Chaos
An Interdisciplinary Journal of Nonlinear Science

RESEARCH ARTICLE | AUGUST 28 2023

Can the PageRank centrality be manipulated to obtain any desired ranking?

Special Collection: Nonlinear dynamics, synchronization and networks: Dedicated to Jürgen Kurths' 70th birthday

Gonzalo Contreras-Aso ; Regino Criado ; Miguel Romance

Chaos 33, 083152 (2023)
https://doi.org/10.1063/5.0156226

Can the PageRank centrality be manipulated to obtain any desired ranking?

Cite as: Chaos 33, 083152 (2023); doi: 10.1063/5.0156226
Submitted: 27 April 2023 · Accepted: 7 August 2023 ·
Published Online: 28 August 2023

Gonzalo Contreras-Aso,[1,2,a)] Regino Criado,[1,2,3] and Miguel Romance[1,2,3]

AFFILIATIONS

[1]Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
[2]Laboratory of Mathematical Computation on Complex Networks and Their Applications, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
[3]Data, Complex networks and Cybersecurity Research Institute, Universidad Rey Juan Carlos, 28028, Madrid, Spain

Note: This paper is part of the Focus Issue on Nonlinear dynamics, synchronization and networks: Dedicated to Juergen Kurths' 70th birthday.

[a)]**Author to whom correspondence should be addressed:** gonzalo.contreras@urjc.es

ABSTRACT

The significance of the PageRank algorithm in shaping the modern Internet cannot be overstated, and its complex network theory foundations continue to be a subject of research. In this article, we carry out a systematic study of the structural and parametric controllability of PageRank's outcomes, translating a spectral graph theory problem into a geometric one, where a natural characterization of its rankings emerges. Furthermore, we show that the change of perspective employed can be applied to the biplex PageRank proposal, performing numerical computations on both real and synthetic network datasets to compare centrality measures used.

The rise of complex network theory owes much to the development of the Internet. While several factors contributed to this, one of the most significant was the introduction of the PageRank algorithm, which became a crucial measure of centrality in the network theory. Borrowing from both spectral theorems and the theory of random walks, this algorithm is both simple and efficient, making it a favorite of the network theory community. This led a myriad of researchers to scrutinize the properties, limitations, and implications of PageRank. In this article, we address one crucial aspect that has not been explored: the parametric controllability of PageRank rankings. By examining this issue, we can provide yet another argument for the reliability of PageRank as a ranking measure.

I. INTRODUCTION

Almost 25 years have passed since the PageRank algorithm was devised.[1] It brought about two revolutions: on the industry side, it shaped the Internet landscape making Google the giant it is today. On the academic side, it triggered an enormous cascade of studies, interested in understanding its properties, its limitations, and its implications.[2–6] Furthermore, it has been shown to be relevant beyond its original goal of webpage ranking: indeed, it has found applications in very diverse fields such as biology, engineering, and even literature (see Ref. 7 for an extensive survey).

The academic research poured in the PageRank algorithm coincided with both the development of the interdisciplinary field of complex networks and the advent of accessible computing resources. This allowed for both theoretical and numerical results[2,4,6,7] that have many direct applications in the economic and social world, since the PageRank algorithm is in the core of most popular web engines. One of these direct implications in marketing and economics is the so-called *Search Engine Optimization Problem* or *Web Positioning Problem*[8–10] that tries to find out the strategies that can be performed in a network in order to maximize PageRank of a specific node (or set of nodes). This theoretical problem has huge real applications with severe economic impact in the global markets. In our nowadays on-line world, for any company not only it is crucial to be present in the WWW, but also to appear highest in the ranking of any web engine; the web-master of a site is, thus, interested in increasing the PageRank of website by connecting it

properly with other webpages, since the highest the ranking of a website the biggest economic revenue the corresponding company gets.[11] This major search engine optimization problem belongs to a more general class of problems related with centrality measures of networks: the control of a centrality measure. This general problem deals with the ability to modify at our wish the centrality of a specific node (or set of nodes) of a given network by slightly changing the link structure of the network or by modifying the intrinsic parameters of the centrality measure. Note that the search engine optimization problem is related to the control of centrality measures by changing the link structure, while in this paper, we will focus on the control by modifying the intrinsic parameters of the centrality measure.

While search engine optimization problem has attracted broad attention by the scientific community, the control of a centrality measure by modifying its intrinsic parameters has been less considered despite the fact that it also has some potential real applications, since, for example, it gives valuable information for a web engine administrator about how to modify the ranking of a webpage (or a set of webpages) simply by tuning the parameters of the centrality measure that is behind his web searcher. It is well known that most of the web engines work with algorithms that modify their ingredients in order to improve the results,[6] so a detailed analysis of the influence and sensibility of each parameter of these centrality measures must be considered. In particular, in the case of PageRank centrality, there are two parameters of this measure to be considered:[6,7] the *damping factor* $\alpha \in (0,1)$ and the *personalization vector* $\boldsymbol{v} \in \mathbb{R}^n$. The damping factor has been extensively studied, discussed, and interpreted (see e.g., Ref. 4), but the role of personalization vector has always remained understudied.[12]

In this article, we attempt to shed some light on the relationship between the centrality vectors resulting from PageRank and the choice of personalization vectors. This is actually intertwined with the subject of centrality control in complex networks:[13] probing the space of possible centrality vectors with suitable changes in either the underlying graph or the centrality measure. There are already a number of studies discussing the possibility of increasing a node's own PageRank score[14–17] as well as some advances regarding PageRank competitors[12] (nodes whose relative ranking position depend on the value of the algorithm's parameters). While these approaches are interesting on their own, they focus on specific nodes and their scores or rankings. In this work, we discuss centrality vectors and their rankings as a whole, without reference to individual improvements or detriments.

This paper is structured as follows: In Sec. II, we establish some notation and basic graph-theoretical concepts as well as introduce the terminology that will be used throughout the paper. Section III presents the mathematical definition of the PageRank algorithm and then explores some routes toward controlling its resulting centrality, with either structural or parametric changes. Theoretical results connecting PageRank and personalization vectors are proven, and network datasets are then used for numerical comparisons and discussion of implications. In Sec. IV, we apply the same techniques to the case of the biplex PageRank,[18] an alternative centrality measure based on the PageRank algorithm. We conclude with a discussion and comparison between the results obtained with each of the different approaches.

II. PRELIMINARIES AND NOTATION

Let $G = (V, E)$ be a graph (irregardless of directionality or weights), with node set $V = \{1, \ldots, n\}$, for some $n \in \mathbb{N}$ and adjacency matrix $A = (a_{ij})$ such that

$$a_{ij} = \begin{cases} w_{ij} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where w_{ij} is the weight corresponding to edge (i,j), by default $w_{ij} = 1$ if unweighted.

