A Shapley-value Approach for Influence Attribution

Panagiotis Papapetrou †, Aristides Gionis ‡, and Heikki Mannila †

†Department of Information and Computer Science, Aalto University,
†Helsinki Institute for Information Technology (HIIT), Finland.
‡Yahoo! Research, Barcelona, Spain.

Abstract. Finding who and what is "important" is an ever-occurring question. Many methods that aim at characterizing important items or influential individuals have been developed in areas such as, bibliometrics, social-network analysis, link analysis, and web search. In this paper we study the problem of attributing influence scores to individuals who accomplish tasks in a collaborative manner. We assume that individuals build small teams, in different and diverse ways, in order to accomplish atomic tasks. For each task we are given an assessment of success or importance score, and the goal is to attribute those team-wise scores to the individuals. The challenge we face is that individuals in strong coalitions are favored against individuals in weaker coalitions, so the objective is to find fair attributions that account for such biasing. We propose an iterative algorithm for solving this problem that is based on the concept of Shapley value. The proposed method is applicable to a variety of scenarios, for example, attributing influence scores to scientists who collaborate in published articles, or employees of a company who participate in projects. Our method is evaluated on two real datasets: ISI Web of Science publication data and the Internet Movie Database.

1 Introduction

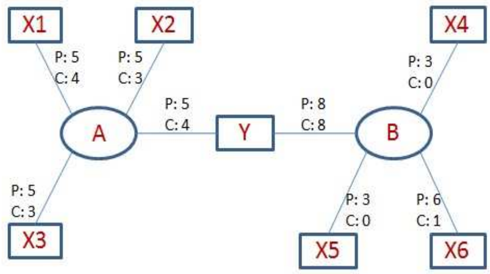
People have always been intrigued by characterizing "influential" ideas, books, inventions, scientists, politicians, art movements, etc. The question of finding important or influential items has attracted a lot of attention in social sciences, computer science, as well as other fields. For instance, the analysis of social networks has developed many methods to identify "important" individuals. Research in bibliometrics has provided many methods, typically based on citations, in order to assess the "impact factor" of publications, conferences, or journals. In information retrieval and web search, link-analysis methods, such as PageRank [18] and HITS [14], aim at identifying "authoritativeness" of web documents; furthermore, those link-analysis ideas have been applied to a wide variety of other scenarios. Recently, with the explosion of user-generated content a lot of emphasis has been placed in identifying influential users in blogs, micro-blogs, question-answering portals, and other social-media sites.

In this paper we address a novel problem in the context of characterizing who is influential. In particular, we focus on the case of attributing influence scores to a set of individuals who accomplish tasks in a collaborative manner. The setting we consider is as follows: we start with a large set of individuals; based on their interests, expertise, skills, and social dynamics, individuals form small teams in order to accomplish various tasks; we assume that for each of the accomplished tasks a value of importance (or influence) can be obtained. The problem we address is how to attribute to the individuals of each team the influence scores that we obtain for the whole team. We call this problem the *influence-attribution problem.*

Influence attribution has applications in many scenarios in which we want to estimate importance of individuals when importance evaluation scores are available at a team level. An obvious application of our setting is estimating importance scores for scientists through analysis of co-authorship information in scientific publications. For the latter problem, assigning importance scores to specific publications has received a lot of attention, and it is typically accomplished via analysis of citation graphs; however, how to distribute those scores to the individual authors of the publications is a much less studied problem. Other application scenarios include finding importance of company employees who participate in various projects, editors of wikipedia articles, bloggers in sites of collective blogging, artists in collaborative art spaces, and in general in any online collaboration community.

A simple approach for attributing influence scores to individuals is to divide the score of a task equally to all team members. However, in many cases, an equal division of importance scores may introduce inaccuracies. For instance, imagine a very talented individual who is capable of creating influential work in a number of different teams with a wide range of collaborators. In such a case, if there is evidence in the data that a large part of the influential work is due to this talented individual, it should be appropriate that she will receive more credit than her collaborators. The following example, in the context of scientific publications, demonstrates in more detail that equal division of importance score may be misleading.

Example. Consider the example shown in Figure 1. Let B be a researcher who, during a postdoctoral period, collaborated with a very distinguished researcher Y, and together they produced a total of 8 publications. Each of these publications has so far acquired 8 citations. Since the end of the postdoctoral period, B produced 12 additional publications, collaborating with 3 other researchers, but not with Y. None of these additional publications, however, received a significant amount of citations. Based on Figure 1, the total number of publications produced by A is 20, the total citation count is 70, while the H-index is 8. In the meantime, researcher A also spent a postdoctoral period collaborating with Y and produced 5 publications. Each of these publications received 4 citations. After the postdoctoral period, A collaborated with 3 other researchers producing 15 additional publications. In particular, each publication produced by the collaboration with *X1* received 4 citations and each publication produced by the collaboration with *X2* and *X3* received 3 citations. Thus, the total number of publications of A is 20, the total citation count is 70, while the H-index is 4.



