Traditional PageRank Versus Network Capacity Bound

Robert A. Kłopotek[1](✉) and Mieczysław A. Kłopotek[2]

[1] Faculty of Mathematics and Natural Sciences, School of Exact Sciences, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland
r.klopotek@uksw.edu.pl

[2] Computer Science Fundamental Research Institute, Polish Academy of Sciences, Warsaw, Poland
mieczyslaw.klopotek@ipipan.waw.pl

Abstract. In a former paper [10] we simplified the proof of a theorem on personalized random walk that is fundamental to graph nodes clustering and generalized it to bipartite graphs for a specific case where the probability of random jump was proportional to the number of links of "personally preferred" nodes. In this paper, we turn to the more complex issue of graphs in which the random jump follows a uniform distribution.

Keywords: Bipartite graphs · PageRank · Uniform jump probability · Flow limits · Graph mining

1 Introduction

The PageRank is widely used as a (primary or supplementary) measure of the importance of a web page since its publication in [15]. Subsequently, the idea was explored with respect to methods of computation [3], application areas (web page ranking, client and seller ranking, clustering, classification of web pages, word sense disambiguation, spam detection, detection of dead pages etc.) and application related variations (personalized PageRank, topical PageRank, Ranking with Back-step, Query-Dependent PageRank, Lazy Walk Pagerank etc.), [11].

The traditional PageRank reflects the probability that a random walker reaches a given webpage. The walker, upon entering a webpage, follows with uniform probability one of the outgoing edges unless he gets bored or there are no outgoing edges. If so, he jumps to any web page with uniform probability.

One of the application areas of PageRank is the creation of new clustering methods especially for graphs, including undirected[1] graphs in which we are interested in this paper. One of the clues for clustering of graphs assumes that a good cluster has low probability to be left by a random walker. Though the concept seems to be plausible, it has been investigated theoretically only for

[1] Unoriented graphs have multiple applications as a means to represent relationships spanned by a network of friends, telecommunication infrastructure or street network.

J. Li et al. (Eds.): ADMA 2019, LNAI 11888, pp. 236–249, 2019.
https://doi.org/10.1007/978-3-030-35231-8_17

a very special case of a random walker (different from the traditional walker), performing the "boring jump" with the probability being proportional to the number of incident edges (and not uniformly) – see e.g. [4,10].

In this paper, we will make an attempt to extend this result to the case when the "boring jump" is performed uniformly (as in case of traditional walker) (Sect. 2 with some variants described in Sect. 3) and to generalize it to bipartite graphs (Sect. 4).

PageRank computation for bipartite graphs was investigated already in the past in the context of social networks, e.g. when concerning mutual evaluations of students and lecturers [13], reviewers and movies in a movie recommender systems, or authors and papers in scientific literature or queries and URLs in query logs [7], recommendations [9], food chain analysis [1], species ranking [8], economy [16], social net analysis [6], or performing image tagging [2]. Akin algorithms like HITS were also generalized for bipartite graphs, [14]. As pointed at in [10], the bipartite graphs have a periodic structure explicitly while PageRank aims at graph aperiodicity. Therefore a suitable generalization of PageRank to a bipartite structure is needed and we will follow here the proposals made in [10].

2 Traditional PageRank

One of the many interpretations of PageRank views it as the probability that a knowledgeable (knowing addresses of all the web pages) but mindless (choosing next page to visit without regard to any content hints) random walker will encounter a given web page. So upon entering a particular web page, if it has no outgoing links, the walker jumps to any web page with uniform probability. If there are outgoing links, he chooses with uniform probability one of the outgoing links and goes to the selected web page, unless he gets bored. If he gets bored (which may happen with a fixed probability ζ on any page), he jumps to any web page with uniform probability. One of the modifications of this behavior (called personalized PageRank) was a mindless page-u-fan random walker who is doing exactly the same, but in case of a jump out of boredom he does not jump to any page, but to the page u. Also, there exist plenty of possibilities of other mindless walkers between these two extremes. An unacquainted reader is warmly referred to [12] for a detailed treatment of these topics.

Let us recall the formalization of these concepts. With $\mathbf{r}$ we will denote a (column) vector of ranks: r_j will mean the PageRank of page j. All elements of $\mathbf{r}$ are non-negative and their sum equals 1.

Let $\mathbf{P} = [p_{ij}]$ be a matrix such that if there is a link from page j to page i, then $p_{i,j} = \frac{1}{outdeg(j)}$, where $outdeg(j)$ is the out-degree of node j[2]. In other words, $\mathbf{P}$ is column-stochastic matrix satisfying $\sum_i p_{ij} = 1$ for each column j. If a node had an out-degree equal 0, then prior to construction of $\mathbf{P}$ the node is replaced by one with edges outgoing to all other nodes of the network. Hence

[2] For some versions of PageRank, like TrustRank $p_{i,j}$ would differ from $\frac{1}{outdeg(j)}$ giving preferences to some outgoing links over the other. We are not interested in such considerations here.

