assigning numbers to entities in the real world conforming to some property under examination. According to this framework, the key point is to understand how real world objects are related to each other since measure properties are then derived from these relations.

Moving to the IR context, being an interval scale is not just a numeric property of an evaluation measure but firstly we need to understand how system runs are ordered, then what *intervals* of system runs are, and finally how these intervals are ordered too. Only at this point, we can verify whether an evaluation measure complies with these notions and determine whether it is an interval scale or not.

More precisely, a **relational structure** [9], [16] is an ordered pair $\mathbf{X} = \langle X, R_X \rangle$ of a domain set X and a set of relations R_X on X, where the relations in R_X may have different arities, i.e. they can be unary, binary, ternary relations and so on. Given two relational structures $\mathbf{X}$ and $\mathbf{Y}$, a *homomorphism* $\mathbf{M} : \mathbf{X} \to \mathbf{Y}$ from $\mathbf{X}$ to $\mathbf{Y}$ is a mapping $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ where: (i) M is a function that maps X into $\mathrm{M}(X) \subseteq Y$, i.e., for each element of the domain set there exists one corresponding image element; (ii) M_R is a function that maps R_X into $\mathrm{M}_R(R_X) \subseteq R_Y$ such that $\forall r \in R_X$, r and $\mathrm{M}_R(r)$ have the same arity, i.e., for each relation on the domain set there exists one (and it is usually, and often implicitly, assumed: and only one) corresponding image relation; (iii) $\forall r \in R_X, \forall x_i \in X$, if $r(x_1, \ldots, x_n)$ then $\mathrm{M}_R(r)\Big(\mathrm{M}(x_1), \ldots, \mathrm{M}(x_n)\Big)$, i.e., if a relation holds for some elements of the domain set then the image relation must hold for the image elements.

A relational structure $\mathbf{E}$ is called *empirical* if its domain set E spans over entities in the real world, i.e. system runs in our case; a relational structure $\mathbf{S}$ is called *symbolic* if its domain set S spans over a given set of numbers. A **measurement (scale)** is the homomorphism $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ from the real world to the symbolic world and a **measure** is the number assigned to an entity by this mapping[1].

3.2 Measurement Scales

There are four major types of measurement scales [10] which can be ordered by their increasing properties and allows for different computations: *nominal scales* allow us to compute the number of cases and the mode; in addition, *ordinal scales* allow us to compute median and percentiles; *interval scales* add the possibility to compute mean, variance, product-moment correlation and rank correlation; finally, *ratio scales*

[1] Note that an evaluation measure like AP should be called a measurement according to this terminology, since AP is the process by which numbers are assigned to system runs; on the other hand, the actual value of AP assigned to a system should be called measure. However, in the rest of the paper, we will keep using the term measure instead of measurement, since this is what is commonly used and understood in the IR field.

add the capability to compute the coefficient of variation. Over the years, there has been debate [17] on whether these rules are too strict or not but they are applied widely.

If we already know that on an empirical structure there is an interval scale M, the uniqueness theorem – see e.g. Theorem 3.18 in [16] – ensures that any other measurement M' on that structure is a linear positive transformation of M, that is $\mathrm{M}' = \alpha\mathrm{M} + \beta,\ \alpha, \beta \in \mathbb{R}$ and $\alpha > 0$.

However, in the case of IR evaluation measures, we lack a known interval scale M which we can use to compare all the other IR measures against. Actually, the core issue is even more severe: we lack any notion of order on the empirical set E of the IR system runs, thus we also lack the notion of interval of system runs and, consequently, we cannot define an interval scale M too. This issue was somehow early pointed out by van Rijsbergen [18]:

> In the physical sciences there is usually an empirical ordering of the quantities we wish to measure [...] Such a situation does not hold for information retrieval. There is no empirical ordering for retrieval effectiveness and therefore any measure of retrieval effectiveness will by necessity be artificial.

In this paper, we overcome these issues and, following [9], [16], we rely on the notion of *difference structure* to introduce a definition of interval among system runs and to ensures the existence of an interval scale.

Given E, a *weakly ordered* empirical structure is a pair $(E, \preceq)$ where, for every $a, b, c \in E$,

- $a \preceq b$ or $b \preceq a$;
- $a \preceq b$ and $b \preceq c \ \Rightarrow\ a \preceq c$ (transitivity).

Note that if $a, b \in E$ are such that $a \preceq b$ and $b \preceq a$, then we write $a \sim b$ and we say that a and b are equivalent elements of E for $\preceq$. This does not necessarily mean that a and b are equal, i.e. $a = b$, since they might be two distinct objects. When the antisymmetric relation holds, that is when $a \preceq b$ and $b \preceq a$ implies that a and b are the same element (namely $a = b$), we talk about a *total order*.

An interval on the empirical structure is an element $(a, b) \in E \times E$ and we introduce a notion of **difference** Δ_{ab} over intervals, to act as a signed distance we exploit to compare intervals. Once we have a notion of difference Δ_{ab}, we can define a weak order $\preceq_d$ between the Δ_{ab} differences and, consequently, among intervals. We can proceed as follows: if two elements $a, b \in E$ are such that $a \sim b$, then the interval $[a, b]$ is null and, consequently, we set $\Delta_{ab} \sim_d \Delta_{ba}$; if $a \prec b$ we agree upon choosing $\Delta_{aa} \prec_d \Delta_{ab}$ which, in turn implies that $\Delta_{aa} \succ_d \Delta_{ba}$, that is there exist a kind of "zero" and the inverse with respect to this "zero".

Note that the symbols $\sim$ and $\preceq$ indicate equivalence and ordering among objects while the symbols $\sim_d$ and $\preceq_d$ indicate equivalence and ordering among intervals of objects, as highlighted by the subscript d standing for the difference Δ_{ab} which induces such ordering.

The following notion of difference structure allows us to verify whether a measurement is an interval scale or not.

Definition 1. Let E be a finite (not empty) set of objects. Let $\preceq_d$ be a binary relation on $E \times E$ that satisfies, for each $a, b, c, d, a', b', c' \in E$, the following axioms:

i. $\preceq_d$ is *weak order*;
ii. if $\Delta_{ab} \preceq_d \Delta_{cd}$, then $\Delta_{dc} \preceq_d \Delta_{ba}$;
iii. *Weak Monotonicity:* if $\Delta_{ab} \preceq_d \Delta_{a'b'}$ and $\Delta_{bc} \preceq_d \Delta_{b'c'}$ then $\Delta_{ac} \preceq_d \Delta_{a'c'}$;
iv. *Solvability Condition*: if $\Delta_{aa} \preceq_d \Delta_{cd} \preceq_d \Delta_{ab}$, then there exists $d', d'' \in R$ such that $\Delta_{ad'} \sim_d \Delta_{cd} \sim_d \Delta_{d''b}$.

Then $(E, \preceq_d)$ is a **difference structure**.

The first condition defines an ordering among intervals while the second one sets a sign for differences. The *Weak Monotonicity* condition gives us a rule to compose adjacent intervals; among other things, it tells us that adding a non-null interval to an interval produces a greater interval. The *Solvability Condition* ensures the existence of an equally spaced gradation between the elements of E, indispensable to construct an interval scale measurement.

The *representation theorem* for difference structures states:

Theorem 1. Let E be a finite (not empty) set of objects and let $(E, \preceq_d)$ be a difference structure. Then, there exist a interval scale measurement $\mathrm{M} : E \to \mathbb{R}$ such that for every $a, b, c, d \in E$

$$\Delta_{ab} \preceq_d \Delta_{cd} \Leftrightarrow \mathrm{M}(b) - \mathrm{M}(a) \leq \mathrm{M}(d) - \mathrm{M}(c)\ .$$

This theorem ensures us that, if there is a difference structure on the empirical set E, then there exists an interval scale M over it.

Please, see the Electronic Appendix for a more detailed description of the representational theory of measurement.

3.3 Overall Approach

The procedure we adopt to study whether IR measures are interval scales or not is as follows:

- we define an ordering among system runs by creating a poset P, which allows us also to introduce a notion of interval among runs;
- if the poset P is graded of rank n, then there exists a unique rank function ρ which assigns a natural number to each run; we have to construct such rank function ρ;
- we define the length of an interval as the natural distance $\Delta_{ab} := \ell(a, b) := \ell([a, b]) = \rho(b) - \rho(a)$, which also corresponds to the length of a maximal chain in $[a, b]$ minus 1. Moreover, the natural length of any interval of a graded poset equals to the number of edges in every shortest path connecting the endpoints of the interval in its Hasse diagram;
- we can check whether the poset P with the above natural length is a difference structure or not;
- in case we have a difference structure, we can define an interval scale M as the rank function ρ itself;
- we can finally check whether IR measures are a linear positive transformation of this interval scale M and determine whether they are an interval scale or not.

Since any notion of interval depends on a notion of order among systems, we investigate two alternatives for defining the poset P by introducing two ways of ordering system runs, namely a total ordering and a partial order. Moreover, as a litmus test, we also consider the ordering of system runs induced by the evaluation measures themselves, considering this as the ultimate chance for them to show to be an interval scale.

Below, we provide an introductory example on how this procedure works by using natural numbers.

Example 1.

Let B be the set of natural numbers from 0 to $N \in \mathbb{N}$, that is $B = \{0, \dots, N\}$, ordered by the *less than or equal to* relation $\leq$. B is a totally ordered set and, as poset, is graded of rank N. Note that, given $a, b \in B$, a covers b if $b = a - 1$. Therefore, the rank function ρ is defined as $\rho(a) = a$ for any $a \in B$, since $\rho(0) = 0$ and $\rho(a) = \rho(a-1) + 1$. For $a, b \in B$ such that $a \leq b$, the *natural distance* is then given by $\ell(a, b) = \rho(b) - \rho(a)$.

We can define the difference as equal to the *natural distance*, that is $\Delta_{ab} = \ell(a, b)$ if $a \leq b$, otherwise $\Delta_{ab} = -\ell(b, a)$. Therefore $\Delta_{ab} = b - a$ for any $a, b \in B$. Since $\Delta_{a,b} \in \mathbb{Z}$ for any $a, b \in B$ by contruction, we choose $\preceq_d$ to be the *less than or equal to* relation defined between natural numbers. Therefore, for any $a, b, c, d \in B$:

$$\Delta_{ab} \preceq_d \Delta_{cd} \;\Leftrightarrow\; b - a \leq d - c \; .$$

Let us prove that $(B, \preceq_d)$ is a difference structure. Let $a, a', b, b', c, c', d \in B$:

i. $\preceq_d$ is a weak order, due to the similar property satisfied by $\leq$ on B. Note that $\preceq_d$ is not a total order, since $\Delta_{ab} \preceq_d \Delta_{cd}$ and $\Delta_{cd} \preceq_d \Delta_{ab}$ imply that the two intervals have the same length and not that they coincide.

ii. $\Delta_{ab} \preceq_d \Delta_{cd} \;\Leftrightarrow\; b - a \leq d - c \;\Leftrightarrow\; c - d \leq a - b \;\Leftrightarrow\; \Delta_{dc} \preceq_d \Delta_{ba}$.

iii. $\Delta_{ab} \preceq_d \Delta_{a'b'}$ and $\Delta_{bc} \preceq_d \Delta_{b'c'} \;\Leftrightarrow\; b-a \leq b'-a'$ and $c-b \leq c'-b' \;\Leftrightarrow\; c-a = c-b+b-a \leq c'-b'+b'-a' = c'-a' \;\Leftrightarrow\; \Delta_{ac} \preceq_d \Delta_{a'c'}$.

iv. Let a, b, c, d be such that $\Delta_{aa} \preceq_d \Delta_{cd} \preceq_d \Delta_{ab}$, then $0 = a - a \leq d - c \leq b - a$. To prove the *Solvability Condition* we need to show that there exist d', $d'' \in B$ such that $\Delta_{ad'} \sim_d \Delta_{cd} \sim_d \Delta_{d''b}$, that is $d' - a = d - c = b - d''$. Clearly such d' and d'' have to be such that $d' = a+d-c$ while $d'' = b+c-d$; moreover, from $0 \leq d-c \leq b-a$, it follows that $N \geq b \geq a+d-c \geq a \geq 0$ and that $0 \leq a \leq b-(d-c) \leq b \leq N$; therefore, d', $d'' \in B$ and the *Solvability Condition* is satisfied.

To further understand the indispensable *Solvability Condition* and the equi-spacing among intervals, let for example a, b, c, d be respectively equal to $1, 6, 7, 9$. $\Delta_{11} \preceq_d \Delta_{79} \preceq_d \Delta_{16}$ since $\Delta_{11} = 1 - 1 = 0$, $\Delta_{79} = 9 - 7 = 2$ and $\Delta_{16} = 6 - 1 = 5$: then $d' = 3 = a + d - c$ and $d'' = 4 = b+c-d$ are such that $\Delta_{13} \sim_d \Delta_{79} \sim_d \Delta_{46} = 2$. Finally, a measurement M given by the rank function, that is $M(a) = \rho(a) = a$, for any $a \in B$ is an interval scale. Indeed, for any $a, b, c, d \in B$:

$$\begin{aligned} \Delta_{ab} \preceq_d \Delta_{cd} &\Leftrightarrow b - a \leq d - c \\ &\Leftrightarrow M(b) - M(a) \leq M(d) - M(c) \; , \end{aligned}$$

as Theorem 1 requires.

4 Basic Formalism

Let $(REL, \preceq)$ be a finite and totally ordered set of **relevance degrees**. We set $REL = \{\mathscr{a}_0, \mathscr{a}_1, \dots, \mathscr{a}_c\}$ with $\mathscr{a}_i \prec \mathscr{a}_{i+1}$ for all $i \in \{0, \dots, c-1\}$; REL has a minimum $\mathscr{a}_0$, called the "not relevant" relevance degree.

Let us consider a finite set of **documents** D and a set of **topics** T. For each pair $(t, d) \in T \times D$, the **ground-truth** GT is a map which assigns a relevance degree $rel \in REL$ to a document d with respect to a topic t.

Let N be a positive natural number called the *length of the run*. We assume that all the runs have same length N, since this is what typically happens in real evaluation settings when you compare IR systems.

We define $D(N)$ as the **set of all the possible** N **retrieved documents**.

A **run** $r : T \to D(N)$ retrieves N documents belonging to $D(N)$ in response to a topic $t \in T$.

Let $R(N)$ be the **set of** N **judged documents**, that is the set of all the N possible combinations of relevance degrees.

We call **judged run** of length N the function $\hat{r}$ from $T \times D(N)$ into $R(N)$ which assigns a relevance degree to each retrieved document, i.e. a judged run $\hat{r}$ is the application of the ground-truth GT function to each element of the run r.

We define the **gain function** $g : REL \to \mathbb{R}_+$ as the map that assigns a positive real number to any relevance degree. We set, without loss of generality, $g(\mathscr{a}_0) = 0$ and we require g to be strictly increasing.

We define the **indicator function** for the relevance degrees as $\delta_{\mathscr{a}}(\mathscr{a}_j) = j \quad \forall j \in \{0, \dots, c\}$. Note that $\delta_{\mathscr{a}}$ is a particular gain function.

Given the gain function g, the **recall base** $RB : T \to \mathbb{R}_+$ is the map defined as $RB(t) = \sum_{j=1}^{|D|} g(GT(t, d_j))$. In the binary relevance case when $c = 1$ and $REL = \{\mathscr{a}_0, \mathscr{a}_1\}$, the gain function usually is $g(\mathscr{a}_1) = \delta_{\mathscr{a}}(\mathscr{a}_1) = 1$ and RB counts the total number of relevant documents for a topic.

An **evaluation measure** is a function $\mathrm{M} : R(N) \to \mathbb{R}_+$ which maps a judged run $\hat{r}$ into a positive real number which quantifies its effectiveness. Note that most of the evaluation measures are normalized and thus the co-domain is the $[0, 1]$ interval.

In the following, we specialize the above definitions to the case of both set-based and rank-based retrieval.

4.1 Set-Based Retrieval

The set of all the possible unordered N retrieved documents is $D(N) = \big\{\{d_1, \dots, d_N\} : d_i \in D\big\}$. A **run** r is given by $r = \{d_1, \dots, d_N\}$. We denote by r_j the j-th element of the set r, i.e. $r_j = d_j$.

A *multiset* (or bag) is a set which may contain the same element several times [19]. The **set of judged documents** is a multiset $(REL, \, m) = \{\mathscr{a}_1, \mathscr{a}_1, \mathscr{a}_0, \dots, \mathscr{a}_c, \mathscr{a}_2, \mathscr{a}_c, \dots\}$, where m is a function from REL into $\mathbb{N}_0$ representing the multiplicity of every relevance degree $\mathscr{a}_j$ [20]; if the multiplicity is 0, a given relevance degree is not present in the multiset. Let $\mathscr{M}$ be the set of all the possible multiplicity functions m, then $REL(\mathscr{M}) := \bigcup_{m \in \mathscr{M}} (REL, m)$ is the **universe set of judged documents**, i.e. the set of all the possible sets of judged documents (REL, m). We can define the set of N judged documents as $R(N) := \{\hat{r} \in REL(\mathscr{M}) : |\hat{r}| = N\}$.

Note that, since each judged run in $R(N)$ is an unordered set of N relevance degrees, $R(N)$ consists of all the N combinations of $c+1 = |REL|$ objects with repetition.

For each run $\hat{r} \in R(N)$, we use the convention to represent it as $\{\hat{r}_1, \dots, \hat{r}_N\}$ where $\hat{r}_i \succeq \hat{r}_{i+1}$, for any $i \in \{1, \dots, N-1\}$. In other terms, among all the possible ways of "displaying" it, we choose the one where the relevance degrees are listed in decreasing order. This choice does not affect the generality of the proposed framework but it just makes proofs a bit easier to follow.

We now introduce the definitions of generalized precision and recall [13], which extend precision and recall to the multi-graded relevance case, and of F-measure.

4.1.1 Generalized Precision (gP)

$$\text{gP}(\hat{r}) = \frac{1}{N} \sum_{i=1}^{N} \frac{g(\hat{r}_i)}{g(\mathcal{a}_c)} \ ,$$

where the factor $1/g(\mathcal{a}_c)$ is needed in order to normalize the gain function to one. Note that gP coincides with *Precision (P)* when binary relevance ($c = 1$) is considered.

4.1.2 Generalized Recall (gR)

$$\text{gR}(\hat{r}) = \frac{1}{RB} \sum_{i=1}^{N} \frac{g(\hat{r}_i)}{g(\mathcal{a}_c)} \ ,$$

where RB is recall base. gR coincides with *Recall (R)* when binary relevance ($c = 1$) is considered.

4.1.3 F-Measure

The *F-measure* is a binary measure, i.e. it works with binary relevance when $REL = \{\mathcal{a}_0, \mathcal{a}_1\}$, and is the harmonic mean of *Precision (P)* and *Recall (R)* given by

$$\text{F}(\hat{r}) = 2 \frac{\text{P}(\hat{r}) \cdot \text{R}(\hat{r})}{\text{P}(\hat{r}) + \text{R}(\hat{r})} \ .$$

To the best of our knowledge, its extension to multi-graded case is not usually considered in the literature but it comes naturally if you use gP and gR.

4.2 Rank-Based Retrieval

The set of all the possible ordered list of N retrieved documents is $D(N) = \{(d_1, \dots, d_N) : d_i \in D, d_i \neq d_j \text{ for any } i \neq j\}$, i.e. a set of ranked lists of retrieved documents without duplicates. A **run** r is the vector $r = (d_1, \dots, d_N)$ and we denote by $r[j]$ its j-th element, i.e. $r[j] = d_j$. Similarly, a **judged run** is the vector $\hat{r} = \big(GT(t, d_1), \dots, GT(t, d_N)\big)$, i.e. an ordered list of relevance degrees, where we denote by $\hat{r}[j]$ its j-th element, i.e. $\hat{r}[j] = GT(t, d_j)$.

4.2.1 Average Precision (AP)

AP is a binary measure given by

$$\text{AP}(\hat{r}) = \frac{1}{RB} \sum_{i=1}^{N} \left(\frac{1}{i} \sum_{j=1}^{i} g(\hat{r}[j]) \right) g(\hat{r}_t[i]) \ ,$$

where the gain function is such that $g(\mathcal{a}_0) = 0$ and $g(\mathcal{a}_1) = 1$, that is g is the indicator function $\delta_{\mathcal{a}}(\cdot)$, and RB is the recall base.

4.2.2 Graded Rank-Biased Precision (gRBP)

Let $p \in (0, 1)$ be the persistence parameter, i.e. how much a user is willing to continue to scan a result list. gRBP [21], [22] is a multi-graded relevance measure given by

$$\text{gRBP}(\hat{r}) = \frac{(1-p)}{g(\mathcal{a}_c)} \sum_{i=1}^{N} p^{i-1} g(\hat{r}[i]) \ .$$

Typical values of p are 0.5 for a very impatient user, 0.8 for a relatively patient user, and 0.95 for a user very persistent in deeply scanning the result list.

gRBP coincides with RBP when binary judgments ($c = 1$) are considered and $g(\mathcal{a}_1) = 1$.

4.2.3 Discounted Cumulated Gain (DCG)

DCG [23] is a multi-graded relevance measure given by

$$\text{DCG}_b(\hat{r}) = \sum_{i=1}^{N} \frac{g(\hat{r}[i])}{\max\{1, \log_b i\}} \ ,$$

where base b of the logarithm indicates the patience of the user in scanning the result list and plays a role somewhat similar to the persistence parameter p of RBP. Typical values for b are 2 for an impatient user and 10 for a patient user.

4.2.4 Expected Reciprocal Rank (ERR)

ERR [24] is a cascaded multi-graded relevance measure, accounting for all the previously seen relevant documents, given by

$$\text{ERR}(\hat{r}) = \sum_{i=1}^{N} \frac{1}{i} x_i \prod_{j=1}^{i-1} (1 - x_j) \ ,$$

with the convention that $\prod_{i=1}^{0} = 1$ and x_k represents the probability that a user leaves their search after considering the document at position k. In this work, we adopt the typical setting $x_k = (2^{g(\hat{r}[k])} - 1)/2^{g(\mathcal{a}_c)}$.

5 Set-based Measures

5.1 Total Ordering

As discussed in Section 3.3, we start by introducing an order relation $\preceq$ on the set of judged runs. Let $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \neq \hat{s}$, and let k be the biggest relevance degree at which the two runs differ for the first time, i.e. $k = \max\{j \leq c : \big|\{i : \hat{r}_i = \mathcal{a}_j\}\big| \neq \big|\{i : \hat{s}_i = \mathcal{a}_j\}\big|\}$. We strictly order any pair of distinct system runs as follows

$$\hat{r} \prec \hat{s} \iff \big|\{i : \hat{r}_i = \mathcal{a}_k\}\big| < \big|\{i : \hat{s}_i = \mathcal{a}_k\}\big| \ . \qquad (1)$$

Note that, with the previously explained convention that elements in any judged run $\hat{r}$ are represented in decreasing order, $\hat{r} \prec \hat{s}$ if and only if exists $j \leq N$ such that $\hat{r}_i = \hat{s}_i$ for any $i < j$ and $\hat{r}_j \prec \hat{s}_j$. From (1) we define the order relation

$$\hat{r} \preceq \hat{s} \iff \hat{r} \prec \hat{s} \text{ or } \hat{r} = \hat{s} \ .$$

For example, let $c = 4$, $\hat{r} = \{\mathcal{a}_4, \mathcal{a}_3, \mathcal{a}_2, \mathcal{a}_2, \mathcal{a}_2, \mathcal{a}_2\}$, and $\hat{s} = \{\mathcal{a}_4, \mathcal{a}_3, \mathcal{a}_3, \mathcal{a}_0, \mathcal{a}_0, \mathcal{a}_0\}$; we have $\hat{r} \prec \hat{s}$ since both runs have one document with relevance degree equal to $\mathcal{a}_4$, but $\hat{s}$ has two documents with relevance $\mathcal{a}_3$ while $\hat{r}$ has just one document with this relevance degree. Note that the number of documents each run has with lowest degrees doesn't matter, since they already differ at $\mathcal{a}_3$.