Author	Total number of papers	Total citation count	H-index
A	20	70	4
B	20	70	8

Fig. 1. An example showing that taking into account author coalitions is important. P denotes the number of papers per author and C the number of citations per paper. Thus, author $X1$ received 20 citations, authors $X2$ and $X3$ received 15 citations, authors $X4$ and $X5$ 0 citations, and author $X6$ 6 citations. Also, author Y received 84 citations. For simplicity, we have kept the citation numbers unrealistically small.

The key question now is the following: which of the two researchers is more "influential?" Taking into account the bibliometric measures, B should be favored for having a higher H-index than B. But looking a bit deeper, one may notice that B managed to obtain all the "fame" only because of collaborating with Y. So, isn't it just that Y is a very strong researcher or has a very strong team that makes B influential? In other words, if Y is dropped from the picture, then the remaining publications of B are insignificant. On the other hand, A has also collaborated with the same number of people as B, though all collaborations were fruitful. Thus, it is more likely that A is a much stronger researcher than B, and should be definitely favored.

One of course could make the following alternative conclusion: A has better selected collaborators but, in general, he/she is weak. However, in this case it is not just that A made a wise selection of collaborators, but these collaborators chose A as well. Thus, A should be strong since he/she has been chosen by strong collaborators.

Our approach. Our solution to the problem of influence attribution is inspired by the concept of *Shapley value*, which was introduced by Lloyd Shapley in 1953 [22]. The Shapley value is a game-theoretic concept devised for fair division of gains obtained by co-operation of n players, namely in a setting very similar to the one we consider in this paper. It can be shown that the division made by the Shapley value satisfies very natural fairness properties, such as efficiency, individual fairness, symmetry, and additivity [23]. However, applying the concept of Shapley value in real scenarios is highly impractical, since the theory assumes that it is possible to assess the expected gain for every possible co-operation, that is, for all 2^N possible teams, where N is the total number of individuals.

To address the inefficiency that stems from the definition of Shapley value, we propose an iterative algorithm that aims to approximate the Shapley value using only co-operations for which an estimate of the expected importance score exists. For example, in the case of co-authorships and scientific publications, the only co-operations used by the algorithm are teams of authors who have written a paper together. Even though the total number of co-operations that occur in real data is very small compared to the number of all possible co-operations, and there is no theoretical guarantee that the scores obtained by our method satisfy any fairness conditions, our approach takes into account the marginal contribution of the individuals to the teams into which they participate. Thus, we argue and experimentally demonstrate that our method produces more fair attributions than the simple baseline of equal division.

Contributions. The main contributions of this paper include:

- the formulation of the problem of finding influential individuals in a collaborative environment,
- an iterative method to solve the influence-attribution problem that is based on the Shapley value,
- an experimental evaluation on two large and commonly used datasets: ISI Web of Science and the Internet Movie Database.

The remainder of this paper is organized as follows: in Section 2 we present the related work, in Section 3 we provide the necessary background along with the problem formulation, Section 4 presents the proposed methodology, in Section 5 we present the experimental results, and finally, in Section 6 we conclude the paper and discuss directions for future work.

2 Related work

Social-network analysis. Researchers in social-network analysis have developed many methods to measure "importance" of individuals in an implicitly- or explicitly-defined social network. In the basic model, the network is represented by a *directed* graph $G = (V, E)$, where the nodes represent individuals, and the edges model "endorsement" from one individual to another. For such a directed graph, the concept of *in-degree* is the simplest measure of importance of an individual. Refinements of the in-degree measure include the Katz-index [12] and the Hubbell index [11]. Notions of importance in social networks have been proposed also for the case of undirected graphs, most of those measures rely on various notions of *centrality*, e.g., degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, etc. For an overview of those notions see the manuscript of Newman [17]. A more interesting centrality concept, more closely related with this paper, is the concept of *game-theoretic centrality* [1, 9].

Bibliometrics. Research in bibliometrics studies the use of citations in order to measure the "impact" of scientific articles (e.g., journal articles, conference proceedings) or publication venues (e.g., peer-reviewed conference proceedings, scientific journals). A large number of measures has been introduced, again starting from the simplistic citation count. The well-known *Garfield's impact factor* [7] is the average number of citations by articles published the previous two years, while the *Crown indicator* [21] is a normalized version of such a citation count. Pinski and Narin [19] and subsequently Geller [8] observe that not all citations are equally important, and using a recursive definition similar to PageRank [18, 20] and HITS [14], they propose that a journal is "influential" if it is cited by other "influential" journals. The above measures are used to assign impact scores on publication "units" (e.g., articles, journals) and not to individual authors. Thus, the use of such measures is orthogonal and complementary to our work, where we aim attribute those impact scores for articles to individual authors. On the other hand, measures such as the *H-index* [10], and the *G-index* [6] (defined in Section 5) have been proposed for assigning impact scores to individual authors, and thus, are directly comparable to our proposed measure. However, both H-index and G-index do not attempt to address the issue of coalition bias as our method does. Overall, any bibliometrics impact index needs to be used with extreme care. As such our method is intended to complement and enrich existing analytics toolboxes, rather than substitute well-established methods.

Web search rankings. A prominent application domain of importance measures is in the area of web search and hypertext ranking. The goal is to assign importance scores to web documents in order to assist users locate the most relevant results for their searches. Not surprisingly many of the importance measures discussed above can also by used in the case of web search ranking, however, the two most well-known techniques are PageRank [18] and HITS [14]. Many variants of those methods have been proposed, as well as adaptations of those basic methods for different objectives; a thorough survey is beyond the scope of this paper.

