Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 170–177
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

On the *h*-index, the size of the Hirsch core and Jin's *A*-index

Quentin L. Burrell

Isle of Man International Business School, The Nunnery, Old Castletown Road, Douglas, Isle of Man IM2 1QB, via United Kingdom

Received 7 November 2006; received in revised form 4 January 2007; accepted 5 January 2007

Abstract

Hirsch's *h*-index seeks to give a single number that in some sense summarizes an author's research output and its impact. Essentially, the *h*-index seeks to identify the most productive core of an author's output in terms of most received citations. This most productive set we refer to as the Hirsch core, or *h*-core. Jin's *A*-index relates to the average impact, as measured by the average number of citations, of this "most productive" core. In this paper, we investigate both the total productivity of the Hirsch core – what we term the size of the *h*-core – and the *A*-index using a previously proposed stochastic model for the publication/citation process, emphasising the importance of the dynamic, or time-dependent, nature of these measures. We also look at the inter-relationships between these measures. Numerical investigations suggest that the *A*-index is a linear function of time and of the *h*-index, while the size of the Hirsch core has an approximate square-law relationship with time, and hence also with the *A*-index and the *h*-index.
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Hirsch *h*-index; Hirsch *h*-core; Jin *A*-index; Stochastic model; Informetric process

1. Introduction

Ever since Hirsch (2005) proposed the *h*-index, a single number to measure both an individual's research output and its impact, it has received much attention, both in the popular domain and in the academic literature; see Burrell (2007a) for some references. Jin (2006) has suggested a supplementary measure, termed the *A*-index – since it relates to an average – by Rousseau (2006). In this note, we investigate the *A*-index and what we term the Hirsch core using the model proposed by Burrell (1992, 2007a) for the publication–citation process. More recently, Egghe (2006) has proposed the *g*-index and Kosmulski (2006) the $h(2)$ index. Although, still seeking to identify a core of an author's published work based upon citation counts, these latter two are suggested as alternatives to the *h*-index and will be discussed elsewhere (Burrell, 2007b,c). All of these measures seek to provide simple summary statistics for an author's impact, in which case it is useful for the scientometrician to know how they might depend upon both the internal – author based – and external – environment based – influences.

2. The stochastic model

Here, we just recap the essentials of the model and refer the reader to Burrell (2007a) for full details. The basic idea is that an author publishes papers at certain times and that these papers subsequently attract citations following their publication, where both the publication and citation accumulation processes are random. We further assume that

E-mail address: q.burrell@ibs.ac.im.

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2007.01.003

some papers are more citable than others so that the citation rate varies between different publications. Much of this was originally described by Burrell (1992). The precise technical assumptions, without the mathematical details, are.

2.1. Assumptions

(1) From the start of his/her publishing career at time zero, an author publishes papers according to a Poisson process of rate θ, which gives the mean number of publications per unit time, called the *publication rate*.
(2) Any particular publication acquires citations according to a Poisson process of rate Λ, where Λ varies from paper to paper. Here, Λ denotes the mean number of citations to the paper per unit time following publication, called the *citation rate*.
(3) The citation rate Λ for this author varies over the set of his/her publications according to a gamma distribution of index $\nu \geq 1$ and scale parameter $\alpha > 0$.

See Burrell (2007a) for the precise details.

Remarks.

(a) Although, the citation rate depends on the two gamma parameters, α and ν, Burrell (2007a) found that his results were fairly robust to changes in the two parameters so long as the mean, i.e. the ratio $\mu = \nu/\alpha$ of the parameters, remains the same. Our numerical investigations for both the size of the h-core and the A-index are similarly fairly robust, i.e. the details change slightly but the general picture is the same. In all that follows we use, for purposes of illustration, citation rates of $\mu = 2$, 5 and 10, where the actual calculations are performed using $\alpha = 1$.
(b) As a referee has pointed out, one can argue over the robustness of the model assumptions – indeed, see Burrell (2007a) for some reservations – but at least they lead to a simple stochastic model that is analytically viable. In particular, the assumption of a fixed publication rate has been questioned by Burrell (2007b), based on Liang (2006) empirical data.

