PageRank’s ability to track webpage quality: reconciling Google’s wisdom-of-crowds justification with the scale-free structure of the web

George Masterton a, Erik J. Olsson b,*

a Department of Culture and Communication, Linköping University, Sweden
b Department of Philosophy, Lund University, Sweden

* Corresponding author.
E-mail address: erik_j.olsson@fil.lu.se (E.J. Olsson).

Abstract

We address the fundamental question why we should use PageRank and similar link-based algorithms in search engines, if at all. In a legendary article from 1998, the Google founders gave an intriguing wisdom-of-crowds justification for PageRank according to which the latter tracks quality online. This striking suggestion stands in contrast to the view that PageRank merely tracks what is popular. However, Masterton and Olsson (2017) showed that web-ecologies generated by Google-like assumptions essentially fail to reflect the scale-free structure of the web. They pointed to attraction to popularity or a rich-get-richer effect being the likely main cause of scalefreeness. In this article, we explore dual models of linking behavior, i.e. models that combine attraction to importance (quality) with attraction to popularity. Our results, obtained through computer simulation, indicate that there exist dual models that give rise both to a wisdom-of-crowds effect for PageRank and to scale-free web-graphs, thus giving a partial vindication of the wisdom-of-crowds thesis for the real web. Future work should explore larger web-graphs as well as other aspects pertaining to the empirical plausibility of dual linking models.
Keywords: Information science, Philosophy

1. Introduction

In their legendary 1998 article, the Google founders — Sergey Brin and Larry Page — gave what can be reasonably understood as an informal wisdom-of-crowds justification for their well-known PageRank ranking algorithm, arguing that the latter, by aggregating in-coming links in a clever way, tracks what is important on the World Wide Web (WWW). We refer to Brin et al. (1998) for a formal definition of PageRank. Drawing an analogy with academic citation, Brin and Page wrote (p. 109):

The citation (link) graph of the Web is an important resource that has largely gone unused in existing Web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow rapid calculation of a Web page’s ‘PageRank’, an objective measure of its citation importance that corresponds well with people’s subjective idea of importance.

Similarly (p. 110),

Another intuitive justification is that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Intuitively, pages that are well cited from many places around the Web are worth looking at. Also, pages that have perhaps only one citation from something like the Yahoo! homepage are also generally worth looking at. If a page was not high quality, or was a broken link, it is quite likely that Yahoo’s homepage would not link to it. PageRank handles both these cases and everything in between by recursively propagating weights through the link structure of the Web.

We will follow the Google founders in using the term “importance” throughout in the sense of “importance based on quality content” as opposed to importance based on mere popularity.

Why is it that “pages that are well cited from many places around the Web are worth looking at”? This is where the idea of wisdom-of-crowds enters. The proposal is that webmasters in charge of webpages are individually somewhat attracted to important webpages in the sense that they have an inclination to link to what is high quality. What PageRank does is aggregating these individual, somewhat “shaky”, judgements into a reliable collective judgement of relative quality. As a prominent commentator explains (Surowiecki, 2004, p. xiv),

[This intelligence, or what I’ll call “the wisdom of crowds” is at work in the world in many different guises. It’s the reason why the internet search engine
Google can scan a billion webpages and find the one page that has the exact piece of information you were looking for.

We refer to the thesis in question — that “[t]he analysis of link structure via PageRank allows Google to evaluate the quality of Web pages” (Brin and Page, 1998, p. 116) and that this is due to the wisdom-of-crowds — as the wisdom-of-crowds-justification-for-PageRank (WCJPR).

We wish to stress that referring to WCJPR is strikingly different from saying that PageRank merely tracks what is popular on the web, or popular with webmasters. What is popular need not be important in the sense of having high quality content. Clearly, the case for using PageRank for ranking webpages in search engines is significantly stronger with WCJPR than without it; indeed, without it, we may wonder why PageRank should be used at all. If PageRank tracks what is really important on the basis of high quality content as opposed to mere popularity, this is a significant fact that counts strongly in favor of using that algorithm in search engines because at the end of the day we want search engines to give us high quality and not merely popular information on various topics of interest. This is why we believe that the topic we are addressing here is fundamental for Google. Indeed, broadly interpreted, the topic concerns the very rationale for using link-based ranking for determining visibility on the web.

Against the backdrop of what we just wrote, it must come as a surprise to the reader that, until recently, there was no rigorous theoretical work on the basis for the WCJPR thesis. Brin and Page went on, famously, to create one of the most successful enterprises the world has ever seen without publishing more on the subject, and for a long time no one in academia followed their lead. One reason why the topic remained underdeveloped may be the fact that the usual way of thinking about webmasters’ linking behavior is not easily combined with the idea of wisdom-of-crowds. This goes in particular for the influential preferential attachment model due to Barabási and Albert (1999) which assumes that webmasters link to what other webmasters have linked to, thus leaving no room for attraction to quality itself to play any role in the linking process. Without such attraction there is no wisdom-of-crowds effect to be had.

In fact, 18 years would pass until someone would demonstrate the formal validity of the Google founders’ reasoning. To the best of our knowledge, this was first done in Masterton et al. (2016) which features a theorem to the effect that PageRank indeed tracks importance perfectly in the limit as the web-graph grows given, roughly, that webmasters are somewhat attracted to important content and link independently (as explained below). Masterton et al. (2016) proved similar theorems for the more elementary In-Degree algorithm which, unlike PageRank, simply counts the number of in-links. As detailed in Masterton et al. (2016), sec. 2, this approach to linking was
inspired by the famous Condorcet jury theorem for majority voting in political science. Recently, strikingly similar models have been studied in computer science under the heading of crowdsourcing. For useful overviews, the reader may consult Doan et al. (2011), Zhang et al. (2016) and Zheng et al. (2017). (We would like to thank an anonymous referee for directing us to this literature.)

It would now seem that WCJPR had been spectacularly vindicated. However, it is one thing to prove the WCJPR thesis as a mathematical theorem, i.e. that the conclusion holds given the assumptions, and quite another to show that the assumptions of the theorem are true or at least plausible for the real web. As Masterton and Olsson (2017) pointed out, the empirical adequacy of the Google-inspired model introduced by Masterton, Olsson and Angere (henceforth, MOA) can be questioned. The Barabási-Albert (BA) model is realistic in the minimalist sense that it gives rise to scale-free networks not dissimilar to the WWW. To be sure, the MOA model, too, can generate scale-free networks of the right kind for the WWW, but the way this is accomplished is, as we will see, trivial and uninteresting. Instead, the fact that webmasters link to what other webmasters have linked to, broadly along the lines of the BA model, is the likely main cause of scalefreeness.

Thus, Masterton and Olsson concluded that there is at present no rigorous demonstration of the WCJPR thesis that is also based on realistic assumptions about the web. The models of linking behavior that exist either do not allow for a rigorous theoretical demonstration (the BA model) or they fail to be faithful to the structure of the WWW and the likely cause of that structure (the MOA model). This means that the very idea of using PageRank to rank webpages on the real web, pace Brin and Page, is left with little or no support from appeal to the wisdom-of-crowds.

As a remedy, Masterton and Olsson suggested that future work should explore “dual models” of the web, i.e. models that combine the BA model and the MOA model into one account of linking behavior. They conjectured that there are dual models that are realistic models of the web in terms of its scale-free structure and at the same time allow for the rigorous demonstration of the WCJPR thesis.

Our main task in the present article is to investigate the theoretical validity of this conjecture by studying the properties of a number of such dual models by means of computer simulations of web-graphs of up to 15,000 pages. As we will see, the simulations indicate that the conjecture is true for several dual models. Thus, taking into account the scalefreeness of the WWW does not preclude PageRank from being justified with reference to its capacity for tracking quality on the web, neither does incorporating a preferential attachment component. This is a partial vindication of the WCJPR thesis, as there are other empirical properties of the web to take into account. Future work should explore larger web-graphs and other aspects of the empirical plausibility of the dual models in question.
2. Background

Here we recapitulate some of the background. The interested reader should consult Masterton and Olsson (2017) for more details on the background and motivation for our research, as well as for additional references. As is easily appreciated, the World Wide Web can be seen as a directed graph of links, sometimes called hyperlinks, connecting web-pages (Broder et al., 2000). Our focus here is on the structure of this web-graph. The degree distribution \( P(N) \) of a graph of order \( n \), for each number \( N \) between 0 and \( n - 1 \), is the proportion of vertices with \( N \) links. Directed graphs have both an in-degree distribution and an out-degree distribution. In this paper, we are interested in directed graphs and in-link based metrics, whence we will refer to in-degree distributions as simply degree distributions. A scale free degree distribution is one that conforms to an inverse power law. For example, the World Wide Web (WWW) has been found to have a scale free degree distribution with an exponent of around 2.1 (Albert et al., 1999; Broder et al. 2000).

