Relationship of Model Parameters and Solution Stability in the Power Walk Page Rank Method.

Ryan Greenup

October 11, 2020

Contents

1 Introduction

Much information is interconnected, consider for example written articles, research papers, citations, movies, books, personal notes, knowledge bases, personal releationships etc., this interconnectivity creates a network which can naturally be visualised as a graph.

Determining the most central vertex of such a graph is useful as it may assist in research, learning and collaboration. The *PageRank* method, which originally underpinned *Google*'s search-engine, essentially asserts, that the

centre-most vertex of a graph could be considered the vertex that has the highest probability of being a destination following a random walk. [24] Although this method is popular, it does not, however, easily adapt to negatively weighted edges (a negative weight in this case measuring dissaproval or an aversion to follow that link as opposed to endorsement). This limitation is increasingly relevant given recent developments in the area of *Sentiment Analysis*. [34]

The *Power Walk* Method is similar to the *Random Surfer* model but takes a different approach in order to accommodate the potential for an edge to have any arbitrary real weight. [34]

This research project will be an investigation into the relationship between model parameters and the solution stability of the *Power Walk* method.

1.1 Motivation

Taking advantages of inter-related ideas allows for exploratory research which can promote both a deeper and broader understading of a topic, an example of this is the concept of a *pathfinder*, which is a list of central sources for a topic that can help in early stages of literature searches, [15] examples of such Pathfinders include:

- Institution examples:
  - *M.I.T.* [28]
  - *Harvard* [16]
  - Berkley University [2]
  - *Library of Congress* [38]
    - The Library of Congress refers to a *PathFinder* as a *Tracer Bullet*, which is both a very descriptive and very American name.
- Wiki Examples:
  - Math Online [26]
  - *Brilliant* Wiki [40]
  - Category Pages within MediaWiki Sites [18]
    - For example the *Wikipedia* Mathematics Category Page [7]

Using *Wikipedia* for example can make for a very effective subject guide by following the various hyperlinks across wikipedia and leveraging Category Pages [18] in order to map out a topic while referring to references as necessary. Such a strategy is even recommended by some Library Study Guides [29], and has even been implemented in recommendation systems to overcome limitations imposed by a lack of data [25] (see section 4.1.2).

Similarly the *Zettelkasten* method of note taking involves, in essence, a collection of small interlinked notes [33], much like a wiki, within the last two decades this method has peaked in popularity [13], despite being a method dating back to atleast the 18th Century. [14] This is likely driven, atleast in part, by the development of the internet, *HTML* and the increasing breadth and depth of human knowledge.

A comprehensive review of the pathfinder approach, however, reveals that many such subject guides are often not consistent with the needs of those engaging in research. [41] This issue is somewhat analogous to the work of Larry Page and Sergey Brin when developing *Google* in that the centrality of an article of information corresponds very greatly to it's relevance. [24]

This issue of relevance in information retrieval can be addressed by using an algorithm to identify what is most central to a topic, ideally the *Power Walk* method will prove to be an effective tool to achieve this.

2 Background and Rationale

Given some graph $G(V, E)$ with $\mid V \mid = n$, a corresponding adjacency matrix $\mathbf{A}$ describes the number of edges between vertex i and j, such that $\mathbf{A}_{i,j}$ represents the number of edges between the vertices.

With respect to a directed graph the element $\mathbf{A}_{i,j}$ may either denote the edge:

- $i \to j$
  - One such example of this implementation being the `igraph` library [12]
- $j \to i$ [32, §2.3], OR

While both definitions appear in the literature, the latter definition is more common/convenient when working with *probability transition matrices* and will hence be adopted here.

2.1 Stationary Distribution

Given this adjacency matrix $\mathbf{A}$, a *probability transition matrix* $\mathbf{T}$ can be produced by scaling each column to 1, such that each element $\mathrm{T}_{i,j}$ would represent the probability of leaving j and travelling to i during a random walk (as opposed to the number of edges as was the case with $\mathbf{A}$), this can be acheived with matrix multiplication as illustrated in (1):

$$\mathbf{T} = \mathbf{A} \ \operatorname{diag}\left(\texttt{colsums}\left(\mathbf{A}\right)\right)^{-1} \tag{1}$$

(See § 5.1 for a note on the math)

The state distribution $\vec{p_k}$ describes the frequency (scaled to 1) of visiting each vertex during a random walk for the k^{th} step of the walk, given this, the stationary distribution $\vec{p}$ is given by (2):

$$\begin{aligned} \vec{p_i} &= \mathbf{T}p_{i-1} \\ \lim_{n\to\infty}[\vec{p_i}] &= \lim_{n\to\infty}[\mathbf{T}\vec{p_{i-1}}] \\ \implies \vec{p} &= \mathbf{T}\vec{p} \end{aligned} \tag{2}$$

If $G(V, E)$ is an ergodic graph (i.e. all vertices may be reached from any initial vertex), this can be solved by iteration by setting some threshold (η) for convergence (which will be referred to as the *Power Method*) or by solving the eigenvalue problem for $\lambda = 1$ as shown in (3):

$$\begin{aligned} \lambda\vec{p} &= \mathbf{T}\vec{p} \\ \lambda = 1 \implies \vec{p} &= \mathbf{T}\vec{p} \end{aligned} \tag{3}$$