The in-degree (number of incoming links) and out-degree (number of outgoing links) of node $i \in V$ are defined as

$$\deg_{in}(i) = \sum_{j=1}^{n} a_{ji}, \quad \deg_{out}(i) = \sum_{j=1}^{n} a_{ij}, \tag{2.2}$$

respectively. For undirected graphs, we clearly have $\deg_{in}(i) = \deg_{out}(i)$. Nodes in a graph with no outgoing links, i.e., such that $\deg_{out}(i) \neq 0$, are called *dangling* nodes. As will be pointed out later, only networks without dangling nodes will be considered, since similar results can be obtained for general settings simply by using some standard techniques.[12]

By using these definitions, we can introduce the first ingredient of PageRank, the row-normalized adjacency matrix P, which is defined as

$$P = (p_{ij}) = \left(\frac{a_{ij}}{\deg_{out}(i)} \right) \in M_{n\times n}(\mathbb{R}). \tag{2.3}$$

In the theory of Markov processes (i.e., memory-less stochastic processes) this matrix is referred to as the "transition matrix" of the random walker, as its component p_{ij} provides the probability of transitioning from state j to state i. Due to the intrinsic random walk nature of the PageRank algorithm (as discussed in Ref. 19, 20), we will use that notation from now on.

We will denote vectors as $\boldsymbol{v} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$ and the canonical basis of $\mathbb{R}^n$ as $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$. The vector with 1 in all components will be $\boldsymbol{e} = (1, \ldots, 1)^T$. I_n will denote the identity matrix. Finally, we will say that a vector is positive if it is positive components-wise, and we will say that it has unit norm if its 1-norm is equal to 1.

III. STANDARD PAGERANK

The best way to introduce the PageRank algorithm[1] is through the lens of a *random walker with random teleportation*. Let us forget about the teleportation step for a while and consider a random walker on a network G: starting at node i, at each step, it will choose an outlink from those available in its current node, with probability proportional to the weight of each outlink. This is a Markovian process, whose steady state gives a measure of the "centrality" of each node. In other words, the more the walker passes through node i, the more important or central it is.

The PageRank algorithm corresponds to a *personalized* version of this centrality measure, consisting of a biased random walker: with probability α, it will follow the previously described rules of standard random walks, and with probability $1 - \alpha$, it will "teleport"

or "jump" to a random node in the network, with associated probabilities given by a distribution $\boldsymbol{v}$, sometimes called the teleportation vector. The mathematical formulation of this idea is the following:

Definition 3.1 (PageRank vector): *Let G be a graph with no dangling nodes, v a positive, unit norm vector, and $\alpha \in (0,1)$. Then, the PageRank vector of G with damping factor α and personalization vector $\boldsymbol{v}$ is the only positive, unit norm vector (i.e. $\boldsymbol{\pi} > 0$, $|\boldsymbol{\pi}|_1 = 1$) such that satisfying*

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \left(\alpha P + (1-\alpha)\boldsymbol{e}\boldsymbol{v}^T\right), \tag{3.1}$$

where P is the transition matrix of the graph.

Note that existence and uniqueness of $\boldsymbol{\pi}$ are guaranteed by the classic Perron Theorem, as $\alpha P + (1-\alpha)ev^T$ is a positive matrix (see, for example, Refs. 21 and 6).

In what follows, we will restrict ourselves to graphs with no dangling nodes. Were there any, they can be dealt with in the usual way.[22] This does not affect the results discussed here, and we will, thus, omit it for the sake of clarity.

We are interested in understanding the conditions under which an arbitrary stochastic vector can be set to be the PageRank centrality of a given graph. We can state this more formally:

Problem 3.2 (PageRank centrality control): *Can we modify the graph $G = (V, E)$ or the components of the PageRank measure (damping factor or personalization vector) such that an arbitrary positive, unit norm vector $\boldsymbol{\pi}_0$ is the PageRank vector?*

Changing the structure of the graph in some way (adding/removing edges, changing weights) would be considered as a *structural* change, whereas changing the parameters of the PageRank measure, such as the damping factor or its personalization vector, would be a *parametric* change.

In the context of the Eigenvector centrality, it was proven[13] that by a rather mild structural change as changing edge weights, one is able to fully fix the resulting centrality vector at will, so long as the network is directed and strongly connected. In the present case, where we instead deal with the PageRank centrality, things are not that simple due to the row-normalization of the adjacency matrix: the construction of P normalizes out any weight placed on out-edges coming from nodes with out-degree equal to 1. The simplest way to see this is considering directed rings, as in Fig. 1.

We could consider controlling the centrality by means of other types of structural changes, such as adding nodes, rewiring edges, etc. However, those are considerably more drastic modifications and go out of the scope of this paper. Instead, we will now focus on parametric changes, i.e., modifications in the parameters of the centrality measure.

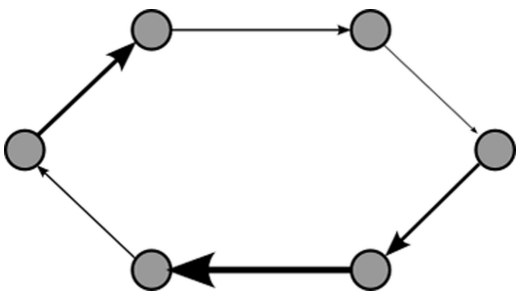

FIG. 1. Simple example of a network (the directed cycle C_6) whose PageRank centrality is unaffected by any modification of edge weights.

A. Constraints on the personalization vector

It is clear that suitable adjustments of the damping factor α and the personalization vector v will be needed in order to fix the PageRank centrality of the given network G (see, for example, Refs. 4 and 12). What we attempt to do is quantifying the balance between the adjustment of both parameters. In other words, we want to understand what ranges of α provide the desired centrality vector for suitable v.

By operating with Eq. (3.1), it is straightforward to obtain the following formula:[4]

$$\boldsymbol{\pi}^T(I_n - \alpha P) = (1-\alpha)\boldsymbol{v}^T. \tag{3.2}$$

Traditionally, this equation can be viewed as an equation for $\boldsymbol{\pi}$ given α, $\boldsymbol{v}$, and P. However, we can also view it as an equation for $\boldsymbol{v}$ given α, $\boldsymbol{\pi}$, and P,

$$\boldsymbol{v}^T = \frac{1}{1-\alpha}\boldsymbol{\pi}^T(I_n - \alpha P). \tag{3.3}$$

This equation tells us which personalization vector is required to obtain a desired PageRank vector for a fixed network and damping factor. This raises a question: can we always find such non-negative personalization vector that gets a prescribed PageRank centrality? This natural question is summarized in the following problem:

Problem 3.3 (Centrality control via personalization vector): *Given a graph G, a damping factor $\alpha \in (0,1)$, and a positive, unit norm vector $\boldsymbol{\pi}_0$, does it always exists a positive, unit norm $\boldsymbol{v}$ such that the $\boldsymbol{\pi}_0$ is the PageRank outcome?*

In other words: can any PageRank vector be set for a given graph and damping factor if we have control over the personalization vector used in the algorithm?

The answer is no, since there is no positive ($v_i > 0, \forall i$) solution in some cases. Nevertheless, we can study the conditions under which $\boldsymbol{\pi}_0$ actually has an associated personalization vector $\boldsymbol{v}$, and the following result give a characterization of the existence of positive personalization vectors that give a prescribed PageRank centrality $\boldsymbol{\pi}_0$ in terms of the size of its components.

