Was der h-Index *wirklich* aussagt

Christian Krattenthaler

Diese Note legt dar, dass der sogenannte h-Index (Hirschs bibliometrischer Index) im Wesentlichen dieselbe Information wiedergibt wie die Gesamtanzahl von Zitationen von Publikationen einer Autorin oder eines Autors, also ein nutzloser bibliometrischer Index ist. Dies basiert auf einem faszinierenden Satz der Wahrscheinlichkeitstheorie, der hier ebenfalls erläutert wird.

1 Präambel

Im Artikel [6] schlug der theoretische Physiker Jorge Eduardo Hirsch 2005 einen neuen bibliometrischen Index vor, der seitdem als *h-Index* bekannt ist und in vielen Naturwissenschaften mit der größten Selbstverständlichkeit und Überzeugung als Maßzahl für den Einfluss und die Relevanz des Publikationsoutputs einer Autorin/eines Autors verwendet wird. Laut Hirsch ist sein Index leicht zu berechnen (das stimmt), er vermeidet (angeblich) die Probleme anderer bibliometrischer Indizes, und er erlaubt (angeblich) insbesondere den fundierten Vergleich auch von Autor(inn)en, die sehr verschiedene Gesamtanzahlen von Publikationen oder sehr verschiedene Gesamtanzahlen von Zitationen aufweisen.

Ich werde im Folgenden darlegen, dass mathematisch nachgewiesen werden kann, dass der h-Index nicht das hält, was er verspricht. Im Gegenteil, er ist im Wesentlichen äquivalent zu einer anderen (einfacheren) bibliometrischen Maßzahl, nämlich der Gesamtanzahl von Zitationen der Publikationen einer Autorin/eines Autors, die im Folgenden mit N_{Zit} bezeichnet werden wird.

Um die Kernaussage vorweg zu nehmen: Grob gesprochen ist es ein *mathematischer Satz*, dass der h-Index einer Autorin/eines Autors (ungefähr) gleich $0.54 \times \sqrt{N_{\text{Zit}}}$ ist! Siehe die genauere Formulierung des Satzes in Korollar 3 im Abschnitt 3.

Die Motivation für diese Note kam von einer (verspäteten) Lektüre des Positionspapiers der DMV zur Verwendung bibliometrischer Daten [3]. Darin ist dem h-Index ein eigener Absatz gewidmet. Es ist auch alles richtig, was dort gesagt wird. Aber: Es fehlt die Begründung für viele der angeführten Fehler des h-Index. Das ist besonders schade, da diese eben *mathematisch* ist! Und es ist auch deswegen schade, weil man – mit dieser Begründung – die Kritik sogar noch verstärken kann.[1]

2 Was ist der h-Index?

Hier ist die Definition des h-Index.

Definition 1. Wir stellen uns vor, ein(e) Autor(in) hat für ihre/seine Publikationen $N_1, N_2, \dots$ Zitationen bekommen, wobei $N_1 \geq N_2 \geq \cdots$. In anderen Worten, wir ordnen die Publikationen einer Autorin/eines Autors absteigend gemäß der Anzahl der Zitationen, die sie bekommen haben, sodass die i-te Publikation N_i Zitationen erhalten hat. Dann ist der h-Index das maximale k, sodass $k \leq N_k$.

Auch wenn das einen einfachen Algorithmus bedingt, wie man den h-Index bestimmen kann, kann man sich darunter ad hoc möglicherweise wenig vorstellen. Es gibt aber einen visuellen Zugang zum h-Index, der – jedenfalls mir – unmittelbar klar macht, worum es geht.

Zur Illustration wähle ich ein Beispiel: Nehmen wir an, dass wir von einer/einem Autor(in) reden, die/der 11 Artikel veröffentlicht hat. Weiters seien die Zitationszahlen der einzelnen Artikel $N_1 = 16$, $N_2 = 8$, $N_3 = 7$, $N_4 = 6$, $N_5 = 3$, $N_6 = 3$, $N_7 = 3$, $N_8 = 1$, $N_9 = 1$, $N_{10} = 1$, $N_{11} = 1$. Wir tragen nun diese Zitationszahlen in einem Balkendiagramm auf, wie wir das aus der Schule kennen, siehe Abbildung 1 links. Im rechten Teil der Abbildung haben wir die vertikalen Unterteilungsstriche „vergessen“. (Das gestrichelte Quadrat möge die/der Leser(in) zu diesem Zeitpunkt ignorieren.)