2.2 Random Surfer

If however a graph is non-ergodic, this random walk will not traverse every vertex, to overcome this, the *Random Surfer* model can be implemented [24]. This method involves, essentially, introducing into the *probability transition matrix* $(\mathbf{T})$, some probability $(\frac{1-\alpha}{n})$ of traversing to a disconnected vertex (V), this is shown in (4):

$$\mathbf{T}_{\text{RS}} = \mathbf{S} = \alpha\mathbf{T} + (1-\alpha)\mathbf{B} \tag{4}$$

where:

- $\mathbf{B}$ is matrix of size $n \times n$ such that $\mathbf{B}_{i,j} = \frac{1}{n}, \ \ \forall i, j \in [1, n] \cap \mathbb{N}$
- In the literature α is often referred to as a damping factor see [1, 5, 11, 20, 3] or a smoothing constant [21] .

2.3 Power Walk

The random surfer model (4), however, assumes that all edges are an edorsement of the target, i.e. they are weighted positively, the power walk method [34], shown in (5), takes a different approach to create a *transition probability matrix* ($\mathbf{W}$) and is compatible with a negatively weighted edges:

$$\mathbf{W}_{i,j} = \frac{\beta^{\mathbf{A}'i,j}}{\sum_{j=1}^{n} [\beta^{a_{i,j}}]} \tag{5}$$

where:

$\mathbf{A}'$ is a weighted adjacency matrix such that $\mathbf{A}_{i,j} \in \mathbb{R}$

β is the ratio of probability between following an edge and making a jump to a vertex for which there is no path

- i.e. βx is the probability of following a path with a weight of 1 where:

  x is the probability of travelling to a vertex for which there is no connection.
  - Similarly to (4) , $x = \frac{1-\alpha}{n}$

2.4 Solving the stationary distribution

Solving the EigenValue problem for a large matrix can be very resource intensive, for example *Wikipedia* currently has over 6, 000, 000 pages [42] which would correspond to an adjacency matrix with over 10^{12} entries, yet even a relatively fast compiled language like *Julia* can struggle to solve the eigen vectors for a matrix of size $(10^4)^2$ as shown in listing 1.

The power method, first mentioned in section 2.1 , is a better suited approach, with respect to performance, because:

1. The method is only looking for one solution
2. The accuracy of the solution (measured by $\exists \eta \in \mathbb{R}$) can be tuned to improve performance.

Listing 1: Time to Solve Eigen Value for matrix of size n

3 Proposed Research

Consider the ordered set of EigenVectors (6) of a positive transition probability matrix such as $\mathbf{S}$ (4) or $\mathbf{T}$ (2):

$$\{\lambda_k \mid \ \ \lambda_k < \lambda_{k-1}, \ \ k \in \mathbb{Z}^+ \leq n\} \tag{6}$$

3.1 Dominant EigenVector

It has been shown that $\lambda_k \leq 1, \ \ \forall k \leq n$ and that the dominant[1] λ can be computed by the *power method.* [9]

3.2 Stability and Convergence

It has also been shown that the stationary distribution $\vec{p}$ (see (2)) can be reached in a limited number of steps (≈ 50) for graphs on the order of a million vertices [3, p. 123], under the assumption that the smoothing constant $\alpha \in [0, 1]$ is not too close to 1 (in which case convergence can become quite slow) [39]

How quickly the *Power Method* converges generally depends on the magnitude of $\mid \lambda_2 \mid$ [6] and with respect to the random surfer model (4), It has been shown that: [20]

- $\mid \lambda_2 \mid \leq \alpha$, and
- if the corresponding graph contains two or more irreducible closed subgraphs then $\mid \lambda_2 \mid = \alpha$

This is demonstrated in listing 2 and figure 1.

It has also been shown that an α value near 1 will imply an unstable stationary distribution [31] that converges slowly [39], this is because λ_2 is bound above by α and a small change to the corresponding graph could lead to $\lambda_1 \leftrightarrow \lambda_2$ and hence different eigenvectors will correspond to the solution as shown in (3) .

It is not clear how similar properties are exhibited with respect to the *Power Walk* method. [36]

3.3 Choosing α

Although section 3.2 might suggest that smaller values of α may be more ideal, it is worth recalling that as α is reduced the probability of a random walk visiting any other vertex will become more and more uniform because $\frac{1-\alpha}{n} \to \frac{1}{n}$ as $\alpha \to 0$. [36]

The value used originally by Page and Brin was $\alpha = 0.85$ See [24, p. 109] and this appears to have been widely adopted, see [20, 4]. Research suggests, however, that modifying the value may be useful in detecting spam. [43, 4]

Listing 2: Implementing the random surfer model for the graph shown in figure

3.4 Research Question

It is not clear how λ_2 behaves with respect to the *Power Walk* method, (5) although it has been shown that under specific circumstances the value of $\mid \lambda_2 \mid$ can be predicted from the method parameters and properties of the graph. [34, §3.4]