Theorem 3.4 (Existence of the personalization vector): *Given a graph G and a positive, unit norm vector $\boldsymbol{\pi}_0$ then there exists a positive, unit norm personalization vector $\boldsymbol{v}$ such that $\boldsymbol{\pi}_0$ is the PageRank vector if and only if $\boldsymbol{\pi}_0^T\boldsymbol{e}_j > \alpha\boldsymbol{\pi}_0^T P\boldsymbol{e}_j$ for all j.*

Proof. First, we prove that Eq. (3.3) leads to unit norm personalization vectors, since

$$|\boldsymbol{v}|_1 = \boldsymbol{v}^T\boldsymbol{e} = \frac{1}{1-\alpha}\boldsymbol{\pi}_0^T(I_n - \alpha P)\boldsymbol{e} = \frac{1}{1-\alpha}\boldsymbol{\pi}_0^T(\boldsymbol{e} - \alpha P\boldsymbol{e}) = \boldsymbol{\pi}_0^T\boldsymbol{e} = |\boldsymbol{\pi}_0|_1 = 1, \tag{3.4}$$

where we used the row-stochasticity in $P\boldsymbol{e} = \boldsymbol{e}$. We now require that all of $\boldsymbol{v}$s components are positive, so

$$v_j = \boldsymbol{v}\boldsymbol{e}_j = \frac{1}{1-\alpha}\boldsymbol{\pi}_0^T(I_n - \alpha P)\boldsymbol{e}_j > 0, \tag{3.5}$$

which completes the proof. □

It is also remarkable to point out that Theorem 3.4 presents some analytical interplay between the damping factor and personalization vectors, since if we take a positive, unit norm $\boldsymbol{\pi}_0$ and $0 < \alpha \leq \min_j \left(\boldsymbol{\pi}_0^T \boldsymbol{e}_j\right)$, then it can be checked that for any graph without dangling nodes there exists a positive, unit norm personalization vector $\boldsymbol{v}$ such that $\boldsymbol{\pi}_0$ is the PageRank vector. In fact, if we consider a graph without dangling nodes, note that $P\boldsymbol{e}_j$ is the jth column of P, that is

$$P\boldsymbol{e}_j = \left(\frac{a_{1j}}{\deg_{out}(1)}, \frac{a_{2j}}{\deg_{out}(2)}, \ldots, \frac{a_{nj}}{\deg_{out}(n)}\right)^T, \tag{3.6}$$

so we have that $\boldsymbol{\pi}_0^T P\boldsymbol{e}_j \leq \boldsymbol{\pi}_0^T \boldsymbol{e} = 1$, since $0 \leq a_{ij}/\deg_{out}(i) \leq 1$, and hence, if we take $\alpha < \min_j \left(\boldsymbol{\pi}_0^T e_j\right)$, then

$$\alpha \boldsymbol{\pi}_0^T P e_j \leq \alpha < \left(\boldsymbol{\pi}_0^T e_j\right), \quad \forall 1 \leq j \leq n; \tag{3.7}$$

hence, there exists a personalization vector $\boldsymbol{v}$ such that $\boldsymbol{\pi}_0$ is the PageRank vector, simply by using Theorem 3.4.

B. The ranking control problem

In this section, we will analyze the centrality control problem by using Theorem 3.4, as seen in Sec. III A.

Centrality measures typically return a list (vector) of centrality scores: numbers between 0 and 1 specifying the importance of each node in the network with respect to the chosen measure. However, for most applications, the actual score of a node is not relevant; instead what matters is its relative position with respect to the rest of the nodes. In other words, the ranking of nodes based on their centrality.

The subject of ranking control has remained fairly unexplored due to its technical complexity (as lifting the constraint of fixing concrete centrality vectors makes the problem harder to tackle), but in the PageRank case, Theorem 3.4 provides us with a valuable tool to investigate in this direction by using some techniques from convex geometry.

Consider the following milder version of Problem 3.3, where we are now only interested in rankings rather than concrete PageRank vectors.

Problem 3.5 (Ranking control via personalization vector): *Given a graph G, a damping factor $\alpha \in (0,1)$ and an ordering of the nodes (allowing for ties), does it always exists a positive, unit norm personalization vector $\boldsymbol{v}$ such that the PageRank outcome follows the prescribed order?*

In order to study this problem we will now change the viewpoint of the discussion to a geometric one: consider the n-simplex defined as

$$\Delta_n = \{\boldsymbol{x} \in \mathbb{R}^n, \quad \text{such that} \quad \boldsymbol{x} > 0,\ |\boldsymbol{x}|_1 = 1\}. \tag{3.8}$$

This set represents the convex span of vectors $\{e_1, \ldots, e_n\}$, and thus, it is the space of all possible personalization vectors and the space of all possible PageRank vectors of graphs with n nodes. Therefore, we can understand Eq. (3.2) as the following map from Δ_n to itself:

$$\begin{aligned} \boldsymbol{\pi}(G,\alpha,\cdot): \quad & \Delta_n \longrightarrow \Delta_n \\ & \boldsymbol{v} \longmapsto \boldsymbol{\pi}(G,\alpha,\boldsymbol{v}). \end{aligned} \tag{3.9}$$

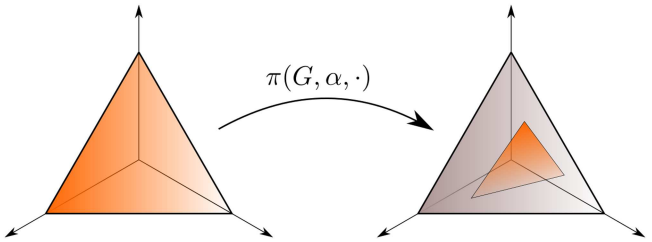

FIG. 2. Depiction of map $\pi(G,\alpha,\cdot)$ for $n = 3$.

This map is injective (however, in general, it is not surjective) and linear in $\boldsymbol{v}$, so $\boldsymbol{\pi}(G,\alpha,\Delta_n)$ is a polytope (i.e., the convex hull of a finite number of points) in $\Delta_n \subset \mathbb{R}^n$. Figure 2 illustrates this geometrical interpretation of $\boldsymbol{\pi}(G,\alpha,\cdot)$ in case $n = 3$.

The key point in this geometric viewpoint is that we can associate each possible ranking to a portion of the simplex. If we consider the center point (barycenter) of the simplex Δ_n, given by the normalization of $\boldsymbol{e}$, i.e., $\boldsymbol{e}_0 \equiv \boldsymbol{e}/n = \sum_{i=1}^n \boldsymbol{e}_i/n$, then we can define the hyperplanes bisecting the simplex through the center $\boldsymbol{e}_0$ and any combination of $n-2$ vertices as

$$\mathcal{H}_n^{i,j} = \left\{ \sum_{\substack{k=0 \\ k \neq i,j}}^{n} \lambda_k \boldsymbol{e}_k, \quad \text{such that} \quad \lambda_k \in \mathbb{R} \right\} \subseteq \mathbb{R}^n. \tag{3.10}$$

The relevance of this construction is that it provides us with a way to classify the points $\boldsymbol{\pi} \in \Delta_n$ according to their ranking. To see this, consider, for instance, the hyperplane $\mathcal{H}_4^{1,2}$. It can be identified as the region of ranking space where $c_1 = c_2$, by definition. If we move away from it in the direction of $\boldsymbol{e}_2$, we will have $c_1 < c_2$ and viceversa.