$$\mathbf{r} = (1-\zeta)\cdot\mathbf{P}\cdot\mathbf{r} + \zeta\cdot\mathbf{s} \tag{1}$$

where $\mathbf{s}$ is the so-called "initial" probability distribution (i.e. a column vector with non-negative elements summing up to 1) that is also interpreted as a vector of web page preferences.[3] For a knowledgeable walker for each node j of the network $s_j = \frac{1}{|N|}$, where $|N|$ is the cardinality of the set of nodes N constituting the network. For a page-u-fan we have $s_u = 1$, and $s_j = 0$ for any other page $j \neq u$. For a uniform-set-U-fan[4] we get

$$s_j = \begin{cases} \dfrac{1}{|U|} & \text{if } j \in U \\ 0 & \text{otherwise} \end{cases}, \ j = 1, \dots |N| \tag{2}$$

and for a hub-page-preferring-set-U-fan we obtain

$$s_j = \begin{cases} \dfrac{outdeg(j)}{\sum_{k\in U} outdeg(k)} & \text{if } j \in U \\ 0 & \text{otherwise} \end{cases}, \ j = 1, \dots |N| \tag{3}$$

The former case is the topic of this paper, the second was considered in our former paper [10].

Instead of a random walker model, we can view a web as a pipe-net through which the authority is flowing in discrete time steps. In single time step a fraction ζ of the authority of a node j flows into so-called *super-node*, and the fraction $\frac{1-\zeta}{outdeg(j)}$ is sent from this node to each of its children in the graph. After the super-node has received authorities from all the nodes, it redistributes the authority to all the nodes in fractions defined in the vector $\mathbf{s}$. Note that the authority circulates lossless (we have a kind of a closed loop here). Besides this, as was proven in many papers, we have to do here with a self-stabilizing process. Starting with any stochastic vector $\mathbf{r}^{(0)}$ and applying the operation

$$\mathbf{r}^{(n+1)} = (1-\zeta)\cdot\mathbf{P}\cdot\mathbf{r}^{(n)} + \zeta\cdot s$$

the series $\{\mathbf{r}^{(n)}\}$ will converge to $\mathbf{r}$ being the solution of the Eq. (1) (i.e. to the main eigenvector corresponding to eigenvalue 1).

Subsequently let us consider only connected graphs (one-component graphs) with symmetric links, i.e. unoriented graphs. Hence for each node j the relationships between in- and out-degrees are: $indeg(j) = outdeg(j) = deg(j)$. In a former paper we have proven [10].

Theorem 1. *For the preferential personalized PageRank we have*

$$p_o\zeta \leq (1-\zeta)\frac{|\partial(U)|}{Vol(U)}$$

where $\partial(U)$ is the set of edges leading from U to the nodes outside of U (the so-called "edge boundary of U"), $|\partial(U)|$ is the its cardinality, and $Vol(U)$, called volume or capacity of U is the sum of out-degrees of all nodes from U.

[3] We will denote the solution to the Eq. (1) with $\mathbf{r}^{(t)}(\mathbf{P}, \mathbf{s}, \zeta)$.

[4] We will call the set U "fan-pages" or "fan-set" or "fan-nodes".

Let us discuss now a uniform-set-U-fan defined in Eq. (2). Consider the situation where U is only a proper subset of N, and assume that

$$r_j^{(t)} = \begin{cases} \dfrac{1}{|U|} & \text{if } j \in U \\ 0 & \text{otherwise} \end{cases}, \; j = 1, \ldots |N| \tag{4}$$

in a moment t. To find the distribution $\mathbf{r}^{(t')}$ for $t' > t$ we state that if in none of the links the passing amount of authority will exceed $\gamma = (1-\zeta)\frac{1}{|U| \min_{k \in U} deg(k)}$, then at any later time point $t' > t$ the inequality $r_j^{(t')} \leq deg(j) \cdot \gamma + \frac{\zeta}{|U|}$ holds at any node $j \in U$, because if a node $j \notin U$ gets via links $l_{j,1}, ..., l_{j,deg(j)}$ the authority amounting to $a_{l_{j,1}} \leq \gamma, ..., a_{l_{j,deg(j)}} \leq \gamma$ then it accumulates

$$\mathfrak{a}_j = \sum_{k=1}^{deg(j)} a_{j,k} \leq \gamma \cdot deg(j)$$

of total authority, and in the next time step the following amount of authority flows out through each of these links:

$$(1-\zeta)\frac{\mathfrak{a}_j}{deg(j)} \leq \gamma(1-\zeta) \leq \gamma$$

If a node $j \in U$ gets via incoming links $l_{j,1}, ..., l_{j,deg(j)}$ the authority amounting to $a_{l_{j,1}} \leq \gamma, ..., a_{l_{j,deg(j)}} \leq \gamma$ then, due to the authority obtained from the super-node equal to $\mathfrak{b}_j = \zeta\frac{1}{|U|} \leq deg(j)\gamma\frac{\zeta}{1-\zeta}$, in the next step through each link the authority amounting to

$$(1-\zeta)\frac{\mathfrak{a}_j}{deg(j)} + (1-\zeta)\frac{\mathfrak{b}_j}{deg(j)} \leq \gamma(1-\zeta) + \gamma\frac{\zeta}{1-\zeta}(1-\zeta) = \gamma$$

flows out. So if already at time point t the authority flowing out through any link from any node did not exceed γ, then this property will hold (by induction) forever, especially for the equation solution $\mathbf{r}$ which is unique. Let us denote by p_o the total mass of authority contained in all the nodes outside of U. We ask: “How much authority from outside of U can flow into U via super-node at the point of stability?” This question concerns the quantity $p_o\zeta$. We claim that