The basic result for the model is the following:

Theorem (Burrell, 2007a). Under the assumptions of the model, the distribution of X_T, the number of citations to a randomly chosen paper by time T, is given by

$$P(X_T = r) = \frac{\alpha}{(\nu - 1)T} B\left(\frac{T}{\alpha + T};\quad r + 1,\quad \nu - 1\right) \text{ for } r = 0, 1, 2, \ldots \tag{1}$$

where $B(x; a, b) = [\Gamma(a + b)/\Gamma(a)\Gamma(b)] \int_0^x y^{a-1}(1 - y)^{b-1}\, dy$ is the cumulative distribution function of a beta distribution (of the first kind) with parameters a and b.

3. Time-dependence of the size of the Hirsch core

According to the preprint of Hirsch (2005), the h-index for an author is that integer h such that h of his/her papers have at least h citations each, while the rest have fewer than h citations. Actually, this is not quite well-defined, see the print version of Hirsch (2005), Glänzel (2006) and Rousseau (2006), since there is ambiguity if there are several papers with the same number of citations at h. To get round this, let us introduce

Notation. Write $f(n; T)$ for the number of an author's papers receiving exactly n citations by time T, and $N(n; T)$ for the number of an author's papers that have received at least n citations by time T so that $N(n; T) = \sum_{r=n}^{\infty} f(r; T)$.

Definition 1. Hirsch's h-index at time T is, for any particular author, the integer $h(T)$ satisfying

$$h(T) = \max\{n : n \leq N(n; T)\}$$

Note that this is an empirical measure, requiring observation of the actual values of $N(n; T)$.

Rousseau (2006) has proposed the idea of the *Hirsch core* as the set consisting of the first $h(T)$ articles, in order of decreasing citations. In case there are more than one articles with $h(T)$ citations, he proposes listing them in anti-chronological order since this rewards the newer articles, which have achieved their $h(T)$ citations in a shorter time. Alternatively, one could argue that the Hirsch core should comprise all of the author's publications that have received at least $h(T)$ citations thus, perhaps, including some older papers. Note that his distinction, although possibly important in empirical studies, will not affect our theoretical development.

Definition 2. The *size* of the Hirsch core at time T is denoted by $C(T)$ and gives the total number of citations accumulated by those papers in the Hirsch core. Thus

$$C(T) = \sum_{n=h(T)}^{\infty} nf(n;T)$$

Note that this is again an empirical definition, requiring the observed numbers of citations $f(n;T)$ of those papers in the core.

Remarks. The h-index relates to the number of *publications* in the core, the size gives the total number of *citations* accumulated by the publications in the core. It has been pointed out by Hirsch (2005), Egghe (2006) and Rousseau (2006) that all one can say definitely is that this total is at least h^2, since there are h papers having at least h citations each. Thus, $C(T) \geq h(T)^2$. Within the limits of our model assumptions, we shall see that we are able to estimate the size, i.e. the (theoretical) number of citations to papers in the Hirsch core.

We calculate the theoretical, i.e. the expected value of, $C(T)$ by means of the following, which gets round the difficulty that the above definition involves an infinite sum:

Proposition. The expected total number of citations accumulated by those papers in the Hirsch core is, for our assumed model

$$E[C(T)] = \theta T \left(\frac{\nu T}{2\alpha} - \frac{\alpha}{(\nu-1)T} \sum_{n=0}^{h(T)-1} nB\left(\frac{T}{\alpha+T}; n+1, \nu-1 \right) \right) \tag{2}$$

Proof. See Appendix A.

Remarks.

(i) Although this expression may look unwieldy, its basic form is intuitively reasonable. An author producing on average θ publications per unit time over a period of length T will (on average) have produced a total of θT papers, each receiving (on average) ν/α citations per unit time. The average time for a paper to be available for citation is $T/2$ so that the total number of accumulated citations will be (on average) the product of these, namely $\theta\nu T^2/2\alpha$. This is the main term on the RHS of (2); the other is the deduction for total citations received by those papers not in the Hirsch core. (The precise mathematical justification of this intuitive argument is found in Appendix A.) Note that, already, this suggests that the size of the core could correlate, at least approximately, with the square of time.

