From Impact to Importance

the Current State of the Wisdom-of-Crowds Justification of Link-based Ranking Algorithms

Masterton, George; Olsson, Erik J

Published in:
Philosophy & Technology

DOI:
10.1007/s13347-017-0274-2

2018

Document Version:
Publisher's PDF, also known as Version of record

Philos. Technol.
DOI 10.1007/s13347-017-0274-2

RESEARCH ARTICLE

From Impact to Importance: The Current State of the Wisdom-of-Crowds Justification of Link-Based Ranking Algorithms

George Masterton[1] · **Erik J. Olsson**[1]

Received: 13 December 2016 / Accepted: 2 August 2017

Abstract In a legendary technical report, the Google founders sketched a wisdom-of-crowds justification for PageRank arguing that the algorithm, by aggregating incoming links to webpages in a sophisticated way, tracks importance (quality, relevance, etc.) on the web. On this reading of the report, webpages that have a high impact as measured by PageRank are supposed to be important webpages in a sense of importance that is not reducible to mere impact or popularity. In this paper, we look at the state of the art regarding the more precise statement of the thesis that PageRank and other similar in-link-based ranking algorithms can be justified by reference to the wisdom of crowds. We argue that neither the influential preferential attachment models due to Barabási and Albert in (Science 286:509-512, 1999) nor the recent model introduced by Masterton et al. in (Scientometrics 106:945–966, 2016) allows for a satisfactory wisdom-of-crowds justification of PageRank. As a remedy, we suggest that future work should explore "dual models" of linking on the web, i.e., models that combine the two previous approaches. Dual models view links as being attracted to both popularity and importance.

Keywords Impact · Importance · Link-based ranking · World Wide Web · Wisdom of crowds · Google · PageRank · Preferential attachment

✉ George Masterton
george.masterton@fil.lu.se

Erik J. Olsson
erik_j.olsson@fil.lu.se

[1] Department of Philosophy, Lund University, Lund, Sweden

Published online: 16 August 2017

1 Introduction

In a legendary technical report, the Google founders gave what looks like an informal wisdom-of-crowds justification for PageRank arguing that the algorithm tracks importance on the web by aggregating in-links in a sophisticated way (Brin et al. 1998). We refer to this thesis as wisdom-of-crowds justification for PageRank (WCJPR). In this paper, we look at the state of the art regarding a precise statement of the WCJPR thesis and its proof.

Our first point is that while the influential preferential attachment model due to Barabási and Albert (1999) is, in a minimalist sense, a realistic model of the web in that it gives rise to scale-free networks not dissimilar to the WWW, it does not allow for a convincing formulation, much less a proof, of the WCJPR thesis. Our second point is that while the recent linking model proposed by Masterton et al. (2016), which was explicitly introduced to account for the Google founders' reasoning, does allow for a formulation, and proof of the WCJPR thesis, it is not a realistic model of the web because although it can generate scale-free networks of the right kind for the WWW, it does so for the wrong reasons. Thus, there is at present, to the best of our knowledge, no fully satisfactory formulation and proof of the WCJPR thesis.

As a remedy, we suggest that future work should explore "dual models" of the web, i.e., models that combine preferential attachment and the Masterton, Olsson, and Angere (MOA) model into one account of the web. We conjecture that there are dual models that are realistic models of the web and at the same time allow for the rigorous formulation and proof of the WCJPR thesis.

2 Background

The PageRank algorithm of Google is a method for evaluating the relative importance of webpages. Everything else being equal, the more webpages that link to a given page, the higher that page's PageRank. A page's PageRank is further increased if the webpages linking to it have higher PageRanks themselves. Finally, a page's PageRank decreases as the profligacy of linking of those pages that link to it increases. Thus, the PageRank of a webpage depends not only on the local topology of the web in which it occurs but also upon the global topology. PageRank should be contrasted with the simpler In-Degree algorithm which simply ranks webpages by counting their respective numbers of incoming links. What we will say in the following about the justification of PageRank, or lack thereof, applies equally to In-Degree. However, for the sake of definiteness, we will focus on PageRank. Much of what we say will also generalize to citation-based algorithms for ranking scientific publications. However, we will leave this obvious parallel unexplored for the purposes of the present investigation. For the details of the PageRank algorithm (US patent 6,285,999), see Brin et al. (1998). Franceschet (2011) and Wills (2006) have useful popular introductions.

PageRank is not the only factor determining the ranking of a given webpage in Google. There are reportedly some 300 further "quality signals" that determine the ranking of a particular webpage. However, all signals are not of the same importance, and PageRank is believed to still play a significant initial role in Google's rankings of search results.

One intuitive motivation for PageRank invites us to consider the case of a "random surfer" (Brin and Page 1998; see also Brin et al. (1998)):

"We assume there is a 'random surfer' who is given a webpage at random and keeps clicking on links, never hitting 'back' but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank."

This however is purely "web-internal" justification of PageRank: a webpage with a high PageRank has a central position in the web seen as a graph structure of nodes and links. There is no claim in the random surfer justification that having a central position in a webgraph should correspond to being of great importance or quality in any more substantial sense of these terms.

A more intriguing attempt at justifying PageRank refers to a proposed analogy with scientific citation (Brin and Page 1998):

"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."

Google's homepage identifies an "underlying assumption" behind the citation analogy:

"PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites."[1]

Similarly, Surowiecki (2004, p. 16) attributes the following quote to Google:

"PageRank capitalizes on the uniquely democratic characteristic of the web by using its vast link structure as an organizational tool. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. Google assesses a page's importance by the votes it receives. But Google looks at more than sheer volume of votes, or links; it also analyzes the page that casts the vote. Votes cast by pages that are themselves 'important' weigh more heavily and help to make other pages 'important'."