Wir „zwängen“ nun das größte Quadrat, das möglich ist, zwischen der oberen/rechten Begrenzung des Balkendiagramms und den Koordinatenachsen hinein. Siehe das strichlierte Quadrat im rechten Teil von Abbildung 1. Die Seitenlänge dieses Quadrates ist der gesuchte h-Index! (Im Beispiel von Abbildung 1 ist das also 4.)

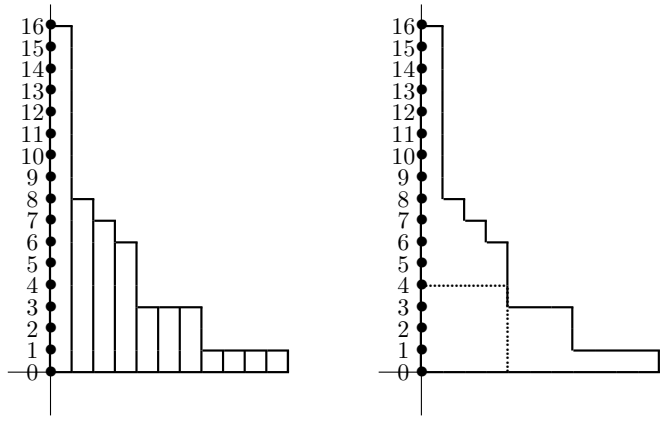

Abbildung 1. Balkendiagramm einer Verteilung von Zitationszahlen (links) und h-Index respektive Durfee-Quadrat (rechts)

DOI 10.1515/dmvm-2021-0050

Kombinatorikern (so wie mir) sind diese Konstruktionen wohlbekannt. Im Beispiel gilt ja

$$N_{\text{Zit}} = 50 = 16+8+7+6+3+3+3+1+1+1+1.$$

Eine solche Summendarstellung einer gegebenen Zahl (im konkreten Beispiel 50), wo die Summanden (schwach) absteigend angeordnet sind, nennen wir (Zahlen-)*Partition* von N_{Zit}. Die Diagrammdarstellung so wie im rechten Teil der Abbildung 1 nennt sich *Young-Diagramm* oder *Ferrers-Diagramm* der Partition. Schließlich wird das größte Quadrat, das man da wie im rechten Teil der Abbildung 1 hineinzwängen kann, *Durfee-Quadrat* genannt.[2]

3 Konzentration der Verteilung: Die „Formel" für den h-Index

Wir kommen nun zum (mathematischen) Kernstück dieser Note. Dieses ist ein Grenzfallsatz.

Wir geben uns ein N_{Zit} vor. Dann wählen wir aus allen möglichen Partitionen von N_{Zit} (= Verteilungen der Gesamtanzahl N_{Zit} an Zitationen auf die einzelnen Publikationen einer Autorin/eines Autors) eine zufällig, wobei wir alle solche Partitionen als gleich wahrscheinlich erachten. Die Frage, die wir uns stellen, ist: Wie sieht eine solche zufällig gewählte Partition, geeignet skaliert, aus, wenn N_{Zit} groß ist?

Die Frage mag auf den ersten Blick unsinnig erscheinen, ist aber eine Standardfrage in der Wahrscheinlichkeitstheorie. Aber es stimmt, es ist von vorneherein nicht klar, ob es darauf eine sinnvolle Antwort gibt.

Nichtsdestotrotz, wir kennen alle ein Beispiel einer solchen Fragestellung, wo es eine sinnvolle Antwort gibt. Wenn man Irrfahrten auf den ganzen Zahlen, die gemäß vorgegebenen Wahrscheinlichkeiten in jedem Zeitschritt entweder zur nächsthöheren oder zur nächstkleineren ganzen Zahl springen und N solche Schritte machen, mit $N^{1/2}$ skaliert, dann erhält man für $N \to \infty$ eine (eindimensionale) *Brownsche Bewegung*. Dabei kommt es auf die Details des diskreten Prozesses (die Irrfahrten) gar nicht so sehr an, im Grenzfall erhält man universell eine Brownsche Bewegung. Diese ist selbst ein Zufallsprozess.[3]

Manchmal jedoch ist der Grenzfallprozess *deterministisch*. Man spricht dann von *Grenzfallformen* (*limit shapes*). Und dieses faszinierende Phänomen liegt bei den zufällig gewählten Partitionen vor.[4]

Der folgende Satz macht die vorangegangenen Bemerkungen präzise. Grob gesprochen besagt er, dass Young-Diagramme von Partitionen von N_{Zit}, wenn sie in x- und y-Richtung um $N_{\text{Zit}}^{1/2}$ skaliert werden (was bedeuten soll, dass sowohl x- als auch y-Koordinate aller Punkte durch $N_{\text{Zit}}^{1/2}$ dividiert werden), mit hoher Wahrscheinlichkeit von der Kurve in (3.1) praktisch ununterscheidbar sind.

