IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 20, 2017, accepted August 28, 2017, date of publication September 22, 2017, date of current version October 12, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2750700

Node Aggregation for Enhancing PageRank

J. M. MAESTRE[1,2], (Member, IEEE), HIDEAKI ISHII[2], (Senior Member, IEEE), AND E. ALGABA[3]
[1]Department of Systems and Automation Engineering, Higher Technical School of Engineering, University of Seville, 41092 Seville, Spain
[2]Department of Computer Science, School of Computing, Tokyo Institute of Technology, Yokohama 226-8502, Japan
[3]Department of Applied Math II, Higher Technical School of Engineering, University of Seville, 41092 Seville, Spain

Corresponding author: J. M. Maestre (pepemaestre@us.es)

This work was supported in part by the FP7- ICT Project DYMASOS under Grant 611281, in part by the Spanish Ministry of Economy and Competitiveness COOPERA Project under Grant DPI2013-46912-C2-1-R, in part by the Spanish Ministry of Education under the *José Castillejo* Grant (CAS14/00277), and in part by the JSPS under Grant-in-Aid for Scientific Research Grant 15H04020 and Fellowship PE16048.

ABSTRACT In this paper, we study the problem of node aggregation under different perspectives for increasing PageRank of some nodes of interest. PageRank is one of the parameters used by the search engine Google to determine the relevance of a web page. We focus our attention to the problem of finding the best nodes in the network from an aggregation viewpoint, i.e., what are the best nodes to merge with for the given nodes. This problem is studied from global and local perspectives. Approximations are proposed to reduce the computation burden and to overcome the limitations resulting from the lack of centralized information. Several examples are presented to illustrate the different approaches that we propose.

INDEX TERMS Networks, game theory, graphs, centrality measures, model reduction.

I. INTRODUCTION

The relationships established among the entities that interact in a large-scale system can be often modeled as a graph. This is particularly visible in the case of networks, where nodes are connected through links that allow exchanging or transmitting different types of flows (e.g., water, electricity, and information). The characterization of the relevance of the nodes and links in a graph is a problem that has been studied under different perspectives and that is generally related to the calculation of *measures*, i.e., numerical values that provide a notion of the relevance according to a specific criterion. While some measures focus on the inherent structure of the network (e.g., the well-known degree and betweenness centrality measures), others aggregate additional information into the calculation.

In this article, we study specifically the centrality measure provided by the PageRank. It has been used by search engine Google, as one of the means to rank the relevance of the results they offer to their users.[1] Specific information about the PageRank value can be found in [4] and [5]. An analysis of this value and algorithms for its computation can be found in [19] and [20]. The rationale behind PageRank is that the relevance of a web page must be greater if there are significant pages pointing to it. Somehow it is as if each page votes the pages that it points to by means of the hyperlinks. Hence, the PageRank value depends only on the link structure of the graph that describes the web and its calculation can be seen as a variation to that of another popular centrality measure: eigenvector centrality.

In particular, we would like to gain insights into the problem of finding the most interesting nodes for given nodes to merge with from a PageRank viewpoint. While the aggregation of several nodes into a new one is beneficial in terms of incoming hyperlinks (as long as the new node is pointed to by all the links that were originally pointing to the merged nodes), it is clear that not all the nodes contribute equally to the fusion. Moreover, some fusions may be super-additive in the sense that the PageRank of the new node can be greater than the sum of the PageRanks of the merged nodes.

This viewpoint can be of interest in several applications where the fusion of nodes is an option to gain relevance, e.g., web pages, journals, conferences and companies. For example, PageRank was used together with other measures to evaluate company's value taking into account the effect of network structures in [32], where data of one million Japanese companies were studied. In [9], the world trade matrix is studied to rank countries by means of PageRank and CheiRank, which is a variation that deals with the transposed link matrix. Another interesting work is [21], where economic influence and contagion propagation over the world economy are

[1] Given that the PageRank was introduced originally in the context of a network of web pages, we will be using the term node to refer to web pages as well. Note however, that all the claims made are applicable to all types of networks in general.

analyzed by means of a Google matrix of economic activities in the period 1995-2009. Other interesting applications can be found in [14], where the social network Twitter is examined by using this approach to detect the most influencing users, and in [22], where PageRank is used as a tool in citation analysis. Finally, it is worthy to mention [2], where the effect of newly created links on PageRank is analyzed to find out how much a web page can control its PageRank.

In order to analyze this problem, we propose two different approaches, from the global and local perspectives. In the global one, we assume that full information about the network structure is available. In order to gain insights into the relevance of the web pages, we propose different cooperative games in which the value of each coalition is characterized as the PageRank received by the new node resulting from aggregation. The Shapley value [31] of the games is related to the PageRank that a certain page should expect from joining randomly to a coalition of web pages. In this way, new centrality measures are obtained. Notice that the Shapley value has been proposed as a mathematical tool to measure node centrality in a network in many works of the literature. For example, in [28], the information diffusion process in a network is modeled as a cooperative game and its Shapley value is used to provide a measure of the node influence. Another related work is [16], where the Shapley value of the difference of a game and its graph-restricted version is proposed as a centrality measure in a network. Likewise, it has been also applied to characterize the relative relevance of links by means of the position value in [3]. Although other solution concepts also appear in the literature (e.g., in [18] a measure degree of centrality in a social network based on an extension of the Banzhaf power index is presented), the Shapley value is by far the most used game theoretical payoff rule used in this context, probably due to its properties and its straightforward interpretation in terms of average marginal contribution. Applications of the Shapley value as a measure to gain an insight into the relevance of the players in this context include social networks [18], [28], wine ranking [15], scientific influence attribution [29], shareholder influence attribution in companies [25], and control systems [24], [27], among many others.

In addition, we provide a mechanism to compute a numerical approximation of these values to overcome the combinatorial explosion issues that arise in this type of problems. It must be noticed that, while there are numerous studies on the Shapley value from a theoretical perspective, from a practical point of view the complexity of its computation is exponential, in general, leading to an NP-complete problem [8]. In the literature, the Shapley value has been computed for special classes of games as weighted voting games restricted by a tree [13] or weighted multiple majority games [1] where it has been calculated in polynomial time by algorithms based on generating functions, or particular cases of operations research games called extended tree games [17], among others. An approximation of the Shapley value for voting games by a randomized polynomial method is presented in [10]. Also, the Shapley value for some centrality-related cooperative games on networks has been solved analytically by exact algorithms in polynomial time [26]. In addition, a polynomial method based on sampling theory which can be applied to approximate the Shapley value for cooperative games is given in [7], since it is not possible to compute the marginal contributions from every player when the number of players is large. In fact, the authors of this work applied sampling as a method to describe the average of the marginal contributions for each player on the set of all possible permutations of the set of players from a representative sample of the set of all permutations. This method is only efficient if the worth of any coalition can be calculated in polynomial time.

The second approach we propose is based solely on local information. In particular, it is assumed that the nodes only have information about their neighbors. Hence, it is not possible to calculate directly the PageRank of any merger, although it can be approximated if certain simplifying assumptions hold, e.g., if the PageRank of the neighbors is not modified after the merger is formed or if the rest of the network can be approximated so that the PageRank of the nodes under study is preserved. Four alternatives are considered depending on different assumptions regarding the way the approximated PageRank is computed, which are assessed by simulation. As it will be seen in the corresponding section, our simulation results show that some of these methods are very accurate for determining super-additivity.

Finally, this work enhances substantially and contributes with more relevant results to the earlier version [23] in several ways: it includes the proofs of the theorems contained in that paper and it presents additional theoretical contributions to calculate analytically equivalent networks in a PageRank sense and estimators for the PageRank of a coalition of nodes when there is no global information available. As it will be shown, the methods proposed allow predicting with high accuracy whether a fusion of nodes will be super-additive from a PageRank viewpoint. Moreover, the limitations derived from the combinatorial explosion problem in large networks, which limited the applicability of our previous results, are solved in this article by means of a randomized method proposed in [6], which is adapted here to the games considered and assessed by means of a large example.

The outline of the paper is as follows. In Section II, grounds on the calculation of the PageRank and the problem setting are given. Section III deals with the centralized approach and introduces cooperative games based on the PageRank whose Shapley values are proposed as centrality measures. Section IV deals with the local information approach. Each of the proposed approaches is illustrated with different examples in the corresponding sections. Finally, Section V concludes the paper with final remarks and comments about future research.

II. PROBLEM SETTING

Let $\mathcal{N}$ be a set of networked web pages. The network can be represented as a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{E}$ is the set of edges representing the hyperlinks among the web pages. In case that a web page i has a link pointing to web page j then $(i, j) \in \mathcal{E}$.

The PageRank is itself a centrality measure related to eigenvector centrality. It provides information regarding the relevance of a certain web page based on the dominant eigenvector of a modified adjacency matrix. In particular, the PageRank of the web page i is given by the number $x_i^* \in [0, 1]$, with $\sum_{i\in\mathcal{N}} x_i^* = 1$. It relies on the assumption that a web page having links from important web pages is also important. Hence, the value of a web page i is the sum of the contributions from all web pages that have links to it, i.e.,

$$x_i^* = \sum_{j\in\mathcal{N}_i} \frac{x_j^*}{n_j},$$

where $\mathcal{N}_i := \{j : (j, i) \in \mathcal{E}\}$ is the set of web pages pointing to the web page i and n_j is the number of outgoing links of web page j.

The calculation of the PageRank can also be stated in matrix form. Let $\mathbf{x}^* := [x_i^*]_{i\in\mathcal{N}}$ be a vector of the PageRanks of all the web pages. The calculation of the PageRank can be performed by solving

$$\mathbf{x}^* = \mathbf{A}\mathbf{x}^*, \quad \mathbf{x}^* \in [0, 1]^{|\mathcal{N}|}, \quad \sum_{i\in\mathcal{N}} x_i^* = 1, \tag{1}$$

where $\mathbf{A}$ is the so-called hyperlink matrix, a variation of the adjacency matrix, given by

$$a_{ij} = \begin{cases} \dfrac{1}{n_j} & j \in \mathcal{N}_i, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix $\mathbf{A}$ is column stochastic, i.e., $\sum_{i\in\mathcal{N}} a_{ij} = 1$ for all $j \in \mathcal{N}$. As it is shown in [20], in order to have a graph of the network that satisfies this property it is necessary to remove the so-called dangling nodes, i.e., the pages with no links to other pages. This can be done by assuming that they have backward links to (some of) the pages pointing to them, which corresponds to the use of the back button in the web browser.

