On the Shapley value and its application to the Italian VQR research assessment exercise

Camil Demetrescu[a], Francesco Lupia[b], Angelo Mendicelli[b], Andrea Ribichini[c],
Francesco Scarcello[b], Marco Schaerf[d]

[a]*Department of Computer, Automation and Management Engineering "Antonio Ruberti"*
Sapienza University of Rome, Via Ariosto 25, 00185 Rome Italy
demetrescu@diag.uniroma1.it
[b]*Department of Computer Science, Modeling, Electronics and Systems Engineering*
University of Calabria, 87036, Rende, Italy
{lupia,a.mendicelli,scarcello}@dimes.unical.it
[c]*Department of Physics*
Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome Italy
ribichini@diag.uniroma1.it
[d]*Department of Computer, Automation and Management Engineering "Antonio Ruberti"*
Sapienza University of Rome, Via Ariosto 25, 00185 Rome Italy
and
Institute of Information Technologies and Telecommunications
North Caucasus Federal University, Stavropol, Russian Federation
marco.schaerf@uniroma1.it

Abstract

Research assessment exercises have now become common evaluation tools in a number of countries. These exercises have the goal of guiding merit-based public funds allocation, stimulating improvement of research productivity through competition and assessing the impact of adopted research support policies. One case in point is Italy's most recent research assessment effort, VQR 2011-2014 (Research Quality Evaluation), which, in addition to research institutions, also evaluated university departments, and individuals in some cases (i.e., recently hired research staff and members of PhD committees). However, the way an institution's score was divided, according to VQR rules, between its constituent departments or its staff members does not enjoy many desirable properties well known from coalitional game theory (e.g., budget balance, fairness, marginality). We propose, instead, an alternative score division rule that is based on the notion of Shapley value, a well known solution concept in coalitional game theory, which enjoys the desirable properties mentioned above. For a significant test case (namely, Sapienza University of Rome, the largest university in Italy), we present a detailed comparison of the scores obtained, for substructures and individuals, by applying the official VQR rules, with those resulting from Shapley value computations. We show that there are significant differences in the resulting scores, making room for improvements in the allocation rules used in research assessment exercises.

Keywords: Research assessment, Bibliometrics, Shapley value, Research productivity

1. Introduction

In recent years, national research evaluation exercises have been adopted by a growing number of countries, including Italy. Objectives of these exercises include guiding merit-based public funds allocation, stimulating improvement of research productivity through comparative analysis between research structures, identifying weaknesses and strengths of a country's research infrastructure, and assessing the impact of adopted support policies Abramo & D'Angelo (2015). Research assessments are conducted in a variety of ways, and methodologies in a given country may even vary from one iteration to the next, based on accumulated experience, theoretical advancements, available resources and policy aims Abramo et al. (2011). Italy's most recent exercise, namely VQR 2011-2014, was based on a hybrid approach (i.e., bibliometric indicators for hard sciences and peer review for social sciences and humanities) and it evaluated a relatively small selection of research products (two per researcher for universities, and three per researcher for other institutions). We refer the reader to Section 2.1 for more details.

While the focus of national assessment exercises remains primarily the evaluation of entire research structures (i.e., universities or other research institutions), it has been argued that the aims of these exercises could be compromised if internal redistribution of government resources within each institution does not follow a consistent logic Abramo & D'Angelo (2011). The original goal of VQR 2011-2014 was the evaluation of structures (e.g., universities) and (partially) substructures (e.g., departments), however, its results have then been used for other assessments that more directly involve individuals. As an example, the assessment of Ph.D. courses has included the VQR performance of the members of the Ph.D. board[1]. A more recent use of the VQR results has been in the selection of the so-called "Dipartimenti di Eccellenza" (Excellent Departments)[2]. All these uses of the VQR results require that the credit allocation for the publications is done in a fair way. However, the way research product scores are currently used for these purposes yields evaluations that do not satisfy properties that are, from a methodological perspective, highly desirable (e.g., budget balance, fairness, marginality; see Section 3 for a detailed discussion).

The Shapley value Shapley (1953) is a well known solution concept in the context of coalitional game theory, and evaluations based on this notion are known to enjoy the desirable properties outlined above. While it can be shown that research evaluation efforts, such as VQR 2011-2014, can be modelled as a coalitional game, exact Shapley value computation remains a difficult problem, requiring time that is exponential in the size of the input instance. This made its application to large universities, with thousands of researchers, impractical. However, recent results have shown the feasibility of computing very good approximations to the Shapley value even for such large input instances (see Section 4 for an overview of these results).

[1] For more details, see (only available in Italian) http://www.anvur.org/index.php?option=com_content&view=article&id=455&Itemid=481&lang=en

[2] For more details, see (only available in Italian) http://hubmiur.pubblica.istruzione.it/web/universita/programmazione/dipartimenti-di-eccellenza

1.1. Contributions

In this article, we consider the main division rules used for allocation games, where there are indivisible goods and monetary compensation is possible. For these division rules, we discuss the properties that are desirable when allocation games are employed for modeling a national research evaluation process, such as the VQR 2011-2014 evaluation by the Italian Ministry of Education, Universities, and Research (MIUR).

We focus in particular on Sapienza University of Rome, the largest university in Italy and one of the largest in Europe, covering almost all Italian research areas. For this significant test case, we present a detailed comparison of the official research product division rules used by MIUR with those resulting from Shapley value computations. The comparison includes individual researchers, research groups, and substructures (e.g., departments). Our findings highlight that there are significant differences in the resulting scores, making room for possible methodological improvements in the approach currently used in Italian research evaluation exercises.

1.2. Structure of the article

Section 2 provides details about the Italian research assessment exercises and introduces coalitional game theory, with an emphasis on the Shapley value. Section 3 models Italian research assessment exercises as an allocation game. It discusses division rules and the properties that they should enjoy. Moreover, it compares the various division rules according to their properties. Section 4 reports on the difficulties of exact Shapley value computation, and on recent advances (e.g., approximation algorithms, input simplification techniques) that make it possible to compute exact results or at least good approximations even for large instances. Section 5 presents, for our test case (namely, Sapienza University of Rome), a critical comparison of the scores assigned to individual researchers, research groups and substructures (e.g., university departments) according to the official VQR 2011-2014 rules, with those resulting from Shapley value computations. Section 6 discusses related work, putting our article into perspective. Finally, Section 7 presents some conclusions and outlines some future research directions.

2. Preliminaries

2.1. The Italian VQR

VQR (Research Quality Evaluation) 2011-2014, the most recent Italian research assessment effort was based on a hybrid evaluation approach. Evaluation of the so called *bibliometric areas* (i.e., hard sciences) relied primarily on bibliometric analysis, while *non-bibliometric areas* (i.e., social sciences and humanities) were subjected to a peer review process.

VQR2011-2014 required every Italian research structure R recognized by the Italian Ministry of Education, University, and Research to select a subset of its *research products* (which could be journal articles, conference papers, books, book chapters, critical reviews, commentaries, book translations, patents, prototypes, exhibitions, works of art, etc.) and submit them to an evaluation agency called ANVUR. While doing so, structure R was in competition with all other Italian research structures, as the outcome of the evaluation was expected to guide the allocation of the merit-based share of the core public funding to Italian research institutions (until the next evaluation is performed). Every structure was therefore interested in selecting and submitting its best research products.

The VQR program was articulated in two phases. During Phase 1, based on the authors' self-evaluations and on ANVUR guidelines, R selected and submitted to ANVUR (at most) the required number of research products for each one of its authors, in such a way that each product was formally associated with exactly one author. The number of products required for each author varied according to the type of research institution. The default value was 2 for universities and 3 for other research structures, with exceptions in specific cases (e.g., recently hired personnel). During Phase 2 ANVUR formulated its independent quality judgment about the submitted research products (the score assigned to each product is currently made known only to its authors). The sum of the scores resulting from ANVUR's evaluation was then taken as the VQR score of R.

ANVUR published an evaluation of all departments, based on the product scores (the score of each department was computed as the sum of the scores of the products formally assigned to the authors in that department). The scores were also used for evaluating new individual researchers hired by R (this also greatly influenced R's funds in subsequent years), as well as members of PhD committees. Scores for recently hired researchers were computed as the sum of the scores of the products formally assigned to them. Data in this respect were published by ANVUR in aggregated form only for each department and for each scientific field. Evaluations for researchers that are members of PhD committees have been computed as the sum of the scores of the best publications each one of them had coauthored, among all the publications submitted for the VQR (for this evaluation, the formal assignment of publications to authors is irrelevant). Data in this respect were published by ANVUR in aggregated form only for each PhD committee and was only accessible to PhD coordinators and university officials. Another use of the VQR results, as already mentioned, was the creation of a ranking of the departments with the goal of identifying "excellent" ones and provide them with a significant extra funding. This ranking used the VQR scores of the members of the department normalized with respect to the average grade of all products in its scientific field (in Italy each professor is assigned to a unique scientific field, called SSD).

VQR 2011-2014 followed in the wake of two previous national research evaluation exercises. The first of these was VTR (Triennial Research Evaluation) 2001-2003. This first research evaluation effort was based solely on a peer review process, and it required each institution to submit a number of publications equal to 25% of its research staff complement (a relatively small sample), in each of the 18 disciplinary areas considered.

