The Evolutionary Dynamics of Incubation Periods

Bertrand Ottino-Löffler
Advisor: Steve Strogatz, Co-Author: Jacob Scott

Cornell University

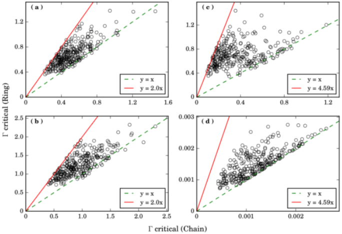
05/04/18

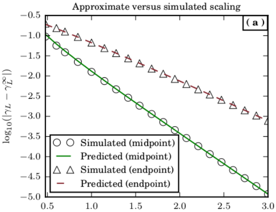
Things I have done

Previously:

Things I have done

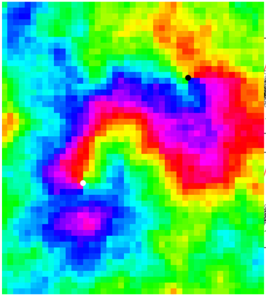
Previously:
Asymptotic Analysis!

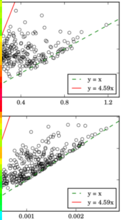


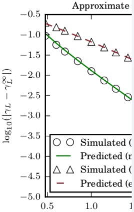


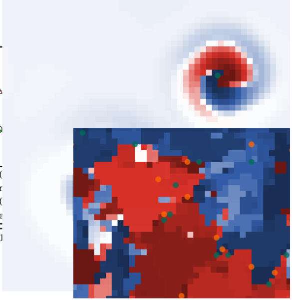
Things I have done

Previously:
Asymptotic Analysis! Coupled Oscillators!



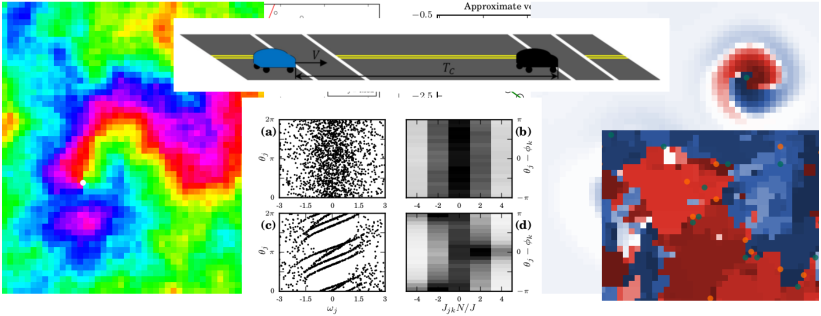






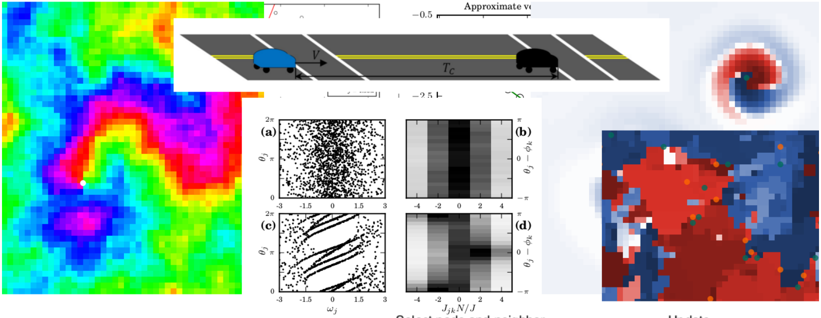
Things I have done

Previously:
Asymptotic Analysis! Coupled Oscillators! Dynamical Systems!

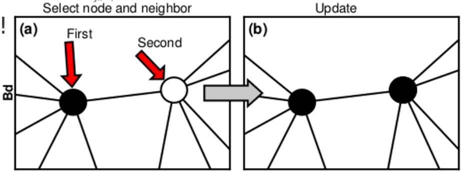


Things I have done

Previously:
Asymptotic Analysis! Coupled Oscillators! Dynamical Systems!