The PageRank vector $\mathbf{x}^*$ is the nonnegative unit eigenvector that corresponds to the eigenvalue 1 of $\mathbf{A}$. However, in order to guarantee its uniqueness, (1) is slightly modified and the PageRank vector $\mathbf{x}^*$ is defined as the solution of

$$\mathbf{x}^* = \mathbf{M}\mathbf{x}^*, \quad \mathbf{x}^* \in [0, 1]^{|\mathcal{N}|}, \quad \sum_{i\in\mathcal{N}} x_i^* = 1 \tag{2}$$

with $\mathbf{M} := (1 - m)\mathbf{A} + m\mathbf{J}$, $m \in (0, 1)$. Matrix $\mathbf{M}$ is a convex combination of $\mathbf{A}$ and $\mathbf{J}$, with $\mathbf{J} = \frac{1}{|\mathcal{N}|}\mathbf{1}^{|\mathcal{N}|\times|\mathcal{N}|}$, which is a *jump* matrix with all its elements equal to $1/|\mathcal{N}|$. The jump matrix models the random jumps between pages that users perform while surfing the web. Notice that matrix $\mathbf{M}$ is a column stochastic matrix with all positive entries. The parameter m is usually set to 0.15 by Google [5], [20].

A. NODE AGGREGATION

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_s\}$ be a set composed of the web pages $s_1, s_2, \ldots s_s$ that are aggregated into a single web page that preserves all the original links of the merged web pages.[2] For convenience we also define $\mathcal{R} = \{r_1, r_2, \ldots, r_r\}$ as the complementary set $\mathcal{N} \setminus \mathcal{S}$. In this paper, we are interested in the value of the resulting web page in the new configuration of the network, which will be denoted as the directed graph $\mathcal{G}' = (\mathcal{N}', \mathcal{E}')$.

The aggregation of web pages can be analytically achieved by means of three auxiliary transformation matrices, namely $\mathbf{P}_\mathcal{S}$, $\mathbf{T}_\mathcal{S}$, and $\mathbf{D}_\mathcal{S}$.[3] Matrix $\mathbf{P}_\mathcal{S}$ is a permutation matrix whose purpose is to rearrange the link matrix so that the web pages inside $\mathcal{S}$ appear next to each other in the first columns and rows. In particular, the permutation is given by

$$\mathbf{P}_\mathcal{S} = \left[\mathbf{e}_{s_1}\mathbf{e}_{s_2}\ldots\mathbf{e}_{s_s}\mathbf{e}_{r_1}\mathbf{e}_{r_2}\ldots\mathbf{e}_{r_r}\right],$$

where $\mathbf{e}_i$ denotes the column vector of length $|\mathcal{N}|$ with 1 in the i-th entry and 0 in all other entries. $\mathbf{T}_\mathcal{S}$ is a transformation matrix whose goal is to aggregate into a single web page the members in $\mathcal{S}$. It is given by

$$\mathbf{T}_\mathcal{S} = \begin{bmatrix} \mathbf{1}^{|\mathcal{S}|\times 1} & \mathbf{0}^{|\mathcal{S}|\times|\mathcal{R}|} \\ \mathbf{0}^{|\mathcal{R}|\times 1} & \mathbf{I}^{|\mathcal{R}|\times|\mathcal{R}|} \end{bmatrix}.$$

Finally, $\mathbf{D}_\mathcal{S}$ guarantees that the new link matrix $\mathbf{A}'$ is also a column stochastic matrix. To this end, it normalizes the column in which the web pages are aggregated. Hence,

$$\mathbf{D}_\mathcal{S} = \begin{bmatrix} \dfrac{1}{|\mathcal{S}|} & \mathbf{0}^{1\times|\mathcal{R}|} \\ \mathbf{0}^{|\mathcal{R}|\times 1} & \mathbf{I}^{|\mathcal{R}|\times|\mathcal{R}|} \end{bmatrix}.$$

The link matrix of $\mathcal{G}'$ can be derived from the original network $\mathcal{G}$ as follows:

$$\mathbf{A}' = \mathbf{T}_\mathcal{S}^\mathrm{T}\mathbf{P}_\mathcal{S}^\mathrm{T}\mathbf{A}\mathbf{P}_\mathcal{S}\mathbf{T}_\mathcal{S}\mathbf{D}_\mathcal{S}.$$

Here, the rows and columns of $\mathbf{A}$ are conveniently rearranged by $\mathbf{P}_\mathcal{S}$, the rows and columns of the web pages in $\mathcal{S}$ are merged by means of $\mathbf{T}_\mathcal{S}$, and $\mathbf{D}_\mathcal{S}$ adjusts the final result to have the column stochastic matrix $\mathbf{A}'$. The corresponding PageRank can be calculated as in (2), with $\mathbf{M}' := (1 - m)\mathbf{A}' + m\mathbf{J}'$ and $\mathbf{J}' = \frac{1}{|\mathcal{N}'|}\mathbf{1}^{|\mathcal{N}'|\times|\mathcal{N}'|}$.

The PageRank of the web page that results from merging the web pages in $\mathcal{S}$ is defined as the PageRank of the aggregated web page in $\mathcal{G}'$, that is, the first component of the new PageRank vector $\mathbf{x}'^*$. This can be calculated as the limit of the sequence generated by the power method as

$$\mathbf{x}'[k+1] = \mathbf{M}'\mathbf{x}'[k] = (1-m)\mathbf{A}'\mathbf{x}'[k] + \frac{m}{|\mathcal{N}'|}\mathbf{1}^{|\mathcal{N}'|\times 1} \tag{3}$$

when $k \to \infty$ and $\mathbf{x}'[0] = \frac{1}{|\mathcal{N}'|}\mathbf{1}^{|\mathcal{N}'|\times\mathbf{1}}$.

Remark 1: It is possible to consider alternatives in the way the matrix $\mathbf{M}'$ *is built. The only strong requirement is that it must be a column stochastic matrix. One key question is whether to assume if the random surfer still jumps between the pages in the merged network* $\mathcal{G}'$ *with the same probability, i.e., if* $\mathbf{J}' = \frac{1}{|\mathcal{N}'|}\mathbf{1}^{|\mathcal{N}'|\times|\mathcal{N}'|}$ *should hold for this case.*

[2] The links between these web pages are transformed into self-pointing links of the resulting web page.

[3] A lower number of auxiliary matrices could be used but the operations performed to calculate the link matrix are clearer in this way.

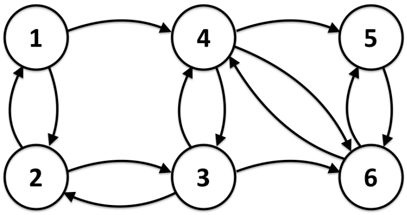

FIGURE 1. Example web with six pages.

This corresponds to the extreme case in which the new node receives the same amount of random jumps as any other one even when it may emerge after the fusion of many nodes.

An alternative case could be to consider that the random surfer jumps with an aggregated probability to the nodes inside $\mathcal{S}$, i.e.,

$$\mathbf{J}' := \frac{1}{|\mathcal{N}|}\begin{bmatrix} |\mathcal{S}|\mathbf{1}^{1\times|\mathcal{N}'|} \\ \mathbf{1}^{|\mathcal{R}|\times|\mathcal{N}'|} \end{bmatrix}.$$

In this case, it is assigned to the merged node an additional probability of random jumps because it comes from the aggregation of several independent nodes. Additional alternatives based on a different rationale are possible and even a convex combination of them could be used for the calculation. However, it should be noticed that strictly speaking the use of any of these alternatives does not lead to the PageRank but to a modified value for the merger.

B. AN ACADEMIC EXAMPLE

In this subsection, we introduce an example taken from [20], which will help us illustrate the main problem studied in this paper.

The network is shown in Fig. 1. As can be seen, it can be modeled by means of a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{E} = \{(1, 2), (2, 1), (1, 4), (2, 3), (3, 2), (4, 3), (3, 4), (4, 6), (6, 4), (4, 5), (5, 6), (6, 5), (3, 6)\}$. The matrix $\mathbf{A}$ of this network is

$$\mathbf{A} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & 1 & 0 \end{bmatrix}$$

and the corresponding matrix $\mathbf{M}$ becomes

$$\mathbf{M} = \begin{bmatrix} 0.025 & 0.450 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.450 & 0.025 & 0.308 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.450 & 0.025 & 0.308 & 0.025 & 0.025 \\ 0.450 & 0.025 & 0.308 & 0.025 & 0.025 & 0.450 \\ 0.025 & 0.025 & 0.025 & 0.308 & 0.025 & 0.450 \\ 0.025 & 0.025 & 0.308 & 0.308 & 0.875 & 0.025 \end{bmatrix}.$$

TABLE 1. PageRank of the example network.

Page	PR
1	0.061
2	0.086
3	0.122
4	0.214
5	0.214
6	0.302

Table 1 provides the values of the PageRank (PR) of this network. As can be seen, web pages 4, 5 and 6 have the highest PageRank values due to their larger numbers of incoming links. The value of web page 6 is greater because its incoming links come from web pages with larger values, namely pages 3, 4 and 5. It is very interesting that pages 4 and 5 share the same value of PageRank, specially since page 5 is a dangling node in [20].[4]

Finally, the aggregation of web pages can be studied by using (3). For example, the PageRank of the merger of web pages 1 and 2 is 0.11, which is greater than the PageRank of any of these web pages, as shown in Table 1. Nevertheless, the PageRank of the merger is lower than the sum of the PageRanks corresponding to the merged nodes. Next, let us examine what happens when nodes 1 and 4 are aggregated. In this case, the PageRank of the merger becomes 0.281, which is greater than the sum of the PageRanks of the nodes aggregated. Hence, the aggregation is super-additive in terms of PageRank.

In the next sections, we explore the generation of super-additive PageRank from different perspectives. In particular, we study which nodes are more likely to generate super-additive PageRank and also different estimators for the PageRank value of the merger when full information is not available.

III. GLOBAL APPROACH: COOPERATIVE GAMES BASED ON THE PAGERANK

Game theory deals with situations in which there are several coupled interacting entities. The outcome of a game depends on the combinations of the decisions taken by its players. Depending on the way the decisions are taken, it is possible to classify the games as cooperative or noncooperative. In the former case, players may attain binding agreements; in the latter case, decisions are taken without guarantees regarding the behavior of the rest of the players. In this work, we focus on cooperative games, which are defined by two basic elements, namely a set of players and a characteristic function. More specifically, we propose to use the PageRank that an aggregation of web pages obtains as the value that corresponds to that coalition. The Shapley value [31] of these aggregation PageRank games is proposed as a centrality measure.