Information diffusion and viral marketing. A different setting for studying influential individuals in social networks is through the concepts of *information diffusion* and *viral marketing*. This setting is described by a dynamic process. It assumes a model of information spread in the social network, and influential individuals are those who can act as good initiators of the information spread. Building on the seminal work of Domingos and Richardson [5] and Kempe et al. [13], a large number of papers has studied variants of the problem and proposed different solutions, including game-theoretic and Shapley-value-based solutions such as the work of Narayanam and Narahari [16]. However, the overall setting of information-diffusion processes is very different from the problem studied in this paper.

Coalitional games and Shapley value. Studying coalitional games is a major area of research in game theory. In contrast to competitive games, to which the standard solution concept is the Nash equilibrium, in coalitional games the players cooperate and receive collective payoff. The Shapley value [22] is the classic solution concept for distributing the payoff "fairly" among the members of the coalition. The model of coalitional game theory provides a rich framework for studying a variety of solution concepts and their properties [2–4]. Finding the allocations described by these concepts is in general computationally intractable, hence, much of the research in this area focuses on a variety of restricted domains in which one can hope to find approximate solutions. An overview of the theory of coalitional games is beyond the scope of the paper. However, a key difference from our work is that in this line of research approximate solutions are searched by allowing to sample random coalitions. Instead, we follow a more pragmatic approach where only a small number of coalitions is given as input, and we seek solutions by probing only those coalitions.

3 Problem Setting

We consider a set $\mathcal{V} = \{V_1, \ldots, V_n\}$ of n individuals and a set $\mathcal{T} = \{T_1, \ldots, T_m\}$ of m tasks. Each individual $V_i \in \mathcal{V}$ has participated in a subset of tasks of $\mathcal{T}$, which we denote by $\mathcal{T}_{V_i}$. Moreover, each task T_j is given an *impact score* $I_j = f(T_j)$; the higher the impact score the more "valuable" the task. Examples for function $f(\cdot)$ will be provided in Section 3.1.

Based on the participation of individuals to tasks, and the impact score of each task, our goal is to estimate the *influence score* ϕ_i of each individual $V_i \in \mathcal{V}$. Here, we assume that influence scores are non-negative numbers such that $\phi_i \geq \phi_j$ implies that V_i is at least as influential as V_j.

Hence, the *influence-attribution problem* can be formalized as follows:

Problem 1 ***(Influence-attribution)*** *Given a set of individuals $\mathcal{V} = \{V_1, \ldots, V_n\}$, a set of tasks $\mathcal{T} = \{T_1, \ldots, T_m\}$, and a set of impact scores $\mathcal{I} = \{I_1, \ldots, I_m\}$, with $I_j = f(T_j)$, compute the set of influence scores $\phi = \{\phi_1, \ldots, \phi_n\}$, where ϕ_i is the influence score of individual $V_i \in \mathcal{V}$.*

3.1 Example: author-publication instantiation

Let us now instantiate the above problem setting by considering a task to be a scientific publication and an individual to be a scientist. We call this the *author-publication* instantiation. In such setting, there may exist citations among different publications, so we define one additional concept.

Definition 1 ***(Incoming Citations)*** *Given a set of publications $\mathcal{T}$, for each publication $T_j \in \mathcal{T}$ we define the set of* incoming citations *to T_j as follows:*

$$\mathcal{C}^{in}_{T_j} = \{T_k \mid T_k \in \mathcal{T} \text{ and } T_k \text{ cites } T_j\}. \quad (1)$$

As mentioned above, $I_j = f(T_j)$ is the impact score of publication $T_j \in \mathcal{T}$. For the author-publication instantiation, we consider two options for $f(\cdot)$:

- `CC`: the citation count of each publication corresponds to the total number of citations received by the publication.

$$f(T_j) = |\mathcal{C}^{in}_{T_j}|, \quad \text{for each } T_j \in \mathcal{T}. \quad (2)$$

– PR: the PageRank score of each publication, which is computed by applying the PageRank algorithm on the citation network.

The definition of CC is simple enough and rates each incoming citation equally. The PR score is computed by the well-known algorithm for ranking web documents [18], which is based on estimating the stationary distribution of a random walk in the citation graph.

4 Methods

We present two methods for solving Problem 1. The first one is a straightforward uniform assignment of impact scores to the corresponding individuals, whereas the second one exploits the Shapley value in order to take into account coalitions of individuals and the impact score in these coalitions.

4.1 Naïve approach

The first approach to solve Problem 1 is to assign individuals with the mean impact score of their tasks.

Definition 2 ***(Mean Impact)*** *Given an individual V_i, the set of tasks $\mathcal{T}_{V_i}$ assigned to V_i, and their corresponding impact scores $\mathcal{I} = \{I_1, \ldots, I_{|\mathcal{T}_{V_i}|}\}$, the mean impact of V_i is defined by*

$$\phi_i = \frac{1}{|\mathcal{T}_{V_i}|} \sum_{T_j \in \mathcal{T}_{V_i}} I_j. \tag{3}$$

For the remainder of the paper, this method will be denoted as Naïve.