$R(N)$ is a totally ordered set with respect to the ordering $\preceq$ defined by (1). Indeed, every pair of runs in $R(N)$ is comparable. The antisymmetry follows since $\hat{r} \preceq \hat{s}$ and $\hat{s} \preceq \hat{r}$ hold true iff $\hat{r} = \hat{s}$. The transitivity, i.e. $\hat{r} \preceq \hat{s}$ and $\hat{s} \preceq \hat{u} \Rightarrow \hat{r} \preceq \hat{u}$, is trivial if $\hat{r} = \hat{s}$ or $\hat{s} = \hat{u}$. If $\hat{r} \prec \hat{s}$ and $\hat{s} \prec \hat{u}$, there exist j_1 and j_2 such that $\hat{r}_i = \hat{s}_i$ for any $i < j_1$, $\hat{r}_{j_1} \prec \hat{s}_{j_1}$, $\hat{s}_i = \hat{u}_i$ for any $i < j_2$ and $\hat{s}_{j_2} \prec \hat{u}_{j_2}$. If $i < j = \min(j_1, j_2)$, then $\hat{r}_i = \hat{u}_i$, while $\hat{r}_j \prec \hat{u}_j$, which implies that $\hat{r} \prec \hat{u}$.

As for any totally order set, $R(N)$ is a poset consisting of only one maximal chain (the whole set); therefore it is *graded of rank* $|R(N)| - 1$, where $|R(N)| = \binom{N+c}{N}$ since it consists of all the N combinations of $c + 1 = |REL|$ objects with repetition.

Since $R(N)$ is graded of rank $|R(N)| - 1$, there exists a unique *rank function* $\rho(\hat{r}) : R(N) \longrightarrow \mathbb{N}$ such that $\rho(\hat{0}) = 0$ and $\rho(\hat{s}) = \rho(\hat{r}) + 1$ if $\hat{s}$ covers $\hat{r}$.

We now show how to construct such unique rank function $\rho(\hat{r})$. Let us agree to set $\binom{n}{m} = 0$ if $n < m$ and consider $\hat{r} = \{\hat{r}_1, \hat{r}_2, \dots, \hat{r}_N\} \in R(N)$. We can construct an ordered chain of runs from $\hat{0}$ to $\hat{r}$ as follows: $\mathscr{R} = \{\hat{r}^0 = \hat{0}, \hat{r}^1, \hat{r}^2, \dots, \hat{r}^N = \hat{r}\} \subset R(N)$ such that for any $j \leq N$

$$\hat{r}^j = \{\hat{r}_1, \hat{r}_2, \dots, \hat{r}_j, 0, \dots 0\} \, .$$

By construction, $\hat{r}^j \preceq \hat{r}^i$ for any $0 \leq j \leq i \leq N$; moreover

$$\rho(\hat{r}) = \rho(\hat{r}^N) = \sum_{j=1}^{N} \left(\rho(\hat{r}^j) - \rho(\hat{r}^{j-1})\right) \, .$$

Since $\hat{r}^j = \{\hat{r}_1, \hat{r}_2, \dots, \hat{r}_{j-1}, \hat{r}_j, 0, \dots 0\}$ and $\hat{r}^{j-1} = \{\hat{r}_1, \hat{r}_2, \dots, \hat{r}_{j-1}, 0, \dots 0\}$, the difference $\rho(\hat{r}^j) - \rho(\hat{r}^{j-1})$ equals $\widetilde{\rho}(\hat{s}^j) - \widetilde{\rho}(\hat{s}^{j-1})$ where: $\widetilde{\rho}(\cdot)$ is the rank function defined on $R(N - (j-1))$; $\hat{s}^j = \{\hat{r}_j, 0, \dots, 0\}$; and $\hat{s}^{j-1} = \{0, 0, \dots, 0\}$, that is we are considering only the tails where the two runs differ. $\widetilde{\rho}(\hat{s}^{j-1}) = 0$; indeed, if $\hat{r}_j = \mathpzc{a}_0$, then also $\widetilde{\rho}(\hat{s}^j) = 0$; otherwise, if $\hat{r}_j = \mathpzc{a}_k \succ \mathpzc{a}_0$, then $\hat{s}^j$ covers $\hat{u} = \{\mathpzc{a}_{k-1}, \dots, \mathpzc{a}_{k-1}\} \in R(N - (j-1))$. This means that, when $\hat{r}_j = \mathpzc{a}_k \succ \mathpzc{a}_0$, $\hat{s}^j$ is the run in the total ordered set $R(N - (j-1))$ that comes after $N - (j-1)$ combinations of $\left|\{\mathpzc{a}_0, \dots, \mathpzc{a}_{k-1}\}\right| = k = \delta_{\mathpzc{a}}(\hat{r}_j)$ objects with repetitions, that is

$$\widetilde{\rho}(\hat{s}^j) = \binom{N - (j-1) + \delta_{\mathpzc{a}}(\hat{r}_j) - 1}{N - (j-1)} = \rho(\hat{r}^j) - \rho(\hat{r}^{j-1}) \, .$$

Therefore, the rank function $\rho : R(N) \longrightarrow \{0, 1, \dots, \binom{N+c}{N}\}$ is uniquely defined as:

$$\rho(\hat{r}) = \sum_{j=1}^{N} \binom{\delta_{\mathpzc{a}}(\hat{r}_j) + N - j}{N - j + 1} \, , \tag{2}$$

where $\hat{r} = \{\hat{r}_1, \dots, \hat{r}_N\} \in R(N)$ with $\hat{r}_i \preceq \hat{r}_{i+1}$ for any $i < N$.

The following example clarifies how the rank function has been constructed.

Example 2. Let us set $c = 3$, $N = 5$ and $\hat{r} = \{\mathpzc{a}_3, \mathpzc{a}_2, \mathpzc{a}_1, \mathpzc{a}_0, \mathpzc{a}_0\}$. By (1), $\hat{t} = \{\mathpzc{a}_3, \mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0\} \preceq \hat{r}$. Then $\hat{t}$ covers $\hat{u} = \{\mathpzc{a}_2, \mathpzc{a}_2, \mathpzc{a}_2, \mathpzc{a}_2, \mathpzc{a}_2\}$, indeed there is no $\hat{v} \in R(N)$ such that $\hat{u} \prec \hat{v} \prec \hat{t}$, and we can easily determine $\rho(\hat{t})$ since $\hat{t}$ is preceded by 5 combinations of $|\{\mathpzc{a}_0, \mathpzc{a}_1, \mathpzc{a}_2\}| = 3$ objects with repetition, thus $\rho(\hat{t}) = \binom{\delta_{\mathpzc{a}}(\hat{r}_1)+N-1}{N-1+1} = \binom{3+5-1}{5} = 21$ (remember that $\rho(\{\mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0\}) = 0$ by definition).

Let us now consider $\hat{w} = \{\mathpzc{a}_3, \mathpzc{a}_2, \mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0\}$: $\hat{t} \prec \hat{w} \prec \hat{r}$. In order to determine $\rho(\hat{w}) - \rho(\hat{t})$, we have just to compute the rank of $\hat{\bar{w}} = \{\mathpzc{a}_2, \mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0\}$. Analogously, $\hat{\bar{w}}$ covers $\{\mathpzc{a}_1, \mathpzc{a}_1, \mathpzc{a}_1, \mathpzc{a}_1\}$, that is $\hat{\bar{w}}$ is preceded by 4 combinations of $|\{\mathpzc{a}_0, \mathpzc{a}_1\}| = 3$ objects with repetition, thus $\rho(\hat{\bar{w}}) = \binom{\delta_{\mathpzc{a}}(\hat{r}_2)+N-2}{N-2+1} = \binom{2+5-2}{5-2+1} = 5 = \rho(\hat{w}) - \rho(\hat{t}) = 5$.

Eventually, we can compute $\rho(\hat{r}) - \rho(\hat{w})$ just as the rank of $\hat{\bar{r}} = \{\mathpzc{a}_1, \mathpzc{a}_0, \mathpzc{a}_0\}$. Since $\hat{\bar{r}}$ covers $\{\mathpzc{a}_0, \mathpzc{a}_0, \mathpzc{a}_0\}$, then $\rho(\hat{\bar{r}}) = 1 = \rho(\hat{r}) - \rho(\hat{w})$.

Therefore $\rho(\hat{r}) = \rho(\hat{t}) + \left(\rho(\hat{w}) - \rho(\hat{t})\right) + \left(\rho(\hat{r}) - \rho(\hat{w})\right) = 21 + 5 + 1 = 27$, as we expected from (2).

The *natural distance* is then given by $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$, for $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \preceq \hat{s}$, and we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$.

Definition 2. Given two runs $\hat{r}, \hat{s} \in R(N)$, the **difference** between $\hat{r}$ and $\hat{s}$ is defined as

$$\Delta_{\hat{r}\hat{s}} = \sum_{j=1}^{N} \left[\binom{\delta_{\mathpzc{a}}(\hat{s}_j) + N - j}{N - j + 1} - \binom{\delta_{\mathpzc{a}}(\hat{r}_j) + N - j}{N - j + 1} \right] \, .$$

Let $\preceq_d$ be the *less than or equal to* relation on $R(N) \times R(N)$, where the subscript d is to highlight its connection with intervals, as described in Section 3.2. We show that $(R(N), \preceq_d)$ is a *difference structure*. The ordering $\preceq_d$ between intervals is given by the well known order $\leq$ among real numbers, since the difference $\Delta_{\hat{r}\hat{s}}$ is an integer number and, therefore, $\preceq_d$ is a weak order. Indeed for every $\hat{r}, \hat{s}, \hat{t}, \hat{u}, \hat{v}, \hat{z} \in R(N)$, from the fact that $\Delta_{\hat{r}\hat{s}}, \Delta_{\hat{t}\hat{u}}, \Delta_{\hat{v}\hat{z}} \in \mathbb{Z}$, it follows $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{t}\hat{u}}$ or $\Delta_{\hat{t}\hat{u}} \preceq_d \Delta_{\hat{r}\hat{s}}$. Moreover if $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{t}\hat{u}}$ and $\Delta_{\hat{t}\hat{u}} \preceq_d \Delta_{\hat{v}\hat{z}}$, then $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{v}\hat{z}}$. Condition ii of Definition 1 follows from the fact that, from its definition, $\Delta_{\hat{r}\hat{s}} = -\Delta_{\hat{s}\hat{r}}$. For example, when $\hat{r} \preceq \hat{s}$ and $\hat{t} \preceq \hat{u}$, $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{t}\hat{u}}$ means that $\ell(\hat{r}, \hat{s}) \leq \ell(\hat{t}, \hat{u})$, which implies that $-\ell(\hat{t}, \hat{u}) \leq -\ell(\hat{r}, \hat{s})$, which in turn implies that $\Delta_{\hat{u}\hat{t}} \preceq_d \Delta_{\hat{s}\hat{r}}$. Condition iii. follows from the fact that $\Delta_{\hat{r}\hat{s}} = \Delta_{\hat{r}\hat{t}} + \Delta_{\hat{t}\hat{s}}$, for any $\hat{r}, \hat{s}, \hat{y} \in R(N)$. Whereas the *Solvability Condition*, i.e. having an equally-spaced gradation on $R(N)$, is satisfied by construction: if $\hat{s}$ covers $\hat{r}$, the difference $\Delta_{\hat{r}\hat{s}}$ is constant and equal to 1, since $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r}) = \rho(\hat{r}) + 1 - \rho(\hat{r}) = 1$.

Let us show how we can construct an interval scale measure M on $R(N)$. Given $\hat{r} \in R(N)$, $\rho(\hat{r})$ computes the total number of runs preceding $\hat{r}$ and, if $\hat{s}$ covers $\hat{r}$, the difference $\Delta_{\hat{r}\hat{s}} = \rho(\hat{s}) - \rho(\hat{r})$ is always equal to 1, by construction. Thus, an interval scale measure M on $(R(N), \preceq_d)$ is given by the rank function itself:

$$\mathrm{M}(\hat{r}) = \rho(\hat{r}) = \sum_{j=1}^{N} \binom{\delta_{\mathpzc{a}}(\hat{r}_j) + N - j}{N - j + 1} \, , \tag{3}$$

which satisfies the condition imposed by Theorem 1: let $\hat{r}, \hat{s}, \hat{u}, \hat{v} \in R(N)$ such that $\Delta_{\hat{r}\hat{s}} \leq \Delta_{\hat{u}\hat{v}}$, then $\Delta_{\hat{r}\hat{s}} \leq \Delta_{\hat{u}\hat{v}} \Leftrightarrow \rho(\hat{s}) - \rho(\hat{r}) \leq \rho(\hat{v}) - \rho(\hat{u}) \leq \sum_{i=1}^{N} \left(\hat{v}_i - \hat{u}_i\right) \Leftrightarrow \mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) \leq \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u})$, since $R(N)$ is totally ordered. Thus, M is an interval scale on $(R(N), \preceq_d)$.

Let us explore more deeply how the measure defined in (3) works. The first relevance degree immediately above not relevant, i.e. a_1, always gives a constant contribution, independently from how many a_1 documents are retrieved, since:

$$\binom{\delta_a(a_1) + N - j}{N - j + 1} = \binom{1 + N - j}{N - j + 1} = 1 \ .$$

However, when we consider higher relevance degrees, i.e. a_k with $k > 1$, the binomial coefficient strictly depends and changes on the basis of how many of them are retrieved. Indeed, $\delta_a(a_k)$ is constant for all the documents with the same relevance degree a_k but the term $N - j$ decreases as the number of a_k retrieved documents increases due N being constant and j increasing, i.e. the binomial coefficient is decreasing in the number of a_k retrieved documents. In other terms, each additional a_k retrieved document gives a contribution smaller than the previously retrieved ones by a discount factor j. This somehow recalls the idea that relevance is a dynamic notion which changes as far as more relevant documents are inspected, see e.g. [25]. As a consequence, given $\hat{r}, \hat{s} \in R(N)$, a replacement in $\hat{r}$ may have a different effect than the same replacement in $\hat{s}$, if the relevance degree of the new document is greater than a_1.

Example 3. Let us consider $REL = \{a_0, a_1, a_2\}$, $N = 5$, $\hat{r} = \{a_2, a_2, a_1, a_1, a_1\}$, and $\hat{s} = \{a_2, a_1, a_1, a_0, a_0\}$. From(3), we have that $M(\hat{r}) = \binom{2+5-1}{5-1+1} + \binom{2+5-2}{5-2+1} + 1 + 1 + 1 = 14$ and $M(\hat{s}) = \binom{2+5-1}{5-1+1} + 1 + 1 = 8$.
Let us now replace one a_1 retrieved document with a a_2 one; we obtain the new runs $\hat{r}^{(1)} = \{a_2, a_2, a_2, a_1, a_1\}$ and $\hat{s}^{(1)} = \{a_2, a_2, a_1, a_0, a_0\}$. Therefore, $M(\hat{r}^{(1)}) = \binom{2+5-1}{5+1-1} + \binom{2+5-2}{5-2+1} + \binom{2+5-3}{5-3+1} + 1 + 1 = 17$ and $M(\hat{s}^{(1)}) = \binom{2+5-1}{5-1+1} + \binom{2+5-2}{5-2+1} + 1 = 12$.
We can note the different contribution of the same replacement in the two runs, since $M(\hat{r}^{(1)}) - M(\hat{r}) = 3 \neq 4 = M(\hat{s}^{(1)}) - M(\hat{s})$.

We can finally proceed with the last step of Section 3.3 and check whether an IR measure uses an interval scale on $(R(N), \preceq_d)$ by looking for a linear positive transformation with M.

5.1.1 Binary Relevance Case

When $c = 1$, i.e. in the binary relevance case, the ordering (1) just orders judged runs by how many relevant documents they retrieve, i.e. by their total mass of relevance:

$$\hat{r} \preceq \hat{s} \ \Leftrightarrow \ \sum_{i=1}^{N} \delta_a(\hat{r}_i) \leq \sum_{i=1}^{N} \delta_a(\hat{s}_i) \ ,$$

since there is only one relevant relevance degree a_1.

Therefore the rank function becomes

$$\rho(\hat{r}) = \sum_{i=1}^{N} \delta_a(\hat{r}_i) = \mathrm{M}(\hat{r}) \ .$$

This follows easily from (3), using the fact that $\delta_a(\hat{r}_i) \in \{0, 1\}$ for any $i \leq N$ when $c = 1$.

Let now g be the gain function, and let us consider *Precision*

$$\mathrm{P}(\hat{r}) = \frac{1}{N} \sum_{i=1}^{N} \frac{g(\hat{r}_i)}{g(a_1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_a(\hat{r}_i) = \frac{\mathrm{M}(\hat{r})}{N} \ ,$$

since $g(a_0) = 0 = \delta_a(a_0)$ and $c = 1$. Thus Precision is an interval scale, as it is a linear positive transformation of M.

Similarly, *Recall*

$$\mathrm{R}(\hat{r}) = \frac{1}{RB} \sum_{i=1}^{N} \frac{g(\hat{r}_i)}{g(a_1)} = \frac{1}{RB} \sum_{i=1}^{N} \delta_a(\hat{r}_i) = \frac{\mathrm{M}(\hat{r})}{RB}$$

is an interval scale.

The *F-measure*, that is the harmonic mean of Precision and Recall,

$$\mathrm{F}(\hat{r}) = 2\frac{\mathrm{P}(\hat{r}) \cdot \mathrm{R}(\hat{r})}{\mathrm{P}(\hat{r}) + \mathrm{R}(\hat{r})} = \frac{2}{N + RB} \sum_{i=1}^{N} \delta_a(\hat{r}_i) = \frac{2\mathrm{M}(\hat{r})}{N + RB}$$

is an interval scale as well.

5.1.2 Multi-graded Relevance Case

Neither Generalized Precision nor Generalized Recall are a positive linear transformation of M defined in (3). Indeed, in these measures, the individual contribution of each retrieved document $\hat{r}_j$ is independent from the contribution of any other retrieved document $\hat{r}_k$. However, the previous discussion on the measure defined in (3) pointed out that, for each relevance degree a_k with $k > 1$, the individual contribution of an a_k retrieved document depends on how many a_k retrieved documents there are in the run. Therefore neither *gP* nor *gR* are an interval scale, since they cannot be a linear transformation of M.

Moreover they are not even an ordinal scale which, again, implies they cannot be an interval scale too. Indeed, a measure M′ is an ordinal scale on $R(N)$ if, for every $\hat{r}, \hat{s} \in R(N)$, the following statement is true:

$$\hat{r} \preceq \hat{s} \ \Leftrightarrow \ \mathrm{M}'(\hat{r}) \leq \mathrm{M}'(\hat{s}) \ .$$

Let us consider $\hat{r} = \{a_1, \dots, a_1\}$ and $\hat{s} = \{a_2, a_0, \dots, a_0\}$, two runs of length N. We have $\hat{r} \prec \hat{s}$. Moreover, since gR and gP are both proportional to $\mathrm{G}(\hat{v}) := \sum_{i=1}^{N} g(\hat{v}_i)/g(a_c)$, for any $\hat{v} \in R(N)$, we can just prove that $\mathrm{G}(\cdot)$ in not on an ordinal scale with respect to the order (1).

Since $g(a_0) = 0$, $\mathrm{G}(\hat{r}) = Ng(a_1)/g(a_c)$ while $\mathrm{G}(\hat{s}) = g(a_2)/g(a_c)$. From the fact that the gain function g is a positive strictly increasing function and it is defined independently from the length N of the runs, by choosing a N big enough we can have $\mathrm{G}(\hat{r}) > \mathrm{G}(\hat{s})$. Therefore, in the multi-graded relevance case when $c > 1$, the measures introduced in Section 4.1 are not even an ordinal scale on $(R(N), \preceq)$ for the ordering $\preceq$ defined in (1).

5.2 Partial Ordering

We now consider a partial order on $R(N)$, based on the following *monotonicity-like* property:

> **Replacement**. A run replacing a document with another one with a higher relevance degree should be greater than the original run.

Let us consider two runs $\hat{r}, \hat{s} \in R(N)$, then

$$\hat{r} \preceq \hat{s} \ \Leftrightarrow \big|\{i : \hat{r}_i \succeq a_j\}\big| \leq \big|\{i : \hat{s}_i \succeq a_j\}\big| \quad \forall j \in \{0, \dots, c\} \tag{4}$$

considers a run greater than another one if, for each relevance degree, it has more documents above that relevance degree than second one. The replacement property consists of replacing a document with another one with higher relevance, i.e. if $\hat{r} = \{\hat{r}_1, \dots, \hat{r}_{i-1}, \hat{r}_i, \hat{r}_{i+1} \dots, \hat{r}_N\}$, we can replace $\hat{r}_i$ with a document with higher relevance degree, that is $\bar{\hat{r}} = \{\hat{r}_1, \dots, \hat{r}_{i-1}, \mathbb{a}_j, \hat{r}_{i+1} \dots, \hat{r}_N\}$, with $\hat{r}_i \preceq \mathbb{a}_j$. It is intuitive that $\hat{r}$ should be *smaller* than $\bar{\hat{r}}$, and this relation is satisfied by the order (4).

The order (4) is a partial order on $R(N)$ since not every pair of runs is comparable. Therefore, $R(N)$ is a poset.

***Example* 4.** Let $\hat{r}, \hat{s} \in R(N)$ be such that $\hat{r} = \{\mathbb{a}_2, \mathbb{a}_2, \mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_0\}$ and $\hat{s} = \{\mathbb{a}_3, \mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_0\}$. These two runs are incomparable with respect to the ordering (4) since $\big|\{i : \hat{r}_i \succeq \mathbb{a}_3\}\big| = 0 < 1 = \big|\{i : \hat{s}_i \succeq \mathbb{a}_3\}\big|$ while $\big|\{i : \hat{r}_i \succeq \mathbb{a}_2\}\big| = 2 > 1 = \big|\{i : \hat{s}_i \succeq \mathbb{a}_2\}\big|$.
As a side note, the two runs are comparable when considering the total order (1) instead. Indeed, $\hat{r} \prec \hat{s}$ since $\hat{r}$ has no documents with relevance $\mathbb{a}_3$, while $\hat{s}$ has one document with such relevance degree.