The claim is that while one link/vote by itself may not be a very strong indicator of importance, the aggregation of many links/votes is. PageRank, moreover, is a (sophisticated) algorithm for aggregating many links/votes which takes into account the impact of the voting webpage. Therefore, it is concluded that webpages with high PageRanks tend to be more important than other webpages.

This is a "web-external" justification of PageRank that motivates the algorithm by referring to its alleged capacity for tracking importance where the latter is understood in a web-external sense (as information quality, truthiness, informativity, authority, and the like). It is also a wisdom-of-crowds justification. Wisdom of crowds here refers to the idea that "[e]ven if most of the people within a group are not especially well-informed or rational, [the group] can still reach a collectively wise decision" (Surowiecki, 2004, p. xiii–xiv). Surowiecki (2004, p. xiv) is explicit about Google drawing on the wisdom of crowds:

[1] Quoted from http://www.google.com/competition/howgooglesearchworks.html, January 3, 2014. This link is no longer valid as of September 12, 2016; however, this passage is independently cited in http://www.wikiweb.com/page-rank/ and http://en.wikipedia.org/wiki/PageRank.

"This intelligence, or what I'll call 'the wisdom of crowds' is at work in the world in many difference 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."

Here is an explicit recent statement of the WCJPR by a prominent scholar of the web (Thelwall 2013, p. 77):

"[W]eb pages/sites attracting many hyperlinks tend to be more important and popular than those attracting fewer. This is exploited by Google's hyperlink-based algorithm PageRank that helps Google to return highly linked sites at the top of its results."

We do not claim that the Google founders can be unambiguously tied to the WCJPR thesis. Some things they write definitely point in this direction. Other quotes suggest that they think of importance as something purely web-internal, i.e., as a measure of a specific type of popularity (with webmasters), and that this is the end of the story. Typically, passages can be read in both ways. Yet, there is huge difference between claiming that PageRank only tracks popularity (Why should we care?) and claiming that it (also) tracks something "out there," something that we really value, such as truth, authority, relevance, or quality. Hence, regardless of the interpretational issues, we think that the WCJPR thesis is a highly interesting one in its own right and that it has by far not been given the attention it deserves, which is why it is the focus of the present article.

Curiously, even though WCJPR has arguably been in circulation for some 20 years, until recently, there existed no rigorous attempt to actually formulate and prove it. The first (and only) attempt that we know of is that of Masterton et al. (2016). Before we look at the MOA model, we will show that the popular preferential attachment model due to Barabási and Albert's (1999) model (henceforth, the Barabási-Albert (BA) model) of the web is unsuitable as a framework within which WCJPR could be even rigorously formulated, much less proven. One reason why we take up BA for consideration in this context is because it is arguably the most influential theory of linking to date. The reason why the BA model fails also turns out to be instructive for the purpose of acquiring a deeper understanding of what WCJPR involves. Finally, we will, in the end, suggest that a fully satisfactory WCJPR needs to incorporate BA as a part of a more accurate linking model.

3 The BA Model and Attraction to Popularity

The influential preferential attachment model of the linking process is motivated as follows according to Albert and Barabási (2002, p. 73):

"[M]ost real networks exhibit preferential connectivity. For example, a newly created webpage will more likely include links to well known, popular documents with already high connectivity. This example indicates that the probability with which a new vertex connects to the existing vertices is not uniform, but there is a higher probability to be linked to a vertex that already has a large number of connections."

According to the BA model, introduced by Barabási and Albert (1999), the probability for entrant page j to link to incumbent page i is equal to the proportion of all the links in the graph that terminate at page i. That is, where $V = \{p_k : k \in [1, j-1] \subseteq \mathbb{N}\}$ is

the vertex set of the webgraph just prior to the addition of page j, and N_k is the number of in-links of the kth page, then the probability for a link from entrant page j ($p_j \notin V$) to incumbent page i ($p_i \in V$) is

$$P_{\mathrm{BA}}(j \to i) = \frac{N_i}{\sum_{k=1}^{j-1} N_k}.$$

To avoid dividing by zero, and to get the ball rolling, the BA model usually assumes a small random starter graph and then "grows" this graph by consecutively adding vertices and their links.[2]

Albert and Barabási (2002) went on to show that a network generated in accordance with their model organizes itself into a scale-free stationary state with a power law In-Degree distribution not too dissimilar from the WWW. Let us expand on the meaning of this significant result.

Consider first the In-Degree distribution of the WWW. Graphs are useful mathematical constructs for modeling how things are related to each other, with the things represented as vertices on the graph and the relations represented as edges. When we are only concerned about how a few, say less than 30, things are related to each other, one can simply draw graphs to appreciate their topology. But when the number of things and relations is large, and some graphical representations can be very large indeed, we face the problem of how to come to grips with the topology of objects that are unimaginably large and complex. The primary way of dealing with this problem is to define graph statistics. One can find many of these in the literature but some principle ones are clustering coefficient, the shortest path length distribution, and degree distribution.

The cluster coefficient of a vertex is the ratio of the number of links between all those vertices linked to that vertex to the maximum number of such links. The clustering coefficient of a graph is the average of the cluster coefficients of its vertices. One definition of the shortest path length distribution for a graph is the number of vertex pairs with the shortest path length of k, for each k between 0 and $n - 1$. 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, where the former is, for each number N between 0 and $n - 1$, the proportion of vertices with N in-links, and the latter is the same for out-links. As our interest herein is solely with directed graphs and in-link based metrics, we shall adopt the common shorthand of referring to In-Degree distributions as degree distributions.