Theorem 2. *For the uniform personalized PageRank we have*

$$p_o\zeta \leq= (1-\zeta)\frac{|\partial(U)|}{|U| \min_{k \in U} deg(k)}$$

Proof. Let us notice first that, due to the closed loop of authority circulation, the amount of authority flowing into U from the nodes belonging to the set $\overline{U} = N \backslash U$ must be identical with the amount flowing out of U to the nodes in $\overline{U}$. But from U only that portion of authority flows out that flows out through the boundary of U because no authority leaves U via super-node (it returns from there immediately). As at most the amount $\gamma|\partial(U)|$ leaves U, then

$$p_o\zeta \leq \gamma|\partial(U)| = (1-\zeta)\frac{1}{|U|min_{k\in U}deg(k)}|\partial(U)| = (1-\zeta)\frac{|\partial(U)|}{|U|min_{k\in U}deg(k)}$$

When you compare the above two Theorems 1 and 2, you will see immediately that the bound in case of "preferential" Theorem 1 is lower than in case of "uniform" Theorem 2. If we look more broadly at the s vector with $s_j > 0\ \forall_{j\in U}$ and $s_j = 0\ \forall_{j\notin U}$, we will derive immediately by analogy the relation.

Theorem 3. *For the personalized PageRank with arbitrary s vector such that $s_j > 0\ \forall_{j\in U}$ and $s_j = 0\ \forall_{j\notin U}$ we have*

$$p_o\zeta \leq= (1-\zeta)\frac{|\partial(U)|}{\min_{k\in U}\frac{deg(k)}{s_k}}$$

3 Variants of the Theorems

In this section, our attention is concentrated on some versions of PageRank related to a random walk with a distinct semantic connotation.

3.1 Lazy Random Walk PageRank

A variant of PageRank, so-called *lazy-random-walk-PageRank* was described e.g. by [5]. It differs from the traditional PageRank in that the random walker before choosing the next page to visit he first tosses a coin and upon heads he visits the next page, and upon tails, he stays in the very same node of the network. Recall that for the lazy walker PageRank we have:

$$\mathbf{r}^{(l)} = (1-\zeta)\cdot(0.5\mathbf{I} + 0.5\mathbf{P})\cdot\mathbf{r}^{(l)} + \zeta\cdot\mathbf{s} \tag{5}$$

where $\mathbf{I}$ is the identity matrix.[5] Rewriting reveals relation to traditional one.

$$\mathbf{r}^{(l)} = \frac{1-\zeta}{1+\zeta}\cdot(\mathbf{P})\cdot\mathbf{r}^{(l)} + \frac{2\zeta}{1+\zeta}\cdot\mathbf{s} \tag{6}$$

So $\mathbf{r}^{(l)}$ for ζ is the same as $\mathbf{r}^{(t)}$ for $\frac{2\zeta}{1+\zeta}$ $(\mathbf{r}^{(l)}(\mathbf{P},\mathbf{s},\zeta) = \mathbf{r}^{(t)}(\mathbf{P},\mathbf{s},\frac{2\zeta}{1+\zeta}))$ Hence

Theorem 4. *For the preferential lazy personalized PageRank we have*

$$p_o\zeta \leq \frac{1-\zeta}{2}\frac{|\partial(U)|}{Vol(U)}$$

Theorem 5. *For the uniform lazy personalized PageRank we have*

$$p_o\zeta \leq \frac{1-\zeta}{2}\frac{|\partial(U)|}{|U|\min_{k\in U}deg(k)}$$

[5] We will denote the solution to the Eq. (5) with $\mathbf{r}^{(l)}(\mathbf{P},\mathbf{s},\zeta)$.

3.2 Generalized Lazy Random Walk

Let us generalize this behavior to *generalized-lazy-random-walk-PageRank* by introducing the laziness degree λ. It means that, upon tossing an unfair coin, probability of tails is λ (and heads $1-\lambda$). For the generalized lazy walker PageRank we have:

$$\mathbf{r}^{(g)} = (1-\zeta)\cdot\left(\lambda\mathbf{I} + (1-\lambda)\mathbf{P}\right)\cdot\mathbf{r}^{(g)} + \zeta\cdot\mathbf{s} \tag{7}$$

where $\mathbf{I}$ is the identity matrix.[6] Rewrite it to relate to the traditional PageRank.

$$\mathbf{r}^{(g)} = \frac{(1-\zeta)\cdot(1-\lambda)}{1-\lambda+\zeta\lambda}\mathbf{P}\cdot\mathbf{r}^{(g)} + \frac{\zeta}{1-\lambda+\zeta\lambda}\cdot\mathbf{s} \tag{8}$$