5.2.1 Binary Relevance Case

The order (4) is a total order, since runs are ordered by the total number of relevant documents they retrieve. Therefore, the order (4) coincides with the total order (1) and the same results of Section 5.1.1 hold, i.e. Precision, Recall, and F-measure are interval scales.

5.2.2 Multi-graded Relevance Case

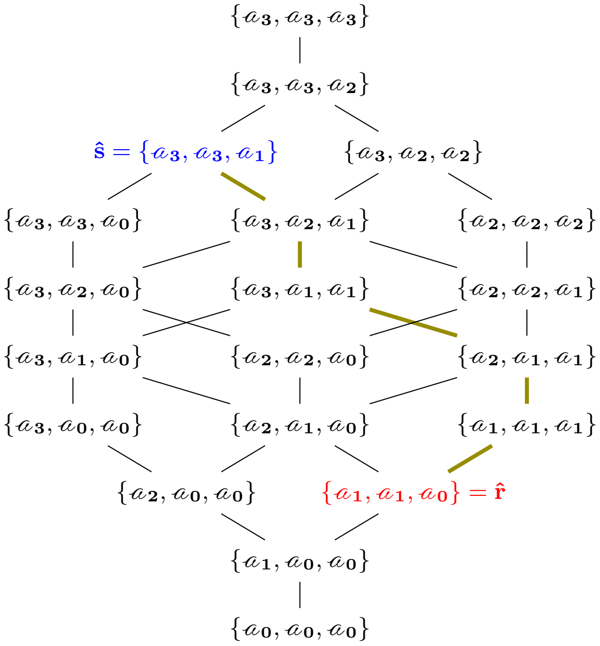
In this case, $R(N)$ is bounded since, for any $\hat{r} \in R(N)$, $\hat{r} \succeq \{\mathbb{a}_0, \dots, \mathbb{a}_0\}$ and $\hat{r} \preceq \{\mathbb{a}_c, \dots, \mathbb{a}_c\}$. Moreover, it is of finite rank since $\big|R(N)\big| = \binom{N+c}{N}$, as shown before.

The following proposition holds:

***Proposition* 2.** Let $N \in \mathbb{N}$ be fixed and let $REL = \{\mathbb{a}_0, \dots, \mathbb{a}_c\}$ with $c > 1$. The poset $R(N)$ is graded, i.e. every maximal chain of $R(N)$ has the same length.

See the Electronic Appendix for a complete proof.

***Example* 5.** Let us fix $N = 3$ and $c = 3$. The Hasse Diagram of $R(N)$ is



$\{\mathbb{a}_3, \mathbb{a}_3, \mathbb{a}_3\}$
$\{\mathbb{a}_3, \mathbb{a}_3, \mathbb{a}_2\}$
$\hat{\mathbf{s}} = \{\mathbb{a}_3, \mathbb{a}_3, \mathbb{a}_1\}$
$\{\mathbb{a}_3, \mathbb{a}_2, \mathbb{a}_2\}$
$\{\mathbb{a}_3, \mathbb{a}_3, \mathbb{a}_0\}$
$\{\mathbb{a}_3, \mathbb{a}_2, \mathbb{a}_1\}$
$\{\mathbb{a}_2, \mathbb{a}_2, \mathbb{a}_2\}$
$\{\mathbb{a}_3, \mathbb{a}_2, \mathbb{a}_0\}$
$\{\mathbb{a}_3, \mathbb{a}_1, \mathbb{a}_1\}$
$\{\mathbb{a}_2, \mathbb{a}_2, \mathbb{a}_1\}$
$\{\mathbb{a}_3, \mathbb{a}_1, \mathbb{a}_0\}$
$\{\mathbb{a}_2, \mathbb{a}_2, \mathbb{a}_0\}$
$\{\mathbb{a}_2, \mathbb{a}_1, \mathbb{a}_1\}$
$\{\mathbb{a}_3, \mathbb{a}_0, \mathbb{a}_0\}$
$\{\mathbb{a}_2, \mathbb{a}_1, \mathbb{a}_0\}$
$\{\mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_1\}$
$\{\mathbb{a}_2, \mathbb{a}_0, \mathbb{a}_0\}$
$\{\mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_0\} = \hat{\mathbf{r}}$
$\{\mathbb{a}_1, \mathbb{a}_0, \mathbb{a}_0\}$
$\{\mathbb{a}_0, \mathbb{a}_0, \mathbb{a}_0\}$

From the diagram you can note that a run covers another one if and only if a relevance degree $\mathbb{a}_j$ is replaced with $\mathbb{a}_{j-1}$, for $j > 0$.
If we consider $\hat{r} = \{\mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_0\}$ (in red) and $\hat{s} = \{\mathbb{a}_3, \mathbb{a}_3, \mathbb{a}_1\}$ (in blue), there are several shortest paths in the diagram connecting these two runs. One of them is, for example, composed by $\hat{r} = \{\mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_0\}$, $\{\mathbb{a}_1, \mathbb{a}_1, \mathbb{a}_1\}$, $\{\mathbb{a}_2, \mathbb{a}_1, \mathbb{a}_1\}$, $\{\mathbb{a}_3, \mathbb{a}_1, \mathbb{a}_1\}$, $\{\mathbb{a}_3, \mathbb{a}_2, \mathbb{a}_1\}$, $\{\mathbb{a}_3, \mathbb{a}_3, \mathbb{a}_1\} = \hat{s}$ (green edges). These paths all have the same length equal to 5 and each of them coincides with a maximal chain on $[\hat{r}, \hat{s}]$, viewed as a subset of $R(N)$.

As detailed in the Electronic Appendix, given two runs, one covers the other one if they differ only for the replacement of a document with another one with relevance degree immediately consecutive in REL. Therefore, in order to compute the rank function $\rho(\hat{r})$, we need to count the number of replacements needed to go from the smallest run possible, i.e. $(\mathbb{a}_0, \dots, \mathbb{a}_0)$ to $\hat{r}$, replacing elements of relevance degree $\mathbb{a}_j$ with relevance degree $\mathbb{a}_{j+1}$, $j < c$. In the Hasse diagram this means to follow one of the shortest paths connecting the two runs and counting the number of edges passed through, since each edge represent a “cover” relation. Therefore, the explicit expression for the unique rank function is

$$\rho(\hat{r}) = \sum_{i=1}^{N} \delta_{\mathbb{a}}(\hat{r}_i) \ ,$$

for $\hat{r} \in R(N)$.

***Example* 6.** Let us consider $\hat{s} = \{\mathbb{a}_3, \mathbb{a}_1, \mathbb{a}_0\}$. Since $\hat{s}_1 = \mathbb{a}_3$, from $\hat{0} = \{\mathbb{a}_0, \mathbb{a}_0, \mathbb{a}_0\}$ we need $3 = \delta_{\mathbb{a}}(\hat{s}_1)$ subsequent replacements in $\hat{0}_1$, leading to the run $\hat{t} = \{\mathbb{a}_3, \mathbb{a}_0, \mathbb{a}_0\}$, plus $1 = \delta_{\mathbb{a}}(\hat{s}_2)$ replacement in $\hat{t}_2$ to reach $\{\mathbb{a}_3, \mathbb{a}_1, \mathbb{a}_0\} = \hat{s}$. In other terms, we have to perform four “cover” operations to go from $\hat{0}$ to $\hat{s}$ and, equivalently, the path in the Hasse diagram has length equal to 4.

Therefore, from the natural distance $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$ if $\hat{r} \preceq \hat{s}$, we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$.

***Definition* 3.** Given two comparable runs $\hat{r}, \hat{s} \in R(N)$, with $REL = \{\mathbb{a}_0, \dots, \mathbb{a}_c\}$, the **difference** between $\hat{r}$ and $\hat{s}$ is $\Delta_{\hat{r}\hat{s}} = \sum_{i=1}^{N} \big(\delta_{\mathbb{a}}(\hat{s}_i) - \delta_{\mathbb{a}}(\hat{r}_i)\big)$.

Note that, contrary to the previous case, since the ordering given by (4) is only partial, in order to compare differences between intervals we need to restrict our study to a maximal chain, i.e. a totally ordered subset of $R(N)$. Thus, denoted with $\mathscr{C}(R(N))$ a maximal chain of $R(N)$, and given the *less than or equal to* $\preceq_d$ relation which, as in the previous case, coincides with the order relation $\leq$ among real numbers, the relational structure $(\mathscr{C}(R(N)), \preceq_d)$ is a difference structure. This follows from the same discussion we have done for the difference structure in the total order case in the previous subsection.

Therefore, an interval scale measure M on $(\mathscr{C}(R(N)), \preceq_d)$ is given by the rank function itself, that is

$$\mathrm{M}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N} \delta_{\mathbb{a}}(\hat{r}_i) \ , \tag{5}$$

for $\hat{r} \in \mathscr{C}(R(N))$.

Let us suppose that the gain function g is such that

$$g(\mathscr{a}_{i+1}) - g(\mathscr{a}_i) = K$$

for a positive K constant and for any $i \in \{0, \dots, c-1\}$. It follows that we can write the gain of each relevance degree as $g(\mathscr{a}_j) = K\delta_{\mathscr{a}}(\mathscr{a}_j) = Kj$, that is the gain function is the indicator function times a positive constant K. In other terms, we are requesting that the gain function itself is a ratio scale on the set REL of the relevance degrees. Under this condition, the measures of retrieval effectiveness defined in Section 4.1 are an interval scale.

Indeed, let us consider Generalized Precision

$$\mathrm{gP}(\hat{r}) = \frac{1}{N}\sum_{i=1}^{N}\frac{g(\hat{r}_i)}{g(\mathscr{a}_c)} = \frac{1}{N}\sum_{i=1}^{N}\frac{K\delta_{\mathscr{a}}(\hat{r}_i)}{K\delta_{\mathscr{a}}(\mathscr{a}_c)} = \frac{\mathrm{M}(\hat{r})}{cN} \; ,$$

since $\delta_{\mathscr{a}}(\mathscr{a}_c) = c$. Thus, it is an interval scale, as it is a linear positive transformation of M defined in (5).

Similarly, Generalized Recall

$$\mathrm{gR}(\hat{r}) = \frac{1}{RB}\sum_{i=1}^{N}\frac{g(\hat{r}_i)}{g(\mathscr{a}_c)} = \frac{1}{RB}\sum_{i=1}^{N}\frac{K\delta_{\mathscr{a}}(\hat{r}_i)}{K\delta_{\mathscr{a}}(\mathscr{a}_c)} = \frac{\mathrm{M}(\hat{r})}{cRB}$$

is an interval scale.

If the gain function g itself is not a linear positive transformation of the indicator function $\delta_{\mathscr{a}}$, i.e. it is not a ratio scale for the relevance degrees, it is not possible to find a linear positive transformation between M in (5) and any of Generalized Precision and Generalized Recall. Indeed, for any of those measures, we can find an example where Theorem 1 does not hold, as it is done in the following.

Example 7. Let us consider four runs in a maximal chain of $R(3)$, for $c = 3$: $\hat{u} = \{\mathscr{a}_0, \mathscr{a}_0, \mathscr{a}_0\}$, $\hat{v} = \{\mathscr{a}_1, \mathscr{a}_1, \mathscr{a}_1\}$, $\hat{r} = \{\mathscr{a}_2, \mathscr{a}_2, \mathscr{a}_1\}$ and $\hat{s} = \{\mathscr{a}_3, \mathscr{a}_2, \mathscr{a}_2\}$. We have $\hat{u} \prec \hat{v} \prec \hat{r} \prec \hat{s}$, and $\Delta_{\hat{r}\hat{s}} = 2 \leq 3 = \Delta_{\hat{u}\hat{v}}$ (see the Hasse diagram in Example 5).

Let us now consider the measure gP. If this measure is an interval scale for a given gain function g, then we should have $\mathrm{gP}(\hat{s}) - \mathrm{gP}(\hat{r}) \leq \mathrm{gP}(\hat{v}) - \mathrm{gP}(\hat{u})$, thanks to Theorem 1. This is true if and only if, multiplying both sides by $Ng(\mathscr{a}_c)$, $\sum_{i=1}^{N}(g(\hat{s}_i) - g(\hat{r}_i)) \leq \sum_{i=1}^{N}(g(\hat{v}_i) - g(\hat{u}_i)) \Leftrightarrow g(\mathscr{a}_3) - g(\mathscr{a}_1) \leq 3g(\mathscr{a}_1) \Leftrightarrow g(\mathscr{a}_3) \leq 4g(\mathscr{a}_1)$.

Therefore, when g is such that $g(\mathscr{a}_3) > 4g(\mathscr{a}_1)$, the measure Generalized Precision (and, similarly, also Generalized Recall) is not an interval scale. Similar examples can be found for every gain function that is not a linear positive transformation of the indicator function.

5.3 Induced Total Ordering

Fixed a measure, e.g. gP, let us consider the total order induced by the measure on the domain, i.e. $\hat{r} \preceq \hat{s}$ if and only if $\mathrm{gP}(\hat{r}) \leq \mathrm{gP}(\hat{s})$. With this ordering, a measure can be interval scale only if it generates an equally spaced graduation on the set $\{\mathrm{gP}(\hat{r}) : \hat{r} \in R(N)\} \subset \mathbb{R}$, i.e. if the *Solvability Condition* holds in the codomain.

We have already proved that the measures defined in Section 4.1 are interval scales in the binary relevance case and in the multi-graded relevance case only when the gain function is a ratio scale for the relevance degrees. In these cases, the induced total ordering coincides with the orders studied in the previous sections and the same results hold.

Let us now deal with the multi-graded case and a gain function g which is not a ratio scale for the relevance degrees, i.e. such that exists $j \in \{1, \dots, c\} : g(\mathscr{a}_j) \neq K\delta_{\mathscr{a}}(\mathscr{a}_j)$.

Looking at the codomain of, e.g., Generalized Precision, we can show that the *Solvability Condition* is not satisfied. Consider Example 7 and let $g(\mathscr{a}_3) - g(\mathscr{a}_2) \neq g(\mathscr{a}_2) - g(\mathscr{a}_1)$, i.e. the relevance degrees are not a ratio scale. The run $\hat{s} = \{\mathscr{a}_3, \mathscr{a}_3, \mathscr{a}_3\}$ covers $\hat{r} = \{\mathscr{a}_3, \mathscr{a}_3, \mathscr{a}_2\}$, which in turn covers $\hat{u} = \{\mathscr{a}_3, \mathscr{a}_3, \mathscr{a}_1\}$. We have $\mathrm{gP}(\hat{s}) - \mathrm{gP}(\hat{r}) = \frac{1}{Ng(\mathscr{a}_c)}(g(\mathscr{a}_3) - g(\mathscr{a}_2)) \neq \frac{1}{Ng(\mathscr{a}_c)}(g(\mathscr{a}_2) - g(\mathscr{a}_1)) = \mathrm{gP}(\hat{r}) - \mathrm{gP}(\hat{u})$, that is the *Solvability Condition* (considered on the codomain) fails. Similar considerations hold for the other cases and the Generalized Recall as well.

Therefore, even using the induced total ordering, these measures are not an interval scale in the multi-graded case, when the gain function g itself is not a ratio scale.

6 Rank-based Measures

6.1 Total Ordering

Top-heaviness is a central property in IR, stating that the higher a system ranks relevant documents the better it is. If we apply this property at each rank position and we take to extremes the importance of having a relevant document ranked higher, we can define a *strong top-heaviness* property which, in turn, will induce a total ordering among runs.

Let $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \neq \hat{s}$, then there exists $k = \min\{j \leq N : \hat{r}[j] \neq \hat{s}[j]\} < \infty$, and we order system runs as follows

$$\hat{r} \prec \hat{s} \;\Leftrightarrow\; \hat{r}[k] \prec \hat{s}[k] \; . \tag{6}$$

This ordering prefers a single relevant document ranked higher to any number of relevant documents, with the same relevance degree or higher, ranked just below it; more formally, $(\hat{u}[1], \dots, \hat{u}[m], \mathscr{a}_j, \mathscr{a}_0, \dots, \mathscr{a}_0)$ is greater than $(\hat{u}[1], \dots, \hat{u}[m], \mathscr{a}_0, \mathscr{a}_c, \dots, \mathscr{a}_c)$, for any $1 \leq j \leq c$, for any length $N \in \mathbb{N}$ and any $m \in \{0, 1, \dots, N-1\}$. This is why we call it *strong top-heaviness*.

$R(N)$ is totally ordered with respect to $\preceq$, since for every pair of runs $\hat{r}, \hat{s} \in R(N)$, if k is the smallest depth at which the two runs differ, we establish which one is the biggest by just looking at the relevance degrees of $r[k]$ and $s[k]$.

Moreover, $R(N)$ is *graded of rank* $(c+1)^N - 1$ since $R(N)$ is the set of the dispositions with repetition of $(c+1)$ elements from REL in a collection of N elemets, hence $|R(N)| = (c+1)^N$. Therefore, there is a unique rank function $\rho : R(N) \longrightarrow \{0, 1, \dots, (c+1)^N - 1\}$ which is given by:

$$\rho(\hat{r}) = \sum_{i=1}^{N}\delta_{\mathscr{a}}(\hat{r}[i])(c+1)^{N-i},$$

where $\delta_{\mathscr{a}}$ is the indicator function.

Let us set $\boldsymbol{\delta}_{\boldsymbol{\mathscr{a}}}(\hat{r}) = (\delta_{\mathscr{a}}(\hat{r}[1]), \dots, \delta_{\mathscr{a}}(\hat{r}[N]))$. If we look at $\boldsymbol{\delta}_{\boldsymbol{\mathscr{a}}}(\hat{r})$ as a string, the rank function is exactly the conversion in base 10 of the number in base $c+1$ identified by $\boldsymbol{\delta}_{\boldsymbol{\mathscr{a}}}(\hat{r})$ and the ordering among runs $\preceq$ corresponds to the ordering $\leq$ among numbers in base $c+1$.

Example 8. Let us firstly consider the binary relevance case, that is $c = 1$ and $REL = \{\mathscr{a}_0, \mathscr{a}_1\}$. Let

$\hat{r}, \hat{s} \in R(5)$ be such that $\hat{r} = (a_0, a_0, a_1, a_1, a_1)$ and $\hat{s} = (a_0, a_1, a_0, a_0, a_0)$. Since $\hat{r}[1] = \hat{s}[1]$, while $\hat{r}[2] = a_0 \prec a_1 = \hat{s}[2]$, we have $\hat{r} \prec \hat{s}$. Moreover $\boldsymbol{\delta}_{a}(\hat{r}) = (0, 0, 1, 1, 1)$ and $\boldsymbol{\delta}_{a}(\hat{s}) = (0, 1, 0, 0, 0)$, thus $\rho(\hat{r}) = 0 * 2^4 + 0 * 2^3 + 1 * 2^2 + 1 * 2^1 + 1 * 2^0 = 7 < 8 = 0 * 2^4 + 1 * 2^3 + 0 * 2^2 + 0 * 2^1 + 0 * 2^0 = \rho(\hat{s})$ and, in particular, $\hat{s}$ covers $\hat{r}$ (indeed $\rho(\hat{s}) = \rho(\hat{r}) + 1$).

Now let us consider the multi-graded relevance case when $c > 1$: for example we set $c = 3$ and $\hat{r}, \hat{s} \in R(5)$ such that $\hat{r} = (a_1, a_3, a_0, a_3, a_2)$ and $\hat{s} = (a_1, a_3, a_1, a_0, a_2)$. Note that $\hat{r} \prec \hat{s}$ since $\hat{r}[1] = \hat{s}[1], \hat{r}[2] = \hat{s}[2]$ but $\hat{r}[3] \prec \hat{s}[3]$, furthermore $\boldsymbol{\delta}_{a}(\hat{r}) = (1, 3, 0, 3, 2)$ and $\boldsymbol{\delta}_{a}(\hat{s}) = (1, 3, 1, 0, 2)$, thus $\rho(\hat{r}) = 1 * 4^4 + 3 * 4^3 + 0 * 4^2 + 3 * 4^1 + 2 * 4^0 = 461 < 466 = 1 * 4^4 + 3 * 4^3 + 4^2 + 0 * 4^1 + 2 * 4^0 = \rho(\hat{s})$.

The *natural distance* is then given by $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$, for $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \preceq \hat{s}$, and we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$.

Definition 4. Given two runs $\hat{r}, \hat{s} \in R(N)$, the **difference** between $\hat{r}$ and $\hat{s}$ is defined as $\Delta_{\hat{r}\hat{s}} = \sum_{i=1}^{N} \big(\delta_a(\hat{s}[i]) - \delta_a(\hat{r}[i])\big)(c+1)^{N-i}$.

Let $\preceq_d$ be the *less than or equal to* relation on $R(N) \times R(N)$ which, similarly to what has been demonstrated in the set-based case, is exactly the order relation $\leq$ among real numbers, then $(R(N), \preceq_d)$ is a difference structure. Indeed, as shown for the set-based case, the first three axioms of Theorem 1 follow immediately from the fact that the ordering $\preceq_d$ between intervals is given by the well known order $\leq$ among real numbers, thanks to the definition of difference. Finally, the *Solvability Condition* is satisfied by construction of the rank function, since $\Delta_{\hat{r}\hat{s}} = \rho(\hat{s}) - \rho(\hat{r}) = 1$ for every $\hat{r}, \hat{s} \in R(N)$ such that $\hat{s}$ covers $\hat{r}$.

Similarly to the set-based case, an interval scale measure M on $(R(N), \preceq_d)$ is given by the rank function itself

$$\mathrm{M}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N} \delta_a(\hat{r}[i])(c+1)^{N-i} \tag{7}$$

which is an interval scale since it satisfies the condition imposed by Theorem 1. To prove it, let $\hat{r}, \hat{s}, \hat{u}, \hat{v} \in R(N)$ such that $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{u}\hat{v}}$; then, $\Delta_{\hat{r}\hat{s}} \preceq_d \Delta_{\hat{u}\hat{v}} \Leftrightarrow \sum_{i=1}^{N}(\delta_a(\hat{s}[i]) - \delta_a(\hat{r}[i]))(c+1)^{N-i} \leq \sum_{i=1}^{N}(\delta_a(\hat{v}[i]) - \delta_a(\hat{u}[i]))(c+1)^{N-i} \Leftrightarrow \mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) \leq \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u})$.

In the following, we first investigate when IR evaluation measures are an ordinal scale, which is a necessary condition for being an interval scale. Remember that a measure M' is an ordinal scale on $R(N)$ if, for every $\hat{r}, \hat{s} \in R(N)$, the following statement is true:

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \mathrm{M}'(\hat{r}) \leq \mathrm{M}'(\hat{s}) \; .$$

Let us recall the definition of gRBP_p from Section 4.2:

$$\text{gRBP}_p(\hat{r}) = \frac{1-p}{g(a_c)} \sum_{i=1}^{N} g(\hat{r}[i]) p^{i-1},$$

and let $G = \min_{j \in \{1,\dots,c\}} (g(a_j) - g(a_{j-1}))/g(a_c) > 0$ be the normalized smallest gap between the gain of two consecutive relevance degrees. Then, gRBP_p is an ordinal scale on $R(N)$ with respect to the total order (6) if and only if $0 < p \leq G/(G+1)$; see the Electronic Appendix for a complete proof. Note that in the binary relevance case, RBP_p is gRBP_p where $c = 1$ and $G = 1$, that is RBP_p is ordinal if and only if $0 < p \leq 1/2$.

gRBP_p with $p > G/(G+1)$ and other IR measures – namely AP, DCG, and ERR – are not even an ordinal scale on $R(N)$, as the following example shows. Therefore, these measures cannot be an interval scale too, since an interval scale measure is also ordinal scale.