4.2 The Shapley-value approach

The second method is based on the concept of Shapley value [15], which is a way to divide goods gained by cooperation among many individuals. Specifically, consider an underlying set $\mathcal{V}$ and assume that for all possible subsets $\mathcal{S} \subseteq \mathcal{V}$, we know the value of $v(\mathcal{S})$, called *gain function*, which expresses the gain achieved by the cooperation of the individuals in $\mathcal{S}$.

Definition 3 ***(Share Allocation)*** *Given a gain function $v(\cdot)$, the* share allocation *$\phi_i(v)$ to individual V_i is defined as:*

$$\phi_i(v) = \sum_{\mathcal{S} \subseteq \mathcal{V}, V_i \notin \mathcal{S}} \frac{|\mathcal{S}|!(|\mathcal{V}| - |\mathcal{S}| - 1)!}{|\mathcal{V}|!} (v(\mathcal{S} \cup \{V_i\}) - v(\mathcal{S})). \tag{4}$$

The intuition is to consider the marginal utility that V_i brings in set $\mathcal{S}$, averaged over all possible sets. Notice that the average is weighted. In fact, the process can be seen not as averaging over sets, but averaging over all possible permutations

– assume that coalitions are generated by adding one individual at a time – so the weight of a set is the number of permutations that produce each coalition.

The definition of Shapley value is very attractive, as it can be shown theoretically that the resulting attribution satisfies natural fairness properties [23]. However, a direct application of Definition (4) in our setting is not possible. Not only it assumes an averaging over exponentially many sets, but also it is not possible to probe arbitrary sets $\mathcal{S}$ and obtain $v(\mathcal{S})$; for example, we do not have available the impact score of papers for every possible subset of authors! To address this difficulty we introduce the following ideas:

- First, we do not compute marginal gains by averaging over all possible coalitions, but only over coalitions for which we have available impact scores. The details are given in the next sections where we describe our iterative algorithm.
- Second, in order to average in a marginal contribution $v(\mathcal{S} \cup \{V_i\}) - v(\mathcal{S})$ we need to have available both values $v(\mathcal{S} \cup \{V_i\})$ and $v(\mathcal{S})$. However, in many cases we have available only one of the two. Simply ignoring those cases would lead to very sparse data. Therefore, we choose to take into account all cases for which $v(\mathcal{S} \cup \{V_i\})$ is available. If for those cases $v(\mathcal{S})$ is not available, we propose to approximate it by composing it from its subsets.

Now, we discuss how to define the values $v(\mathcal{S})$ and reasonable approximations of $v(\mathcal{S})$ when those values are not directly available. The first issue is that in some cases the same coalition has accomplished many different tasks, and each task has different impact factors; for example, a given set of authors may have many different publications. We then define the value of the coalition as the average of the impact scores of all the tasks into which the coalition has participated.

Definition 4 ***(Shared Impact Factor)*** *Consider a set of individuals $\mathcal{S}$ and let $\mathcal{T}_\mathcal{S}$ be the set of their common tasks. Let I_j be the impact factor of each common task $T_j \in \mathcal{T}_\mathcal{S}$. Then, the* shared impact factor *of $\mathcal{S}$ can be defined as:*

$$v(\mathcal{S}) = \frac{1}{|\mathcal{T}_\mathcal{S}|} \sum_{j=1}^{|\mathcal{T}_\mathcal{S}|} I_j. \tag{5}$$

The next issue that we need to address is that, during the execution of our iterative algorithm, we need to use the value of $v(\mathcal{S})$, for cases that we do not have any information about coalition $\mathcal{S}$ (e.g., there is no paper with that exact set of authors). Thus, we propose to approximate the value of $v(\mathcal{S})$ by taking into account subsets $\mathcal{S}' \subseteq \mathcal{S}$ for which we have information about the impact of their coalition. We suggest to average over all subsets of $\mathcal{S}$ for which we have information, and also include a term to capture the contribution of the individuals who do not form any coalition in subsets of $\mathcal{S}$.

Definition 5 ***(Approximated Shared Impact Factor)*** *Consider a set of individuals $\mathcal{S}$ and let $\mathcal{S}^c = \bigcup_i \mathcal{S}_i^c$ be the set of all subsets $\mathcal{S}_i^c$ of $\mathcal{S}$ where all individuals have at least one common task. Then, the* approximated shared impact factor *of the individuals in* $\mathcal{S}$ *is:*

$$v'(\mathcal{S}) = \frac{1}{|\mathcal{S}^c| + 1} \left(\sum_{i=1}^{|\mathcal{S}^c|} v(\mathcal{S}_i^c) + \bar{v}(\mathcal{S} \setminus \mathcal{S}_i^c) \right), \tag{6}$$

where $v(\mathcal{S}_i^c)$ *can be computed using Equation (5) and* $\bar{v}(\mathcal{S} \setminus \mathcal{S}_i^c)$ *is an estimation of the impact of individuals who do not form any coalition in subsets of* $\mathcal{S}$.