Degree distributions are of particular interest, because many networks exhibit scale-free degree distributions. A scale-free degree distribution is one that conforms to an inverse power law. One such network is the Internet at the Autonomous Systems (AS) level, another is the WWW. In the former, a vertex represents a subnet roughly corresponding to an Internet service provider (ISP). The links are then inter-ISP connections covered by the border gateway protocol (Vázquez et al. 2002). This network's degree distribution has been empirically found to follow a power law. There have been various studies done to determine the exponent of that power law.

[2] The value of $\sum_{k=1}^{j-1} N_k$ may, or may not, be updated while the links from the jth page are assigned.

The exponent was determined by Albert and Barabási (2002) to be 2.1 ± 0.1 and by Maxim Zhukovskiy et al. (2012) to be 2.276 ± 0.001. That is, the log of the number of ASs with N in-links on the Internet is a linear function of the log of ASs with a gradient of about − 2.2. The webgraph of the WWW represents URLs (webpages) as vertices and hypertext links as directed edges. The WWW is accessed via the Internet, so the two networks are deeply intertwined, but conceptually, one can disassociate them. The In-Degree distribution of the WWW is also scale free and has been empirically found to have an exponent of around 2.1 (Albert et al. 1999; Broder et al. 2000). Figure 1 shows a typical power law ("broomstick") distribution for a portion of the web. As one can see from this figure, strictly speaking, only the tails of WWW degree distributions typically conform to power laws; at higher numbers of in-links, such distributions often deviate from being scale free. This is a common feature of empirical phenomena that exhibit adherence to power laws and has led some, e.g., Clauset et al. (2009), to claim that most of the time, the most we can say is that our data is consistent with the phenomenon in question being governed by a power law up to a certain cut off threshold. We shall conform to the common, though somewhat sloppy, practice of referring to distributions that are consistent with a governing power law up to a certain threshold as "power law" or "scale free" distributions. Where we claim that some model of linking can account for such distributions, this claim is tacitly restricted to the tail of such distributions.

Albert and Barabási (2002) note that the fact that the BA model gives rise to the right degree distribution makes the model a minimally realistic model of linking on the WWW (p. 75):

"It is far from us to suggest that the scale-free model introduced above describes faithfully the topology of the www…Nevertheless, we believe that our model captures in a minimalist way the main ingredients that are responsible for the development of the scale free state observed for the www."

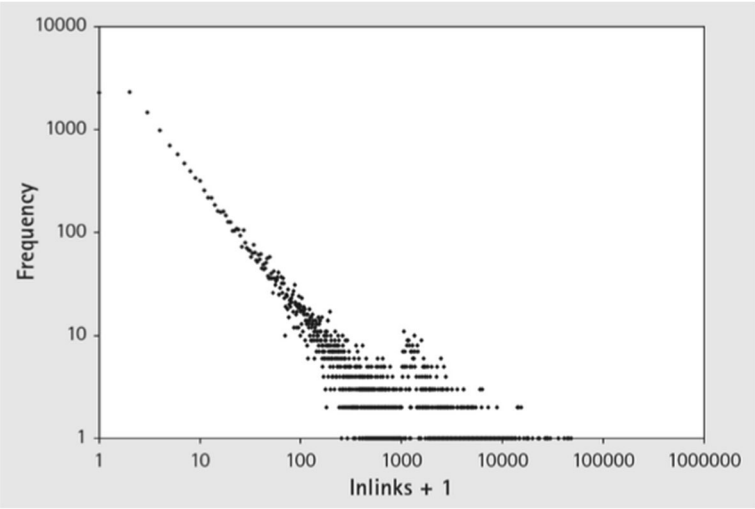

Fig. 1 A typical power law link distribution for a portion of the web (log-log plot, adapted from Thelwall (2013), p. 73)

However, Albert and Barabási (2002) also register a number of limitations of their model (p. 76). For instance, the model assumes that new links appear only when new nodes are added to the network, where as in the WWW, new links are added continuously. They suggest that their model can be extended to incorporate the addition of new links without the network reducing to a fully connected network.

More important for our purposes is the fact that their preferential attachment model is a purely "internal" model of link creation in the sense that the probability of new links to a given node is solely dependent on structural features of the webgraph. Thus, the Google founders' "fundamental assumption" that links are attracted to important webpages—assuming "importance" to refer to web-external qualities such as truthfulness and comprehensiveness—is not valid, or indeed even expressible, in the model. Hence, the links that are created in the process described by the model cannot be interpreted as "votes" for important pages, which means that there is nothing that PageRank can aggregate so as to produce an importance-tracking ranking of webpages reflecting the wisdom of crowds. More precisely, the BA model is compatible with the interpretation that the initial links have been generated in a way that reflects an attraction to importance in the relevant sense. However, links that are added as the webgraph grows cannot reasonably be thus interpreted.

Preferential attachment models are generally unsuitable as frameworks within which a wisdom-of-crowds justification of PageRank (WCJPR) can be rigorously formulated, much less demonstrated. The critical observation is that such models make linking a wholly web-internal affair that can be defined and understood solely on the basis of structural features of the webgraph. This goes not only for the original BA model but also for all relevantly similar models, by which we mean models that take attraction to popularity to be the fundamental mechanism behind linking on the web. One could add that it is counterintuitive to view linking merely as a "sociological" phenomenon, as it were, without linking having any contact with an "external" world outside the network, a point that we will return to in Sect. 5.

