Mechanisms for Generating Power-Law Distributions

Principles of Complex Systems

CSYS/MATH 300, Fall, 2010

Prof. Peter Dodds

Department of Mathematics & Statistics
Center for Complex Systems
Vermont Advanced Computing Center
University of Vermont

Outline

Random Walks
The First Return Problem
Examples
Variable transformation
Holtsmark’s Distribution
PLIPLO

Growth Mechanisms
Random Copying
Words, Cities, and the Web

References

Random walks

The essential random walk:

▶ One spatial dimension.
▶ Time and space are discrete
▶ Random walker (e.g., a drunk) starts at origin $x = 0$.

- Step at time $t$ is $\epsilon_t$:

$$\epsilon_t = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

Random walks

Displacement after $t$ steps:

$$x_t = \sum_{i=1}^{t} \epsilon_i$$

Expected displacement:

$$\langle x_t \rangle = \left\langle \sum_{i=1}^{t} \epsilon_i \right\rangle = \sum_{i=1}^{t} \langle \epsilon_i \rangle = 0$$

Random walks

Variances sum: $(\text{V})^*$

$$\text{Var}(x_t) = \text{Var} \left( \sum_{i=1}^{t} \epsilon_i \right) = \sum_{i=1}^{t} \text{Var}(\epsilon_i) = \sum_{i=1}^{t} 1 = t$$

$^*$ Sum rule = a good reason for using the variance to measure spread; only works for independent distributions.
Random walks

Power-Law Mechanisms

Random Walks

Variable transformation

Holtsmark's Distribution

PLIPLO Growth Mechanisms

Random Copying

Words, Cities, and the Web

References

Random walks

Power-Law Mechanisms

Random Walks

Examples

Variable transformation

Holtsmark's Distribution

PLIPLO Growth Mechanisms

Random Copying

Words, Cities, and the Web

References

The problem of first return:

> What is the probability that a random walker in one dimension returns to the origin for the first time after $t$ steps?
> Will our drunkard always return to the origin?
> What about higher dimensions?

See Feller, [3] Intro to Probability Theory, Volume I

Random walks

Power-Law Mechanisms

Random Walks

Examples

Variable transformation

Holtsmark's Distribution

PLIPLO Growth Mechanisms

Random Copying

Words, Cities, and the Web

References

Random walks—are weierder than you might think...

For example:

- $\xi_{r,t}$ is the probability that by time step $t$, a random walk has crossed the origin $r$ times.
- Think of a coin flip game with ten thousand tosses.
- If you are behind early on, what are the chances you will make a comeback?
- The most likely number of lead changes is... 0.

In fact:

$\xi_{r,1} > \xi_{r,2} > \cdots$

Even crazier:

The expected time between tied scores = $\infty$!
First returns

Reasons for caring:
1. We will find a power-law size distribution with an interesting exponent
2. Some physical structures may result from random walks
3. We’ll start to see how different scalings relate to each other
First Returns

Key observation:

# of $t$-step paths starting and ending at $x = 1$
and hitting $x = 0$ at least once

= # of $t$-step paths starting at $x = -1$ and ending at $x = 1$

= $N(-1, 1, t)$

So $N_{\text{first return}}(2n) = N(1, 1, 2n - 2) - N(-1, 1, 2n - 2)$

See this 1-1 correspondence visually...

Next problem: what is $N(i, j, t)$?

- # positive steps + # negative steps = $t$.
- Random walk must displace by $j - i$ after $t$ steps.
- # positive steps - # negative steps = $j - i$.
- # positive steps = $(t + j - i)/2$.

$N(i, j, t) = \left(\frac{t}{\text{# positive steps}}\right) = \left(\frac{t + j - i}{2}\right)$

We now have

$N_{\text{first return}}(2n) = N(1, 1, 2n - 2) - N(-1, 1, 2n - 2)$

where

$N(i, j, t) = \left(\frac{t}{(t + j - i)/2}\right)$
First Returns

Insert question from assignment 4 (III)

Find $N_{\text{first return}}(2n) \sim \frac{2n^{3/2}}{\sqrt{2\pi n}}$

- Normalized Number of Paths gives Probability
- Total number of possible paths = $2^{2n}$

$$P_{\text{first return}}(2n) = \frac{1}{2^{2n}} N_{\text{first return}}(2n)$$

$$\approx \frac{1}{\sqrt{2\pi n}} \frac{2^{2n}}{2^{2n}} \frac{3/2}{2n^{3/2}}$$

$$= \frac{1}{\sqrt{2\pi n}} (2n)^{-3/2}$$

Random walks

On finite spaces:

- In any finite volume, a random walker will visit every site with equal probability
- Random walking $\Rightarrow$ Diffusion
- Call this probability the Invariant Density of a dynamical system
- Non-trivial Invariant Densities arise in chaotic systems.