Example 9. Let $\hat{r} = (a_1, a_0, a_2, a_0, a_1)$ and $\hat{s} = (a_1, a_1, a_0, a_0, a_0)$ be two runs on $R(5)$ with $c = 2$ and let us use the indicator function δ as gain function g. We have $\hat{r} \preceq \hat{s}$. Then $\text{DCG}_2(\hat{r}) = 1 + 2/\log_2 3 + 1/\log_2 5 > 1 + 1 = \text{DCG}_2(\hat{s})$; $\text{ERR}(\hat{r}) = 1/4 + 3/16 + 3/320 > 1/4 + 3/32 = \text{ERR}(\hat{s})$; finally, since $g(a_2) = \delta_a(a_2) = 2$, $2\text{gRBP}_p(\hat{r}) = (1-p)(1 + 2p^2 + p^4) > (1-p)(1+p) = 2\text{gRBP}(\hat{s})$ for $p \gtrsim 0.454$, and such an example can be found for any other values of $p > G/(G+1)$, where $G = 1/2$.

AP is a binary measure and, just to stay with the same data above, we adopt a lenient mapping of multi-graded to binary relevance degrees setting $g(a_1) = g(a_2) = 1$ and thus $RB{\cdot}\text{AP}(\hat{r}) = 1 + 2/3 + 3/5 > 1 + 1 = RB{\cdot}\text{AP}(\hat{s})$, where RB is the recall base.

Therefore, only gRBP_p with $p \leq G/(G+1)$ may be on interval scale. Firstly note that, given $\hat{r} \in R(N)$, in order for gRBP_p to be a linear positive transformation of (7), the gain function g has to be such that $g(a_i) = K\delta_a(a_i)$, for any $i \in \{0, \dots, \mathbb{N}\}$ and for any $K > 0$ fixed. In other terms, g is a ratio scale with respect to the relevance degrees.

With such gain function, G is equal to $1/c$ and gRBP_p is interval scale if and only if $p = G/(1+G) = (c+1)^{-1}$, since

$$\begin{aligned} \text{gRBP}_{(c+1)^{-1}}(\hat{r}) &= \frac{c}{c+1} \frac{1}{Kc} \sum_{i=1}^{N} K\delta_a(\hat{r}[i])(c+1)^{-i+1} \\ &= \frac{1}{(c+1)^N} \sum_{i=1}^{N} \delta_a(\hat{r}[i])(c+1)^{N-i} \\ &= \frac{1}{(c+1)^N} \mathrm{M}(\hat{r}) \; . \end{aligned}$$

On the other hand gRBP_p with $p < G/(G+1)$ is not a linear positive transformation of M, since it does not preserve the equivalence between differences. Indeed, let us consider $\hat{r} = (a_0, a_0, a_0, a_0, a_c)$, $\hat{s} = (a_0, a_0, a_0, a_1, a_0)$, $\hat{u} = (a_0, a_0, a_0, a_c, a_c)$ and $\hat{v} = (a_0, a_0, a_1, a_0, a_0)$, four runs on $R(5)$. Note that $\hat{s}$ covers $\hat{r}$ and $\hat{v}$ covers $\hat{u}$ but $\text{RBP}_p(\hat{s}) - \text{RBP}_p(\hat{r}) = \text{RBP}_p(\hat{v}) - \text{RBP}_p(\hat{u}) \Leftrightarrow p^3 - cp^4 = p^2 - cp^3 - cp^4$, that is if and only if $p = G/(1+G) = (c+1)^{-1}$.

Therefore, given the total order (6) induced by the strong top-heaviness, $\text{gRBP}_{(c+1)^{-1}}$ with the gain function such that $g(a_j) = Kj$ for all $j \in \{0, \dots, c\}$ and for any $K > 0$ fixed, is the only one among the considered IR measures that is an interval scale with respect to the difference structure defined above. Note that in the binary relevance case, this means that RBP is an interval scale only for $p = 1/2$.

Finally, note that when $g(a_j) \neq Kj$ not even $\text{gRBP}_{(c+1)^{-1}}$ is an interval scale measure; see the Electronic Appendix for a complete proof.

6.2 Partial Ordering

We abandon the total ordering induced by the strong top-heaviness and we explore a partial ordering, induced by a weaker form of top-heaviness. This ordering is based on two *monotonicity-like* properties which extend the replacement property defined in the previous section:

> **Replacement**. A run replacing a document with another one **in the same rank position** with a higher relevance degree should be greater than the original run.
> **Swap**. A run swapping a less relevant document with a more relevant one in a lower rank position should be greater than the original run.

Let us consider two runs $\hat{r}, \hat{s} \in R(N)$, then

$$\hat{r} \preceq \hat{s} \ \Leftrightarrow \big|\{i \le k : \hat{r}[i] \succeq a_j\}\big| \le \big|\{i \le k : \hat{s}[i] \succeq a_j\}\big| \quad \forall j \in \{0, \dots, c\} \ \text{ and } \ k \in \{1, \dots, N\} \tag{8}$$

defines a partial ordering among system runs, which considers a run bigger than another one when, for each rank position, it has more relevant documents than the other one up to that rank for every relevance degree. With respect to the strong top-heaviness (6), this ordering is less extreme because it is sensitive to the total mass of relevance accumulated at the different rank positions instead of “cutting” everything just because of a single relevant document ranked higher. This is why we call it *weak top-heaviness*.

Similarly to the set-based case, the replacement property consists of replacing a document of a judged run with another one with higher relevance in the same rank position, i.e. if $\hat{r} = (\hat{r}[1], \dots, \hat{r}[i-1], \hat{r}[i], \hat{r}[i+1] \dots, \hat{r}[N])$, we can replace $\hat{r}[i]$ with a document with higher relevance degree, that is $\hat{\hat{r}} = (\hat{r}[1], \dots, \hat{r}[i-1], a_j, \hat{r}[i+1] \dots, \hat{r}[N])$, with $\hat{r}[i] \preceq a_j$. We can thus agree on $\hat{r}$ being *smaller* run than $\hat{\hat{r}}$ and this relation is satisfied by the partial ordering just introduced. The swap property, instead, consists of swapping two documents with different relevance degrees as follows: let for example $\hat{r} = (\hat{r}[1], \dots, \hat{r}[i], \dots, \hat{r}[j], \dots, \hat{r}[N])$ with $\hat{r}[i] \preceq \hat{r}[j]$, then $\hat{\hat{r}} = (\hat{r}[1], \dots, \hat{r}[i-1], \hat{r}[j], \hat{r}[i+1], \dots, \hat{r}[j-1], \hat{r}[i], \hat{r}[j+1], \dots, \hat{r}[N])$. According to the ordering (8) the two runs are such that $\hat{r} \preceq \hat{\hat{r}}$.

The ordering $\preceq$ is a partial ordering on $R(N)$, since not every pair of runs is comparable, as pointed out in the next example. Thus $R(N)$ is a poset.

Example 10. Let $\hat{r}, \hat{s} \in R(N)$ be such that $\hat{r} = (a_1, a_1, a_3, a_0, a_2)$ and $\hat{s} = (a_1, a_2, a_0, a_0, a_3)$. These two runs are incomparable for the above ordering (8) since $|\{i \le 2 : \hat{r}[i] \succeq a_2\}| = 0 < 1 = |\{i \le 2 : \hat{s}[i] \succeq a_2\}|$, while $|\{i \le 3 : \hat{r}[i] \succeq a_1\}| = 3 > 2 = |\{i \le 3 : \hat{s}[i] \succeq a_1\}|$.
As a side note, on the contrary, according to the ordering (6) and characterized by the strong top-heaviness, $\hat{r} \preceq \hat{s}$ since $\hat{r}[1] = \hat{s}[1]$ while $\hat{r}[2] \prec \hat{s}[2]$ regardless of the relevance of the documents ranked in lower positions.

In addition, $R(N)$ is bounded since $\hat{r} \succeq (a_0, \dots, a_0)$ and $\hat{r} \preceq (a_c, \dots, a_c)$ for every $\hat{r} \in R(N)$, and it is of finite rank since $|R(N)| = (c+1)^N < \infty$, as shown above.

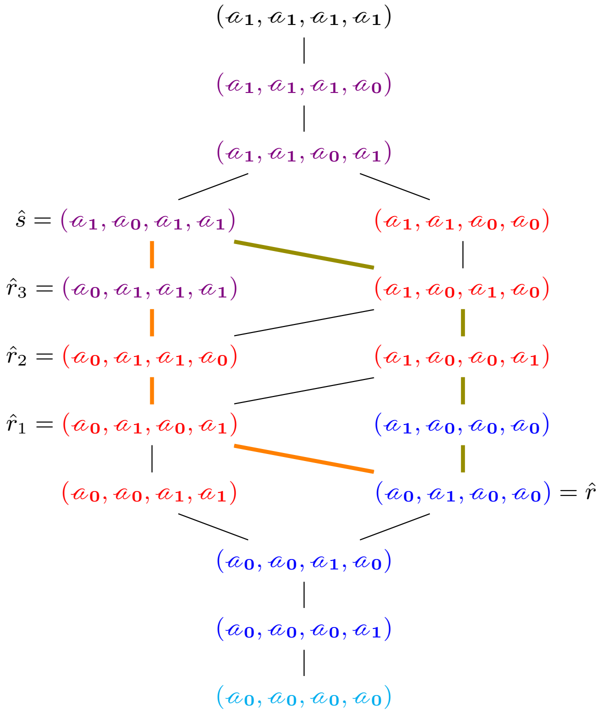
6.2.1 *Binary Relevance Case*

The following proposition holds:

Proposition 3. Let $N \in \mathbb{N}$ be fixed and let $REL = \{a_0, a_1\}$. The poset $R(N)$ is graded, i.e. every maximal chain of $R(N)$ has the same length.

See the Electronic Appendix for a complete proof.

Example 11. Let $N = 4$. The Hasse Diagram of $R(N)$ is



where different colours of the runs correspond to different total numbers of relevant retrieved documents. Given $\hat{r} = (a_0, a_1, a_0, a_0)$, $\hat{s} = (a_1, a_0, a_1, a_1)$, let us consider one of the shortest paths between the two runs, for example the one with orange edges: it starts from $\hat{r}$, goes through $\hat{r}_1, \hat{r}_2, \hat{r}_3$, and ends in $\hat{s}$. Note that $\{\hat{r}_0 := \hat{r}, \hat{r}_1, \hat{r}_2, \hat{r}_3, \hat{r}_4 := \hat{s}\}$ is also a maximal chain, since there does not exist $\hat{u} \in R(4)$ such that $\hat{r}_i \prec \hat{u} \prec \hat{r}_{i+1}$ for any $i \in \{0, 1, 2, 3\}$. Moreover this shortest path has length 4, and every other shortest path between $\hat{r}$ and $\hat{s}$ has the same length, e.g. the one with dark green edges. Thus the natural length of $[\hat{r}, \hat{s}]$ is 4.

Given two runs one covers the other one if they differ only for a swap of length one or a replacement in the last position (see the Electronic Appendix for the details). Therefore, in order to compute the rank function $\rho(\hat{r})$, we need to count the number of replacements and swaps needed to go from the smallest run possible, i.e. $(a_0, \dots, a_0)$ to $\hat{r}$ along a path in the Hasse diagram where the edges are the “cover” relations. Thus, the rank function is

$$\rho(\hat{r}) = \sum_{i=1}^{N} (N - i + 1)\delta_a(\hat{r}[i]) \; .$$

Example 12. Let us consider $\hat{s} = (a_1, a_0, a_1, a_1)$ from the previous example. Since $\hat{s}[1] = a_1$, from $\hat{0} = (a_0, a_0, a_0, a_0)$ we need a replacement in $\hat{0}[4]$ plus three swaps to reach $\hat{s}[1]$, that is, we have to perform four “cover” operations to go from $\hat{0}$ to (a_1, a_0, a_0, a_0)

and the path in the Hasse diagram has length equal to 4. Since $\hat{s}[3] = a_1$, from (a_1, a_0, a_0, a_0) to (a_1, a_0, a_1, a_0) we need a replacement in the last position plus a swap, that is 2 more "cover" operations. Eventually, with another replacement, we reach $\hat{s}$. Hence $\rho(\hat{s}) = 1 + 2 + 4 = 7 = \sum_{i=1}^{4}(4 - i + 1)\hat{s}[i]$, as stated above.

Therefore, from the natural distance $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$ if $\hat{r} \preceq \hat{s}$, we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$.

Definition 5. Given two comparable runs $\hat{r}, \hat{s} \in R(N)$, with $REL = \{a_0, a_1\}$, the **difference** between $\hat{r}$ and $\hat{s}$ is $\Delta_{\hat{r}\hat{s}} = \sum_{i=1}^{N}(N - i + 1)\big(\delta_a(\hat{s}[i]) - \delta_a(\hat{r}[i])\big)$.

Since the order (8) is only partial, in order to compare differences between intervals we need to restrict our study to a maximal chain, i.e. a totally ordered subset of $R(N)$. Thus, denoted with $\mathscr{C}(R(N))$ a maximal chain of $R(N)$, and given the *less than or equal to* $\preceq_d$ relation, which as in the previous cases coincides with the order relation $\leq$ among real numbers, the relational structure $(\mathscr{C}(R(N)), \preceq_d)$ is a difference structure. This follows from the same discussion we have done for the difference structure in the strong top-heaviness case in the previous section.

Thus, an interval scale measure M on $(\mathscr{C}(R(N)), \preceq_d)$ is given by the rank function itself

$$\mathrm{M}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N}(N - i + 1)\delta_a(\hat{r}[i]) \ , \tag{9}$$

for $\hat{r} \in \mathscr{C}(R(N))$.

AP, RBP, DCG, and ERR are not on an interval scale since there does not exist any positive linear transformation between the above M and any of them. In particular, the next example shows how each of them fails on intervals with same length, i.e. the *Solvability Condition*.

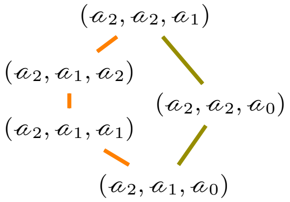
Example 13. Consider the following runs on $R(4)$: $\hat{r} = (a_0, a_1, a_0, a_0), \hat{s} = (a_1, a_0, a_0, a_0), \hat{u} = (a_0, a_0, a_0, a_1)$ and $\hat{v} = (a_0, a_0, a_1, a_0)$. These runs are comparable, that is they belong to the same maximal chain on $R(4)$. Moreover, $\hat{s}$ covers $\hat{r}$ and $\hat{v}$ covers $\hat{u}$, that is the differences $\Delta_{\hat{r}\hat{s}}$ and $\Delta_{\hat{u}\hat{v}}$ are equal.
Hence, an interval scale measure M should satisfy $\mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) = \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u})$, as a consequence of Theorem 1. We use the indicator function δ as gain function g. However, in the case of AP, we have that $RB \cdot (\mathrm{AP}(\hat{s}) - \mathrm{AP}(\hat{r})) = 1 - 1/2 > 1/3 - 1/4 = (\mathrm{AP}(\hat{v}) - \mathrm{AP}(\hat{u})) \cdot RB$, where RB is the recall base. In the case of RBP we have that $\mathrm{RBP}_p(\hat{s}) - \mathrm{RBP}_p(\hat{r}) = (1-p)^2 > (1-p)^2p^2 = \mathrm{RBP}_p(\hat{v}) - \mathrm{RBP}_p(\hat{u})$ since $p < 1$. In the case of DCG we have that $\mathrm{DCG}_2(\hat{s}) - \mathrm{DCG}_2(\hat{r}) = 1 - 1 < 1/\log_2 3 - 1/\log_2 4 = \mathrm{DCG}_2(\hat{v}) - \mathrm{DCG}_2(\hat{u})$. Finally, in the case of ERR we have that $\mathrm{ERR}(\hat{s}) - \mathrm{ERR}(\hat{r}) = 1/2 - 1/4 > 1/6 - 1/8 = \mathrm{ERR}(\hat{v}) - \mathrm{ERR}(\hat{u})$. This proves that none of these measures is an interval scale on $(\mathscr{C}(R(N)), \preceq_d)$.

However, note that AP, RBP, DCG, and ERR are on a ordinal scale with respect the partial order (8) induced by the weak top-heaviness, as demonstrated by [26].

6.2.2 *Multi-graded Relevance Case*

Unfortunately, the multi-graded relevance case is very complicated since the poset $R(N)$ is not graded, as the following example shows.

Example 14. Let, for example, $c = 2$, $N = 3$, $\hat{r} = (a_2, a_1, a_0)$, and $\hat{s} = (a_2, a_2, a_1)$. According to the above partial ordering (8), $\hat{r} \preceq \hat{s}$ and the Hasse diagram associated to $[\hat{r}, \hat{s}]$ is



The Hasse diagram shows that there are two maximal chains on $[\hat{r}, \hat{s}]$ but with different length: the orange one on the left has length 3 while the dark green one on the right has length 2. Thus, $\{a_0, a_1, a_2\}^3$ is not graded.

$R(N)$ not graded means that the length of an interval is not uniquely defined, leading to a very tough case to address. Therefore, it is left for future work to study whether there exist alternative ways, based on more sophisticated algebraic structures, to deal with it.

6.3 Induced Total Ordering

Similarly to the induced total ordering in the set-based case in Section 5.3, we consider the order among runs induced by a measure itself to understand if the measure generates an equally spaced graduation between the codomain values which, in turn, induces an interval scale.

As it happened for Section 5.3, when the induced total ordering coincides with those considered in Sections 6.1 and 6.2, the same results hold.

Let us now consider the multi-graded relevance case and the measures gRBP_p, with $p \neq G/(G+1)$, ERR and DCG: none of these are interval scale measure with respect to the order induced by each of them on $R(N)$. Indeed, given any of these measures, which we denote with M, for any pair $\hat{r}, \hat{s} \in R(N)$ such that $\hat{s}$ covers $\hat{r}$, it can be always found another pair of runs $\hat{u}, \hat{v} \in R(N)$ such that $\hat{v}$ covers $\hat{u}$ but $\mathrm{M}(\hat{s}) - \mathrm{M}(\hat{r}) \neq \mathrm{M}(\hat{v}) - \mathrm{M}(\hat{u})$, that is the *Solvability condition* (considered on the codomain) fails.

Take for example DCG_2, $N = 3$ and $c = 2$: for any possible true multi graded gain function g, i.e. with $g(1) \neq g(2)$, the three biggest runs, with respect to the order induced by DCG, in decreasing order are $\hat{u} = (a_2, a_2, a_2)$, $\hat{s} = (a_2, a_2, a_1)$ and $\hat{r} = (a_2, a_1, a_2)$. We get $\mathrm{DCG}_2(\hat{u}) - \mathrm{DCG}_2(\hat{s}) = \mathrm{DCG}_2(\hat{s}) - \mathrm{DCG}_2(\hat{r})$ iff $g(2)(1 - 2/\log_2 3) = g(1)(1 - 2/\log_2 3)$ which leads to a contradiction. Similar examples can be provided for the other multi-graded measures.

7 Related Work

The relation between the representational theory of measurements and IR evaluation measures has been early investigated by van Rijsbergen [5], [27] in the context of set-based IR measures. In particular, van Rijsbergen [5] exploited *conjoint structures* [9] to study Precision and Recall by considering all the possible Precision and Recall pairs, i.e. $R \times P$, as the empirical set E and then creating a kind of "second order" measure on this set E whose properties are examined, e.g. if this "second order" measure is interval based. In this sense, it resembles what we do in the induced total ordering case, even if in our case we look at a single measure at time and not pairs of them relying on conjoint structures. In general, we take a different approach since we consider system runs as the empirical set E and not the set of all the possible Precision and Recall pairs; moreover, we directly determine if an IR measure is on an interval scale by exploiting the ordering and difference among system runs.

Bollmann and Cherniavsky [4] introduced the *MZ-metric* and, following the example of van Rijsbergen [5], they defined a conjoint structure on the contingency table relevant/not relevant and retrieved/not retrieved in order to determine under which transformations the MZ-metric was on an interval scale. Instead of a conjoint structure on the contingency table, we directly created a difference structure on the set of system runs that can be used to determine if any set-based IR measure is on an interval scale. Moreover, the MZ-metric is not on a interval scale if we use the structure we defined above, as shown in [28]. In addition, Bollmann [3] studied set-based measures by showing that measures complying with a monotonicity and an Archimedean axiom are a linear combination of the number of relevant retrieved documents and the number of not relevant not retrieved documents. We address a completely different issue, that is determining which scales are used by IR measures.

Both Amigó et al. [6], [29] and Moffat [30] studied the properties of rank-based IR measures, in a formal and a numeric way respectively, defining, e.g., how an IR measure should behave when a relevant document is added or removed from a system run. All the identified properties could be exploited to introduce some sort of structure among the system runs but these authors did not do that explicitly. They also did not study what scales are adopted by IR measures, which is the topic of this paper instead.

Mizzaro et al. [7], [31] used the notion of scale and mapping among scales to model different kinds of similarity and to introduce constraints and axioms over them. However, they did not address the problem of determining the actual scales used by an IR measure.

We introduced the partial order based on the replacements and swap properties in [26], where we used it to demonstrate when both set-based and rank-based IR measures are ordinal scales in the binary relevance case. In this work, we go beyond by considering the multi-graded relevance case and investigating interval scales instead.

We started to explore interval scales in the binary relevance case in [28], where we investigated the total order for the set-based IR measures and the total and partial orders for the rank-based IR measures. In this work, we go beyond by considering the multi-graded relevance case, by investigating the partial order for the set-based IR measures.

Overall, this work not only extends in several respects our previous works [26], [28] but it also provides a single coherent framework where all the different types of relevance types and ordering among runs are dealt with. Moreover, we also provide a better and more abstract model, expressing the whole framework and its proofs in a fully symbolic way, while in previous works [26], [28] we relied on specific numerical values of the gain function g to state the definitions and demonstrate the properties.

8 Conclusions and Future Work

We developed a theory of IR evaluation measures to explore whether and when both set-based and rank-based IR measures are interval scales. This is a fundamental question since the validity of the descriptive statistics, such as mean and variance, and the statistical significance tests we daily use to compare IR systems depends on its answer.