4 The MOA Models and Attraction to Importance

Masterton et al. (2016) present two models designed to model the web ecology assumed by the Google founders in their sketch of a wisdom-of-crowds justification for PageRank. A central concept in their model is that links are attracted, not to popularity, but to importance. We know of no other mathematically precise models of this kind. We will spend some time describing these models, as they are less well known than the BA model.

The models are based on two assumptions. First, there is the Google founders' fundamental assumption already alluded to that those responsible for assigning links from a page (source page) 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 be higher the more important that target page is. This is why links from one page to another can be viewed as the webmaster of the source page "voting" for the target page. Second, there is also an assumption that the strength of this attraction of importance varies with the competence of the webmaster of the source page with more competent webmasters administrating more important pages. In particular, the more important the source page, the greater the tendency

of its webmaster to link to other important pages, while the less important the source page, the more random the webmaster will be in her linking behavior. The *basic* MOA model implements the first of these assumptions in a model of Internet ecology, and the *extended* MOA model implements both. (For simplicity, the extended MOA model assumes that each webmaster is administrating a website with only one webpage.)

In both models, the web is modeled as a directed graph, with webpages represented by vertices and links represented by directed edges. The vertices are endowed with a single attribute: importance. Page importance ($I \in [0, 1]$) is sampled from distributions truncated to the unit interval. Any type of truncated distribution is permissible, though herein, we have sampled importance from negative exponential and Pareto distributions.

In both models, the model parameters will include the size of the webgraph ($n \in \mathbb{N}^+$) and the parameters determining the importance sampling; in the case of negative exponential distributions over the unit interval, they are characterized by their expectation (expected page importance = $\alpha \in (0, 0.5)$), while in the case of the Pareto distributions, they are characterized by a minimum value (minimum page importance = $mpi \in (0, 0.2]$) and scale ($\gamma \in \mathbb{R}^+$). Beyond 0-importance indicating the complete lack of all qualities that go towards making a page important and 1-importance indicating their maximal presence, we leave the interpretation of page importance deliberately vague.

Where the models differ is in how the importance of the pages determines the link structure of the webgraph. In the basic model, the probability that the *j*th page links to the *i*-th page is a function of the *i*th page's importance; $g(I_i) : [0, 1] \rightarrow [0, 1]$. The parameters are probability scaling ($Ps \in [0, 1]$), which determines overall link density, and the probability weighting ($Pw \in \mathbb{R}^+$), which determines the linearity of the dependence of linking probability on target page importance. Thus, the probability for any page j ($j \neq i$) to have a link to page i is given by

$$P_{\mathrm{B}}(j \rightarrow \mathrm{i}) = g(I_i) = Ps(I_i)^{Pw}.$$

In the extended model, the probability of page j linking to page i is to be dependent on the importance of both these pages. There are innumerable ways this could be done. What are needed are some constraints on the function $P(j \rightarrow i) = f(I_i, I_j) : [0, 1] \times [0, 1] \rightarrow [0, 1]$, to wit

1. $f(I_i, 1) = Ps(I_i)^{Pw}$. Our basic model holds when a webmaster is fully competent and the webmasters of maximally important pages are fully competent.
2. $f(I_i, 0) = c$, where c is a constant. A totally incompetent webmaster should link randomly, and the webmasters of utterly unimportant pages are totally incompetent.
3. $EP = f(a, a) = g(a)$, Ideally, the expected link density (EP) for a given parameter configuration should be equal to the expected link density for that configuration in our basic model to allow direct comparison of degrees of correlation in web metrics across models without risk of differences in link density skewing the results.
4. The basic and extended models should have the same parameters to make configuration comparison possible.
5. The more incompetent the webmasters, the more they link randomly; and the more competent the webmasters, the more their linking is determined by target page importance.

The following function satisfies these five desiderata:

$$P_{\mathrm{E}}(j \rightarrow i) = f\left(I_i, I_j\right) = Ps(\alpha)^{Pw\left(1-I_j\right)}(I_i)^{Pw\left(I_j\right)}.$$

Proof of 1–3:

$$f(I_i, 1) = Ps(\alpha)^{Pw(1-1)}(I_i)^{Pw(1)} = Ps(I_i)^{Pw}$$

$$f(I_i, 0) = Ps(\alpha)^{Pw(1-0)}(I_i)^{Pw(0)} = Ps(\alpha)^{Pw} = c$$

$$f(\alpha, \alpha) = Ps(\alpha)^{Pw}(\alpha)^{Pw(-\alpha)}(\alpha)^{Pw(\alpha)} = Ps(\alpha)^{Pw} = g(\alpha)$$

Indeed, the following generalization of the previous function also satisfies these five desiderata:

$$P_{\mathrm{E}}(j \rightarrow i) = f\left(I_i, I_j\right) = Ps(\alpha)^{Pw\left(1-h\left(I_j\right)\right)}(I_i)^{Pw.h\left(I_j\right)}$$

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

The function $h(I_j)$ is referred to as the *linking competence* function. A candidate for this function is $h(I_j) = (I_j)^C$ where C is the *competence factor*. We shall herein 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 ($0 \leq C < 1$). Then, the basic model can be viewed as being valid in the limit where all webmasters, irrespective of the importance of the webpage in their charge, are fully competent in their linking ($C = 0$). As C increases, the linking probability becomes less and less dependent upon the target page's importance for any given source page importance until, ultimately, only the webmasters of very important pages will link in a manner dependent upon the target page's importance. Indeed, in the limit 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, the linking probability in the extended model is

$$P_{\mathrm{E}}(j \rightarrow i) = Ps(a)^{Pw\left(1-I_j\right)}(I_i)^{Pw(Ij)}.$$

Figure 2 shows a contour plot of the probability of a link from the jth to the ith page. Note how $P(j \rightarrow i)$ is constant but non-zero where $I_j = y = 0$ and how $P(j \rightarrow i) = Ps(I_i)^{Pw}$ is recovered when $I_j = y = 1$.