(ii) The $h(T)$ in the above is now the theoretical h-index, determined as described in Burrell (2007a).

(iii) The actual calculation of the RHS of (2) is now straightforward, for any given set of parameter values, with any computer package allowing evaluation of the cumulative beta distribution.

(iv) At first sight, it might appear from (2) that $E[C(T)]$ is directly proportional to the publication rate θ. However, the range of the summation that appears on the RHS of (2) involves $h(T)$ and this in turn depends on θ. In fact, Burrell (2007a) conjectures that $h(T)$ is approximately linear in $\ln\theta$.

Although, it does not seem possible to give a straightforward analytical description of the time-dependence of the size of the h-core, the model does allow numerical investigation to at least suggest the form of the dependence. Purely for purposes of illustration, we have chosen to fix the publication parameter at $\theta = 5$, representing a "moderate" rate.

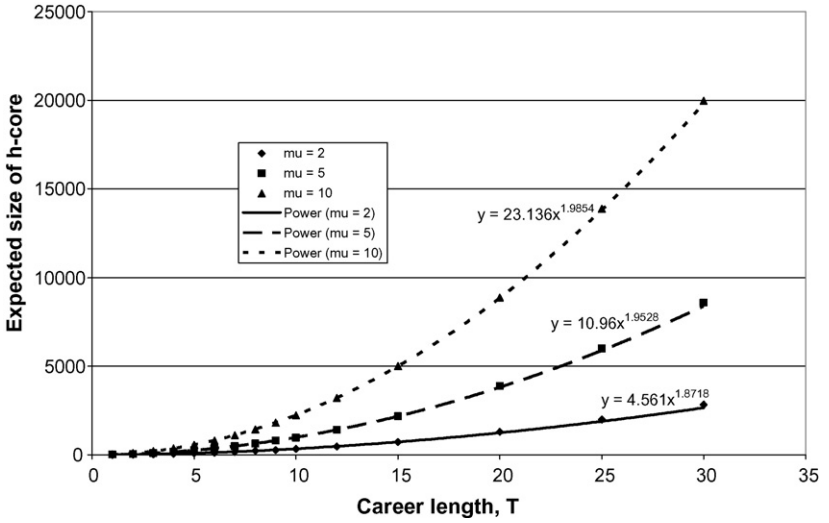

Fig. 1. Growth of h-core with time. Mean publication rate = 5.

Of course an average of five publications per year would be viewed as high in fields such as mathematics and possibly as low in other fields reporting a high level of collaborative work. For the citation rate, which can be thought of as the environmental or external factor, we use $\mu = 2$, 5 and 10 which correspond to low, medium and high rates of citation. The relationship between time and the size of the h-core for these scenarios is illustrated in Fig. 1, where the plotted points correspond to the time points or (current) career lengths $T = 1$–10, 12, 15, 20, 25 and 30.

Remarks. From Fig. 1, we can see quite clearly a very close power-law relationship between the size of the h-core and time. In all cases, the fit is extremely good, with $R^2 > 0.99$, and, in particular, note that the actual power is approximately (but always slightly less than) two. This last point should not be too surprising. We have already established that the expected *total* number of citations is proportional to T^2, it then seems reasonable that this should at least approximately be the case also when restricting attention to the core sources.

4. Time-dependence of Jin's A-index

According to Rousseau (2006), Jin's idea of an A-index (Jin, 2006) is that it should be the average number of citations received by those publications in an author's Hirsch core. In our notation, then, Jin's time-dependent A-index is as in:

Definition 3. Jin's A-index at time T is given by

$$A(T) = \frac{C(T)}{h(T)} = \frac{1}{h(T)} \sum_{n=h(T)}^{\infty} n f(n;T)$$

Again, this is an empirical measure, whereas we are investigating a theoretical model so we modify this to:

Definition 4. The theoretical A-index at time T is given by

$$A(T) = \frac{E[C(T)]}{h(T)} = \frac{1}{h(T)} \sum_{n=h(T)}^{\infty} nE[f(n;T)] = \frac{\theta T}{h(T)} \sum_{n=h(T)}^{\infty} nP(X_T = n)$$

where $h(T)$ is now the theoretical h-index, calculated as in Burrell (2007a).

