Mechanisms for Generating Power-Law Size Distributions, Part 3

Last updated: 2021/10/06, 23:35:55 EDT

Principles of Complex Systems, Vols. 1 & 2
CSYS/MATH 300 and 303, 2021–2022| @pocsvox

Prof. Peter Sheridan Dodds | @peterdodds

Computational Story Lab | Vermont Complex Systems Center
Vermont Advanced Computing Core | University of Vermont

Outline

Aggregation:

- Random walks represent additive aggregation
- Mechanism: Random addition and subtraction
- Compare across realizations, no competition.
- Next: Random Additive/Copying Processes involving Competition.
- Widespread: Words, Cities, the Web, Wealth, Productivity (Lotka), Popularity (Books, People, ...)
- Competing mechanisms (trickiness)

Pre-Zipf's law observations of Zipf's law

- 1910s: Word frequency examined re Stenography (or shorthand or brachygraphy or tachygraphy), Jean-Baptiste Estoup [6].
- 1910s: Felix Auerbach pointed out the Zipfitude of city sizes in "Das Gesetz der Bevölkerungskonzentration" ("The Law of Population Concentration") [1].
- 1924: G. Udny Yule [15]:
  # Species per Genus (offers first theoretical mechanism)
- 1926: Lotka [9]:
  # Scientific papers per author (Lotka's law)

Theoretical Work of Yore:

- 1949: Zipf's "Human Behaviour and the Principle of Least-Effort" is published. [16]
- 1953: Mandelbrot [10]:
  Optimality argument for Zipf's law; focus on language.
- 1955: Herbert Simon [14, 16]:
  Zipf's law for word frequency, city size, income, publications, and species per genus.
- 1965/1976: Derek de Solla Price [4, 13]:
  Network of Scientific Citations.
- 1999: Barabasi and Albert [2]:
  The World Wide Web, networks-at-large.

Herbert Simon (1916–2001):

- Political scientist (and much more)
- Involved in Cognitive Psychology, Computer Science, Public Administration, Economics, Management, Sociology
- Coined 'bounded rationality' and 'satisficing'
- Nearly 1000 publications (see Google Scholar)
- An early leader in Artificial Intelligence, Information Processing, Decision-Making, Problem-Solving, Attention Economics, Organization Theory, Complex Systems, And Computer Simulation Of Scientific Discovery.
- 1978 Nobel Laureate in Economics (his Nobel bio is here).

Essential Extract of a Growth Model:

Random Competitive Replication (RCR):

1. Start with 1 elephant (or element) of a particular flavor at $t = 1$
2. At time $t = 2, 3, 4, \ldots$, add a new elephant in one of two ways:
   - With probability ρ, create a new elephant with a new flavor
     = Mutation/Innovation
   - With probability $1 - \rho$, randomly choose from all existing elephants, and make a copy.
     = Replication/Imitation
   - Elephants of the same flavor form a group

Random Competitive Replication:

Example: Words appearing in a language

- Consider words as they appear sequentially.
- With probability ρ, the next word has not previously appeared
  = Mutation/Innovation
- With probability $1 - \rho$, randomly choose one word from all words that have come before, and reuse this word
  = Replication/Imitation

Note: This is a terrible way to write a novel.

For example:

Some observations:

- Fundamental Rich-get-Richer story;
- Competition for replication between individual elephants is random;
- Competition for growth between groups of matching elephants is not random;
- Selection on groups is biased by size;
- Random selection sounds easy;
- Possible that no great knowledge of system needed (but more later ...).

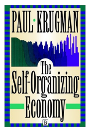
Your free set of tofu knives:

- Related to Pólya's Urn Model, a special case of problems involving urns and colored balls.
- Sampling with super-duper replacement and sneaky sneaking in of new colors.

Random Competitive Replication:

Some observations:

- Steady growth of system: +1 elephant per unit time.
- Steady growth of distinct flavors at rate ρ
- We can incorporate
  1. Elephant elimination
  2. Elephants moving between groups
  3. Variable innovation rate ρ
  4. Different selection based on group size (But mechanism for selection is not as simple...)



"The Self-Organizing Economy" by Paul Krugman (1996).[8]

Ch. 3: An Urban Mystery, p. 46

"...Simon showed—in a completely impenetrable exposition!—that the exponent of the power law distribution should be ..."[1, 2]

[1] Krugman's book was handed to the Deliverator by a certain Álvaro Cartea many years ago at the Santa Fe Institute Summer School.

[2] Let's use π for probability because π's not special, right guys?