In general, the $\binom{n}{2}$ planes $\mathcal{H}_n^{i,j}$ uniquely determine the pairwise inequalities between components i,j of the PageRank vector. The original simplex Δ_n is then divided into $n!$ regions (the number of permutations of the components of the PageRank vector), each of them determining a different ranking. A depiction of these regions for the $n = 3$ case can be seen in Fig. 3.

In this light, we can see that there is Ranking control if and only if

$$\boldsymbol{e}_0 = \frac{1}{n}\boldsymbol{e} \in \mathrm{Im}(\boldsymbol{\pi}) \quad \text{and} \quad \boldsymbol{e}_0 = \frac{1}{n}\boldsymbol{e} \notin \partial \mathrm{Im}(\boldsymbol{\pi}). \tag{3.11}$$

The argument here is identical to that of the hyperplanes: $\boldsymbol{\pi} = \boldsymbol{e}_0$ is the point in ranking space where $c_1 = c_2 = \cdots = c_n$. Given that all hyperplanes $\mathcal{H}_n^{i,j}$ pass through $\boldsymbol{e}_0$ by construction, all ranking regions are $\varepsilon > 0$ away from it. Thus, moving $\varepsilon > 0$ away in any direction will lead to different rankings.

This idea may be easier to visualize if we take into account Fig. 2. Notice that in that case, the resulting triangle (right) contains points associated to any ranking (as shown in Fig. 3). The condition necessary and sufficient for this to happen is for it to contain the centerpoint of the bigger triangle.

Next, we can give an analytical characterization of the existence of a prescribed ranking of nodes in terms of the relationship between

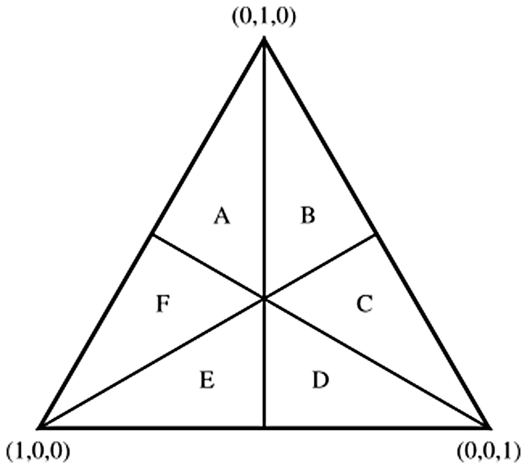

FIG. 3. Different ranking regions in the $n = 3$ case. For instance, if ∂A denotes the (topological) boundary of $A \subseteq \Delta_n$ in the containing plane of A, then $\pi = (\pi_1, \pi_2, \pi_3) \in A \setminus \partial A$ corresponds to $\pi_2 > \pi_1 > \pi_3$, while the intersection between triangles would lead to equal scores, e.g., $\pi \in B \cap C$ would correspond to $\pi_2 = \pi_3 > \pi_1$.

the damping factor and the column sums of P, which is analogous to Theorem 3.4 but for the ranking problem.

Theorem 3.6 (Characterization of ranking control): *Given a graph G and damping factor $\alpha = (0, 1)$, then it is possible to obtain any ranking of the nodes under the PageRank if and only if*

$$\frac{1}{\alpha} > \max_j \left(\sum_{i=1}^{n} P_{ij} \right). \tag{3.12}$$

Proof. Using $\boldsymbol{\pi}_0 = \boldsymbol{e}_0$ in Theorem 3.4 yields

$$\boldsymbol{e}_0^T \boldsymbol{e}_j = \frac{1}{n} > \alpha \boldsymbol{e}_0^T P \boldsymbol{e}_j = \frac{1}{n} \alpha \sum_{i=1}^{n} P_{ij}, \tag{3.13}$$

for all $1 \leq j \leq n$, which already gives us the characterization of the existence of a personalization vector that gives any prescribed ranking of nodes. By virtue of the aforementioned theorem, we can also conclude it to be a sufficient condition for the existence of a personalization vector allowing for any desired ranking. □

Given that $\sum_i P_{ij}$ is the total probability that a random walker visits node j, this theorem can be interpreted as an upper bound for α in terms of the maximum of these total probabilities. This upper bound tells us that, provided we have $\alpha < 1/\max_j \sum_i P_{ij}$, we can always find any desired ranking with an appropriate choice of personalization vector. It is important to note that this is not a statistical result, in the sense that as long as there is one node targeted by many others with low out-degree, there will be almost no room for ranking control, regardless of the topology of the rest of the network. As we will see later, this is very reminiscent of the scale-free[23] network paradigm: indeed, scale-free networks present these high in-degree nodes pointed to by low out-degree ones.

It is also remarkable to point out the fact that if we denote

$$\alpha_0 = \frac{1}{\max_j \sum_i P_{ij}}, \tag{3.14}$$

then $\alpha_0 \in (0, 1]$ is a measure of the *controllability* of the PageRank in graph G, since the bigger α_0 is the wider range of damping factors allow Ranking control of PageRank in G.

C. Real network datasets

Having found a network-specific upper bound for the value of the damping factor α, which would allow the PageRank of the network to be ranking-controllable tinkering with the personalization vector, it is left for us to find out whether it is a hard or soft bound.

The standard value considered for the damping is $\alpha = 0.85$,[6] whose interpretation in terms of Internet hyperlink networks is that of a surfer clicking on hyperlinks ∼8 times before losing interest and searching for something else; this value corresponds to constraining the maximum of the column sum of P to around 1.17. This is clearly a very strict condition.

In fact, we have computed the maximum of the column sums of P for a variety of networks,[24] publicly available from different Internet sources (all fetch from the KONECT network repository[25] and the CASOS network repository[26]). We can extract the maximum value of the damping factor α which would enable us to have ranking control over each network's PageRank rankings. This is shown in Fig. 4.

As expected from the above discussion, the maximum values of the damping factor are generally small compared to the standard $\alpha = 0.85$, regardless of the network size. There are a couple of exceptionally high values but still lower than such value. We see, on the other hand, that the smaller the network the more controllable it is. This can also be understood from Theorem 3.6: a higher number of nodes means that the maximum column sum of P is likely to be higher (specially due to the number of edges growing also linearly with the number of nodes), hindering controllability.

While our results are of a theoretical nature and, thus, are not related to any specific implementation or application of PageRank, it might also be interesting to address the implications of this bound in some specific use cases of PageRank (and more concretely, understanding the teleportation vector in them).

- World Wide Web and similar data: Here, the purpose of PageRank is mainly identifying websites of interest for a given user. The teleportation, therefore, allows for tweaking the preferences of the user, providing different rankings to a user with a different personalization vector. The bound (3.12) in this case tells us that the ordering is robust: the ranking cannot be completely altered by the choice of personalization.
- Genetical or Protein–Protein Interaction networks: As explained in Ref. 7, PageRank has been applied in a variety of biological networks.[27,28] In these applications, the teleportation vector is designed to focus the search on specific areas of the network. Given that, as discussed in the aforementioned paper, the damping used in these applications is high ($\alpha \approx 0.8$), we can also to conclude that the ranking will also be robust with respect to changes in the personalization vector.