The main findings and contributions of the paper are:

- a fully formal framework, based on the representational theory of measurement, accounting for both set-based and rank-based IR evaluation measures as well as both binary and multi-graded relevance, exploring three kinds of ordering: a total order, a partial order, and the order induced by the measures themselves;
- in the case of set-based IR measures:
  - binary relevance: Precision, Recall, and F-measure are interval scales, independently from the adopted order;
  - multi-graded relevance: gP and gR are interval scales only when using a partial order and when the relevance degrees themselves are a ratio scale; when using a total order instead, they are not interval scales;
- in the case of rank-based IR measures
  - binary relevance: when using a total order, RBP is an interval scale only if $p = \frac{1}{2}$ while all the other measures – namely AP, DCG, ERR, and RBP for other values of p – are not. When using a partial order, none of these measures is an interval scale;
  - multi-graded relevance: when using a total order, RBP is an interval scale only if $p = (c + 1)^{-1}$ and when the relevance degrees themselves are a ratio scale while all the other measures – namely AP, DCG, ERR, and RBP for other values of p – are not. In the case of the partial order, we have shown that the multi-graded relevance case produces an overwhelmingly complex structure on the set of runs which cannot be dealt with in the current framework;
- brand new IR measures which guarantee to interval scales – see eq. (3), (5), (7), and (9) – according to the developed framework.

The impact of the findings of this paper goes beyond IR itself since many of the studied measures are widely used also in neighboring fields, such as databases and data mining, which can benefit from a better understanding of the measures they daily use in benchmarking.

As future work, we will investigate these new interval scale measures from an experimental point of view, e.g. by performing an analysis of their robustness to pool downsampling or of their discriminative power, as well as the exploration of how they behave in statistical significance testing with respect to the traditional IR measures which, instead, violate the interval scale assumption.

Furthermore, we will explore alternative notions of distance to possibly achieve interval-like properties. We just started this exploration in the binary relevance case [32] by introducing a notion of distance as a vector instead of as a number, as done in this paper, but we still have to understand what happens in the general multi-graded case.

Finally, a completely new line a research stemming from the foundations laid in this paper could be investigating and formalizing what happens when you have multiple possible disagreeing orderings, as it happens in the case of interassessor agreement and crowd-sourcing.

REFERENCES

[1] M. Sanderson, "Test Collection Based Evaluation of Information Retrieval Systems," *FnTIR*, vol. 4, no. 4, pp. 247–375, 2010.

[2] C. W. Cleverdon, "The Cranfield Tests on Index Languages Devices," *Aslib Proceedings*, vol. 19, no. 6, pp. 173–194, 1967.

[3] P. Bollmann, "Two Axioms for Evaluation Measures in Information Retrieval," in *Proc. of the Third Joint BCS and ACM Symposium on Research and Development in Information Retrieval*, Cambridge University Press, UK, 1984, pp. 233–245.

[4] P. Bollmann and V. S. Cherniavsky, "Measurement-theoretical investigation of the MZ-metric," in *SIGIR 1980)*, 1980, pp. 256–267.

[5] C. J. van Rijsbergen, "Foundations of Evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[6] E. Amigó, J. Gonzalo, and M. F. Verdejo, "A General Evaluation Measure for Document Organization Tasks," in *SIGIR 2013*, 2013, pp. 643–652.

[7] L. Busin and S. Mizzaro, "Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics," in *ICTIR 2013*, 2013, pp. 22–29.

[8] F. Sebastiani, "An Axiomatically Derived Measure for the Evaluation of Classification Algorithms," in *ICTIR 2015*, 2015, pp. 11–20.

[9] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement. Additive and Polynomial Representations*. Academic Press, New York, USA, 1971, vol. 1.

[10] S. S. Stevens, "On the Theory of Scales of Measurement," *Science, New Series*, vol. 103, no. 2684, pp. 677–680, 1946.

[11] B. A. Carterette, "Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments," *ACM TOIS*, vol. 30, no. 1, pp. 4:1–4:34, 2012.

[12] S. Robertson, "On GMAP: and Other Transformations," in *CIKM 2006*, 2006, pp. 78–83.

[13] J. Kekäläinen and K. Järvelin, "Using Graded Relevance Assessments in IR Evaluation," *JASIST*, vol. 53, no. 13, pp. 1120—1129, 2002.

[14] R. P. Stanley, *Enumerative Combinatorics – Volume 1*, 2nd ed., ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, UK, 2012, vol. 49.

[15] S. Foldes, "On distances and metrics in discrete ordered sets," *arXiv.org, Combinatorics (math.CO)*, vol. arXiv:1307.0244, 2013.

[16] G. B. Rossi, *Measurement and Probability. A Probabilistic Theory of Measurement with Applications*. Springer-Verlag, New York, USA, 2014.

[17] P. F. Velleman and L. Wilkinson, "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading," *The American Statistician*, vol. 47, no. 1, pp. 65–72, 1993.

[18] C. J. van Rijsbergen, "Retrieval effectiveness," in *Information Retrieval Experiment*, Butterworths, London, United Kingdom, 1981, pp. 32–43.

[19] D. E. Knuth, *The Art of Computer Programming – Volume 2: Seminumerical Algorithms*, 2nd ed. Addison-Wesley, USA, 1981.

[20] S. Miyamoto, "Generalizations of Multisets and Rough Approximations," *International Journal of Intelligent Systems*, vol. 19, no. 7, pp. 639–652, 2004.

[21] A. Moffat and J. Zobel, "Rank-biased Precision for Measurement of Retrieval Effectiveness," *ACM TOIS*, vol. 27, no. 1, pp. 2:1–2:27, 2008.

[22] T. Sakai and N. Kando, "On information retrieval metrics designed for evaluation with incomplete relevance assessments," *Information Retrieval*, vol. 11, no. 5, pp. 447–470, 2008.

[23] K. Järvelin and J. Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," *ACM TOIS*, vol. 20, no. 4, pp. 422–446, 2002.

[24] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan, "Expected Reciprocal Rank for Graded Relevance," in *CIKM 2009*, 2009, pp. 621–630.

[25] S. Mizzaro, "Relevance: The Whole History," *JASIST*, vol. 48, no. 9, pp. 810–832, 1997.

[26] M. Ferrante, N. Ferro, and M. Maistro, "Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness," in *ICTIR 2015*, 2015, pp. 21–30.

[27] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. Butterworths, London, England, 1979.

[28] M. Ferrante, N. Ferro, and S. Pontarollo, "Are IR Evaluation Measures on an Interval Scale?" in *ICTIR 2017*, 2017, pp. 67–74.

[29] E. Amigó, J. Gonzalo, J. Artiles, and M. F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information Retrieval*, vol. 12, no. 4, pp. 461–486, 2009.

[30] A. Moffat, "Seven Numeric Properties of Effectiveness Metrics," in *AIRS 2013*. LNCS 8281, Springer, Germany, 2013, pp. 1–12.

[31] E. Maddalena and S. Mizzaro, "Axiometrics: Axioms of Information Retrieval Effectiveness Metrics," in *Proc. 6th International Workshop on Evaluating Information Access (EVIA 2014)*, National Institute of Informatics, Tokyo, Japan, 2014, pp. 17–24.

[32] M. Ferrante, N. Ferro, and S. Pontarollo, "An Interval-Like Scale Property for IR Evaluation Measures," in *Proc. 8th International Workshop on Evaluating Information Access (EVIA 2017)*, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2008/, 2017, pp. 10–15.

Marco Ferrante received the PhD degree in Functional Analysis and Applications from the SISSA in 1993. He is professor of Probability Theory at the University of Padua. His research interests include probability theory, stochastic calculus, stochastic processes and their applications to biological models and information retrieval. More about his research at http://www.math.unipd.it/~ferrante/.

Nicola Ferro is associate professor in computer science at the University of Padua. His research interests include information retrieval, its experimental evaluation, multilingual information access and digital libraries. He is the coordinator of the CLEF evaluation initiative. More about his research at http://www.dei.unipd.it/~ferro/.

Silvia Pontarollo received the Master degree in Mathematics from the University of Padua in 2017. At present she is a early stage researcher at the Department of Mathematics "Tullio Levi-Civita" of the University of Padua. More about her research at http://www.math.unipd.it/~spontaro/.

Electronic Appendix

A General Theory of IR Evaluation Measures

Marco Ferrante, Nicola Ferro and Silvia Pontarollo

1 Poset, Lattice, and Hasse Diagram

In this section, following [1], we recall some definitions and results about posets.

A partially ordered set P, **poset** for short, is a set with a partial order $\preceq$ defined on it. A **partial order** $\preceq$ is a binary relation over P which is reflexive, antisymmetric and transitive (see Table 1). Given $s, t \in P$, we say that s and t are *comparable* if $s \preceq t$ or $t \preceq s$, otherwise they are *incomparable*. P is called **bounded** if it has a maximum and a minimum element, namely $\hat{1}, \hat{0} \in P$ such that for every $s \in P$, $s \preceq \hat{1}$ and $\hat{0} \preceq s$.

Example 1. Given a set A, let us consider the power set $\{E : E \subseteq A\} = 2^A$ and then define the following ordering: given $E, F \in 2^A$, we say that $E \preceq F$ if $E \subseteq F$. 2^A is the set of all subsets of A *ordered by inclusion* and it is a poset.

A **total order** over a set P is a partial order where every pair of elements are comparable, whereas a **weak order** is a total order without the antisymmetric relation (see Table 1).

Example 2. The power set 2^A with the ordering defined before is not totally ordered unless $|A| = 1$, i.e. it contains just one element.

A further example of a weak order is given by the order induced on a set A by a non injective real function f, for example the height of a set of people. Indeed, if we define the ordering on A: $a \preceq b$ iff $f(a) \leq f(b)$, the antisymmetry holds true iff f is injective. In the height case, the ordering is weak iff there are at least two individuals with the same height.

Table 1
Characterization of possible ordering, namely weak, partial and total orders, on a given set P, using $a, b, c \in P$.

	Weak	Partial	Total
Reflexivity $a \preceq a \ \forall a \in P$	✓	✓	✓
Antisymmetry $a \preceq b$ and $b \preceq a \Rightarrow a = b$	X	✓	✓
Transitivity $a \preceq b, b \preceq c \Rightarrow a \preceq c$	✓	✓	✓
Comparability for any $a, b \in P, \ a \preceq b$ or $b \preceq a$	✓	X	✓

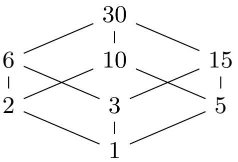
A **(closed) interval** is a subset of P defined as $[s, t] := \{u \in P : s \preceq u \preceq t\}$, where $s, t \in P$ and $s \preceq t$. Moreover, we say that t **covers** s if $s \preceq t$ and $[s, t] = \{s, t\}$, that is there does not exist $u \in P$ such that $s \prec u \prec t$.

We can represent a finite poset P by using the **Hasse diagram** which is a graph where vertices are the elements of P, edges represent the *covers* relations, and if $s \prec t$ then s is below t in the diagram. Note that if $s, t \in P$ lie on the same horizontal level of the diagram, then they are incomparable by construction. Furthermore, elements on different levels may be incomparable as well.

Example 3. Let $N = 30$ and P the set of all divisors of N, that is $P = \{1, 2, 3, 5, 6, 10, 15, 30\}$. Let us define the following ordering on P: given $a, b \in P$ we say that $a \preceq b$ if a divide b. P is a poset with respect to the ordering $\preceq$, and its Hasse diagram is:

- M. Ferrante and S. Pontarollo are with Department of Mathematics, University of Padua, Italy E-mail: {ferrante, spontaro}@math.unipd.it
- N. Ferro is with Department of Information Engineering, University of Padua, Italy E-mail: ferro@dei.unipd.it

Manuscript received January 16, 2018; revised April 10, 2018; accepted May 10, 2018



2 covers 1, since $[1,2] = \{1,2\}$, while 6 doesn't, since $[1,6] = \{1,2,6\}$. $2,\ 3$ and 5 are on the same horizontal level and they are incomparable since, for example, neither 2 divides 3 nor 3 divides 2. Moreover 3 and 10 lie on different levels and they are incomparable.

A subset C of a poset P is a **chain** if any two elements of C are comparable: a chain is a totally ordered subset of a poset. If C is a finite chain, the **length** of C, $\ell(C)$, is defined by $\ell(C) = |C| - 1$. A **maximal chain** of P is a chain that is not a proper subset of any other chain of P. P is a poset of **finite rank** if the length of the longest maximal chain of P is finite. Referring to the previous example, a chain is the subset $\{1, 10, 30\}$, while an example of maximal chain is the subset $\{1, 2, 10, 30\}$; moreover P is of finite rank since every maximal chain has length equal to 3.

If every maximal chain of P has the same length n, we say that P is **graded of rank n**; in particular there exists a unique function $\rho : P \to \{0, 1, \ldots, n\}$, called the **rank function**, such that $\rho(s) = 0$, if s is a minimal element of P, and $\rho(t) = \rho(s) + 1$, if t covers s.

Example 4. The poset $P = \{1, 2, 3, 5, 6, 10, 15, 30\}$ defined in the previous example is graded of rank 3 since any maximal chain of P, which are $\{1, 2, 6, 30\}, \{1, 2, 10, 30\}, \{1, 3, 6, 30\}, \{1, 3, 15, 30\}, \{1, 5, 10, 30\}\{1, 5, 15, 30\}$, has length equal to 3. The rank function ρ is defined as follows: $\rho(1) = 0$, $\rho(2) = \rho(3) = \rho(5) = 1$, $\rho(6) = \rho(10) = \rho(15) = 2$ and $\rho(30) = 3$.

Finally, since any interval on a graded poset is graded, the **length of an interval** $[s, t]$ is given by $\ell(s, t) := \ell([s, t]) = \rho(t) - \rho(s)$.

Given $s, t \in P$, an upper bound of {s,t} is any $u \in P$ such that $s \preceq u$ and $t \preceq u$. A *least upper bound* (or supremum) of {s,t}, denoted by $s \vee t$, is an upper bound u such that every other upper bound $v \in P$ of {s,t} satisfies $v \succeq u$. Dually it is defined the *greatest lower bound* (or infimum) $s \wedge t$. Note that not every pair of elements in a poset has necessarily the infimum or the supremum. A poset L for which every pair of elements has a least upper bound and a greatest lower bound is called **lattice**.

Example 5. The poset P from the previous example is a lattice: for example given the elements $2, 15 \in P$, we get that $2 \vee 15 = 30$ and $2 \wedge 15 = 1$, and for any other pair of elements $s, t \in P$ one has $s \vee t =$ *least common multiple* and $s \wedge t =$ *greatest common divisor*.

Thanks to the following Lemma, see [2] for its proof, we have a easy-to-prove sufficient condition for a poset to be a lattice.

Lemma 1. Let P be a bounded poset of finite rank such that, for any $s, t \in P$, if both s and t cover an element u, then $s \vee t$ exists in P. Then P is a lattice.

Moreover, as Stanley [1] shows, the Proposition below give us a necessary and sufficient condition for a finite lattice to be graded.

Proposition 2. Let L be a finite lattice. The following two conditions are equivalent:

i. L is graded, and the rank function ρ of L satisfies

$$\rho(s) + \rho(t) \geq \rho(s \wedge t) + \rho(s \vee t),$$

for all $s, t \in L$.

ii. If s and t both covers $s \wedge t$, then $s \vee t$ covers both s and t.

Finally, Foldes [3] proves that in a graded poset P the length $\ell(\cdot, \cdot)$ of any interval, also called the **natural distance**, equals the length of the shortest path connecting the two endpoints of the interval in its Hasse diagram.

2 Representational Theory of Measurement

2.1 Measurement

> **Measurement** is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them accordingly to clearly defined rules.

The above definition of measurement [4] highlights several important facts about it. An *entity* is an object or an event existing in the real world, which is described by means of its identifying characteristics – the *attributes* – that allow us to distinguish one entity from another. Therefore, neither we measure things nor we measure attributes but rather we measure attributes of things. In order to make it easier to work with and process them, we often define the attributes in terms of *numbers* or, more in general, *symbols*. In doing this, we have to take care of preserving in the numerical domain the

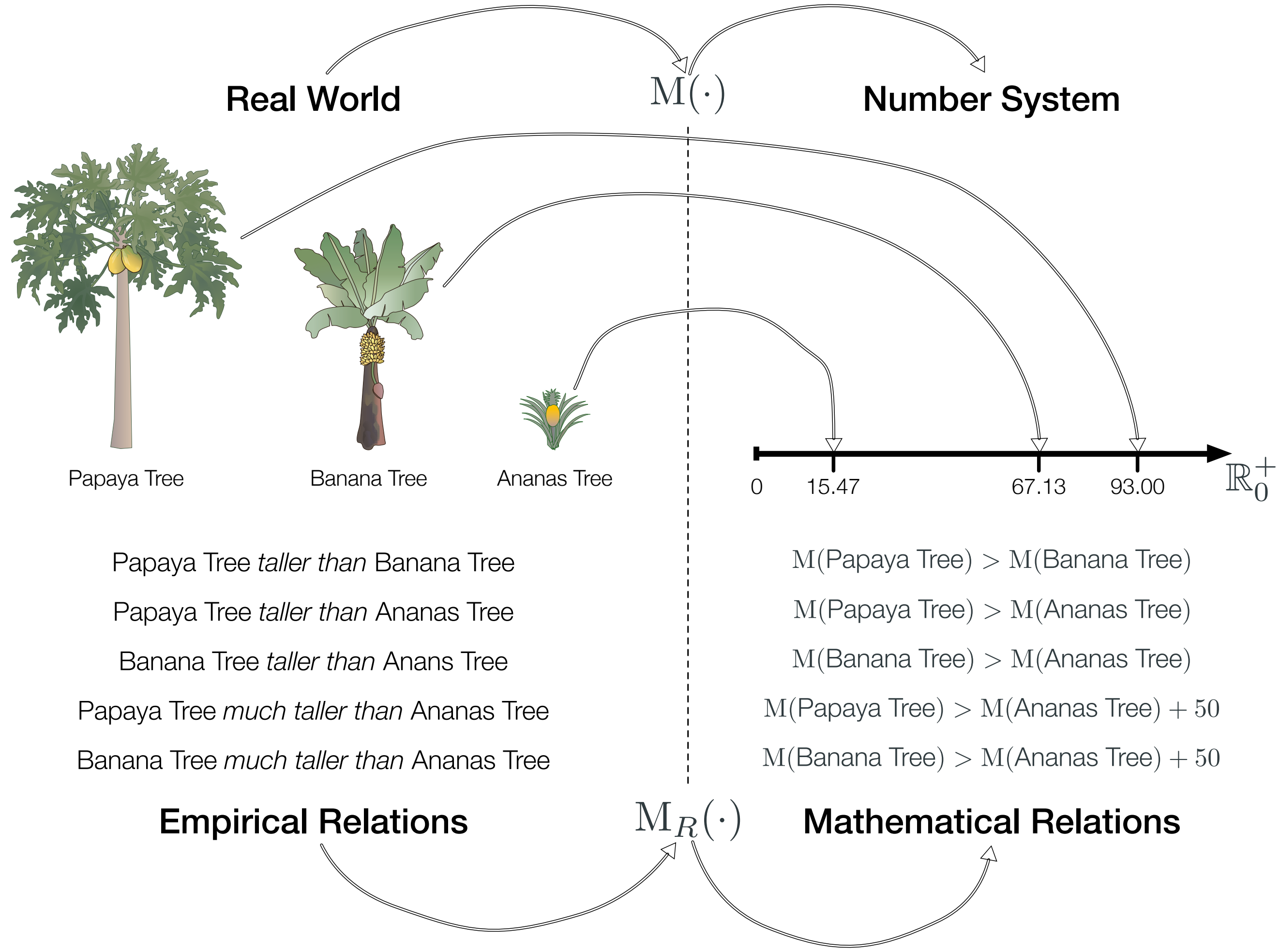

Real World
M(·)
Number System
Papaya Tree
Banana Tree
Ananas Tree
0
15.47
67.13
93.00
$\mathbb{R}_0^+$
Papaya Tree *taller than* Banana Tree
Papaya Tree *taller than* Ananas Tree
Banana Tree *taller than* Anans Tree
Papaya Tree *much taller than* Ananas Tree
Banana Tree *much taller than* Ananas Tree
M(Papaya Tree) > M(Banana Tree)
M(Papaya Tree) > M(Ananas Tree)
M(Banana Tree) > M(Ananas Tree)
M(Papaya Tree) > M(Ananas Tree) + 50
M(Banana Tree) > M(Ananas Tree) + 50
Empirical Relations
$\mathrm{M}_R(\cdot)$
Mathematical Relations

Figure 1. Example of an empirical relations for the attribute "height" of a tree and its representation condition.

relationships we see among the (attributes of the) entities in the real world. In other terms, consider that a number or symbol is assigned by measurement to the attribute of an entity, and other number or symbols are assigned by the same process to other manifestations of that attribute. Then the logical relations between the numbers or symbols, in the symbolisation system adopted, have to imply and to be implied by empirical relations between the attribute manifestations [5]. This allows us to reason about and work with entities by means of the numeric (symbolic) representation of their attributes and to guarantee that the conclusions we draw about the entities via this numeric processing are valid. The construction of such a suitable mapping between the real and symbolic worlds is known as the *representation problem*.

The *representational theory of measurement* [4], [6] aims at providing a formal basis to our intuition about the way the world works. According to the above definition of measurement, the numbers or symbols we collect as measures about the attributes of the entities we examine should be such that their processing and manipulation maintain the relationships among the actual entities under examination in the real world. Therefore, at the basis of measurement, there are the relationships among entities and how we empirically observe them. This way of proceeding frees us from dealing directly with real world entities but we can manipulate the measures associated to them in order to understand and learn about them and it is part of that "measurement as knowledge advancement" discussed in the previous section.

Consider the example shown in Figure 1 about the attribute "heigth" of a tree, where the real world is constituted by just three entities: a Papaya tree, a Banana tree, and an Ananas tree. We can easily see that some trees are "taller than" others: for example, we can see that the Papaya tree is "taller than" the Banana and Ananas ones while the the Banana tree is "taller than" the Ananas one. Moreover, we can have multiple relations on the same set of entities. For example, we can see that both a Papaya and a Banana tree are "much taller than" an Ananas one.

"Taller than" and "much taller than" are *empirical relations* for height (of a tree) and we can think at them as a *mappings* from the real world to a formal mathematical world. Indeed, they can be considered as a mapping from the set of trees to the set of real numbers, provided that, for example, whenever a Papaya tree is "taller than" a Banana one, any measure of height assigns a higher number to the Banana tree than to the Papaya one.

This is the so called *representation condition* which ensures that a measurement must map attributes of entities into numbers (symbols) and empirical relations into numerical (symbolic)ones so that the empirical relations imply and are implied by the numerical (symbolic) ones.

More formally [6], [7], a *relational structure* is an ordered pair $\mathbf{X} = \langle X, R_X \rangle$ of a domain set X and a set of relations R_X on X, where the relations in R_X may have different arities, i.e. they can be unary, binary, ternary relations and so on. Given two relational structures $\mathbf{X}$ and $\mathbf{Y}$, a *homomorphism* $\mathbf{M} : \mathbf{X} \to \mathbf{Y}$ from $\mathbf{X}$ to $\mathbf{Y}$ is a mapping $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ where:

- M is a function that maps X into $\mathrm{M}(X) \subseteq Y$, i.e. for each element of the domain set there exists at least one, but not necessarily only one, corresponding image element;
- M_R is a function that maps R_X into $\mathrm{M}_R(R_X) \subseteq R_Y$ such that $\forall r \in R_X$, r and $\mathrm{M}_R(r)$ have the same arity, i.e. for each relation on the domain set there exists one (and it is usually, and often implicitly, assumed: and only one) corresponding image relation,

with the condition that $\forall r \in R_X, \forall x_i \in X$, if $r(x_1, \ldots, x_n)$ then $\mathrm{M}_R(r)\Big(\mathrm{M}(x_1), \ldots, \mathrm{M}(x_n)\Big)$, i.e. if a relation holds for some elements of the domain set then the image relation must hold for the image elements.