Random walks on networks:

- On networks, a random walker visits each node with frequency $\propto$ node degree
- Equal probability still present: walkers traverse edges with equal frequency.

Higher dimensions:

- Walker in $d = 2$ dimensions must also return
- Walker may not return in $d \geq 3$ dimensions
- For $d = 1, \gamma = 3/2 \Rightarrow \langle t \rangle = \infty$
- Even though walker must return, expect a long wait...

Scheidegger Networks [11, 2]

- Triangular lattice
- 'Flow' is southeast or southwest with equal probability.
Observed for real river networks
Larger basins more isometric (Typically: $1 < h < 1.5$, $1.5 < \gamma < 2$)
Smaller basins more allometric ($h > 1/2$)
Larger basins more isometric ($h = 1/2$)

Connections between Exponents

For a basin of length $\ell$, width $\propto \ell^{1/2}$

- $P(\ell) \propto \ell^{-3/2}$
- $P(a) \propto a^{-\gamma}$, and $P(\ell) \propto \ell^{-\tau}$

 HACK’s law [4] $\ell \propto a^h$

where $0.5 < h < 0.7$
Redo calc with $\gamma$, $\tau$, and $h$. 

Connections between Exponents

Both basin area and length obey power law distributions
Observed for real river networks
Typically: $1.3 < \tau < 1.5$ and $1.5 < \gamma < 2$
Smaller basins more allometric ($h > 1/2$)
Larger basins more isometric ($h = 1/2$)
### Other First Returns

**Failure**
- A very simple model of failure/death:
  - \( x_t = \text{entity's 'health' at time } t \)
  - \( x_0 \) could be > 0.
  - Entity fails when \( x \) hits 0.

**Streams**
- Dispersion of suspended sediments in streams.
- Long times for clearing.

### More than randomness

- Can generalize to Fractional Random Walks
- Levy flights, Fractional Brownian Motion
- In 1-d, \( \sigma \sim t^\alpha \)
  - \( \alpha > 1/2 \) — superdiffusive
  - \( \alpha < 1/2 \) — subdiffusive
- Extensive memory of path now matters...

### Variable Transformation

**Understand power laws as arising from**
1. elementary distributions (e.g., exponentials)
2. variables connected by power relationships
Example

Exponential distribution

Given \( P_\lambda(x) = \frac{1}{\lambda} e^{-x/\lambda} \) and \( y = cx^{-\alpha} \), then

\[
P(y) \propto y^{-1-1/\alpha} + O \left( y^{-1-2/\alpha} \right)
\]

- Exponentials arise from randomness...
- More later when we cover robustness.

Transformation

Using \( F \propto F^{-1/2} \) and \( P_\lambda(r) \propto r^\alpha \), then

\[
\int F \propto dF \propto d(\gamma - 1/2)
\]

\[
P_\lambda(F) dF = P_\lambda(r) dr
\]

\[
\propto P_\lambda(F^{-1/2}) F^{-3/2} dF
\]

\[
\propto (F^{-1/2})^2 F^{-3/2} dF
\]

\[
= F^{-1-3/2} dF
\]

\[
= F^{-5/2} dF
\]

Gravity

- Select a random point in the universe \( \vec{x} \)
- (possible all of space-time)
- Measure the force of gravity \( F(\vec{x}) \)
- Observe that \( P_F(F) \sim F^{-5/2} \).

Ingredients

Matter is concentrated in stars:

- \( F \) is distributed unevenly
- Probability of being a distance \( r \) from a single star at \( \vec{x} = 0 \):
  \[
P_\lambda(r) dr \propto r^\alpha dr
\]
- Assume stars are distributed randomly in space (oops?)
- Assume only one star has significant effect at \( \vec{x} \).
- Law of gravity:
  \[
  F \propto r^{-2}
  \]
- invert:
  \[
  r \propto F^{-1/2}
  \]

Transformation

- Mean is finite
- Variance = \( \infty \)
- A wild distribution
- Random sampling of space usually safe but can end badly...
Caution!

► PLIPLO = Power law in, power law out
► Explain a power law as resulting from another unexplained power law.
► Yet another homunculus argument... (slap, slap)
► Don't do this!!! (slap, slap)
► We need mechanisms!