In this section, we first provide some grounds about cooperative game theory. Next we define how the PageRank can

[4] Web page 5 is a dangling node and the link going from 5 to 6 was introduced to guarantee that the link matrix is stochastic.

be computed. Finally, cooperative games based on the PageRank are defined.

A. COOPERATIVE GAME THEORY

Cooperative game theory studies situations in which players can negotiate and attain binding agreements. Which coalitions of players should be expected and how to divide the benefits/costs derived from cooperation are major concerns in this branch of game theory. Formally, a cooperative game with transferable utility is a pair $(\mathcal{N}, v)$, where $\mathcal{N}$ is the set of players and v is a function that assigns a value to each possible coalition $\mathcal{S} \subseteq \mathcal{N}$, with $v(\emptyset) = 0$. A payoff rule is a vector that assigns a certain amount to each player according to its contribution to the game. In this work we use the Shapley value, which assigns to each player $i \in \mathcal{N}$ the value

$$\phi_i(\mathcal{N}, v) = \sum_{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}} \frac{|\mathcal{S}|!(|\mathcal{N}|-|\mathcal{S}|-1)!}{|\mathcal{N}|!}[v(\mathcal{S}\cup\{i\})-v(\mathcal{S})]. \tag{4}$$

This can be interpreted as the value that each player gets according to a weighted average of the contributions he makes to the different coalitions. The Shapley value on the class of transferable utility games satisfies some interesting properties such as *Linearity*, *Efficiency*, *Dummy player* and *Symmetry* [31].

B. THE PAGERANK AGGREGATION GAME

The definition of a cooperative game requires a characteristic function that assigns a value to each of the possible coalitions of players that can be formed. In the PageRank aggregation game, the players are the web pages contained in $\mathcal{G}$ and the term *coalition* is understood in the sense of aggregation. A coalition of web pages $\mathcal{S} = \{s_1, s_2, \ldots, s_s\}$ is the aggregation of these web pages into a single web page that preserves all the original links. The value of coalition $\mathcal{S}$ is defined as the PageRank of the aggregated web page in $\mathcal{G}'$, that is, the first component of the new PageRank vector $\mathbf{x}'^{*}$. Hence

$$v(\mathcal{S}) = [1\ 0 \ldots 0]\mathbf{x}'^{*} = \mathbf{e}_1^{\mathrm{T}}\mathbf{x}'^{*}. \tag{5}$$

Note that (5) depends only on the structure of the network of web pages $\mathcal{G}$ and hence the Shapley value of this game provides us with a centrality measure of the web pages.

C. THE PAGERANK DIFFERENCE AGGREGATION GAME

An allocation vector of the previous game provides information regarding the players that are expected to provide more PageRank when they are aggregated into a coalition. While this information is valuable, it does not consider the net effect of the aggregation. It is interesting to know whether the PageRank of the merger is greater, lower or equal to the sum of the individual PageRanks of the players in the coalition. To this end, we define the PageRank difference aggregation game over the same set of players but with the following characteristic function:

$$v_{\mathrm{dif}}(\mathcal{S}) = v(\mathcal{S}) - \sum_{j\in\mathcal{S}} v(j),$$

where for simplicity we have denoted $v(\{i\})$ by $v(i)$. The characteristic function $v_{\mathrm{dif}}(\mathcal{S})$ measures the gain or loss of PageRank derived from the aggregation. If $v_{\mathrm{dif}}(\mathcal{S}) > 0$ for a certain coalition $\mathcal{S}$ then the aggregation is rational and the players have a strong incentive to perform the coalition. The Shapley value of the difference game provides information about the best players to be integrated from this perspective.

Remark 2: The fact that the PageRank obtained by a coalition is lower than the sum of the PageRanks of its members should not be taken as if it is not rational to carry out the integration. A certain loss of PageRank can be acceptable in order to gain relevance in terms of centrality.

Finally, the following proposition simplifies the calculation of the Shapley value of the difference aggregation game.

Proposition 1: The Shapley value of the difference aggregation game can be calculated as

$$\phi_i(\mathcal{N}, v_{\mathrm{dif}}) = \phi_i(\mathcal{N}, v) - v(i).$$

Proof: The Shapley value is linear by construction. Hence,

$$\begin{aligned}\phi_i(\mathcal{N}, v_{\mathrm{dif}}) &= \phi_i(\mathcal{N}, v(\mathcal{S}) - \sum_{j\in\mathcal{S}} v(j)) \\ &= \phi_i(\mathcal{N}, v(\mathcal{S})) - \phi_i(\mathcal{N}, \sum_{j\in\mathcal{S}} v(j)).\end{aligned}$$

The Shapley value of the game $\sum_{j\in\mathcal{S}} v(j)$ is simply $\phi_i(\mathcal{N}, \sum_{j\in\mathcal{S}} v(j)) = v(i)$ because the marginal contribution in this game of adding any player to a random coalition is constant and equal to its own value $v(i)$. Q.E.D.

Remark 3: Notice that the computation of $\phi_i(\mathcal{N}, v_{\mathrm{dif}})$ can be performed immediately after $\phi_i(\mathcal{N}, v)$ is computed as it only requires to subtract the PageRank of the corresponding player to this amount.

D. AN ACADEMIC EXAMPLE (CONT.)

In this subsection, we work again with the example in Section II-B, which will help us show the differences between the PageRank of the web pages in the network and the Shapley value of the games proposed.

From the viewpoint of game theory, we model the web given by the directed graph $\mathcal{G}$ as a cooperative game $(\mathcal{N}, v)$ where $\mathcal{N}$ is the set of web pages and v is defined as the PageRank of the aggregated coalition. In Table 2, the value of the characteristic function v is shown for each coalition of web pages.

Table 3 provides the values of the PageRank (PR) and the Shapley values of the PageRank aggregation games of this network. As can be seen, the Shapley value of the aggregation game is fairly close to the PageRank, but it does recognize a difference on the relevance on nodes 4 and 5. In particular,

TABLE 2. Characteristic function for the example network.

$\mathcal{S}$	$\sum_i v(i)$	$v(\mathcal{S})$	$v_{\text{dif}}(\mathcal{S})$
{1}	0.061	0.061	0
{2}	0.086	0.086	0
{3}	0.122	0.122	0
{4}	0.214	0.214	0
{5}	0.214	0.214	0
{6}	0.302	0.302	0
{1,2}	0.147	0.110	−0.038
{1,3}	0.184	0.165	−0.019
{1,4}	0.276	0.281	0.006
{1,5}	0.276	0.249	−0.027
{1,6}	0.364	0.331	−0.033
{2,3}	0.208	0.179	−0.029
{2,4}	0.300	0.295	−0.004
{2,5}	0.300	0.281	−0.019
{2,6}	0.388	0.365	−0.023
{3,4}	0.336	0.324	−0.012
{3,5}	0.336	0.316	−0.020
{3,6}	0.424	0.410	−0.015
{4,5}	0.428	0.405	−0.023
{4,6}	0.517	0.473	−0.044
{5,6}	0.517	0.460	−0.056
{1,2,3}	0.269	0.198	−0.071
{1,2,4}	0.361	0.370	0.008
{1,2,5}	0.361	0.373	0.011
{1,2,6}	0.449	0.460	0.010
{1,3,4}	0.398	0.386	−0.012
{1,3,5}	0.398	0.393	−0.005
{1,3,6}	0.486	0.479	−0.007
{1,4,5}	0.490	0.465	−0.025
{1,4,6}	0.578	0.524	−0.054
{1,5,6}	0.578	0.434	−0.144
{2,3,4}	0.422	0.408	−0.014
{2,3,5}	0.422	0.435	0.013
{2,3,6}	0.510	0.528	0.018
{2,4,5}	0.514	0.490	−0.024
{2,4,6}	0.602	0.552	−0.050
{2,5,6}	0.602	0.479	−0.123
{3,4,5}	0.551	0.528	−0.022
{3,4,6}	0.639	0.595	−0.043
{3,5,6}	0.639	0.515	−0.123
{4,5,6}	0.731	0.605	−0.125
{1,2,3,4}	0.483	0.458	−0.026
{1,2,3,5}	0.483	0.511	0.028
{1,2,3,6}	0.572	0.608	0.036
{1,2,4,5}	0.576	0.596	0.021
{1,2,4,6}	0.664	0.653	−0.010
{1,2,5,6}	0.664	0.570	−0.093
{1,3,4,5}	0.612	0.611	−0.001
{1,3,4,6}	0.700	0.672	−0.028
{1,3,5,6}	0.700	0.588	−0.112
{1,4,5,6}	0.792	0.637	−0.155
{2,3,4,5}	0.636	0.653	0.017
{2,3,4,6}	0.724	0.717	−0.007
{2,3,5,6}	0.724	0.644	−0.080
{2,4,5,6}	0.816	0.682	−0.134
{3,4,5,6}	0.853	0.710	−0.143
{1,2,3,4,5}	0.698	0.751	0.053
{1,2,3,4,6}	0.786	0.810	0.024
{1,2,3,5,6}	0.786	0.737	−0.049
{1,2,4,5,6}	0.878	0.791	−0.087
{1,3,4,5,6}	0.914	0.801	−0.113
{2,3,4,5,6}	0.939	0.859	−0.080
{1,2,3,4,5,6}	1	1	0

node 4 is more important because of its greater impact when it is aggregated into a random coalition of web pages. For example, in all the coalitions with 5 web pages, the one with the lowest value is that in which web page 4 is not included. As a direct consequence, the relative relevance of page 6 is also reduced. The Shapley value of the difference aggregation game shows that nodes 1 to 4 are more likely to produce mergers with a higher PageRank value. Indeed, a closer look at Table 2 shows that most coalitions that produce a positive difference contain at least some of these players. Likewise, notice that the Shapley values of this game correspond to that of the aggregation game minus the PageRank of each player.[5]

TABLE 3. Comparison of the values.

Page	PR	$\phi_i(\mathcal{N}, v)$	$\phi_i(\mathcal{N}, v_{\text{dif}})$
1	0.061	0.079	0.017
2	0.086	0.116	0.030
3	0.122	0.144	0.021
4	0.214	0.218	0.004
5	0.214	0.184	−0.030
6	0.302	0.259	−0.043

This simple academic example shows the potential of the Shapley value of the aggregation PageRank games as an alternative centrality measure for the nodes. In addition, this perspective is also useful when evaluating the potential fusion of nodes inside the network. For example, several web owners that plan to integrate their web pages would be very interested in this type of information, specially since revenues are related to the number of visitors and hence PageRank. However, a problem arises at this point due to the combinatorial explosion and the need of centralized information. While there are efficient methods for the distributed computation of the PageRank [20], the computational and informational requirements to calculate this value are very demanding.