Note that we talk about a homomorphism rather than an isomorphism because M is generally not one-to-one; in general $M(a) = M(b)$ does not mean that two trees are identical but merely of equal height.

A relational structure $\mathbf{E}$ is called *empirical* if its domain set E spans over the entities under consideration, e.g. the set of trees in Figure 1; a relational structure $\mathbf{S}$ is called *symbolic* if its domain set S spans over a given set of symbols, e.g. the set of positive real numbers $\mathbb{R}_0^+ = \{x \in \mathbb{R} \mid x \geq 0\}$ in Figure 1.

We can now provide a more precise definition of measurement on the basis of the just introduced concepts [4].

> We define **measurement** as a homomorphism $\mathbf{M} = \langle \mathrm{M}, \mathrm{M}_R \rangle$ from the real world to a symbolic world. Consequently, a **measure** is the number or symbol assigned to an entity by this mapping in order to characterize an attribute.

In the case of the example of Figure 1 we have the following measurement homomorphism:

- *empirical relational structure* $\mathbf{E} = \langle E, R_E \rangle$, where $E = \{\text{Ananas Tree}, \text{Banana Tree}, \text{Papaya Tree}\}$ and $R_E = \{R_{E_1}, R_{E_2}\}$ with the relation "taller than" expressed by the relation $R_{E_1} = \{(\text{Papaya Tree}, \text{Banana Tree}), (\text{Papaya Tree}, \text{Ananas Tree}), (\text{Banana Tree}, \text{Ananas Tree})\}$ and the relation "much taller than" expressed by the relation $R_{E_2} = \{(\text{Papaya Tree}, \text{Ananas Tree}), (\text{Banana Tree}, \text{Ananas Tree})\}$;
- *symbolic relational structure* $\mathbf{S} = \langle S, R_S \rangle$, where $S = \mathbb{R}_0^+ = \{x \in \mathbb{R} \mid x \geq 0\}$ and and $R_S = \{R_{S_1}, R_{S_2}\}$ with $R_{S_1} = \{(x, y) \in S \times S \mid x > y\}$ and $R_{S_2} = \{(x, y) \in S \times S \mid x > y + 50\}$;
- *representation condition* with M such that $\mathrm{M}(\text{Ananas Tree}) = 15.47$, $\mathrm{M}(\text{Banana Tree}) = 67.13$, and $\mathrm{M}(\text{Papaya Tree}) = 93.00$; and with $\mathrm{M_R}$ such that $\mathrm{M}_R(R_{E_1}) = R_{S_1}$ and $\mathrm{M}_R(R_{E_2}) = R_{S_2}$.

Note that the representation condition complies with the additional constraint $\forall r \in R_X, \forall x_i \in X$, if $r(x_1, \ldots, x_n)$ then $\mathrm{M}_R(r)\Big(\mathrm{M}(x_1), \ldots, \mathrm{M}(x_n)\Big)$; for example $r_{E_1} = (\text{Papaya Tree}, \text{Banana Tree}) \in R_{E_1}$ corresponds to $93.00 > 67.13$, i.e. $r_{S_1} = (93.00, 67.13) \in R_{S_1}$ and $r_{E_2} = (\text{Papaya Tree}, \text{Ananas Tree}) \in R_{E_2}$ corresponds to $93.00 > 15.47 + 50.00$, i.e. $r_{S_2} = (93.00, 15.47) \in R_{S_2}$.

As an additional example, consider a set of rods A [6] where a comparison relation $\succ$ and a concatenation operation $\circ$ among rods exist. Note that $\succ$ is a binary relation on the set of rods A while $\circ$ is a ternary one which assigns to each pair of rods a third rod representing their concatenation. Then, the empirical relational structure $\mathbf{E} = \langle A, \succ, \circ \rangle$ can be mapped into the symbolic relational structure $\mathbf{S} = \langle \mathbb{R}_0^+, >, + \rangle$, using as mapping function $M(\cdot)$ the length of a rod so that $a \succ b \Leftrightarrow M(a) > M(b)$ and $M(a \circ b) = M(a) + M(b)$.

2.2 Scales

As discussed in the previous section, the goal of measurement is to be able to process and manipulate data in the numerical system in order to understand and learn about attributes of entities in real world.

It is thus natural to wonder whether all the measurements are the same from the point of view of the processing and manipulation you can perform with them and how they affect the kind of analyses that can be conducted. To this end, we refer to the previously introduced measurement homomorphism $\mathbf{M}$ as a **measurement scale** [4]; when the empirical and symbolic relational structures are well known and understood, they are often left implicit and we talk about $\mathrm{M}(\cdot)$ as a (measurement) scale.

The idea behind the formal definition of scale types is that if we have a measurement homomorphism $\mathbf{M}$ for an attribute of an entity with respect to an empirical relational structure, we want to know what other measurements, i.e. what other homomorphisms, exist that are also acceptable. In order to clearly define what "acceptable" means, we introduce the notion of *permissible transformations*.

A permissible transformation [6] is a mapping $\mathbf{M} \to \mathbf{M}'$ where $\mathbf{M}$ and $\mathbf{M}'$ are both homomorphisms of an empirical relational structure $\mathbf{E} = \langle E, R_E \rangle$ into the *same* symbolic relational structure $\mathbf{S} = \langle S, R_S \rangle$. In other terms we seek out a transformation $\mathrm{M} \to \mathrm{M}'$ between two mappings (scales) M and M' of the attributes of real world entities into numbers/symbols which preserves the correspondence M_R between the empirical and symbolic relational structures.

This is known as the *uniqueness problem* since it boils down to identify the family of measurement homomorphisms between an empirical relational structure and a symbolic one which can be univocally mapped one into another by means of a permissible transformation.

Therefore, the complete and formal definition of a measurement usually consists of two parts: a *representation theorem* which asserts the existence of a homomorphism $\mathbf{M}$ into a particular symbolic relational structure (and usually provides also the means to construct it) and a *uniqueness theorem* which sets forth the permissible transformations $\mathbf{M} \rightarrow \mathbf{M}'$ that also yield homomorphisms into the same symbolic relational structure [6].

> A statement involving measurement scales is **meaningful** if and only if its truth value is unchanged whenever permissible transformations are applied to all scales in question.

Meaningfulness [8], [9], [10] is a central concept to clearly shape and define the questions discussed above: according to the adopted measurement scales, what processing, manipulation, and analyses can be conducted and what can we tell about the conclusions drawn from such processing?

For example, the statement "A banana weights more than an ananas" is meaningful even if it is clearly false; indeed, its truth value, i.e. false, does not change whatever weight scale you use (kilograms, pounds, and so on). On the other hand, the statement "Today, the temperature in Rome is twice as the one in Oslo" is not meaningful even it might be true on some days; indeed, if, on the Celsius scale, the temperature today in Rome is 20 °C and in Oslo is 10 °C then it is true but the same temperature becomes, on the Fahrenheit scale, 68 °F in Rome and 50 °F in Oslo and so the statement becomes false; therefore it is not meaningful since it changes its truth value passing from a scale to another one.

Therefore, meaningfulness is a distinct concept from the one of truth of a statement and it is somehow close to the notion of invariance in geometry, since the truth value of a statement stays the same independently from the permissible scales used to express it.

We can classify five major types of scales [4], [11]:

- nominal;
- ordinal;
- interval;
- ratio.

which will be discussed in detail in the following sections. As we will see, these scales introduce increasing constraints in terms of permissible transformations among them, i.e. the functions mapping from one to another are lesser and lesser generic. On the other hand, we can consider these scales to be in increasing order of richness, since they allow us to perform more and more sophisticated operations on them.

2.2.1 Nominal Scale

A *nominal scale* is used when entities of the real world can be placed into different classes or categories on the basis of their attribute under examination. The main characteristics of the nominal scale are [4]:

- the empirical relational structure consists only of different classes without any notion of ordering among them;
- any distinct numeric (symbolic) representation of the classes is an acceptable measure but there is no notion of magnitude associated with numbers or symbols;
- the class of permissible transformations is the set of all *one-to-one mappings*, i.e. bijective functions

$$\mathrm{M}' = \mathrm{f}(\mathrm{M})$$

From the properties of the nominal scales, it turns out that any arithmetic operation on the numeric representation is not meaningful and the only allowable operations basically are counting number of items in each class that is, in statistical terms, mode and frequency.

As an example, consider a classification of people by their spoken language (English, French, German, Italian, and so on). We could define the two following measurements:

$$\mathrm{M}_1 = \begin{cases} 1 & \text{if English} \\ 2 & \text{if French} \\ 3 & \text{if German} \\ 4 & \text{if Italian} \\ \cdots & \text{if } \cdots \end{cases} \qquad \mathrm{M}_2 = \begin{cases} 20 & \text{if English} \\ 30 & \text{if French} \\ 13 & \text{if German} \\ 11 & \text{if Italian} \\ \cdots & \text{if } \cdots \end{cases}$$

both M_1 and M_2 are valid measures, which can be related with a one-to-one mapping.

Suppose now that we have observe a set of 10 people, where 5 people speak English, 3 German, 1 French, and 1 Italian. According to M_1 we would have $P_1 = [1\ 1\ 1\ 1\ 1\ 3\ 3\ 3\ 2\ 4]$ while according to M_2 we would have $P_2 = [20\ 20\ 20\ 20\ 20\ 13\ 13\ 13\ 30\ 11]$. In both cases, the statement "The most spoken language is English" is meaningful since, if we compute the mode of the values, it is 1 in the case of M_1 and 20 in the case of M_2 which both correspond to

English speaking people. On the other hand, the statement "The highest quartile consist of Italian speaking people" is not meaningful, since it is true with 4 corresponding to Italian in the case of M_1 but is is false with 30 corresponding to French in the case of M_2. Indeed, the first statement about the mode involves just counting, which is an allowable operation for a nominal scale, while the second statement about the highest quartile requires a notion of ordering not present in a nominal scale.

2.2.2 Ordinal Scale

The *ordinal scale* can be considered as a nominal scale where, in addition, there is a notion of ordering among the different classes or categories. Its main characteristics are [4]:

- the empirical relational structure consists of classes that are ordered with respect to the attribute under examination;
- any distinct numeric (symbolic) representation which preserves the ordering is acceptable but numbers (symbols) represent only ranking. Therefore, addition, subtraction or other operations have no meaning;
- the class of permissible transformations is the set of all the *monotonic increasing functions*, since they preserve ordering

$$\mathrm{M}' = \mathrm{g}(\mathrm{M})$$

From the properties of the ordinal scales, it turns out that, besides the operations already allowed for nominal scales, i.e. mode and frequency, also quantiles and percentiles are appropriate, since there is a notion of ordering.

As an example, the European Commission Regulation 607/2009 [12] sets the following increasing scale of taste to classify sparkling wines on the basis of their sugar content: *pas dosè* (brut nature), *extra brut*, *brut*, *extra dry*, *sec* (dry), *demi-sec* (medium dry), and *doux* (sweet).

We could introduce two alternative mappings M_1 and M_2 of the above wine scale which can be transformed one into another by means of a monotonic transformation $\mathrm{M}_2 = \mathrm{fibonacci}(\mathrm{M}_1 + 1)$.

$$\mathrm{M}_1 = \begin{cases} 1 & \text{if } \textit{pas dosè} \\ 2 & \text{if } \textit{extra brut} \\ 3 & \text{if } \textit{brut} \\ 4 & \text{if } \textit{extra dry} \\ 5 & \text{if } \textit{sec} \\ 6 & \text{if } \textit{demi-sec} \\ 7 & \text{if } \textit{doux} \end{cases} \qquad \mathrm{M}_2 = \begin{cases} 1 & \text{if } \textit{pas dosè} \\ 2 & \text{if } \textit{extra brut} \\ 3 & \text{if } \textit{brut} \\ 5 & \text{if } \textit{extra dry} \\ 8 & \text{if } \textit{sec} \\ 13 & \text{if } \textit{demi-sec} \\ 21 & \text{if } \textit{doux} \end{cases}$$

Suppose now that we have two wineries. The first winery W^A produced five bottles as follows: *pas dosè*, *extra brut*, *brut*, *extra dry*, and *doux*; the second one W^B produced five bottles as follows: *pas dosè*, *pas dosè*, *pas dosè*, *demi-sec*, and *doux*. Therefore, according to the scale M_1, we have $W_1^A = [1\ 2\ 3\ 4\ 7]$ and $W_1^B = [1\ 1\ 1\ 6\ 7\ ]$; while according to the scale M_2, we have $W_2^A = [1\ 2\ 3\ 5\ 21]$ and $W_2^B = [1\ 1\ 1\ 13\ 21\ ]$. The statement "The median of the first winery is greater than the one of the second winery" is meaningful since $3 > 1$ with M_1 is true as well as $3 > 1$ with M_2; so we could safely say that the first winery tends to produce little more sweet wines than the second one. On the other hand, the statement "The average of the first winery is greater than the one of the second winery" is not meaningful since $3.40 > 3.20$ with M_1 is true but $6.40 > 7.40$ with M_2 is false, which would lead us to draw basically opposite conclusions based on the scale we use. Indeed, the first statement about the median involves just a notion of ordering which is present in an ordinal scale, while the second statement about the average requires to sum values, which is not an allowable operation for an ordinal scale.

2.2.3 Interval Scale

Besides relying on ordered classes, the *interval scale* also captures information about the size of the intervals that separate the classes. Its main characteristics are [4]:

- the empirical relational structure consists of classes that are ordered with respect to the attribute under examination and where the size of the "gap" among two classes is somehow understood;
- it preserves order, as an ordinal scale, and differences among classes but not ratios are meaningful. Therefore, addition and subtraction are acceptable operations but not multiplication and division;
- the class of permissible transformations is the set of all *affine transformations*

$$\mathrm{M}' = \alpha\mathrm{M} + \beta,\ \alpha > 0$$

which is equivalent to say that ratios of intervals are invariant:

$$\frac{\mathrm{M}'(a) - \mathrm{M}'(b)}{\mathrm{M}'(c) - \mathrm{M}'(d)} = \frac{[\alpha\mathrm{M}(a) + \beta] - [\alpha\mathrm{M}(b) + \beta]}{[\alpha\mathrm{M}(c) + \beta] - [\alpha\mathrm{M}(d) + \beta]} = \frac{\mathrm{M}(a) - \mathrm{M}(b)}{\mathrm{M}(c) - \mathrm{M}(d)}$$

From the properties of the interval scales, it turns out that, besides the operations of nominal and ordinal scales, also mean and standard deviation are allowable since they depend on sum and subtraction.

To show this in the case of the mean, let $\{x_1, x_2, \dots, x_n\}$ and $\{y_1, y_2, \dots, y_n\}$ be two sets of entities for which some attribute can be measured on an interval scale. Our goal is to demonstrate that the statement "The mean of x_i is greater than the mean of y_i" is meaningful, as it is shown below:

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{M}'(x_i) > \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}'(y_i) \;\Leftrightarrow\; \frac{1}{n}\sum_{i=1}^{n}\left[\alpha\mathrm{M}(x_i)+\beta\right] > \frac{1}{n}\sum_{i=1}^{n}\left[\alpha\mathrm{M}(y_i)+\beta\right] \;\Leftrightarrow$$
$$\alpha\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{M}(x_i)\right)+\beta > \alpha\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{M}(y_i)\right)+\beta \;\Leftrightarrow\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}(x_i) > \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}(y_i)$$

A typical example of interval scale is temperature on the Fahrenheit or Celsius scales where the affine transformation $F = \frac{9}{5}C + 32$ allows us to pass from one to the another. As previously discussed the statement 'Today, the temperature in Rome is twice as the one in Oslo" is not meaningful. Nevertheless, interval scales preserve ratios among intervals, so the statement 'Today the difference in temperature between Rome and Oslo is twice as the one month ago" is meaningful. Indeed, if, on the Celsius scale, the temperature today in Rome is 20 °C and in Oslo is 10 °C while one month ago it was 12 °C and 7 °C, leading to $20 - 10 = 10$ which is twice $12 - 7 = 5$, on the Fahrenheit scale we would have $68 - 50 = 18$ which is twice $53.6 - 44.6 = 9$.

2.2.4 Ratio Scale

The *ratio scale* is the most powerful one since it allows us to compute ratios among the different classes to say statements like "entity x is twice as entity y". Its main characteristics are [4]:

- the empirical relational structure consists of classes that are ordered, where there is a notion of "gap" among two classes and where the "proportion" among two classes is somehow understood;
- there is a zero element, representing the total lack of the attribute;
- it preserves order and differences as well as ratios are meaningful. Therefore, all the arithmetic operations are allowed;
- the class of permissible transformations is the set of all *similarity transformations*

$$\mathrm{M}' = \alpha\mathrm{M},\ \alpha > 0$$

From the properties of the ratio scales, it turns out that, besides the operations of nominal, ordinal and interval scales, also geometric and harmonic mean are allowable since they depend on multiplication and division.

To show this in the case of the geometric mean, let $\{x_1, x_2, \dots, x_n\}$ and $\{y_1, y_2, \dots, y_n\}$ be two sets of entities for which some attribute can be measured on an interval scale. Our goal is to demonstrate that the statement "The geometric mean of x_i is greater than the mean of y_i" is meaningful, as it is shown below:

$$\sqrt[n]{\prod_{i=1}^{n}\mathrm{M}'(x_i)} > \sqrt[n]{\prod_{i=1}^{n}\mathrm{M}'(y_i)} \;\Leftrightarrow\; \sqrt[n]{\prod_{i=1}^{n}\alpha\mathrm{M}(x_i)} > \sqrt[n]{\prod_{i=1}^{n}\alpha\mathrm{M}(y_i)} \;\Leftrightarrow$$
$$\alpha\sqrt[n]{\prod_{i=1}^{n}\mathrm{M}(x_i)} > \alpha\sqrt[n]{\prod_{i=1}^{n}\mathrm{M}(y_i)} \;\Leftrightarrow\; \sqrt[n]{\prod_{i=1}^{n}\mathrm{M}(x_i)} > \sqrt[n]{\prod_{i=1}^{n}\mathrm{M}(y_i)}$$

A typical example of ratio scale is length which can be expressed on different scales (kilometers, meters, centimeters, miles, feet, inches, and so on) which can all be mapped one into another via a similarity transformation. For example, to pass from kilometers (M_1) to miles (M_2), we have the following transformation $\mathrm{M}_2 = 0.62\mathrm{M}_1$. Moreover, there is a zero element, somewhat abstract, which is an entity with zero-length intended as the limit for things that get smaller and smaller.

For example, if the air distance between Rome and Padua is (about) 400 kilometers and the air distance among Rome and Oslo is (about) $2,000$ kilometers, the statement "Rome and Oslo are five times as distant as Rome and Padua" is meaningful, even expressed in miles, since $248.54 \backsim 5 \cdot 1,242.74$.

Another example of ratio scale is the absolute temperature on the Kelvin scale where there is a zero element, which represents the absence of any thermal motion. On this scale, it does make sense to say that a thing is twice as hot as another thing if, for example, the first one is 273 K (almost 0 °C, 32 °F) and the second one is 546 K (almost 273 °C, 523.4 °F).

2.2.5 Overall Considerations

Table 2 summarizes the discussion about measurement scales reporting the empirical relations they are based on, the permissible transformation for each of them as well as the appropriate statistics and statistical test for each. The scales of Table 2 are ordered by their increasing richness, i.e. by their capability of allowing more and more sophisticated operations; therefore, all the appropriate statistics reported for a lower richness scale can be applied to a higher richness one, even if they are not explicitly repeated in the table to save space.

Table 2
Summary of measurement scales, empirical relations they are based on, permissible transformations, and statistics relevant to each [4].

Scale Type	**Empirical Relations**	**Permissible Transformation**	**Appropriate Statistics**	**Appropriate Statistical Tests**
Nominal	Equivalence	bijective function f	Mode Frequency	Non-parametric
Ordinal	Equivalence Greater than	monotonic increasing function g	Median Percentile Spearman ρ Kendall τ Kendall W	Non-parameteric
Interval	Equivalence Greater than Ratio of any interval	$\mathrm{M}' = \alpha\mathrm{M} + \beta, \ \alpha > 0$	Mean Standard deviation Pearson product-moment correlation Multiple product-moment correlation	Non-parametric
Ratio	Equivalence Greater than Ratio of any interval Ratio of any values	$\mathrm{M}' = \alpha\mathrm{M}, \ \alpha > 0$	Geometric mean Harmonic mean Coefficient of variation	Non-parametric Parametric

It should be noted that the indications about which arithmetic operations, statistics and statistical tests are allowable for each scale type should not be taken in the most restrictive way possible. Stevens, who proposed the above distinctions for the first time, already noted [11]:

> *In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these [ordinal] scales, for these statistics imply a knowledge of something more than the relative rank-order of data. On the other hand, for this "illegal" statisticizing there can be invoked a kind of pragmatic sanction: in numerous instances it leads to fruitful results.*

This and similar lines of reasoning have led to a whole debate about whether measurement scales should or should not be used to prescribe and proscribe statistics and statistical tests [13]. It is well beyond the scope of this paper to get into the details of this discussion.

However, it should be noted that the notions of scale types and meaningfulness should be constantly kept present, at least, in order to always wonder and check whether the manipulation and analyses we are performing are licit as well as the conclusions we draw from them. On the other hand, careful thinking and consideration can indicate us when it is actually the case to go beyond the strict rules of measurement scales and meaningfulness for our experimentation, even if there should be clear motivations, which need to be explicitly reported and explained.

Coming back to the case of *Information Retrieval (IR)* and the claim of Robertson [14] that the assumption of *Average Precision (AP)* being an interval scale is somehow arbitrary, according to the above discussion, it means that we may find even disagreeing conclusions when, to compare systems, we perform operations which assume such kind of scale, such as the arithmetic mean, i.e. when we compute the *Mean Average Precision (MAP)*, or when we perform operations which assume even a ratio scale, such as the geometric mean, i.e. when we compute the *Geometric Mean Average Precision (GMAP)*. However, being the interval scale assumption violated in both cases, we should consider both conclusions equally valid and we should to resort to external considerations to understand which conclusions are actually meaningful for our needs.

3 Summary of Used Symbols

Table 3 summarizes the main symbols used through the paper. For each symbol it is reported its meaning and where it has been defined in the main paper.

4 Set-Based Measures

4.1 Partial Ordering

Proposition 3. Let $N \in \mathbb{N}$ be fixed and let $REL = \{\mathscr{a}_0, \dots, \mathscr{a}_c\}$ with $c > 1$. The poset $R(N)$ is graded, i.e. every maximal chain of $R(N)$ has the same length.

Proof: Let us prove that $R(N)$ is a lattice using Lemma 1. Firstly. let us study the *cover* relation, i.e. the operation which passes from a run in $R(N)$ to a new run that covers the first one.

