A random model of publication activity

Ágnes Backhausz

Eötvös Loránd University, Budapest

agnes@cs.elte.hu

YEP VIII, 18 March 2011, Eindhoven

Acknowledgement

This work is a joint research with Tamás F. Móri.

It is supported by the Hungarian National Foundation for Scientific Research, Grant No. K67961.

1. Definition of a random model of publication activity
2. Conditions
3. Continuous weight distribution
4. Discrete weight distribution

Definitions

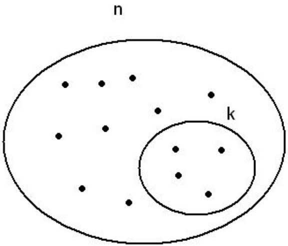
The model consists of objects (researchers), which have positive weight; it evolves in time, step by step.
It starts with a single researcher, which has random initial weight.
Every step consists of the following phases.

1. A new publication is born. Its authors are randomly chosen.
2. The new publication contributes to the weights of its authors. The bonuses are random variables.
3. A new researcher is added to the system with a random initial weight.

There are n researchers after $n-1$ steps; each of them has positive weight.

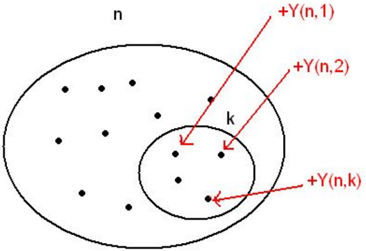
Step n, phase 1: new publication

The number of coauthors, k, is decided at random, independently of the past. The probability of selecting a given team is proportional to the total weight of the group.



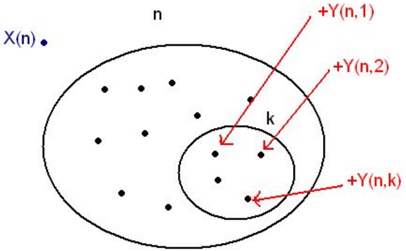
Step n, phase 2: bonuses for the authors

The new publication contributes to the weights of the authors. The bonuses are interchangeable random variables with a joint distribution only depending on the size of the group.



Step n, phase 3: new researcher

A new researcher is added to the system with a random initial weight. The initial weight is independent of every other random variable in the past.



Particular cases: recursive trees

If every paper is produced by a single author, connect him and the new researcher with an edge – random recursive tree

- Albert–Barabási tree: initial weight = paper bonus = const.
  Bollobás–Riordan–Spencer–Tusnády (RSA, 2001)

  preferential attachment model with scale free property

- generalized PORT: initial weight = const., paper bonus = 1
  Móri (Studia, 2002)

- uniform recursive tree: initial weight = const., no paper bonus
  Tapia–Myers (1967)

Object

How does the weight distribution behave as the number of steps tends to infinity?

It can be studied by monitoring the proportion of researchers with weight above level t ($t > 0$).

Notations

- $X_0, X_1, \ldots$ – i.i.d. initial weights of researchers
- ν_n – number of authors of the nth paper $(\nu_n \leq n)$
- $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,k}$ – authors' bonuses when $\nu_n = k$;
  Y_n – one dimensional marginal
- $Z_n = \sum_{i=1}^{\nu_n} Y_{n,i}$ – total weight of the nth paper
- $\xi_n(t)$ – number of authors with weight $> t$ after $n-1$ steps

Notations

- $X_0, X_1, \ldots$ – i.i.d. initial weights of researchers
- ν_n – number of authors of the nth paper ($\nu_n \leq n$)
- $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,k}$ – authors' bonuses when $\nu_n = k$;
  Y_n – one dimensional marginal
- $Z_n = \sum_{i=1}^{\nu_n} Y_{n,i}$ – total weight of the nth paper
- $\xi_n(t)$ – number of authors with weight $> t$ after $n-1$ steps

Examples

- $\nu_n \stackrel{d}{=} \nu \mid \nu \leq n$,
- $\nu_n \stackrel{d}{=} \min\{\nu, n\}$
- $Y_1, Y_2, \ldots$ i.i.d, and $Y_{n,1} = \cdots = Y_{n,\nu_n} = Y_n$
- $Z_1, Z_2, \ldots$ i.i.d, and $Y_{n,1} = \cdots = Y_{n,\nu_n} = Z_n/\nu_n$

Conditions

- the initial weights X_n, and the pairs $((Y_{n,1}, \dots, Y_{n,\nu_n}), \nu_n)$, $n = 1, 2, \dots$ are independent
- $\nu_n \to \nu$ in distribution, and $\mathbb{E}\nu_n^2 \to \mathbb{E}\nu^2 < \infty$
- the conditional distribution of $(Y_{n,1}, \dots, Y_{n,\nu_n})$, given $\nu_n = k$, does not depend on n
- hence Y_n and Z_n converge in distribution to some Y and Z
- X_n and Y_n are positive with positive probability
- X, Y, Z have finite moment generating functions

ν_n: number of authors; X_n: initial weight of the nth author;
$Y_{i,n}$: author bonus; Z_n: total weight of the nth paper