E. THE SHAPLEY VALUE OF LARGE NETWORKS

As previously pointed out, one of the most important solution concepts in cooperative games is the Shapley value [30]. The Shapley value allocates the worth of the grand coalition when all agents in the set of players decide to cooperate. However, a well-known problem of the Shapley value is its computation. The explosion on the number of coalitions that have to be computed hinders its calculation for large-scale games. For a set of players $\mathcal{N}$, $2^{|\mathcal{N}|}$ different coalitions need to be evaluated. For example, a network with only 30 nodes requires 2^{30} coalitions to be evaluated, i.e., a billion of value computations for a relatively small size network. To overcome this issue, in this work we employ the numerical approximation of the Shapley value proposed in [6], which is based on a randomized method. Next, we describe briefly this method. It is important to emphasize that for the games defined previously the computation of the value of each coalition can be realized in polynomial time.

In the sampling method given in [6], an alternative definition of the Shapley value is used. Indeed, the Shapley value can be expressed in terms of all possible orders of the players $\mathcal{N}$, assuming that all different orders have the same

[5] The small differences are due to the precision employed for the presentation of the results.

probability, in the following way:

$$\phi_i(\mathcal{N}, v) = \frac{1}{|\mathcal{N}|!} \sum_{\pi \in \Pi(\mathcal{N})} m_i^\pi(\mathcal{N}, v), \text{ for all } i \in \mathcal{N},$$

where $\Pi(\mathcal{N})$ is the collection of all permutations $\pi : \mathcal{N} \to \mathcal{N}$ on $\mathcal{N}$, and for every permutation $\pi \in \Pi(\mathcal{N})$,

$$m_i^\pi(\mathcal{N}, v) = v(\{j \in \mathcal{N} \mid \pi(j) \leq \pi(i)\}) - v(\{j \in \mathcal{N} \mid \pi(j) < \pi(i)\}), \quad (6)$$

is the marginal contribution of player i to the players that are ranked before him in the order π. Therefore, the Shapley value assigns to every game the average over all marginal vectors associated to all permutations of the player set $\mathcal{N}$.

Next, some terminology corresponding to the sampling process to approximate the Shapley value is given as follows:

(1) The population set $\mathcal{P} = \Pi(\mathcal{N})$ from which the sample is taken is represented by the set of all possible orders of the set of players $\mathcal{N}$. The sample $\mathcal{Q}$ is an element of

$$\underbrace{\mathcal{P} \times \mathcal{P} \times \cdots \times \mathcal{P}}_{q \text{ times}},$$

i.e., the sample size is q and it is obtained with replacement.

(2) The parameter vector $\phi = (\phi_1, \ldots, \phi_n)$ under consideration consists of the Shapley value for each $i \in \mathcal{N}$.

(3) The characteristics observed for each sampling unit $\pi \in \Pi(\mathcal{N})$ correspond to the marginal contributions of the players in the order π, i.e.,

$$x(\pi) = (x_1(\pi), \ldots, x_n(\pi)) \quad \text{with } x_i(\pi) = m_i^\pi(\mathcal{N}, v).$$

(4) The estimate of the Shapley value, $\widetilde{\phi}_i(\mathcal{N}, v)$, will be given by the average of the marginal contributions over the sample Q, i.e.,

$$\widetilde{\phi}_i(\mathcal{N}, v) = \frac{1}{q} \sum_{\pi \in Q} m_i^\pi(\mathcal{N}, v), \quad \text{for all } i \in \mathcal{N}.$$

(5) Any order $\pi \in \Pi(\mathcal{N})$ will be taken with equal probability to determine the sample $\mathcal{Q}$ in the process of selection.

The sampling process described above gives an approximation of the Shapley value with desirable properties. For instance, it is possible to calculate the theoretical error in a probabilistic way. In particular,

$$\widetilde{\phi}_i(\mathcal{N}, v) \sim N(\phi_i(\mathcal{N}, v), \sigma^2/q),$$

i.e., the estimator is unbiased and its variance is given by σ^2/q [6]. Hence, if the sample size satisfies $q \geq Z^2_{\alpha/2}\sigma^2/e^2$, then $P(|\phi_i(\mathcal{N}, v) - \widetilde{\phi}_i(\mathcal{N}, v)| \leq e) \geq 1 - \alpha$, with e being the approximation error, $Z \sim N(0, 1)$, and $Z^2_{\alpha/2}$ being the value such that $P(Z \geq Z^2_{\alpha/2}) = \alpha/2$. Given that σ^2 is unknown, it is necessary to provide an upper bound, which becomes $\sigma^2 \leq (x^i_{\max} - x^i_{\min})/4$ for any random variable bounded in a range $[x^i_{\min}, x^i_{\max}]$ [6]. If we take into account that $\phi_i(\mathcal{N}, v) \in [0, 1]$, then the bound simply becomes $\sigma^2 \leq 0.25$ in our case.

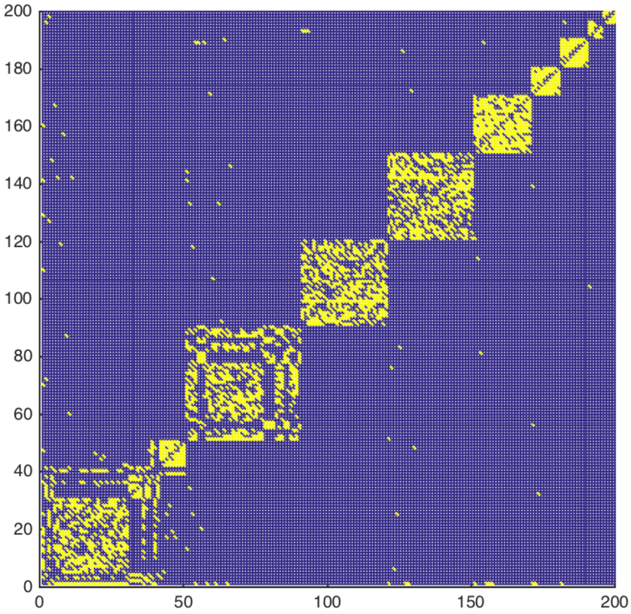

FIGURE 2. Example web with 200 pages.

F. EXAMPLE

In this subsection we deal with a web composed of 200 nodes. The links between the nodes can be seen in Figure 2, where yellow and blue are used to denote respectively the presence and absence of links between two given nodes. Notice that the direct computation of the Shapley value of any of the games proposed in this work requires the computation of the value of 2^{200} coalitions (1.60×10^{60}), which is not feasible.

The application of the method from the previous subsection is particularly interesting in this case. In particular, the fact that the PageRank of any given node is limited between 0 and 1 allows calculating accurate results with a reduced set of samples. For example, a maximum error of 0.01 in the PR requires the evaluation of 829400 coalitions. A maximum error ten times lower requires to compute 1353600 coalitions. As can be seen, the resulting computation burden is feasible within the limits imposed by current technology. In addition, it must be noticed that the computation can be easily parallelized.

Figure 3 shows the PageRank and Shapley values of the aggregation and difference games of the example calculated for a maximum error of 5×10^{-4} and a 99% confidence level. In general, there is a strong correlation between the PageRank and the Shapley value of the aggregation game. However, the difference in the value of the last nodes is remarkable. As can be seen in Figure 2, these nodes are almost isolated. Hence, it is difficult for the random surfer to get out of this set of nodes once he is there. As a consequence, the PageRank of these nodes is high. However, this effect is corrected in the Shapley value of the aggregation game. From this perspective, there is not much to gain by merging with the nodes in this set. Moreover, the Shapley value of the difference game shows losses in terms of PageRank for any merger containing one

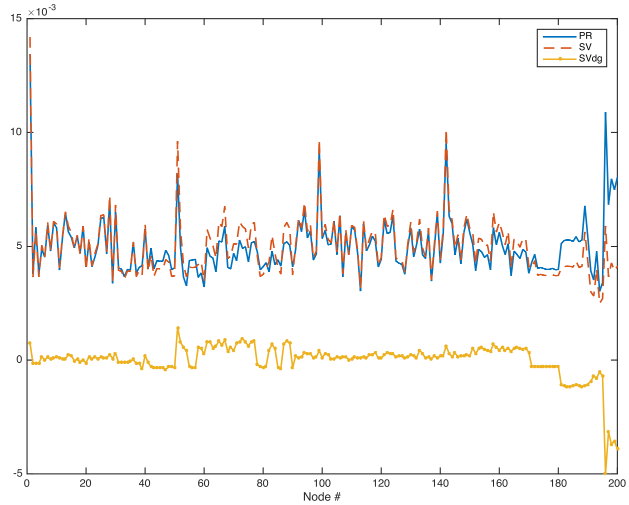

FIGURE 3. PageRank (PR), Shapley values of the aggregation (SV) and difference games (SVdg) of the 200 nodes example.

of the nodes in this set. Merging breaks the isolation of the nodes, and hence the gain of PageRank may be lost. Finally, the Shapley value of the difference game shows that nodes more likely to provide an increased PageRank are those between 50 and 90.

IV. LOCAL APPROACH

Even when we have shown that obtaining information regarding the suitability of merging a set of nodes $\mathcal{S}$ is feasible, full information is not always available. In such cases, it is possible to assess the potential of the merger by means of an approximation of $v(\mathcal{S})$, denoted by $v'(\mathcal{S})$, before making any decision. To this end, let us partition the nodes $\mathcal{N}$ in $\mathcal{G}$ in three disjoint sets:

- $\mathcal{S}$: These are the core nodes, i.e., those whose fusion is to be assessed.
- $\mathcal{I}$: Interface nodes, which are the nodes that point to or are pointed to by the nodes in $\mathcal{S}$. These are the neighbors of the core nodes, i.e., their interface with the rest of the network. We assume that information from this nodes is also available.
- $\mathcal{O}$: The rest of the nodes are grouped here and are called the outer nodes.

