# Measures for Measures

## Supplementary Online Information

S. Lehmann[1]
A. D. Jackson
B. E. Lautrup

[1]Electronic Address: lehmann@nbi.dk

# Contents

# 1 Data

## 1.1 Acquisition

This section provides a short description of the acquisition and processing of data from the SPIRES (Stanford Public Information REtrieval System) data base. Ultimo 2003, the database manager[1] provided us with a text file containing the following information for each paper in spires: Title, List of authors, Publication information, References, Sub-field classification, and Keywords. We added this information to a relational data base (MySQL) in order to create a network of authors and papers. The data used here was generated by querying the resulting data base. Thus, only citations from within the data base are counted. In order to ensure the validity of the data, we have also used an independent route to generate the data; we employed the programming language Perl to extract the relevant information from the main text file.

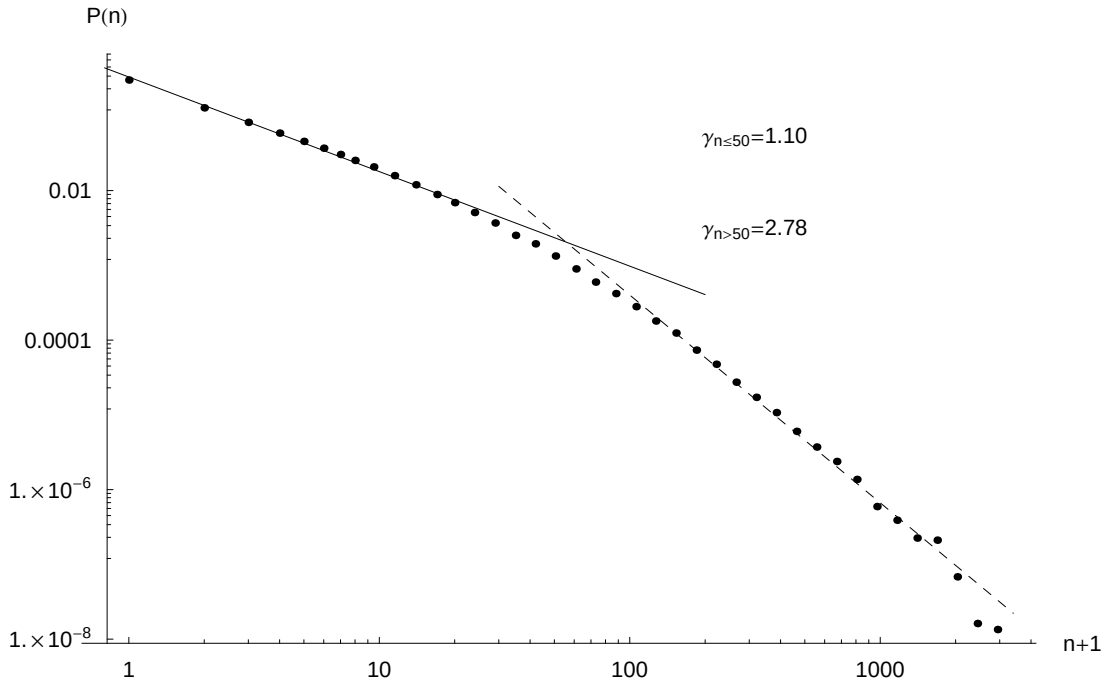
One main problem in processing this data is identifying authors uniquely, since the same author can represent his name in may different ways (e.g. John James Smith, John J. Smith, J. J. Smith, J. Smith, etc.). For the data shown, authors were identified by last name and first two initials. Checks were performed using (i) last name and all initials and (ii) last name and first initial only. These two cases represent approximate upper and lower bounds on the number of unique authors in the data base. No significant changes were found in either case.