Random Competitive Replication:

Definitions:

- k_i = size of a group i
- $N_{k,t}$ = # groups containing k elephants at time t.

Basic question: How does $N_{k,t}$ evolve with time?

First: $\sum_k kN_{k,t} = t =$ number of elephants at time t

Random Competitive Replication:

$P_k(t)$ = Probability of choosing an elephant that belongs to a group of size k:

- $N_{k,t}$ size k groups
- $\Rightarrow kN_{k,t}$ elephants in size k groups
- t elephants overall

$$P_k(t) = \frac{kN_{k,t}}{t}.$$

Random Competitive Replication:

$N_{k,t}$, the number of groups with k elephants, changes at time t if

1. An elephant belonging to a group with k elephants is replicated:
   $N_{k,t+1} = N_{k,t} - 1$
   Happens with probability $(1-\rho)kN_{k,t}/t$
2. An elephant belonging to a group with $k-1$ elephants is replicated:
   $N_{k,t+1} = N_{k,t} + 1$
   Happens with probability $(1-\rho)(k-1)N_{k-1,t}/t$

Random Competitive Replication:

Special case for $N_{1,t}$:

1. The new elephant is a new flavor:
   $N_{1,t+1} = N_{1,t} + 1$
   Happens with probability ρ
2. A unique elephant is replicated:
   $N_{1,t+1} = N_{1,t} - 1$
   Happens with probability $(1-\rho)N_{1,t}/t$

Random Competitive Replication:

Putting everything together:

For $k > 1$:

$$\langle N_{k,t+1} - N_{k,t} \rangle = (1-\rho)\left((+1)(k-1)\frac{N_{k-1,t}}{t} + (-1)k\frac{N_{k,t}}{t}\right)$$

For $k = 1$:

$$\langle N_{1,t+1} - N_{1,t} \rangle = (+1)\rho + (-1)(1-\rho)1 \cdot \frac{N_{1,t}}{t}$$

Random Competitive Replication:

Assume distribution stabilizes: $N_{k,t} = n_k t$
(Reasonable for t large)

- Drop expectations
- Numbers of elephants now fractional
- Okay over large time scales

- For later: the fraction of groups that have size k is n_k/ρ since
$$\frac{N_{k,t}}{\rho t} = \frac{n_k t}{\rho t} = \frac{n_k}{\rho}.$$

Random Competitive Replication:

Stochastic difference equation:

$$\langle N_{k,t+1} - N_{k,t}\rangle = (1-\rho)\left((k-1)\frac{N_{k-1,t}}{t} - k\frac{N_{k,t}}{t}\right)$$

becomes

$$n_k(t+1) - n_k t = (1-\rho)\left((k-1)\frac{n_{k-1}t}{t} - k\frac{n_k t}{t}\right)$$

$$n_k(\cancel{t}+1-\cancel{t}) = (1-\rho)\left((k-1)\frac{n_{k-1}\cancel{t}}{\cancel{t}} - k\frac{n_k \cancel{t}}{\cancel{t}}\right)$$

$$\Rightarrow n_k = (1-\rho)\left((k-1)n_{k-1} - kn_k\right)$$

$$\Rightarrow n_k\left(1+(1-\rho)k\right) = (1-\rho)(k-1)n_{k-1}$$

Random Competitive Replication:

We have a simple recursion:

$$\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1+(1-\rho)k}$$

- Interested in k large (the tail of the distribution)
- Can be solved exactly.
  Insert question from assignment 4
- For just the tail: Expand as a series of powers of $1/k$
  Insert question from assignment 4
  We (okay, you) find

$$n_k \propto k^{-\frac{(2-\rho)}{(1-\rho)}} = k^{-\gamma}$$

$$\boxed{\gamma = \frac{(2-\rho)}{(1-\rho)} = 1 + \frac{1}{(1-\rho)}}$$

- Micro-to-Macro story with ρ and γ measurable.

$$\boxed{\gamma = \frac{(2-\rho)}{(1-\rho)} = 1 + \frac{1}{(1-\rho)}}$$

- Observe $2 < \gamma < \infty$ for $0 < \rho < 1$.
- For $\rho \simeq 0$ (low innovation rate):

$$\gamma \simeq 2$$

- 'Wild' power-law size distribution of group sizes, bordering on 'infinite' mean.
- For $\rho \simeq 1$ (high innovation rate):

$$\gamma \simeq \infty$$

- All elephants have different flavors.
- Upshot: Tunable mechanism producing a family of universality classes.