A natural question to ask is what accounts for this topology of the WWW. In other words, what kind of linking-dispositions of webmasters give rise to the observed structural features? A common answer in the literature is that the degree distribution of the WWW arises as the effect of preferential attachment. According to this model, the degree distribution of the internet is a consequence of webmasters being attracted to popular and already linked-to pages when creating new links, a phenomenon also known as the “rich get richer” effect.

One such preferential attachment model is the Barabási-Albert model (BA model). We refer to Barabási and Albert (1999) for details. That webmasters are more likely to link to pages with lots of in-links to them is as a plausible assumption for several reasons. First, these pages are more easily found by search engines (Thelwall, 2013). Second, if one randomly navigates the WWW, then the probability of visiting a page in this manner is equal to its PageRank, and pages with high PageRanks will tend to have high numbers of in-links (Brin et al., 1998). Third, it is plausible that we are in part guided by the opinion of others as to what is worth linking to, so that if many are linking to something, we are also likely to do so. Finally, the fact that many link to a given page may be a sign of its real importance. If so, our search for important information may be what explains our attraction to popularity.

However, in Masterton and Olsson (2017), we concluded that, if in-link-based metrics such as PageRank and In-Degree (just counting the number of in-links) are to be useful as signals as to what is important, then there must be more to the linking dynamics than preferential attachment, in particular, there must be an attraction-to-importance component to those dynamics. Masterton et al. (2016), henceforth MOA, developed two vertex affinity models tailored to the web-ecology assumed
by Google in their wisdom-of-crowds justification of PageRank’s role in their search engine, as expressed in our introductory quotes.

MOA argued that there are two assumptions in play. First, there is the assumption that webmasters are, to some degree, attracted to importance in the sense that *ceteris paribus* the probability of them assigning a link to a page (target page) will depend positively on the importance of the target page. Second, it is also assumed that the strength of this attraction to importance varies with the competence of the webmaster. In particular, the more important the source page, the greater the tendency of its webmaster to link to other important pages, whereas a webmaster administering a less important page will be more random in her linking behavior. The basic model introduced by MOA is governed by the first assumption. An extended model was introduced to do justice to both.

In both models, the web is represented as a directed graph, where vertices correspond to webpages and directed edges to hyperlinks. The vertices have a single attribute: importance. Page importance \( p_i \in [0, 1] \) is sampled from distributions truncated to the unit interval. While the model is completely flexible as to what distribution to sample from, our choice will be to sample from negative exponential and Pareto distributions.

The parameters in the models include the size of the web-graph \( n \in \mathbb{N}^+ \) and the parameters determining the importance sampling; in the case of negative exponential distributions over the unit interval the expectation (expected page importance = \( \alpha \in (0, 0.5) \)) and in the case of the Pareto distributions a minimum value (minimum page importance = \( m_p \in (0, 0.2) \)) and scale (\( \gamma \in \mathbb{R}^+ \)). Here, 0-importance indicates the complete lack of importance-determining qualities and 1-importance their complete presence. The interpretation of page importance based on quality content is deliberately vague in the interest of maximum generality. What we have in mind are factors like “relevance”, “truthfulness”, “completeness of coverage” and the like.

In the basic model, the probability that the \( j \)-th page links to the \( i \)-th page is a function of the importance of the \( i \)-th page, i.e. \( g(I_i) : [0, 1] \mapsto [0, 1] \). The parameters are probability scaling (\( P_s \in [0, 1] \)), which determines overall link density, and the probability weighting (\( P_w \in \mathbb{R}^+ \)), which determines the linearity of the dependence of linking probability on target page importance. Hence, the probability for any page \( j \) to link to a distinct page \( i \) is given by Eq. (1):

\[
P_B(j \rightarrow i) = g(I_i) = P_s(I_i)^{P_w}.
\] (1)

In the extended model, the probability of page \( j \) linking to page \( i \) depends on the importance of both pages. There are many ways in which this dependence can be spelled out, as explained in Masterton et al. (2016). For instance, the function in Eq. (2) satisfies some reasonable desiderata:
\[ P_E(j \rightarrow i) = f(I_i, I_j) = P_S(\alpha) P_w(I_j^{1-h(I_j)}) (I_j)^{P_w (I_j)} \]  \hspace{1cm} (2)

so long as \( h(0) = 0 \) and \( h(1) = 1 \) for \( h(I_j) \): [0, 1] \( \rightarrow \) [0, 1].

Following MOA, we shall refer to \( h(I_j) \) as the linking competence function. A plausible candidate is \( h(I_j) = (I_j)^C \) where \( C \) is the competence factor. We will assume that linking competence scales linearly with page importance \((C = 1)\), but competence might trail page importance \((C > 1)\) or it might advance on page importance \((0 \leq C < 1)\). Consequently, the basic model can be seen as corresponding to the limit case in which all webmasters are fully competent in their linking \((C = 0)\). When \( C \) increases, the linking probability decreasingly dependent upon the importance of the target page to the point where only the webmasters of very important pages link in a manner dependent upon the importance of the target page. Indeed, where \( C \) goes to infinity, the linking probability becomes constant and equal to the expected linking probability.

As stated previously, we here assume that linking competence scales linearly with page importance \((h(I_j) = I_j)\), so that for the purposes of this article Eq. (3) gives the linking probability in the extended model:

\[ P_E(j \rightarrow i) = P_S(\alpha) P_w(I_j^{1-h(I_j)}) (I_j)^{P_w (I_j)} . \]  \hspace{1cm} (3)

In either model, links are created by metaphorically flipping a coin for each prospective link, with a bias for heads equal to the probability for that link, adding the link if the coin lands heads. As observed in Masterton et al. (2016), this makes link creation for a given page a Bernoulli trial in the basic model and a Poisson trial in the extended model. Sampling a web-graph means completing this process for the web-ecology jointly specified by the model and parameter configuration in question.

Masterton and Olsson (2017) prove that in the basic model, given linear \((P_w = 1)\) attraction to importance and maximal \((P_s = 1)\) link density, the degree distribution of a web-graph will almost surely converge on the importance distribution from which webpage importance was sampled. This result entails that if importance conforms to a power law, then the degree distribution of a web-graph generated in the basic model with \( P_w = 1 \) will converge on being scale-free in the limit. However, Masterton and Olsson (2017) concluded that this is an unconvincing way to obtain scalefreeness and that the more plausible explanation is that provided by the preferential attachment model. Hence, while it is reasonable to assume that attraction to importance plays a role in a realistic model of linking behavior, it is less plausible to claim that it is the sole factor. Rather, attraction to popularity will also play a role when explaining webmasters’ decisions to add links to other webpages.
3. Model

The upshot of the previous discussion is that dual models combining the two factors — attraction to popularity and attraction to importance — are desirable not only because they are more empirically adequate, but also because they more faithfully represent the balance of preferences plausibly in play when webmasters decide what links to assign from the webpages in their charge. Thus, even though one can parameterize the basic and extended models in such a way that large web-graphs with the same degree distribution as the WWW will be generated with practical certainty, arguably these models could be made more faithful by the inclusion of attraction to popularity elements.

Masterton and Olsson (2017) suggested combining the PA and MOA models linearly in the manner of Lee et al. (2015). Since there are two MOA models — the basic and extended versions — this gives rise to two linear models each characterized by its own linking probability function. For each model the linking probability function holds for entrant page $j$ with respect to incumbent page $i$. The preference that the webmaster of the $j$-th page has for important, over popular, pages $\left( \mu_j \in [0, 1] \right)$ is the weight of the linear sum and is an additional attribute of page $j$ in these models that is sampled from some distribution. We used truncated exponential distributions characterized by the expected preference (for important over popular pages) weight $(\epsilon_{Pw} (0, 0.5) \cap (0.5, 1))$

Linear model 1: $P_{B+BA}(j \rightarrow i) = \mu_j \cdot P_{B}(j \rightarrow i) + (1 - \mu_j) \cdot P_{BA}(j \rightarrow i)$

Linear model 2: $P_{E+BA}(j \rightarrow i) = \mu_j \cdot P_{E}(j \rightarrow i) + (1 - \mu_j) \cdot P_{BA}(j \rightarrow i)$

These models are illustrated in Figs. 1 and 2, respectively.

**Fig. 1.** To the left is a typical web-graph generated by the basic model of Masterton et al. (2016) in a particular parameter configuration ($\alpha = 0.25$, $P_{s} = 1$, $P_{w} = 2$, $n = 32$), and to the right is a typical web-graph generated by Linear model 1 in the same parameter configuration with preferences split evenly between importance and popularity ($\epsilon_{Pw} = 0.5$).
In these dual models, starter graphs are not required. So long as \( \sum_{k=1}^{N_k} N_k \) is prohibited from taking the value 0 when there are no links in the graph \( \sum_{k=1}^{N_k} N_k \) is assigned a value of 1, otherwise it takes its appropriate value — one can “grow” the graph from a single page in both models. If one proceeds in this fashion, then, until a page has acquired a non-zero In-Degree, only the attraction to importance component of the linking probability will determine linking to that page. The reciprocal probability for linking from incumbent pages to the new page is just \( \mu_i \cdot P_B(i \rightarrow j) \) in linear model 1 and \( \mu_i \cdot P_E(i \rightarrow j) \) in linear model 2. The idea here is that as the new page has not had any time to gain popularity, the only reason to link to it is its importance, which the webmaster of the \( i \)-th page values to extent \( \mu_i \).