Satz 2. *Sei γ die Kurve*

$$\gamma = \left\{(x,y) : x,y > 0 \quad \textit{und} \quad e^{-\pi x/\sqrt{6}} + e^{-\pi y/\sqrt{6}} = 1\right\}, \qquad (3.1)$$

und sei $\varepsilon > 0$ vorgegeben. Dann strebt die Wahrscheinlichkeit, dass die Treppenfunktion, die durch das Young-Diagramm einer zufällig gewählten Partition von N (bezüglich Gleichverteilung), skaliert um $N^{1/2}$ in x- und y-Richtung, in einer ε-Umgebung von γ bleibt, gegen 1 für $N \to \infty$.

Abbildung 2 illustriert diesen Satz. Die Kurve γ (blau) ist zusammen mit der skalierten Zitationsverteilung/Partition (gelb) aus Abbildung 1 (sprich: x- und y-Koordinaten aller Punkte wurden durch $50^{1/2}$ dividiert) dargestellt. An dieser Stelle sei verraten, dass ich die Partition in Abbildung 1 mit der Eingabe `RandomPartition[50]` in *Mathematica* (unter Zuhilfenahme des `Combinatorics`-Pakets) erzeugt hatte. Man kann sehen, dass sich die Treppenfunktion relativ eng an γ anschmiegt. Der Satz besagt, dass das kein Zufall ist.

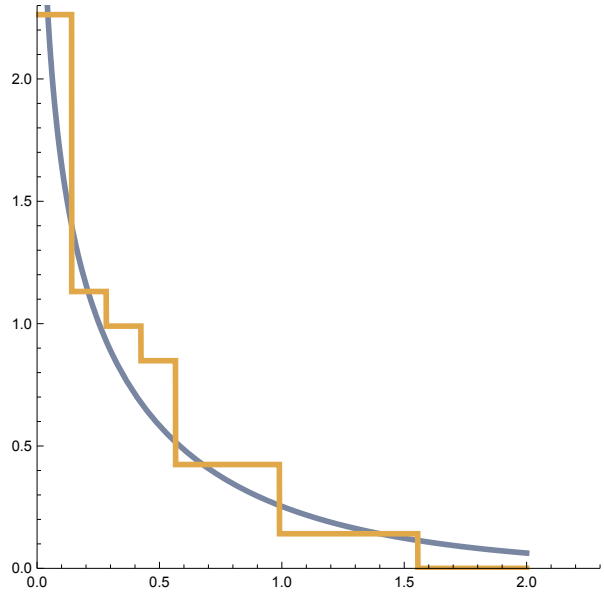

Abbildung 2. Die Kurve γ aus Satz 2 (blau) und die skalierte Partition aus Abbildung 1 (gelb)

Ich werde weiter unten etwas über die Geschichte des Satzes sagen. Zuvor sollten wir aber zum h-Index zurückkehren. Denn wenn es so ist, dass eine skalierte zufällige Partition nahe an der Kurve γ ist, dann muss auch der h-Index der Partition nahe dem „h-Index" der Kurve γ (deskaliert) sein. Klarerweise erhält man den „h-Index" von γ, indem man $x = y$ in (3.1) setzt: Man erhält $x = y = 6^{1/2}\log(2)/\pi$.

Korollar 3. *Sei $\varepsilon > 0$ vorgegeben. Dann strebt die Wahrscheinlichkeit, dass der h-Index einer zufällig gewählten Partition von N (bezüglich Gleichverteilung) um weniger als ε Prozent von*

$$\frac{6^{1/2}\log(2)}{\pi}\sqrt{N} \qquad (3.2)$$

abweicht, gegen 1 für $N \to \infty$. Hier ist $\frac{6^{1/2}\log(2)}{\pi} = 0.540445\ldots$.

Die (naheliegenden) Folgerungen aus dem Korollar werden im nächsten Abschnitt besprochen.