So $\mathbf{r}^{(g)}$ for ζ is the same as $\mathbf{r}^{(t)}$ for $\frac{\zeta}{1-\lambda+\zeta\lambda}$ $(\mathbf{r}^{(g)}(\mathbf{P},\mathbf{s},\zeta,\lambda) = \mathbf{r}^{(t)}(\mathbf{P},\mathbf{s},\frac{\zeta}{1-\lambda+\zeta\lambda}))$ Therefore

Theorem 6. *For the preferential generalized lazy personalized PageRank we have*

$$p_o\zeta \leq (1-\lambda)(1-\zeta)\frac{|\partial(U)|}{Vol(U)}$$

Theorem 7. *For the uniform generalized lazy personalized PageRank we have*

$$p_o\zeta \leq (1-\lambda)(1-\zeta)\frac{|\partial(U)|}{|U|\min_{k\in U} deg(k)}$$

4 Bipartite PageRank

Some non-directed graphs occurring e.g., in social networks are in a natural way bipartite graphs. That is there exist nodes of two modalities, and meaningful links may occur only between nodes of distinct modalities (e.g., clients and items purchased by them). Literature exists already for such networks attempting to adapt PageRank to the specific nature of bipartite graphs, e.g., [7]. Regrettably, no generalization of Theorem 2 was formulated. The one seemingly obvious choice would be to use the traditional PageRank like it was done in papers [2,13]. However, this would be conceptually wrong because the nature of the super-node would cause authority flowing between nodes of the same modality, which is prohibited by the definition of these networks. Therefore in this paper, we intend to close this conceptual gap using Bipartite PageRank concept created in our former paper [10] and will extend the Theorem 2 to this case.

So let us consider the flow of authority in a bipartite network with two distinct super-nodes: one collecting the authority from items and passing them to clients, and the other the authority from clients and passing them to items.

$$\mathbf{r}^p = (1-\zeta^{kp})\cdot\mathbf{P}^{kp}\cdot\mathbf{r}^k + \zeta^{kp}\cdot\mathbf{s}^p \tag{9}$$

$$\mathbf{r}^k = (1-\zeta^{pk})\cdot\mathbf{P}^{pk}\cdot\mathbf{r}^p + \zeta^{pk}\cdot\mathbf{s}^k \tag{10}$$

The following notation is used in these formulas

[6] We will denote the solution to the Eq. (7) with $\mathbf{r}^{(g)}(\mathbf{P},\mathbf{s},\zeta,\lambda)$.

- $\mathbf{r}^p$, $\mathbf{r}^k$, $\mathbf{s}^p$, and $\mathbf{s}^k$ are stochastic vectors, i.e. the non-negative elements of these vectors sum to 1;
- the elements of matrix $\mathbf{P}^{kp}$ are: if there is a link from page j in the set of *Clients* to a page i in the set of *Items*, then $p^{kp}_{ij} = \frac{1}{outdeg(j)}$, otherwise $p^{kp}_{ij} = 0$;
- the elements of matrix $\mathbf{P}^{pk}$ are: if there is a link from page j in the set of *Items* to page i in the set of *Clients*, then $p^{pk}_{ij} = \frac{1}{outdeg(j)}$, otherwise $p^{pk}_{ij} = 0$;
- $\zeta^{kp} \in [0, 1]$ is the boring factor when jumping from *Clients* to *Items*;
- $\zeta^{pk} \in [0, 1]$ is the boring factor when jumping from Items to Clients.

Definition 1. *The solutions* $\mathbf{r}^p$ *and* $\mathbf{r}^k$ *of the equation system (9) and (10) will be called item-oriented and client-oriented bipartite PageRanks, resp.*

Let us assume first that $\zeta^{pk} = \zeta^{kp} = 0$ i.e. that the super-nodes have no impact. Let $K = \sum_{j\in Clients} outdeg(j) = \sum_{j\in Items} outdeg(j)$ mean the number of edges leaving one of the modalities. Then for any $j \in Clients$ we have $r^k_j = \frac{outdeg(j)}{K}$, and for any $j \in Items$ we get $r^p_j = \frac{outdeg(j)}{K}$. Because the same amount of $\frac{1}{K}$ authority is passed through each channel, within each bidirectional link the amounts passed cancel out each other. So the $\mathbf{r}$'s defined this way are a fix-point (and solution) of the Eqs. (9) and (10). For the other extreme, when $\zeta^{kp} = \zeta^{pk} = 1$ one obtains, that $\mathbf{r}^p = \mathbf{s}^p$, $\mathbf{r}^k = \mathbf{s}^k$.

In analogy to the traditional PageRank let us note at this point that for $\zeta^{kp}, \zeta^{pk} > 0$ the "fan"-nodes of both the modalities (the sets of them being denoted with U^p for items and U^k for clients), will obtain in each time step from the super-nodes the amount of authority equal to ζ^{pk} for clients and ζ^{kp} for items, resp. Let us now think about a fan of the group of nodes U^p, U^k who jumps uniformly, Assume further that at the moment t we have the following state of authority distribution: node j contains $r^k_j(t) = \frac{1}{|U^k|}, r^p_j(t) = \frac{1}{|U^p|}$ (meaning analogous formulas for r^p and r^k). Let us consider now the moment $t+1$. From the item node j to the first super-node the authority $\zeta^{pk}\frac{1}{|U^p|}$ flows, and into each outgoing link $(1 - \zeta^{pk})\frac{1}{|U^p|deg(j)}$ is passed. On the other hand the client node c obtains from the same super-node authority $\zeta^{pk}\frac{1}{|U^k|}$, while from link ingoing from j $(1 - \zeta^{pk})\frac{1}{|U^p|deg(j)}$. The authority from clients to items passes in the very same way.