In either model, one populates the network with links by, for each prospective link, (metaphorically) flipping a coin with a heads bias equal to the linking probability for

that link and assigning the link if the coin lands heads. As noted in Masterton et al. (2016), this makes link assignment for a given page a Bernoulli trial in the basic model and a Poisson trial in the extended model. Completing this process samples a webgraph for the web ecology specified by the pair of model and parameter configuration.

The striking fact about the MOA models is that they allow for a rigorous statement and proof of the WCJPR thesis and also for a similar thesis about In-Degree: impact, as measured by PageRank or In-Degree, is perfectly correlated with importance "in the limit," i.e., as the webgraph grows. In particular, impact implies importance in a statistical sense if the webgraph is sufficiently large. Specifically, Masterton et al. (2016), pp. 962–964, proved the following three theorems (in the case of Theorem 2 drawing on a theorem from Fortunato et al. (2008)):

Theorem 1: If linking probability is a monotonically increasing function of target page importance, then as the number of pages in a webgraph goes to infinity, the probability that In-Degree is perfectly correlated with page importance in that webgraph tends to one.

Theorem 2: If linking probability is a monotonically increasing function of target page importance, then as the number of pages in a webgraph goes to infinity, the probability that PageRank is perfectly correlated with page importance in that webgraph tends to one.

Theorem 3: If linking probability is a monotonically increasing function of target and source page importance, then as the number of pages in a webgraph goes to infinity, the probability that In-Degree is perfectly correlated with page importance in that webgraph tends to one.

The first theorem states that the ranking induced by In-Degree is perfectly correlated with webpage importance in the limit in the basic MOA model. The second theorem states the same result but for PageRank instead of In-Degree. The third theorem states that the result generalizes to In-Degree in the extended MOA model. Thus, if the

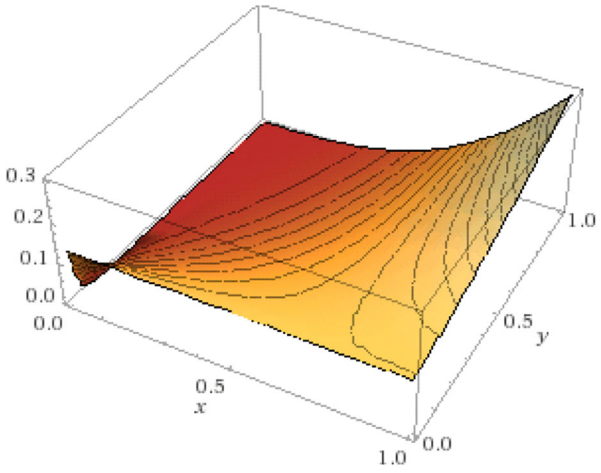

Fig. 2 A 3D contour plot of $P(j \rightarrow i) = Ps(\alpha)^{Pw(1-I_j)}(I_i)^{Pw(pi_j)}$, where $I_j = y$, $pi_i = x$, $\alpha = 0.75$, $Ps = 0.3$, $Pw = 3$, and $P(j \rightarrow i) = z$

ecology of the WWW is accurately described by the basic MOA model, then, given the WWW's size, we can be practically certain that rankings of webpages by PageRank—such as those performed by Google—will perfectly agree with rankings of webpages by their importance.

We now turn to the important question whether the MOA models satisfy the constraint of being minimally realistic models of the WWW. In other words, do they generate webgraphs reflecting the degree distribution of the web? Unsurprisingly, given the linking probability functions characterizing the basic and extended models, the degree distribution of the webgraph generated by the MOA models for a particular parameter configuration is entirely dependent upon how importance is distributed across the webpages of that webgraph. Indeed, one can prove the following theorem.

Theorem 4: In the basic MOA model, where attraction to importance is linear ($Pw = 1$) and link density is maximal ($Ps = 1$), the degree distribution of a graph will almost surely converge on the importance distribution from which webpage importance was sampled.

Proof: see Appendix.

As an immediate consequence of Theorem 4, we get the following result:

Corollary: If importance is distributed according to a power law, then the degree distribution of a graph generated in the basic model with $Pw = 1$, will converge on being scale free in the limit.

We can easily confirm this result in a computer simulation of the basic MOA model (Fig. 3).

As shown in Fig. 3, we get the characteristic broomstick distribution of links by selecting a corresponding importance distribution.

While the connection between importance distribution and degree distribution is demonstrably direct in the basic model where attraction to importance is linear, the general point holds in both the extended and the basic models irrespective of parameterization: for a particular model parameterization and webgraph degree distribution, there will be an importance distribution such that that parameterized model almost surely generates that degree distribution in the limit. That importance distribution may differ markedly from the target degree distribution (Fig. 4), but there will still be some importance distribution that does the job.

The upshot of all this is that for any parameterization of our basic or extended models, one can specify a distribution of importance so that the webgraphs generated are practically certain, if sufficiently large, to have the same degree distribution as the WWW. For instance, if importance is Pareto distributed with a minimum level of importance of 0.0001 and an expected level of importance of 0.00066, then the degree distribution of the resultant webgraphs in the basic model with $Ps = Pw = 1$ is roughly that of the WWW (Fig. 5).