These linear models can be thought of as modeling web-ecologies where webmasters have two distinct preferences that they weigh against each other. That is, in these models webmasters have a preference for important pages and a preference for popular pages, and each finds his/her own balance between the two. But perhaps webmasters have a single complex preference for important and popular pages. How would this best be modeled? We propose that this situation could be represented by geometric models such as those below:

Geometric model 1: \( P_{B \rightarrow BA}(j \rightarrow i) = P_B(j \rightarrow i) \cdot P_{BA}(j \rightarrow i) \).

Geometric model 2: \( P_{E \rightarrow BA}(j \rightarrow i) = P_E(j \rightarrow i) \cdot P_{BA}(j \rightarrow i) \).

These models are illustrated in Figs. 3 and 4, respectively.

Fig. 2. To the left is a typical web-graph generated in the extended model of Masterton et al. (2016) in a particular parameter configuration (\( \alpha = 0.25, P_s = 1, P_w = 2, n = 32 \), and to the right is a typical web-graph generated by Linear model 2 in the same parameter configuration with preferences split evenly between importance and popularity (\( \epsilon P_w = 0.5 \)).
To prevent problems related to dividing and multiplying by zero, we stipulate that until \( \sum_{k=1}^{j-1} N_k \) is greater than 0 and \( N_i \) is greater than some threshold \( T \in [1, n-1] \), we can use \( P_{B,BA}(j \rightarrow i) = P_{B}(j \rightarrow i) \) and \( P_{E,BA}(j \rightarrow i) = P_{E}(j \rightarrow i) \). The results were obtained with \( T = 1 \), which is to say that the popularity of a page was assumed to play a role in determining links to it as soon as that page’s popularity exceeded null. The reciprocal probabilities for links from the incumbent pages to the new page are \( P_{B}(j \rightarrow i) \) in Geometric model 1 and \( P_{E}(j \rightarrow i) \) in Geometric model 2. Again, the idea is that, as the new page has not had any time to gain popularity, the only reason to link to it, in the first instance, is its importance. This means that in the geometric models the probability for an incumbent linking to an entrant is higher than the

**Fig. 3.** To the left is again the typical web-graph generated in the basic model of Masterton et al. (2016) in parameter configuration \((\alpha = 0.25, P_s = 1, P_w = 2, n = 32)\), and to the right is a typical web-graph generated by Geometric model 1 in the same parameter configuration.

**Fig. 4.** To the left is the typical web-graph generated in the extended model of Masterton et al. (2016) in parameter configuration \((\alpha = 0.25, P_s = 1, P_w = 2, n = 32)\), and to the right is a typical web-graph generated by Geometric model 2 in the same parameter configuration.

To prevent problems related to dividing and multiplying by zero, we stipulate that until \( \sum_{k=1}^{j-1} N_k \) is greater than 0 and \( N_i \) is greater than some threshold \( T \in [1, n-1] \), we can use \( P_{B,BA}(j \rightarrow i) = P_{B}(j \rightarrow i) \) and \( P_{E,BA}(j \rightarrow i) = P_{E}(j \rightarrow i) \). The results were obtained with \( T = 1 \), which is to say that the popularity of a page was assumed to play a role in determining links to it as soon as that page’s popularity exceeded null. The reciprocal probabilities for links from the incumbent pages to the new page are \( P_{B}(j \rightarrow i) \) in Geometric model 1 and \( P_{E}(j \rightarrow i) \) in Geometric model 2. Again, the idea is that, as the new page has not had any time to gain popularity, the only reason to link to it, in the first instance, is its importance. This means that in the geometric models the probability for an incumbent linking to an entrant is higher than the
probability for the entrant linking to the incumbent. Thus the geometric models almost certainly give entrants too many links. This is an issue that should be corrected were our intention to produce realistic models of the ecology of the internet, but that is not our goal. We merely wish to demonstrate that there are models of web-ecology incorporating preferential attachment that produce web-graphs with the right kind of degree distribution for the internet where link-based metrics are useful indicators of page quality/importance. For this more limited goal, some deviation from realism is acceptable. Note, however, that this same concern does not afflict the linear models.

Another idealization in these models is that incumbents do not revise links between each other (see Krapivsky et al., 2001, for a model that incorporates this feature). Again one could add such a feature to these models easily — possibly by making the probability for addition/deletion of a link directly/inversely proportional to the target’s importance and popularity — but as we are simulating the growth of the graphs over 1,000 steps, and as such revision is rare in real life, an idealization seemed justified.

Unlike the pure BA-model, these dual models all ensure that the importance of the pages involved is, in some way, reflected in the linking structure of the network concerned. This means that wisdom-of-crowds justification for using link-based metrics as measures of intrinsic importance has at least some prima facie plausibility in the web-ecologies characterized by these linking-probability functions.

Our main research question is whether any of these models is viable as a model of the WWW, given what we know about the degree distribution of that web-graph, and the extent to which there is overlap between those parameter configurations and models that produce the right kind of topology and those parameter configurations and models where PageRank or In-Degree is well correlated with webpage importance. In other words: do any of these four models generate scale-free web-graph topologies with exponents in the region of 2.1? If so, how robustly (for how wide and reasonable a region of the parameter space of the model) do they do so? Finally: how well do PageRank and In-Degree perform when/if they do?

4. Methodology

Our method of choice was computer simulation in NetLogo, a multi-agent programmable modeling environment due to Uri Wilenski. After exploring a wide parameter space it was found that scale-free networks did not arise in the dual models when important pages are common. For this reason we limited ourselves to distributions of importance with expectations that are less than 0.5. We considered both negative exponential distributions of importance and power law (Pareto) distributions.

Sampling importance values for the web-pages was performed by inverse transform sampling. To sample an importance value from such distributions in this way one
samples a real from a uniform distribution on the unit interval and inputs this real into the quantile function for the intended distribution. The value of that quantile function for that random seed is the sampled value.

For sampling from negative exponential distributions we assume that the probability that a page has an importance within \([a, b]\) is given by the integral, between these limits, of a negative exponential probability density function (pdf) truncated to \([0, 1]\). Let \(I_i\) be the page importance for the \(i\)-th page, let \(x_i\) be the \(i\)-th random number in \((0,1)\) and let \(\alpha \in (0, 0.5)\) be the expected page importance, then Eq. (4) gives the quantile function:

\[
I_i = -\alpha \times \ln \left(1 - x_i \left(1 - e^{-\frac{1}{\alpha}}\right)\right)
\]  

A nice property of this form of distribution is that it is fully specified by its expectation. Things are not this easy when it comes to sampling from Pareto distributions.

For mathematical simplicity we initially assume a Pareto distribution truncated to \([1, 10000]\). Sampling from such a distribution will return a real in this interval that we then convert into an importance value by dividing by 10,000. With the lower bound set to 1, no page ever has an importance of less than 0.0001. This is a small change from sampling from truncated exponential pdf’s but not one that will significantly alter topological results in graphs of 1,000 pages.

Eq. (5) gives the Pareto pdf truncated to \([1, 10000]\):

\[
\rho(x) = \frac{x^{\gamma-1}}{1 - \left(\frac{1}{10000}\right)^\gamma}
\]  

Where \(x_i\) is the random seed for the \(i\)-th page, the importance of the \(i\)-th page is given by 0.0001 times the value of the quantile function for the above pdf, as shown in Eq. (6):

\[
10000I_i = \left(1 - x_i \frac{10000^\gamma - 1}{10000^\gamma}\right)^{-\frac{1}{\gamma}}
\]  

Eq. (7) shows the expected importance for a given \(\gamma\):

\[
10000\alpha = \left(\frac{\gamma}{\gamma - 1}\right) \left(1 - \frac{1}{10000^\gamma}\right)
\]  

From this we can determine (numerically) the value of \(\gamma\) required to obtain a desired \(\alpha\) where the least importance a page can have is 0.0001. In particular, to have \(\alpha \in \{0.001, 0.01, 0.1, 0.25, 0.4\}\) requires \(\gamma \in \{0.97, 0.5, 0.017, -0.31, -0.66\}\). For a fixed minimum importance of 0.0001, these values of \(\gamma\) in the quantile function will produce distributions of page importance whose mean will almost surely...
converge on the desired expected importance in the infinite limit of sampling. Note that the scale factor $\gamma$ of truncated Pareto distributions can be negative.

By studying two distinct, reasonably plausible, distribution types for importance we hope to be able to check the stability of our results. Also, we wish to see whether having importance distributed according to a power law increases the tendency in our dual models to have power law degree distributions.