Without loss of generality, let us assume that the set $\mathcal{N}$ in $\mathcal{G}$ has its elements arranged in the following order: outer nodes ($\mathcal{O}$), interface nodes ($\mathcal{I}$), and core nodes ($\mathcal{S}$). The link-matrix $\mathbf{A}$ has this structure as well:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{\mathcal{OO}} & \mathbf{A}_{\mathcal{OI}} & \mathbf{A}_{\mathcal{OS}} \\ \mathbf{A}_{\mathcal{IO}} & \mathbf{A}_{\mathcal{II}} & \mathbf{A}_{\mathcal{IS}} \\ \mathbf{A}_{\mathcal{SO}} & \mathbf{A}_{\mathcal{SI}} & \mathbf{A}_{\mathcal{SS}} \end{bmatrix}. \tag{7}$$

Let us define the function PR$(\cdot)$ that provides us with the PageRank vector of the nodes in its argument, e.g., PR$(\mathcal{S})$ stands for the PageRank of the core nodes. As we saw, the PageRank is the eigenvector corresponding to the unit eigenvalue of the matrix $\mathbf{M}$ in (2). Hence, the PageRank of the core nodes has to satisfy the following equation:

$$PR(\mathcal{S}) = (1-m)\,[\mathbf{A}_{\mathcal{SO}}\ \ \mathbf{A}_{\mathcal{SI}}\ \ \mathbf{A}_{\mathcal{SS}}]\,PR(\mathcal{N}) + \frac{m}{|\mathcal{N}|}\mathbf{1}^{|\mathcal{S}|\times 1}.$$

Taking into account that $\mathbf{A}_{\mathcal{SO}} = \mathbf{0}$ by construction and that we can decompose the PageRank vector as $PR(\mathcal{N}) = [PR(\mathcal{O})^{\mathrm{T}}, PR(\mathcal{I})^{\mathrm{T}}, PR(\mathcal{S})^{\mathrm{T}}]^{\mathrm{T}}$, it is possible to rewrite the equation as

$$\begin{aligned} PR(\mathcal{S}) = {} & \left[(m-1)\mathbf{A}_{\mathcal{SS}} + \mathbf{I}^{|\mathcal{S}|\times|\mathcal{S}|}\right]^{-1} \\ & \cdot \left[(1-m)\mathbf{A}_{\mathcal{SI}}PR(\mathcal{I}) + \frac{m}{|\mathcal{N}|}\mathbf{1}^{|\mathcal{S}|\times 1}\right]. \end{aligned} \tag{8}$$

Remark 4: A more general version of (8) can be built by considering a non-homogeneous jump matrix $\mathbf{J}$*. In this case, (8) becomes*

$$\begin{aligned} PR(\mathcal{S}) = {} & \left[(m-1)\mathbf{A}_{\mathcal{SS}} - m\mathbf{J}_{\mathcal{SS}} + \mathbf{I}^{|\mathcal{S}|\times|\mathcal{S}|}\right]^{-1} \\ & \cdot [(1-m)\mathbf{A}_{\mathcal{SI}}PR(\mathcal{I}) \\ & + m(\mathbf{J}_{\mathcal{SO}}PR(\mathcal{O}) + \mathbf{J}_{\mathcal{SI}}PR(\mathcal{I}))]\,. \end{aligned} \tag{9}$$

A. NAIVE COALITION VALUE ESTIMATION

Equation (8) allows us to calculate the PageRank of the set of core nodes from the PageRank of its neighbors. This also motivates us to introduce our first and simplest estimator for $v(\mathcal{S})$ as the sum of the PageRank of the corresponding nodes in $\mathcal{S}$ as

$$v^{\mathrm{SPR}}(\mathcal{S}) = \mathbf{1}^{1\times|\mathcal{S}|}PR(\mathcal{S}),$$

where $PR(\mathcal{S})$ is obtained as in (8). Despite its simplicity, $v^{\mathrm{SPR}}(\mathcal{S})$ has a problem: it cannot be used to obtain information regarding the super-additivity of the merger regarding PageRank. This is important because we would like to know whether the PageRank of the merger will be greater than or equal to the sum of the PageRanks of the merged nodes.

B. CETERIS PARIBUS APPROACH

Equation (8) can also be applied to the link-matrix of the network resulting from merging the nodes in the coalition $\mathcal{S}$, denoted by $\mathbf{A}'$, i.e.,

$$PR'(\mathcal{S}) = \frac{(1-m)\mathbf{A}'_{\mathcal{SI}}PR'(\mathcal{I}) + \frac{m}{|\mathcal{N}|-|\mathcal{S}|+1}}{(m-1)\mathbf{A}'_{\mathcal{SS}} + 1}, \tag{10}$$

where the matrices $\mathbf{A}'_{\mathcal{SS}}$ and $\mathbf{A}'_{\mathcal{SI}}$ can be easily obtained from $\mathbf{A}$ as

$$\begin{aligned} \mathbf{A}'_{\mathcal{SS}} &= \frac{\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{SS}}\mathbf{1}^{|\mathcal{S}|\times 1}}{|\mathcal{S}|}, \\ \mathbf{A}'_{\mathcal{SI}} &= \mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{SI}}. \end{aligned}$$

Here, the superscript $'$ denotes the PageRank and the matrices that correspond to the network after merging the core nodes. Note that this is consistent with Section II-A. Also, everything in (10) can be known in advance with the exception of $PR'(\mathcal{I})$, i.e., the PageRank of neighbors after the merging. Note moreover that Equation (10) provides also the actual value of $v(\mathcal{S})$, i.e., $v(\mathcal{S}) = PR'(\mathcal{S})$. As a consequence, and taking into account that super-additive coalitions must satisfy

$$v(\mathcal{S}) \geq v^{\mathrm{SPR}}(\mathcal{S}),$$

it is possible to write the condition that has to be satisfied to have a super-additive coalition.

Proposition 2: A coalition $\mathcal{S}$ generates super-additivity in terms of PageRank if and only if the following condition holds:

$$\frac{(1-m)\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{I}}PR'(\mathcal{I}) + \frac{m}{|\mathcal{N}|-|\mathcal{S}|}}{(m-1)\frac{\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{S}}\mathbf{1}^{|\mathcal{S}|\times 1}}{|\mathcal{S}|}+1} \geq \mathbf{1}^{1\times|\mathcal{S}|}\left[(m-1)\mathbf{A}_{\mathcal{S}\mathcal{S}} + \mathbf{I}^{|\mathcal{S}|\times|\mathcal{S}|}\right]^{-1} \cdot \left[(1-m)\mathbf{A}_{\mathcal{S}\mathcal{I}}PR(\mathcal{I}) + \frac{m}{|\mathcal{N}|}\mathbf{1}^{|\mathcal{S}|\times 1}\right]. \tag{11}$$

No proof is provided for the proposition since it is a simple consequence of the definitions of $v(\mathcal{S})$ and $v^{\mathrm{SPR}}(\mathcal{S})$.

Equation (11) highlights two key ideas for a coalition to be super-additive:

- Coalitions with nodes with a strong internal link interaction with respect to their interaction with rest of the network have more chances due to the term $\mathbf{A}_{\mathcal{S}\mathcal{S}}$ in the denominator of (11). Given that $m-1 < 0$, the greater the elements of this matrix are, the better.
- Coalitions making their neighbors better off after the merging have better chances due to the influence of $PR'(\mathcal{I})$ in (11). That is, nodes forming a coalition should not only pay attention to their PageRank. Their neighbors' PageRank is very important in this regard.

Note also that it is not possible to evaluate Equation (11) before actually computing $\mathbf{A}'$. Nevertheless, if we take the classical *ceteris paribus* approach and assume that everything holds constant despite merging the nodes, we can evaluate Equation (11) by taking $PR'(\mathcal{I}) = PR(\mathcal{I})$. This assumption provides us with a reasonable guess of whether we can expect a super-additive coalition as long as the PageRank of neighbors is not much altered after the merging, which is likely to happen in a large-scale network. As a consequence, we arrive at our *ceteris paribus* estimator for $v(\mathcal{S})$, which is defined as

$$v^{\mathrm{CP}}(\mathcal{S}) = \frac{(1-m)\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{I}}PR(\mathcal{I}) + \frac{m}{|\mathcal{N}|-|\mathcal{S}|+1}}{(m-1)\frac{\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{S}}\mathbf{1}^{|\mathcal{S}|\times 1}}{|\mathcal{S}|}+1}. \tag{12}$$

A second version of this estimator can be defined in the line of Equation (9), i.e., for the case in which the jump matrix is also affected by the node aggregation. In this case, the probability of randomly jumping to the merger becomes the sum of the probabilities of the merged nodes before the jump took place, hence giving

$$v^{\mathrm{CP2}}(\mathcal{S}) = \frac{(1-m)\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{I}}PR(\mathcal{I}) + m\frac{|\mathcal{S}|}{|\mathcal{N}|}}{(m-1)\frac{\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{S}}\mathbf{1}^{|\mathcal{S}|\times 1}}{|\mathcal{S}|}+1}. \tag{13}$$

C. EQUIVALENT SUBNETWORKS IN A PAGERANK SENSE

In this subsection, we take a slightly different approach and pay attention to the following problem: given a network $\mathcal{G}$ that contains a subnetwork $\mathcal{G}_{\mathrm{sub}}$, we would like to find another network $\mathcal{G}'$ that also contains $\mathcal{G}_{\mathrm{sub}}$ so that the PageRank of the nodes in $\mathcal{G}_{\mathrm{sub}}$ is equal in both networks $\mathcal{G}$ and $\mathcal{G}'$.

Definition 1: Let $\mathcal{G}$ and $\mathcal{G}'$ be two networks containing a subnetwork $\mathcal{G}_{\mathrm{sub}}$. We say that they are equivalent in a strict PageRank sense for the subnetwork $\mathcal{G}_{\mathrm{sub}}$ if the PageRank of the nodes in $\mathcal{G}_{\mathrm{sub}}$ is the same in both networks.

Our goal is to find a network $\mathcal{G}'$ simpler than $\mathcal{G}$ that allows assessing the convenience of merging some of the elements in $\mathcal{G}_{\mathrm{sub}}$. In particular, we assume that $\mathcal{G}_{\mathrm{sub}}$ contains the nodes in $\mathcal{S}$ and $\mathcal{I}$ and their internal links.

Proposition 3: Given a subnetwork $\mathcal{G}_{\mathrm{sub}}$ of $\mathcal{G}$ with a nonempty set $\mathcal{S}$ of core nodes, a simplified network $\mathcal{G}'$ that also contains $\mathcal{G}_{\mathrm{sub}}$ must have the same number of nodes with $\mathcal{G}$ to provide the same PageRank to the nodes in $\mathcal{G}_{\mathrm{sub}}$, i.e., to provide strict PageRank equivalence for subnetwork $\mathcal{G}_{\mathrm{sub}}$.