Thus, our basic and extended models can account for the degree distribution of the WWW. However, one may still doubt that the MOA models are minimally realistic despite their capacity to account for the degree distribution of the WWW. There is something quite ad hoc about choosing the importance distribution in the model in

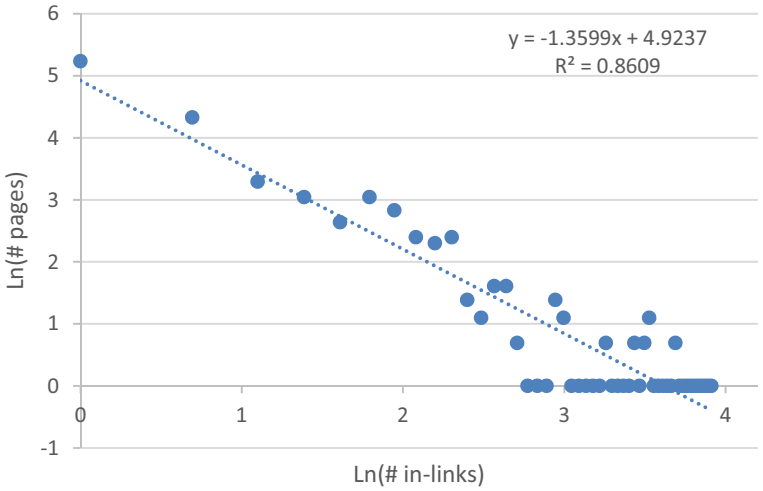

Fig. 3 Ln/Ln plot of the degree distribution of a 1000-page webgraph generated in the basic model ($Ps = Pw = 1$) with importance Pareto distributed with minimum importance level set at 0.0001 and expected importance at 0.01

order to get the right webgraph topologies being generated. Indeed, in the extended model, the importance distributions would have to be quite peculiar to result in the desired topologies and the only reason for adopting such peculiar distributions would be to get those topologies. This concern is arguably not as serious for the basic MOA model; one can argue that importance being distributed according to a power law and important webpages being generally rare are natural assumptions and so argue that the basic model really accounts for the degree distribution on the web. However, this would be very much a non-standard explanation of the cause of this degree distribution and one that is peculiarly sensitive to the characteristic parameters of the cited importance distribution. We now turn to a more detailed discussion of these and related concerns.

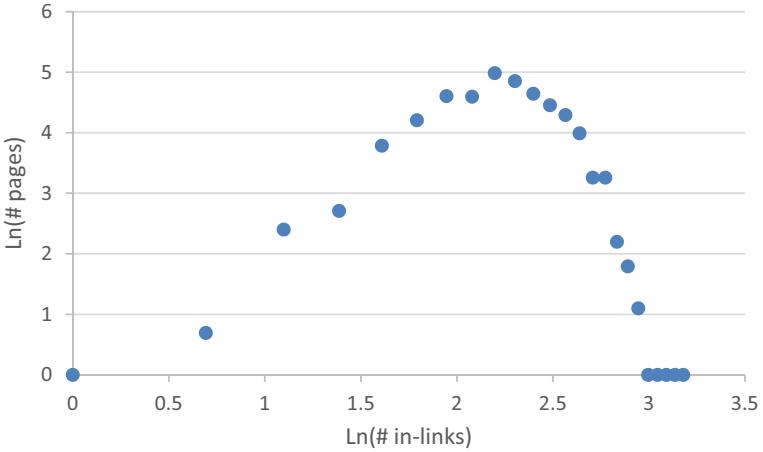

Fig. 4 Ln/Ln plots of the degree distribution of a 1000-page webgraph generated in the extended model ($Ps = Pw = 1$) with importance Pareto distributed with minimum importance level set at 0.0001 and expected importance at 0.01

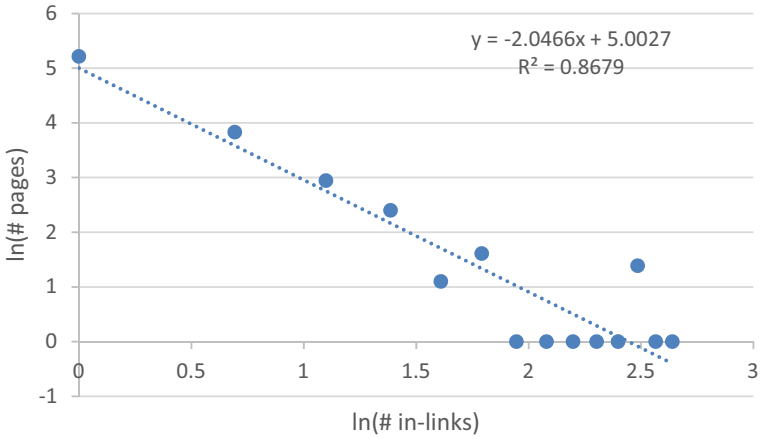

Fig. 5 Ln/Ln plot of the degree distribution of a 1000-page webgraph generated in the basic model ($Ps = Pw = 1$) with importance Pareto distributed with minimum importance level set at 0.0001 and expected importance at 0.00066. Note that the power of the degree distribution is 2.05 which is close to the value given by Albert et al. (1999) for the WWW