Nun also zur Geschichte von Satz 2. Wie es in [2, Sec. 12.1] so schön heißt: "It is difficult to credit limit shape results like $\langle\cdots\rangle$ precisely ...". Die Schwierigkeit besteht einerseits darin, ein Resultat zu finden, das so stark wie Satz 2

ist, und andererseits dann auch noch mit einem Beweis einhergeht, der heutigen Ansprüchen an Rigorosität standhält. (Autoren, die sich mit solchen Fragen beschäftigen, haben oft einen Physik-Hintergrund.) Wie in [2, Sec. 12.1] ausgeführt, geht ein solches Resultat auf Temperley [8] zurück, allerdings bloß auf heuristischen Argumenten basierend. Später kommen Vershik und Kerov (siehe [9] und die darin angegebenen Referenzen) in ihrer Arbeit ebenfalls auf Grenzfallformen von Partitionen zu sprechen, scheinen aber schwächere Resultate nicht wirklich rigoros herzuleiten. Möglicherweise ist es Fristedt [5, Theorem 2.9], der der erste ist, der ein Resultat rigoros beweist, aus dem Satz 2 leicht folgt. Eine Quelle, wo Satz 2 genau so wie oben formuliert ist (mit einer stärkeren Aussage über die Konvergenzgeschwindigkeit) ist [7, Theorem 1].

4 Schlussfolgerungen

Profan gesprochen sagt Korollar 3, dass *der h-Index genau dieselbe Information enthält* (ja sicher, mit Wahrscheinlichkeit nahe bei 1) *wie die Gesamtanzahl der Zitationen einer Autorin/eines Autors*. (Potentielle Einwände gegen diese Aussage werden im folgenden Abschnitt besprochen und – argumentativ – vom Tisch gewischt.) Er tut also genau das nicht, was Hirsch von ihm behauptet, nämlich dass er es erlauben würde, Vergleiche zwischen Autor(inn)en anzustellen, die verschiedene Anzahlen von Zitationen haben, in verschiedenen Karrierestadien wären, usw. (Es kommt mir so vor wie in der Kryptographie, wenn man etwas Raffiniertes austüftelt, um immun gegen einen bestimmten Entschlüsselungsangriff zu sein, und dafür in die Falle eines anderen tappt, in diesem Fall die Konzentration der Verteilung.)

Es ist also unsinnig, h-Indizes verschiedener Forscher(innen) miteinander zu vergleichen. Wenn schon, dann muss man einen h-Index einer Forscherin/eines Forschers mit dem gemäß (3.2) „erwarteten“ Wert vergleichen. Ist der h-Index in etwa genau gleich diesem Wert, so besagt er überhaupt nichts. Das war ja erwartet. Einzig und allein wenn der h-Index wesentlich davon abweicht, dann *könnte* das irgendetwas besagen. Aber was?

Bei etablierteren Forscher(inne)n kann man häufig beobachten, dass der h-Index doch wesentlich kleiner als die Formel (3.2) ist. Dafür gibt es aber eine sehr einfache Erklärung: Solche Autor(inn)en haben meist ein, zwei, drei extrem viel zitierte Publikationen (meist Bücher oder Überblicksartikel). Nimmt man diese dann aus der Zählung heraus, dann „stimmt“ alles wieder (sprich: der „reduzierte“ h-Index liegt nahe bei der Formel).

Auf der anderen Seite, wenn der h-Index wesentlich größer als der „erwartete“ Wert in (3.2) ist, dann ist es wohl so, dass diese(r) Forscher(in) wenig tatsächlich viel zitierte Publikationen aufweisen kann und die anderen mehr oder weniger gleich (mäßig?) zitiert wurden.

Ist es also so, dass es eher positiv zu bewerten ist, wenn der h-Index wesentlich *unter* dem Wert der Formel (3.2) liegt?

Wir merken schon, das Niveau der Diskussion gleitet zunehmend ab. Statt alle möglichen Dinge in den h-Index hineinzugeheimnissen, ist es wohl sinnvoller, sich das gesamte Publikationsprofil einer Autorin/eines Autors anzusehen. (Und noch besser ist es, sich tatsächlich den *Inhalt und Gehalt* der Publikationen anzusehen … Aber das ist wohl eine andere Diskussion.)

5 Einwände

Es gibt zwei potentielle Einwände gegen die im vorangegangenen Abschnitt vorgetragenen Schlussfolgerungen:

1. Korollar 3 ist ein asymptotisches Resultat. In der Praxis der Berechnung von h-Indizes von Autor(inn)en haben wir es aber mit „sehr endlichen“ Zahlen N_{Zit} zu tun.
2. Ist denn die Annahme der Gleichverteilung aller Partitionen (= Zitationsverteilungen) realitätsnahe?