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)

Work of Yore

► 1924: G. Udny Yule\textsuperscript{[14]}:
  # Species per Genus
► 1926: Lotka\textsuperscript{[13]}:
  # Scientific papers per author (Lotka's law)
► 1953: Mandelbrot\textsuperscript{[18]}:
  Optimality argument for Zipf's law; focus on language.
► 1955: Herbert Simon\textsuperscript{[12, 15]}:
  Zipf's law for word frequency, city size, income, publications, and species per genus.
► 1965/1976: Derek de Solla Price\textsuperscript{[9, 10]}:
  Network of Scientific Citations.
► 1999: Barabasi and Albert\textsuperscript{[1]}
<table>
<thead>
<tr>
<th>Random Competitive Replication</th>
<th>Power-Law Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Competition for replication between elements is random</td>
<td>Random Walks</td>
</tr>
<tr>
<td>- Competition for growth between groups is not random</td>
<td>The First Return Problem</td>
</tr>
<tr>
<td>- Selection on groups is biased by size</td>
<td>Examples</td>
</tr>
<tr>
<td>- Rich-gets-richer story</td>
<td>Variable</td>
</tr>
<tr>
<td>- Random selection is easy</td>
<td>innovation</td>
</tr>
<tr>
<td>- No great knowledge of system needed</td>
<td>rate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Competitive Replication</th>
<th>Power-Law Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Steady growth of system: +1 element per unit time.</td>
<td>Random Walks</td>
</tr>
<tr>
<td>- Steady growth of distinct flavors at rate $\rho$</td>
<td>The First Return Problem</td>
</tr>
<tr>
<td>- We can incorporate</td>
<td>Examples</td>
</tr>
<tr>
<td>1. Element elimination</td>
<td>Variable</td>
</tr>
<tr>
<td>2. Elements moving between groups</td>
<td>innovation</td>
</tr>
<tr>
<td>3. Variable innovation rate $\rho$</td>
<td>rate</td>
</tr>
<tr>
<td>4. Different selection based on group size</td>
<td>Power-Law</td>
</tr>
<tr>
<td>(But mechanism for selection is not as simple...)</td>
<td>Growth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Competitive Replication</th>
<th>Power-Law Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definitions:</td>
<td>Random Walks</td>
</tr>
<tr>
<td>- $k_i = \text{size of a group } i$</td>
<td>The First Return Problem</td>
</tr>
<tr>
<td>- $N_k(t) = # \text{groups containing } k \text{ elements at time } t$.</td>
<td>Examples</td>
</tr>
</tbody>
</table>

**Basic question:** How does $N_k(t)$ evolve with time?

First: $\sum_k kN_k(t) = t = \text{number of elements at time } t$

<table>
<thead>
<tr>
<th>Random Competitive Replication</th>
<th>Power-Law Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_k(t)$, the number of groups with $k$ elements, changes at time $t$ if</td>
<td>Random Walks</td>
</tr>
<tr>
<td>1. An element belonging to a group with $k$ elements is replicated</td>
<td>The First Return Problem</td>
</tr>
<tr>
<td>$N_k(t+1) = N_k(t) + 1$</td>
<td>Examples</td>
</tr>
<tr>
<td>Happens with probability $(1-\rho)kN_k(t)/t$</td>
<td>Variable</td>
</tr>
<tr>
<td>2. An element belonging to a group with $k-1$ elements is replicated</td>
<td>innovation</td>
</tr>
<tr>
<td>$N_k(t+1) = N_k(t) + 1$</td>
<td>rate</td>
</tr>
<tr>
<td>Happens with probability $(1-\rho)(k-1)N_{k-1}(t)/t$</td>
<td>Power-Law</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Competitive Replication</th>
<th>Power-Law Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special case for $N_k(t)$:</td>
<td>Random Walks</td>
</tr>
<tr>
<td>1. The new element is a new flavor:</td>
<td>The First Return Problem</td>
</tr>
<tr>
<td>$N_i(t+1) = N_i(t) + 1$</td>
<td>Examples</td>
</tr>
<tr>
<td>Happens with probability $\rho$</td>
<td>Variable</td>
</tr>
<tr>
<td>2. A unique element is replicated.</td>
<td>innovation</td>
</tr>
<tr>
<td>$N_i(t+1) = N_i(t)$</td>
<td>rate</td>
</tr>
<tr>
<td>Happens with probability $(1-\rho)N_i(t)/t$</td>
<td>Power-Law</td>
</tr>
</tbody>
</table>

$P_k(t) = \text{Probability of choosing an element that belongs to a group of size } k$:  
$N_k(t)$ size $k$ groups  
⇒ $kN_k(t)$ elements in size $k$ groups  
$t$ elements overall  

$P_k(t) = \frac{kN_k(t)}{t}$
Numbers of elements now fractional

Okay over large time scales

Drop expectations

Random Competitive Replication

For $k > 1$:

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

For $k = 1$:

$$
\langle N_1(t+1) - N_1(t) \rangle = \rho - (1 - \rho) 1 \cdot \frac{N_1(t)}{t}
$$

Random Competitive Replication

Assume distribution stabilizes: $N_k(t) = n_k t$

(Reasonable for $t$ large)

Drop expectations

Numbers of elements now fractional

Okay over large time scales

$n_k / \rho$ = the fraction of groups that have size $k$.