Continuous weight distribution

Suppose that the distributions of $Y_n \mid \nu_n = k\,(k = 1, 2, \ldots, n)$ and X are continuous. Introduce $F(t) = \mathbb{P}(Y > t)$, $H(t) = \mathbb{E}\left((\nu - 1)\mathbb{I}(Y > t)\right)$,

$$L\left(t, s\right) = \frac{sF\left(s\right) + t\left(1 - F\left(s\right)\right)}{\mathbb{E}X + \mathbb{E}Z} - H\left(s\right) \quad \left(0 \le s \le t\right).$$

Continuous weight distribution

Suppose that the distributions of $Y_n \mid \nu_n = k\,(k = 1, 2, \ldots, n)$ and X are continuous. Introduce $F(t) = \mathbb{P}(Y > t)$, $H(t) = \mathbb{E}\left((\nu - 1)\mathbb{I}(Y > t)\right)$,

$$L\,(t, s) = \frac{sF\,(s) + t\,(1 - F\,(s))}{\mathbb{E}X + \mathbb{E}Z} - H\,(s) \quad (0 \leq s \leq t)\,.$$

Theorem

$\frac{\xi_n(t)}{n} \to G(t)$ a.s., where $G(t)$ satisfies

$$G(t) = \left[\int_0^t G(t-s)\, d_s L\,(t, s) + H(t) + \mathbb{P}(X > t)\right]\left[\frac{t}{(\mathbb{E}X + \mathbb{E}Z)} + \mathbb{E}\nu\right]^{-1},$$

for $t > 0$, and $G(0) = 1$.

ν_n: number of authors; X: initial weight of the new author;
Y: author bonus; Z: total weight of a paper;
$\xi_n(t)$: number of authors with weight $> t$

A renewal-like integral equation

$$G(t) = \int_0^t G(t-s)\, w_{t,s} ds + r(t),$$

$$w_{t,s} = a(s) + \frac{b(s)}{t+d} + c(t,s), \quad 0 \le s \le t.$$

Here a is a probability density function, $0 < d$, and

$$\int_0^\infty \left(a(s) + |b(s)| + \int_0^s |c(s,u)|\, du \right) z^s ds < \infty$$

for some $z > 1$.
Under suitable conditions either $G(t) = 0$ for all t large enough, or $G(t)\, t^\gamma \to C$ holds as $t \to \infty$, where $0 < C < \infty$,

$$\gamma = -\frac{\int\limits_0^\infty b(s) ds}{\int\limits_0^\infty s a(s) ds}.$$

Asymptotics

Suppose that in the publication model all random variables concerning weigths are absolutely continuous, moment generating functions exist, and $\limsup_{t\to\infty} G(t) > 0$.

Corollary

Under suitable conditions, in the absolutely continuous publication model $G(t)t^{\gamma} \to C$, *where*

$$\gamma = \frac{\mathbb{E}X + \mathbb{E}Z}{\mathbb{E}Y}.$$

Limit distribution of the initial weights: X; author's bonus: Y; total weight of a paper: Z.

Discrete weight function

$X \geq 0$, $Y \geq 0$ integer valued
$\xi_n(k)$ – number of researchers with weight k

Discrete weight function

$X \geq 0$, $Y \geq 0$ integer valued
$\xi_n(k)$ – number of researchers with weight k

Theorem

$\frac{\xi_n(k)}{n} \to x_k$ *a.s., where* x_k *satisfies the recursion*

$$x_k = \frac{\sum_{i=1}^{k-1} x_{k-i} \left[\frac{(k-i)\mathbb{P}(Y=i)}{\mathbb{E}X+\mathbb{E}Z} + \mathbb{E}((\nu-1)\mathbb{I}(Y=i)) \right] + \mathbb{P}(X=k)}{\alpha k + \beta + 1}$$

where $\alpha = \frac{\mathbb{P}(Y>0)}{\mathbb{E}X+\mathbb{E}Z}$, $\beta = \mathbb{E}((\nu-1)\mathbb{I}(Y>0))$.

ν: number of authors; X: initial weight of the new author;
Y: author bonus; Z: total weight of a paper

A renewal-type recursion

This is a renewal-type equation of the form

$$x_k = \sum_{i=1}^{k-1} x_{k-i} w_{k,i} + r_k,$$

where $\lim_{k\to\infty} w_{k,i} = a_i$, and $(a_1, a_2, \dots)$ is a probability distribution.
[c.f. *Cooper–Frieze* (RSA, 2003), *Milne-Thompson* (1933)]

If $w_{k,i} = a_i$, then $x_k \to C$. Does that $o(1)$ make any difference?

A renewal-type recursion

This is a renewal-type equation of the form

$$x_k = \sum_{i=1}^{k-1} x_{k-i} w_{k,i} + r_k,$$

where $\lim_{k\to\infty} w_{k,i} = a_i$, and $(a_1, a_2, \dots)$ is a probability distribution.