## 1.2 Statistics

Our data set consists of all publications by "academic scientists"—defined as those with 25 or more published papers—in the theory subfield of SPIRES. The resulting data set contains 274 470 papers written by 6 737 authors; this data set is highly homogeneous [1]. One possible description of the distribution of citations of papers is a double power-law structure[2]. Specifically the probability that a paper will recieve $n$ citations is approximately proportional to $(n+1)^{-\gamma}$ with $\gamma = 1.10$ for $n \leq 50$ and $\gamma = 2.78$ for $n > 50$. These features of the global distribution are also present in the conditional probabilities for subgroups of authors binned according to most measures of quality. In virtually all cases, the conditional probabilities can also be described accurately by separate power-laws in each of two regions with a relatively sharp transition between the regions. As one might expect, authors with more citations are described by flatter distributions (i.e., smaller values of $\gamma$) and a somewhat higher transition point. Supplementary Figure 1 displays the total distribution of citations as a binned and normalized histogram.

[1] Travis C. Brooks from the SLAC Library.

[2] The double power-law description is only one of many possible parameterizations of the data; better fits to the data can certainly be made, but any increase in the number of parameters demands a justification.




Supplementary Figure 1: Logarithmically binned histogram of the citations counts of all papers by authors with more than 25 publications in the theory subsection of SPIRES. The data is normalized and the axes are logarithmic.

# 2 The Bayesian Method

We have binned the SPIRES authors and their citation records according to each of the four tentative measures, $m$, described in the main paper. Studies performed on the first 25, first 50 and all papers of authors with a given value of $m$ indicate the absence of temporal correlations in the citation distributions of individual authors. In practice, we bin authors in deciles according to their value of $m$ and papers logarithmically, due to the asymptotic power law behavior noted above. We have confirmed that the results here are relatively insensitive to binning effects.

We have constructed the prior distribution, $p(\alpha)$, that an author is in author bin $\alpha$ (in the case of decile bins $p(\alpha) = 1/10$ for all bins) and the conditional probability, $P(i|\alpha)$, that a paper by an author in bin $\alpha$ will fall in citation bin $i$. For each bin $\alpha$, the $P(i|\alpha)$'s are simply citation distributions analogous to the normalized histogram displayed in Supplementary Figure 1, but constructed using only papers written by authors in bin $\alpha$.

Now, we wish to calculate the probability, $P(\{n_i\}|\alpha)$, that an author in bin $\alpha$ will have a citation record with $n_i$ papers in each citation bin $i$. To do this, we assume[3] that citations for the $M$ papers written by a given author with $n_i$ papers in citation bin $i$ are obtained

[3]The argument here is based on the additional simplifying assumption that the distribution of total papers per author is the same in all author bins. This assumption, which is readily relaxed, has no significant effect on the results presented here.

from $M$ independent random draws on the appropriate distribution, $P(i|\alpha)$. Thus,

$$P(\{n_i\}|\alpha) = M! \prod_i \frac{P(i|\alpha)^{n_i}}{(n_i)!} \ . \tag{1}$$

We have already noted the absence of large-scale temporal variations in $P(i|\alpha)$ during an author's scientific life. Other correlations could be present. For example, one particularly well-cited paper could lead to an increased probability of high citations for its immediate successor(s). While it is difficult to demonstrate the presence or absence of such correlations, the results below provide *a posteriori* indications that such correlations, if present, are not overly important. We can invert the probability $P(\{n_i\}|\alpha)$ using Bayes' Theorem to obtain

$$\begin{aligned} P(\alpha|\{n_i\}) &= \frac{P(\{n_i\}|\alpha)\, p(\alpha)}{p(\{n_i\})} \\ &= \frac{p(\alpha) \prod_k P(k|\alpha)^{n_k}}{\sum_{\alpha'} p(\alpha') \prod_{k'} P(k'|\alpha')^{n_{k'}}} \ . \end{aligned} \tag{2}$$