VQR 2004-2010 and the more recent VQR 2011-2014 closely resembled one another. The main differences were in the default number of products per researcher that had to be submitted (3 for universities and 6 for other research institutions in VQR 2004-2010, which extended over a longer time interval than VQR 2011-2014) and in the methodology for computing product scores based on bibliometric indicators. Both VQRs used a combination of citation counts and journal impact factors, as they were derived from international databases such as Scopus and WoS, in order to rate articles in bibliometric areas, resorting to informed peer review only in cases of significant discrepancies between the two measures (e.g., highly cited articles in poorly-ranked journals, or vice versa). However, the way these indicators were combined differed between the various versions of VQRs. We wish to stress that these details are immaterial for our article, as we focus not on computing product scores, but on how such scores are distributed among the respective authors (and their subsets).

2.2. Coalitional games

Coalitional games provide a rich mathematical framework to analyze interactions between intelligent agents (see, e.g., Osborne & Rubinstein (1994)). We consider coalitional games of the form $\mathsf{G} = \langle N, v\rangle$, consisting of a set N of n agents and a characteristic function v. The latter maps each coalition $C \subseteq N$ to the worth that agents in C can obtain by collaborating with each other. In this context, the crucial problem is to find a mechanism to allocate the worth $v(N)$, i.e., the value of the grand coalition N, in a way that is fair for all players, and that satisfies some additional important properties such as efficiency (i.e., distributing precisely the available budget $v(N)$ to the players, not more and not less). Moreover, for stability reasons, it is usually required that every group of agents C gets at least the worth $v(C)$ that it can guarantee to the game.

Several solution concepts have been considered in the literature as “fair allocation” schemes and, among them, a prominent one is the *Shapley value* Shapley (1953). According to this notion, the worth of any agent i is determined by considering its actual contribution to all possible coalitions of agents. More precisely, we consider the so-called *marginal contribution* to any coalition C, that is, the difference between what can be obtained when i collaborates with the agents in C and what can be obtained without the contribution of i. More formally, the Shapley value of a player $i \in N$ is defined by the following weighted average of all such marginal contributions:

$$\phi_i(\mathsf{G}) = \sum_{C \subseteq N\setminus\{i\}} \frac{|C|!(n-|C|-1)!}{n!}\Big(v(C\cup\{i\}) - v(C)\Big).$$

Allocation games. Among the various classes of coalitional games, we focus in this article on *allocation games*, which is a setting for analyzing fair division problems where monetary compensations are allowed and utilities are quasi-linear Moulin (1992). Allocation games naturally arise in various application domains, ranging from house allocation to room assignment and rent division, to (cooperative) scheduling and task allocation, to protocols for wireless communication networks, and to queuing problems (see, e.g., Greco & Scarcello (2014); Maniquet (2003); Mishra & Rangarajan (2007); Moulin (1992); Iera et al. (2011) and the references therein).

Computing the Shapley value of such games is a difficult problem, indeed it is #P-hard even if goods can only have two different possible values Greco et al. (2015). In this article we focus on large instances of this problem, involving thousands of agents and goods, for which no algorithm described in the literature is able to provide an exact solution. Fortunately, several approximation algorithms and input instance simplification techniques have been developed in recent years, that allow us to attack even such large instances. We refer to Section 4 for a detailed discussion of these results.

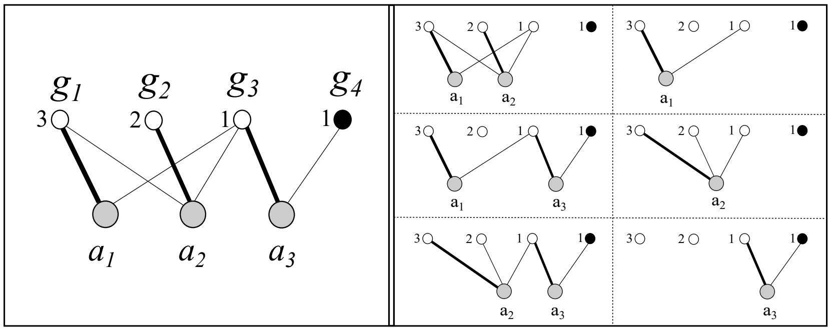

Figure 1: Allocation scenario $\mathcal{A}_0$ in Example 1.

In the setting considered in this article, an *allocation game* is defined by a tuple $\langle N, \mathbb{G}, \Omega, \texttt{val}, k\rangle$ comprising a set of agents N and a set of goods $\mathbb{G}$, whose values are given by the function `val` mapping each good to a non-negative real number.[3] The function Ω associates each agent with the set of goods he/she is interested in. Moreover, the natural number k provides the maximum number of goods that can be assigned to each agent. Each good is indivisible and can be assigned at most to one player.

An allocation scenario over this game is (determined by) a selected set of goods $\mathcal{A} \subseteq \mathbb{G}$ that have to be allocated to the agents.

For a coalition of agents $C \subseteq N$, a (feasible) allocation $\pi^C_{\mathcal{A}}$ is a mapping from C to sets of goods from $\mathcal{A}$ such that: each agent $i \in C$ gets a set of goods $\pi^C_{\mathcal{A}}(i) \subseteq \Omega(i)$ with $|\pi^C_{\mathcal{A}}(i)| \leq k$, and $\pi^C_{\mathcal{A}}(i) \cap \pi^C_{\mathcal{A}}(j) = \emptyset$ for any other agent $j \in C$ (each good can be assigned to one agent at most).

With a slight abuse of notation, we denote by $\texttt{val}(S)$ the sum of all the values of a set of goods $S \subseteq \mathbb{G}$, and by $\texttt{val}(\pi^C_{\mathcal{A}})$ the value of the goods allocated to C, that is, the goods in $\bigcup_{i\in C} \pi^C_{\mathcal{A}}(i)$. An allocation $\pi^C_{\mathcal{A}}$ is optimal if there exists no allocation $\bar{\pi}^C_{\mathcal{A}}$ with $\texttt{val}(\bar{\pi}^C_{\mathcal{A}}) > \texttt{val}(\pi^C_{\mathcal{A}})$. The total value of such an optimal allocation for the coalition C is denoted by $\texttt{opt}_{\mathcal{A}}(C)$. The budget available for $\mathcal{A}$, also called the (maximum) social welfare, is $\texttt{opt}_{\mathcal{A}}(N)$, that is, the value of any optimal allocation for the whole set of agents N (the grand coalition).

The *coalitional game* defined by the scenario $\mathcal{A}$ is the pair $\langle N, \texttt{opt}_{\mathcal{A}}\rangle$, that is, the game where the worth of any coalition is given by the value of any of its optimal allocations (w.r.t. the scenario $\mathcal{A}$). By using solution concepts from coalitional game theory, it is possible to distribute the available worth to agents an a way that is budget-balanced and that is perceived as a fair one by agents. Note that $\texttt{opt}_{\mathcal{A}}(C) \geq 0$ holds for each $C \subseteq N$, since the allocation where no agent receives any good is a feasible one (the value of an empty set of goods is 0). The definition trivializes for $C = \emptyset$, with $\texttt{opt}_{\mathcal{A}}(\emptyset) = 0$.

Example 1. Consider the allocation game $\langle\{a_1, a_2, a_3\}, \{g_1, g_2, g_3, g_4\}, \Omega, \texttt{val}, 1\rangle$ and its scenario with goods $\mathcal{A}_0 = \{g_1, g_2, g_3\}$, depicted in a graphical way in Figure 1,

[3] For the purpose of this article the setting is slightly simplified with respect to Greco et al. (2015).

where each edge connects an agent to a good he/she is interested in, and it is possible to allocate just one good to each agent ($k = 1$). Note that the good g_4, shown in black in the figure, is not used in the scenario $\mathcal{A}_0$ and thus cannot be assigned to any agent in allocations based on this selection of goods. The figure shows on the left an allocation for all the agents (with respect to $\mathcal{A}_0$), with the edges in bold identifying the allocation of goods to agents. Note that this is an optimal allocation, i.e., a feasible allocation whose sum of values of the allocated goods is the maximum possible one. The value of this allocation is $\mathtt{val}(g_1) + \mathtt{val}(g_2) + \mathtt{val}(g_3) = 3 + 2 + 1 = 6$.