As this again involves, the evaluation of an infinite sum, for purposes of calculation we use the following straightforward:

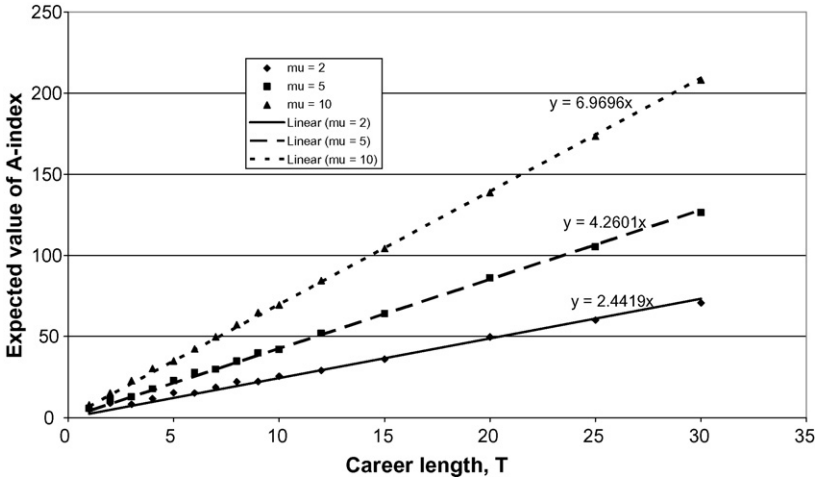

Fig. 2. Growth of A-index with time. Mean publication rate = 5.

Corollary to the Proposition. Under the assumptions of the model, the theoretical A-index is given by substituting the expression for $E[C(T)]$ from (2) into the basic definition of $A(T)$ above.

$$A(T) = \frac{\theta T}{h(T)} \left(\frac{\nu T}{2\alpha} - \frac{\alpha}{(\nu - 1)T} \sum_{n=0}^{h(T)-1} nB\left(\frac{T}{\alpha + T}; n + 1, \nu - 1 \right) \right)$$

Although, again, it does not seem possible to give a direct analytic expression of the dependence between the A-index and time, evaluation of the above expression for any given parameter values is routine given a computer package including the cumulative beta distribution function. Using the same combinations of parameter values as before, we illustrate the results of such calculations in Fig. 2.

Note that, as it seems intuitively reasonable that at time $T = 0$ the A-index should also be equal to zero, the fitted line – and the displayed regression equation – has the constraint that it should pass through the origin. In all cases, the approximate linearity is evident; indeed, we again have $R^2 > 0.99$. The reason for the linearity can be explained as follows. By its definition, Jin's A-index is given by $A(T) = C(T)/h(T)$. But we have just argued that $C(T)$ is (approximately) proportional to T^2 and Burrell (2007a) investigations strongly suggest that, with this model, $h(T)$ is approximately proportional to T. Taking these together, we should expect that A, as their ratio, is also approximately proportional to T.

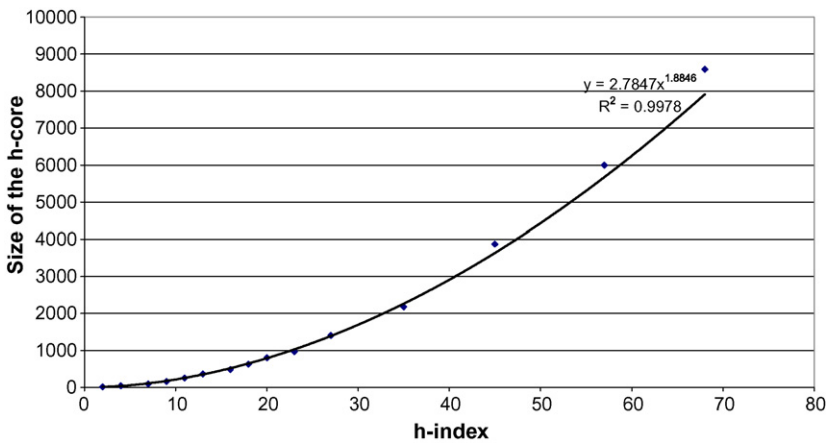

Fig. 3. Growth over time of the h-core with the h-index.