To complete the definition of function $v'(\cdot)$ we need to define function $\bar{v}(\cdot)$, which we use when individuals do not participate in any common coalition. We do this by using a recursive definition. We choose to compute $\bar{v}(\mathcal{S})$ using the influence factors ϕ_i of the individuals who form coalition $\mathcal{S}$. Assuming a monotonic behavior, that is, assuming that teams are at least as good as the best individual who composes the team, we define $\bar{v}(\cdot)$ using the *maximum* operator.

Definition 6 ***(Approximated Gain Function)*** *The* approximated gain function $\bar{v}(\mathcal{S})$ *is defined as follows:*

$$\bar{v}(\mathcal{S}) = \max_{V_i \in \mathcal{S}} \phi_i(v). \tag{7}$$

We note that Definition (6) is recursive since our goal is to compute the scores ϕ_i. Therefore, our definition suggests an iterative algorithm, described in more detail in the next section: the algorithm starts with initial estimates of the scores ϕ_i, and then it iteratively adjusts the estimate of those scores until convergence.

4.3 The iterative algorithm

We propose an iterative algorithm to compute the influence score ϕ_i of each individual $V_i \in \mathcal{V}$. At each iteration t, the value of the influence score is denoted as ϕ_i^t. At first, the influence scores of each individual are initialized to the average impact factor his/her tasks. Then, at each iteration, the Shapley value is computed using Definition (4). The key idea behind our algorithm is to compute the Shapley value averaging over a small set of coalitions, for which the impact factor is available. Whenever the algorithm needs to probe a coalition whose impact factor is not available, then the approximated shared impact factor is used (Definition 5).

During the main loop of the algorithm, for each available task $T_j \in \mathcal{T}$ and each individual V_i of T_j, we update the influence score $\phi_i^{t+1}(v')$ by adding the difference of the gain function of sets $\mathcal{T}_{V_j}$ and $\mathcal{T}_{V_j} \setminus V_i$. The same weighting factor is used, as in Equation (4), that is,

$$\phi_i^{t+1}(v') = \sum_{\mathcal{T}_{V_j} | V_i \in T_j} \frac{|\mathcal{T}_{V_j} \setminus V_i|!(|\mathcal{V}| - |\mathcal{T}_{V_j} \setminus V_i| - 1)!}{|\mathcal{V}|!} (v'(\mathcal{T}_{V_j}) - v'(\mathcal{T}_{V_j} \setminus V_i)). \tag{8}$$

This procedure continues until convergence, which is defined by the following criterion.

$$\frac{\sum_{i=1}^{|\mathcal{V}|} |\phi_i^t - \phi_i^{t-1}|}{\sum_{i=1}^{|\mathcal{V}|} \phi_i^{t-1}} \leq \epsilon \in (0, 1). \tag{9}$$

Algorithm 1 The `Shapley` Algorithm

Input: a set of individuals $\mathcal{V}$, a set of tasks $\mathcal{T}$, and the corresponding set of impact scores $\mathcal{I}$.
Output: the influence score ϕ_i of each individual $V_i \in \mathcal{V}$
// Initialization: $\forall T_i, i = 1, \ldots, m$ assigned to individual V_i:
for $j = 1 : |\mathcal{V}|$ **do**
  $\phi_i^0 = \sum_{i=j}^{m} I_j$
end for
while convergence **do**
  Initialize $\phi_i^{t+1}(v') = 0$
  for $T_j \in \mathcal{T}$ **do**
    for $V_i \in \mathcal{T}_{V_j}$ such that V_i is assigned with task T_j **do**
      $\phi_i^{t+1}(v') = \phi_i^{t+1}(v') + \frac{|\mathcal{T}_{V_j}|!(|\mathcal{V}|-|\mathcal{T}_{V_j}|-1)!}{|\mathcal{V}|!}(v'(\mathcal{T}_{V_j}) - v'(\mathcal{T}_{V_j} \setminus V_i))$
    end for
  end for
end while

The pseudocode of this method is given in Algorithm 1. For the remainder of this paper, this method will be denoted as `Shapley`.

4.4 Enforcing monotonicity of the gain function

In the theory of coalition games, the gain function is assumed to be monotone and non-negative. Here monotonicity means that if $\mathcal{S}_1 \subseteq \mathcal{S}_2$ then $v(\mathcal{S}_1) \leq v(\mathcal{S}_2)$. Those assumptions are also desirable in our setting. Thus, we enforce those conditions by first computing all payoffs—$v(\mathcal{S})$—that are available for each coalition $\mathcal{S}$. Then, we identify all pairs of sets $\mathcal{S}_1$ and $\mathcal{S}_2$, such that $\mathcal{S}_1 \subseteq \mathcal{S}_2$ and $v(\mathcal{S}_1) > v(\mathcal{S}_2)$. For each such pair we increase $v(\mathcal{S}_2)$ by setting $v(\mathcal{S}_1) = v(\mathcal{S}_2)$. This is repeated until all violations in these pairs are eliminated. In other words, we define the partial order of all coalition sets, and then we guarantee that while paths of set inclusion are followed, the payoffs defined by the gain function do not increase. Once the monotonicity property is satisfied, the computations performed by our algorithm ensure that the gain function is also non-negative.