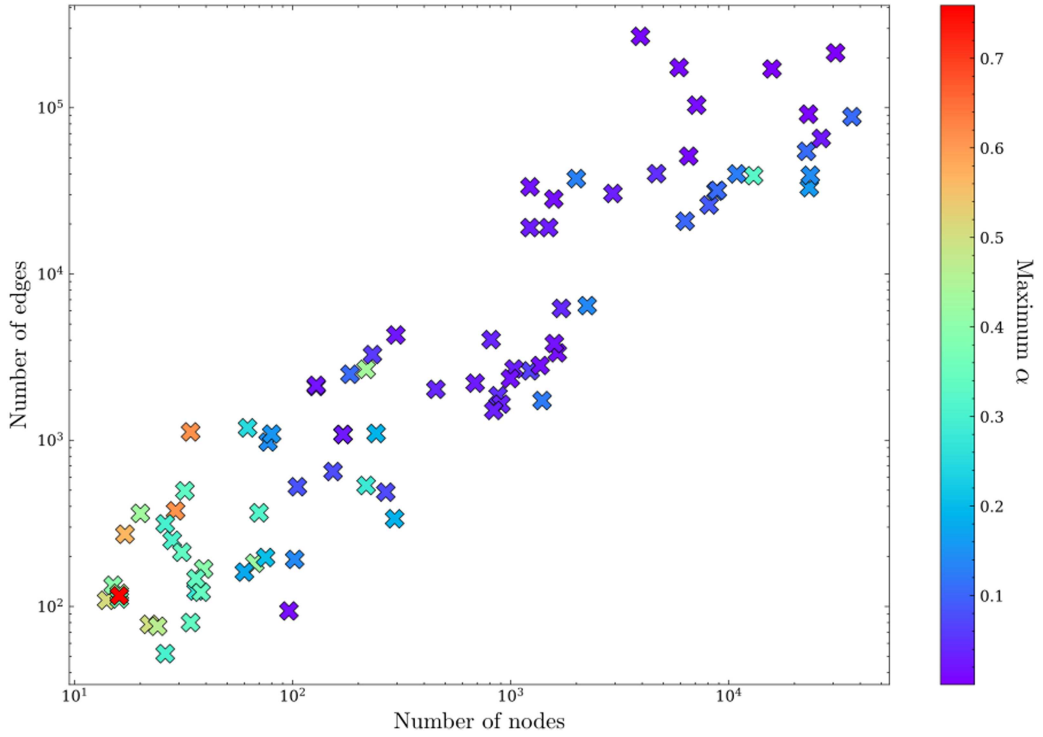

FIG. 4. Scatterplot showing the number of edges against the number of nodes for 84 different real networks obtained from the KONECT network repository[25] and the CASOS network repository,[26] with datapoints colored based on the maximum value of α providing ranking control.

- Knowledge information systems: It is also discussed in Ref. 7 that PageRank has found a way to be used in semi-supervised learning tasks (for instance, in graphs where each node is an image and two are connected if they share a caption label[29]). These studies present a different phenomenology to the previously discussed cases, as they employ very low values of $\alpha \approx 0.1$. Although we have no datasets of this type, our results force us to conclude that it is quite likely that the rankings obtained will be highly dependent on the personalization vector used.
- There are plenty other applications (see Ref. 7), most of them using high α. We can draw similar conclusions to the previous cases, the robustness of the ranking.

Some of the datasets used in Fig. 4 fall into these or other categories. The specific data used in each of them can be found in our GitHub repository (which can be found in the Data Availability section of this article), and the information about each dataset is in Ref. 25 and 26.

IV. BIPLEX PAGERANK

In Ref. 18, a novel version of the PageRank vector was put forward by establishing an analogy between the standard PageRank algorithm and a random surfer on a "virtual" biplex network, constructed from the initial graph G.

Although we will not discuss it here, this algorithm was shown to be useful in order to extend the notion of PageRank centralities to multiplex networks. Multiplex networks are networks where the interactions between nodes fall into different categories.[30] Hence, they can be represented as different layers, each containing the same nodes but with a particular set of connections. Standard complex network algorithms (such as centrality measures, community detection algorithms, and others) need to be extended to account for these more intricate structures, and the biplex PageRank[18] is one of such proposals.

Nevertheless, the application of this algorithm to monoplex networks provides yet another extension of PageRank, which can serve as a playground for novel ideas related to centrality. In our case, it will be clear that the geometrical solution to the ranking problem described in Sec. III B is not restricted to just the vanilla PageRank algorithm: it can serve as a guiding principle in more complicated, although related, measures.

In the biplex PageRank algorithm, the auxiliar biplex network considered consists of two layers: one with the actual edge connections between the n nodes (this layer essentially accounts for the teleportation-less random walk), while the other contains a fully connected graph between them (this is the "teleportation layer"). The biased random walker with teleportation then chooses, at each step, whether to follow the links in the usual transition layer or the teleportation layer.

This construction led to the following definition:

Definition 4.1 (Biplex PageRank centrality[18]): *Let G be a graph with no dangling nodes, with transition matrix P. Let $\mathbf{v}$ be a*

positive, unit norm vector and $\beta \in (0,1)$. Then, the biplex PageRank vector of G with damping factor β and personalization vector $\boldsymbol{v}$ is the vector,

$$\boldsymbol{\pi}_{\mathrm{BPR}} = \boldsymbol{\pi}_u + \boldsymbol{\pi}_d, \tag{4.1}$$

where $[\boldsymbol{\pi}_u^T, \boldsymbol{\pi}_d^T] \in \mathbb{R}^{2n}$ is the only positive, unit norm eigenvector of

$$M_{\mathrm{BPR}} = \begin{pmatrix} \beta P & (1-\beta) I_n \\ \beta I_n & (1-\beta)\boldsymbol{e}\boldsymbol{v}^T \end{pmatrix}. \tag{4.2}$$

Note that π_u corresponds to the centrality of the nodes in the transition layer, while π_d corresponds to the centrality of the nodes in the teleportation layer.

It is remarkable to point out that existence and uniqueness of the Biplex PageRank centrality are granted by the Perron–Frobenius theorem. This alternative version of the biased walker leads to a different centrality measure, whose technical details we will skip, only keeping the necessary ones and referring the interested reader to Ref. 18 for them.

Vectors $\boldsymbol{\pi}_u$ and $\boldsymbol{\pi}_d$ satisfy the following relations: $\boldsymbol{\pi}_u \boldsymbol{e} = \beta$ and $\boldsymbol{\pi}_d \boldsymbol{e} = 1-\beta$.

Later in Ref. 31, a closed form formula for the Biplex PageRank vector, in resemblance to formula (3.2), was found as

$$\boldsymbol{\pi}_{\mathrm{BPR}}^T = (1-\beta)^2 \boldsymbol{v}^T (\beta I_n + Y)\, Z^{-1}, \tag{4.3}$$

where $Y = I_n - \beta P$, $Z = \gamma I_n - \beta P$, and $\gamma = 1 - \beta(1-\beta)$. It is straightforward to check that $(\beta I_n + Y)$ is invertible in the $\beta \in (0,1)$ range, so we also have the formula

$$\boldsymbol{v}^T = \frac{1}{(1-\beta)^2} \boldsymbol{\pi}_{\mathrm{BPR}}^T Z (\beta I_n + Y)^{-1}. \tag{4.4}$$

With this, we can state the following theorem that characterizes when a personalization vector exists for a prescribed biplex PageRank centrality.