[1] Dominant in this case refers to the the largest $\mid \lambda_k \mid$

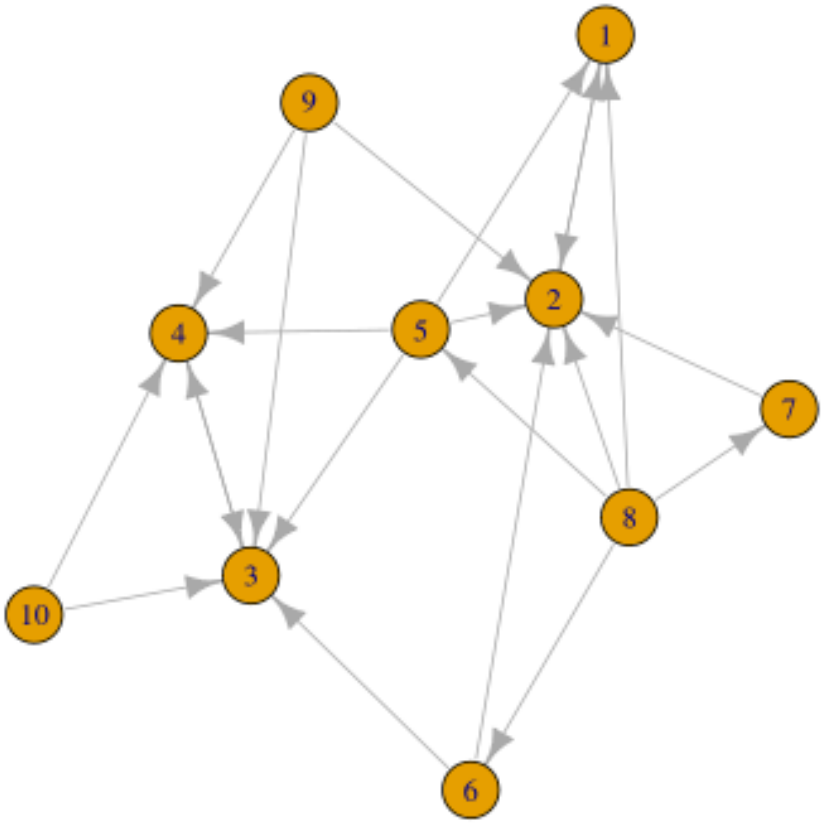

Figure 1: Graph with two closed irreducible subgraphs

This research will involve investigating the relationship between the second eigenvalue of the *Power Walk* transition matrix and the features of a graph corresponding to some type of network (e.g. a social network, webpages, wiki, etc.)

In particular, open questions are whether or not the value of the second eigenvalue can:

- be predicted from the parameters of the model and/or features of the graph
  - e.g. some function of α / β
- indicate the stability of the stationary distribution of a graph
- indicate how quickly the *Power Method* will converge to a solution

4 Literature Review

The proposed research (see section 3.4) relates broadly to the *PageRank* method, Random-Surfer model, sentiment analysis and graph centrality, for which material is quite abundant, although much of the literature is concerned with either:

1. The original *PageRank* method developed by Page and Brin [24]
2. Modifying the *PageRank* method to improve upon:
   - Precision and accuracy see [31, 1, 30, 11]
   - Performance with respect to:
     - Rate of convergence in terms of iterations and time, see [39, 22]
     - Stability of any given solution, see [31]

Although neither of these points are a direct analogue for the proposed research, which relates in itself to an alternative *PageRank* algorithm, much of the work will be very similar in approach and hopefully offer much insight upon closer inspection.

4.1 Building on Literature Referred to in Primary Resource

This research is focused primarily on the *Power Walk* method proposed by Park and Simoff in a 2013 conference paper, [34] this paper contained some discussion of relevant research.

4.1.1 Stability and Convergence

Haveliwala and Kamvar [17] proved that λ_2 (see (6)) is bounded above by the smoothing constant α and in the case that the corresponding graph has more than 1 closed subgraph is equal to α. This is an important revelation because it has been shown that the further the second eigenvalue is from 1, the more resistant the stationary distriubtion of the *PageRank* is to perturbations in the corresponding graph, [31] and the faster the *PageRank* will converge [6].

It has been shown that the *power method* (see section 2.4) will always converge $\forall \alpha < 1$ [3] and that an α closer to the value of 1 does not necessarily correspond to a more meaningful ranking, [4] hence, given the upper bound of $\lambda_2 \leq \alpha$, the value of α can be tuned away from 1 in order to improve the convergence and stability of the *PageRank* (however a value of α that is too small will indeed be meaningless as discussed in section 3.3). [36]

This works provides a framework for considering the method parameters and λ_2 with respect to the convergence and stability of the *Power Walk* method.

4.1.2 Building on the Random Surfer

Related work referred to in the paper has involved using community ratings of web pages to improve upon the *PageRank* method [35], similar work has also been undertaken more recently that found replacing the background probability $\frac{1}{n}$ with a combination of usage statistics and content quality scores can significantly improve the precision and accuracy of the page rank method. [30]

Such a strategy is however limited to websites that make usage statistics public, such as wikis.