Today: Evolutionary dynamics!
Probability!
Stochastic Processes!
Typhoid!



This talk

eLIFE
elifesciences.org

RESEARCH ARTICLE

Evolutionary dynamics of incubation periods

Bertrand Ottino-Loffler[1], Jacob G Scott[2,3], Steven H Strogatz[1]*

[1]Center for Applied Mathematics, Cornell University, Ithaca, United States; [2]Department of Translational Hematology and Oncology Research , Cleveland Clinic, Cleveland, United States; [3]Department of Radiation Oncology, Cleveland Clinic, Cleveland, United States

Personal: people.cam.cornell.edu/∼bjo34/
Twitter: @OttinoLoffler

This talk

Personal: people.cam.cornell.edu/∼bjo34/
Twitter: @OttinoLoffler

Outline

1. Incubation periods, Sartwell's law
2. Moran models, Bd and Db at infinite fitness
3. The complete graph
4. The star graph
5. Lattices, critical dimensions
6. Neutral fitness
7. Summary and closing
8. Bonus: relaxation of assumptions

The Incubation Period

Definition

The **Incubation Period** of a disease is defined to be the time between first exposure to a contagion and observation of first symptoms.

The Incubation Period

Then incubation period of a disease is important for...

- ... individual diagnosis.
- ... deciding quarantine policy.
- ... predicting secondary outbreaks of epidemics.

However, they are difficult to measure.

The Incubation Period

The Incubation Period

NINETY-THREE PERSONS INFECTED BY A TYPHOID CARRIER AT A PUBLIC DINNER

WILBUR A. SAWYER, M.D.
Director of the Hygienic Laboratory of the California State Board of Health
BERKELEY, CAL.

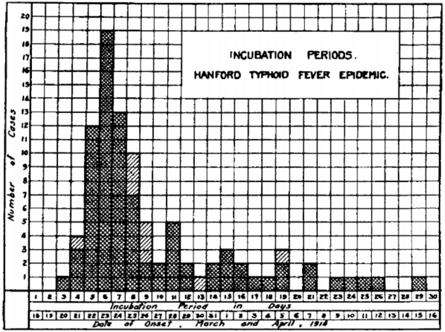


Chart of the cases in the Hanford typhoid fever epidemic, showing incubation periods and dates of onset. The heavily shaded areas represent definite cases of typhoid fever. The lightly shaded areas represent the doubtful cases.

The Incubation Period

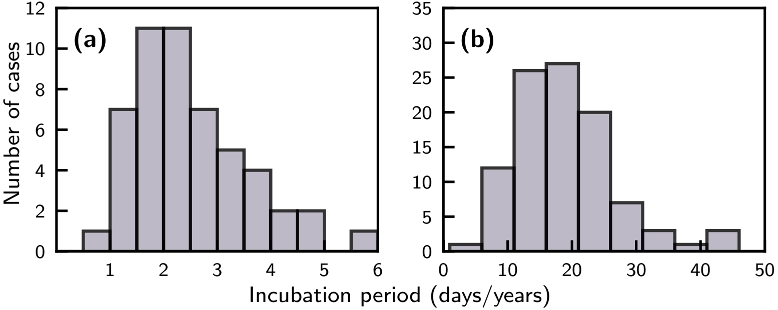
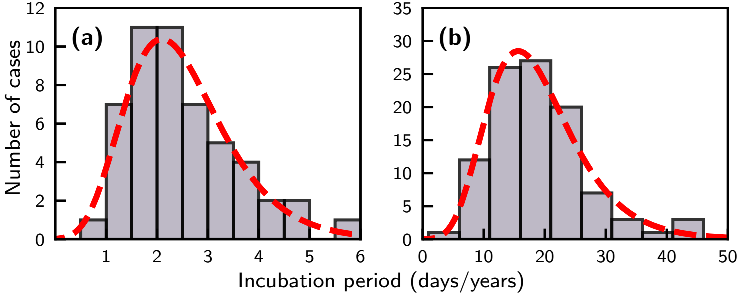


Figure: (a) Data from an outbreak of food-borne streptococcal sore throat, reported in 1950 (Sartwell, 1950). (b) Occupation-induced bladder tumors (Goldblatt, 1949).