The coalitional game associated with this scenario is $\mathsf{G}_{\mathcal{A}_0} = \langle\{a_1, a_2, a_3\}, v_{\mathcal{A}_0}\rangle$, where the worth function $v_{\mathcal{A}_0}$ is precisely $\mathtt{opt}_{\mathcal{A}_0}$. In particular, we have seen that, for the grand coalition, $v_{\mathcal{A}_0}(\{a_1, a_2, a_3\}) = 6$ holds. For each $C \subset \{a_1, a_2, a_3\}$ with $C \neq \emptyset$, an optimal allocation restricted to the agents in C is also reported in Figure 1. It follows that the other values of the worth function are $v_{\mathcal{A}_0}(\{a_1, a_2\}) = 5$, $v_{\mathcal{A}_0}(\{a_1, a_3\})$ = $v_{\mathcal{A}_0}(\{a_2, a_3\}) = 4$, $v_{\mathcal{A}_0}(\{a_1\}) = v_{\mathcal{A}_0}(\{a_2\}) = 3$, and $v_{\mathcal{A}_0}(\{a_3\}) = 1$. $\lhd$

3. The Italian VQR modelled as an allocation game

The VQR research assessment exercise for a certain university R can be naturally modeled as an allocation game $\langle\mathcal{R}, \mathbb{G}, \mathit{products}, \mathtt{val}, 2\rangle$ where $\mathcal{R}$ is the set of researchers affiliated with R, $\mathbb{G}$ is the set of all publications produced by the researchers in $\mathcal{R}$, *products* maps authors to the set of publications they have written, and $\mathtt{val}$ assigns a value to each publication. In the VQR 2011-2014 program, the range of $\mathtt{val}$ was $\{0, 0.1, 0.4, 0.7, 1\}$, with the latter value reserved to *excellent* products.

3.1. Selecting an optimal allocation scenario

In the submission phase, publication values are estimated by R according to the authors' self-evaluations, and to the reference tables published by ANVUR (only available for some research areas). Then, R selects for the assessment exercise a set of publications $\mathcal{P} \subseteq \mathbb{G}$ such that $|\mathcal{P}| \leq 2|\mathcal{R}|$ and there exists a feasible allocation $\pi_{\mathcal{P}}^{\mathcal{R}}$ to allocate these products to its researchers. To maximize its outcome, a structure R may solve a weighted matching problem and select the $\mathcal{P}$ having the maximum possible total value among all those authored by researchers in $\mathcal{R}$ in the considered time frame. At the end of the program, R receives an amount of funds proportional to $V_R = \mathtt{val}(\mathcal{P})$, that is, to the considered measure of the quality of these products.

Example 2. Consider the weighted bipartite graph in Figure 2, whose vertices are the researchers $\mathcal{R} = \{r_1, r_2, r_3\}$ of a university R and all the publications they have written. Edges encode the authorship relation *products*, and weights encode the mapping $\mathtt{val}$ providing the values of the publications. Consider the optimal allocation ψ such that $\psi(r_1) = \{p_1, p_3\}$, $\psi(r_2) = \{p_2, p_4\}$, and $\psi(r_3) = \{p_6, p_7\}$, encoded by the solid lines in the figure. Based on this allocation, an optimal allocation scenario for the evaluation is obtained by selecting $\mathcal{P}_\psi = \{p_1, p_2, p_3, p_4, p_6, p_7\}$. The publications that do not belong to the scenario, i.e., that are not submitted to ANVUR, are shown in black in the figure.

Note that p_2 is co-authored by r_1, r_2, and r_3, while p_3 is co-authored by r_1 and r_2. The total value of the grand coalition is $\mathtt{opt}(\mathcal{R}) = 45$. $\lhd$

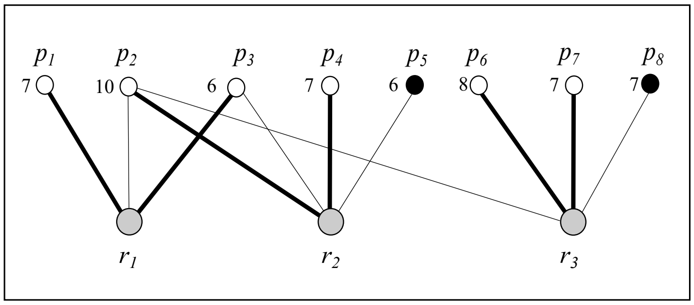

Figure 2: Authors and products in Example 2.

3.2. Division rules

The main issue in allocation problems is to compute, for a given allocation scenario, a fair distribution of the worth to agents. In particular, for the VQR case, we would like to compute a fair score for individual researchers, or groups, or departments, and so on.

Let $\mathcal{P}$ be an allocation scenario (the selected products), which is the image of an (optimal) allocation ψ to all agents in $\mathcal{R}$. In the VQR application, ψ is the formal association between researchers and products used for the submission to ANVUR.

Definition 1. A *division rule* γ_ψ (with respect to the allocation ψ) is a real-valued function that, given an agent $r \in N$, returns its score $\gamma_\psi(r) \geq 0$ with respect to ψ.

By slightly abusing notation, for any coalition $\mathcal{S} \subseteq \mathcal{R}$ we denote by $\gamma_\psi(\mathcal{S})$ the value $\sum_{r\in\mathcal{S}} \gamma_\psi(r)$. □

Some examples of division rules. The naive division rule, called `proj`, assigns to any agent r the sum of values of the products allocated to her according to ψ, that is,

$$\texttt{proj}_\psi(r) = \sum_{p\in\psi(r)} \texttt{val}(p).$$

This is the basic division rule used in the VQR to evaluate newly hired personnel and the substructures. For instance, the VQR evaluation of a department with a set of researchers $\mathcal{S}$ is precisely $\texttt{proj}_\psi(\mathcal{S})$.

The division rule that was used for evaluating PhD committees considers the overall VQR score that any subset $\mathcal{S}$ of agents would achieve if the research structure was constituted by them only (i.e., without caring about their co-authors outside $\mathcal{S}$ and considering all the products in $\mathcal{P}$):

$$\texttt{best}_\psi(\mathcal{S}) = \texttt{opt}_\mathcal{P}(\mathcal{S}).$$

Note that in the extreme case where $\mathcal{S} = \mathcal{R}$ we just obtain the overall VQR score of the research structure R.

Another possible division rule described in the literature (Karpov, 2014) and used in some university, called `owner`, is based on equal weights co-authorship sharing. This division rule assigns to each author the sum of the "normalized" scores of the submitted products she has co-authored, where by normalization we just mean here dividing the score of any product p by the cardinality of the set of its authors in $\mathcal{R}$, denoted by $authors(p)$:

$$\texttt{owner}_{\psi}(r) = \sum_{p \in products(r) \cap \mathcal{P}} \frac{\texttt{val}(p)}{|authors(p)|}.$$

Finally, we can use as division rule the Shapley value of the coalitional game $\langle \mathcal{R}, \texttt{opt}_{\mathcal{P}} \rangle$, defined by the scenario where the products in $\mathcal{P}$ are selected. Let $n = |\mathcal{R}|$ and define the marginal contribution of a set of agents S to a set of agents C as the difference $marg(S, C) = \texttt{opt}_{\mathcal{P}}(C) - \texttt{opt}_{\mathcal{P}}(C \setminus S)$. Then, this division rule can be written as follows:

$$\texttt{Shapley}_{\psi}(r) = \sum_{C \subseteq \mathcal{R} \setminus \{i\}} \frac{|C|!(n - |C| - 1)!}{n!} marg(\{i\}, C).$$

Desirable properties of division rules. We next recall the main desirable properties of any division rule γ for allocation problems. We refer the interested reader to Greco & Scarcello (2013) for a more detailed description of these properties.

(P1) Budget-balance. *The division rule precisely distributes the available worth over all agents, i.e.,* $\sum_{r \in \mathcal{R}} \gamma_{\psi}(r) = \texttt{val}(\mathcal{P})$.

(P2) Marginality. *For any set of researchers* $S \subseteq \mathcal{R}$, $\sum_{r \in S} \gamma_{\psi}(r) \geq marg(S, \mathcal{R})$. *That is, every group is granted at least its marginal contribution to the performance of the grand coalition* $\mathcal{R}$.

(P3) Independence of the specific allocation. *The division rule must be independent of the specific allocation used for the submission, that is, for any pair of optimal allocations* ψ *and* $\bar{\psi}$ *over the same allocation scenario* $\mathcal{P}$, *for each* $r \in \mathcal{R}$, $\gamma_{\psi}(r) = \gamma_{\bar{\psi}}(r)$. *This implies that, for each set of agents* S, $\gamma_{\psi}(S) = \gamma_{\bar{\psi}}(S)$ *also holds.*

(P4) Independence of the selected allocation scenario. *The division rule is indifferent w.r.t. the selected allocation scenario, as long as the selection is any optimal one: Consider two sets of products* $\mathcal{P}$ *and* $\bar{\mathcal{P}}$ *such that* $\texttt{val}(\mathcal{P}) = \texttt{val}(\bar{\mathcal{P}}) = \texttt{opt}_{\mathbb{G}}(\mathcal{R})$. *Then, for every optimal allocation* ψ *over the scenario* $\mathcal{P}$ *and every optimal allocation* $\bar{\psi}$ *over the scenario* $\bar{\mathcal{P}}$, *for each* $r \in \mathcal{R}$, $\gamma_{\psi}(r) = \gamma_{\bar{\psi}}(r)$. *Note that (P4) clearly entails (P3). For the VQR application, this means that the worth of any agent does not depend on the specific (optimal) set of research products submitted to ANVUR.*

Greco & Scarcello (2013) describe further properties to guarantee a correct self-evaluation in the preliminary phase of products selection, such as the "truthfulness" property: *A division rule must provide no incentive in misreporting the score of the research products*.