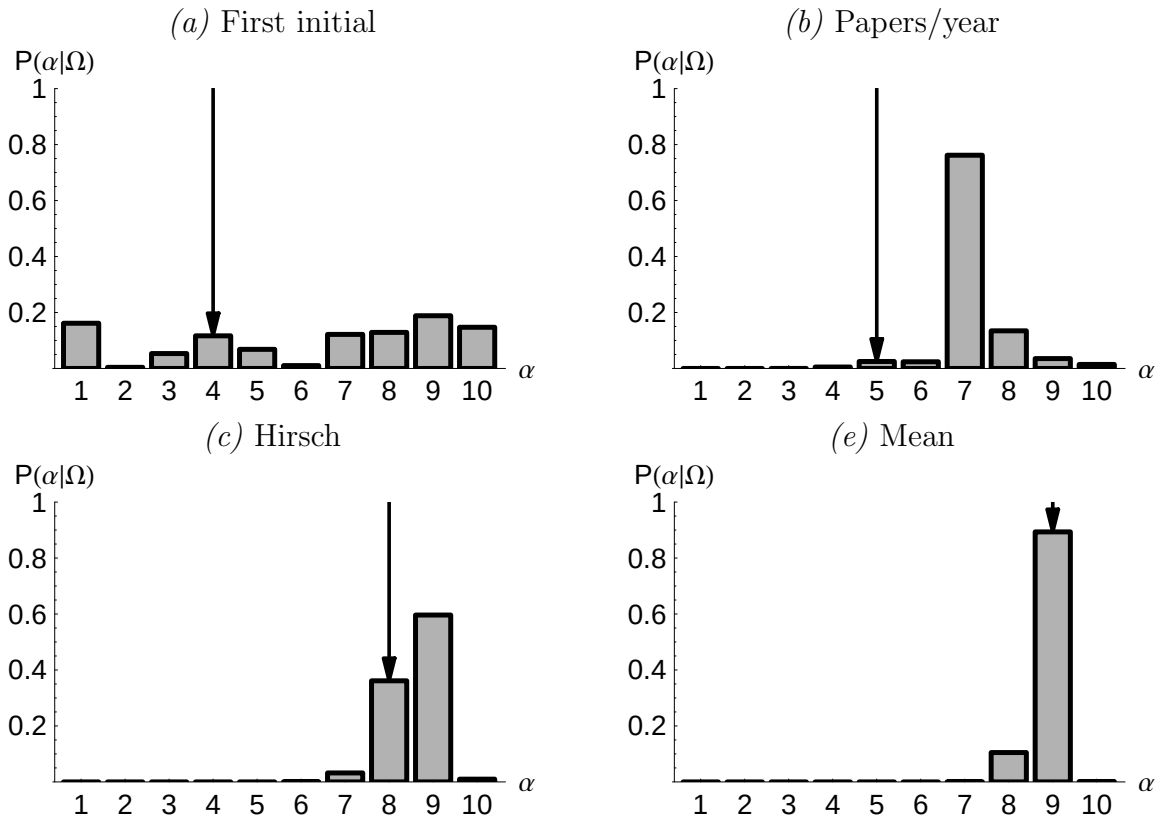
Note that the combinatoric factors cancel.

The quantity $P(\alpha|\{n_i\})$, which represents the probability that an author with citation record $\{n_i\}$ belongs in quality bin (i.e., decile) $\alpha$, is of primary interest. While any given measure (e.g., the mean number of citations per paper) can be calculated immediately from an author's citation record $\{n_i\}$, the calculated values of $P(\alpha|\{n_i\})$ provide more detailed and reliable information. By exploiting differences between the various conditional probabilities, $P(\{n_i\}|\alpha)$, as a function of $\alpha$, Supplementary Equation (2) determines the appropriate decile value of $m$ (or its most probable value) using all statistical information in the data. By using the an author's full citation record, the large fluctuations which are inevitable in e.g. the number of citations of the author's maximally cited paper are thereby materially reduced. Further, by providing us with values of $P(\alpha|\{n_i\})$ for all $\alpha$, we have a statistically trustworthy gauge of whether the resulting uncertainties in the assigned value of $m$ are sufficiently small for it to be a reliable measure of author quality.