The parameter spaces we explored were those that, after preliminary examination, proved most interesting. For instance, we know from our previous results and theorem that the basic model only produces scale free web-graphs with the right scaling for the internet where importance is sampled from a Pareto distribution with an expectation of around 0.001. Further, through a cursory analysis of some simulation data it was clear that nothing interesting happens to the degree distribution of graphs generated in that model where attraction to importance is non-linear. As the basic model is a limit of Linear model 1 as $epw \rightarrow 1$, this determined the appropriate parameter space to explore for Linear model 1. While there were no theorems supporting the other parameter space choices, they were chosen by trial and error with an eye toward them being as similar as possible. For instance, it was discovered that unless expected importance was sufficiently high, Linear model 2 produced web-graphs with a link density too low to be assessed and it produced no scale-free web-graphs when attraction to importance was linear. While this piecemeal approach limited inter-model comparison, it enabled us to carry out the most instructive intra-model analysis. The parameter spaces explored were as follows:

**Linear model 1:** $Pw \in \{1\}$, $\alpha \in \{0.001, 0.01, 0.1\}$, $epw \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$

**Linear model 2:** $Pw \in \{2, 3, 4\}$, $\alpha \in \{0.1, 0.25, 0.4\}$, $epw \in \{0.01, 0.3, 0.5, 0.7, 0.9\}$

**Geometric model 1:** $Pw \in \{1, 2, 3, 4\}$, $\alpha \in \{0.001, 0.01, 0.1, 0.25\}$

**Geometric model 2:** $Pw \in \{1, 2, 3, 4\}$, $\alpha \in \{0.1, 0.25\}$

The model parameters held fixed in both cases were $Ps = 1$ and $n = 1000$.

When sampling from Pareto distributions the minimum importance a page can have was fixed at 0.0001. This meant that $\gamma$ had to be sampled from $\{0.97, 0.5, 0.017, -0.31, -0.66\}$ to effectively sample $\alpha$ from $\{0.001, 0.01, 0.1, 0.25, 0.4\}$

For each model, for each configuration, a web-graph was generated of order 1,000. The number of pages with 1 in-link, the number of pages with 2 in links up to the number of pages with 999 in-links were all determined and then exported. The log of the number of in-links was then plotted against the log of the number of pages with that number of in-links. If the topology of the network is scale-free — if the degree distribution follows a power law — then affine interpolation is a good fit (as measured by correlation coefficients) to the data in the log/log plot. The (negative)
gradient of the line of best fit is the power of the law for the degree distribution best supported by the data. The gradients and their correlation coefficients were recorded and can be found in section 6.

Once the extent to which a web-ecology generates a scale-free web-graph of the right kind for the internet was determined, the extent to which In-Degree and PageRank track page importance in those ecologies was investigated. The method here was to generate 20 web-graphs for each model parameterization specifying a unique web-ecology, determine the PageRank for each of the pages and check the degree of correlation between PageRank and importance in each of those 20 web-graphs. Finally, the average of those degrees of correlation was taken as the measure of how well PageRank tracks importance in the web-ecology specified by the model parameter configuration pair in question. This process was repeated for In-Degree. The results can be found in section 6. Degree of correlation between link-based metric and page importance in each web-graph was measured by the coefficient of determination relative to 3rd order polynomial regression, this order of regression having previously been established as generally neither grossly under- nor over-fitting the data (Masterton et al., 2016).

5. Results

5.1. Results for degree distribution

5.1.1. Linear model 1

As we noted, it has been established that the importance distribution is decisive in determining degree distribution in the basic model (Masterton and Olsson, 2017). Furthermore, where attraction to importance is linear and maximal ($P_s = P_w = 1$) a power law importance distribution will give rise to a power law degree distribution. Finally, if the scale factor of a Pareto importance distribution with minimum set at 0.0001 was set so that expected importance was in the region of 0.001, the ecology would generate web-graphs of 1,000 pages with scale-free degree distributions with exponents in the ball park of the accepted value of 2.1. All of this data suggested that the most interesting parameter space to explore for Linear model 1 — where the basic model is linearly combined with the BA preferential attachment model — would be in the region of where the basic model is known to produce the right topologies.

Perhaps unsurprisingly given what we know of the basic model, we found that where attraction to importance is linear and maximal, a power law distribution of importance ensures a power law degree distribution in the Linear model 1 web-graphs. The most obvious trend in the data is that the lower the expected importance and
the higher the expected preference for important pages, the higher the power of the degree distribution (Fig. 5).

This means that while none of the original configurations considered herein yielded a degree distribution with the right power for the World Wide Web, there should be plenty of parameter configurations for this model with the right balance of low expected importance and high preference for importance that will give topologies with the right degree distribution for the web. Moreover, those parameterizations will be fairly reasonable, if one allows that important pages are very rare.

To confirm these results we did a special series of simulations. Due to the very low expected page importance required to get into the relevant region of the parameter space, and the consequently very low link density, we needed much larger graphs (15,000 vertices) to be able to resolve degree distributions and a lower minimum value \((mpi = 1 \times 10^{-5})\) for pareto-distributed importance. With the limitations of the NetLogo models we are working with we needed increased processing power to run the necessary simulations. Therefore, we ran the simulations on LUNARC’s Aurora parallel processing facility. The results are in section 6 but are summarized in Fig. 6. As can be clearly seen, the simulations confirmed that Linear Model 1 can indeed generate a web-graph with the right degree distribution for the web. What is more, it can do so for a wide range of preference weights between attraction to importance and attraction to popularity.

In short, Linear model 1 seems to meet the desideratum of being able to account for the scale-free topology of the internet while accommodating both attraction to importance and preferential attachment linking dynamics given that we assume
importance to be both rare and distributed according to a power law and, finally, that attraction to importance is linear.

Where importance was exponentially distributed, degree distributions were scale-free only when expected importance was low, and there was some evidence that expected preference for popular pages also had to be relatively high. That all of these distributions would cease to be plausibly scale-free in larger web-graphs is likely to be the case, given the theorem mentioned above for the basic model.

5.1.2. Linear model 2

In this model, when importance is sampled from a Pareto distribution, only one of the 45 parameter configurations supported degree distributions that are close to scale-free with a power close to that of the web. These results indicate that, where importance is Pareto distributed, Linear model 2 does not support the right kind of topology. Where importance is exponentially distributed, Linear model 2 has 12 configurations that produce web-graphs with close to scale-free topologies, and 7 of these have exponents in the goldilocks zone, though again there is deviation at lower numbers of in-links. However, most of these ecologies represent cases where attraction to importance is heavily non-linear, which makes them somewhat implausible as candidates for the ecology of the WWW.

All in all, while Linear model 2 can support web-ecologies conducive to web-graphs with topologies that approach being scale-free with the right sort of exponents, there are reasons to doubt its appropriateness as a model of the ecology of the web. In particular, only over a relatively narrow and implausible range of parameterizations...
does one get the right topologies. Furthermore, there is generally some deviation from a power law distribution at lower numbers of in-links and this would likely become more pronounced as one moved to larger graphs permitting higher resolutions of degree distributions.

5.1.3. Geometric model 1

This model supported power law degree distributions irrespective of whether importance is sampled from Pareto or exponential distributions. Moreover, it is certainly the case that where attraction to importance is linear and importance sufficiently rare ($\alpha \approx 0.001$), this model produced scale-free degree distributions with similar exponents to that of the web (Fig. 7). This might also be the case, as the data for higher expected importance suggests, where attraction to importance is non-linear, but the link density in such ecologies where importance is very rare was too low to allow this issue to be resolved in the graph sizes we considered. While all the above was true where importance was Pareto distributed or exponentially distributed, the tendency is arguably stronger where the former is the case.

All of this indicates that Geometric model 1 produces web-ecologies that overwhelmingly produce web-graphs with power law degree distributions. Furthermore, where attraction to importance is linear and importance sufficiently rare, the power of the degree distributions is within, or close to, the empirically established figures. In short, Geometric model 1 seems to meet the desideratum of being able to account for the scale-free topology of the web while accommodating both attraction to importance and preferential attachment linking dynamics, given that we assume importance to be sufficiently rare.

5.1.4. Geometric model 2

Only 2 of 16 configurations considered for Geometric model 2 exhibited anything approaching scale free topology. In all cases, however, the exponents are too high

![Log/Log plot of degree distribution for parameter configuration 1 of Geometric model 1 with importance Pareto distributed.](https://doi.org/10.1016/j.heliyon.2018.e00978)
and the attraction to importance required too non-linear to be plausible. This indicates that this model broadly fails to produce the right kind of web-graphs for the web, and so is unfit as a model of actual web-ecology.

5.2. Results for linked-based algorithm performance

In this section, we focus on our second research question concerning the performance of In-Degree and PageRank as importance-trackers in the models under consideration. In particular, we investigate the extent to which there are regions of the parameter space where the web-ecologies generated by the models not only reflect the degree-distribution of the WWW, but where in addition the rankings induced by the link-based algorithms are well-correlated with the importance of web-pages, under various assumptions about the distribution of importance.