Random Competitive Replication

Stochastic difference equation:

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

becomes

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

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

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

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

Random Competitive Replication

We have a simple recursion:

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

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

- Interested in $k$ large (the tail of the distribution)
- Can be solved exactly.

Random Competitive Replication

- To get at tail: Expand as a series of powers of $1/k$

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

Random Competitive Replication

- We (okay, you) find

$$
\frac{n_k}{n_{k-1}} \approx (k - 1)^{2 - \rho}
$$

$$
\frac{n_k}{n_{k-1}} \approx \frac{k - 1}{k}^{2 - \rho}
$$

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

Random Competitive Replication

- Observe $2 < \gamma < \infty$ as $\rho$ varies.
- For $\rho \approx 0$ (low innovation rate):

$$
\gamma \approx 2
$$

Recalls Zipf’s law: $s_r \sim r^{-\gamma}$

(s_r - size of the rth largest element)

- We found $\alpha = 1 / (\gamma - 1)$
- $\gamma = 2$ corresponds to $\alpha = 1$
Random Competitive Replication

- We (roughly) see Zipfian exponent $^{[5]}$ of $\alpha = 1$ for many real systems: city sizes, word distributions, ...
- Corresponds to $\rho \to 0$ (Krugman doesn’t like it) $^{[5]}$
- But still other mechanisms are possible...
- Must look at the details to see if mechanism makes sense... more later.

Random Competitive Replication

We had one other equation:

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

- As before, set $N_t(t) = n_1 t$ and drop expectations

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

- Rearrange:

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

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

Random Competitive Replication

So...

$$N_t(t) = n_1 t = \frac{\rho t}{2 - \rho}$$

- Recall number of distinct elements $= \rho t$.
- Fraction of distinct elements that are unique (belong to groups of size 1):

$$\frac{N_t(t)}{\rho t} = \frac{1}{2 - \rho}$$

(also = fraction of groups of size 1)

- For $\rho$ small, fraction of unique elements $\sim 1/2$
- Roughly observed for real distributions
- $\rho$ increases, fraction increases
- Can show fraction of groups with two elements $\sim 1/6$
- Model does well at both ends of the distribution

Words

From Simon $^{[10]}$:

Estimate $\rho_{\text{est}} = \# \text{ unique words}/\# \text{ all words}$

For Joyce’s Ulysses: $\rho_{\text{est}} \approx 0.115$

<table>
<thead>
<tr>
<th>$N_1$ (real)</th>
<th>$N_1$ (est)</th>
<th>$N_2$ (real)</th>
<th>$N_2$ (est)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16,432</td>
<td>15,850</td>
<td>4,776</td>
<td>4,870</td>
</tr>
</tbody>
</table>

Evolution of catch phrases

- Yule’s paper (1924) $^{[14]}$:

“A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S.”

- Simon’s paper (1955) $^{[12]}$:

“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.

More on Herbert Simon (1916–2001):

- Political scientist
- Involved in Cognitive Psychology, Computer Science, Public Administration, Economics, Management, Sociology
- Coined ‘bounded rationality’ and ‘satisficing’
- Nearly 1000 publications
- An early leader in Artificial Intelligence, Information Processing, Decision-Making, Problem-Solving, Attention Economics, Organization Theory, Complex Systems, And Computer Simulation Of Scientific Discovery
- Nobel Laureate in Economics
New papers have no citations
Selection mechanism is more complicated

Self-fulfilling prophecy
Role model
Focused interview
Unintended (or unanticipated) consequences

Robert K. Merton:
The citation network of scientific papers

Derek de Solla Price was the first to study network evolution with these kinds of models.

Citation network of scientific papers
Price’s term:
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

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.”

(Matthew = unit of purchasing power.)
Matilda effect: women’s scientific achievements are often overlooked

Merton was a catchphrase machine:
1. Self-fulfilling prophecy
2. Role model
3. Unintended (or unanticipated) consequences
4. Focused interview → focus group
And just to be clear...
Merton’s son, Robert C. Merton, won the Nobel Prize for Economics in 1997.

Barabasi and Albert [1]—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)
+ Randomly connect to the node’s friends (also easy)
Scale-free networks = food on the table for physicists

References I


References II

References III

Empirical tests of Zipf's law mechanism in open source Linux distribution.

An informational theory of the statistical structure of languages.

Networks of scientific papers.

References IV

A general theory of bibliometric and other cumulative advantage processes.

The algebra of stream-order numbers.

On a class of skew distribution functions.

Critical Phenomena in Natural Sciences.

References V

A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S.

Human Behaviour and the Principle of Least-Effort.
Addison-Wesley, Cambridge, MA, 1949.