We have a painful surprise this time. In general, we cannot define a useful state of the authority of nodes, analogous to that of traditional PageRank from Sect. 2, so that in both directions between U^p and U^k nodes the same upper limit of authority would apply. This is due to the fact that in general capacities of U^k and U^p may differ. Therefore a broader generalization is required.

To find such a generalization let us reconsider the way how we can limit the flow of authority in a single channel. The amount of authority passed consists of two parts: a variable one being a share of the authority at the feeding end of the channel and a fixed one coming from a super-node. So, by increasing the variable part, we come to the point that the receiving end gets less authority that was there on the other end of the channel.

Let us seek the amount of authority d such that multiplied by the number of out-links of a sending node will be not lower than the authority of this node and that after the time step its receiving node would have also amount of authority equal or lower than d multiplied by the number of its in-links. That is we want to have that:

$$d\cdot(1-\zeta^{pk}) + \frac{\zeta^{pk}}{\sum_{v\in U^k} outdeg(v)} \le d$$

The above relationship corresponds to the situation that on the one hand if a node in *Items* has at most d amount of authority per link, then it sends to a node in *Clients* at most $d\cdot(1-\zeta^{pk})$ authority via the link. The receiving node j on the other hand, if it belongs to U^k, then it gets additionally from the super-node exactly $\frac{\zeta^{pk}}{|U^k|deg(j)}$ authority per its link. We seek a d such that these two components do not exceed d together.

If we look from the perspective of passing authority from *Clients* to *Items*, then, for similar reasons at the same time we have

$$d\cdot(1-\zeta^{kp}) + \frac{\zeta^{kp}}{|U^p|deg(j)} \le d$$

This implies immediately, that

$$d \ge \frac{1}{|U^k| \min_{j\in U^k} deg(j)} \text{ and } d \ge \frac{1}{|U^p| \min_{j\in U^p} deg(j)}$$

so we come to a satisfactory d when

$$d = \max(\frac{1}{|U^k| \min_{j\in U^k} deg(j)}, \frac{1}{|U^p| min_{j\in U^p} deg(j)})$$

$$= \frac{1}{\min(|U^k| \min_{j\in U^k} deg(j), |U^p| min_{j\in U^p} deg(j))}$$

Now we are ready to formulate a theorem for bipartite PageRank analogous to the preceding Theorem 2.

Theorem 8. *For the uniform personalized bipartite PageRank we have*

$$p_{k,o}\zeta^{kp} \le \frac{(1-\zeta^{pk})\partial(\frac{U^p}{U^k})}{min(|U^k|min_{j\in U^k}deg(j), |U^p|min_{j\in U^p}deg(j))}$$

and

$$p_{p,o}\zeta^{pk} \le \frac{(1-\zeta^{kp})\partial(\frac{U^k}{U^p})}{min(|U^k|min_{j\in U^k}deg(j), |U^p|min_{j\in U^p}deg(j))}$$

where

- $p_{k,o}$ *is the sum of authorities from the set* $Clients\backslash U^k$,
- $p_{p,o}$ *is the sum of authorities from the set* $Items\backslash U^p$,
- $\partial(\frac{U^k}{U^p})$ *is the set of edges outgoing from* U_k *into nodes from* $Items - U_p$ *(that is "fan's border" of* U^k*),*
- $\partial(\frac{U^p}{U^k})$ *is the set of edges outgoing from* U^p *into nodes from* $Clients\backslash U^k$ *(that is "fan's border" of* U^p*),*

□

The proof is analogous as in case of classical PageRank, using now the quantity d we have just introduced.

Proof. Let us notice first that, due to the closed loop of authority circulation, the amount of authority flowing into U^k from the nodes belonging to the set $\overline{U^p} = Items\backslash U^p$ must be identical with the amount flowing out of U^p to the nodes in $\overline{U^k}$. The same holds when we exchange the indices $p < - > k$.

But from U^p only that portion of authority flows out to $\overline{U^k}$ that flows out through the boundary of U^p because no authority leaves the tandem U^p, U^k via super-nodes (it returns from there immediately). As the amount $d|\partial(\frac{U^p}{U^k})|$ leaves at most the U^p not going into U^k, then

$$p_{k,o}\zeta^{kp} \leq d(1-\zeta^{pk})\partial(\frac{U^p}{U^k}) = \frac{(1-\zeta^{pk})\partial(\frac{U^p}{U^k})}{min(|U^k|min_{j\in U^k}deg(j), |U^p|min_{j\in U^p}deg(j))}$$

The convergence can be verified in an analogous way as done for the HITS (consult e.g., [12, Ch. 11]).