Sartwell's Law (1966)

Sartwell's Law

Incubation periods for diseases tend to be distributed as lognormals; more generally, they will be right-skewed.



Explanations?

Explanations?

Traditionally: **Population-level heterogeneity** of ...

- ... The inoculum of contagion.
- ... The contagion's fitness.
- ... The host's immunosensitivity.

Noise in ODE

$$\theta = De^{rT}, \text{ where}$$

- T is the incubation period.
- D is the inoculum.
- r is the pathogen growth rate.
- θ is the host tolerance.

A common misconception

Adding normal randomness to the parameters in

$$T = \frac{1}{r} \log \left(\frac{\theta}{D} \right)$$

does **not** induce lognormals. Expand

Alternate explanations?

- Optimal virulence theory
- Immune system diversity
- Nonnegativity arguments

Alternate explanations?

Q: Can Sartwell's Law arise from *just* the intrinsic randomness of disease incubation?

An evolutionary graph theory approach

An evolutionary graph theory approach

Many illnesses consist of spreading on a within-host network

The Illness ...	... Takes over the ...	... Which is a ...
Typhoid	well-mixed gut microbiome	Complete graph
Leukemia	healthy bone marrow cells	3D lattice
Influenza	uncompromised tracheal cells	2D lattice

Evolutionary graph theory

- We will emulate the **Incubation Period** of a disease via the **The Takeover/Fixation Time** of an evolutionary takeover process on a network. Expand

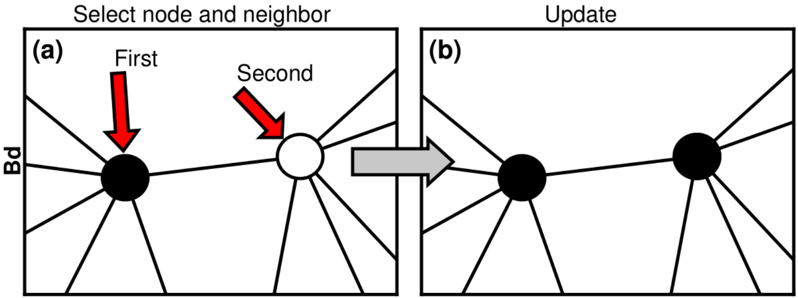
The Moran Model

Definition

The **Moran Birth-death (Bd) model** consists of three steps:

1. With probability proportional to fitness (r), randomly select a node on the network to give birth.
2. Uniformly randomly select a neighbor of the first node to die.
3. The dying node takes on the type of the birthing node.

The Moran Model (Family?) Expand



$r = \infty$?