5 Towards Dual Models of the Linking Process

The upshot of our discussion so far is a dilemma for anyone who finds a wisdom-of-crowds justification of PageRank (WCJPR) and similar in-link-based ranking algorithms plausible: there seems to be no model on the market which both allows for a precise statement and proof of the WCJPR thesis and at the same time is minimally realistic in the sense of naturally giving rise to scale-free webgraphs with the same degree distribution as the real WWW. The MOA models of Masterton, Olsson, and Angere satisfy the former condition by allowing for a precise statement and proof of the thesis in question, but they fail to generate the right kind of webgraphs. As we saw, the latter claim is in need of some qualification. The MOA models can in fact generate any degree distribution, including a degree distribution that corresponds to the WWW, but the way this is accomplished seems entirely ad hoc, though perhaps slightly less so for the basic model. The BA model of Barabási and Albert, by contrast, satisfies the latter condition of giving rise to the right kind of webgraphs (for the right reasons), but does not allow for a precise statement, much less proof, of the WCJPR thesis.

As we noted, there are independent reasons to think that the BA model is inaccurate as a model of the linking process. It is implausible to view the linking process as a wholly web-internal affair. Surely, people link to other webpages not only because others have linked to them, thereby making those pages more visible in search engines and other web services.[3] They must also link to what they themselves consider important and perhaps find by chance or through offline friends. Similarly, there are independent reasons to think that the MOA models are not completely faithful to the

[3] Cf. Thelwall (2013, p. 72): "Search engines repeatedly claim that they do not manipulate their results for money, so how do they decide which sites to prioritize? The primary data that they use to identify popular websites is the structure of the Web itself in the form of hyperlinks: the more links point to a website, the more likely it is to have a large audience (Brin and Page, 1998). This creates a rich-get-richer effect, because popular websites attract more visitors from commercial search engines, making them even more popular and likely to attract even more links."

phenomena that they attempt to represent. Surely, people link to other webpages not only because of the intrinsic qualities of those pages but also because others have linked to them, thereby making those pages more visible.

Since the BA and MOA models seem to reflect complementary rather than contrasting ways of looking at the linking process, the obvious move would be to combine them. We will call such combined models dual models. Dual models recognize two mechanisms behind linking on the web: attraction to popularity, as in the BA model, and attraction to importance, as in the MOA model.

Now, there are two main dual models arising from combining the BA model with either the basic or the extended MOA model. For instance, a linear combination of the BA model with the basic (B) MOA model gives rise to the following dual model ($0 < \mu_j < 1$):

$$P_{\mathrm{B+BA}}(j \rightarrow i) = \mu_j \cdot P_{\mathrm{B}}(j \rightarrow i) + \left(1 - \mu_j\right) \cdot P_{\mathrm{BA}}(j \rightarrow i)$$

We get different variations of this model by choosing the weight μ_j differently. A high value makes attraction to importance the main factor; a low value makes attraction to popularity the dominant mechanism. Moreover, since "combining" can mean a lot of different things and does not necessarily have to be interpreted linearly, we would expect there to be more than two plausible main dual models. Indeed, even if we fix on a particular linear combination of models and relative weights, there are a lot of parameters that can be given different values.

The fundamental question now is whether there are dual models that give rise to a power law distribution of in-links corresponding to the web for reasons that are not ad hoc and such that PageRank (and In-Degree) are well correlated with importance. Such a model would ideally allow for an exact statement and proof of the WCJPR thesis while satisfying the requirement of minimal realism with regard to the degree distribution of the WWW. To wit, the exhibition of such a model would be a strong argument for the rationality of using link-based ranking on the real web. If, by contrast, no such model can be found, we would have reason to doubt the rationality of such ranking on the real web. Either way, the importance of the question can hardly be exaggerated. We conjecture that there are dual models of the kind in question, but we have to leave a detailed inquiry into the matter for future work.

6 Conclusion

In this paper, we attempted to identify the state of the art regarding the more precise statement of the claim that PageRank (and other in-link-based ranking algorithms) can be justified with reference to the wisdom of crowds. Our first point was that while current preferential attachment models are, in a minimalist sense, realistic models of the (complete) web as they naturally give rise to scale-free networks, they do not allow for formulation, much less proof, of the wisdom-of-crowds thesis.

Our second point was that while the recent linking models proposed by Masterton et al. (2016) do allow for the formulation and proof of the thesis in question, they are

not minimally realistic models of the web because, as we demonstrated, although they can give rise to scale-free networks of the required kind, they do so in an ad hoc manner. We concluded that there is, to the best of our knowledge, at present no fully satisfactory wisdom-of-crowds justification for PageRank or similar in-link-based algorithms.

Finally, we proposed, as a remedy, that future work should explore dual models of the linking process, i.e., models that combine preferential attachment models with the kind of models explored by Masterton et al. into one unified account of the linking process. We conjectured that there are dual models that are realistic models of the web and at the same time allow for the rigorous formulation and proof of the wisdom-of-crowds thesis. We left a detailed investigation into the validity of this conjecture for future work.

Acknowledgements We would like to thank two anonymous reviewers for their helpful comments and suggestions.

Appendix

Theorem 4: In the basic model, where attraction to importance is linear ($Pw = 1$) and link density is maximal ($Ps = 1$), the degree distribution of a graph will almost surely converge on the importance distribution from which webpage importance was sampled.

Proof: In the basic model, where $Pw = 1$ and $Ps = 1$, the linking probability function is

$$P_B(j \rightarrow i) = I_i$$

By the law of large numbers—due to link assignment being a Bernoulli trial—and every page being able to link to all others but once and never to itself, the probability is one that as $n \longrightarrow \infty$, so $\frac{N_i}{n-1} \longrightarrow I_i$, where n is the size of the webgraph and N_i is the number of in-links to the ith page. As this is true of all pages, so $\frac{N}{n-1} \longrightarrow I$ in the limit almost surely (see Masterton et al. (2016) for the full proof of this).