Zum ersten Einwand: Es ist eine Tatsache, dass die Konzentration der Verteilung um den erwarteten Wert, der in der Formel (3.2) angegeben ist, sehr stark ist, auch schon für kleine N_{Zit}. Somit ist dieser Einwand gegenstandslos. Hier sind zwei konkrete Beispiele: Für $N_{\mathrm{Zit}} = 50$ gibt es 204 226 Partitionen, und die „Formel“ (3.2) liefert 3.82 für den „erwarteten“ h-Index. Eine nicht schwierige Rechnung[5] zeigt, dass 77 % der Partitionen von 50 einen h-Index von 3 oder 4 haben, und dass 97 % einen h-Index 3, 4 oder 5 haben. Wenn wir andererseits $N_{\mathrm{Zit}} = 1000$ annehmen, dann gibt es etwa 24×10^{30} Partitionen von 1000, und der „erwartete“ h-Index aus (3.2) ist 17.1. Hier zeigt sich, dass 88 % der Partitionen von 1000 einen h-Index zwischen 15 und 19 haben, und dass 97 % einen h-Index zwischen 15 und 20 haben.

Nun zum zweiten Einwand. Ist das Modell der Gleichverteilung ein valides Modell? Zu diesem Punkt kann man selbstverständlich nur empirisch, nicht mathematisch, argumentieren. Ich behaupte, dass das Gleichverteilungsmodell jedenfalls für die Mathematik sehr plausibel ist. Eine der wesentlichen Charakteristika mathematischer Publikationspraxis ist, dass auch ältere (und alte) Publikationen weiterhin zitiert werden, da sie ihre Gültigkeit nicht verlieren, und auch deswegen, da Mathematiker(innen) den Ehrgeiz haben (jedenfalls haben sollten …), die Originalquelle eines Resultats zu zitieren (so es sich nicht um Basiswissen oder ein „Folklore“-Resultat handelt). Der Effekt, der sich daraus für die Zitationszahlen von Publikationen einer Autorin/eines Autors ablesen lässt, ist, dass es darunter im Normalfall einige wenige geben wird, die Aufsehen erregt haben, und die weiterhin Zitationen bekommen werden, und dass es viele geben wird, die mäßig bis gar nicht rezipiert werden, und dadurch wenige (bis gar keine …) Zitationen bekommen werden, und auch über die Zeit nicht so viele dazukommen werden. Das implizierte Profil des Balkendiagramms einer solchen Zitationsverteilung entspricht genau dem Profil einer zufälligen Partition, das in Satz 2 präzise gefasst wird. Die Konfidenz wird praktisch zur Gewissheit, wenn man sich Beispiele ansieht, siehe Abschnitt 6.

Ich bin überzeugt, dass das Gleichverteilungsmodell auch für einige andere Naturwissenschaften äußerst tauglich ist (etwa Theoretische Physik, aber nicht nur). Wegen

seiner Konzentrationseigenschaft ist es sehr robust gegen Verzerrungen, die nicht zu groß sind. Sollte es aber auf Grund der Gepflogenheiten einer Disziplin so sein, dass Publikationen „notwendigerweise“ nach einer gewissen (kurzen) Zeit obsolet werden, da sie etwa durch neuere (technische) Entwicklungen zwangsläufig „überholt“ werden, dann ist das Gleichverteilungsmodell wohl kein taugliches Modell mehr. Man wird es dann mit Zitationsverteilungen zu tun haben, die ein wenig „verformt“, „dünner“ sind, sodass es viel mehr mäßig oft zitierte Publikationen gibt, aber nach wie vor einige wenige besonders oft zitierte. Man müsste sich eine solche Disziplin näher ansehen, um ein taugliches Partitionsmodell dafür zu entwickeln; es werden dann eben nicht mehr alle Partitionen gleich wahrscheinlich sein. Es ist aber nicht so, dass solche Szenarien nicht auch untersucht worden wären, siehe [2] und die dort angegebenen Referenzen. Es ist nicht weiter überraschend, dass es auch für diese Szenarien praktisch immer Grenzfallsätze ähnlichen Charakters wie Satz 2 gibt. Dann ist es aber auch wieder so, dass es einen „erwarteten“ h-Index der Form $c\sqrt{N_{\text{Zit}}}$ gibt, mit starker Konzentration, nur dass die Konstante c nicht die in (3.2) ist.