A side-issue that we also investigate is whether the two link-based algorithms are equally good at tracking importance or whether one is superior to the other. The convergence theorems proven in Masterton et al. (2016) provide the backdrop of this investigation. There, it was shown that both In-Degree and PageRank converge to full correlation with importance in the limit as the size of the web-graph grows. This implies that, in the basic model, these two link-based algorithms are “equivalent in the limit”, although as was observed in the same article there can still be systematic differences in importance-tracking performance for finite web-graphs.

The question concerning the relative performance of PageRank and In-Degree is even more interesting in the dual models. This is due to the fact that their convergence to full correlation with importance in the limit is not guaranteed in those models. Recall that we have two mechanisms at work: attraction to importance and attraction to popularity. The former would have the rankings induced converge on full correlation with importance in the limit in the basic model, but this is no longer guaranteed when the second mechanism is added. In cases where there is no convergence, the two link-based metrics we consider may not be equivalent even in the limit. Thus, there can be more fundamental differences in performance between In-Degree and PageRank in the dual models than in the original MOA models.

5.2.1. Linear model 1

Here we were genuinely surprised by our results. Our previous experience with the basic model has been that where expected importance is low, the degree of correlation between page importance and link-based metrics is low. As we just observed, we know from the theorems in Masterton et al. (2016) that the correlation will eventually converge toward being perfect as web-graph size increases, but our experience is that for expected importance less than 0.1, 1,000 page web-graphs are not large
enough for this effect to be sufficiently pronounced. It came as a surprise that in Linear model 1 the performance of link-based metrics was markedly improved against this baseline. While this is so whether importance is Pareto or exponentially distributed, the results are particular impressive in the former case. It is as if the preferential attachment component of the Linear model is amplifying the convergence effect previously demonstrated for the basic model. With this evidence we can safely say that if the ecology of the web is that of Linear model 1 — and what we know of the web’s degree distribution is consistent with this as we saw in the previous section — then link-based metrics, particularly In-Degree, are surely appropriate guides as to what is important on the web.

Given the construction of the basic model, it is hardly surprising that In-Degree performs a little better than PageRank in Linear model 1, as confirmed in Figs. 8 and 9.

Not recorded in this document is the data gathered over the same model parameterizations for graphs of 500 pages. As one might expect, given the convergence theorems for the basic model, in nearly all model configurations, increasing graph size increased degree of correlation between link-based metric and importance (Fig. 10). In those configurations where expected importance is 0.1 (and presumably greater) this trend appears to reverse, but this is presumably the least plausible region of the parameter space we considered.

**Fig. 8.** The coefficients of determination (= degree of correlation with page importance) for PageRank (Blue) and In-Degree (Red) in Linear model 1 for the parameter space \((\alpha \in [0.001, 0.01, 0.1], Pw \in \{1\}, epw \in \{0.1, 0.3, 0.5, 0.7, 0.9\})\) where importance is Pareto distributed in web-graphs with 1,000 webpages. One can see that, in all configurations In-Degree correlates better with page importance than does PageRank. Also the absolute degree of correlation is impressive, even in configurations 1, 6 and 11 where expected preference for importance is a ninth of expected preference for popularity.
Fig. 9. The coefficients of determination (= degree of correlation with page importance) for PageRank (blue) and In-Degree (red) in Linear model 1 for the parameter space $(\alpha \in [0.001, 0.01, 0.1], P_w \in \{1\}, \epsilon_{prw} \in \{0.1, 0.3, 0.5, 0.7, 0.9\})$ where importance is exponentially distributed in web-graphs with 1,000 webpages. Again one can see that, in all configurations In-Degree correlates better with page importance than PageRank.

Fig. 10. The coefficients of determination (= degree of correlation with page importance) for PageRank in 500-page web-graphs (blue), PageRank in 1000-page web-graphs (red), In-Degree in 500-page web-graphs (purple) and In-Degree in 1000-page web-graphs (turquoise) in Linear model 1 for the parameter space $(\alpha \in [0.001, 0.01, 0.1], P_w \in \{1\}, \epsilon_{prw} \in \{0.1, 0.3, 0.5, 0.7, 0.9\})$ where importance is Pareto. In the first 10 configurations, where expected importance is <0.01, increasing web-graph size increases the correlation of metric with importance in all configurations save where the metrics correlation with importance is already almost maximal (In-Degree in configurations 9 and 10).
Crucially, we similarly determined the degree of correlation between these two metrics and importance in the region of the parameter space where Linear Model 1 reproduces the degree distribution of the web and confirmed high degrees of correlation, particularly for In-Degree, where this is so and attraction to importance is greater than attraction to popularity (Fig. 11).

5.2.2. Linear model 2

In absolute terms, where importance is exponentially distributed In-Degree and PageRank perform similarly well across the parameter space of this model. That performance generally increases with expected importance, but is negligible where expected importance is around 0.1 or lower. As this is the most plausible region of the parameter space, this raises a concern that were the ecology of the WWW to be as described in Linear model 2, link-based metrics would be woeful guides as to what is important on line; so woeful is their performance in this region of the parameter space that their employment in search engines would arguably be unwarranted. Perhaps the most noteworthy thing about this model is that even though the attraction to importance component is designed to satisfy all of Google’s assumptions, and so one could think that they would favor PageRank over In-Degree, In-Degree performed better than PageRank in the vast majority of configurations where their performance was non-negligible (Fig. 12).

5.2.3. Geometric models

Web-metric performance is relatively poor across the board in Geometric model 2, so not only does this model not produce the right kind of degree distributions for the

![Figure 11](https://doi.org/10.1016/j.heliyon.2018.e00978)

**Fig. 11.** The coefficients of determination (= degree of correlation with page importance) for PageRank in 7,500-page web-graphs (blue), and In-Degree in 7,500-page web-graphs (red) in Linear model 1 for the parameter space \( \{\alpha \in \{8.2 \times 10^{-6}, 2.6 \times 10^{-5}, 1.4 \times 10^{-4}\}, mp_{\infty} \{1 \times 10^{-6}\}, Pri_{\infty} \{1\}, ep_{\infty} \{0.1, 0.5, 0.9\}\} \) where importance is Pareto distributed. The configurations where the exponents of the degree distribution of the generated graphs are similar to the WWW are 6, 7 and 8.
internet, it also represents an ecology where there is little reason to suppose PageRank or In-Degree to be reliable indicators of page importance. Geometric model 1, on the other hand, is not only promising in terms of the web-graph topologies it generates but it is also the case that In-Degree performs well here. Strikingly, PageRank’s performance in this model is substantially poorer than In-Degree’s. We judge these results to be strong enough to warrant the conclusion that, if the web’s ecology is a Geometric model 1 ecology, then In-Degree is a superior link-based metric for search algorithms to PageRank. Furthermore, there is no clear trend that increasing web-graph size leads to increasing link-based metric performance (Fig. 13). While it appears that in some parameter configurations, increasing web-graph size increases the degree of correlation between PageRank/In-Degree and page importance, there are just as many configurations where this does not appear to be the case. However, this does appear to be so where attraction to importance is linear and importance is Pareto distributed so that the web-graphs generated are scale-free with exponents near to the established 2.1 (configuration 1 in Fig. 13), which is arguably the most plausible region of the parameter space.

6. Calculation

This section details the calculations referred to in section 5. In the tables below, Num is a number denoting a parameter configuration of a model representing a particular web-ecology; Pw a positive real number that induces non-linearity on attraction to importance for values other than 1; α the level of importance one expects of a
Page sampled at random from a web-graph produced by the web-ecology concerned; epw the preference for importance over popularity one expects of a page sampled at random from a web-graph produced by the web-ecology concerned; and Pow(par) the power of the degree distribution of a randomly selected web-graph from those produced by the web-ecology concerned where importance is sampled from a Pareto distribution, if the graph’s degree distribution is scale-free (if the degree-distribution is not scale-free then such a value is non-applicable (N.A.) to web-graphs produced by such an ecology). This value is indicative only.

Furthermore, $R^2_{p(par)}$ is the correlation coefficient of the linear regression in the log/log plot of number of pages against number of in-links whose gradient is Pow(par).

A rough measure of the degree of confidence that is reasonable in the Pow(par) value given the data. This value is indicative only. $R^2_{par PR}$ is the average coefficient of determination, relative to a third order polynomial regression for the importance of pages against their PageRanks (as determined by the diffusion algorithm over a 20 step simulation), over 20 web-graphs produced by the web-ecology concerned where importance is sampled from a Pareto distribution. This is a measure of the degree to which PageRank is correlated with page importance in the web-graphs produced by the ecology specified by the parameterized model concerned. $R^2_{par ID}$ is the average coefficient of determination, relative to a third order polynomial regression for the importance of pages against their In-Degrees, over 20 web-graphs produced by the web-ecology concerned where importance is sampled from a Pareto distribution.