- Recall Zipf's law: $s_r \sim r^{-\alpha}$
  (s_r = size of the rth largest group of elephants)
- We found $\alpha = 1/(\gamma - 1)$ so:

$$\boxed{\alpha = \frac{1}{\gamma-1} = \frac{1}{\cancel{1} + \frac{1}{(1-\rho)} - \cancel{1}} = 1-\rho.}$$

- $\gamma = 2$ corresponds to $\alpha = 1$
- We (roughly) see Zipfian exponent [16] of $\alpha = 1$ for many real systems: city sizes, word distributions, ...
- Corresponds to $\rho \to 0$, low innovation.
- Still, other quite different mechanisms are possible...
- Must look at the details to see if mechanism makes sense... more later.

What about small k?:

We had one other equation:

$$\langle N_{1,t+1} - N_{1,t}\rangle = \rho - (1-\rho)1\cdot\frac{N_{1,t}}{t}$$

- As before, set $N_{1,t} = n_1 t$ and drop expectations

$$n_1(t+1) - n_1 t = \rho - (1-\rho)1\cdot\frac{n_1 t}{t}$$

$$n_1 = \rho - (1-\rho)n_1$$

- Rearrange:

$$n_1 + (1-\rho)n_1 = \rho$$

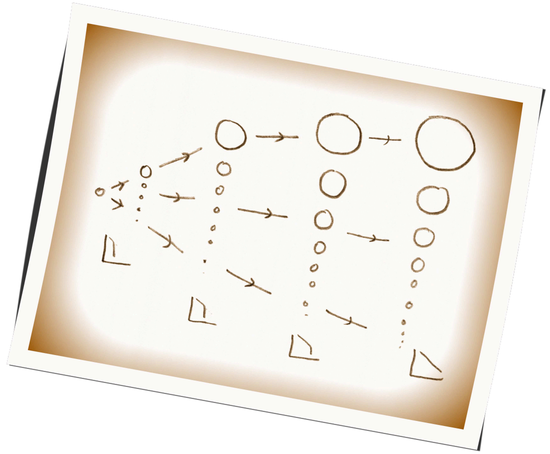
$$\boxed{n_1 = \frac{\rho}{2-\rho}}$$

So... $\quad N_{1,t} = n_1 t = \frac{\rho t}{2-\rho}$

- Recall number of distinct elephants = ρt.
- Fraction of distinct elephants that are unique (belong to groups of size 1):

$$\frac{1}{\rho t}N_{1,t} = \frac{1}{\cancel{\rho t}}\frac{\cancel{\rho t}}{2-\rho} = \frac{1}{2-\rho}$$

  (also = fraction of groups of size 1)
- For ρ small, fraction of unique elephants $\sim 1/2$
- Roughly observed for real distributions
- ρ increases, fraction increases
- Can show fraction of groups with two elephants $\sim 1/6$
- Model works well for large and small k #awesome



Words:

From Simon [14]:

Estimate ρ_{est} = # unique words/# all words

For Joyce's Ulysses: $\rho_{est} \simeq 0.115$

N_1 (real)	N_1 (est)	N_2 (real)	N_2 (est)
16,432	15,850	4,776	4,870

Evolution of catch phrases:

- Yule's paper (1924) [15]:
  "A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S."
- Simon's paper (1955) [14]:
  "On a class of skew distribution functions" (snore)

From Simon's introduction:

It is the purpose of this paper to analyse a class of distribution functions that appear in a wide range of empirical data—particularly data describing sociological, biological and economic phenomena.

Its appearance is so frequent, and the phenomena so diverse, that one is led to conjecture that if these phenomena have any property in common it can only be a similarity in the structure of the underlying probability mechanisms.

Evolution of catch phrases:

Derek de Solla Price:

- First to study network evolution with these kinds of models.
- Citation network of scientific papers
- Price's term: Cumulative Advantage
- Idea: papers receive new citations with probability proportional to their existing # of citations
- Directed network
- Two (surmountable) problems:
  1. New papers have no citations
  2. Selection mechanism is more complicated

Evolution of catch phrases:

Robert K. Merton: the Matthew Effect

- Studied careers of scientists and found credit flowed disproportionately to the already famous

  From the Gospel of Matthew:
  "For to every one that hath shall be given...
  (Wait! There's more....)
  but from him that hath not, that also which he seemeth to have shall be taken away.
  And cast the worthless servant into the outer darkness; there men will weep and gnash their teeth."
- (Hath = suggested unit of purchasing power.)
- Matilda effect: women's scientific achievements are often overlooked

Evolution of catch phrases:

Merton was a catchphrase machine:

1. Self-fulfilling prophecy
2. Role model
3. Unintended (or unanticipated) consequences
4. Focused interview → focus group
5. Obliteration by incorporation (includes above examples from Merton himself)

And just to be clear...