Wo die Gleichverteilungsannahme sicher deplatziert ist, ist bei Autor(inn)en, die schon lange nicht mehr publizieren. Da ist es ja dann so, dass die vorhandenen Publikationen weiterhin Zitationen erhalten werden, aber keine neuen Publikationen (mit am Anfang sehr niedrigen Zitationszahlen – am Anfang eben 0) dazu kommen. Das verträgt sich nicht mit dem Gleichverteilungsprofil aus Satz 2.

6 Einige Beispiele zur Illustration

Ich beginne mit mir selbst. Wie sieht mein h-Index aus? Zuerst: Ich habe den Punkt bisher nicht berührt, da er für das Fazit dieser Note unerheblich ist: Verwendet man verschiedene Datenbanken (etwa Web of Science, Google-Scholar, Scopus, MathSciNet, Zentralblatt, ...), dann bekommt man jedes Mal ganz verschiedene Zitierungszahlen, da jede Datenbank Dinge auf verschiedene Art erhebt (und auch tatsächlich verschiedene Dinge erhebt ...), und damit muss auch der h-Index jedes Mal ein anderer sein. Ich verwende hier jedenfalls MathSciNet. Dieses weist für mich (am 20. Juli 2021) $N_{\text{Zit}} = 1990$ aus, und der h-Index stellt sich als 21 heraus. Mit $N_{\text{Zit}} = 1990$ erhält man in der „Formel“ (3.2) den Wert 24.09. Ich würde sagen: Passt ziemlich gut. Aber es ist eigentlich noch besser. Sieht man sich meine beiden meistzitierten Publikationen an, dann stellt man fest, dass es sich dabei um Übersichtsartikel handelt, keine richtigen Forschungsartikel. Klar werden die – sozusagen – „überproportional“ zitiert. Wir sollten diese also aus der „Zählung“ herausnehmen. Dann bleiben 1600 Zitate übrig und der „reduzierte“ h-Index ist 20. Setzt man nun $N_{\text{Zit}} = 1600$ in (3.2) ein, dann erhält man 21.60.

Ich habe mir weiters erlaubt, N_{Zit} und h-Index der Präsidentin und des Vizepräsidenten der DMV zu erheben (wiederum unter Verwendung von MathSciNet). Ich hoffe, sie verzeihen mir das. Jedenfalls ist für Ilka Agricola $N_{\text{Zit}} = 414$, und ihr h-Index ist 12. Setzt man $N_{\text{Zit}} = 414$ in (3.2) ein, erhält man 10.99. Auch hier müsste man eigentlich den meistzitierten Artikel herausnehmen, da er ein Übersichtsartikel ist. Die reduzierte Zitationsanzahl ist dann $N_{\text{Zit}} = 342$, der reduzierte h-Index ist 11, und die „Formel“ (3.2) ergibt nun 9.99. Für Joachim Escher weist MathSciNet $N_{\text{Zit}} = 5900$ und einen h-Index von 38 aus. Mit $N_{\text{Zit}} = 5900$ erhält man in (3.2) den Wert 41.48. Auch das ist relativ nahe am tatsächlichen Wert des h-Index. Hier ist es nicht so, dass der/die meistzitierte(n) Artikel Bücher oder Übersichtsartikel wären. Hingegen ist der meistzitierte Artikel ein offensichtlich besonders fundamentaler Artikel über Wellenbrechung, der aus diesem Grund besonders zahlreich zitiert wird. Nimmt man diesen aus der „Wertung“, dann ergibt sich für die reduzierte Zitationenanzahl $N_{\text{Zit}} = 5113$, für den reduzierten h-Index 37, und die „Formel“ (3.2) produziert 38.61.

Ich weise darauf hin, dass der Artikel [10] zahlreiche weitere Daten und Beispiele enthält, insbesondere eine Auswertung und Vergleich mit (3.2) der h-Indizes von Abel-Preis-Rezipienten, von Mitgliedern der National Academy of Sciences der USA, und von Associate Professors dreier Mathematics Departments von amerikanischen Forschungsuniversitäten. Alle diese Daten bestätigen die in Abschnitt 4 präsentierten Schlussfolgerungen.

Ich lade Sie ein, Ihren h-Index zu „überprüfen“ und mit der „Formel“ (3.2) zu vergleichen!