An extension to this research could involve an investigation into the precision of the *Power Walk* method in conjuction with usage statistics compared with the *Power Walk* method.

There is literature suggesting that the network structure of wiki articles can be an important feature in the emergence of quality [19], related work also shows that *Wikipedia* can be used to improve performance of recommender systems when there is limited data [25] and it would be very interesting to see how the *Power Walk* method would perform compared to the *PageRank* method in those situations.

4.2 Page Rank

4.2.1 Building on the *PageRank* Method

The *PageRank* method is a relatively versatile approach[2] that is relatively robust to manipulation compared with other methods for dealing with information retrieval, [23] perhaps for this reason there is much literature on modifying the *PageRank* method to improve upon it as discussed generally in section 4 .

Choosing a smoothing constant, however, is a somewhat difficult task because it can have an impact on the behaviour of the model (see [11] and section 4.1.1) but also because without empirical guidance it can feel somewhat arbitrary, there is an approach in the literature that involves using input/output ratios to determine an appropriate value [11] and another that seeks to use structural network dynamics to provide a score distribution and obviate the need for a smoothing constant entirely. [1]

It is not entirely clear if this approach will offer much to this method but a more careful inspection may reveal helpful perspectives.

4.2.2 Stability and Convergence

Improving the rate of convergence of the *PowerRank* is obviously desirable and there has been considerable mathematical resarch to develop better algorithms.

As previously mentioned in 3.2 , the stability and convergence of the *Power Rank* method is poor when the smoothing constant α is close to 1, a 2016 paper published in the *Journal of Computational and Applied Mathematics* [39] found that the trace of a matrix can be used to produce a considerably more efficient approach to solve the *PageRank* for values of α near 1. It is not clear how relevant this is given that α values near 1 offer no improvement in precision [4] and that the solution is unstable [31] (see sections 3.3 and 3.2), but, it is yet to be shown if these characteristics necessarily apply to the *Power Walk* method and such an approach may prove to be insighful nonetheless.

Another approach involves involves reordering the problem and taking advantage of the fact that the transition probability matrix is sparse [3] in order to produce a new algorithm which cannot perform worse than the *power method* but has been shown to improve the rate of convergence in certain cases. [22].

[2] The approach has even been used in conjuction with linear regression to map gene expresseions, see [45]

[3] if an adjacency matrix and/or corresponding probability transition matrix were not sparse each vertex would be like an index, which is unlikely

4.2.3 Insightful Miscellaneous Work

PageRank as a Power Series

Research has shown that the *PageRank* Method can be expressed as a power series and an algorithm for calculating the page rank derived, [5] the solution corresponds to the *power method* but a slightly faster algorithm is also presented. Seperate work has been undertaken to similarly express the PageRank in terms of a *McLaurin Series*, finding that each partial sum of the series corresponds to an iteration of the *power method*. [4] This work is extremely relevant to the *Power Walk* method because the exponent in that method (see (5)) suggests that an generating function such as $f(x) = \sum_{i=0}^{n} \left[x^n \frac{a}{n!}\right]$ may be able to show a more direct relationship between the *PowerRank* and *Power Walk* approaches.

Modelling

The *PageRank* method has been leveraged as a value to assist in building artificial networks in order to model real-world networks, such networks have been shown to have upper and lower bounds on there diamaters. [27] This is a very interesting area of research and it would be interesting to see whether or not the use of the *Power Walk* method in such an approach produces graphs that are more consistent with social networks.

Pure Mathematics

One very interesting piece of work in the literature was an application of the *PageRank* method to a graph of integers with edges based on divisors, as shown in figure 2 and listing 3. [10]

This is well outside the scope of this research, but if the precision of the power walk method is found to be reasonably good, it would make for a very interesting exercise to measure it's performance at predicting integers and attempting to find relationships between the two.

Another paper outside the scope of this paper is work by Ding & Li concerned with extending the *PageRank* method to *multi-plex* graphs[4], although very interesting and quite practical, such research is beyond the scope of this work.

Listing 3: *PageRank* Probability Transition Matrix of network based on divisibility of $\mathbb{Z}^+ \leq 30$

4.3 Search Engine Optimisation

There is a considerable amount of work in the literature concerning the relationship between the *PageRank* method and Search Engine optimisation, such as:

- Using decision trees with machine learning to inductively model search engines [37]

[4]Multi-plex, in this case, refers to edges between vertices accross different dimensions, for example a link from a webpage to a food outlet could be made by way of a hyperlink, a phone number and a street address, this would be 3 different types of edges between two vertecies and so would be multi-plex.