**Fig. 13.** The coefficients of determination (= degree of correlation with page importance) for PageRank in graphs of 1,000 pages (blue) and 500 pages (red), and the same for In-Degree in graphs of 1,000 pages (purple) and 500 pages (turquoise) in Geometric model 1 for the parameter space $(\alpha \in \{0.001, 0.01, 0.1\}, \beta \in \{1, 2, 3, 4\})$ where importance is Pareto distributed. The only really clear conclusions supported by the data are that, in general, In-Degree tracks importance better than PageRank in these ecologies and increasing non-linearity in attraction to importance generally decreases link-based metric performance.
distribution. This is a measure of the degree to which In-Degree is correlated with page importance in the web-graphs produced by the ecology specified by the parameterized model concerned.

Moreover, Pow(exp) is the power of the degree distribution of a randomly selected web-graph from those produced by the web-ecology concerned where importance is sampled from an exponential distribution, if the graph’s degree distribution is scale-free. If the degree-distribution is not scale-free then such a value is non-applicable (N.A.) to web-graphs produced by such an ecology. This value is indicative only. \(R^2\text{p(exp)}\) is the correlation coefficient of the linear regression in the log/log plot of number of pages against number of in-links whose gradient is Pow(exp). A rough measure of the degree of confidence that is reasonable in the Pow(exp) value given the data. This value is indicative only. \(R^2\text{(exp)PR}\) is the The average coefficient of determination, relative to a third order polynomial regression for the importance of pages against their PageRanks (as determined by the diffusion algorithm over a 20 step simulation), over 20 web-graphs produced by the web-ecology concerned where importance is sampled from an exponential distribution. This is a measure of the degree to which PageRank is correlated with page importance in the web-graphs produced by the ecology specified by the parameterized model concerned.

Finally, \(R^2\text{(exp)ID}\) is the The average coefficient of determination, relative to a third order polynomial regression for the importance of pages against their In-Degrees, over 20 web-graphs produced by the web-ecology concerned where importance is sampled from an exponential distribution. This is a measure of the degree to which In-Degree is correlated with page importance in the web-graphs produced by the ecology specified by the parameterized model concerned. Table 1 shows calculated values for Linear model 1, Table 2 for Linear model 2, Table 3 for Geometric model 1 and Table 4 for Geometric model 2.

7. Discussion

Geometric model 2 is simply not a candidate for the ecology of the WWW given the topological results above, and if it were, this would be bad news for the wisdom-of-crowds justification of link-based metrics in search engines. By contrast, Linear model 2 can generate, over small regions of its parameter space, web-graphs with topologies that are close to scale-free in their degree distributions and with exponents in the ball park of what is observed for the web. However, these regions are small, implausible, and one can see a tendency in many of the degree distributions toward deviation from a true power law at low numbers of in-links that would likely be exacerbated as one moved to larger web-graphs. While link-based metric performance is acceptable in this model, and improving with increasing web-graph size, this is only so where importance is relatively common which we deem to be an implausible
Table 1. Calculated values for Linear model 1. The number of webpages for determining the degree distribution power law exponent was 15,000. The number of webpages for determining coefficients of determination for In-Degree and PageRank was 7,500. Coefficients of determination were determined on the 20 webpages with the highest In-Degrees/PageRanks. Minimum page importance on the sampling distribution is $1 \times 10^{-6}$.

| Num | Pw | $\alpha$ | epw | Pow (par) | $R^2_{p(par)}$ | $R^2_{(par)PR}$ | $R^2_{(par)ID}$ | Pow (exp) | $R^2_{p(exp)}$ | $R^2_{(exp)PR}$ | $R^2_{(exp)ID}$ |
|-----|----|---------|-----|----------|----------------|----------------|----------------|-----------|----------------|----------------|----------------|
| 1   | 1  | 0.001   | 0.1 | 0.81     | 0.57           | 0.658          | 0.872          | N.A.      | 0.004          | 0.007           |                 |
| 2   | 1  | 0.001   | 0.3 | 1.03     | 0.82           | 0.835          | 0.963          | 1.6       | 0.88           | 0.017           | 0.020           |
| 3   | 1  | 0.001   | 0.5 | 1.42     | 0.88           | 0.959          | 0.987          | 1.83      | 0.97           | 0.090           | 0.136           |
| 4   | 1  | 0.001   | 0.7 | 1.4      | 0.8            | 0.972          | 0.988          | 1.87      | 0.93           | 0.151           | 0.210           |
| 5   | 1  | 0.001   | 0.9 | 1.64     | 0.87           | 0.984          | 0.991          | N.A.      | 0.244          | 0.387           |                 |
| 6   | 1  | 0.01    | 0.1 | 1.07     | 0.66           | 0.672          | 0.874          | 1.72      | 0.92           | 0.074           | 0.119           |
| 7   | 1  | 0.01    | 0.3 | 1.19     | 0.83           | 0.837          | 0.963          | N.A.      | 0.310          | 0.437           |                 |
| 8   | 1  | 0.01    | 0.5 | 1.37     | 0.91           | 0.961          | 0.987          | N.A.      | 0.701          | 0.845           |                 |
| 9   | 1  | 0.01    | 0.7 | 1.35     | 0.86           | 0.970          | 0.988          | N.A.      | 0.745          | 0.871           |                 |
| 10  | 1  | 0.01    | 0.9 | 1.41     | 0.89           | 0.984          | 0.991          | N.A.      | 0.809          | 0.901           |                 |
| 11  | 1  | 0.1     | 0.1 | 0.98     | 0.79           | 0.661          | 0.872          | N.A.      | 0.657          | 0.871           |                 |
| 12  | 1  | 0.1     | 0.3 | 0.91     | 0.70           | 0.834          | 0.964          | N.A.      | 0.841          | 0.964           |                 |
| 13  | 1  | 0.1     | 0.5 | 0.99     | 0.80           | 0.959          | 0.988          | N.A.      | 0.961          | 0.987           |                 |
| 14  | 1  | 0.1     | 0.7 | 0.98     | 0.82           | 0.971          | 0.988          | N.A.      | 0.971          | 0.988           |                 |
| 15  | 1  | 0.1     | 0.9 | 0.93     | 0.74           | 0.984          | 0.991          | N.A.      | 0.984          | 0.991           |                 |

(continued on next page)
| Num | Pow | Pow (par) | R² (par) | ID |
|-----|-----|-----------|---------|----|
| 1   | 1.4 | 0.9       | 0.9     | 1  |
| 2   | 8.2 | 0.5       | 0.95    | 0.03 |
| 3   | 1.4 | 0.9       | 0.9     | 0.03 |
| 4   | 2.6 | 0.9       | 0.9     | 0.03 |
| 5   | 8.2 | 0.5       | 0.95    | 0.03 |
| 6   | 1.4 | 0.9       | 0.9     | 0.03 |

*Table 1. (Continued)*
Table 2. Calculated values for Linear model 2.

| Num | Pw | α  | e pw | Pow (par) | R²p(par) | R²(par)PR | R²(par)ID | Pow (exp) | R²p(exp) | R²(exp)PR | R²(exp)ID |
|-----|----|----|------|----------|----------|-----------|-----------|----------|----------|-----------|-----------|
| 1   | 2  | 0.1| 0.1  | 1.30     | 0.93     | 0.013     | 0.019     | 1.8      | 0.95     | 0.015     | 0.019     |
| 2   | 2  | 0.1| 0.3  | 1.17     | 0.86     | 0.005     | 0.003     | N.A.     | 0.005    | 0.003     | 0.003     |
| 3   | 2  | 0.1| 0.5  | 1.2      | 0.88     | 0.007     | 0.003     | N.A.     | 0.013    | 0.019     | 0.013     |
| 4   | 2  | 0.1| 0.7  | 1.23     | 0.92     | 0.015     | 0.017     | N.A.     | 0.005    | 0.005     | 0.005     |
| 5   | 2  | 0.1| 0.9  | 1.14     | 0.88     | 0.002     | 0.003     | N.A.     | 0.003    | 0.002     | 0.002     |
| 6   | 3  | 0.1| 0.1  | 0.89     | 0.86     | 0.007     | 0.004     | N.A.     | 0.003    | 0.002     | 0.002     |
| 7   | 3  | 0.1| 0.3  | 0.84     | 0.76     | 0.013     | 0.014     | 1.5      | 0.96     | 0.016     | 0.021     |
| 8   | 3  | 0.1| 0.5  | 0.79     | 0.76     | 0.004     | 0.004     | 1.9      | 0.9      | 0.005     | 0.005     |
| 9   | 3  | 0.1| 0.7  | 0.88     | 0.8      | 0.008     | 0.005     | 1.9      | 0.9      | 0.005     | 0.004     |
| 10  | 3  | 0.1| 0.9  | 0.88     | 0.81     | 0.008     | 0.016     | 2.5      | 0.93     | 0.013     | 0.024     |
| 11  | 4  | 0.1| 0.1  | N.A.     | 0.005    | 0.006     | N.A.      | 0.006    | 0.004    | 0.004     | 0.004     |
| 12  | 4  | 0.1| 0.3  | N.A.     | 0.002    | 0.002     | N.A.      | 0.003    | 0.006    | 0.006     | 0.006     |
| 13  | 4  | 0.1| 0.5  | N.A.     | 0.014    | 0.017     | N.A.      | 0.016    | 0.015    | 0.015     | 0.015     |
| 14  | 4  | 0.1| 0.7  | N.A.     | 0.004    | 0.007     | N.A.      | 0.002    | 0.009    | 0.009     | 0.009     |
| 15  | 4  | 0.1| 0.9  | N.A.     | 0.005    | 0.002     | N.A.      | 0.003    | 0.003    | 0.003     | 0.003     |
| 16  | 2  | 0.25| 0.1 | 0.97     | 0.77     | 0.099     | 0.950     | N.A.     | 0.693    | 0.866     | 0.866     |
| 17  | 2  | 0.25| 0.3 | 0.98     | 0.85     | 0.826     | 0.842     | N.A.     | 0.647    | 0.701     | 0.701     |
| 18  | 2  | 0.25| 0.5 | 1.0      | 0.73     | 0.724     | 0.711     | N.A.     | 0.484    | 0.397     | 0.397     |