## 2.1 A Single Author Example

In short, Supplementary Equation (2) provides us with a measure of an author's expected lifetime quality along with information which allows us to assess the reliability of this determination. The confidence with which we can assign a value of $m$ approaches 100% exponentially with the total number of published papers. As we shall see, it is also sensitive to the quality measure chosen. To gain an understanding of $P(\alpha|\{n_i\})$, let us consider a concrete example.

We will investigate the (real) citation record of author $A$ with citation record $\Omega$. Supplementary Figure 2 shows the probabilities that $A$ will lie in each of the deciles using the four different measures defined in the main text. It is clear from the figure that there are significant differences in the results obtained, both in the apparent accuracy of their




Supplementary Figure 2: A single author example. We analyze the citation record of author $A$ with respect to four different measures. Author $A$ has written a total of 88 papers. The mean number of citations per paper is 26, Hirsch's $h$-index is 29 for this author, the maximally cited paper has 187 citations, and papers have been published at the average rate of 2.5 papers per year. The various panels give the probability that author $A$ belongs to each of the ten deciles based on the corresponding measure; the vertical arrow shows the decile bin to which author A is assigned by direct calculation of each measure.

predictions and, more importantly, in the corresponding uncertainties. In all cases, large uncertainties are due to the fact that the conditional probabilities, $P(i|\alpha)$ are largely independent of $\alpha$. Such independence is to be expected in the case of the alphabetic binning of authors, and the inability of the citation record to identify the first initial of author $A$'s name is hardly surprising. The figure also suggests that, although this distribution has a peak, the number of papers published per year is unable to determine to which bin author $A$ was assigned. The mean number of citations per paper provides an accurate determination with a small uncertainty, thus the use of Supplementary Equation (2) has compensated for the large fluctuations which might have been expected from the use of mean citation rate as a measure of quality. Hirsch's measure falls somewhere between the best and worst choice of measures.

## 2.2 Construction of Figure 1 in Main Paper

Measures of quality are of value only to the extent that they can be assigned to individual authors with high confidence. The methods described above allow us to determine this confidence for any choice of measure in a manner which is value-free and completely quantitative. In order to perform this evaluation, we repeat the calculations leading to Supplementary Figure 2 for all authors in the SPIRES database. We calculate the probability, $P(\beta|\alpha)$, which is the probability, averaged over the authors in author bin $\alpha$, that the full citation record of an author initially assigned to bin $\alpha$ by the measure under consideration was drawn at random on the distribution $P(i|\beta)$, appropriate for author bin $\beta$. Stated simply, $P(\beta|\alpha)$ is the probability that an author assigned to be in bin $\alpha$ is predicted to lie in bin $\beta$. Thus, $P(\beta|\alpha)$ is the average

$$P(\beta|\alpha) = \frac{1}{N_\alpha} \sum_{\{n_i\}\in\alpha} P(\beta|\{n_i\}), \tag{3}$$

where $N_\alpha$ is the number of authors in bin $\alpha$. The figure in the main paper is simply the "stacked" results of this calculation, that is, for each measure, we plot the array of probabilities

$$\begin{array}{cccc} P(1|10) & P(2|10) & & P(10|10) \\ \vdots & & \iddots & \\ P(1|2) & P(2|2) & & P(10|2) \\ P(1|1) & P(2|1) & \cdots & P(10|1) \end{array}, \tag{4}$$

where each probability $P(\beta|\alpha)$ is represented as a black square with area proportional to the corresponding probability.