On the desirability of the Shapley value. We next discuss the properties of the various division rules mentioned above, in particular with regard to the application to research quality assessments. Table 1 summarizes the properties enjoyed by the division rules. All of them satisfy the minimal requirement that guarantees to every agent at least her marginal contribution to the grand coalition (property P2). However, the division rules behave quite differently with respect to the other properties.

	`proj`	`best`	`owner`	`Shapley`
(P1) budget-balance	■	□	■	■
(P2) marginality	■	■	□	■
(P3) independence of the allocation	□	■	■	■
(P4) independence of the allocation scenario	□	■	□	■

Table 1: Summary of desiderable properties of division rules.

Note that the basic `proj` division rule is trivially budget-balanced (Property P1). However it fails to satisfy property P3 (and hence P4), because it is completely based on the specific allocation used for submitting the products to the ANVUR agency. This means that this rule is definitely not adequate for evaluating single researchers, as well as groups of researchers and substructures. Indeed, the optimal allocations are chosen to maximize (only) the social welfare for the structure R, regardless of what happens to groups and researchers. Consider for instance Example 2, with the allocation scenario and its allocation ψ depicted in Figure 2, where $\texttt{proj}_{\psi}(r_1) = 13$ and $\texttt{proj}_{\psi}(r_2) = 17$. Then, consider an alternative allocation ψ' over the same scenario where agents r_1 and r_2 exchange their goods p_2 and p_3. In this case, nothing changes for the university, but the worth of these researchers is quite different, since $\texttt{proj}_{\psi'}(r_1) = 17$ and $\texttt{proj}_{\psi'}(r_2) = 13$.

Contrasted with the previous notion, `best` does not depend on the specific allocation (a new optimal allocation is computed for each set of agents to be considered), so that it satisfies P3 and, from the results in Greco & Scarcello (2014), it can be shown that it satisfies P4, too. For this reason, we simplify the notation and do not report the allocation (i.e., we just write `best` rather than $\texttt{best}_{\psi}$). We do the same for the subsequent division rules, which are independent of the allocation, too.

Note however that the division rule `best` is not budget-balanced, because it would distribute more resources than those that are actually available. In the same Example 2, we have $\texttt{best}(r_1) = 17$, $\texttt{best}(r_2) = 17$, and $\texttt{best}(r_3) = 18$, which means that we would need a total worth 52, while only $45 = \texttt{val}(\mathcal{P})$ is available. Besides the fact that it is not budget-balanced, if used for single researchers or small groups, this division rule clearly advantages those researchers whose research products have many co-authors, because each of them can independently use her best products. As a matter of fact, using this notion means getting rid of the main constraint of the VQR procedure, which states that the same product cannot be used simultaneously by more authors.

The division rule `owner` is instead budget-balanced and satisfies property P3, because for each agent all its products available in the allocation scenario are considered, regardless of the specific allocation used in the submission phase to ANVUR. However even this division rule is perceived as an unfair one, because it does not satisfy marginality and because it is sensitive to the different possible selection of products, that is, it strongly depends on the specific allocation scenario that has been chosen.

As far as marginality is concerned, consider the coalition $\{r_1, r_2\}$ and note that $marg(\{r_1, r_2\}, \{r_1, r_2, r_3\}) = 45 - 18 = 27$. However, $\texttt{owner}(r_1) = \texttt{owner}(r_2) = 7 + 10/3 + 6/2$, so that the overall worth of this coalition is $80/3$, which is $1/3$ less than its marginal contribution to the grand-coalition. Therefore, `owner` does not satisfy

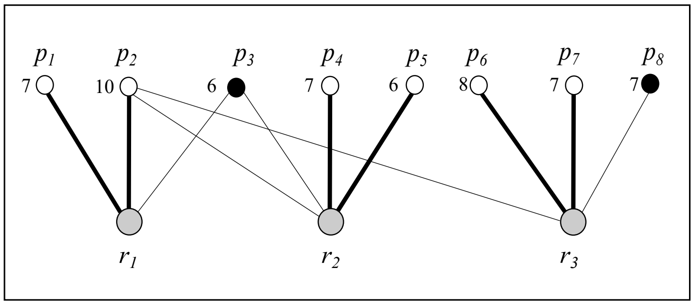

Figure 3: An alternative allocation scenario for authors and products in Example 2.

the basic requirement P2.

Concering Property P4, consider the alternative allocation scenario based on the products $\mathcal{P}' = \{p_1, p_2, p_4, p_5, p_6, p_7\}$, depicted in Figure 3. The figure also shows a possible optimal allocation ψ'' that gives the same (maximum) social welfare 45 for the grand-coalition. Now focus on agent r_2: by using the former allocation scenario $\mathcal{P}$, we get $\texttt{owner}(r_2) = 10/3 + 6/2 + 7 = 40/3$; by using the alternative allocation scenario $\mathcal{P}'$, r_2 increases her worth of the quantity 3, since we get $\texttt{owner}(r_2) = 10/3 + 7 + 6 = 49/3$. Indeed, in the latter case the scenario comprises the product p_5 that is authored only by r_2, instead of p_3 that r_1 and r_2 have in common. On the other hand, with the allocation scenario $\mathcal{P}'$, r_1 loses a quantity $6/2$, which came from her share in the product $p_3 \notin \mathcal{P}'$.

Therefore, `owner` fails to satisfy property P4, which we sometimes call the *fairness* property, because it is the very property that legitimates university R to freely look for any optimal selection of the research products, without caring about the personal and usually contrasting preferences of groups and researchers. Indeed, we have seen that, without property P4, the choice of a specific optimal set of products may lead to quite different scores. This is true not only for individuals but even for their aggregations: think, e.g., of researchers r_1 and r_2 belonging to different departments, and assume these departments compete with each other (or even with departments from other universities) for some funds depending on the VQR outcome.

We remark that other elementary division rules that do not depend on the specific allocation have the same problem, too. For instance, consider the alternative division rule `average`, which is based on the average value of all products authored by an agent in the considered scenario. It can be checked easily that this division rule does not meet property P4, either. For instance, contrasting the scenario $\mathcal{P}'$ with the scenario $\mathcal{P}$, the average value for r_1 is modified by the selection in the latter scenario of the product p_3, which reduces its average value from $17/2$ to $23/3$.

Moreover, observe that it is unfeasible to consider division rules that possibly take into account products outside the allocation scenario. Indeed, products outside the scenario are not part of any allocation and cannot be used. Think of the VQR application: these products have not been submitted to ANVUR and hence none of them has actually been evaluated by the committee. Therefore, none of them has a final, validated and official VQR value.

On the other hand, it has been shown that the division rule `Shapley` satisfies all the considered properties (and even more) Greco & Scarcello (2013): it is well known that the Shapley value is a budget-balanced notion and it is clearly independent of the specific allocation; moreover, in Greco & Scarcello (2014) it is proved that it is independent of the selection of the products in the scenario, and that satisfies marginality (because it is in the core of the "dual" coalitional game $\langle \mathcal{R}, marg(\cdot, \mathcal{R}) \rangle$). Unlike the above proposals, the worth of any agent is computed by considering the marginal contributions of all the possible coalitions she may join, and then by taking the weighted average of these contributions (where the weight depends on the cardinality of each coalition, precisely on the number of permutations of agents in C and in its complement). In the subsequent sections, we face the problem of computing this allocation, and discuss the differences with the division rules used for the VQR by ANVUR, by considering the actual values in our case study at Sapienza University of Rome.

Because the notion is both independent of the scenario and of the allocation, we next denote $\texttt{Shapley}_{\psi}(r)$ simply by ϕ_r, as it is usual in game theory.

4. Computational issues and how to deal with them

Computing the Shapley value is a #P-hard problem for many classes of games (see, e.g., Aziz & de Keijzer (2014); Bachrach & Rosenschein (2009); Deng & Papadimitriou (1994); Nagamochi et al. (1997)), including allocation games, even if goods may have only two possible values Greco & Scarcello (2014). In our application scenario, where agents are researchers and goods are research products, for large universities we may have to deal with thousands of agents and goods. The brute-force approach to the Shapley value would need to solve, for each agent $i \in N$, $2^{|N|}$ optimization problems, which is clearly infeasible for our test cases.

In order to mitigate the complexity of this problem, various input simplification techniques and approximation algorithms have been proposed and experimentally tested (see, e.g., Liben-Nowell et al. (2012); Maleki et al. (2013); Lupia et al. (2018)): we summarize them in the remainder of this section.

4.1. Input instance simplification

Several properties have been outlined in Lupia et al. (2018), which may be useful in simplifying allocation games without altering the Shapley value of the involved players. First, it has been proven that each connected component of the agents graph can be treated as a separate coalitional game. Next, it has been shown that goods having value 0 may safely be removed from the game (these goods, while not impacting on the computation of the optimal allocation, induce connections among agents that complicate the structure of the input graph). A third property states that, for any coalition $Z \subseteq N$ such that $\texttt{opt}(Z) + \texttt{opt}(N \setminus Z) \leq \texttt{opt}(N)$, the Shapley value may be calculated separately for subsets Z and $N \setminus Z$. Intuitively, this means that any subset of agents Z that do not exhibit an effective synergy with the rest, can be removed from the game and their Shapley value computed independently. Finally, a mathematical condition has been given to identify goods that are provably useless for some agents. Edges between goods and agents that do not need them may be safely dropped without altering the Shapley value.