5 Experiments

5.1 Setup

We evaluated the performance of the proposed method on two real data sets: one bibliographic data set and one movie data set.

ISI Web of Science. The first data set is part of the Thomson Reuters ISI Web of Science data. ISI covers mainly journal publications. We sampled data related to our institutions published within years 2003 and 2009. Our dataset contained data about 1212 authors.

We used two very common bibliometric indicators as the baseline:

– **H-Index**: a scientist's H-index is h, if h of his/her publications have at least h citations and the rest of his/her publications have at most h citations each.
– **G-Index**: a scientist's G-index is g, if g of his/her highly cited publications received, together, at least g^2 citations.

We investigated both types of impact scores discussed earlier (Section 3.1) for the author-publication scenario.

Internet Movie DataBase. The second data set is part of the Internet Movie Database (IMDB) data.[1] We sampled a total of 2 000 male actors and 4 560 movies. We restricted the movie genre type to "comedy" or "action", and did not include "TV series". For each actor we considered only the movies where his credit position was among the top 3. Each movie T_j was assigned with an impact score I_j defined to be $I_j = R_j A_j$, where R_j is the average rating received by movie T_j and A_j is the number of people who evaluated this movie.

To evaluate the performance of our methods we used the following measure:

Definition 7 ***(rank of individual)*** *Given a set of individuals* $\mathcal{V} = \{V_1, \ldots, V_n\}$ *and their influence scores* $\phi = \{\phi_1, \ldots, \phi_n\}$*, the rank of individual* V_i *is* r*, if*

$$|V_j|\phi_j \geq \phi_i, j \in 1, \ldots, n| = r. \tag{10}$$

In other words, the rank of an individual V_i measures the number of individuals who are at least as influential as V_i.

5.2 Experimental Results

ISI Web of Science. We compared the performance of `Naïve` and `Shapley` on the ISI Web of Science data sample described in Section 5.1. Each of the two methods computed an impact score per author. We then compared the author ranks of the proposed methods with the ranks obtained when computing the `H-index` and `G-index`. In Figure 2, we can see the performance of the proposed methods with respect to author ranks, compared to `H-index`. It is evident that there is a high correlation between `H-index` and `Naïve-CC` (the latter denotes the version of `Naïve` that uses the average citation score for defining function $f(\cdot)$). This correlation becomes much weaker in the case of `Naïve-PR`, where PageRank is used, implying that considering higher-order citations (as done by PageRank) can cause a significant change in the ranking of publications and, consequently, in the ranking of authors. In the case of `Shapley`, we see that as the `H-index` rank increases, the variance in the rank obtained by `Shapley` also increases. The results with respect to `G-index` are very similar but omitted due to space limitations.

Next, we compared the performance of `Naïve` and `Shapley` with respect to author ranks. Again, we investigated both cases of impact scores for the publications: citation count and PageRank. In Figure 3 we see a comparison of the

[1] `http://www.imdb.com/`

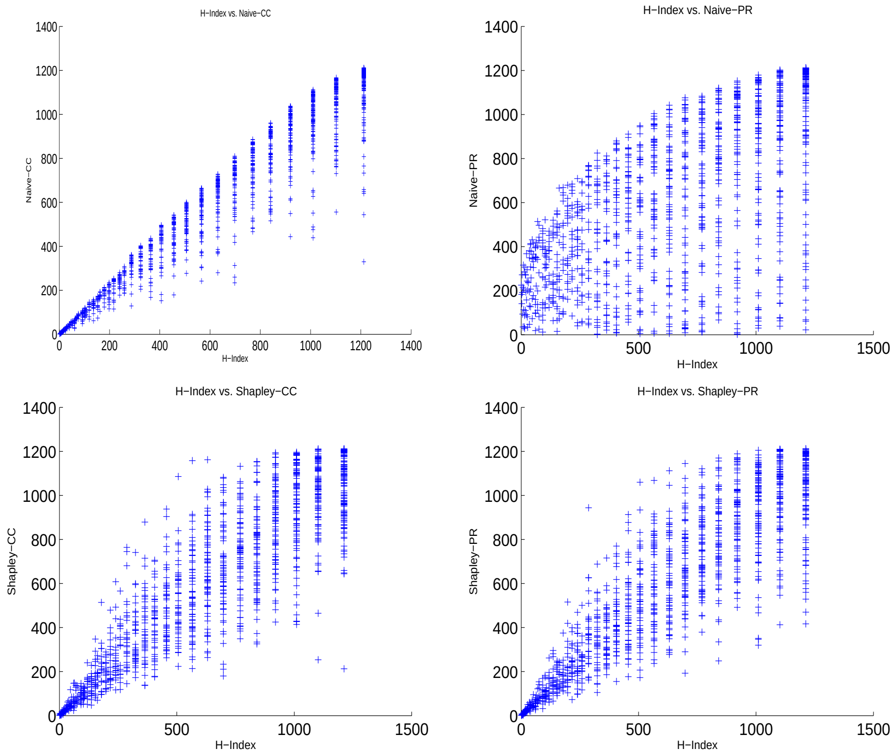


Fig. 2. Performance of `Naïve` and `Shapley` compared to H-index, with respect to author ranks, for the ISI Web of Science data. Note that the values in the x and y-axis of the figures correspond to ranks.

ranks produced by the four methods. When comparing `Naïve` and `Shapley` using citation count the two methods deviate as ranks increase. The same behavior is noticed when comparing `Shapley` using citation count vs. using PageRank. This effect is more intense for `Naïve`. In the latter case we note that `Naïve-CC` generally increases the ranking of the publications, compared to `Naïve-PR`. The results with respect to `G-index` are very similar.