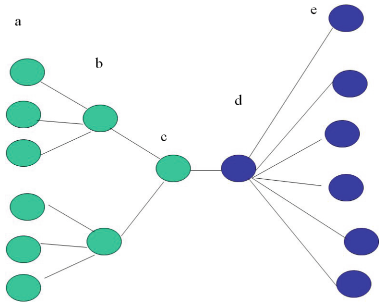

Fig. 1. Unoriented tree-like network

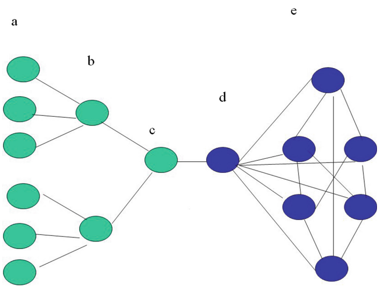

Fig. 2. Unoriented complex network

5 Experimental Exploration of the Limits

With the established limits, we can pose the question how tight the limits are or rather whether we can construct networks for which the limits are approached sufficiently close.

For this purpose we will use a family of networks depicted in Figs. 1 and 2. Each network is divided into three *zones* of nodes. Zones d and e belong to the set of fan-nodes.

Zones a, b, c are not fan-sets. There is only one node in zones c and d so that the edge connecting d to c is the channel through which the authority flows out

Table 1. PageRanks for network Fig. 1. Boring factor = 0.1

	Zone a	Zone b	Zone c	Zone d	Zone e
Traditional uniform	0.012479	0.055464	0.072565	0.370274	0.061892
Traditional preferential	0.013019	0.057864	0.075705	0.386296	0.057358
	Outflow	Limit	Rel.left		
Traditional uniform	0.025837	0.128571	0.799		
Traditional preferential	0.026955	0.069230	0.610		

Table 2. PageRanks for network Fig. 2. Boring factor = 0.1

	Zone a	Zone b	Zone c	Zone d	Zone e
Traditional uniform	0.006094806	0.027088026	0.0354401	0.1808376	0.1154962
Traditional preferential	0.006306085	0.028027043	0.0366687	0.1871064	0.1137223
	Outflow	Limit	Rel.left		
Traditional uniform	0.0126185	0.032142	0.6074242		
Traditional preferential	0.013055	0.029032	0.5502957		

Table 3. PageRanks for enlarged network Fig. 2 by factor in the first column. Boring factor = 0.1. Traditional PageRank with preferential authority re-distribution.

Factor	Zone a	Zone b	Zone c	Zone d	Zone e
10	1.737833e−05	7.723703e−05	7.073624e−04	2.438616e−02	1.620532e−02
100	1.969998e−08	8.755548e−08	7.674967e−06	2.494130e−03	1.662448e−03
1000	1.995495e−11	8.868865e−11	7.739489e−08	2.499416e−04	1.666249e−04
10000	1.998069e−14	8.880307e−14	7.745988e−10	2.499942e−05	1.666625e−05
100000	1.998323e−17	8.881436e−17	7.746628e−12	2.499994e−06	1.666662e−06
Factor	Outflow	Limit	Rel.left		
10	0.0003294802	0.0003657049979	0.0990544634		
100	3.700605208e−06	3.740632831e−06	0.01070076246		
1000	3.745018664e−08	3.749062578e−08	0.0010786466		
10000	3.749501576e−10	3.749906250e−10	0.000107915		
100000	3.749943936e−12	3.749990625e−12	1.245027547e−05		

Table 4. PageRanks for enlarged network Fig. 2 by factor in the first column. Boring factor = 0.1. Traditional PageRank with uniform authority re-distribution.

Factor	Zone a	Zone b	Zone c	Zone d	Zone e
10	1.679439e−05	7.464174e−05	6.835939e−04	2.356675e−02	1.622082e−02
100	1.904272e−08	8.463432e−08	7.418903e−06	2.410918e−03	1.662589e−03
1000	1.928972e−11	8.573209e−11	7.481482e−08	2.416095e−04	1.666263e−04
10000	1.931466e−14	8.584294e−14	7.487786e−10	2.416610e−05	1.666626e−05
100000	1.931710e−17	8.585376e−17	7.488401e−12	2.416661e−06	1.666663e−06
1000000	1.931896e−20	8.586206e−20	7.488419e−14	2.416666e−07	1.666666e−07
Factor	Outflow	Limit	Rel.left		
10	0.0003184092239	0.0003688524590	0.1367572		
100	3.577140044e−06	3.743760399e−06	0.04450614810		
1000	3.620173279e−08	3.749375104e−08	0.03445956209		
10000	3.624516929e−10	3.749937501e−10	0.03344604329		
100000	3.624940904e−12	3.749993750e−12	0.033347481127		
1000000	3.625220796e−14	3.74999937e−14	0.033274293139		