4.2. Approximating the Shapley value

In order to approximate the Shapley value, one possibility is to use the fully polynomial randomized approximation scheme (FPRAS) described in Liben-Nowell et al. (2012). An alternative, when a range for the marginal contribution of each agent is known, is provided by the sampling method presented in Maleki et al. (2013). We will now briefly discuss both approaches.

FPRAS for supermodular and monotone coalitional games. As stated in Liben-Nowell et al. (2012), it is possible to compute, in polynomial-time, an $\epsilon-$approximation of the Shapley value, with probability of failure at most δ, for supermodular and monotone coalitional games (it can be shown that our allocation games are indeed of this type Greco & Scarcello (2014)).

The method is based on generating a certain number of permutations (of all agents) and computing the marginal contribution of each agent to the coalition of agents occurring before her in the considered permutation. Then the Shapley value of each player is computed as the average of all such marginal contributions. The above procedure (see Figure 4 for detailed pseudo-code) is repeated $O(\log(1/\delta))$ times, in independent runs, with the result for each agent consisting of the median of all computed values for her. Finally, the obtained values are scaled (i.e., they are all multiplied by a common numerical factor) to ensure that the budget-balance property is not violated. Clearly, the more permutations are considered, the closer to the Shapley value the result will be, the relation between the number m of required random permutations in each run and ϵ being $m = 4n(n-1)/\epsilon^2$, where n is the number of agents.

The algorithm runs in $O(m \times n \times margComp)$ time, where $margComp$ denotes the cost of computing each marginal contribution. This requires the computation of an optimal weighted matching in a bipartite graph, which is feasible in $O(n^3)$, via the classical Hungarian algorithm Kuhn (1955).

Sampling algorithm when the range of marginal contributions is known. Maleki et al. (2013) propose a bound, based on Hoeffding's inequality Hoeffding (1963), on the samples required to estimate an agent's Shapley value, when the range of his/her contributions is known. They prove that, in order to approximate the Shapley value of agent i within an absolute value ϵ_{abs} with probability larger than $1-\delta_i$, i.e., in order to get

$$Prob\{|\tilde{\phi}_i - \phi_i| \geq \epsilon_{abs}\} \leq \delta_i \tag{1}$$

at least m_i samples are required, where:

$$m_i = \left\lceil \frac{\ln\left(\frac{2}{\delta_i}\right) \cdot r_i^2}{2 \cdot \epsilon_{abs}^2} \right\rceil \tag{2}$$

In the above expression, r_i denotes the range of i's marginal contributions (i.e., $r_i = \texttt{opt}(\{i\}) - marg(\{i\}, N)$), where N is the set of agents that partecipate in the allocation game).

Given this bound, once ϵ_{abs} and δ_i are fixed, it is possible determine the number of required random samples for each agent i. Assuming an overall failure probability δ is

```
**Input:** An allocation game $\mathsf{G}_{\mathcal{A}} = \langle N, v_{\mathcal{A}} \rangle$;
**Parameters:** Real number $0 < \epsilon < 1$;
**Output:** A vector $\tilde{\phi}$ that is an approximation of the Shapley value of $\mathsf{G}_{\mathcal{A}}$;
$m = \frac{4 \cdot |N| \cdot (|N|-1)}{\epsilon^2}$;
$i = 0$;
$\tilde{\phi} = [0]$;
while $i < m$ do
    $\mathit{shuffle}(N)$;
    $C := \{\emptyset\}$;
    for all $j \in N$ do
        $\tilde{\phi}_j \mathrel{+}= v_{\mathcal{A}}(C \cup \{j\}) - v_{\mathcal{A}}(C)$;
        $C := C \cup \{j\}$;
        $i = i + 1$;
    end for
end while
for all $j \in N$ do
    $\tilde{\phi}_j = \frac{\tilde{\phi}_j}{m}$;
end for
return $\tilde{\phi}$;
```

Figure 4: FPRAS core procedure.

desired, each agent $i \in N$ could be assigned a failure probability $\delta_i = \delta/|N|$. Alternatively, a higher failure probability could be tolerated for agents with larger ranges, at the expense of a lower one for agents with smaller ranges.

Once the number of required samples for each agent has been computed, the approximate Shapley value, with the desired guarantees, can easily be computed by a randomized sampling algorithm, such as the one shown in Figure 4.

4.3. Lower and upper bounds for the Shapley value

In Lupia et al. (2018), techniques are introduced for the computation of lower and upper bounds for the Shapley value. The availability of such bounds can be helpful to provide a more accurate estimation of the approximation error in randomized algorithms. Besides, whenever the two bounds coincide for some agent, we clearly get the precise Shapley value for that agent. Furthermore, in some applications, when the Shapley value cannot be computed exactly, one may be more interested in a guaranteed interval, rather than one probably good point, as it is yielded by approximation algorithms.

Trivially, for each agent i in our monotone allocation games, his/her Shapley value will be in the interval between i's marginal contribution to the grand coalition, and i's optimal allocation. To obtain tighter bounds, Lupia et al. observe that the neighbors of i in a coalition C are the agents having the highest influence on the marginal contribution of i to C. Indeed, they are precisely those agents interested in using the goods of i when he/she does not belong to the coalition. On the other hand, in the extreme case that no neighbors are present, i contributes with all her/his best goods. The idea is to consider the power-set of $Neigh(i)$ as the only relevant sets of agents, and assign upper and lower bounds to the marginal contribution of i to a coalition C as a function only of $C \cap Neigh(i)$. A detailed algorithm is provided in Lupia et al. (2018), where it is shown that, for each element of the power set of $Neigh(i)$, only a constant number of optimal allocations have to be computed, leading to an $O(2^{|Neigh(i)|}|N|^3)$ running time.

4.4. Experimental results

Lupia et al. (2018) report an experimental study concerning the feasibility of computing the Shapley value for the researchers of Sapienza University of Rome (one of the largest universities in Europe) who participated in the most recent Italian research assessment exercise (namely, VQR2011-2014). We will now briefly summarize those results.

Parallel implementations of the algorithms described above were run on a machine equipped with two Intel Xeon E5-4610 v2 @ 2.30GHz with 8 cores and 16 logical processors each, for a total of 32 logical processors, 128 GB of RAM, and operating system Linux Debian Wheezy. The code was written in Java and run on the OpenJDK Runtime Environment (IcedTea 2.6.7). The input instance consisted of 3562 researchers and 5909 research products. The scores for the research products were computed by applying, when available, the bibliographic assessment tables provided by ANVUR.

The dataset described above was preliminarily simplified by applying the results outlined in Section 4.1. After the whole preprocessing phase, a total of 159 connected components (CCs henceforth) were obtained, with the largest one containing 685 authors, and all others having size at most 15. All components in the simplified graph, except for the largest one, could be completely solved by an exponential running time, exact Shapley value algorithm, in a matter of seconds. However, it was estimated that even the parallel implementation of the FPRAS method, in spite of its high efficiency when compared to the exact algorithm, would have taken, with $\epsilon = 0.05$ and $\delta = 0.01$, roughly 3.33 years to fully analyze the largest CC.

The approach provided by Maleki et al. proved much more efficient than the FPRAS method, at least for the case study considered, in which the range of marginal contributions for each agent is fairly limited. In its standard form, the algorithm by Maleki et al. guarantees an absolute error relative to the Shapley value (see eq. (1) in Section 4.2), as opposed to the FPRAS method, where ϵ represents a percentage relative to the correct value. Computing the Shapley value of all authors in the largest CC of the simplified graph, with 99% probability within an absolute error $\epsilon_{abs} = 5$, took approximately 5.5 hours on the considered experimental platform, while the same calculation on the largest CC in the unsimplified graph took approximately 20 hours.

In order to consider, for the Maleki-based algorithm, the classical percentage expression for the approximation error, as opposed to the absolute error provided by its standard form, ϵ_{abs} should be replaced by $\epsilon \cdot \phi_i$ in eq. (1). First observe that, in the case of the simplified graph, it will be $\phi_i \neq 0$ for any agents i considered by the algorithm, as the simplification techniques preliminarily identify and remove from the game all agents having Shapley value equal to 0 (these agents will be interested only in goods with value 0). In fact, the value of ϕ_i that would appear in eq. (2) may be replaced

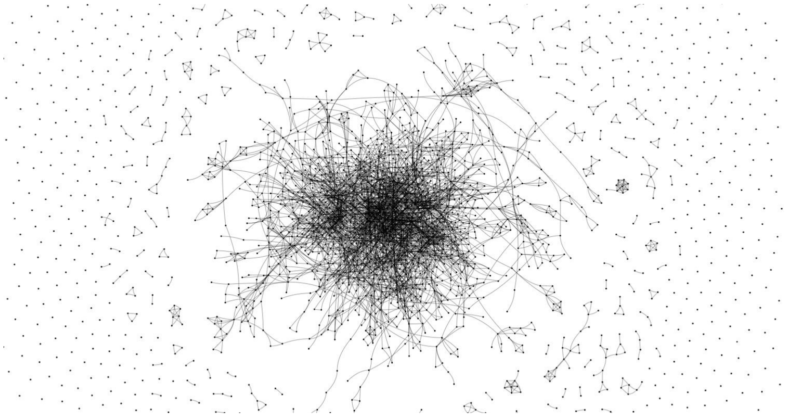

Figure 5: Sapienza Coautorship graph.

by any known (positive) lower bound $\ell_i \leq \phi_i$, at the expense of taking more samples than strictly necessary. For the largest CC of the simplified VQR2011-2014 graph, a parallel implementation of the techniques described in Section 4.3 allowed Lupia et al. to compute lower bounds for all agents in approximately 160 hours. These in turn allowed the Maleki-based algorithm to approximate the Shapley value of all agents, with 99% probability within 5% of the correct value, in approximately 12 hours.