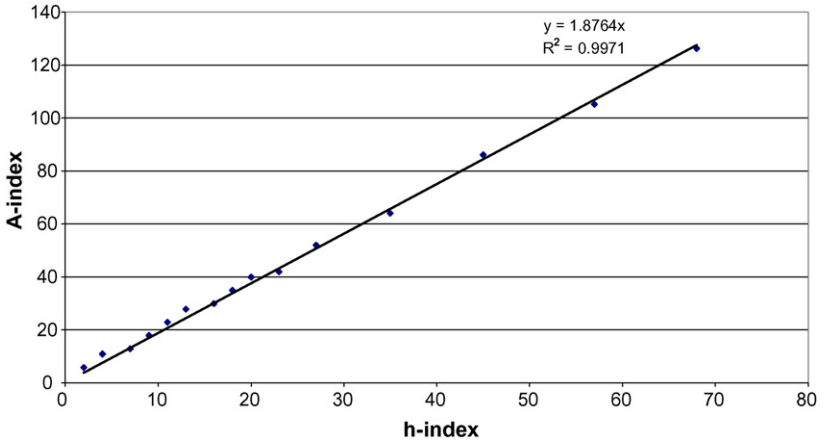

Fig. 4. Growth over time of A-index with h-index.

5. Relationships with the h-index

Given that $C(T)$ is approximately proportional to T^2 and that $h(T)$ is approximately proportional to T (Burrell, 2007a), it is not hard to see that, according to this model, the size of the core should be, at least approximately, proportional to h^2. Similarly, since $A(T)$ is approximately proportional to T, we would expect that the A-index is approximately proportional to the h-index. We illustrate these relationships in Figs. 3 and 4, which confirm these arguments. (Note that we have used a publication rate of $\theta = 5$ and citation rate $\mu = 5$ for these graphs. Other values produce similar results, but with different constants of proportionality.)

In Fig. 3, the reader might argue that, although the reported value of R^2 is very high, visual inspection suggests that the divergence increases with increasing h. In fact this results from our original (restricted and unequal) choice of values for the time parameter/career length T.

6. Concluding remarks

We have shown that Burrell's (1992, 2007a) model for the publication–citation process allows analytic and numerical investigation of the time-dependent behaviour of both the size of the h-core and the A-index (for an individual author). The main results are that the model suggests that the size of an author's Hirsch core should be approximately proportional to the square of the h-index (and of time), while for the A-index we should expect approximate direct proportionality to h and time. We await empirical studies to see to what extent these general findings agree with practice. Such studies would be analogous to Liang (2006) work on the time evolution of an author's h-index, but working forwards from the beginning of an author's active career, not backwards, see Burrell (2007b) for comments on this. Our own feeling is that all of these measures are indicative, supplementary scientometric measures. It will take much more empirical as well as theoretical research before anyone can claim a single definitive measure.

Appendix A

Proof of proposition. From its definition, we have

$$E[C(T)] = E\left[\sum_{n=h(T)}^{\infty} n f(n;T)\right] = \sum_{n=h(T)}^{\infty} n E[f(n;T)] = \theta T \sum_{n=h(T)}^{\infty} n P(X_T = n) \tag{A1}$$

$$\text{But} \quad \sum_{n=h(T)}^{\infty} n P(X_T = n) = E[X_T] - \sum_{n=1}^{h(T)-1} n P(X_T = n) \tag{A2}$$

So far as $P(X_T = n)$ is concerned, this follows from Eq. (1). Hence, all we need to complete the proof is an expression for the mean of X_T. This is given by the following:

Lemma.

$$E[X_T] = \frac{\nu T}{2\alpha}$$

(i) Standard proof

By definition

$$E[X_T] = \sum_{r=0}^{\infty} rP(X_T = r) = \sum_{r=1}^{\infty} rP(X_T = r) = \frac{\alpha}{(\nu - 1)T} \sum_{r=1}^{\infty} rB\left(\frac{T}{\alpha + T}; r + 1, \nu - 1\right)$$

making use of Eq. (1).