Merton's son, Robert C. Merton, won the Nobel Prize for Economics in 1997.

Evolution of catch phrases:

- Barabasi and Albert [2]—thinking about the Web
- Independent reinvention of a version of Simon and Price's theory for networks
- Another term: "Preferential Attachment"
- Considered undirected networks (not realistic but avoids 0 citation problem)
- Still have selection problem based on size (non-random)
- Solution: Randomly connect to a node (easy) ...
- ...and then randomly connect to the node's friends (also easy)
- "Scale-free networks" = food on the table for physicists

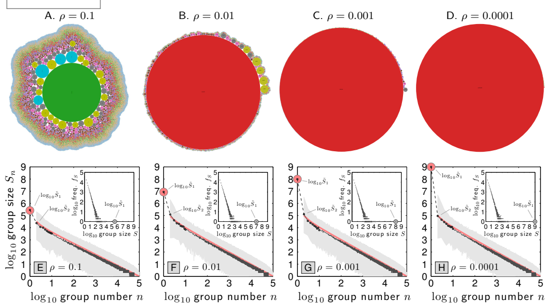
Another analytic approach: [5]

- Focus on how the nth arriving group typically grows.
- Analysis gives:

$$S_{n,t} \sim \begin{cases} \frac{1}{\Gamma(2-\rho)} \left[\frac{1}{t}\right]^{-(1-\rho)} & \text{for } n = 1, \\ \rho^{1-\rho} \left[\frac{n-1}{t}\right]^{-(1-\rho)} & \text{for } n \geq 2. \end{cases}$$

- First mover is a factor $1/\rho$ greater than expected.
- Because ρ is usually close to 0, the first element is truly an elephant in the room.
- Appears that this has been missed for 60 years ...

"Simon's fundamental rich-get-richer model entails a dominant first-mover advantage"
Dodds et al.,
Physical Review E, **95**, 052301, 2017. [5]



- See visualization at paper's online app-endices

Alternate analysis:

- Evolution of the nth arriving group's size:

$$\langle S_{n,t+1} - S_{n,t} \rangle = (1-\rho_t) \cdot \frac{S_{n,t}}{t} \cdot (+1).$$

- For $t \geq t_n^{\text{init}}$, fix $\rho_t = \rho$ and shift t to $t-1$:

$$S_{n,t} = \left[1 + \frac{(1-\rho)}{t-1}\right] S_{n,t-1}.$$

where $S_{n,t_n^{\text{init}}} = 1$.

Betafication ensues:

$$S_{n,t} = \left[1 + \frac{(1-\rho)}{t-1}\right] \left[1 + \frac{(1-\rho)}{t-2}\right] \cdots \left[1 + \frac{(1-\rho)}{t_n^{\text{init}}}\right] \cdot 1$$

$$= \left[\frac{t+1-\rho}{t-1}\right] \left[\frac{t-\rho}{t-2}\right] \cdots \left[\frac{t_n^{\text{init}}+1-\rho}{t_n^{\text{init}}}\right]$$

$$= \frac{\Gamma(t+1-\rho)\Gamma(t_n^{\text{init}})}{\Gamma(t_n^{\text{init}}+1-\rho)\Gamma(t)}$$

$$= \frac{\mathrm{B}(t_n^{\text{init}}, 1-\rho)}{\mathrm{B}(t, 1-\rho)}.$$

The first mover is really different:

- The issue is t_n^{init} in

$$S_{n,t} = \frac{\mathrm{B}(t_n^{\text{init}}, 1-\rho)}{\mathrm{B}(t, 1-\rho)}$$

- For $n \geq 2$ and $\rho \ll 1$, the nth group typically arrives at $t_n^{\text{init}} \simeq \left[\frac{n-1}{\rho}\right]$
- But $t_1^{\text{init}} = 1$ and the scaling is distinct in form.
- Simon missed the first mover by working on the size distribution.
- Contribution to $P_{k,t}$ of the first element vanishes as $t \to \infty$.
- Note: Does not apply to Barabási-Albert model.