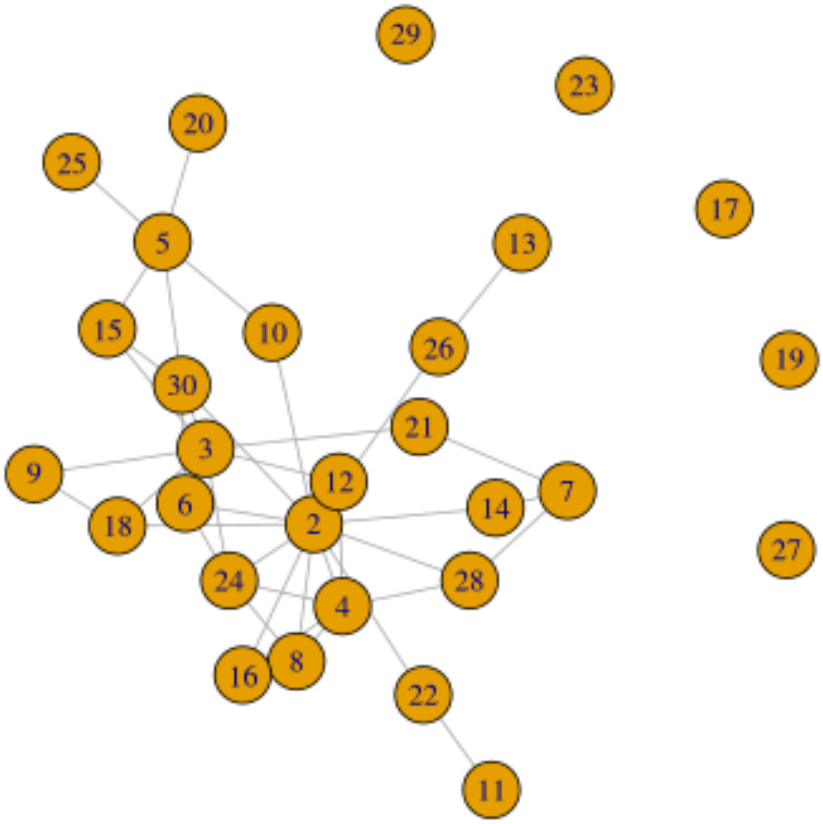

Figure 2: Graph of $\mathbb{Z}^+ \leq 30$ with edges based on divisors

- Methods to solve the optimisation problem involved in centring a vertex by creating a limited number of edges [20, 8]
  - Consider a website trying to maximise exposure for example
  - Related papers consider also keyword frequency, see for example [44]

Such literature however is suited to an ex post facto study and is hence not terribly relevant to the proposed research.

5 Reviewing the Material

5.1 Scaling Matrix from section 2.1

So what we're trying to do is to scale each column such that it adds to 1, all though we could just do:

$$\mathbf{T} = \frac{\mathbf{A}}{\text{diag}\left(\vec{1}\mathbf{A}\right)} \tag{7}$$

The problem with this method is that if of the columns in $\mathbf{A}$ are 0 it's not going to evaluate when we really want that column to just stay zero, so what we do instead is:

$$\mathbf{T} = \mathbf{A}\mathbf{D}_{\mathbf{A}}^{-1} \tag{8}$$

Where we have:

$$\mathbf{D_A}^{-1} \quad : \quad \left[\mathbf{D}_{\mathbf{A}}^{-1}\right]_i = \begin{cases} 0 & , \quad [\mathbf{D_A}]_i = 0 \\ \frac{1}{\mathbf{D_A}}, & [\mathbf{D_A}]_i \neq 0 \end{cases} \tag{9}$$

Observe also that:

- $i = j$ in the above example
- $\mathbf{D_A} = \frac{1}{\mathbf{D_A}} \iff \mathbf{A}$ is ergodic
  - i.e. in this case the use of $[\quad]^{-1}$ is non standard in this case for want of notation.

References

[1] Joost Berkhout and Bernd F. Heidergott. "Ranking Nodes in General Networks: A Markov Multi-Chain Approach". In: *Discrete Event Dyn Syst* 28.1 (Mar. 1, 2018). The choice of damping factor of Googles page rank might have a large impact on the values given to vertices.
This suggests an approach that uses structural network dynatims to provide an appropriate score distribution.
The method implemented is not something I have come yet to understand, but it could be very interesting to see:
- how it relates to the power walk method
- whether or not it could offer insightts into the convergence and stability of the power walk method
- Whether or not the method would be compatible with negatively weighted edges., pp. 3–33. ISSN: 1573-7594. DOI: 10.1007/s10626-017-0248-7. URL: https://doi.org/10.1007/s10626-017-0248-7 (visited on 08/19/2020) (cit. on pp. 4, 7, 8).

[2] Berkley University. *Business Library | UC Berkeley Library*. URL: https://www.lib.berkeley.edu/libraries/business-library (visited on 08/19/2020) (cit. on p. 2).

[3] Monica Bianchini, Marco Gori, and Franco Scarselli. "Inside PageRank". In: *ACM Trans. Inter. Tech.* 5.1 (Feb. 1, 2005). This is a discussion on the stability, complexity and critical role of parameters involved in the computation., pp. 92–128. ISSN: 15335399. DOI: 10.1145/1052934.1052938. URL: http://portal.acm.org/citation.cfm?doid=1052934.1052938 (visited on 08/18/2020) (cit. on pp. 4, 5, 7).