Table 5. PageRanks for densified network from last line of previous table - the zone e node degrees as in the first column

e node deg.	Zone a	Zone b	Zone c	Zone d	Zone e
5000000	1.572210e−20	6.987602e−20	6.094048e−14	1.966666e−07	1.666666e−07
5500000	1.440820e−20	6.403644e−20	5.585813e−14	1.803030e−07	1.666666e−07
5900000	1.352272e−20	6.010099e−20	5.242307e−14	1.692090e−07	1.666666e−07
5990000	1.333886e−20	5.928383e−20	5.171155e−14	1.669171e−07	1.666666e−07
5999000	1.331915e−20	5.919623e−20	5.163832e−14	1.666916e−07	1.666666e−07
5999900	1.332161e−20	5.920716e−20	5.163986e−14	1.666691e−07	1.666666e−07
e node deg.	Outflow	Limit	Rel.left		
5000000	2.950251336e−14	2.999999500e−14	0.01658272402		
5500000	2.703801851e−14	2.727272272e−14	0.0086058227330		
5900000	2.537613978e−14	2.542372457e−14	0.00187166895		
5990000	2.503123889e−14	2.504173205e−14	0.00041902692587		
5999000	2.4994569e−14	2.500416319e−14	0.0003836900900		
5999900	2.49983829e−14	2.500041250e−14	8.117929671e−05		

of the fan-node set and we seek the upper limit of authority lost via this link. The zones are symmetrically constructed. The number of nodes in a is a multiple of the number of nodes in b. All nodes in e are connected to d, and otherwise, they constitute a regular subgraph. In Fig. 1 this subgraph is of degree zero, and in Fig. 2 it is of degree 3. Because of symmetry, the PageRanks in each of the zones are identical.

Table 1 shows the PageRanks for the graph in Fig. 1. Table 2 shows the PageRanks for the graph in Fig. 2. In each table the columns *zone a*,...,*zone e* show the PageRank attained by each node in the respective zone. *outflow* column shows the amount of authority flowing out from the fan-set of nodes to the rest of the network. *limit* column is the upper limit derived theoretically in the

Table 6. PageRanks for various network structures with the same upper limit of authority passing - the preferential redistribution. Zone a and b both 60000 nodes each.

e node deg./count	Zone a	Zone b	Zone c	Zone d	Zone e
511/1024	6.092727e−11	1.353939e−10	5.370716e−06	1.953285e−03	9.746382e−04
255/2048	6.095403e−11	1.354534e−10	5.373075e−06	3.906379e−03	4.863655e−04
127/4096	6.096742e−11	1.354832e−10	5.374255e−06	7.812567e−03	2.422291e−04
63/8192	6.097412e−11	1.354980e−10	5.374845e−06	1.562494e−02	1.201609e−04
31/16384	6.097746e−11	1.355055e−10	5.375140e−06	3.124970e−02	5.912678e−05
15/32768	6.097914e−11	1.355092e−10	5.375287e−06	6.249920e−02	2.860973e−05
7/65536	6.097997e−11	1.355110e−10	5.375361e−06	1.249982e−01	1.335121e−05
3/131072	6.098038e−11	1.355119e−10	5.375397e−06	2.499961e−01	5.721944e−06
1/262144	6.098056e−11	1.355124e−10	5.375413e−06	4.999919e−01	1.907314e−06
e node deg./count	Outflow	Limit	Rel.left		
511/1024	1.714998913e−06	1.716610495e−06	0.0009388162095		
255/2048	1.715752081e−06	1.716610495e−06	0.0005000635487		
127/4096	1.716128933e−06	1.716610495e−06	0.0002805306660		
63/8192	1.71631741e−06	1.716610495e−06	0.0001707321858		
31/16384	1.716411644e−06	1.716610495e−06	0.0001158392913		
15/32768	1.716458707e−06	1.716610495e−06	8.842327961e−05		
7/65536	1.716482125e−06	1.716610495e−06	7.478124133e−05		
3/131072	1.716493600e−06	1.716610495e−06	6.809630427e−05		
1/262144	1.716498849e−06	1.716610495e−06	6.503878422e−05		

previous sections for the respective case. *rel.left* is computed as 1-*outflow*/*limit*. The lower the value, the closer the actual outflow to the theoretical limit.

The obvious tendency to keep authority is observed when the network of connections is densified between fan nodes. Also, the outflow of authority gets closer to the theoretical bound.

How close can it go? In Tables 3 and 4 we increase by the factor of 10,100 etc. the number of nodes in zones *a*, *b* and *e* and also the number of connections between the nodes in zone *e* (enlarging the network of Fig. 2, results in Table 5).

We see that in case of preferential attachment, we quickly approach the bounds. In case of uniform authority redistribution, we get a stabilization. The situation changes for the uniform case, however, if we densify the connections in zone *e*. For the network of the last line, we increase the density of connections in zone *e*. Last not least, let us observe that the relationship between the upper limit and the actual amount of authority passed is a function of the structure of the network. In the Tables 6 (for preferential redistribution) and 7 (for uniform redistribution) we see this effect. For preferential redistribution, we see that the lower degrees the nodes are, the bigger part of the authority is flowing out. For the uniform redistribution, the tendency is in the other direction.