5. Using Shapley values for publication credit allocation in VQR 2011-2014

In the previous section, we have summarized a number of technical results that allow for the computation of an approximation of the Shapley value, even for large instances including thousands of researchers and publications. As shown in Lupia et al. (2018), by using the simplification and approximation techniques described in Section 4, it is possible to solve the publication credit allocation problem for large instances using the Shapley value.

We wish to point out that the Sapienza test case is not only a large instance of a publication credit allocation problem, comprising 3562 researchers and 5909 research products, but it is also a complex one, due to the very high number of collaborations among Sapienza researchers. In Figure 5, we show the complex interaction of Sapienza researchers, where each node represents a Sapienza researcher and an arc denotes the existence of a coauthored article (among the set of articles selected by Sapienza for the VQR evaluation). As you can see, the graph has an extremely dense core, thus making the credit allocation problem very difficult to solve.

Unfortunately, we do not have access to the individual scores assigned by ANVUR to the submitted articles, but at least for the scientific areas that adopted a bibliometric assessment of the publications, we have been able to estimate the score of most of the submitted papers. Therefore, from now on, we will only concentrate on the authors and publications whose scores we could reasonably estimate. Moreover, we excluded the authors that refused to partecipate to the VQR (over 300 researchers and professors of

Sapienza joined a boycott of the VQR, one of the largest percentages in Italy). After this simplification, the remaining set of researchers and professors is composed of 2346 individuals who contributed 4638 articles. Notice that the simplification does not really decrease the complexity of the problem, since (almost) all of the excluded persons do not belong to the tightly connected core shown in Figure 5.

In order to assess the improvement in the fairness of the credit allocation using the Shapley value we compare, at the level of individuals, scientific fields (SSDs) and departments, the rankings currently used by ANVUR for different purposes and the rankings based on the Shapley value, namely, we consider the following division rules:

1. `proj`, also called "the VQR score";
2. `Best`;
3. `Shapley`.

We do not provide a comparison with the `owner` allocation because this allocation rule, due to the very limited number of papers submitted to (and evaluated by) ANVUR, would greatly penalize authors with large sets of coauthors. In order to make the scores more easily comparable, ANVUR published the average score for every SSD in its VQR2011-2014 report[4] and the average score for every SSD using the "Best" assignment[5]. In the sequel, following the ANVUR approach, we will compare the ratio of the individual scores to the average score of the SSD the reasercher belongs to (ANVUR calls it the R indicator). The VQR score and the Shapley value will be normalized using the average score, while the "Best" assignment will be normalized by using the average score for every SSD using the "Best" assignment. More precisely, given the allocation ψ used for the submission to ANVUR, we compute, for every researcher r,

- $R_{VQR}(r)$, as the ratio between $\texttt{proj}_{\psi}(r)$ and the average score of her SSD multiplied by the number of assigned publications;

- $R_{Shapley}(r)$, as the ratio between ϕ_r and the average score of her SSD multiplied by the number of assigned publications;

- $R_{Best}(r)$, as the ratio between $\texttt{Best}(r)$ and the average score of her SSD using the "Best" assignment multiplied by the number of assigned publications.

To summarize our results, we computed for every researcher the discrepancy between R_{VQR} and $R_{Shapley}$, and the discrepancy between R_{Best} and $R_{Shapley}$. More precisely, we define the percentual discrepancy between R_{VQR} and $R_{Shapley}$ as follows:

$$\frac{R_{VQR}(r) - R_{Shapley}(r)}{R_{Shapley}(r)} \cdot 100$$

[4] Available here (in both English and Italian) http://www.anvur.org/index.php?option=com_content&view=article&id=1206:vqr&catid=2:non-categorizzato&lang=it&Itemid=789

[5] Available here (in Italian only) http://www.anvur.org/index.php?option=com_content&view=article&id=455&Itemid=481&lang=it

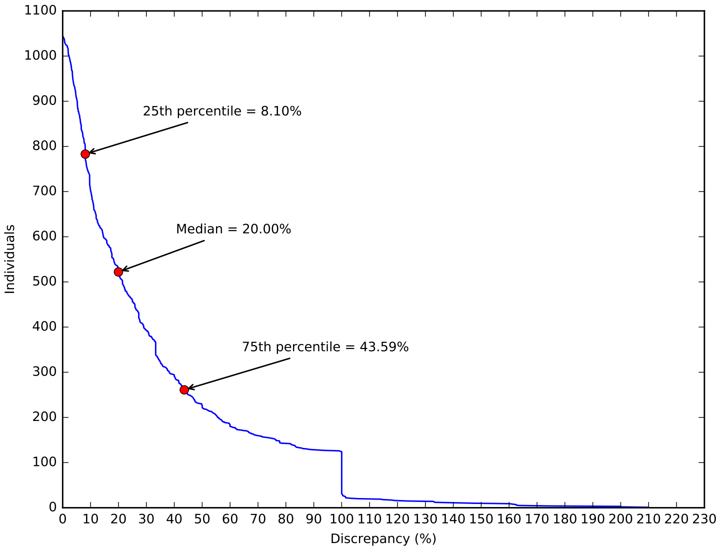

Figure 6: Number of researchers with at least a given percentage discrepancy between R_{VQR} and $R_{Shapley}$ (absolute value).

Analogously, the percentual the discrepancy between R_{Best} and $R_{Shapley}$ is defined as follows:

$$\frac{R_{Best}(r) - R_{Shapley}(r)}{R_{Shapley}(r)} \cdot 100$$

Figure 6 shows, for each absolute value v of discrepancy between R_{VQR} and $R_{Shapley}$, the number of researchers having discrepancy at least v. Moreover, the figure shows the median and the main percentiles evaluated over the population of the 1022 researchers having a discrepancy different than zero. Note that half of these researchers exhibits an absolute value discrepancy of at least 20% and there are 124 researchers with a discrepancy above 100%. See Figure 7, for a quick view of some notable ranges of (absolute value) discrepancy.

Table 2 reports also the discrepancies between R_{VQR} and $R_{Shapley}$ at SSD and at department level, and provides also positive and negative discrepancies, besides their absolute values (frequencies are cumulative, e.g., the number of researchers with discrepancies of 5% or higher also includes those with discrepancies of 10% or higher, and so forth).

At the scientific discipline level (SSD), 5 SSDs have a positive discrepancy of over 20%. Things are better at the department level (discrepancies tend to compensate within departments), but there is at least one department that has a discrepancy of over 15%. It is evident that the VQR assignment is not always a good indicator of the performances of departments, scientific disciplines, and individuals. Nevertheless, it has been used both for compiling a ranking of all departments, and for assigning significant funds through the "Excellent Departments" program. Notice that the ranking used in this funding program is highly sensitive to small changes in R_{VQR} A depart-

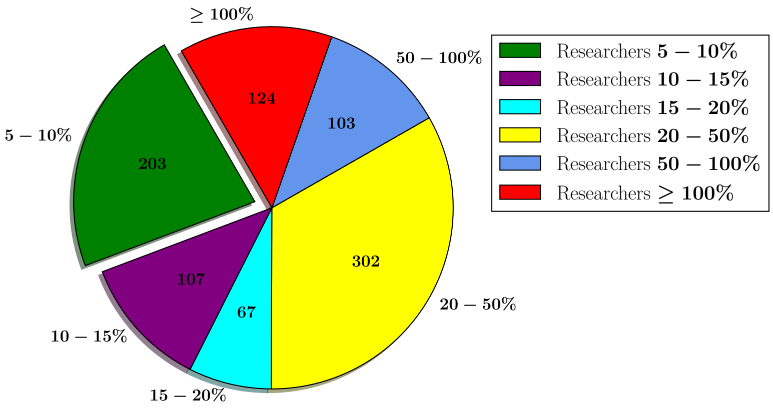

Figure 7: Number of researchers with ranges of absolute value discrepancies between R_{VQR} and $R_{Shapley}$.