Proof: Let $\mathbf{A}$ and $\mathbf{A}'$ be the link-matrices that correspond respectively to $\mathcal{G}$ and $\mathcal{G}'$. Strict PageRank equivalence requires that the condition $PR(\mathcal{S}) = PR'(\mathcal{S})$ holds. Now note that:

- $\mathbf{A}'_{\mathcal{S}\mathcal{O}}$ is a zero block because there is no link relationship between core and outer nodes.
- Likewise, $\mathbf{A}'_{\mathcal{S}\mathcal{I}}$ and $\mathbf{A}'_{\mathcal{S}\mathcal{S}}$ must be equal in $\mathbf{A}'$ and $\mathbf{A}$ to preserve the link relationship between these groups of nodes.

Since Equation (8) can be applied to both networks, we can observe that $PR(\mathcal{S}) = PR'(\mathcal{S})$ holds only if

$$\frac{m}{|\mathcal{N}|}\mathbf{1}^{|\mathcal{S}|\times 1} = \frac{m}{|\mathcal{N}'|}\mathbf{1}^{|\mathcal{S}|\times 1},$$

which is equivalent to $|\mathcal{N}| = |\mathcal{N}'|$. Consequently, a different number of nodes will lead to a different value of the PageRank in the core nodes. Q.E.D.

According to Proposition 3, it is possible to obtain a network $\mathcal{G}'$ with a simpler structure but the number of nodes must remain constant whenever there exist core nodes. Overcoming this issue requires to introduce changes in the way the PageRank is computed following the ideas of Remark 1. If the structure of $\mathbf{M}$ is changed, then it may be possible to find a network $\mathcal{G}'$ with a reduced number of nodes that still provides the same PageRank value for the nodes of the subnetwork under study according to the modified version of the PageRank computation. To this end, we introduce the following definition:

Definition 2: Let $\mathcal{G}$ and $\mathcal{G}'$ be two networks containing a subnetwork $\mathcal{G}_{\mathrm{sub}}$. For $\mathcal{G}'$, let $\mathbf{M}'$ be a column stochastic matrix built using a non-uniform distribution of the random surfer jumps, which is denoted by $\mathbf{J}'$, i.e., $\mathbf{M}' = (1-m)\mathbf{A}' + m\mathbf{J}'$. Let PR' be the eigenvector that corresponds to the eigenvalue 1 of this matrix. We say that $\mathcal{G}$ and $\mathcal{G}'$ are equivalent in a wide PageRank sense for the subnetwork $\mathcal{G}_{\mathrm{sub}}$ if the PageRank PR *of $\mathcal{G}$ and the modified PageRank* PR′ *of $\mathcal{G}'$ are the same for the nodes in $\mathcal{G}_{\mathrm{sub}}$.*

From the definition, it is clear that wide-sense PageRank equivalence is less restrictive than strict PageRank equivalence. The reduced network $\mathcal{G}'$ can be ultimately described by $\mathbf{A}'$, which must be calculated.

Proposition 4: Let $\mathbf{A}'$ *and* $\mathbf{J}'$ *be nonnegative column stochastic matrices representing respectively the link-matrix and the jump matrix of network* $\mathcal{G}'$ *so that*

$$\left((1-m)\mathbf{A}' + m\mathbf{J}'\right)\mathrm{PR}'(\mathcal{N}') = \mathrm{PR}'(\mathcal{N}'). \tag{14}$$

For the networks $\mathcal{G}$ *and* $\mathcal{G}'$ *to be equivalent in a wide PageRank sense for a subnetwork* $\mathcal{G}_{\text{sub}}$*, it holds that*

$$\mathbf{A}'_{\mathcal{I}\mathcal{I}} = \mathbf{A}_{\mathcal{I}\mathcal{I}}, \tag{15}$$

$$\mathbf{A}'_{\mathcal{I}\mathcal{S}} = \mathbf{A}_{\mathcal{I}\mathcal{S}}, \tag{16}$$

$$\mathbf{A}'_{\mathcal{S}\mathcal{I}} = \mathbf{A}_{\mathcal{S}\mathcal{I}}, \tag{17}$$

$$\mathbf{A}'_{\mathcal{S}\mathcal{S}} = \mathbf{A}_{\mathcal{S}\mathcal{S}}, \tag{18}$$

$$\mathbf{A}'_{\mathcal{S}\mathcal{O}} = \mathbf{0}, \tag{19}$$

$$\mathbf{A}'_{\mathcal{O}\mathcal{S}} = \mathbf{0}, \tag{20}$$

$$PR'(\mathcal{I}) = PR(\mathcal{I}), \tag{21}$$

$$PR'(\mathcal{S}) = PR(\mathcal{S}). \tag{22}$$

Proof: The inner structure of $\mathcal{G}_{\text{sub}}$ is preserved by means of the equality constraints of $\mathbf{A}'$ (15)-(20). Both $\mathbf{A}'_{\mathcal{S}\mathcal{O}}$ and $\mathbf{A}'_{\mathcal{O}\mathcal{S}}$ are zero blocks of the corresponding sizes and the elements in $\mathbf{A}'_{\mathcal{I}\mathcal{I}}$, $\mathbf{A}'_{\mathcal{I}\mathcal{S}}$, $\mathbf{A}'_{\mathcal{S}\mathcal{I}}$ and $\mathbf{A}'_{\mathcal{S}\mathcal{S}}$ have the same values as those in the corresponding submatrices in $\mathbf{A}$. Finally, Equations (21) and (22) require that the core and interface nodes receive the same value in PR', which stems from the definition of equivalent network in a wide PageRank sense.Q.E.D.

Notice that the conditions given are necessary but they do not lead in general to a unique solution. In particular, the following elements have to be designed: $\mathbf{A}'_{\mathcal{O}\mathcal{O}}$, $\mathbf{A}'_{\mathcal{I}\mathcal{O}}$, $\mathbf{A}'_{\mathcal{O}\mathcal{I}}$, $\mathbf{J}'$, and $PR'(\mathcal{O}')$. The direct computation of these elements based on (14) leads us to a set of bilinear equations because there are cross multiplications in these variables. Alternatively, it may be preferable to solve a set of linear equations for different fixed values of the modified PageRank value of the external nodes, especially because there is an incentive to keep a low number of them for the sake of simplicity.

Another noteworthy point is that there may not be a feasible solution for the set of constraints given, i.e., the set of equations may be incompatible, or even when a solution can be found $\mathbf{A}'$ may be non-stochastic, etc. In that case, a new attempt can be made with an increase in the number of external nodes. Consecutive increments can be made until $\mathbf{A}'$ is found. Finally, notice that the increase of external elements is bounded because in the worst case a solution for the problem is guaranteed trivially for $\mathcal{G} = \mathcal{G}'$. In that case $\mathcal{G}$ would not be reducible for the considered subnetwork.

Once this merged network $\mathcal{G}'$ has been calculated, an estimation of $v(\mathcal{S})$ is given by $v'(\mathcal{S})$, i.e., the PageRank that corresponds to the fusion of the nodes in the reduced network.

Next, we present a particular case of this approach and the corresponding estimator.

1) DIRECT APPROXIMATION

Here, we introduce a direct approximation for $v'(\mathcal{S})$ based on a simplification of the original network $\mathcal{G}$ in which we aggregate all the nodes outside $\mathcal{G}_{\text{sub}}$ into a single node. As we will see, even if full information is not available, it is possible to calculate this approximation if the links between the nodes in $\mathcal{I}$ and $\mathcal{O}$ are known. To this end, we construct $\mathbf{A}'$ as follows:

$$\mathbf{A}' = \begin{bmatrix} 1 - \mathbf{1}^{1\times|\mathcal{I}|}\mathbf{A}'_{\mathcal{I}\mathcal{O}} & \mathbf{A}'_{\mathcal{O}\mathcal{I}} & 0 \\ \mathbf{A}'_{\mathcal{I}\mathcal{O}} & \mathbf{A}_{\mathcal{I}\mathcal{I}} & \mathbf{A}_{\mathcal{I}\mathcal{S}} \\ 0 & \mathbf{A}_{\mathcal{S}\mathcal{I}} & \mathbf{A}_{\mathcal{S}\mathcal{S}} \end{bmatrix}, \tag{23}$$

with

$$\mathbf{A}'_{\mathcal{O}\mathcal{I}} = \mathbf{1}^{1\times|\mathcal{I}|} - \mathbf{1}^{1\times|\mathcal{I}|}\mathbf{A}_{\mathcal{I}\mathcal{I}} - \mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{I}} \tag{24}$$

and

$$\mathbf{A}'_{\mathcal{I}\mathcal{O}} = \frac{\mathbf{A}_{\mathcal{I}\mathcal{O}}PR(\mathcal{O})}{1 - \sum\limits_{i\in\mathcal{S}\cup\mathcal{I}} PR(i)}. \tag{25}$$

Likewise, the matrix $\mathbf{J}'$ is built by accumulating the jump probability of the nodes aggregated into the new node while it remains constant for the rest of the nodes in the network, i.e.,

$$\mathbf{J}' = \begin{bmatrix} (1 - \dfrac{|\mathcal{I}| + |\mathcal{S}|}{|\mathcal{N}|})\mathbf{1}^{1\times(1+|\mathcal{I}|+|\mathcal{S}|)} \\ \dfrac{1}{|\mathcal{N}|}\mathbf{1}^{(|\mathcal{I}|+|\mathcal{S}|)\times(1+|\mathcal{I}|+|\mathcal{S}|)} \end{bmatrix}. \tag{26}$$

Once the simplified network is calculated, it is possible to assess the performance of the merger by using (10). To this end, the nodes in $\mathcal{S}$ are merged and the PageRank of the resulting merger $PR''(\mathcal{S})$ provides us with the direct approximation estimator $v^{\text{DA}}(\mathcal{S})$. It is also possible to calculate an analytical expression for this estimator based on the PageRank of the interface nodes as follows:

$$v^{\text{DA}}(\mathcal{S}) = \frac{(1-m)\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{I}}PR''(\mathcal{I}) + \frac{m}{|\mathcal{I}|+2}}{(m-1)\frac{\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{S}}\mathbf{1}^{|\mathcal{S}|\times 1}}{|\mathcal{S}|} + 1}, \tag{27}$$

where $PR''(\mathcal{I})$ is the PageRank that the interface nodes $\mathcal{I}$ receive in the new network. Here, the double apostrophe $''$ denotes that the network has been transformed twice before the calculation of the PageRank of these nodes. As can be seen, the rest of the elements in (27) belong to the original link matrix $\mathbf{A}$.