Table 3
Summary of the main symbols use through the paper. For each symbol it is reported its meaning and where it has been defined in the main paper.

Symbol	Meaning	Definition
P	A poset	Section 2
$\preceq$	An ordering	Section 2
C	A subset of a poset P that is a chain	Section 2
$\ell(C)$	The length of a chain C	Section 2
ρ	The rank function of a poset P	Section 2
$\ell([s,t])$	The length of an interval, which is the natural distance in case of a graded poset	Section 2
M	A measure	Section 3.1 and 4
E	A set of empirical objectes, whic is the doamin of a measure	Section 3.1
Δ_{ab}	A difference over intervals	Section 3.2
$\preceq_d$	An ordering among intervals	Section 3.2
$\leq$	An ordering among real numbers	Section 3.3
REL	The (totally ordered) set of relevance degrees	Section 4
a_i	A relevance degree	Section 4
D	The set of documents	Section 4
d	A document	Section 4
T	The set of topics	Section 4
t	A topic	Section 4
GT	The ground-truth	Section 4
N	The length of an IR system run	Section 4
$D(N)$	The set of all the possible N retrieved documents	Section 4, 4.1, and 4.2
r	A run of an IR system	Section 4, 4.1, and 4.2
r_j	the j-th element of the a run in the set-based retrieval case, i.e. $r_j = d_j$	Section 4.1
$r[j]$	the j-th element of the a run in the rank-based retrieval case, i.e. $r[j] = d_j$	Section 4.2
$R(N)$	The set of N judged documents	Section 4, 4.1, and 4.2
$\hat{r}$	A judged run of an IR system	Section 4, 4.1, and 4.2
$\hat{r}_j$	the j-th element of the a judged run in the set-based retrieval case, i.e. $\hat{r}_j = a_j$	Section 4.1
$\hat{r}[j]$	the j-th element of the a judged run in the rank-based retrieval case, i.e. $\hat{r}[j] = a_j$	Section 4.2
g	The gain function	Section 4
δ_a	The indicator function for relevance degrees	Section 4
RB	The recall base	Section 4
m	the multiplicity of a relevance degree in the case of set-based retrieval	Section 4.1
p	The persistence parameter in *Rank-Biased Precision (RBP)*	Section 4.2.2
b	The logarithm base in *Discounted Cumulated Gain (DCG)*	Section 4.2.3
x_k	The probability that a user leaves their search after considering the document at position k in *Expected Reciprocal Rank (ERR)*	Section 4.2.4

The partial order defined in Section 5.2 of the main paper is:

$$\hat{r} \preceq \hat{s} \;\Leftrightarrow\; \big|\{i : \hat{r}_i \succeq a_j\}\big| \leq \big|\{i : \hat{s}_i \succeq a_j\}\big| \qquad \forall j \in \{0, \ldots, c\} \tag{1}$$

for $\hat{r}, \hat{s} \in R(N)$.

Recall that the replacement property consists in replacing an element of a judged run with one of higher relevance, i.e. if $\hat{r} = \{\hat{r}_1, \ldots, \hat{r}_{i-1}, \hat{r}_i, \hat{r}_{i+1} \ldots, \hat{r}_N\}$, we can replace $\hat{r}_i$ with a document with higher relevance degree, that is $\hat{\hat{r}} = \{\hat{r}_1, \ldots, \hat{r}_{i-1}, a_j, \hat{r}_{i+1} \ldots, \hat{r}_N\}$, with $\hat{r}_i \preceq a_j$. Note that:

- for every $0 \leq k \leq \delta_a(\hat{r}_i)$, $\big|\{n : \hat{r}_n \succeq a_k\}\big| = \big|\{n : \hat{\hat{r}}_n \succeq a_k\}\big|$, by construction of $\hat{r}$ and $\hat{\hat{r}}$;
- for $\delta_a(\hat{r}_i) < k \leq j$, $\big|\{n : \hat{r}_n \succeq a_j\}\big| \leq \big|\{n : \hat{\hat{r}}_n \succeq a_j\}\big|$, since $\hat{r}_i \preceq a_j$;
- for every $j < k \leq c$, $\big|\{n : \hat{r}_n \succeq a_k\}\big| = \big|\{n : \hat{\hat{r}}_n \succeq a_k\}\big|$, since $\hat{r}$ and $\hat{\hat{r}}$ have the same number of elements with relevance degree above a_j.

Therefore we proved that $\hat{r} \preceq \hat{\hat{r}}$.

We can iterate the replacement operation to pass from any run $\hat{r} \in R(N)$ to any another run $\hat{s}$ in the same set such that $\hat{r} \preceq \hat{s}$. Let us show that the "cover" relations are given by the replacement of an element of a run with relevance degree a_i with the relevance degree a_{i+1}; remember that since a_i and a_{i+1} are two successive relevance degrees, by definition, it holds $a_i \prec a_{i+1}$.

Given two runs $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \prec \hat{s}$, $\hat{s}$ can different from $\hat{r}$ by one or more replacements.

Let us suppose that only one replacement has been performed, that is an element of $\hat{r}$ – let's say $\hat{r}_k$ with relevance degree a_i – has been exchanged with another one with relevance degree of a_j such that $a_i \prec a_j$; in other terms, we have $\hat{r} = \{\hat{r}_1, \ldots, \hat{r}_{k-1}, a_i, \hat{r}_{k+1}, \ldots, \hat{r}_N\}$ and $\hat{s} = \{\hat{r}_1, \ldots, \hat{r}_{k-1}, a_j, \hat{r}_{k+1}, \ldots, \hat{r}_N\}$.

Recall that the rank function, defined in the main paper, is given by $\rho(\hat{t}) = \sum_{n=1}^{N} \delta_a(\hat{t}_n)$, for $\hat{t} \in R(N)$. Therefore $\rho(\hat{s}) - \rho(\hat{r}) = \sum_{n=1}^{N} \left(\delta_a(\hat{s}_n) - \delta_a(\hat{r}_n)\right) = \sum_{\substack{n=1 \\ n \neq k}}^{N} \left(\delta_a(\hat{r}_n) - \delta_a(\hat{r}_n)\right) + \left(\delta_a(a_j) - \delta_a(a_i)\right) = j - i$. Therefore the replacement is a "cover" relation if and only if $\rho(\hat{s}) - \rho(\hat{r}) = 1$, by definition of rank function, that is if $j = i + 1$.

Therefore, if $\hat{r}$ and $\hat{s}$ are such that $\hat{r} \prec \hat{s}$ and the two runs differ by only one replacement, than the two relevance degrees exchanged have to be consecutive in order for $\hat{s}$ to cover $\hat{r}$.

If $\hat{r}$ and $\hat{s}$ differ for more than one replacement, then we can always find a third run $\hat{v}$ such that $\hat{r} \prec \hat{v} \prec \hat{s}$. Let us suppose that they differ for exactly two replacements, one in $\hat{r}_{k_1}$ and one in $\hat{r}_{k_2}$; the case with more replacements follows analogously. Hence we can write the two runs as

$$\begin{aligned}\hat{s} &= \{\hat{r}_1, \ldots, \hat{r}_{k_1-1}, a_j, \hat{r}_{k_1+1}, \ldots, \hat{r}_{k_2-1}, a_m, \hat{r}_{k_2+1}, \ldots, \hat{r}_N\} ,\\ \hat{r} &= \{\hat{r}_1, \ldots, \hat{r}_{k_1-1}, a_i, \hat{r}_{k+1}, \ldots, \hat{r}_{k_2-1}, a_l, \hat{r}_{k_2+1}, \ldots \hat{r}_N\} ,\end{aligned}$$

with $a_i \prec a_j$ and $a_l \prec a_m$. By construction (and the above discussion) follows that $\hat{r} \prec \hat{s}$.

Therefore the run $\hat{v}$ such that

$$\hat{v} = \{\hat{r}_1, \ldots, \hat{r}_{k_1-1}, a_j, \hat{r}_{k+1}, \ldots, \hat{r}_{k_2-1}, a_l, \hat{r}_{k_2+1}, \ldots \hat{r}_N\} ,$$

is clearly such that $\hat{r} \prec \hat{v}$, since the two runs differs by only one replacement. Moreover $\hat{v} \prec \hat{s}$ since $\hat{v}$ is $\hat{s}$ where a replacement with a higher relevance degree has been performed.

Hence $\hat{s}$ covers $\hat{r}$ if and only if one replacement has been performed between two consecutive relevance degrees. In the following, we call this operation replacement of *weight* one, for simplicity.

Then, we show that $\hat{r}, \hat{s} \in REL^N$ such that both cover a run $\hat{u} \in R(N)$ are uniquely identified using $\hat{u}$.

Since $\hat{r}$ and $\hat{s}$ both cover $\hat{u}$, $\hat{r} \neq \hat{s}$ and $R(N)$ has a partial order, $\hat{r}$ and $\hat{s}$ are incomparable, and it does not exists a run $\hat{z} \in R(N)$ such that $\hat{u} \prec \hat{z} \prec \hat{r}$ nor $\hat{u} \prec \hat{z} \prec \hat{s}$. Let us set $\hat{u} := (\hat{u}[1], \ldots, \hat{u}[N])$. Since $\hat{r}$ and $\hat{s}$ covers $\hat{u}$ and replacement of weight one is the only "cover" relation, then $\hat{u}$ differs from both by a replacement of weight one made between different relevance degrees for $\hat{r}$ and $\hat{s}$.

Thus, if $\hat{u} = (\hat{u}[1], \ldots, \hat{u}[N])$, there exist two distinct indexes $k_1, k_2 \in \{1, \ldots, N\}$ such that $u[k_1] = a_i$ and $u[k_2] = a_j$ with $a_j \prec a_i \prec a_c$, that is

$$\hat{u} = \{\hat{u}_1, \ldots, \hat{u}_{k_1-1}, a_i, \hat{s}_{k_1+1}, \ldots, \hat{u}_{k_2-1}, a_j, \hat{u}_{k_2+1}, \ldots, \hat{u}_N\} ,$$

Then, the two runs $\hat{r}$ and $\hat{s}$ can only have the following expression:

$$\begin{aligned}\hat{s} &= \{\hat{u}_1, \ldots, \hat{u}_{k_1-1}, a_{i+1}, \hat{s}_{k_1+1}, \ldots, \hat{u}_{k_2-1}, a_j, \hat{u}_{k_2+1}, \ldots, \hat{u}_N\} ,\\ \hat{r} &= \{\hat{u}_1, \ldots, \hat{u}_{k_1-1}, a_i, \hat{s}_{k_1+1}, \ldots, \hat{u}_{k_2-1}, a_{j+1}, \hat{u}_{k_2+1}, \ldots, \hat{u}_N\} .\end{aligned}$$

Note that $\hat{r}$ and $\hat{s}$ are incomparable for the ordering (1) since $a_j \prec a_i$, indeed

- $\big|\{n \leq N : \hat{s}_n \succeq a_{i+1}\}\big| = \big|\{n \leq N : \hat{r}_n \succeq a_{i+1}\}\big| + 1 \geq \big|\{n \leq N : \hat{r}_n \succeq a_{i+1}\}\big|$;
- while $\big|\{n \leq N : \hat{s}_n \succeq a_{j+1}\}\big| = \big|\{n \leq N : \hat{r}_n \succeq a_{j+1}\}\big| - 1 \leq \big|\{n \leq N : \hat{r}_n \succeq a_{j+1}\}\big|$.

Now, given $\hat{r}$ and $\hat{s}$ as defined above such that both covers $\hat{u}$, we show that the joint $\hat{r} \vee \hat{s}$ exists.

Therefore let us define $\hat{t} \in R(N)$ as

$$\hat{t} = \{\hat{u}_1, \ldots, \hat{u}_{k_1-1}, a_{i+1}, \hat{s}_{k_1+1}, \ldots, \hat{u}_{k_2-1}, a_{j+1}, \hat{u}_{k_2+1}, \ldots, \hat{u}_N\} .$$

The above observations entail that $\hat{t}$ covers both $\hat{r}$ and $\hat{s}$, since $\hat{t}$ differs from both by just a replacement of weight one. The question is: is $\hat{t}$ the least upper bound, that is $\hat{t} = \hat{r} \vee \hat{s}$?

Let $\hat{v} \in R(N)$ be such that $\hat{r} \prec \hat{v}$ and $\hat{s} \prec \hat{v}$. Thanks to the ordering (1), we can note that $\hat{v}$ has to be such that

- for any $1 \leq k \leq i$, $\big|\{n \leq N : \hat{r}_n \succeq a_k\}\big| \leq \big|\{n \leq N : \hat{v}_n \succeq a_k\}\big|$;
- for any $i+1 \leq k \leq c$, $\big|\{n \leq N : \hat{s}_n \succeq a_k\}\big| \leq \big|\{n \leq N : \hat{v}_n \succeq a_k\}\big|$;

since $\hat{r}, \hat{s} \prec \hat{v}$. But $\hat{t}$ is built in such a way that

- for any $1 \leq k \leq i$, $\big|\{n \leq N : \hat{r}_n \succeq a_k\}\big| = \big|\{n \leq N : \hat{t}_n \succeq a_k\}\big|$;
- for any $i+1 \leq k \leq c$, $\big|\{n \leq N : \hat{s}_n \succeq a_k\}\big| = \big|\{n \leq N : \hat{t}_n \succeq a_k\}\big|$;

Therefore $\big|\{n \leq N : \hat{t}_n \succeq a_k\}\big| \leq \big|\{n \leq N : \hat{v}_n \succeq a_k\}\big|$ for any $k \in \{1, \ldots, c\}$, that is $\hat{v} \succeq \hat{t}$ and this holds for any judged run $\hat{v} \in R(N)$ such that $\hat{r}, \hat{s} \prec \hat{v}$. Hence $\hat{t} = \hat{r} \vee \hat{s}$, and $R(N)$ is a lattice since $\hat{r}, \hat{s}$ where chosen arbitrarily.

Moreover $\hat{t} = \hat{r} \vee \hat{s}$ covers, by construction as showed above, both $\hat{r}$ and $\hat{s}$. Then Proposition 2 let us state that $R(N)$ is graded, and the proof of Proposition 3 is complete. □

5 Rank-Based Measures

5.1 Total Ordering

Proposition 4. Let $REL = \{a_0, \ldots, a_c\}$ and let $G = \min_{i \in \{1,\ldots,c\}}(g(a_i) - g(a_{i-1}))/g(a_c)$. Given the ordering characterized by the strong top-heaviness, *Graded Rank-Biased Precision (gRBP)*$_p$ is ordinal scale on $R(N)$ if and only if $p \leq G/(G+1)$.

Proof:

Even though we work with N fixed, we want gRBP_p to be ordinal for some $p \geq 0$ regardless of the chosen value of N. Moreover, we assume $p \in (0,1)$ in order to remove trivial cases, since for $p \in \{0,1\}$, gRBP_p is constantly equal to 0. Recall that gRBP_p is an ordinal scale for a given value of p if and only if

$$\hat{r} \preceq \hat{s} \Leftrightarrow \mathrm{gRBP}_p(\hat{r}) \leq \mathrm{gRBP}_p(\hat{s}),$$

for any $\hat{r}, \hat{s} \in R(N)$.

Let us consider $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \prec \hat{s}$. Since the ordering is given by the strong top-heaviness, there exists $k \in \{1, \ldots, N\}$ such that

$$\hat{r} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathscr{a}_i, \hat{r}[k+1], \ldots, \hat{r}[N]) \text{ and } \hat{s} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathscr{a}_j, \hat{s}[k+1], \ldots, \hat{s}[N]) \ ,$$

with $i < j$ (that is $\mathscr{a}_i \prec \mathscr{a}_j$).

Let moreover $\hat{\bar{r}}, \hat{\bar{s}} \in R(N)$ be such that

$$\hat{\bar{r}} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathscr{a}_i, \mathscr{a}_c, \ldots, \mathscr{a}_c) \text{ and } \hat{\bar{s}} = (\hat{r}[1], \ldots, \hat{r}[k-1], \mathscr{a}_j, \mathscr{a}_0, \ldots, \mathscr{a}_0) \ .$$

Clearly $\hat{r} \preceq \hat{\bar{r}} \prec \hat{\bar{s}} \preceq \hat{s}$, let us prove that $\mathrm{gRBP}_p(\hat{\bar{r}}) < \mathrm{gRBP}_p(\hat{\bar{s}})$ iff $p \leq G/(G+1)$.

$$\begin{aligned}
\mathrm{gRBP}_p(\hat{\bar{r}}) - \mathrm{gRBP}_p(\hat{\bar{s}}) &= \frac{1-p}{g(\mathscr{a}_c)} \left(\sum_{n=1}^{k-1} g(\hat{r}[n]) p^{n-1} + g(\mathscr{a}_i) p^{k-1} + \sum_{n=k+1}^{N} g(\mathscr{a}_c) p^{n-1} \right) \\
&\quad - \frac{1-p}{g(\mathscr{a}_c)} \left(\sum_{n=1}^{k-1} g(\hat{r}[n]) p^{n-1} + g(\mathscr{a}_j) p^{k-1} + \sum_{n=k+1}^{N} g(\mathscr{a}_0) p^{n-1} \right) \\
&= \frac{1-p}{g(\mathscr{a}_c)} \left(\left(g(\mathscr{a}_i) - g(\mathscr{a}_j) \right) p^{k-1} + \sum_{n=k+1}^{N} g(\mathscr{a}_c) p^{n-1} \right) \ ,
\end{aligned}$$

where we used $g(\mathscr{a}_0) = 0$.

Note that $\mathrm{gRBP}_p(\hat{\bar{r}}) - \mathrm{gRBP}_p(\hat{\bar{s}}) < 0$ iff

$$\frac{g(\mathscr{a}_i) - g(\mathscr{a}_j)}{g(\mathscr{a}_c)} p^{k-1} + \sum_{n=k+1}^{N} p^{n-1} < 0 \ . \tag{2}$$

Since the proof has to be independent from any pair of chosen runs, as long as $\hat{r} \prec \hat{s}$, we can maximize the *Left Hand Side (LHS)* of the inequality above by choosing

$$\max_{i,j \in \{0,\ldots,c\}, i<j} \frac{g(\mathscr{a}_i) - g(\mathscr{a}_j)}{g(\mathscr{a}_c)} = - \min_{i,j \in \{0,\ldots,c\}, i<j} \frac{g(\mathscr{a}_j) - g(\mathscr{a}_i)}{g(\mathscr{a}_c)} = - \min_{i \in \{1,\ldots,c\}} \frac{g(\mathscr{a}_{i+1}) - g(\mathscr{a}_i)}{g(\mathscr{a}_c)} =: -G \ .$$

Therefore, (2) holds iff

$$G p^{k-1} - \sum_{n=k+1}^{N} p^{n-1} > 0 \ \Leftrightarrow \ G p^{k-1} - \frac{p^k - p^N}{1-p} > 0 \ \Leftrightarrow p^{N-k+1} - (1+G) p + G > 0 \ .$$

Clearly, if $p \leq G/(G+1)$

$$p^{N-k+1} - (1+G) p + G \geq p^{N-k+1} > 0 \ ,$$

for any N and $k \leq N$. On the other hand, if $p > G/(G+1)$, then there exists N big enough that

$$p^{N-k+1} - (1+G) p + G < 0 \ .$$

Indeed letting N go to infinity, the LHS of the above inequality becomes strictly smaller than 0 since $-(1+G)p + G < 0$ if $p < G/(G+1)$, and this means that there exists N big enough that $\mathrm{gRBP}_p(\hat{\bar{r}}) - \mathrm{gRBP}_p(\hat{\bar{s}}) > 0$. Therefore, we proved that $\mathrm{gRBP}_p(\hat{\bar{r}}) < \mathrm{gRBP}_p(\hat{\bar{s}}) \ \Leftrightarrow \ p \leq G/(G+1)$.

Moreover, $\mathrm{gRBP}_p(\hat{r}) \leq \mathrm{gRBP}_p(\hat{\bar{r}})$, since $g(\hat{r}[i]) \leq g(\hat{\bar{r}}[i])$ for all $i \in \{1, \ldots, N\}$, and analogously $\mathrm{gRBP}_p(\hat{\bar{s}}) \leq \mathrm{gRBP}_p(\hat{s})$. Thus, we can conclude that only when $p \leq G/(G+1)$, $\hat{r} \prec \hat{s} \Rightarrow \mathrm{gRBP}_p(\hat{r}) < \mathrm{gRBP}_p(\hat{s})$. Furthermore, since if $\hat{r} = \hat{s}$, then simply $\mathrm{gRBP}_p(\hat{r}) = \mathrm{gRBP}_p(\hat{s})$ and we proved that for $p \leq G/(G+1)$, $\hat{r} \preceq \hat{s} \Rightarrow \mathrm{gRBP}_p(\hat{r}) < \mathrm{gRBP}_p(\hat{s})$.

To show the other implication of the *iff*, that is $\mathrm{gRBP}_p(\hat{r}) \leq \mathrm{gRBP}_p(\hat{s}) \Rightarrow \hat{r} \preceq \hat{s}$, we just prove that $\mathrm{not}\{\hat{r} \preceq \hat{s}\} \Rightarrow \mathrm{not}\{\mathrm{gRBP}_p(\hat{r}) \leq \mathrm{gRBP}_p(\hat{s})\}$, i.e. we need to prove that $\hat{r} \succ \hat{s} \Rightarrow \mathrm{gRBP}_p(\hat{r}) > \mathrm{gRBP}_p(\hat{s})$. But this last relation is exactly what we have already proved above, exchanging $\hat{r}$ with $\hat{s}$. Hence, the proof is complete. □

Proposition 5. Let $REL = \{\mathscr{a}_0, \ldots, \mathscr{a}_c\}$ and $g(\mathscr{a}_j) \neq K \delta_{\mathscr{a}}(\mathscr{a}_j)$ for at least one $j \in \{1, \ldots, c\}$, with $K > 0$. Given the ordering characterized by the strong top-heaviness, $\mathrm{gRBP}_{(c+1)^{-1}}$ is not an interval scale on $R(N)$.

Proof: In order for $\text{gRBP}_{(c+1)^{-1}}$ to be on an interval scale we should prove that there exist a positive constant $\alpha \in \mathbb{R}$ and a constant $\beta \in \mathbb{R}$ such that for every $\hat{r} \in R(N)$, $\text{gRBP}_{(c+1)^{-1}}(\hat{r}) = \alpha M(\hat{r}) + \beta$, where $M(\hat{r}) = \sum_{i=1}^{N} \delta_a(\hat{r}[i])(c+1)^{N-i}$. Note that $\beta = 0$ since $\text{gRBP}_{(c+1)^{-1}}(\hat{0}) = 0 = M(\hat{0})$. Therefore, since

$$\begin{aligned}\text{gRBP}_{(c+1)^{-1}}(\hat{r}) &= \frac{c}{c+1}\frac{1}{g(a_c)}\sum_{i=1}^{N} g(\hat{r}[i])(c+1)^{-i+1} \\ &= \frac{c}{g(a_c)}\sum_{i=1}^{N} g(\hat{r}[i])(c+1)^{-i} \ ,\end{aligned}$$

and

$$M(\hat{r}) = \sum_{i=1}^{N} \delta_a(\hat{r}[i])(c+1)^{N-i} = (c+1)^N \sum_{i=1}^{N} \delta_a(\hat{r}[i])(c+1)^{-i} \ ,$$

we should prove that there exist $\alpha > 0$ such that

$$\frac{c}{g(a_c)}\sum_{i=1}^{N} g(\hat{r}[i])(c+1)^{-i} = \alpha * (c+1)^N \sum_{i=1}^{N} \delta_a(\hat{r}[i])(c+1)^{-i} \ ,$$

that is

$$\sum_{i=1}^{N} g(\hat{r}[i])(c+1)^{-i} = \widetilde{\alpha} * \sum_{i=1}^{N} \delta_a(\hat{r}[i])(c+1)^{-i} \ ,$$

where $\widetilde{\alpha} = \alpha * (c+1)^N g(a_c)/c$.