Table 7. PageRanks for various network structures with the same upper limit of authority passing - the uniform redistribution Zone a and b both 60000 nodes each.

e node deg./count	Zone a	Zone b	Zone c	Zone d	Zone e
4/131071	4.480253e−11	9.956117e−11	3.949326e−06	1.836718e−01	6.228041e−06
8/65535	4.933356e−11	1.096301e−10	4.348734e−06	1.011236e−01	1.371576e−05
16/32767	5.196287e−11	1.154730e−10	4.580507e−06	5.325655e−02	2.889275e−05
32/16383	5.339301e−11	1.186511e−10	4.706573e−06	2.736115e−02	5.936787e−05
64/8191	5.416866e−11	1.203748e−10	4.774947e−06	1.387931e−02	1.203889e−04
128/4095	5.468880e−11	1.215307e−10	4.820796e−06	7.006293e−03	2.424855e−04
256/2047	5.545008e−11	1.232224e−10	4.887903e−06	3.551911e−03	4.867770e−04
512/1023	5.783015e−11	1.285114e−10	5.097705e−06	1.852184e−03	9.756907e−04
e node deg./count	Outflow	Limit	Rel.left		
4/131071	1.261114759e−06	1.716613769e−06	0.2653474055		
8/65535	1.388655573e−06	1.716613769e−06	0.1910494961		
16/ 32767	1.462666077e−06	1.716613769e−06	0.147935252		
32/16383	1.502922183e−06	1.716613769e−06	0.1244843712		
64/8191	1.524755444e−06	1.716613769e−06	0.1117655749		
128/4095	1.539396343e−06	1.716613769e−06	0.1032366329		
256/2047	1.560825131e−06	1.716613769e−06	0.09075345911		
512/1023	1.627819998e−06	1.716613769e−06	0.05172612086		

6 Concluding Remarks

In this paper, we have proposed limits for the flow of authority in ordinary unoriented and in the bipartite graph under uniform random jumps. We have empirically demonstrated tightness of some of these limits.

The obtained limits can be used for example, when verifying the validity of clusters in such graphs. It is quite common to assume that the better the cluster, the less authority flows out of it when treating the cluster as the set on which a fan concentrates while a personalized PageRank is computed. The theorem says that the outgoing authority has a natural upper limit dropping with the growth of the size of the sub-network so that the outgoing authority cluster validity criterion cannot be used because it will generate meaningless large clusters. So a proper validity criterion should make a correction related to the established limits in order to be of practical use.

As a further research direction, it is obvious that finding tighter limits are needed. This would improve the evaluation of e.g., cluster quality.

References

1. Allesina, S., Pascual, M.: Googling food webs: can an eigenvector measure species' importance for coextinctions? PLoS Comput. Biol. **5**, e1000494 (2009)
2. Bauckhage, C.: Image tagging using PageRank over bipartite graphs. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 426–435. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69321-5_43
3. Berkhin, P.: A survey on PageRank computing. Internet Math. **2**, 73–120 (2005)

4. Chung, F.: PageRank as a discrete green's function. In: Ji, L. (ed.) Geometry and Analysis, I, Advanced Lectures in Mathematics (ALM), vol. 17, pp. 285–302. International Press of Boston, 15 July 2011
5. Chung, F., Zhao, W.: PageRank and random walks on graphs (2008). http://www.math.ucsd.edu/~fan/wp/lov.pdf
6. De Domenico, M., Sole-Ribalta, A., Omodei, E., Gomez, S., Arenas, A.: Ranking in interconnected multilayer networks reveals versatile nodes. Nat. Commun. **6**, 6868 (2015)
7. Deng, H., Lyu, M.R., King, I.: A generalized co-hits algorithm and its application to bipartite graphs. In: Proceedings of the 15th ACM SIGKDD, pp. 239–248, Paris (2009)
8. Dominguez-Garcia, V., Munoz, M.A.: Ranking species in mutualistic networks. Sci. Rep. **5**, 8182 (2015)
9. He, X., Gao, M., Kan, M.Y., Wang, D.: Birank: towards ranking on bipartite graphs. IEEE Trans. Knowl. Data Eng. **29**(1), 57–71 (2017). https://doi.org/10.1109/TKDE.2016.2611584
10. Kłopotek, M.A., Wierzchoń, S.T., Kłopotek, R.A., Kłopotek, E.A.: Network capacity bound for personalized bipartite PageRank. In: Matwin, S., Mielniczuk, J. (eds.) Challenges in Computational Statistics and Data Mining. SCI, vol. 605, pp. 189–204. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-18781-5_11
11. Langville, A.N.: An annotated bibliography of papers about Markov chains and information retrieval (2005). http://www.cofc.edu/~langvillea/bibtexpractice.pdf
12. Langville, A.N., Meyer, C.D.: Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, Princeton (2006)
13. Link, S.: Eigenvalue-based bipartite ranking. Bachelorarbeit/bachelor thesis (2011)
14. Liu, C., Tang, L., Shan, W.: An extended hits algorithm on bipartite network for features extraction of online customer reviews. Sustain. MDPI Open Access J. **5**(10), 1–15 (2018)
15. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical Report 1999–66, Stanford InfoLab, November 1999. http://ilpubs.stanford.edu:8090/422/
16. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., Pietronero, L.: A new metrics for countries' fitness and products' complexity. Sci. Rep. **2**, 723 (2012)