## 2.3 Scaling

In this section, we will consider the question of how many published papers are required in order to make a reliable prediction of the lifetime quality measure for a given author. (Here, we will consider only results using the mean citation rate as a measure.) Obviously, if this number is sufficiently small, analysis along the lines presented here can provide a practical tool of potential value in predicting long-term scientific accomplishment. In order to address this question, we will look at how $P(m|\{n_i\})$ scales as a function of the the number of papers in each bin for an average author. Assume that an average author belonging to bin $\alpha$ draws $M$ papers at random from the distribution of $P(n|\alpha)$. The most most probable number of papers in each citation bin will thus be given as $n_i = MP(i|\alpha)$. Inserting this result into Supplementary Equation (2) and discarding all fixed factors, we find that

$$P(\alpha|\{n_i\}) \sim p(\alpha) \left( \prod_i P(i|\alpha)^{P(i|\alpha)} \right)^M . \tag{5}$$

For the same citation record, $\{n_i\}$, a similar expression permits determination of the probability that this average author will be assigned to any bin. It is clear from Supplementary

Equation (5) that the probability of assigning this average author to the wrong bin will ultimately vanish exponentially with $M$. Given enough papers, the bin with the largest probability will ultimately dominate. To correctly assign the most probable to outer deciles 1, 2, 3 and 8, 9, 10 at the 90% confidence level requires respectively $M = 10$, 40, 50, and 50, 50, 30 papers.

All quality measures have difficulty in making correct assignments to deciles 4–7. This apparent difficulty is due to our decision to group authors by deciles. It can be understood by assuming that the distribution of intrinsic author quality has a maximum at some non-zero value. Such an assumption seems reasonable if we imagine that there is a natural high-end cutoff and that academic appointment procedures filter out the least able. For any such distribution, the probability density will be highest for authors in the vicinity of this maximum. The binning of authors by deciles or percentiles then invites us to make distinctions where no material quality difference exists. The results of the main figure in the actual commentary remind us that we cannot do so. On the other hand, the probability that an author can be correctly assigned to the bins $4, 5, 6, 7$ collectively on the basis of 50 publications is higher than 90%.

# 3 The Median

Here, we show that the median of $\mathcal{N} = (2N + 1)$ random draws on *any* normalized probability distribution, $q(x)$, is normally distributed in the limit $\mathcal{N} \to \infty$. To this end we define the integral of $q(x)$ as

$$Q(x) = \int^x q(x')dx' \tag{6}$$

Evidently, $Q(x)$ grows monotonically from 0 to 1 independent of $q(x)$. The 'median' of this sample is defined as that value of $x$ such that *(i)* one draw has the value $x$, *(ii)* $N$ draws have a value less than or equal to $x$, and *(iii)* N draws have a value greater than or equal to $x$. The probability that the median is at $x$ is now given as

$$P_{x_{1/2}}(x) = \frac{(2N+1)!}{1!N!N!} q(x) Q(x)^N [1 - Q(x)]^N \ . \tag{7}$$

For large $N$, the maximum of $P_{x_{1/2}}(x)$ occurs at $x = x_{1/2}$ where $Q(x_{1/2}) = 1/2$. Expanding the logarithm of $P_{x_{1/2}}(x)$ about its maximum value, we see that

$$P_{x_{1/2}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x - x_{1/2})^2}{2\sigma^2}] , \qquad \sigma^2 = \frac{1}{4\mathcal{N} q(x_{1/2})^2} \ . \tag{8}$$

An identical argument applies for any percentile—not just the median. E.g., for constructing the distribution of the 90th percentile, we would construct the the probability that $9N$ draws have a value less than $x$, $N$ draws have a value greater than $x$, and one draw has the value of $x$. The distribution of *any* percentile, $0 \leq z \leq 1$ measured with $\mathcal{N}$ random draws on any distribution is a Gaussian with a maximum at some $x_z$ such that $Q(x_z) = z$ and $\sigma^2 \sim (\mathcal{N} q(x_z)^2)^{-1}$.

# 4 Explicit $P(\beta|\alpha)$

In this section we attach the actual probabilities behind the figure in the main text; the numbers below correspond to the array in Supplementary Equation 4. As a visual help, the diagonals are set in bold face.