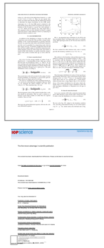
Variability:

- The probability that the nth arriving group, if of size $S_{n,t} = k$ at time t, first replicates at time $t+\tau$:

$$\mathbf{Pr}(S_{n,t+\tau} = k+1 \,|\, S_{n,t+i} = k \text{ for } i = 0, \dots, \tau - 1)$$
$$= \prod_{i=0}^{\tau-1} \left[1 - (1-\rho)\frac{k}{t+i}\right] \cdot (1-\rho)\frac{k}{t+\tau}$$
$$= k\frac{B(\tau, t)}{B(\tau, t-(1-\rho))}\frac{1-\rho}{t+\tau} \propto \frac{\tau^{-(1-\rho)k}}{t+\tau} \sim \tau^{-(2-\rho)k}.$$

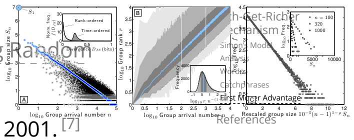
- Upshot: nth arriving group starting at size 1 will on average wait for an infinite time to replicate.

Related papers:



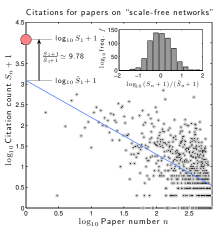
"Organization of Growing Random Networks"
Krapivsky and Redner,
Phys. Rev. E, **63**, 066123, 2001. [7]

Arrival variability:



- Any one simulation shows a high amount of disorder.
- Two orders of magnitude variation in possible rank.
- Rank ordering creates a smooth Zipf distribution.
- Size distribution for the nth arriving group show exponential decay.

Self-referential citation data:



More mattering:

Rich-get-richerness in social contagion:

- We love to rank everyone, everything: Top n lists.
- People, wealth, sports, music, movies, books, schools, cities, countries, dogs (13/10), ...
- Gameable: payola, astroturfing, sockpuppetry, John Barron (the sockpuppet hype man), ...
- Black-box ranking algorithms make ranking opaque.
- Black boxes are gameable but takes money and commensurate skill.
- Black box algorithms can make things spread rampantly.[1]
- No "regramming" is a positive feature of Instagram (also: Pratchett the Cat)
- What if a healthier Facebook is just ...

References I

[1] F. Auerbach.
Das gesetz der bevölkerungskonzentration.
Petermanns Geogr. Mitteilungen, 59:73–76, 1913.

[2] A.-L. Barabási and R. Albert.
Emergence of scaling in random networks.
Science, 286:509–511, 1999. pdf

[3] Y. Berset and M. Medo.
The effect of the initial network configuration on preferential attachment.
The European Physical Journal B, 86(6):1–7, 2013. pdf

[4] D. J. de Solla Price.
Networks of scientific papers.
Science, 149:510–515, 1965. pdf

References II

[5] P. S. Dodds, D. R. Dewhurst, F. F. Hazlehurst, C. M. Van Oort, L. Mitchell, A. J. Reagan, J. R. Williams, and C. M. Danforth.
Simon's fundamental rich-get-richer model entails a dominant first-mover advantage.
Physical Review E, 95:052301, 2017. pdf

[6] J.-B. Estoup.
Gammes sténographiques: méthode et exercices pour l'acquisition de la vitesse.
Institut Sténographique, 1916.

[7] P. L. Krapivsky and S. Redner.
Organization of growing random networks.
Phys. Rev. E, 63:066123, 2001. pdf

References III

[8] P. Krugman.
The Self-Organizing Economy.
Blackwell Publishers, Cambridge, Massachusetts, 1996.

[9] A. J. Lotka.
The frequency distribution of scientific productivity.
Journal of the Washington Academy of Science, 16:317–323, 1926.

[10] B. B. Mandelbrot.
An informational theory of the statistical structure of languages.
In W. Jackson, editor, Communication Theory, pages 486–502. Butterworth, Woburn, MA, 1953. pdf

Related papers:

References IV

[11] M. E. J. Newman.
The first-mover advantage in scientific publication.
Europhysics Letters, 86:68001, 2009. pdf

[12] M. E. J. Newman.
Prediction of highly cited papers.
Europhysics Letters, 105:28002, 2014. pdf

[13] D. D. S. Price.
A general theory of bibliometric and other cumulative advantage processes.
Journal of the American Society for Information Science, pages 292–306, 1976. pdf

[14] H. A. Simon.
On a class of skew distribution functions.
Biometrika, 42(3-4):425–440, 12 1955. pdf

References V

[15] G. U. Yule.
A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S.
Phil. Trans. B, 213:21–87, 1925. pdf

[16] G. K. Zipf.
Human Behaviour and the Principle of Least-Effort.
Addison-Wesley, Cambridge, MA, 1949.