We compared the ranks of the top-10 authors retrieved by `Shapley-CC`, Table 1(left) and `Shapley-PR`, Table 1(right), to the ranks produced by the other methods. To preserve the anonymity of the scientists we do not show their names in the table. An interesting observation is that the deviation in the ranks between the methods is much higher for `Shapley-PR` than for `Shapley-CC`.

Internet Movie Database. The performance of `Naïve` and `Shapley` was also evaluated on the IMDB data sample described in Section 5.1. In this case there are no baseline indicators, such as `H-index`, hence we only compared `Naïve` and `Shapley`. In Table 2(left) we see the names of the top-10 actors given by

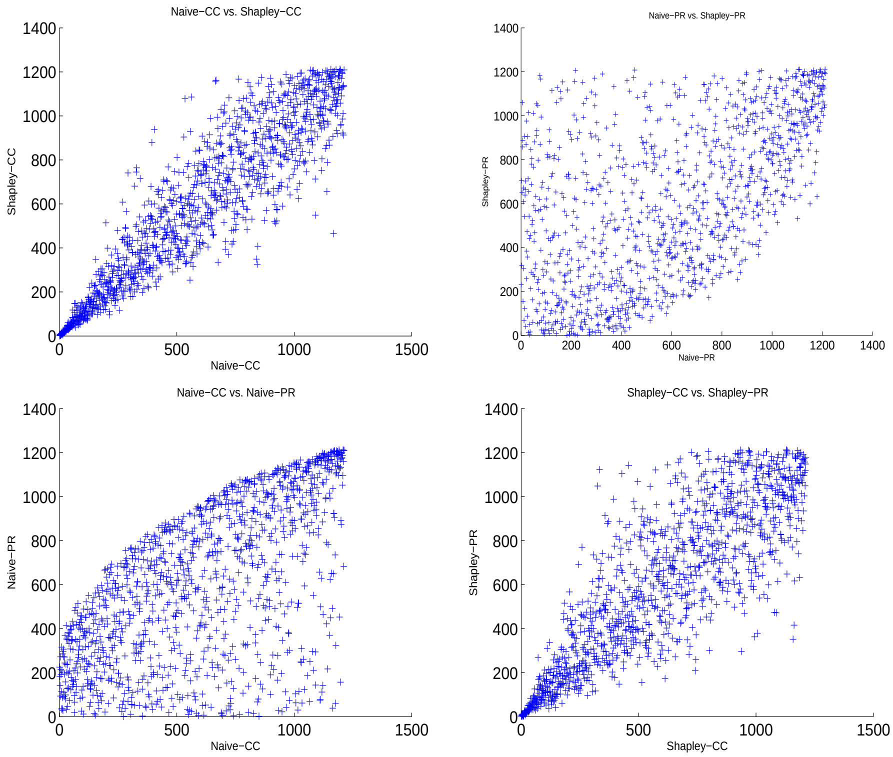


Fig. 3. Comparison of the `Naïve` and `Shapley` methods with respect to author ranks, for the ISI Web of Science data. Note that the values in the x and y-axis of the figures correspond to ranks.

`Shapley-CC`	`Naïve-CC`	H-index	G-index	`Shapley-PR`	`Naïve-PR`	H-index	G-index
1	1	1	2	1	183	2	1
2	3	3	4	2	225	1	2
3	6	8	11	3	35	8	9
4	5	4	4	4	215	8	11
5	4	6	5	5	192	5	7
6	12	25	22	6	272	4	4
7	10	9	11	7	94	46	25
8	2	2	1	8	141	23	14
9	8	8	9	9	208	11	13
10	15	15	14	10	114	11	7

Table 1. List of top-10 authors given by `Shapley-CC` (left) and by `Shapley-PR` (right). For `Shapley-CC` we used the citation count to assign impact scores to the publications, while for `Shapley-PR` we used the PageRank score. All values in the table correspond to ranks.

Shapley and the corresponding ranks given by Naïve. It appears that actors who participated in the casting of very highly rated movies are favored against those who also participated in low rated movies. It is interesting to see that Shapley managed to detect "big" names in *comedy* and *action*. In Table 2(right) we show the top-10 actors discovered by Naïve and their corresponding ranks by Shapley. Deviations between the two methods suggest that actors may manage to achieve a high mean movie rating (Naïve) but still there are several poorly rated movies with highly rated co-actors that cause a drop in their Shapley score.

Four more examples of deviating ranks are shown in Table 3. We can see that several famous actors show a high deviation in their ranking. For example, "Adam Sandler" is ranked 59^{th} by Shapley while Naïve ranks him 4^{th}. Based on the definition of the Shapley value, this implies that he may have highly ranked movies, but he also has movies where he is co-acting with other "strong" actors; these movies have a higher ranking than the ones where he is co-acting with "weaker" actors. Similar conclusions can be made for "Jim Carrey". On the one hand, he has a significant amount of highly rated movies, but also some low rated movies (which drops the score attributed to him by Naïve); on the other hand, in almost all of his movies he has been the main and "strongest" actor (which results in Shapley attributing him a high score).