[4] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. "PageRank as a Function of the Damping Factor". In: *Proceedings of the 14th International Conference on World Wide Web*. WWW '05. New York, NY, USA: Association for Computing Machinery, May 10, 2005, pp. 557–566. ISBN: 978-1-59593-046-0. DOI: 10.1145/1060745.1060827. URL: http://doi.org/10.1145/1060745.1060827 (visited on 08/19/2020) (cit. on pp. 5, 7–9).

[5] Michael Brinkmeier. "PageRank Revisited". In: *ACM Transactions on Internet Technology* 6.3 (Aug. 2006), pp. 282–301. ISSN: 15335399. DOI: 10.1145/1151087.1151090. URL: http://ezproxy.uws.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=22173011&site=ehost-live&scope=site (visited on 08/19/2020) (cit. on pp. 4, 9).

[6] Kurt Bryan and Tanya Leise. "The $25,000,000,000 Eigenvector: The Linear Algebra behind Google". In: *SIAM Review* 48.3 (2006), pp. 569–581. ISSN: 0036-1445. JSTOR: 20453840 (cit. on pp. 5, 7).

[7] *Category:Mathematics*. In: *Wikipedia*. Dec. 4, 2019. URL: https://en.wikipedia.org/w/index.php?title=Category:Mathematics&oldid=929215996 (visited on 08/19/2020) (cit. on p. 2).

[8] Cristobald de Kerchove, Laure Ninove, and Paul van Dooren. "Maximizing PageRank via Outlinks". In: *Linear Algebra and its Applications* 429.5-6 (Sept. 2008), pp. 1254–1276. ISSN: 00243795. DOI: 10.1016/j.laa.2008.01.023. URL: https://linkinghub.elsevier.com/retrieve/pii/S0024379508000529 (visited on 08/21/2020) (cit. on p. 11).

[9] Ayman Farahat et al. "Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization". In: *SIAM Journal on Scientific Computing; Philadelphia* 27.4 (2006), p. 21. ISSN: 10648275. DOI: http://dx.doi.org.ezproxy.uws.edu.au/10.1137/S1064827502412875. URL: http://search.proquest.com/docview/921166362/abstract/9367723222824064PQ/1 (visited on 08/19/2020) (cit. on p. 5).

[10] K M Frahm, A D Chepelianskii, and D L Shepelyansky. "PageRank of Integers". In: *J. Phys. A: Math. Theor.* 45.40 (Oct. 12, 2012), p. 405101. ISSN: 1751-8113, 1751-8121. DOI: 10.1088/1751-8113/45/40/405101. URL: https://iopscience.iop.org/article/10.1088/1751-8113/45/40/405101 (visited on 08/21/2020) (cit. on p. 9).

[11] Hwai-Hui Fu, Dennis K. J. Lin, and Hsien-Tang Tsai. "Damping Factor in Google Page Ranking". In: *Applied Stochastic Models in Business and Industry* 22.5-6 (2006), pp. 431–444. ISSN: 1526-4025. DOI: 10.1002/asmb.656. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.656 (visited on 08/19/2020) (cit. on pp. 4, 7, 8).

[12] Gabor Csardi et al. *Igraph R Manual Pages*. May 9, 2019. URL: https://igraph.org/r/doc/as_adjacency_matrix.html (visited on 08/19/2020) (cit. on p. 3).

[13] *Google Books Ngram Viewer*. URL: https://books.google.com/ngrams/graph?content=Zettelkasten&corpus=26&year_end=2019&year_start=1800&smoothing=3&direct_url=t1%3B%2CZettelkasten%3B%2Cc0#t1%3B%2CZettelkasten%3B%2Cc0 (visited on 08/19/2020) (cit. on p. 2).

[14] Hektor Haarkötter. *"Alles Wesentliche findet sich im Zettelkasten"*. URL: https://www.heise.de/tp/features/Alles-Wesentliche-findet-sich-im-Zettelkasten-3398418.html (visited on 08/19/2020) (cit. on p. 2).

[15] Eloise L. Harbeson. "Teaching Reference and Bibliography: The Pathfinder Approach". In: *Journal of Education for Librarianship* 13.2 (1972), p. 111. ISSN: 00220604. DOI: 10.2307/40322211. JSTOR: 10.2307/40322211 (cit. on p. 2).

[16] Harvard University. *Research Guides*. URL: https://guides.library.harvard.edu/ (visited on 08/19/2020) (cit. on p. 2).

[17] Taher Haveliwala and Sepander Kamvar. "The Second Eigenvalue of the Google Matrix". In: *Stanford Technical Report* (2003). URL: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwi5x7iBhqnrAhVTfisKHQp8CrYQFjAAegQIBRAB&url=https%3A%2F%2Fnlp.stanford.edu%2Fpubs%2Fsecondeigenvalue.pdf&usg=AOvVaw3Em9lm2qOuWEN23PXhUS8J (cit. on p. 7).

[18] *Help:Categories - MediaWiki*. URL: https://www.mediawiki.org/wiki/Help:Categories (visited on 08/19/2020) (cit. on p. 2).