Theorem 4.2 (Existence of the personalization vector–biplex case): *Given a graph G and a positive, unit norm $\boldsymbol{\pi}_{BPR}$, then there exists a positive, unit norm personalization vector $\boldsymbol{v}$ such that $\boldsymbol{\pi}_0$ is the biplex PageRank vector if and only if $\boldsymbol{\pi}_{BPR}^T \boldsymbol{e}_j > \beta \boldsymbol{\pi}_{BPR}^T \mathcal{P} \boldsymbol{e}_j$ for all j, where $\mathcal{P} = (2-\beta)(\beta I_n + Y)^{-1}$.*

Proof. First we prove that Eq. (4.4) leads to unit-norm personalization vectors. Note that $P^n \boldsymbol{e} = \boldsymbol{e}$ due to row-stochasticity; therefore, if we use the resolvent expansion

$$(\beta I_n + Y)^{-1} = \frac{1}{\beta+1} \sum_{m=0}^{\infty} \left(\frac{\beta}{1+\beta} P \right)^m,$$

we end up with

$$\begin{aligned} |\boldsymbol{v}|_1 = \boldsymbol{v}^T \boldsymbol{e} &= \frac{1}{(1-\beta)^2} \boldsymbol{\pi}_{\mathrm{BPR}}^T Z (\beta I_n + Y)^{-1} \boldsymbol{e} \\ &= \frac{1}{(1-\beta)^2} \boldsymbol{\pi}_{\mathrm{BPR}}^T Z \frac{1}{\beta+1} \sum_{m=0}^{\infty} \left(\frac{\beta}{1+\beta} P \right)^m e \\ &= \frac{1}{(1-\beta)^2} \boldsymbol{\pi}_{\mathrm{BPR}}^T (\gamma I_n - \beta P) \boldsymbol{e} \frac{1}{\beta+1} \sum_{m=0}^{\infty} \left(\frac{\beta}{1+\beta} \right)^m \\ &= \frac{1}{(1-\beta)^2} \boldsymbol{\pi}_{\mathrm{BPR}}^T (1-\beta)^2 \boldsymbol{e} = \boldsymbol{\pi}_{\mathrm{BPR}}^T \boldsymbol{e} = 1. \end{aligned}$$

We now require that all components of the required personalization vector are positive,

$$v_j = \boldsymbol{v}^T \boldsymbol{e}_j = \frac{1}{(1-\beta)^2} \boldsymbol{\pi}_{\mathrm{BPR}}^T Z (\beta I_n + Y)^{-1} \boldsymbol{e}_j > 0. \tag{4.5}$$

It will now be convenient expanding the $Z(\beta I_n + Y)^{-1}$ expression in with the previously mentioned resolvent series, multiplying and re-summing. Doing so, we find

$$\begin{aligned} & Z(\beta I_n + Y)^{-1} \\ &= \frac{1}{\beta+1} (\gamma I_n - \beta P) \sum_{m=0} \left(\frac{\beta}{\beta+1} \right)^m P^m \\ &= \frac{1}{\beta+1} \sum_{m=0} \left[\gamma \left(\frac{\beta}{\beta+1} \right)^m P^m - \beta \left(\frac{\beta}{\beta+1} \right)^m P^{m+1} \right] \\ &= \frac{1}{\beta+1} \left[\gamma I_n - \beta(\beta-2) I_n + \beta(\beta-2) \sum_{m=0}^{\infty} \left(\frac{\beta}{\beta+1} \right)^m P^m \right] \\ &= I_n + \beta(\beta-2)(\beta+Y)^{-1}. \end{aligned}$$

Plugging this in the above equation, we find the condition

$$\left[\boldsymbol{\pi}_{\mathrm{BPR}} + \beta(\beta-2) \boldsymbol{\pi}_{\mathrm{BPR}} (\beta I_n + Y)^{-1} \right] \boldsymbol{e}_j > 0, \tag{4.6}$$

which with the identification $\mathcal{P} = (2-\beta)(\beta I_n + Y)^{-1}$ concludes the proof. □

Again, by using the geometric approach proposed in Sec. III B, we can interpret the biplex PageRank vector as the linear map between simplices (3.8),

$$\begin{aligned} \boldsymbol{\pi}_{\mathrm{BPR}}(G, \beta, \cdot): \quad & \Delta_n \longrightarrow \Delta_n \\ & \boldsymbol{v} \longmapsto \boldsymbol{\pi}_{\mathrm{BPR}}(G, \beta, \boldsymbol{v}). \end{aligned} \tag{4.7}$$

This map is again injective and linear in $\boldsymbol{v}$ and, consequently, allows us to employ the same kind of argument for the existence of ranking controllability,

$$\boldsymbol{e}_0 = \frac{1}{n} \boldsymbol{e} \in \mathrm{Im}(\boldsymbol{\pi}_{\mathrm{BPR}}), \quad \boldsymbol{e}_0 = \frac{1}{n} \boldsymbol{e} \notin \partial \mathrm{Im}(\boldsymbol{\pi}_{\mathrm{BPR}}). \tag{4.8}$$

It is straightforward to find an analytic characterization of ranking control in the biplex PageRank case in terms of the relationship between β and the column sums of matrix P, simply by following the same reasoning used in the standard PageRank setting. In fact, following similar arguments that those used in the proof of Theorem 3.6, it can be easily proved the following result:

Theorem 4.3 (Characterization of biplex ranking control): *Given a graph G and a damping factor $\beta = (0,1)$, then it is possible to obtain any ranking of the nodes under the biplex PageRank if and only if*

$$\frac{1}{\beta} > \max_j \left(\sum_{i=1}^{N} \mathcal{P}_{ij} \right). \tag{4.9}$$

By using the definition of $\mathcal{P}$, the condition that appears in Theorem 4.3 can be rewritten as

$$\frac{1}{\beta} > \max_j \left(\sum_{i=1}^{N} \mathcal{P}_{ij} \right) = \left(\frac{2-\beta}{1+\beta} \right) \max_j \sum_{i=1}^{N} \left[\left(I_n - \frac{\beta}{1+\beta} P \right)^{-1} \right]_{ij}, \tag{4.10}$$

but we cannot expect a more simplified expression of the maximal β in terms of P_{ij} since matrix $\mathcal{P}$ depends itself on the damping factor, unlike what happened in the standard PageRank case.