A second version for this estimator is based on the calculation of the PageRank by means of the corresponding aggregated jump matrix

$$\mathbf{J}'' = \begin{bmatrix} (1 - \dfrac{|\mathcal{I}| + |\mathcal{S}|}{|\mathcal{N}|})\mathbf{1}^{1\times(2+|\mathcal{I}|)} \\ \dfrac{1}{|\mathcal{N}|}\mathbf{1}^{|\mathcal{I}|\times(2+|\mathcal{I}|)} \\ \dfrac{|\mathcal{S}|}{|\mathcal{N}|}\mathbf{1}^{1\times(2+|\mathcal{I}|)} \end{bmatrix}.$$

$$\mathbf{A}' = \begin{bmatrix} 0.00 & 0.10 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.62 & 0.50 & 0.50 & 1.00 & 0.40 & 0.60 & 0.50 & 0.80 & 0.00 & 0.00 & 0.60 & 0.00 & 0.00 \\ 1.00 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.06 & 0.00 & 0.00 & 0.00 & 0.00 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.25 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.25 & 0.00 \\ 0.00 & 0.03 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.25 & 0.33 \\ 0.00 & 0.02 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.20 & 0.25 & 0.00 \\ 0.00 & 0.03 & 0.00 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.03 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.04 & 0.25 & 0.00 & 0.00 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.06 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.00 & 0.50 & 0.00 & 0.20 & 0.00 & 0.33 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.33 \\ 0.00 & 0.00 & 0.25 & 0.00 & 0.00 & 0.00 & 0.20 & 0.00 & 0.20 & 0.50 & 1.00 & 0.00 & 0.00 & 0.00 \end{bmatrix} \tag{29}$$

We denote the estimator by $v^{\mathrm{DA2}}(\mathcal{S})$. We can also provide an analytical expression for this estimator based on the contribution of interface nodes by using (9) as follows:

$$v^{\mathrm{DA2}}(\mathcal{S}) = \frac{\left[(1-m)\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{I}}\right] PR''(\mathcal{I}) + m\frac{|\mathcal{S}|}{|\mathcal{N}|}}{(m-1)\frac{\mathbf{1}^{1\times|\mathcal{S}|}\mathbf{A}_{\mathcal{S}\mathcal{S}}\mathbf{1}^{|\mathcal{S}|\times 1}}{|\mathcal{S}|}+1}. \tag{28}$$

D. EXPERIMENTAL RESULTS

In this subsection, we show some experimental results to illustrate the different local information approaches proposed.

1) EQUIVALENT NETWORK EXAMPLE

In Figure 4 a randomly generated 30 node network is depicted. Let's suppose that we would like to assess whether nodes 4 and 8 should be merged, which would be our core nodes. We will take the 1-hop neighborhood as interface nodes, i.e., nodes 1, 2, 9, 11, 16, 19, 23, 24, 25, and 26. Consequently, nodes 3, 5, 6, 7, 10, 12, 13, 14, 15, 17, 18, 20, 21, 22, 27, 28, 29 and 30 are external nodes.

The method described after Proposition 4 can be applied to find a link-matrix equivalent in a wide PageRank sense by solving the set of bilinear equations that stems from (14). The result is given in (29), as shown at the top of this page and corresponds to a 14-*node* network that provides the same PageRank value for all the nodes of interest.

2) ESTIMATOR ASSESSMENT

Finally, we have assessed our estimators. To this end, 10000 simulations with randomly generated networks were carried out for different network sizes in the range [25, 200]. The probability that a link is established between any given two nodes was set to 0.1. The coalition size was also randomly set in the range [2, 5].

The results are shown in Tables 4-7. There, we can see data regarding the average absolute ($|\epsilon|$) and relative ($|\epsilon_r|$) approximation error for two cases, namely, standard PageRank calculation (PR) and PageRank when an aggregated jump matrix is used (PR (aggr. **J**)). As for the latter, it corresponds to the case where the merger receives a probability of jump that is equal to the sum of the probabilities of the nodes before merging. The standard deviations for these errors are also provided ($|\epsilon|_\sigma$ and $|\epsilon_r|_\sigma$). In addition, two additional columns are added to show the rates of correct predictions of PageRank super-additivity/subadditivity based on the estimator turns to be true. These are denoted by CFR, which stands for correct forecast rate. Note that the performance of $v^{\mathrm{SPR}}(\mathcal{S})$ in this regard is omitted because this estimator cannot be used to predict super-additivity.

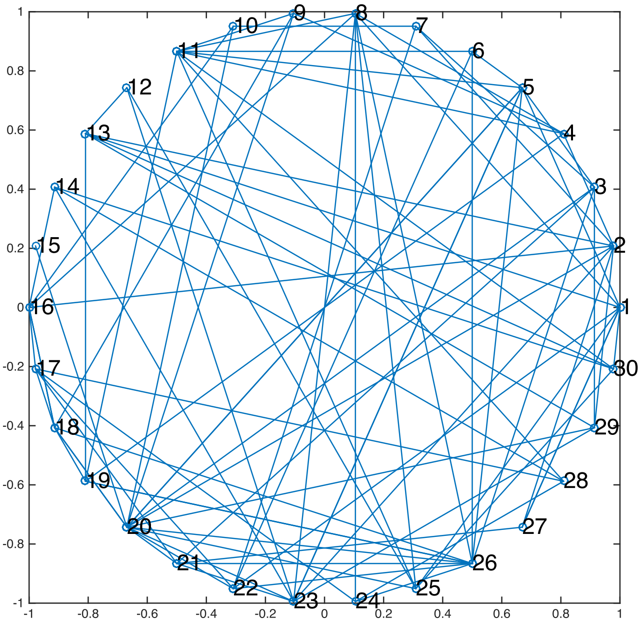

FIGURE 4. Example web with thirty pages.

Clearly, the estimators work better for the cases they are designed for. It is remarkable that some of them offer much better results than the mere sum of PageRank of the merged nodes, specially regarding the prediction of super-additivity. In particular, $v^{\mathrm{CP}}(\mathcal{S})$ and $v^{\mathrm{DA2}}(\mathcal{S})$ are capable of providing very accurate predictions on this matter. Moreover, these estimators become better as the size of the network grows, which

TABLE 4. Assessment of the estimations ($|\mathcal{N}| = 25$).

	PR					PR (aggr. $\mathbf{J}$)				
Method	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR
$v^{\text{SPR}}(\mathcal{S})$	0.0757	0.0553	2.1323	1.3541	-	0.0113	0.0203	0.2052	0.3529	-
$v^{\text{CP}}(\mathcal{S})$	0.0025	0.0100	0.0293	0.0785	0.9883	0.0693	0.0561	0.5810	0.1788	0.6898
$v^{\text{CP2}}(\mathcal{S})$	0.0686	0.0560	1.8003	1.0529	0.6765	0.0018	0.0097	0.0137	0.0496	0.8926
$v^{\text{DA}}(\mathcal{S})$	0.2246	0.1581	5.9453	2.6092	0.0293	0.1558	0.1068	1.5056	0.5792	0.3484
$v^{\text{DA2}}(\mathcal{S})$	0.0688	0.0554	1.8063	1.0460	0.6934	0.0001	0.0005	0.0006	0.0035	0.9057

TABLE 5. Assessment of the estimations ($|\mathcal{N}| = 50$).

	PR					PR (aggr. $\mathbf{J}$)				
Method	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR
$v^{\text{SPR}}(\mathcal{S})$	0.0147	0.0170	0.7701	0.9888	-	0.0055	0.0122	0.1390	0.2873	-
$v^{\text{CP}}(\mathcal{S})$	0.0008	0.0025	0.0251	0.0478	0.9956	0.0115	0.0096	0.3170	0.1659	0.5219
$v^{\text{CP2}}(\mathcal{S})$	0.0114	0.0095	0.6067	0.6307	0.6165	0.0007	0.0024	0.0139	0.0319	0.8580
$v^{\text{DA}}(\mathcal{S})$	0.0302	0.0494	1.9042	3.0676	0.2648	0.0199	0.0426	0.6354	1.0966	0.6039
$v^{\text{DA2}}(\mathcal{S})$	0.0115	0.0093	0.6092	0.6303	0.5174	0.0000	0.0001	0.0009	0.0015	0.9769

TABLE 6. Assessment of the estimations ($|\mathcal{N}| = 100$).

	PR					PR (aggr. $\mathbf{J}$)				
Method	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR
$v^{\text{SPR}}(\mathcal{S})$	0.0052	0.0050	0.5184	0.7495	-	0.0010	0.0031	0.0602	0.1984	-
$v^{\text{CP}}(\mathcal{S})$	0.0001	0.0002	0.0073	0.0133	1.0000	0.0047	0.0032	0.2595	0.1455	0.5046
$v^{\text{CP2}}(\mathcal{S})$	0.0046	0.0032	0.4411	0.4860	0.6158	0.0001	0.0001	0.0031	0.0053	0.8308
$v^{\text{DA}}(\mathcal{S})$	0.0128	0.0319	1.8146	4.4072	0.3790	0.0094	0.0294	0.6756	1.6691	0.5284
$v^{\text{DA2}}(\mathcal{S})$	0.0047	0.0033	0.4420	0.4869	0.5046	0.0000	0.0000	0.0001	0.0001	0.9916

TABLE 7. Assessment of the estimations ($|\mathcal{N}| = 200$).

	PR					PR (aggr. $\mathbf{J}$)				
Method	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR	$\lvert\epsilon\rvert$	$\lvert\epsilon\rvert_\sigma$	$\lvert\epsilon_r\rvert$	$\lvert\epsilon_r\rvert_\sigma$	CFR
$v^{\text{SPR}}(\mathcal{S})$	0.0022	0.0017	0.4371	0.6329	-	0.0002	0.0008	0.0244	0.1244	-
$v^{\text{CP}}(\mathcal{S})$	0.0000	0.0000	0.0031	0.0068	1.0000	0.0021	0.0014	0.2420	0.1435	0.4938
$v^{\text{CP2}}(\mathcal{S})$	0.0021	0.0014	0.4034	0.4699	0.6396	0.0000	0.0000	0.0007	0.0016	0.8212
$v^{\text{DA}}(\mathcal{S})$	0.0065	0.0228	2.2127	6.9114	0.4264	0.0053	0.0218	0.8301	2.4574	0.5136
$v^{\text{DA2}}(\mathcal{S})$	0.0021	0.0014	0.4036	0.4703	0.4944	0.0000	0.0000	0.0000	0.0000	0.9944

make them very suitable for real large-scale networks. Likewise, the estimations provided by $v^{\text{CP2}}(\mathcal{S})$ are good, although they are not as accurate as those of $v^{\text{DA2}}(\mathcal{S})$. Surprisingly, $v^{\text{DA}}(\mathcal{S})$ offers very poor results, specially in terms of absolute relative error. In fact, the results of this estimator are even worse than those of $v^{\text{SPR}}(\mathcal{S})$.