## 4.1 First Initial

$$\begin{pmatrix}
0.0761 & 0.2104 & 0.0686 & 0.0709 & 0.0819 & 0.1148 & 0.1010 & 0.0747 & 0.0801 & \mathbf{0.1216} \\
0.0839 & 0.1869 & 0.0710 & 0.0772 & 0.0866 & 0.1100 & 0.1107 & 0.0818 & \mathbf{0.0876} & 0.1042 \\
0.0820 & 0.1902 & 0.0700 & 0.0760 & 0.0857 & 0.1071 & 0.1147 & \mathbf{0.0820} & 0.0868 & 0.1054 \\
0.0851 & 0.1698 & 0.0715 & 0.0781 & 0.0927 & 0.1080 & \mathbf{0.1248} & 0.0841 & 0.0887 & 0.0972 \\
0.0790 & 0.2127 & 0.0695 & 0.0736 & 0.0847 & \mathbf{0.1142} & 0.1113 & 0.0776 & 0.0817 & 0.0958 \\
0.0814 & 0.1886 & 0.0713 & 0.0757 & \mathbf{0.0887} & 0.1099 & 0.1144 & 0.0817 & 0.0857 & 0.1025 \\
0.0814 & 0.1986 & 0.0680 & \mathbf{0.0751} & 0.0851 & 0.1057 & 0.1154 & 0.0802 & 0.0858 & 0.1048 \\
0.0791 & 0.2029 & \mathbf{0.0719} & 0.0728 & 0.0831 & 0.1096 & 0.1052 & 0.0779 & 0.0826 & 0.1150 \\
0.0776 & \mathbf{0.2276} & 0.0703 & 0.0724 & 0.0822 & 0.1161 & 0.1028 & 0.0757 & 0.0800 & 0.0953 \\
\mathbf{0.0841} & 0.1885 & 0.0712 & 0.0770 & 0.0857 & 0.1089 & 0.1129 & 0.0816 & 0.0876 & 0.1025
\end{pmatrix}.$$

## 4.2 Papers Per Year

$$\begin{pmatrix}
0.4493 & 0.0979 & 0.0347 & 0.0319 & 0.0462 & 0.0415 & 0.2412 & 0.0276 & 0.0169 & \mathbf{0.0128} \\
0.3591 & 0.1180 & 0.0452 & 0.0453 & 0.0637 & 0.0565 & 0.2204 & 0.0437 & \mathbf{0.0273} & 0.0208 \\
0.3134 & 0.1118 & 0.0484 & 0.0503 & 0.0674 & 0.0614 & 0.2388 & \mathbf{0.0536} & 0.0320 & 0.0228 \\
0.2321 & 0.1018 & 0.0518 & 0.0616 & 0.0839 & 0.0758 & \mathbf{0.2547} & 0.0683 & 0.0407 & 0.0292 \\
0.2321 & 0.1280 & 0.0672 & 0.0674 & 0.0861 & \mathbf{0.0780} & 0.1994 & 0.0649 & 0.0436 & 0.0332 \\
0.2130 & 0.1256 & 0.0679 & 0.0711 & \mathbf{0.0891} & 0.0792 & 0.2051 & 0.0699 & 0.0455 & 0.0336 \\
0.2024 & 0.1308 & 0.0768 & \mathbf{0.0746} & 0.0885 & 0.0811 & 0.1855 & 0.0734 & 0.0492 & 0.0378 \\
0.2747 & 0.1563 & \mathbf{0.0805} & 0.0665 & 0.0750 & 0.0692 & 0.1335 & 0.0621 & 0.0452 & 0.0369 \\
0.3077 & \mathbf{0.1741} & 0.0852 & 0.0642 & 0.0699 & 0.0661 & 0.0946 & 0.0529 & 0.0465 & 0.0388 \\
\mathbf{0.3406} & 0.1751 & 0.0841 & 0.0576 & 0.0590 & 0.0573 & 0.0774 & 0.0538 & 0.0482 & 0.0469
\end{pmatrix}.$$