	Researchers	SSDs	Departments
Total	2347	181	57
discr $\geq 100\%$	56	0	0
discr $\geq 50\%$	115	0	0
discr $\geq 20\%$	284	2	0
discr $\geq 15\%$	319	2	1
discr $\geq 10\%$	380	6	1
discr $\geq 5\%$	500	14	1
discr $\leq -100\%$	68	0	0
discr $\leq -50\%$	112	0	0
discr $\leq -20\%$	245	3	0
discr $\leq -15\%$	277	4	0
discr $\leq -10\%$	323	7	0
discr $\leq -5\%$	406	20	2
$\lvert$discr$\rvert \geq 100\%$	124	0	0
$\lvert$discr$\rvert \geq 50\%$	227	0	0
$\lvert$discr$\rvert \geq 20\%$	529	5	0
$\lvert$discr$\rvert \geq 15\%$	596	6	1
$\lvert$discr$\rvert \geq 10\%$	703	13	1
$\lvert$discr$\rvert \geq 5\%$	906	34	3

Table 2: Discrepancies between R_{VQR} and $R_{Shapley}$.

ment may lose some millions of euros because of a difference of a few decimals in its

	Researchers	SSDs	Departments
Total	2347	181	57
discr $\geq 100\%$	42	0	0
discr $\geq 50\%$	191	4	0
discr $\geq 20\%$	388	7	0
discr $\geq 15\%$	458	12	1
discr $\geq 10\%$	538	22	1
discr $\geq 5\%$	614	43	1
discr $\leq -100\%$	0	0	0
discr $\leq -50\%$	0	0	0
discr $\leq -20\%$	8	3	0
discr $\leq -15\%$	91	4	0
discr $\leq -10\%$	426	18	0
discr $\leq -5\%$	914	42	2
$\lvert$discr$\rvert \geq 100\%$	42	0	0
$\lvert$discr$\rvert \geq 50\%$	191	4	0
$\lvert$discr$\rvert \geq 20\%$	396	10	0
$\lvert$discr$\rvert \geq 15\%$	549	16	1
$\lvert$discr$\rvert \geq 10\%$	964	40	1
$\lvert$discr$\rvert \geq 5\%$	1573	85	3

Table 3: Discrepancies between R_{Best} and $R_{Shapley}$.

evaluation.[6].

ANVUR is aware of the limitations of R_{VQR} to assess individuals and, in order to overcome these limitations, it proposed the use of R_{Best} to evaluate smaller groups, such as PhD committees. However, as we show in Table 3, the discrepancies between R_{Best} and $R_{Shapley}$ are even higher than those between R_{VQR} and $R_{Shapley}$.

The disadvantages of R_{Best} are the following:

- It penalizes top performers. Everyone who obtained the top scores has an R_{Best} that is lower than R_{VQR}, since the numerator cannot improve but the denominator gets larger.

- It favors researchers who work in larger groups with respect to researchers who work in smaller groups. Since these results have been used to distribute funding among the universities and departments, the effect is to make large groups stronger and small ones weaker. In the long run, such a policy would force all departments to concentrate all researchers in large groups working in few research areas, thus reducing the breadth of the research areas.

[6] Details (in Italian only) can be found here `http://hubmiur.pubblica.istruzione.it/alfresco/d/d/workspace/SpacesStore/a8a56378-d9f4-44cd-b0dd-7bce0d2f1b7d/Nota_metodologica_ISPD_ANVUR.pdf`

As we already mentioned, the VQR scores have been used in a number of additional assessments. We briefly discuss the impact of the reported discrepancies on the assessments of PhD committees and the selection and ranking of departments of excellence.

1. PhD committees: the ANVUR guidelines dictated, among other requirements, that the member of PhD committees had values of the indicator $R_{Best} \geq 1$. Since the computation is done at the level of single members of the committees, we argue that the error introduced by the use of R_{Best} instead of $R_{Shapley}$ is rather significant.
2. Departments of excellence: the ranking computed by the Ministry of Education, Universities, and Research (MIUR) compared the performance of each department with respect to a "Virtual Department" with the same composition, but where all the products of the researchers obtain the average value of the SSD to which they belong. Moreover, the individual R_{VQR} was divided by the variance of the distribution of scores in the corresponding SSD. Since in many SSDs, variances can be quite small, the normalization actually corresponds to a multiplier that also increases the error in the indicator, thus multiplying the discrepancy.

6. Related work

In Section 6.1, we address various schemes related to the Shapley value, discussing their features in the context of coauthorship analysis. We also put our contributions into perspective by considering related work on the current and previous Italian research assessment exercises in Section 6.2.

6.1. Shapley value-related articles

Papapetrou et al. (2011) approximate the Shapley value by considering not all subsets of authors, but only those including only authors who have actually coauthored at least one publication. Whenever a value for $v(S \cup i)$ is known (where S is a coalition and i is an author not in S), if a value for $v(S)$ is not known (i.e., there is not even a single publication with author set S), it is composed from subsets. This procedure leads to an iterative algorithm, which does not provide any theoretical guarantees that the obtained scores satisfy any of the fairness properties enjoyed by the exact Shapley value. However, the authors argue, and experimentally demonstrate, that their approximation yields fairer results than a simple approach based on the equal division of publication scores among coauthors.

Tol (2012) computes a pseudo-Shapley value taking into account not all possible coalitions, but only the coalitions that correspond to existing research institutions (the article focuses on business schools in Ireland). The value of a coalition is defined as either the average number of publications by the authors in it, or the average number of citations obtained by publications coauthored by the authors in it. Tol divides his pseudo-Shapley value into two terms, one representing the contribution of an author to her institution (a measure of her "power" over the institution), and the other representing her contribution to the other institutions (i.e., a measure of her "market value").

Rankings, of both authors and institutions, based on these measures are either identical, or very similar, to those obtained using more conventional performance indicators, at least for the considered test cases. Insights into the robustness (or fragility) of institutions' ranks may be gained from the power measures of their authors. It should be noted that, analogously to the approach by Papapetrou et al., also in this case there are no theoretical guarantees that the obtained scores satisfy any of the fairness properties enjoyed by the exact Shapley value.

Karpov (2014) proves the equivalence of `owner` and the Shapley value, for specific forms of the characteristic function v that describes the cooperative game. Recall that `owner` is based on equal weights coauthorship sharing, and simply divides equally between coauthors the score of each publication. The score of each author i, according to this division rule, can be computed in polynomial time. Following Karpov's notation, it can be written as

$$y_i = \sum_{j=1}^{m} \frac{1_{i \in S_j}}{|S_j|} \cdot q_j$$

where q_j and S_j are, respectively, the score and the set of coauthors of publication j, m is the total number of publications, and $1_{i \in S_j} = 1$ if $i \in S_j$ and $1_{i \in S_j} = 0$ otherwise. Karpov proves that y_i coincides with the Shapley value of author i, as long as the characteristic function v, used to compute it, takes one of the following three forms:

$$v_1(S) = \sum_{j=1}^{m} 1_{S_j \subseteq S} \cdot q_j \quad \text{(full obligation game)}$$

$$v_2(S) = \sum_{j=1}^{m} 1_{S_j \cap S \neq \emptyset} \cdot q_j \quad \text{(full credit game)}$$

$$v_3(S) = \sum_{i \in S} \sum_{j=1}^{m} \frac{1_{i \in S_j}}{|S_j|} \cdot q_j \quad \text{(equal weights game)}$$

where S denotes any subset (coalition) of authors. In the full obligation game, a publication's score is added to the coalition score only if all its coauthors are members of the coalition. In the full credit game, each publication score is added possibly multiple times, once for every coauthor present in the coalition. In the equal weights game, each coauthor is assigned an equal portion of the publication's score, and only the score fractions of coauthors in the coalition are added. We note that, as sensible as the above criteria might appear, none of them reflects the operations that each Italian research institution should perform, according to VQR rules, in order to compute an optimal assignment of research products to its staff members. Moreover, we have shown in Section 3 that this division rule is unfair, as far as the research evaluation process is concerned, because it does not satisfy the marginality property and because it depends on the specific set of (optimal) products submitted for the evaluation.

6.2. Italian research assessment exercises-related articles

Franceschini & Maisano (2017) present a structured discussion of VQR 2011-2014, collecting several critical remarks, and developing them in detail. Identified methodological vulnerabilities include the following. (1) The number of products evaluated for each researcher is too small to allow identification of the level of excellence, or even the average quality of research institutions. It is argued that all that can be identified are the less virtuous institutions. (2) The use of journal metrics to evaluate individual articles can be misleading, as the variability in the number of citations received by articles published in the same journal is generally high. (3) VQR 2011-2014 normalizes citational data and journal impact factors using percentile ranks (per year and per area of research) and then combines them in a weighted sum. It is argued that such a combination of only apparently compatible indicators may lead to rank distortions. (4) Several "calibration" operations, relative to the bibliometric evaluation procedures (for example, choosing relative weights for citation counts and journal impact factors), are entrusted to panels of experts nominated for each research area. It is suggested that, in the absence of solid guidelines, this additional degree of freedom may be counterproductive. (5) VQR 2011-2014 uses a hybrid approach based both on peer review and on blibliometric evaluation. It is assumed that such evaluations are mutually compatible, an assumption that does not seem to be supported by adequate empirical evidence.