[19] Myshkin Ingawale et al. "Network Analysis of User Generated Content Quality in Wikipedia". In: *Online Information Review* 37.4 (Jan. 1, 2013). Is there a relationship between content quality and the structure of connections? Can high quality Wikipedia pages be used as a benchmark for the structure of connections. The network structure of interactions between articles plays an important role in the emergence of quality. High quality articles clusture in hubs., pp. 602–619. ISSN: 1468-4527. DOI: 10.1108/OIR-03-2011-0182. URL: https://doi.org/10.1108/OIR-03-2011-0182 (visited on 08/17/2020) (cit. on p. 8).

[20] Sepandar Kamvar, Taher Haveliwala, and Gene Golub. "Adaptive Methods for the Computation of PageRank". In: *Linear Algebra and its Applications*. Special Issue on the Conference on the Numerical Solution of Markov Chains 2003 386 (July 15, 2004), pp. 51–65. ISSN: 0024-3795. DOI: 10.1016/j.laa.2003.12.008. URL: http://www.sciencedirect.com/science/article/pii/S0024379504000023 (visited on 08/19/2020) (cit. on pp. 4, 5, 11).

[21] Moshe Koppel and Nadav Schweitzer. "Measuring Direct and Indirect Authorial Influence in Historical Corpora". In: *Journal of the Association for Information Science and Technology* 65.10 (2014), pp. 2138–2144. ISSN: 2330-1643. DOI: 10.1002/asi.23118. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23118 (visited on 08/21/2020) (cit. on p. 4).

[22] Amy N. Langville and Carl D. Meyer. "A Reordering for the PageRank Problem". In: *SIAM Journal on Scientific Computing; Philadelphia* 27.6 (2006), p. 9. ISSN: 10648275. DOI: http://dx.doi.org.ezproxy.uws.edu.au/10.1137/040607551. URL: http://search.proquest.com/docview/921138313/abstract/24AFC1417CF6412BPQ/1 (visited on 08/19/2020) (cit. on pp. 7, 8).

[23] Amy N. Langville and Carl D. Meyer. "A Survey of Eigenvector Methods for Web Information Retrieval". In: *SIAM Review* 47.1 (2005), pp. 135–161. ISSN: 0036-1445. JSTOR: 20453606 (cit. on p. 8).

[24] Larry Page and Sergey Brin. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Computer Networks and ISDN Systems* 30.1-7 (Apr. 1, 1998), pp. 107–117. ISSN: 0169-7552. DOI: 10.1016/S0169-7552(98)00110-X. URL: http://www.sciencedirect.com/science/article/pii/S016975529800110X (visited on 08/19/2020) (cit. on pp. 2, 3, 5, 7).

[25] Antonis Loizou and Srinandan Dasmahapatra. "Using Wikipedia to Alleviate Data Sparsity Issues in Recommender Systems". In: *2010 Fifth International Workshop Semantic Media Adaptation and Personalization*. 2010 5th International Workshop on Semantic Media Adaptation and Personalization (SMAP). For Recommender systems with limited access to data, Wikipedia can be used as an analogue with respect to connections to significantly improve performance. Limassol, Cyprus: IEEE, Dec. 2010, pp. 104–111. ISBN: 978-1-4244-8603-8. DOI: 10.1109/SMAP.2010.5706870. URL: http://ieeexplore.ieee.org/document/5706870/ (visited on 08/17/2020) (cit. on pp. 2, 8).

[26] *Mathonline*. URL: http://mathonline.wikidot.com/ (visited on 08/19/2020) (cit. on p. 2).

[27] Abbas Mehrabian and Nick Wormald. "It's a Small World for Random Surfers". In: *Algorithmica* 76.2 (Oct. 1, 2016). Graphs can be generated in order to model real world networks, these models can use the degree or page rank of a given vertex as a parameter to create the next vertex in generating the graph.

This paper discusses upper and lower bounds for the diameter of a graph generated using random-surfer web-graph model., pp. 344–380. ISSN: 1432-0541. DOI: 10.1007/s00453-015-0034-6. URL: https://doi.org/10.1007/s00453-015-0034-6 (visited on 08/18/2020) (cit. on p. 9).

[28] MIT. *Research Guides & Expert Librarians | MIT Libraries*. URL: https://libraries.mit.edu/experts/ (visited on 08/19/2020) (cit. on p. 2).

[29] Paula Moskowitz. *Library Guides: Wikipedia: Should You Use Wikipedia?* URL: https://mville.libguides.com/c.php?g=370066&p=2500344 (visited on 08/19/2020) (cit. on p. 2).

[30] Waleed Nema and Yinshan Tang. "Consensus-Based Ranking of Wikipedia Topics". In: *Proceedings of the International Conference on Web Intelligence*. WI '17. New York, NY, USA: Association for Computing Machinery, Aug. 23, 2017, pp. 114–124. ISBN: 978-1-4503-4951-2. DOI: 10.1145/3106426.3106529. URL: http://doi.org/10.1145/3106426.3106529 (visited on 08/19/2020) (cit. on pp. 7, 8).