7 Abschließende Bemerkungen

Die „Euphorie“ um die Erfindung des h-Index inspirierte die Erfindung zahlreicher weiterer solcher Indizes, ebenso mit behaupteter Aussagekraft über Einfluss und Relevanz der Publikationen von Autor(inn)en. Exemplarisch sei darunter der *g-Index* von Leo Egghe [4] herausgegriffen, der – laut seinem Erfinder – dem h-Index überlegen wäre. Mit der Notation und den Annahmen von Definition 1 ist der g-Index einer Autorin/eines Autors per Definition das maximale k, sodass $N_1 + N_2 + \cdots + N_k$ (die Anzahl der Zitationen der k meistzitierten Artikel der Autorin/des Autors) mindestens so groß wie k^2 ist. So wie alle anderen Indizes, die versuchen aus der Zitationsverteilung irgendetwas herauszulesen, scheitert auch er daran, dass die (skalierte) Zitationsverteilung wegen Satz 2 – im Wesentlichen – „deterministisch“ ist, und somit auch der entsprechende Index. Im Fall des g-Index müssen wir also die Gleichung

$$\int_0^g \left(-\tfrac{\sqrt{6}}{\pi}\log\left(1-e^{-\pi x/\sqrt{6}}\right)\right)dx = g^2 \qquad (7.1)$$

lösen und erhalten dann (mit Wahrscheinlichkeit nahe bei 1) den g-Index in Abhängigkeit von der Gesamtanzahl N_{Zit} von Zitationen.

Korollar 4. *Sei $\varepsilon > 0$ vorgegeben. Dann strebt die Wahrscheinlichkeit, dass der g-Index einer zufällig gewählten Partition von N (bezüglich Gleichverteilung) um weniger als ε Prozent von*

$$g\sqrt{N} \qquad (7.2)$$

abweicht, wobei g die positive Lösung der Gleichung (7.1) ist, gegen 1 für $N \to \infty$. Hier ist $g = 0.88699\ldots$.

Um auch dies mit Daten zu „unterfüttern“: Mein g-Index ist 37, und die „Formel“ (7.2) liefert mit $N_{\mathrm{Zit}} = 1990$ den Wert 39.70. Der g-Index von Ilka Agricola ist 18, während die „Formel“ (7.2) für $N_{\mathrm{Zit}} = 414$ den Wert 18.10 ergibt. Schließlich ist Joachim Eschers g-Index 73, verglichen mit 68.36, das man mit $N_{\mathrm{Zit}} = 5900$ aus der „Formel“ (7.2) erhält.

Hirschs Artikel [6] ist durchaus sehr mathematisch abgefasst. In der Tat ist sich Hirsch einer Korrelation zwischen h-Index und Gesamtanzahl der Zitationen bewusst. In [6, Eq. (1)] nimmt er die Beziehung $h = c\sqrt{N_{\mathrm{Zit}}}$ an[6] und hypothetisiert, dass sich $1/c^2$ zwischen 3 und 5 bewegt. Er beruft sich dabei auf empirische Daten. An dieser Stelle (wir befinden uns auf der ersten Seite des Artikels) begeht er bereits den fundamentalen Fehler, der alles weitere im Artikel entwertet: Es kommt ihm nicht in den Sinn, dass es eine „feste“ Beziehung zwischen h-Index und N_{Zit} geben könnte (wie in Korollar 3 ausgedrückt), und dass die Schwankungsbreite in der Konstante c mit statistischen Schwankungen und anderen Effekten zu tun haben könnte. Es ist geradezu amüsant festzustellen, dass wir in dieser Bandbreite, $5^{-1/2} = 0.44\cdots < c < 3^{-1/2} = 0.57\ldots$, den „tatsächlichen“ Wert $c = \frac{6^{1/2}\log(2)}{\pi} = 0.54\ldots$ am oberen Ende finden, sodass statistische Schwankung zusammen mit dem in Abschnitt 4 beschriebenen „dämpfenden“ Effekt für etabliertere Forscher(innen) diese Bandbreite hinreichend erklärt. Da eine eventuelle Aussagekraft des h-Index darauf beruht, dass die Konstante c nicht „fest“ ist, ist alles Weitere im Artikel [6] (der großteils aus etwas naiven heuristischen Annahmen und Argumenten und daraus abgeleiteten Rechnungen besteht) Makulatur.

Ich denke, dass die in dieser Note dargelegten Tatsachen besser bekannt werden sollten. Wenn schon andere Wissenschaften an solchen „Hokus-Pokus“ glauben (ich habe einmal versucht, einen Geowissenschafter zu überzeugen, dass der h-Index Unsinn ist; ohne Erfolg), dann sollten wenigstens wir Mathematiker(innen) wissen – und verbreiten, dass der h-Index unter den verschiedenen bibliometrischen Indizes einer der dümmsten ist; in dem Sinn, dass er etwas verspricht, was er nicht hält, und tatsächlich im Wesentlichen zu einem viel einfacheren Index (Anzahl der Zitationen) äquivalent ist.