Actor Name	Shapley	Naïve	Actor Name	Naïve	Shapley
Robert De Niro	1	3	Peter Sellers	1	14
Al Pacino	2	8	Jack Nicholson	2	11
Brad Pitt	3	15	Robert De Niro	3	1
Bruce Willis	4	7	Adam Sandler	4	59
Arnold Schwarzenegger	5	24	Daniel Day-Lewis	5	36
Will Smith	6	13	Chris Farley	6	20
Eddie Murphy	7	10	Bruce Willis	7	4
Robin Williams	8	9	Al Pacino	8	2
Morgan Freeman	9	17	Robin Williams	9	8
Ben Stiller	10	29	Eddie Murphy	10	7

Table 2. List of actors ranked as top-10 by Shapley (left) and by Naïve (right), and their corresponding ranks obtained by the other methods. All values in the table correspond to ranks.

6 Conclusions

We addressed the problem of attributing influence to a set of individuals who participate in a set of tasks. The method we proposed employs the game-theoretic concept of Shapley value. We showed that the proposed methodology can be applied in many real scenarios, for example, in the author-publication scenario for assigning influence scores to scientists who collaborate in published articles, or in the movie-actor scenario for evaluating actors who collaborate in different

Actor Name	Shapley	Naïve	# of Movies in IMDB	Average Movie Rating
Jim Carrey	11	79	34	5.2
Sylvester Stallone	12	41	46	6.4
Daniel Day-Lewis	36	5	27	7.1
Adam Sandler	59	4	39	5.4

Table 3. A list of 4 examples of actors with high deviation between `Shapley` and `Naïve`. All values in the table correspond to ranks.

movies. We evaluated our method on two real datasets, and showed that it highly differs with respect to ranking when compared to a naïve approach of equal division of influence.

In the author-publication scenario, we illustrated by an example (Figure 1), and also experimentally, that existing bibliometrics indicators, such as `H-index` and `G-index`, do not take into account the average strength of author coalitions and thus may favor authors with a few highly cited publications that resulted from a single (or very few) strong collaborators. Similar conclusions were drawn from the movie data, where very famous actors were given a lower ranking by `Shapley` than by `Naïve`. These actors were disfavored since their "fame" was caused only by a few specific movies and co-actors.

Directions for future work include the investigation of other domains such as user-blogs and social-media sites. Moreover, one could study how to use additional information regarding the involved tasks or individuals. For example, in the actor-movie scenario, several movie features are available as well as information about the actors. This additional information may or should affect the share allocation function. Finally, a challenging task is to further evaluate the quality of the rankings obtained, possibly by performing user studies.

Acknowledgements: This work was supported in part by the Academy of Finland ALGODAN Centre of Excellence. Additionally, this work was partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN20101037, "Social Media" [2].

References

1. K. V. Aadithya, B. Ravindran, T. P. Michalak, and N. R. Jennings. Efficient computation of the shapley value for centrality in networks. In *WINE*, 2010.
2. H. Aziz. *Algorithmic and complexity aspects of simple coalitional games*. PhD thesis, University of Warwick, 2009.
3. J. Deegan and E. W. Packel. A new index of power for simple n-person games. *International Journal of Game Theory*, 7(2), 1978.
4. X. Deng and Q. Fang. Algorithmic cooperative game theory. *Pareto Optimality, Game Theory and Equilibria*, 17(1):159185, 2008.

[2] www.cenitsocialmedia.es

5. P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.
6. L. Egghe. Theory and practise of the G-index. *Scientometrics*, 69, 2006.
7. E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.
8. N. Geller. COn the citation influence methodology of Pinski and Narin. *Information Processing & Management*, 14:93–95, 1978.
9. D. Gomez, E. Gonzalez-Aranguena, C. Manuel, G. Owen, M. del Pozo, and J. Tejada. Centrality and power in social networks: a game theoretic approach. *Mathematical Social Sciences*, 46(1):27–54, 2003.
10. J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.
11. C. H. Hubbell. An input-output approach to clique identification. *Sociometry*, 28(4):377–399, 1965.
12. L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.
13. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.
14. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
15. A. MasColell, M. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, 1995.
16. R. Narayanam and Y. Narahari. A shapley value-based approach to discover influential nodes in social networks. *IEEE Transactions on Automation Science and Engineering*, 8(1), 2011.
17. M. E. J. Newman. The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2007.
18. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1999.
19. G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12:297–312, 1976.
20. F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E*, 80(5):056103, Nov 2009.
21. C. Rehn, U. Kronman, and D. Wadskog. Bibliometric indicators – definitions and usage at Karolinska Institutet. Technical report, 2007.
22. L. S. Shapley. A value for n-person games. *Annals of Mathematical Studies*, 28:307–317, 1953.
23. E. Winter. The Shapley value. In *Handbook of Game Theory with Economic Applications*, volume 3. 2002.