A. Comparison to the monoplex result

After presenting the analytic result for the biplex ranking problem (Theorem 4.3), we now turn to numerics in order to develop some understanding for this result. In particular, we are interested in how it compares to the usual PageRank, whether it is *more controllable* or not, in terms of the maximal damping factor that allows full ranking control. A similar comparative analysis was performed in terms of the controllability based on the personalization vector between the (classic) PageRank and the Biplex PageRank by Flores *et al.*[32]

As we have pointed out before, Theorem 3.6 shows that the value α_0 introduced in Eq. (3.14) is a measure of the controllability of the PageRank in G. Similarly, if we consider β_0 the maximal value that verifies Eq. (4.10), then it is also a measure of the controllability of the Biplex PageRank in G since the bigger β_0 is the wider range of damping factors that allow Biplex Ranking control of PageRank in G. A numerical comparison between α_0 and β_0 for the same real network datasets used in the standard PageRank case (all fetch from the KONECT network repository[25] and the CASOS network repository[26]) is presented in Fig. 5. Note that, for most cases, the maximal value of the damping factor α_0 is smaller than the corresponding maximal β_0 for the Biplex PageRank, so we see that in these cases Biplex PageRank is more controllable than (classic) PageRank, which is consistent with the results obtained in Ref. 32 for the controllability related to personalization vectors.

It is also interesting to point out that although the datasets come from very heterogeneous sources, there is a clear tendency in the data, following a curve that we found to be (via a quadratic polynomial fit) $y = 1.014x^2 + 0.492x - 0.041$. This is perhaps more surprising when we take into account that some of the sampled networks are weighted, yet the quadratic behaviour remains unchanged.

In order to delve deeper in this result, we will consider another batch of network data, this time synthetic networks. We have generated, with the aid of the NetworkX library in Python, two distinct sets of networks: some directed random networks (constructed in the same vein as the undirected Erdös–Renyi version) and some directed scale-free ones (constructed based on the procedure prescribed in Ref. 33). In both cases, we generated networks with the number of nodes ranging from 100 to 20 000, with different edge creation probabilities (see the GitHub repository for specific details

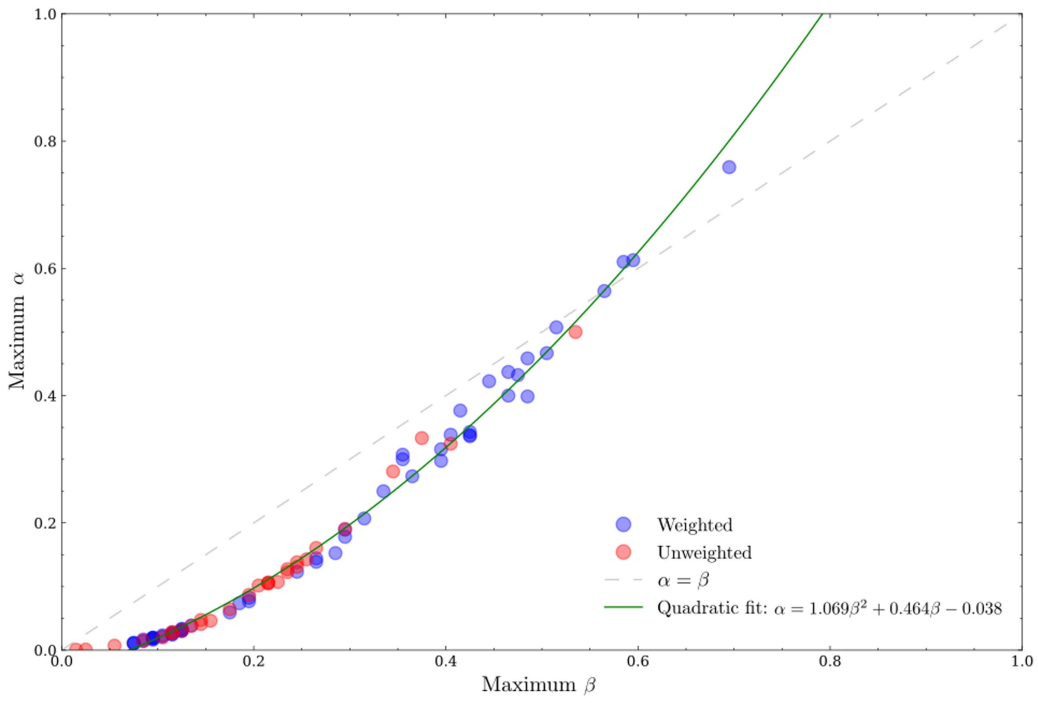

FIG. 5. Comparison between maximum α and β saturating their respective bounds in the cases of Standard and Biplex PageRank for 84 different real networks obtained from the KONECT network repository[25] and the CASOS network repository.[26] Red datapoints represent weighted networks and blue represents unweighted ones.

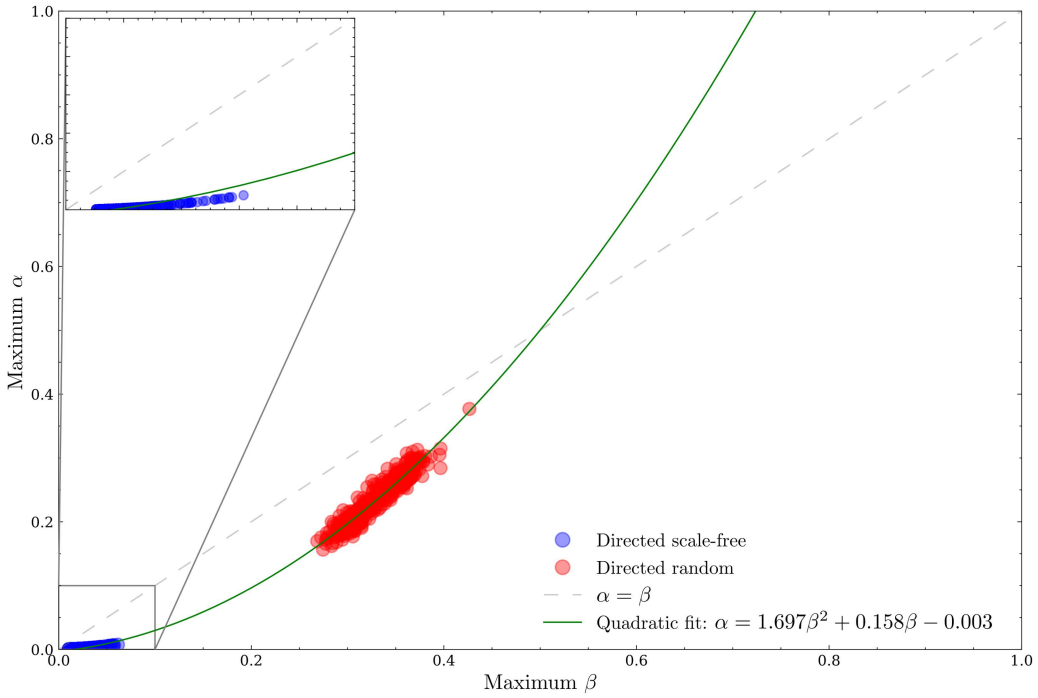

FIG. 6. Comparison of maximum α and β saturating their respective bounds in the cases of Standard and Biplex PageRank for synthetic networks. Here, red datapoints represent random networks and blue represents scale-free ones.

of implementation). We again compute their maximum values of α, β and plot one against the other, obtaining Fig. 6.