[31] Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. "Stable Algorithms for Link Analysis". In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New York, NY, USA: Association for Computing Machinery, Sept. 1, 2001, pp. 258–266. ISBN: 978-1-58113-331-8. DOI: 10.1145/383952.384003. URL: http://doi.org/10.1145/383952.384003 (visited on 08/19/2020) (cit. on pp. 5, 7, 8).

[32] W. Keith Nicholson. *Linear Algebra with Applications*. Toronto: McGraw-Hill Ryerson, 2009. ISBN: 978-0-07-098510-0 (cit. on p. 3).

[33] *Overview • Zettelkasten Method*. URL: https://zettelkasten.de/posts/overview/ (visited on 08/19/2020) (cit. on p. 2).

[34] laurence Park and Simeon Simoff. *Power Walk | Proceedings of the 18th Australasian Document Computing Symposium*. Dec. 5, 2013. URL: https://dl-acm-org.ezproxy.uws.edu.au/doi/10.1145/2537734.2537749 (visited on 08/01/2020) (cit. on pp. 2, 4, 5, 7).

[35] Laurence A. F. Park and Kotagiri Ramamohanarao. "Mining Web Multi-Resolution Community-Based Popularity for Information Retrieval". In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management - CIKM '07*. The Sixteenth ACM Conference. Lisbon, Portugal: ACM Press, 2007, p. 545. ISBN: 978-1-59593-803-9. DOI: 10.1145/1321440.1321517. URL: http://portal.acm.org/citation.cfm?doid=1321440.1321517 (visited on 08/21/2020) (cit. on p. 8).

[36] Laurence A. F. Park and Simeon Simoff. "Power Walk: Revisiting the Random Surfer". In: *Proceedings of the 18th Australasian Document Computing Symposium*. ADCS '13. Brisbane, Queensland, Australia: Association for Computing Machinery, Dec. 5, 2013, pp. 50–57. ISBN: 978-1-4503-2524-0. DOI: 10.1145/2537734.2537749. URL: http://doi.org/10.1145/2537734.2537749 (visited on 07/31/2020) (cit. on pp. 5, 7).

[37] Glen Pringle, Lloyd Allison, and David L. Dowe. "What Is a Tall Poppy among Web Pages?" In: *Computer Networks and ISDN Systems* 30.1-7 (Apr. 1998), pp. 369–377. ISSN: 01697552. DOI: 10.1016/S0169-7552(98)00061-0. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169755298000610 (visited on 08/21/2020) (cit. on p. 9).

[38] *Science Tracer Bullets (Science Reference Services, Science, Technology, and Business Division, Library of Congress)*. URL: https://www.loc.gov/rr/scitech/tracer-bullets/ (visited on 08/19/2020) (cit. on p. 2).

[39] Xueyuan Tan. "A New Extrapolation Method for PageRank Computations". In: *Journal of Computational and Applied Mathematics* 313 (Mar. 15, 2017). This is exactly what laurence was saying and also what Bryan and Leise said., pp. 383–392. ISSN: 0377-0427. DOI: 10.1016/j.cam.2016.08.034. URL: http://www.sciencedirect.com/science/article/pii/S0377042716304034 (visited on 08/19/2020) (cit. on pp. 5, 7, 8).

[40] *Top 100 Wiki Pages on Brilliant | Brilliant Math & Science Wiki*. URL: https://brilliant.org/wiki/best/ (visited on 08/19/2020) (cit. on p. 2).

[41] Luigina Vileno. “From Paper to Electronic, the Evolution of Pathfinders: A Review of the Literature”. In: *Reference Services Review* 35.3 (Jan. 1, 2007), pp. 434–451. ISSN: 0090-7324. DOI: 10.1108/00907320710774300. URL: https://doi.org/10.1108/00907320710774300 (visited on 08/19/2020) (cit. on p. 2).

[42] *Wikipedia:Size of Wikipedia*. In: *Wikipedia*. Aug. 1, 2020. URL: https://en.wikipedia.org/w/index.php?title=Wikipedia:Size_of_Wikipedia&oldid=970572970 (visited on 08/20/2020) (cit. on p. 4).

[43] Hui Zhang et al. “Making Eigenvector-Based Reputation Systems Robust to Collusion”. In: *Algorithms and Models for the Web-Graph*. Ed. by Stefano Leonardi. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 92–104. ISBN: 978-3-540-30216-2. DOI: 10.1007/978-3-540-30216-2_8 (cit. on p. 5).

[44] Jin Zhang and Alexandra Dimitroff. “The Impact of Webpage Content Characteristics on Webpage Visibility in Search Engine Results (Part I)”. In: *Information Processing & Management*. Cross-Language Information Retrieval 41.3 (May 1, 2005), pp. 665–690. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2003.12.001. URL: http://www.sciencedirect.com/science/article/pii/S0306457303001122 (visited on 08/19/2020) (cit. on p. 11).

[45] Qingyang Zhang. “A Modified PageRank Algorithm for Biological Pathway Ranking”. In: *Stat* 7.1 (2018). e204 sta4.204
The page rank method can be used for modelling gene expression., e204. ISSN: 2049-1573. DOI: 10.1002/sta4.204. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.204 (visited on 08/19/2020) (cit. on p. 8).