## 4.3 Hirsch

$$\begin{pmatrix}
0.0000 & 0.0000 & 0.0010 & 0.0051 & 0.0124 & 0.0375 & 0.0805 & 0.1457 & 0.2298 & \mathbf{0.4881} \\
0.0000 & 0.0004 & 0.0105 & 0.0325 & 0.0593 & 0.1145 & 0.1703 & 0.2169 & \mathbf{0.2205} & 0.1752 \\
0.0000 & 0.0048 & 0.0503 & 0.0930 & 0.1292 & 0.1585 & 0.1671 & \mathbf{0.1671} & 0.1498 & 0.0801 \\
0.0003 & 0.0277 & 0.1150 & 0.1541 & 0.1789 & 0.1658 & \mathbf{0.1294} & 0.1041 & 0.0811 & 0.0435 \\
0.0046 & 0.0945 & 0.1787 & 0.1747 & 0.1745 & \mathbf{0.1413} & 0.1011 & 0.0704 & 0.0459 & 0.0142 \\
0.0248 & 0.2102 & 0.2253 & 0.1682 & \mathbf{0.1499} & 0.0957 & 0.0605 & 0.0405 & 0.0190 & 0.0059 \\
0.0711 & 0.3251 & 0.2157 & \mathbf{0.1322} & 0.1118 & 0.0665 & 0.0356 & 0.0211 & 0.0181 & 0.0027 \\
0.2243 & 0.4026 & \mathbf{0.1656} & 0.0768 & 0.0592 & 0.0352 & 0.0195 & 0.0101 & 0.0038 & 0.0029 \\
0.5417 & \mathbf{0.3180} & 0.0761 & 0.0315 & 0.0196 & 0.0071 & 0.0030 & 0.0030 & 0.0000 & 0.0000 \\
\mathbf{0.8844} & 0.0981 & 0.0104 & 0.0039 & 0.0032 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000
\end{pmatrix}.$$

### 4.4 Mean

$$\begin{pmatrix}
0.0000 & 0.0000 & 0.0000 & 0.0005 & 0.0000 & 0.0039 & 0.0049 & 0.0253 & 0.2087 & \mathbf{0.7567} \\
0.0000 & 0.0000 & 0.0006 & 0.0081 & 0.0038 & 0.0337 & 0.0493 & 0.2089 & \mathbf{0.6062} & 0.0895 \\
0.0000 & 0.0000 & 0.0015 & 0.0157 & 0.0185 & 0.0747 & 0.2037 & \mathbf{0.4388} & 0.2434 & 0.0036 \\
0.0000 & 0.0000 & 0.0104 & 0.0224 & 0.0563 & 0.2039 & \mathbf{0.4086} & 0.2566 & 0.0414 & 0.0003 \\
0.0000 & 0.0005 & 0.0257 & 0.0656 & 0.1873 & \mathbf{0.3843} & 0.2648 & 0.0654 & 0.0063 & 0.0000 \\
0.0000 & 0.0026 & 0.0619 & 0.1915 & \mathbf{0.4041} & 0.2600 & 0.0697 & 0.0096 & 0.0005 & 0.0000 \\
0.0000 & 0.0322 & 0.2127 & \mathbf{0.4104} & 0.2706 & 0.0646 & 0.0086 & 0.0007 & 0.0000 & 0.0000 \\
0.0028 & 0.1826 & \mathbf{0.5034} & 0.2542 & 0.0505 & 0.0060 & 0.0004 & 0.0000 & 0.0000 & 0.0000 \\
0.1037 & \mathbf{0.6462} & 0.2212 & 0.0266 & 0.0022 & 0.0001 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
\mathbf{0.8044} & 0.1882 & 0.0071 & 0.0002 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000
\end{pmatrix}.$$

# References

[1] Lehmann, S., Lautrup, B. E., and Jackson, A. D. *Physical Review E* **68**, 026113 (2003).