By convergence in distribution it follows that if importance in a population of n webpages is sampled from a PDF $\rho(I)$, then as $n \longrightarrow \infty$, so $\frac{n_{I \in [a,b]}}{n} \longrightarrow \int_a^b \rho(I) dI$, where $n_{I \in [a,b]}$ is the number of pages with importance $I \in [a, b] \subseteq [0, 1]$.

As $\frac{n_{I \in [a,b]}}{n}$ is the proportion of pages with importance in $[a, b]$, and as $\frac{N}{(n-1)} \longrightarrow I$ in the limit almost surely, so $\frac{n_{N \in [a(n-1), b(n-1)]}}{n} \longrightarrow \frac{n_{I \in [a,b]}}{n}$ in the limit almost surely, where $n_{N \in [a(n-1), b(n-1)]}$ is the number of webpages with between $a(n-1)$ and $b(n-1)$ in-links. As there can only be a whole number of in-links to a page, so $n_{N \in [a(n-1), b(n-1)]} = n_{N \in [s, t]}$, where $a(n-1) \leq s \leq a(n-1) + 1$, $(b)(n-1) - 1 \leq t \leq (b)(n-1)$ and $s, t \in \mathbb{N}$. Further, given these restrictions on s and t:

$$n_{I \in [a,b]} = n_{I \in \left[\frac{s}{(n-1)}, \frac{t}{(n-1)}\right]}.$$

It follows that almost surely in the limit

$$\frac{n_{N\in[s,t]}}{n}\longrightarrow\frac{n_{I\in\left[\frac{s}{(n-1)},\frac{t}{(n-1)}\right]}}{n}.$$

As $\frac{n_{I\in\left[\frac{s}{(n-1)},\frac{t}{(n-1)}\right]}}{n}\longrightarrow\int_{\frac{s}{(n-1)}}^{\frac{t}{(n-1)}}\rho(I)dI$ in the same limit by substitution into $\frac{n_{I\in[a,b]}}{n}\longrightarrow\int_a^b\rho(I)dI$, so in the limit almost surely

$$\frac{n_{N\in[s,t]}}{n}\longrightarrow\int_{\frac{s}{(n-1)}}^{\frac{t}{(n-1)}}\rho(I)dI.$$

The In-Degree distribution of a webgraph is defined as $P(N):=\frac{n_N}{n}$, where n_N is the number of webpages with N in-links. Naturally,

$$\sum_s^t P(N):=\frac{n_{N\in[s,t]}}{n}.$$

It immediately follows that almost surely in the limit

$$\sum_s^t P(N)\longrightarrow\int_{\frac{s}{(n-1)}}^{\frac{t}{(n-1)}}\rho(I)dI.$$

This is so for all values $s,t\in[0,n-1]\subseteq\mathbb{N}, t>s$. By definition, both sides of the convergence equal one where $s=0$ and $t=n-1$.

Thus, in this way, degree distribution almost surely converges on the distribution from which page importance is sampled as $n\longrightarrow\infty$ in the basic model where attraction to importance is linear and link density is maximal.

QED

References

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47–97.

Albert, R., Jeong, H., & Barabási, A. L. (1999). Diameter of the world-wide web. *Nature, 401*, 130–131.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509–512.

Brin S, and Page L (1998) The anatomy of a large-scale hypertextual web search engine, WWW 1998 (Seventh International World-Wide Web Conference), Brisbane, Australia.

Brin S, Page L, Motwami R, and Winograd T. (1998) The PageRank citation ranking: bringing order to the web, Stanford University Technical Report.

Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, and Wiener J (2000) Graph structure in the web, In Proceedings of the Ninth International World Wide Web Conference, Amsterdam, The Netherlands, May 15–19. URL: http://www.immorlica.com/socNet/broder.pdf

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review, 51*(4), 661–703.

Fortunato S, Boguñá M, Flammini A, and Menczer F (2008) Approximating PageRank from in-degree. In: Eds. Aiello, W, Broder A, Janssen J, Milios E (eds.) Algorithms and Models for the Web-Graph.; 59–71.

Franceschet, M. (2011). PageRank: standing on the shoulders of giants. *Communications of the ACM, 54*(6), 92–101.

Masterton, G., Olsson, E. J., & Angere, S. (2016). Linking as voting: how the Condorcet jury theory in political science is relevant to webometrics. *Scientometrics, 106*, 945–966.

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*. London: Little Brown.

Thelwall, M. (2013). Society on the web. In W. H. Dutton (Ed.), *The Oxford handbook of internet studies* (pp. 69–85). Oxford: Oxford University Press.

Vázquez, A., Pastor-Satorras, R., & Vespignani, A. (2002). Large-scale topological and dynamical properties of the Internet. *Physical Review E, 65*(6), 066130.

Wills, R. S. (2006). Google's PageRank: the maths behind the search engine. *The Mathematical Intelligencer, 28*(4), 6–11.

Zhukovskiy, M., Vinogradov, D., Pritykin, Y., Ostroumova, L., Grechnikov, E., Gusev, G., Serdyukov, P., & Raigorodskii, A. (2012). Empirical validation of the Buckley-Osthus model for the web host graph: degree and edge distributions. In: Proceedings of the 21st ACM international conference of information and knowledge management; 1577–158.