For the summation, making use of the integral representation of the cdf of the beta distribution, we have

$$\begin{aligned}\sum_{r=1}^{\infty} rB\left(\frac{T}{\alpha + T}; r + 1, \nu - 1\right) &= \sum_{r=1}^{\infty} r\left(\frac{\Gamma(r + \nu)}{\Gamma(r + 1)\Gamma(\nu - 1)} \int_0^{T/(\alpha+T)} y^r (1 - y)^{\nu-2}\, \mathrm{d}y\right) \\ &= \int_0^{T/(\alpha+T)} \left(\sum_{n=0}^{\infty} \frac{\Gamma(n + 1 + \nu)}{n!\Gamma(\nu - 1)} y^{n+1}(1 - y)^{\nu-2}\right) \mathrm{d}y, \quad \text{where} \quad n = r - 1 \\ &= \nu(\nu - 1) \int_0^{T/(\alpha+T)} \frac{y}{(1 - y)^3} \left(\sum_{n=0}^{\infty} \frac{\Gamma(n + (\nu + 1))}{n!\Gamma(\nu + 1)} y^n (1 - y)^{\nu+1}\right) \mathrm{d}y\end{aligned}$$

Now recognise the inner summation as the total sum of the probability mass function of NBD$(1 - y, \nu + 1)$ random variable and hence is equal to 1 for any y in [0,1].

It is then routine calculus to show that

$$\int_0^{T/(\alpha+T)} \frac{y}{(1 - y)^3}\, \mathrm{d}y = \frac{T^2}{2\alpha^2}$$

Substituting back, then, we find that

$$E[X_T] = \frac{\alpha}{(\nu - 1)T} \nu(\nu - 1) \frac{T^2}{2\alpha^2} = \frac{\nu T}{2\alpha} = \frac{\nu}{\alpha} \frac{T}{2}$$

(ii) Smart proof

$E[X_T] = E_t E[X_T|t]$, where t denotes the (random) time at which the typical paper was published. Now, given the publication time t, the paper has been in the public domain for a time $T - t$ gathering citations at expected rate ν/α per unit time so that $E[X_T|t] = (T - t)\ \nu/\alpha$

$$\text{Thus,} \quad E[X_T] = E_t E[X_T|t] = E[(T - t)\nu/\alpha] = (T - E[t])\nu/\alpha$$

But according to the model, publications appear as a Poisson process and hence publication times are uniformly distributed over [0,T], see for instance, Ross (1996, Chapter 2, p. 67) or Stirzaker (2005, Chapter 2, p. 75). In particular, then, $E[t] = T/2$ and the result follows.

It is now a straightforward matter to substitute this into (A2) and hence (A1). Making use of (1) we establish the result as in (2).

References

Burrell, Q. L. (1992). A simple model for linked informetric processes. *Information Processing and Management*, *28*, 637–645.

Burrell, Q. L. (2007a). Hirsch's h-index: A stochastic model. *Journal of Informetrics*, *1*(1), 16–25.

Burrell, Q. L. (2007b). Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics*, *73*(1), submitted for publication.

Burrell, Q. L. (2007c). Hirsch's h-index and Egghe's g-index. To be presented at the 11th ISSI Conference, 25–27 June, Madrid, Spain.

Egghe, L. (2006). Theory and practice of the *g*-index. *Scientometrics*, *69*(1), 131–152.

Glänzel, W. (2006). On the *h*-index—A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*(2), 315–321.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. (Also available in preprint form as ar Xiv: physica/0508113, accessible at http://xxx.arxiv.org/abs/physics/0508025).

Jin, B. H. (2006). *h*-Index: An evaluation indicator proposed by scientist. *Science Focus*, *1*(1), 8–9. (In Chinese)

Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original *h*-index. *ISSI Newsletter*, *2*(3), 4–6.

Liang, L. (2006). *h*-Index sequence and *h*-index matrix: Constructions and applications. *Scientometrics*, *69*(1), 153–159.

Ross, S. (1996). *Stochastic processes* (2nd ed.). New York: John Wiley.

Rousseau, R. (2006). New developments related to the Hirsch index. *Science Focus*, *1*(4), 23–25. (In Chinese).

Stirzaker, D. (2005). *Stochastic processes and models*. Oxford: Oxford University Press.