V. CONCLUSIONS

In this work, we have studied the problem of merging nodes in a network from a PageRank viewpoint. A global perspective analysis has allowed us to define games of interest in this context. Two measures have been given, one of which is directly related to the amount of additional PageRank that can be expected when merging with a node. A method for the computation of the new measures in polynomial time for large networks has been also applied.

The same problem has also been addressed from a local perspective. The lack of full information is a strong limitation that only allows obtaining estimates of the expected PageRank value. Different estimators have been introduced and experiments have been carried out to show the remarkable accuracy of some of these approximations, very particularly in their capability of assessing the potential of the coalition to generate additional PageRank.

Future work should deal with the utilization of models of restricted cooperation to integrate the structure of the directed graph in the solution. Likewise, the application of these values to coalitional control schemes [11], [12], [24] will be also studied.

REFERENCES

[1] E. Algaba, J. M. Bilbao, J. R. F. García, and J. J. López, "Computing power indices in weighted multiple majority games," *Math. Soc. Sci.*, vol. 46, no. 1, pp. 63–80, 2003.

[2] K. Avrachenkov and N. Litvak, "The effect of new links on Google PageRank," *Stochastic Models*, vol. 22, no. 2, pp. 319–331, 2006.

[3] P. Borm, G. Owen, and S. Tijs, "On the position value for communication situations," *SIAM J. Discrete Math.*, vol. 5, no. 3, pp. 305–320, Aug. 1992.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.

[5] K. Bryan and T. Leise, "The $25,000,000,000 eigenvector: The linear algebra behind Google," *SIAM Rev.*, vol. 48, no. 3, pp. 569–581, 2006.

[6] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," *Comput. Oper. Res.*, vol. 36, no. 5, pp. 1726–1730, 2009.

[7] W. G. Cochran, *Sampling Techniques*. New York, NY, USA: Wiley, 1977.

[8] X. Deng and C. H. Papadimitriou, "On the complexity of cooperative solution concepts," *Math. Oper. Res.*, vol. 19, no. 2, pp. 257–266, May 1994.

[9] L. Ermann and D. L. Shepelyansky, "Google matrix of the world trade network," *Acta Phys. Polonica A*, vol. 120, no. 6A, pp. 158–171, 2011.

[10] S. S. Fatima, M. Wooldridge, and N. R. Jennings, "An analysis of the Shapley value and its uncertainty for the voting game," in *Agent-Mediated Electronic Commerce. Designing Trading Agents and Mechanisms*. Berlin, Germany: Springer, 2006, pp. 85–98.

[11] F. Fele, J. M. Maestre, and E. F. Camacho, "Coalitional control: Cooperative game theory and control," *IEEE Control Syst.*, vol. 37, no. 1, pp. 53–69, Feb. 2017.

[12] F. Fele, J. M. Maestre, M. Hashemy, D. M. de la Peña, and E. F. Camacho, "Coalitional model predictive control of an irrigation canal," *J. Process Control*, vol. 24, no. 4, pp. 314–325, 2014.

[13] J. R. Fernández, E. Algaba, J. M. Bilbao, A. Jiménez, N. Jiménez, and J. J. López, "Generating functions for computing the Myerson value," *Ann. Oper. Res.*, vol. 109, nos. 1–4, pp. 143–158, 2002.

[14] K. M. Frahm and D. L. Shepelyansky, "Google matrix of Twitter," *Eur. Phys. J. B*, vol. 85, no. 10, p. 335, 2012.

[15] V. Ginsburgh and I. Zang, "Shapley ranking of wines," *J. Wine Econ.*, vol. 7, pp. 169–180, Nov. 2012.

[16] D. Gómez, E. González-Arangüena, C. Manuel, G. Owen, M. del Pozo, and J. Tejada, "Centrality and power in social networks: A game theoretic approach," *Math. Soc. Sci.*, vol. 46, no. 1, pp. 27–54, 2003.

[17] D. Granot, J. Kuipers, and S. Chopra, "Cost allocation for a tree network with heterogeneous customers," *Math. Oper. Res.*, vol. 27, no. 4, pp. 647–661, 2002.

[18] B. Grofman and G. Owen, "A game theoretic approach to measuring degree of centrality in social networks," *Soc. Netw.*, vol. 4, no. 3, pp. 213–224, 1982.

[19] H. Ishii and R. Tempo, "Distributed randomized algorithms for the PageRank computation," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 1987–2002, Sep. 2010.

[20] H. Ishii and R. Tempo, "The PageRank problem, multiagent consensus, and Web aggregation: A systems and control viewpoint," *IEEE Control Syst.*, vol. 34, no. 3, pp. 34–53, Jun. 2014.

[21] V. Kandiah, H. Escaith, and D. L. Shepelyansky. (Jul. 2015). "Contagion effects in the world network of economic activities." [Online]. Available: https://arxiv.org/abs/1507.03278

[22] N. Ma, J. Guan, and Y. Zhao, "Bringing PageRank to the citation analysis," *Inf. Process. Manage.*, vol. 44, no. 2, pp. 800–810, 2008.

[23] J. M. Maestre and H. Ishii, "A cooperative game theory approach to the PageRank problem," in *Proc. Amer. Control Conf. (ACC)*, 2016, pp. 3820–3825.

[24] J. M. Maestre, D. M. de la Peña, A. J. Losada, E. Algaba, and E. F. Camacho, "A coalitional control scheme with applications to cooperative game theory," *Optim. Control Appl. Methods*, vol. 35, no. 5, pp. 592–608, 2014.

[25] B. Maury and A. Pajuste, "Multiple large shareholders and firm value," *J. Banking Finance*, vol. 29, no. 7, pp. 1813–1834, 2005.

[26] T. P. Michalak, K. V. Aadithya, P. L. Szczepanski, B. Ravindran, and N. R. Jennings, "Efficient computation of the shapley value for game-theoretic network centrality," *J. Artif. Intell. Res.*, vol. 46, pp. 607–650, Jan./Apr. 2013.

[27] F. J. Muros, J. M. Maestre, E. Algaba, T. Alamo, and E. F. Camacho, "Networked control design for coalitional schemes using game-theoretic methods," *Automatica*, vol. 78, pp. 320–332, Apr. 2017.

[28] R. Narayanam and Y. Narahari, "A Shapley value-based approach to discover influential nodes in social networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 8, no. 1, pp. 130–147, Jan. 2010.

[29] P. Papapetrou, A. Gionis, and H. Mannila, "A Shapley value approach for influence attribution," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2011, pp. 549–564.

[30] L. S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci. USA*, vol. 39, no. 10, pp. 1095–1100, 1953.

[31] L. S. Shapley, "A value for *n*-person games," *Contributions to the Theory of Games: Annals of Mathematics Studies*, vol. 28, H. W. Kuhn and A. W. Tucker, Eds. Princenton, NJ, USA: Princenton Univ. Press, 1953, pp. 307–317.

[32] M. Takayasu, S. Sameshima, T. Ohnishi, Y. Ikeda, H. Takayasu, and K. Watanabe, "Massive economics data analysis by econophysics methods—The case of companies' network structure," in *Proc. Annu. Rep. Earth Simulator Center*, Apr. 2007, pp. 237–242.

J. M. MAESTRE received the master's degree in telecommunications from the University of Seville in 2005, the master's degree in smart home from the Universidad Politcnica de Madrid in 2006, the master's degree in telecommunications economics from the Universidad Nacional de Educacin a Distancia in 2010, and the Ph.D. degree in automation and robotics and the master's degree in economics and development from the University of Seville, in 2010 and 2017, respectively. He was with the University of Seville as an Associate Professor. He was also with LTH at Lund University as a Guest Researcher, TU Delft, the University of Cadiz, and the Tokyo Institute of Technology. Besides his Ph.D., he was awarded with the Extra-Ordinary Prize of the University of Seville. He has authored and co-authored over 100 publications in his research fields, including the books *Distributed Model Predictive Control Made Easy* (Springer, 2014), *Domtica para Ingenieros* (Paraninfo, 2015), and *A Programar se Aprende Jugando* (Paraninfo, 2017). His main research interests are the control of distributed systems and the integration of service robots in the smart home.

HIDEAKI ISHII (M'02–SM'12) received the M.Eng. degree in applied systems science from Kyoto University, Kyoto, Japan, in 1998, and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 2002. He was a Post-Doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL, USA, from 2001 to 2004, and a Research Associate with the Department of Information Physics and Computing, The University of Tokyo, Tokyo, Japan, from 2004 to 2007. He is currently an Associate Professor with the Department of Computer Science, Tokyo Institute of Technology, Yokohama, Japan. His research interests are in networked control systems, multiagent systems, hybrid systems, cyber security of power systems, and probabilistic algorithms.

He received the *IEEE Control Systems Magazine* Outstanding Paper Award in 2015. He has served as an Associate Editor of the IEEE Transactions on Control of Network Systems, the IEEE Control Systems Letters, and *Mathematics of Control, Signals, and Systems* and previously for *Automatica* and the IEEE Transactions on Automatic Control. He was the Chair of the IFAC Technical Committee on Networked Systems from 2011 to 2017

E. ALGABA received the Ph.D. degree in mathematical sciences from the University of Seville. She is an Associate Professor with the Department of Applied Mathematics II, Advanced Engineering School, University of Seville, and a member of the Mathematics Research Institute with the University of Seville. Her main research line is about cooperative games on combinatorial structures with relevant contributions in the area. She is one of the Spanish representatives in the SING (Spain, Italy, The Netherlands Game Theory) Committee from 2012 and member of the Scientific Committee in SING 10 (Cracow, Poland 2014), European meeting on Game Theory SING 11-GTM 2015 (St. Petersburg, Russia 2015), SING 12 (Odense, Denmark 2016), and SING 13 (Paris, France 2017). She has acted as an evaluator of projects at the national and European level and she belongs to the national expert evaluators board from 2014. She has organized diverse courses, streams in international conferences, and workshops on Game Theory. She has been invited speaker and guest researcher in various universities, has given invited summer courses and numerous talks in international conferences.