[c.f. *Cooper–Frieze* (RSA, 2003), *Milne-Thompson* (1933)]

If $w_{k,i} = a_i$, then $x_k \to C$. Does that $o(1)$ make any difference?

Example

Let (a_i) be arbitrary, $x_k = 2 + \sin\big(\log(k+1)\big)$, and

$$w_{k,i} = a_i + \left(x_k - \sum_{i=1}^{k-1} x_{k-i} a_i\right)\left(\sum_{i=1}^{k-1} x_i\right)^{-1}.$$

Then $w_{k,i} = a_i + o(1)$, and $x_k = \sum_{i=1}^{k-1} x_{k-i} w_{k,i} + 2\delta_{k,1}$.

Asymptotics

Consider the recursion

$$x_k = \sum_{i=1}^{k-1} x_{k-i} w_{k,i} + r_k, \quad w_{k,i} = a_i + \frac{b_i}{k} + c_{k,i}, \quad 1 \le i \le k.$$

Suppose that $a_k \ge 0$ and the greatest common divisor of the positive terms is 1; $r_k \ge 0$ and not all of them is 0; there exists $z > 0$ such that

$$1 < \sum_{k=1}^{\infty} a_k z^k < \infty, \qquad \sum_{k=1}^{\infty} |b_k| z^k < \infty,$$

$$\sum_{k=1}^{\infty} \sum_{i=1}^{k-1} |c_{k,i}| z^i < \infty, \qquad \sum_{k=1}^{\infty} r_k z^k < \infty.$$

Asymptotics

Theorem

Under these conditions the following holds. Either $x_k = 0$ if k is large enough, or $x_k k^\gamma q^k \to C$ as $k \to \infty$, where $0 < C < \infty$, q is the positive solution of the equation $\sum_{k=1}^{\infty} a_k q^k = 1$, and

$$\gamma = -\left(\sum_{k=1}^{\infty} b_k q^k\right)\left(\sum_{k=1}^{\infty} k a_k q^k\right)^{-1}.$$

In the most important particular case (a_k) is a probability distribution, $a_k, b_k, c_{k,i}$ and r_k vanish exponentially fast, hence $q = 1$, and x_k decays at a polynomial rate.

Corollary

In the discrete publication model $x_k k^\gamma \to C$, where $\gamma = \frac{\mathbb{E}X + \mathbb{E}Z}{\mathbb{E}Y} + 1$.

Method of the stochastic part

Discrete parameter martingales are used to prove the following

Main lemma

$(\mathcal{F}_n)$ filtration, (ξ_n) nonnegative adapted, $\mathbb{E}\big((\xi_n - \xi_{n-1})^2 \bigm| \mathcal{F}_{n-1}\big)$ $= O\left(n^{1-\delta}\right)$, $\delta > 0$, (u_n), (v_n) *nonnegative predictable,* $u_n < n$.

(1) *Suppose* $\mathbb{E}(\xi_n \mid \mathcal{F}_{n-1}) \le \left(1 - \frac{u_n}{n}\right)\xi_{n-1} + v_n$,
where $u_n \to u$, $\limsup v_n \le v$, *and* $u, v > 0$. *Then*

$$\limsup_{n\to\infty} \frac{\xi_n}{n} \le \frac{v}{u+1} \quad a.s.$$

(2) *Suppose* $\mathbb{E}(\xi_n \mid \mathcal{F}_{n-1}) \ge \left(1 - \frac{u_n}{n}\right)\xi_{n-1} + v_n$,
where $u_n \to u$, $\liminf v_n \ge v$, *and* $u, v > 0$. *Then*

$$\liminf_{n\to\infty} \frac{\xi_n}{n} \ge \frac{v}{u+1} \quad a.s.$$

This is a stochastic counterpart of a lemma of *Chung and Lu* (2006).

Asymptotics of submartingales

Proposition

Let $(M_n, \mathcal{G}_n)$ *be a square integrable nonnegative submartingale, and*

$$A_n = EM_1 + \sum_{i=2}^{n} \left(E\left(M_i \middle| \mathcal{G}_{i-1} \right) - M_{i-1} \right),$$

$$B_n = \sum_{i=2}^{n} \mathrm{Var}\left(M_i \middle| \mathcal{G}_{i-1} \right).$$

If $B_n^{1/2} \log B_n = O(A_n)$, *then* $M_n/A_n \to 1$ *holds almost everywhere on the event* $\{A_n \to \infty\}$.

This is a consequence of Neveu [3], Propositions VII-2-3 and VII-2-4.

References

- Barabási, Albert-László and Albert, Réka. Emergence of scaling in random networks.*Science*, 286:509–512, 1999.
- Cooper, C. and Frieze, A. A general model of web graphs. *Random Structures Algorithms*, 22:311–335, 2003.
- Neveu, J. (1975), *Discrete-Parameter Martingales*, North-Holland, Amsterdam.
- Pittel, B. Note on the heights of random recursive trees and random *m*-ary search trees. *Random Structures & Algorithms*, 5:337–347, 1994.