(continued on next page)
Table 2. (Continued)

| Num | Pw  | epw | Pow (par) | R^2(par) | PR | R^2(par)ID | Pow (exp) | R^2(exp) | PR | R^2(exp)ID |
|-----|-----|-----|-----------|---------|-----|--------|-----------|---------|-----|--------|
| 19  | 2   | 0.25| 0.7       | 0.95    | 0.72| 0.907  | 0.950     | N.A.    | 0.705| 0.873  |
| 20  | 2   | 0.25| 0.9       | 1.16    | 0.75| 0.826  | 0.843     | N.A.    | 0.622| 0.704  |
| 21  | 3   | 0.25| 0.1       | 1.21    | 0.79| 0.724  | 0.720     | 1.6     | 0.9  | 0.501  |
|     |     |     |           |         |     |        |           |         |     |        |
| 22  | 3   | 0.25| 0.3       | 1.19    | 0.80| 0.908  | 0.951     | 1.2     | 0.9  | 0.698  |
| 23  | 3   | 0.25| 0.5       | 1.1     | 0.75| 0.832  | 0.842     | N.A.    | 0.649| 0.694  |
|     |     |     |           |         |     |        |           |         |     |        |
| 24  | 3   | 0.25| 0.7       | 1.25    | 0.84| 0.726  | 0.714     | N.A.    | 0.522| 0.437  |
| 25  | 3   | 0.25| 0.9       | 1.03    | 0.81| 0.905  | 0.950     | N.A.    | 0.711| 0.872  |
|     |     |     |           |         |     |        |           |         |     |        |
| 26  | 4   | 0.25| 0.1       | 1.05    | 0.88| 0.829  | 0.841     | 1.4     | 0.92 | 0.642  |
|     |     |     |           |         |     |        |           |         |     |        |
| 27  | 4   | 0.25| 0.3       | 1.01    | 0.83| 0.728  | 0.722     | 1.7     | 0.92 | 0.463  |
|     |     |     |           |         |     |        |           |         |     |        |
| 28  | 4   | 0.25| 0.5       | 0.97    | 0.81| 0.910  | 0.952     | 1.8     | 0.91 | 0.708  |
|     |     |     |           |         |     |        |           |         |     |        |
| 29  | 4   | 0.25| 0.7       | 0.98    | 0.79| 0.830  | 0.837     | 1.8     | 0.94 | 0.651  |
|     |     |     |           |         |     |        |           |         |     |        |
| 30  | 4   | 0.25| 0.9       | 0.95    | 0.81| 0.727  | 0.715     | N.A.    | 0.483| 0.420  |
| 31  | 2   | 0.4  | 0.1       | N.A.    | 0.947| 0.969  | N.A.      | 0.900  | 0.960|
|     |     |     |           |         |     |        |           |         |     |        |
| 32  | 2   | 0.4  | 0.3       | N.A.    | 0.883| 0.891  | N.A.      | 0.857  | 0.889|
| 33  | 2   | 0.4  | 0.5       | N.A.    | 0.793| 0.801  | N.A.      | 0.756  | 0.790|
|     |     |     |           |         |     |        |           |         |     |        |
| 34  | 2   | 0.4  | 0.7       | N.A.    | 0.910| 0.929  | N.A.      | 0.891  | 0.915|
| 35  | 2   | 0.4  | 0.9       | N.A.    | 0.832| 0.842  | N.A.      | 0.872  | 0.856|
| 36  | 3   | 0.4  | 0.1       | N.A.    | 0.936| 0.951  | N.A.      | 0.901  | 0.927|
|     |     |     |           |         |     |        |           |         |     |        |
| 37  | 3   | 0.4  | 0.3       | N.A.    | 0.793| 0.801  | N.A.      | 0.756  | 0.790|

(continued on next page)
Table 2. (Continued)

| Num | Pw | α  | epw | Pow (par) | R²(p(par)) | R²(par)PR | R²(par)ID | Pow (exp) | R²(p(exp)) | R²(exp)PR | R²(exp)ID |
|-----|----|----|-----|-----------|------------|-----------|-----------|-----------|------------|-----------|-----------|
| 38  | 3  | 0.4| 0.5 | N.A.      | 0.836      | 0.893     | N.A.      | 0.857     | 0.888      |           |           |
| 39  | 3  | 0.4| 0.7 | N.A.      | 0.789      | 0.799     | N.A.      | 0.751     | 0.781      |           |           |
| 40  | 3  | 0.4| 0.9 | 0.86      | 0.71       | 0.945     | 0.969     | N.A.      | 0.897      | 0.961     |           |
| 41  | 4  | 0.4| 0.1 | 0.91      | 0.74       | 0.883     | 0.892     | 1.7       | 0.93       | 0.856     | 0.889     |
| 42  | 4  | 0.4| 0.3 | 0.91      | 0.72       | 0.791     | 0.800     | N.A.      | 0.754      | 0.779     |           |
| 43  | 4  | 0.4| 0.5 | 0.94      | 0.77       | 0.945     | 0.970     | N.A.      | 0.900      | 0.960     |           |
| 44  | 4  | 0.4| 0.7 | 0.92      | 0.76       | 0.881     | 0.892     | N.A.      | 0.855      | 0.889     |           |
| 45  | 4  | 0.4| 0.9 | 0.92      | 0.71       | 0.789     | 0.801     | N.A.      | 0.754      | 0.783     |           |
assumption (although, in the final analysis, its plausibility might depend on a specific qualitative understanding of importance). All this leads us to conclude that it is doubtful that the ecology of the internet is that of either of these models, but were this the case, it is even more doubtful that link-based metrics would be useful in identifying what is worth looking at on the web.

Linear model 1 and Geometric model 1, however, are a markedly different story. Not only do they produce the right kind of web-graph topology for a range of

### Table 3. Calculated values for Geometric model 1.

| Num | Pw  | α   | Pow (par) | R^2(p(par)) | R^2(par)PR | Pow (exp) | R^2(p(exp)) | R^2(exp)PR | R^2(exp)ID |
|-----|-----|-----|-----------|-------------|------------|-----------|-------------|------------|------------|
| 1   | 1   | 0.001 | 2.22   | 0.95       | 0.636      | 2.71      | 0.98        | 0.256      | 0.362      |
| 2   | 2   | 0.001 | N.A.   | 0.746      | 0.793      | N.A.      | 0.013       | 0.012      |            |
| 3   | 3   | 0.001 | N.A.   | 0.726      | 0.739      | N.A.      | 0           | 0          |            |
| 4   | 4   | 0.001 | N.A.   | 0.361      | 0.762      | N.A.      | 0           | 0          |            |
| 5   | 1   | 0.01  | 1.58   | 0.91       | 0.522      | 0.798     | N.A.        | 0.254      | 0.574      |
| 6   | 2   | 0.01  | N.A.   | 0.563      | 0.697      | N.A.      | 0.331       | 0.356      |            |
| 7   | 3   | 0.01  | N.A.   | 0.488      | 0.591      | N.A.      | 0.045       | 0.047      |            |
| 8   | 4   | 0.01  | N.A.   | 0.466      | 0.526      | N.A.      | 0.002       | 0.020      |            |
| 9   | 1   | 0.1   | 1.26   | 0.84       | 0.409      | 0.886     | N.A.        | 0.236      | 0.742      |
| 10  | 2   | 0.1   | 0.83   | 0.73       | 0.356      | 0.673     | 1.49        | 0.89       | 0.312      | 0.657      |
| 11  | 3   | 0.1   | 0.98   | 0.80       | 0.303      | 0.537     | 1.33        | 0.92       | 0.303      | 0.467      |
| 12  | 4   | 0.1   | 1.23   | 0.92       | 0.207      | 0.454     | 1.4         | 0.92       | 0.239      | 0.307      |
| 13  | 1   | 0.25  | 1.11   | 0.88       | 0.300      | 0.818     | N.A.        | 0.250      | 0.752      |
| 14  | 2   | 0.25  | 1      | 0.78       | 0.341      | 0.715     | 1.29        | 0.86       | 0.327      | 0.710      |
| 15  | 3   | 0.25  | 0.87   | 0.69       | 0.282      | 0.597     | 1.22        | 0.84       | 0.294      | 0.569      |
| 16  | 4   | 0.25  | 0.95   | 0.73       | 0.237      | 0.511     | 1.12        | 0.85       | 0.251      | 0.447      |