We can clearly see that the polynomial fit is very similar to that of the real networks case, and we could expect them to match even better had we sampled more networks. Apart from that, it is quite noticeable that the distinct nature of the generated networks also separates them in their behavior with respect to both centrality measures. This was already hinted at in Sec. III: the presence of high in-degree nodes pointed to by low out-degree ones is very common in scale-free networks; thus, their maximum values of α and β are specially low. For random networks, all nodes have, on average, the same connectivity; thus, they allow for more flexibility in ranking control. Another byproduct of the randomness in the corresponding synthetic networks is the higher spread in α for fixed β compared to that of scale-free ones.

V. CONCLUSIONS

Our research has focused on the controllability of the PageRank algorithm as a centrality measure in complex networks. Through our study, we have concluded that full control through weight changes is impossible. Instead, we have investigated the necessary conditions to achieve full control through parametric changes, which involve modifying both the damping factor and the personalization vector. By shifting our focus to centrality rankings rather than centrality scores, we found that a less stringent requirement is sufficient for both standard PageRank and biplex PageRank. However, when we tested this condition on real or synthetic networks, we found it to be a challenging constraint. These findings offer further evidence of the stability of PageRank as an indexing tool.

ACKNOWLEDGMENTS

This work was partially supported by Project Nos. PGC2018-101625-B-I00 (Spanish Ministry, AEI/FEDER, UE), M1993, M2978, and M3033 (URJC Grants). G.C.-A. was funded by the URJC fellowship No. PREDOC-21-026-2164.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Gonzalo Contreras-Aso: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Regino Criado:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Miguel Romance:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in Github at https://github.com/LaComarca-Lab/PageRank_CentralityControl, Ref. 34.

REFERENCES

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," in *Proceedings of the 7th International World Wide Web Conference* (Elsevier, 1998), pp. 161–172.
[2] M. Bianchini, M. Gori, and F. Scarselli, "Inside PageRank," ACM Trans. Int. Tech. **5**, 92–128 (2005).
[3] A. Langville and C. Meyer, "Deeper inside PageRank," Internet Math. **1**, 335–380 (2004).
[4] P. Boldi, M. Santini, and S. Vigna, "PageRank as a function of the damping factor," in *Proceedings of the 14th International Conference on World Wide Web* (Association for Computing Machinery, New York, NY, 2005), pp. 557–566.
[5] P. Boldi, M. Santini, and S. Vigna, "PageRank: Functional dependencies," ACM Trans. Inf. Syst. **27**, 19 (2009).
[6] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond* (Princeton University Press, 2006).
[7] D. F. Gleich, "PageRank beyond the web," SIAM Rev. **57**, 321–363 (2015).
[8] D. Chaffey, C. Lake, and A. Friedlein, *Search Engine Optimization—Best Practice Guide* (Econsultancy.com Ltd, 2009).
[9] J. L. Ledford, *Search Engine Optimization Bible* (John Wiley & Sons, 2015), Vol. 584.
[10] M. R. Henzinger, "Hyperlink analysis for the web," IEEE Internet Comput. **5**, 45–50 (2001).
[11] P. R. Olsen, "A future in directing online traffic," The New York Times 10 (2009).
[12] E. García, F. Pedroche, and M. Romance, "On the localization of the personalized PageRank of complex networks," Linear Algebra Appl. **439**, 640–652 (2013).
[13] V. Nicosia, R. Criado, M. Romance, G. Russo, and V. Latora, "Controlling centrality in complex networks," Sci. Rep. **2**, 218 (2012), arXiv:1109.4521.
[14] K. Avrachenkov and N. Litvak, "The effect of new links on Google PageRank," Stoch. Models **22**, 319–331 (2006).
[15] C. de Kerchove, L. Ninove, and P. van Dooren, "Maximizing PageRank via outlinks," Linear Algebra Appl. **429**, 1254–1276 (2008).
[16] M. Olsen, "Maximizing PageRank with new backlinks," in *Algorithms and Complexity*, edited by T. Calamoneri and J. Diaz (Springer Berlin Heidelberg, 2010), pp. 37–48.
[17] V. Carchiolo, M. Grassia, A. Longheu, M. Malgeri, and G. Mangioni, "Long distance in-links for ranking enhancement," in *Intelligent Distributed Computing XII*, edited by J. Del Ser, E. Osaba, M. N. Bilbao, J. J. Sanchez-Medina, M. Vecchio, and X.-S. Yang (Springer International Publishing, Cham, 2018), pp. 3–10.
[18] F. Pedroche, M. Romance, and R. Criado, "A biplex approach to PageRank centrality: From classic to multiplex networks," Chaos **26**, 065301 (2016).
[19] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," Proc. Natl. Acad. Sci. U. S. A. **105**, 1118–1123 (2008).
[20] R. Lambiotte and M. Rosvall, "Ranking and clustering of nodes in networks with smart teleportation," Phys. Rev. E **85**, 056107 (2012).
[21] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra* (SIAM, 2001).
[22] When there are dangling nodes involved, one can resort to the standard trickery of substituting $P \to P + \boldsymbol{d}^T \boldsymbol{u}$, where $\boldsymbol{d} \in \mathbb{R}^n$ is the distribution of dangling nodes and $\boldsymbol{u} \in \mathbb{R}^n$ is the distribution of imposed outgoing links from them (see, for instance, Ref. 12).
[23] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," Science **286**, 509–512 (1999), arXiv:cond-mat/9910332.
[24] It should be noted that in order to perform this computation we had to deal with the issue of the dangling nodes, which we glossed over at the beginning of this section. We deemed adding a single, random connection from each dangling node to another, non-danging node as the simplest, least intrusive way to remove this issue.
[25] J. Kunegis, "KONECT—The Koblenz Network Collection," in *Proceedings of the Internatioinal Conference on World Wide Web Companion* (International World Wide Web Conferences Steering Committee, 2013), pp. 1343–1350.
[26] See http://www.casos.cs.cmu.edu/tools/data2.php for "Casos network datasets."
[27] J. Morrison, R. Breitling, D. Higham, and D. Gilbert, "Generank: Using search engine technology for the analysis of microarray experiments," BMC Bioinf. **6**, 233 (2005).
[28] V. Freschi, "Protein function prediction from interaction networks using a random walk ranking algorithm," in *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering* (IEEE, 2007), pp. 42–48.
[29] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04 (Association for Computing Machinery, New York, NY, 2004), pp. 653–658.
[30] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," Phys. Rep. **544**, 1–122 (2014).
[31] F. Pedroche, E. García, M. Romance, and R. Criado, "Sharp estimates for the personalized Multiplex PageRank," J. Comput. Appl. Math. **330**, 1030–1040 (2018).
[32] J. Flores, E. García, F. Pedroche, and M. Romance, "Parametric controllability of the personalized PageRank: Classic model vs biplex approach," Chaos **30**, 023115 (2020).
[33] B. Bollobás, C. Borgs, J. T. Chayes, and O. Riordan, "Directed scale-free graphs," in *ACM-SIAM Symposium on Discrete Algorithms* (Society for Industrial and Applied Mathematics, 2003).
[34] G. Contreras-Aso, R. Criado, and M. Romance (2022). "PageRank centrality control," GitHub. https://github.com/LaComarca-Lab/PageRank_CentralityControl.