Let us consider for example $N = 5$, $c = 2$ and $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} = (a_1, a_0, a_0, a_0, a_0)$ and $\hat{s} = (a_2, a_0, a_0, a_0, a_0)$. Clearly $\sum_{i=1}^{N} g(\hat{r}[i])(c+1)^{-i} = g(a_1)/(c+1) = \widetilde{\alpha} * 1/(c+1)$ and it implies that $\widetilde{\alpha} = g(a_1)$. But $\sum_{i=1}^{N} g(\hat{s}[i])(c+1)^{-i} = g(a_2)/(c+1) = \widetilde{\alpha} * 2/(c+1) = (g(a_1)/(c+1)) * 2$. Therefore, $\text{gRBP}_{(c+1)^{-1}}$ is interval scale for $c = 2$ only if $g(a_2) = 2g(a_1)$ that is only if the gain function is the indicator function. For every $c \geq 1$ and for every N we can find an example that shows that $\text{gRBP}_{(c+1)^{-1}}$ is interval scale measure only when the $g(a_j) = Kj$.

□

5.2 Partial Ordering

Proposition 6. Let $N \in \mathbb{N}$ be fixed and let $REL = \{a_0, a_1\}$. The poset $R(N)$ is graded, i.e. every maximal chain of $R(N)$ has the same length.

Proof: Note that $R(N)$ is a bounded, since for every $\hat{r} \in R(N)$, $\hat{r} \succeq (a_0, \ldots, a_0)$ and $\hat{r} \preceq (a_1, \ldots, a_1)$.; moreover it is of finite rank since $|R(N)| = 2^N < \infty$.

Let us prove that $R(N)$ is a lattice using Lemma 1, firstly studying the "cover" relation, i.e. the operation which passes from a run in $R(N)$ to a new run that covers the first one.

Note that, since in the binary case REL has only two relevance degrees – the not relevant degree a_0 and the relevant one a_1, the partial ordering defined in Section 6.2 of the main paper can be expressed as:

$$\hat{r} \preceq \hat{s} \ \Leftrightarrow \big|\{i \leq k : \hat{r}[i] = a_1\}\big| \leq \big|\{i \leq k : \hat{s}[i] = a_1\}\big| \qquad \forall k \in \{1, \ldots, N\} \ . \tag{3}$$

Let $\hat{r}, \hat{s} \in R(N)$ be such that $\hat{r} \prec \hat{s}$, define $m = \big|\{k \leq N : |\{i \leq k : \hat{r}[i] = a_1\}| < |\{i \leq k : \hat{s}[i] = a_1\}|\}\big|$ and denote with $k_1 < \cdots < k_m$ the depths where the strict inequality (3) holds. Firstly, note that if $\hat{r} \prec \hat{s}$ then $m \geq 1$.

If $m = 1$ and $k_1 < N$, then $\hat{s}$ and $\hat{r}$ have to be such that;

- $\big|\{i \leq k : \hat{r}[i] = a_1\}\big| = \big|\{i \leq k : \hat{s}[i] = a_1\}\big|$ for any $k < k_1$, that is $\hat{r}[k] = \hat{s}[k]$ for any $k < k_1$;
- $\big|\{i \leq k_1 : \hat{r}[i] = a_1\}\big| < \big|\{i \leq k_1 : \hat{s}[i] = a_1\}\big|$, that is $\hat{r}[k_1] = a_0$ and $\hat{s}[k_1] = a_1$;
- $\big|\{i \leq k : \hat{r}[i] = a_1\}\big| = \big|\{i \leq k : \hat{s}[i] = a_1\}\big|$ for any $k_1 < k \leq N$, and this is possible only if $\hat{r}[k_1+1] = a_1$, $\hat{s}[k_1+1] = a_0$ and the two runs are equal from depth $k_1 + 2$ to N, since $m = 1$.

In particular, $\hat{r}$ and $\hat{s}$ have to differ by a **swap** of length one:

$$\hat{s} = (\ldots, \hat{s}[k_1 - 1], a_1, a_0, \hat{s}[k_1 + 2], \ldots) \text{ and } \hat{r} = (\ldots, \hat{s}[k_1 - 1], a_0, a_1, \hat{s}[k_1 + 2], \ldots) \ .$$

Similarly, if $m = 1$ and $k_1 = N$:

- $\big|\{i \leq k : \hat{r}[i] = a_1\}\big| = \big|\{i \leq k : \hat{s}[i] = a_1\}\big|$ for any $k < N$, that is $\hat{r}[k] = \hat{s}[k]$ for any $k < N$;
- $\big|\{i \leq N : \hat{r}[i] = a_1\}\big| < \big|\{i \leq N : \hat{s}[i] = a_1\}\big|$, that is $\hat{r}[N] = a_0$ and $\hat{s}[N] = a_1$.

Therefore, $\hat{s}$ and $\hat{r}$ differ by a **replacement** in the last position:

$$\hat{s} = (\hat{s}[1], \ldots, \hat{s}[k_1 - 1], a_1) \text{ and } \hat{r} = (\hat{s}[1], \ldots, \hat{s}[k_1 - 1], a_0) \ .$$

In both cases, namely the swap and the replacement, for every $\hat{u} \in R(N)$ such that $\hat{r} \preceq \hat{u} \preceq \hat{s}$, then $\hat{u} = \hat{r}$ or $\hat{u} = \hat{s}$, that is $\hat{s}$ covers $\hat{r}$. This follows immediately from the partial order recalled above and from the fact that $m = 1$. Indeed, in the case of a replacement in the last position, a run $\hat{u} \in R(N)$ such that $\hat{r} \prec \hat{u} \prec \hat{s}$ has to be such that

- $\big|\{i \leq k : \hat{r}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k : \hat{u}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k : \hat{s}[i] = \mathscr{a}_1\}\big|$ for any $k < N$, that is $\hat{r}[k] = \hat{u}[k] = \hat{s}[k]$ for any $k < N$;
- $\big|\{i \leq N : \hat{r}[i] = \mathscr{a}_1\}\big| < \big|\{i \leq N : \hat{u}[i] = \mathscr{a}_1\}\big| < \big|\{i \leq N : \hat{s}[i] = \mathscr{a}_1\}\big|$, and this is impossible since $\big|\{i \leq N : \hat{r}[i] = \mathscr{a}_1\}\big| = 1 + \big|\{i \leq N : \hat{s}[i] = \mathscr{a}_1\}\big|$.

Then $\hat{u} = \hat{r}$ or $\hat{u} = \hat{s}$, and the proof in the case of a swap of length one follows analogously.

If $m > 1$, there are two cases to study: $k_2 > k_1 + 1$ or $k_2 = k_1 + 1$.

In the first case, $k_2 > k_1 + 1$, we have the following situation:

- $\big|\{i \leq k : \hat{r}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k : \hat{s}[i] = \mathscr{a}_1\}\big|$ for any $k < k_1$, that is $\hat{r}[k] = \hat{s}[k]$ for any $k < k_1$;
- $\big|\{i \leq k_1 : \hat{r}[i] = \mathscr{a}_1\}\big| < \big|\{i \leq k_1 : \hat{s}[i] = \mathscr{a}_1\}\big|$, that is $\hat{r}[k_1] = \mathscr{a}_0$ and $\hat{s}[k_1] = \mathscr{a}_1$;
- $\big|\{i \leq k : \hat{r}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k : \hat{s}[i] = \mathscr{a}_1\}\big|$ for any $k_1 < k < k_2$, and this is possible only if $\hat{r}[k_1 + 1] = \mathscr{a}_1$, $\hat{s}[k_1 + 1] = \mathscr{a}_0$ and the two runs are equal from depth $k_1 + 2$ to $k_2 - 1$, since $m > 1$ and $k_2 > k_1 + 1$;
- $\big|\{i \leq k_2 : \hat{r}[i] = \mathscr{a}_1\}\big| < \big|\{i \leq k_2 : \hat{s}[i] = \mathscr{a}_1\}\big|$, that is $\hat{r}[k_2] = \mathscr{a}_0$ and $\hat{s}[k_2] = \mathscr{a}_1$;
- $\big|\{i \leq k : \hat{r}[i] = \mathscr{a}_1\}\big| \leq \big|\{i \leq k : \hat{s}[i] = \mathscr{a}_1\}\big|$, for any $k_2 < k \leq N$, since $\hat{r} \preceq \hat{s}$.

Therefore, $\hat{r}$ and $\hat{s}$ are such that:

$$\begin{aligned}\hat{s} &= (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_1, \mathscr{a}_0, \hat{s}[k_1 + 2], \ldots, \hat{s}[k_2 - 1], \mathscr{a}_1, \hat{s}[k_2 + 1], \ldots) \ ,\\ \hat{r} &= (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_0, \mathscr{a}_1, \hat{s}[k_1 + 2], \ldots, \hat{s}[k_2 - 1], \mathscr{a}_0, \hat{r}[k_2 + 1], \ldots) \ .\end{aligned}$$

Then, the following run

$$\hat{u} = (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_0, \mathscr{a}_1, \hat{s}[k_1 + 2], \ldots, \hat{s}[k_2 - 1], \mathscr{a}_1, \hat{s}[k_2 + 1], \ldots)$$

is such that $\hat{r} \prec \hat{u} \prec \hat{s}$. Indeed $\hat{u} \prec \hat{s}$ since

- $\hat{u}[k] = \hat{s}[k]$ for $k < k_1$, then $\big|\{i \leq k : \hat{u}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k : \hat{r}[i] = \mathscr{a}_1\}\big|$ for $k < k_1$;
- $\hat{u}[k_1] = \mathscr{a}_0$ and $\hat{s}[k_1] = \mathscr{a}_1$, then $\big|\{i \leq k_1 : \hat{u}[i] = \mathscr{a}_1\}\big| = 1 + \big|\{i \leq k_1 : \hat{s}[i] = \mathscr{a}_1\}\big|$;
- $\hat{u}[k_1 + 1] = \mathscr{a}_1$ and $\hat{s}[k_1 + 1] = \mathscr{a}_0$, then $\big|\{i \leq k_1 + 1 : \hat{u}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k_1 + 1 : \hat{s}[i] = \mathscr{a}_1\}\big|$;
- $\hat{u}[k] = \hat{s}[k]$ for $k_1 + 1 < k \leq N$, and hence $\big|\{i \leq k : \hat{u}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k : \hat{s}[i] = \mathscr{a}_1\}\big|$ for $k_1 + 1 < k \leq N$.

On the other hand, $\hat{r} \prec \hat{u}$ since

- $\hat{u}[k] = \hat{r}[k]$ for $k < k_2$, then $\big|\{i \leq k : \hat{r}[i] = \mathscr{a}_1\}\big| = \big|\{i \leq k : \hat{u}[i] = \mathscr{a}_1\}\big|$ for $k < k_2$;
- $\hat{u}[k_2] = \mathscr{a}_1$ while $\hat{r}[k_2] = \mathscr{a}_0$, then $\big|\{i \leq k_2 : \hat{r}[i] = \mathscr{a}_1\}\big| = 1 + \big|\{i \leq k_2 : \hat{u}[i] = \mathscr{a}_1\}\big|$;
- $\big|\{i \leq k : \hat{r}[i] = \mathscr{a}_1\}\big| = 1 + \big|\{i \leq k : \hat{u}[i] = \mathscr{a}_1\}\big|$ for any $k_2 < k \leq N$ since $\hat{u}[k] = \hat{r}[k]$ for such values of k.

Therefore, if $m > 1$ and $k_2 > k_1 + 1$, $\hat{s}$ cannot cover $\hat{r}$.

When $k_2 = k_1 + 1$, we can have two cases:

$$\begin{aligned}s &= (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_1, \mathscr{a}_0, \hat{s}[k_2 + 1], \ldots) \ ,\\ \hat{r} &= (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_0, \mathscr{a}_0, \hat{r}[k_2 + 1], \ldots) \ ,\end{aligned}$$

or

$$\begin{aligned}\hat{s} &= (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_1, \mathscr{a}_1, \hat{s}[k_2 + 1], \ldots) \ ,\\ \hat{r} &= (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_0, \mathscr{a}_0, \hat{r}[k_2 + 1], \ldots) \ .\end{aligned}$$

In both cases, $\hat{u}$ given by

$$\hat{u} = (\ldots, \hat{s}[k_1 - 1], \mathscr{a}_0, \mathscr{a}_1, \hat{s}[k_2 + 1], \ldots)$$

is such that $\hat{r} \prec \hat{u} \prec \hat{s}$, and it follows analogously at what was done for $k_2 > k_1 + 1$. Then, also when $k_2 = k_1 + 1$, $\hat{s}$ cannot cover $\hat{r}$.

Thus, we have shown that the "cover" relations are justs swaps of length one and replacements in the last position.

Then we show that $\hat{r}, \hat{s} \in R(N)$ such that both cover a run $\hat{u} \in R(N)$ are uniquely identified using $\hat{u}$.

Since $\hat{r}$ and $\hat{s}$ both cover $\hat{u}$, $\hat{r} \neq \hat{s}$ and $R(N)$ has a partial order, $\hat{r}$ and $\hat{s}$ are incomparable, and it does not exists a run $\hat{z} \in R(N)$ such that $\hat{u} \prec \hat{z} \prec \hat{r}$ nor $\hat{u} \prec \hat{z} \prec \hat{s}$. Let us set $\hat{u} \coloneqq (\hat{u}[1], \ldots, \hat{u}[N])$. Since $\hat{r}$ and $\hat{s}$ cover $\hat{u}$ and replacement at depth N and swaps of length one are the only two "cover" relations, then $\hat{u}$ differs from, e,g,, $\hat{r}$ by a swap and from $\hat{s}$ by a swap made at a different depth or by a replacement in the last position.

Thus, if $\hat{u} = (\hat{u}[1], \ldots, \hat{u}[N])$, there exists at least an index $i \in \{1, \ldots, N - 1\}$ such that $u[i] = \mathscr{a}_0$ and $u[i + 1] = \mathscr{a}_1$, since $\hat{u}$ differs from one of the two runs (or both) by a swap of length one (a "cover" relation), and this is possible only when a not relevant degree is followed by a relevant one.

According to these observations, there exists at least an index $i \in \{1, \dots, N-1\}$ such that $u[i] = \mathpzc{a}_0$ and $u[i+1] = \mathpzc{a}_1$, precisely

$$\hat{u} = (\hat{u}[1], \dots, \hat{u}[i-1], \mathpzc{a}_0, \mathpzc{a}_1, \hat{u}[i+2], \dots, \widehat{[N]}) \ ,$$

and we have two possibilities (up to symmetries) for $\hat{r}$ and $\hat{s}$:

i. if there exists an index $j \in \{1, \dots, N-1\} \setminus \{i-1, i, i+1\}$ such that $\hat{u}[j] = \mathpzc{a}_0$ and $\hat{u}[j+1] = \mathpzc{a}_1$ (we choose $j > i+1$, but the case $j < i-1$ is symmetric), then

$$\begin{aligned}\hat{s} &= (\hat{u}[1], \dots, \hat{u}[i-1], \mathpzc{a}_0, \mathpzc{a}_1, \hat{u}[i+2], \dots, \hat{u}[j-1], \mathpzc{a}_1, \mathpzc{a}_0, \hat{u}[j+2], \dots, \hat{u}[N]) \ , \\ \hat{r} &= (\hat{u}[1], \dots, \hat{u}[i-1], \mathpzc{a}_1, \mathpzc{a}_0, \hat{u}[i+2], \dots, \hat{u}[j-1], \mathpzc{a}_0, \mathpzc{a}_1, \hat{u}[j+2], \dots, \hat{u}[N]) \ ,\end{aligned}$$

since $\hat{s}$ differs from $\hat{u}$ by a swap of length one made from depth $j+1$ to depth j and $\hat{r}$ differs from $\hat{u}$ by a swap of length one made from depth $i+1$ to depth i;

ii. if $\hat{u}[N] = \mathpzc{a}_0$, then

$$\begin{aligned}\hat{s} &= (\hat{u}[1], \dots, \hat{u}[i-1], \mathpzc{a}_0, \mathpzc{a}_1, \hat{u}[i+2], \dots, \hat{u}[N-1], \mathpzc{a}_1) \ , \\ \hat{r} &= (\hat{u}[1], \dots, \hat{u}[i-1], \mathpzc{a}_1, \mathpzc{a}_0, \hat{u}[i+2], \dots, \hat{u}[N-1], \mathpzc{a}_0) \ ,\end{aligned}$$

since $\hat{s}$ differs from $\hat{u}$ by a replacement made at depth N and $\hat{r}$ differs from $\hat{u}$ by a swap of length one made from depth $i+1$ to depth i.

Now, given $\hat{r}$ and $\hat{s}$ as defined above such that both cover $\hat{u}$, we show that the joint $\hat{r} \vee \hat{s}$ exists.

Therefore, let us define $\hat{t} \in R(N)$ as

i.

$$\hat{t} = (\hat{u}[1], \dots, \hat{u}[i-1], \mathpzc{a}_1, \mathpzc{a}_0, \hat{u}[i+2], \dots, \hat{u}[j-1], \mathpzc{a}_1, \mathpzc{a}_0, \hat{u}[j+2], \dots, \hat{u}[N-1], \hat{u}[N]) \ ;$$

ii.

$$\hat{t} = (\hat{u}[1], \dots, \hat{u}[i-1], \mathpzc{a}_1, \mathpzc{a}_0, \hat{u}[i+2], \dots, \hat{u}[N-1], \mathpzc{a}_1) \ .$$

The above observations entail that $\hat{t}$ covers both $\hat{r}$ and $\hat{s}$, since $\hat{t}$ differs from $\hat{r}$ by a swap of length one made at depth i, and from $\hat{s}$ respectively by a swap of length one at depth j or a replacement in the last position. The question is: is $\hat{t}$ the least upper bound, that is $\hat{t} = \hat{r} \vee \hat{s}$?

Let $\hat{v} \in R(N)$ be such that $\hat{r} \prec \hat{v}$ and $\hat{s} \prec \hat{v}$. Thanks to the ordering (3), $\hat{v}$ has to be such that

- for any $1 \le k \le i$, $\big|\{i \le k : \hat{r}[i] = \mathpzc{a}_1\}\big| \le \big|\{i \le k : \hat{v}[i] = \mathpzc{a}_1\}\big|$;
- for any $i+1 \le k \le N$, $\big|\{i \le k : \hat{s}[i] = \mathpzc{a}_1\}\big| \le \big|\{i \le k : \hat{v}[i] = \mathpzc{a}_1\}\big|$;

since $\hat{r}, \hat{s} \prec \hat{v}$. But $\hat{t}$ is built in such a way that

- for any $1 \le k \le i$, $\big|\{i \le k : \hat{r}[i] = \mathpzc{a}_1\}\big| = \big|\{i \le k : \hat{t}[i] = \mathpzc{a}_1\}\big|$;
- for any $i+1 \le k \le N$, $\big|\{i \le k : \hat{s}[i] = \mathpzc{a}_1\}\big| = \big|\{i \le k : \hat{t}[i] = \mathpzc{a}_1\}\big|$.

Therefore, $\big|\{i \le k : \hat{t}[i] = \mathpzc{a}_1\}\big| \le \big|\{i \le k : \hat{v}[i] = \mathpzc{a}_1\}\big|$ for any $k \in \{1, \dots, N\}$, that is $\hat{v} \succeq \hat{t}$ and this holds for any judged run $\hat{v} \in R(N)$ such that $\hat{r}, \hat{s} \prec \hat{v}$. Hence, $\hat{t} = \hat{r} \vee \hat{s}$, and $R(N)$ is a lattice since $\hat{r}, \hat{s}$ where chosen arbitrarily.

Moreover $\hat{t} = \hat{r} \vee \hat{s}$ covers, by construction as showed above, both $\hat{r}$ and $\hat{s}$. Then Proposition 2 let us state that $R(N)$ is graded, and the proof of Proposition 6 is complete. □

References

[1] R. P. Stanley, *Enumerative Combinatorics – Volume 1*, 2nd ed., ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, UK, 2012, vol. 49.

[2] A. Björner, P. H. Edelman, and G. M. Ziegler, "Hyperplane arrangements with a lattice of regions," *Discrete & Computational Geometry*, vol. 5, no. 3, pp. 263–288, June 1990.

[3] S. Foldes, "On distances and metrics in discrete ordered sets," *arXiv.org, Combinatorics (math.CO)*, vol. arXiv:1307.0244, June 2013.

[4] N. E. Fenton and J. Bieman, *Software Metrics: A Rigorous & Practical Approach*, 3rd ed. Chapman and Hall/CRC, USA, 2014.

[5] L. Finkelstein, "Widely, Strongly and Weakly Defined Measurement," *Measurement*, vol. 34, no. 1, pp. 39–48, July 2003.

[6] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement. Additive and Polynomial Representations*. Academic Press, New York, USA, 1971, vol. 1.

[7] L. Mari, "Beyond the Representational Viewpoint: a New Formalization of Measurement," *Measurement*, vol. 27, no. 2, pp. 71–84, March 2000.

[8] J. C. Falmagne and L. Narens, "Scales and Meaningfulness of Quantitative Laws," *Synthese*, vol. 55, no. 3, pp. 287–325, June 1983.

[9] L. Narens, *Theories of Meaningfullness*. Lawrence Erlbaum Associates, Mahwah (NJ), USA, 2002.

[10] F. S. Roberts, "Applications of the Theory of Meaningfulness to Psychology," *Journal of Mathematical Psychology*, vol. 29, no. 3, pp. 311–332, September 1985.

[11] S. S. Stevens, "On the Theory of Scales of Measurement," *Science, New Series*, vol. 103, no. 2684, pp. 677–680, June 1946.

[12] European Commission, "Commission Regulation (EC) No 607/2009 of 14 July 2009 laying down certain detailed rules for the implementation of Council Regulation (EC) No 479/2008 as regards protected designations of origin and geographical indications, traditional terms, labelling and presentation of certain wine sector products," *Official Journal of the European Union, OJ L 193, 24.7.2009*, vol. 52, pp. 60–139, July 2009.

[13] P. F. Velleman and L. Wilkinson, "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading," *The American Statistician*, vol. 47, no. 1, pp. 65–72, February 1993.

[14] S. Robertson, "On GMAP: and Other Transformations," in *Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006)*, P. S. Yu, V. Tsotras, E. A. Fox, and C.-B. Liu, Eds. ACM Press, New York, USA, 2006, pp. 78–83.