Point (3) of the above list is also addressed in Abramo & D'Angelo (2016), where it is argued that the combination of citational data and journal metrics in VQR 2011-2014 is not justified on the basis of current scientific literature, and that a simple citation count mechanism would have yielded a better prediction on the long term impact of a publication (and therefore a more accurate score), beginning from a citation window of at least two years.

Methodological objections have also been raised in the context of VQR 2004-2010, some related to those concerning VQR 2011-2014, which is hardly surprising, given the similarities between the two evaluation exercises (see Section 2.1). In particular, Abramo & D'Angelo (2015) identify limits of the bibliometric criteria employed for hard sciences in VQR 2004-2010, including the following. (1) The use of journal metrics as an indication of a publication's impact; the authors argue that citation counts are much more reliable indicators, except for very short citation windows. (2) The failure to consider product quality scores as a continuous range, as the introduction of discrete classes, along with value caps, may have led to the same overall evaluation for publications with quite different bibliometric indicators. (3) The full counting of a publication's score, regardless of the number of coauthors. (4) The limited number of products evaluated.

Rebora & Turri (2013) focus on the way in which research evaluation exercises in the UK and Italy have been acknowledged by universities, the changes that they have undergone over time, and the kind of public debate which accompanied these changes. The authors argue that British RAE (Research Assessment Exercise) has influenced the development of research in the UK by favoring mono-disciplinary publications over multi-disciplinary ones. At the same time, however, it is generally acknowledged that RAE gave a positive contribution to research in the UK, as it is reflected in world university rankings. The most notable change from RAE to its successor REF (Research Excellence Framework), which, very much like its predecessor, is based on peer review, is an attempt to evaluate not only research quality, but also its impact on economy, society and culture. Italy's first research assessment exercise, VTR 2001-2003, which was conducted 15 years later than the earliest RAE, was inspired by it. Its successor, VQR 2004-2010, diverged instead from its British counterpart due to the use of bibliometric indicators for many scientific areas. The authors claim that, compared to the UK, research assessment exercises in Italy were characterized by less debate and a more passive reception. They also note that, differently from the British case, resource allocation for universities, in the years that followed VTR 2001-2003, was determined by a complex algorithm in which the relevance of the assessment itself was reduced, as a large number of other indicators were considered as well. It is also observed that the decision, in VQR 2004-2010, to make wide use of citation data for the scientific areas subject to bibliometric evaluation, and its effects at the micro (i.e., individual) level have not been the subject of in-depth and shared discussions.

7. Conclusions and future work

While the primary goal of national research assessment exercises remains the evaluation of structures (e.g., universities) and substructures (e.g., departments), their results may also be used for other assessments that more directly involve individuals. This has been the case for Italy's most recent research assessment effort, namely VQR 2011-2014, whose data have been used for the evaluation of members of PhD committees as well as for the determination of the so called "departments of excellence". These uses of VQR data, however, require that credit allocation for the submitted research products be done in a fair way. Unfortunately, the way credit allocation is currently performed by ANVUR (the national agency entrusted with VQR execution) yields evaluations that do not satisfy properties that, from a methodological perspective, are highly desirable. We have pointed out that, due to recent algorithmic advances, it is possible to compute reasonably good approximations for the Shapley value (a game-theoretic solution concept for the credit allocation problem that enjoys several "fairness" properties) of individual authors, even for large input instances. We have presented a detailed comparison of the scores assigned by the official VQR rules to individual researchers, research groups and departments belonging to Sapienza University of Rome (one of the largest universities in Italy), with those resulting from our Shapley value computations. Significant discrepancies have emerged at both the individual level and at the level of scientific disciplinary sectors, while score variations seem to more or less compensate within departments, albeit with some exceptions.

As detailed in Section 4.4, our results are based on the exact Shapley value for the vast majority of researchers (2877 out of 3562), while for the remaining 685 (19.23% of the total) we were able to compute quite good approximations (i.e., within 5% of the correct value with 99% probability) in roughly 172 hours. We are currently working on identifying further theoretical properties of the Shapley value in allocation games, with the aim of developing more sophisticated tools for simplifying real-world VQR instances, and consequently being able to compute the exact Shapley value for all researchers.

Our input preprocessing and approximation techniques can be applied to other instances of VQR 2011-2014, referring to different universities, or even to other types of research institutions, assuming that the relevant data were available. This paves the way for an interesting comparative study.

Our techniques can also be applied to previous research evaluation exercises (e.g., VQR 2004-2010), even though with some caveats (the number of required products per researcher was larger in VQR 2004-2010, leading to more complex instances, and negative scores were also used, e.g., for missing products, implying slight differences in how this research evaluation exercise should be modeled as an allocation game). This could provide unique insights on how, over the years, evaluation efforts have contributed changing the way scientific research is conducted in Italy.

References

Abramo, G., & D'Angelo, C. (2011). National-scale research performance assessment at the individual level. *Scientometrics*, *86*, 347–364.

Abramo, G., & D'Angelo, C. (2015). The VQR, Italy's second national research assessment: Methodological failures and ranking distortions. *Journal of the Association for Information Science and Technology*, *66*, 2202–2214.

Abramo, G., & D'Angelo, C. (2016). Refrain from adopting the combination of citation and journal metrics to grade publications, as used in the italian national research assessment exercise (vqr 2011-2014). *Scientometrics*, (pp. 1–13).

Abramo, G., D'Angelo, C., & Di Costa, F. (2011). National research assessment exercises: the effects of changing the rules of the game during the game. *Scientometrics*, *88*, 229–238.

Aziz, H., & de Keijzer, B. (2014). Shapley meets shapley. In *31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014), STACS 2014, March 5-8, 2014, Lyon, France* (pp. 99–111).

Bachrach, Y., & Rosenschein, J. S. (2009). Power in threshold network flow games. In *Autonomous Agents and Multi-Agent Systems* (pp. 106–132).

Blasi, B., Romagnosi, S., & Bonaccorsi, A. (2016). Playing the ranking game: media coverage of the evaluation of the quality of research in italy. *Higher Education*, (pp. 1–17).

Deng, X., & Papadimitriou, C. H. (1994). On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, *19*, 257–266.

Franceschini, F., & Maisano, D. (2017). Critical remarks on the italian research assessment exercise vqr 2011-2014. *Journal of Informetrics*, *11*, 337 – 357.

Greco, G., Lupia, F., & Scarcello, F. (2015). Structural tractability of shapley and banzhaf values in allocation games. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015* (pp. 547–553).

Greco, G., & Scarcello, F. (2013). Fair division rules for funds distribution: The case of the italian research assessment program (vqr 2004-2010). *Intelligenza Artificiale*, *7*, 45–56.

Greco, G., & Scarcello, F. (2014). Mechanisms for fair allocation problems: No-punishment payment rules in verifiable settings. *J. Artif. Intell. Res. (JAIR)*, *49*, 403–449.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*, 13–30.

Iera, A., Militano, L., Romeo, L., & Scarcello, F. (2011). Fair Cost Allocation in Cellular-Bluetooth Cooperation Scenarios. *IEEE Transactions on Wireless Communications*, *10*, 2566–2576.

Karpov, A. (2014). Equal weights coauthorship sharing and the shapley value are equivalent. *Journal of Informetrics*, *8*, 71–76.

Kuhn, H. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*, 83–97.

Liben-Nowell, D., Sharp, A., Wexler, T., & Woods, K. (2012). Computing Shapley Value in Supermodular Coalitional Games. In J. Gudmundsson, J. Mestre, & T. Viglas (Eds.), *Computing and Combinatorics* (pp. 568–579). Springer Berlin Heidelberg volume 7434 of *Lecture Notes in Computer Science*.

Lupia, F., Mendicelli, A., Ribichini, A., Scarcello, F., & Schaerf, M. (2018). Computing the shapley value in allocation problems: approximations and bounds, with an application to the italian vqr research assessment program. *Journal of Experimental & Theoretical Artificial Intelligence*, *0*, 1–20.

Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., & Rogers, A. (2013). Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. *CoRR*, *abs/1306.4265*.

Maniquet, F. (2003). A characterization of the Shapley value in queueing problems. *Journal of Economic Theory*, *109*, 90–103.

Mishra, D., & Rangarajan, B. (2007). Cost sharing in a job scheduling problem. *Social Choice and Welfare*, *29*, 369–382.

Moulin, H. (1992). An application of the Shapley value to fair division with money. *Econometrica*, *60*, 1331–49.

Nagamochi, H., Zeng, D.-Z., Kabutoya, N., & Ibaraki, T. (1997). Complexity of the minimum base game on matroids. *Mathematics of Operations Research*, *22*, 146–164.

Osborne, M. J., & Rubinstein, A. (1994). *A Course in Game Theory*. Cambridge, MA, USA: The MIT Press.

Papapetrou, P., Gionis, A., & Mannila, H. (2011). A shapley value approach for influence attribution. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6912 LNAI*, 549–564.

Rebora, G., & Turri, M. (2013). The uk and italian research assessment exercises face to face. *Research Policy*, *42*, 1657–1666.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the theory of games*, *2*, 307–317.

Tol, R. (2012). Shapley values for assessing research production and impact of schools and scholars. *Scientometrics*, *90*, 763–780.