Anmerkungen

1. Eine Note mit ähnlichem Inhalt ist vor ein paar Jahren in [10] auf Englisch erschienen. Diese stützt sich jedoch auf ein schwächeres Resultat und ist deswegen zurückhaltender formuliert.
2. Der/dem interessierten Leser(in), die/der mehr über die (äußerst gehaltreiche) Theorie von (Zahlen-)Partitionen erfahren will, sei der Klassiker [1] empfohlen.
3. Es gibt in der Wahrscheinlichkeitstheorie zahlreiche weitere solche (universelle) Grenzfallprozesse, so wie etwa Aldous' „continuous random tree“, die Brownsche Karte (eine „Zufallsoberfläche“) oder die Brownsche Schlange von Marckert und Mokkadem.
4. Andere Beispiele von Grenzfallformen sind die Grenzfalloberflächen von sogenannten „Plane Partitions“ (zweidimensionale Verallgemeinerungen von Partitionen) und allgemeiner die Höhenfunktionen zufällig gewählter perfekter Matchings in periodischen bipartiten Graphen.
5. Man benutzt erzeugende Funktionen, siehe [10, Sec. 3].
6. Die Gleichung [1] in [6] lautet tatsächlich $N_{\mathrm{Zit}} = ah^2$, wo h den h-Index bezeichnet. Das übersetzt sich in $a = 1/c^2$. Für diesen Proportionalitätsfaktor a stellt Hirsch empirisch eine Bandbreite von 3 bis 5 fest.

Literatur

[1] G. E. Andrews, *The Theory of Partitions*, Encyclopedia of Math. and its Applications, vol. 2, Addison-Wesley, Reading, 1976.

[2] S. DeSalvo und I. Pak, *Limit shapes via bijections*, Combin. Probab. Comput. **28** (2019), 187–240.

[3] Deutsche Mathematiker-Vereinigung, *Positionspapier zur Verwendung bibliometrischer Daten*, Mitteilungen Deut. Math.-Ver. **27** (2019), 112–117.

[4] L. Egghe, *Theory and practise of the g-index*, Scientometrics **69** (2006), 131–152.

[5] B. Fristedt, *The structure of random partitions of large integers*, Trans. Amer. Math. Soc. **337** (1993), 703–735.

[6] J. E. Hirsch, *An index to quantify an individual's scientific research output*, Proc. Natl. Acad. Sci. USA **102** (2005), 16569–16572.

[7] F. Petrov, *Two elementary approaches to the limit shapes of Young diagrams*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov **370** (2009), Kraevye Zadachi Matematicheskoĭ Fiziki i Smezhnye Voprosy Teorii Funktsiĭ. **40**, 111–131, 221; englische Übersetzung in J. Math. Sci. (N.Y.) **166** (2010), 63–74.

[8] H. N. V. Temperley, *Statistical mechanics and the partition of numbers, II. The form of crystal surfaces*, Proc. Cambridge Philos. Soc. **48** (1952), 683–697.

[9] A. M. Vershik, *Statistical mechanics of combinatorial partitions, and their limit configurations*, Funktsional. Anal. i Prilozhen. **30** (1996), 19–39, 96; englische Übersetzung in Funct. Anal. Appl. **30** (1996), 90–105.

[10] A. Yong, *Critique of Hirsch's citation index: a combinatorial Fermi problem*, Notices Amer. Math. Soc. **61** (2014), 1040–1050.

Prof. Dr. Christian Krattenthaler, Fakultät für Mathematik,
Universität Wien, Oskar-Morgenstern-Platz 1, 1090 Wien, Österreich
christian.krattenthaler@univie.ac.at

Christian Krattenthaler studierte Mathematik an der Universität Wien und Klavier Konzertfach an der Hochschule für Musik und Darstellende Kunst in Wien. Nach Abschluss der Studien (Mathematik 1984, Klavier 1986) war er als Lektor an der Universität Wien tätig, später als Universitätsassistent. 2002 bis 2005 war er Professor an der Université „Claude Bernard“ Lyon 1, ehe er als Professor für Diskrete Mathematik an die Universität Wien zurückkehrte. Für seine wissenschaftlichen Leistungen wurde er 2007 mit dem Wittgenstein-Preis ausgezeichnet.