### Table 4. Calculated values for Geometric model 2.

| Num | Pw  | α   | Pow (par) | R^2(p(par)) | R^2(par)ID | Pow (exp) | R^2(p(exp)) | R^2(exp)ID |
|-----|-----|-----|-----------|-------------|------------|-----------|-------------|------------|
| 1   | 1   | 0.1  | N.A.      | 0.013       | 0.010      | N.A.      | 0.112       | 0.012      |
| 2   | 2   | 0.1  | N.A.      | 0.061       | 0.058      | N.A.      | 0.211       | 0.026      |
| 3   | 3   | 0.1  | N.A.      | 0.128       | 0.165      | 3.56      | 0.96        | 0.276      | 0.055      |
| 4   | 4   | 0.1  | N.A.      | 0.230       | 0.322      | N.A.      | 0.353       | 0.019      |
| 5   | 1   | 0.25 | N.A.      | 0.083       | 0.077      | N.A.      | 0.119       | 0.052      |
| 6   | 2   | 0.25 | N.A.      | 0.174       | 0.161      | N.A.      | 0.210       | 0.112      |
| 7   | 3   | 0.25 | N.A.      | 0.305       | 0.261      | N.A.      | 0.280       | 0.193      |
| 8   | 4   | 0.25 | N.A.      | 0.403       | 0.387      | 3         | 0.9         | 0.367      | 0.316      |
parameterizations, given that importance is power law distributed and sufficiently rare, but the preferential attachment component of the model seems to accelerate the convergence of link-based metrics on perfect correlation with importance that we have previously proven for the basic model. This is apparently so even for PageRank, where the convergence proof in Masterton et al. (2016) assumed that linking from a page is independent of linking to that page, which is an assumption that manifestly holds in neither Linear model 1 nor Geometric model 1 due to the BA-component of those models. Not only can these models account for the degree distribution in the web on plausible parameterizations, but we have strong evidence that there is wisdom-of-crowds in effect for link-based metrics in such web-ecologies, and that this effect is strengthened by the preferential attachment components of these models. It should be noted that in both models, In-Degree was superior to PageRank in its correlation with importance in almost all web-ecologies.

Hence, dual models that model attraction-to-importance linking dynamics in the fashion of the basic model of Masterton et al. (2016) have revealed themselves to be viable models of web-ecology. That the same is not so for dual models based on the extended model of the same authors provides some evidence against Google’s justifying assumption for using PageRank over In-Degree in their search engine — the assumption that webmasters of important pages are more attracted to other important pages than webmasters of less important pages, who are more likely to link randomly. In both Linear model 1 and Geometric model 1, every page is equally attracted to the importance of other pages. Where this is so it is hardly surprising that In-Degree should perform better than PageRank, as borne out in our data, and so one has reason to doubt the superiority of PageRank over In-Degree as a guide to what is important online. Of course, our extended model is just one model conforming to Google’s assumptions about linking dynamics justifying their preference for PageRank. There may be others, and these may be more conducive to scale-free web-topologies and superior PageRank performance. Still, our results provide a reason to doubt the truth of Google’s second assumption, a reason that should be countered by those who champion the relative superiority of PageRank as a guide to what is worth looking at despite its well-known relative disadvantage in computing time.

Our models were implemented in NetLogo which does not easily handle web-graphs of orders much greater than 1,000 on a normal desktop computer. To reduce this limitation, we ran some simulations on web-graphs of size 15,000 using a parallel computing facility. As the form of the degree distribution of a network becomes clearer as the web-graphs grows, our results should be seen as indicative only for the whole web. Were one to look at the degree distributions of the networks generated by the same model configurations as we have entertained herein, save that the number of pages was increased to millions or billions, one might well find that the clearer resolution achievable would indicate that a subset of those configurations we have identified as supporting scale-free topologies in fact do so. Also, one might well
determine that the exponents are slightly different from what we have recorded here. On the other hand, the metric results (regarding the importance-tracking performance of PageRank and In-Degree) are less susceptible to error and even more likely to be representative of the true values. Future work should explore larger web-graphs. Finally, there are, of course, other graph statistics for the web than degree distribution, such as clustering coefficient. We intend to explore the question whether our preferred models can account for them as well in future research.

8. Conclusion

We addressed the fundamental question why we should use PageRank or other similar link-based algorithms for ranking web-pages in search engines. A striking proposal, due to the Google founders themselves, is that such ranking can be justified with reference to the wisdom-of-crowds, the idea being that PageRank aggregates the somewhat reliable individual quality judgements of webmasters into a much more reliable collective judgement. We noted that existing attracting-to-importance models supporting the wisdom-of-crowds justification for PageRank generate web-graphs that fail to account for the scale-free structure of the World Wide Web, where that structure is likely due to webmasters linking to popular pages, producing a rich-get-richer effect. To overcome this problem for the wisdom-of-crowds thesis, we explored, following a proposal in Masterton and Olsson (2017), dual models of the web in which both attraction to popularity and attraction to importance qua quality play a role. Our simulation results for web-graphs of up to 15,000 pages indicate that the wisdom-of-crowd justifications for using PageRank need not be foreclosed by empirical findings about the scale-free structure of the web, nor by arguments to the effect that there must be some preferential attachment component to the linking dynamics. While there are several limitations to our results, we believe that they give a partial vindication of the wisdom-of-crowds justification for PageRank. Future work should explore larger web-graphs as well as other aspects pertaining to the empirical plausibility of the dual linking models under consideration.

Declarations

Author contribution statement

George Masterton: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Erik J. Olsson: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.
Funding statement

This work was supported by Swedish Research Council through the framework grant Knowledge in a Digital World: Trust, Credibility and Relevance on the Web (Olsson, PI).

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Albert, R., Jeong, H., Barabási, A., 1999. Diameter of the world-wide web. Nature 401, 130–131. https://www.nature.com/articles/43601.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30, 107–117.

Brin, S., Page, L., Motwami, R., Winograd, T., 1998. The PageRank Citation Ranking: Bringing Order to the Web. Stanford University Technical Report. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512. http://science.sciencemag.org/content/286/5439/509.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J., 2000. Graph structure in the web: experiments and models. In: Proceedings of the Ninth International World Wide Web Conference, Amsterdam, The Netherlands, May 15–19.

Doan, Anhai, Ramakrishnan, Raghu, Halevy, Alon Y., 2011. Crowdsourcing systems on the world-wide web. Commun. ACM 54 (4), 86–96. https://cacm.acm.org/magazines/2011/4/106563-crowdsourcing-systems-on-the-world-wide-web/fulltext.

Krapivsky, P.L., Rodgers, G.J., Redner, S., 2001. Degree distributions of growing networks. Phys. Rev. Lett. 86 (23), 5401–5404. https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.86.5401.

Lee, J.-Y., Zaheer, M., Günne, Smola, A.J., 2015. Preferential attachment in graphs with affinities. In: Proceedings of the 18th International Conference on
Artificial Intelligence and Statistics, vol. 38. JMLR: W&CP, San Diego, CA, USA. In: http://proceedings.mlr.press/v38/lee15b.html.

Masterton, G., Olsson, E.J., Angere, S., 2016. Linking as voting: how the condorcet jury theorem in political science is relevant to webometrics. Scientometrics 106 (3), 945–966. https://link.springer.com/article/10.1007/s11192-016-1837-1.

Masterton, G., Olsson, E.J., 2017. From impact to importance: the current state of the wisdom of crowds justification of link-based ranking algorithms. Philos. Technol.

Surowiecki, J., 2004. The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. Little Brown, London.

Thelwall, M., 2013. Society on the web. In: Dutton, W.H. (Ed.), The Oxford Handbook of Internet Studies. Oxford University Press, Oxford, pp. 69–85.

Zhang, J., Wu, X., Sheng, V.S., 2016. Learning from crowdsourced labeled data: a survey. Artif. Intell. Rev. 46 (4), 543–576. https://link.springer.com/article/10.1007/s10462-016-9491-9.

Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R., 2017. Truth inference in crowdsourcing: is the problem solved? Proc. VLDB Endow. 10 (5), 541–552. http://www.vldb.org/pvldb/vol10/p541-zheng.pdf.