Bd for $r = \infty$

1. Choose an invader, uniformly at random.

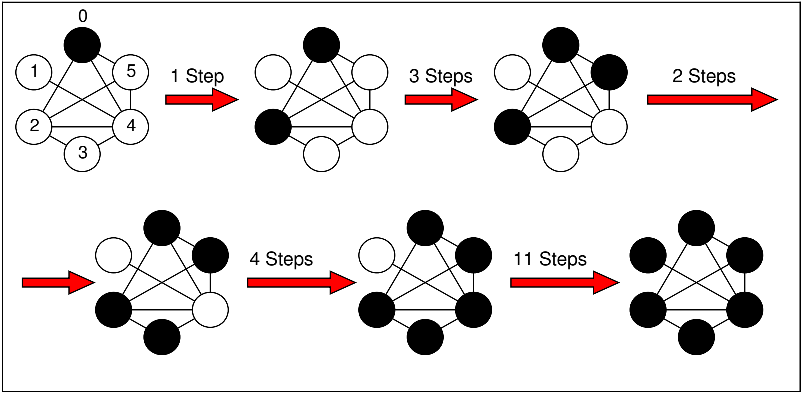
Bd for $r = \infty$

1. Choose an invader, uniformly at random.
2. Choose a neighbor of that invader, uniformly at random.

Bd for $r = \infty$

1. Choose an invader, uniformly at random.
2. Choose a neighbor of that invader, uniformly at random.
3. That neighbor is now an invader.

A path to takeover: $r = \infty$



Complete graph first

Complete graph: Bd

1. Toss out the original invader node, and put the $N - 1$ remaining nodes into a bag.
2. Select one node uniformly at random from the bag.
3. If the resident wasn't before, it is now an invader.
4. The node is returned to the bag, and we repeat.

The Coupon Collector's Problem

The Coupon Collector's Problem

Each day, a kid gets one trading card, uniformly at random. Given that there are N distinct cards, what is the distribution of times T required to form a complete set?

General proof outline

1. Write down the probability of invader addition p_m.
2. The total takeover time becomes a sum of **Geometric** random variables.
3. Translate a sum of geometric random variables T_G into a sum of **Exponential** random variables T_E.
4. Deduce properties of the distribution from T_E.

Proof outline: complete graph

1. A new invader gets added with probability $p_m = (N - m)/(N - 1)$.
2. Use Proposition to translate T_G into T_E.
3. Calculate the distribution via induction and take limits.

The Complete Graph Expand

Complete graph result

The takeover times of a complete graph go to a Gumbel.

Specifically,

$$\frac{T_G - E[T_G]}{N} \xrightarrow{d} \text{Gumbel}(-\gamma, 1), \tag{1}$$

where $\gamma \approx 0.5772$ *is the Euler-Mascheroni constant,* $\xrightarrow{d}$ *denotes convergence in distribution.*

Star graph: summary

Star graph: summary

WLOG take the hub node to be an invader. Given N spokes, m of which are invaders, then

$$\begin{aligned} p_m =& \text{(Prob. of chosing the hub invader)} \\ & \times \text{(Prob of choosing a resident spoke)} \\ =& \frac{1}{m+1}\frac{N-m}{N}, \end{aligned}$$

for $m = 0, \ldots, N-1$.

Star graph: summary

In the limit of large m and N,

$$p_m = \frac{1}{m+1}\frac{N-m}{N} \approx \frac{N-m}{N},$$

so the star graph resembles the coupon collector!

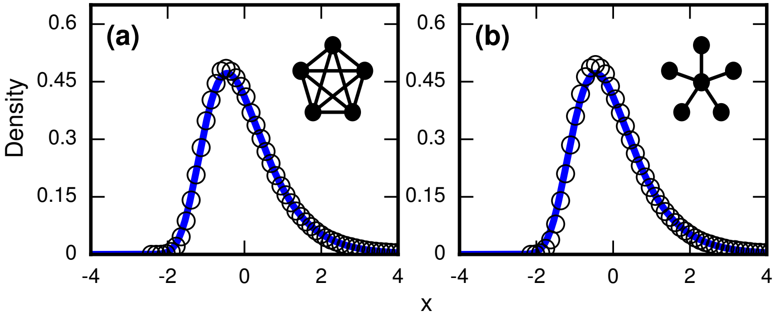
Star graph: comparing to coupons Expand

Star graph result

Using $\mu = N^2 \log(N) + N^2\gamma - N^2 + O(N \log(N))$, *we have*

$$\frac{T_G - \mu}{N^2} \sim \frac{T_E - \mu}{N^2} \sim \frac{T_C - E[T_C]}{N} \xrightarrow{d} Gumbel(-\gamma, 1)$$

Complete graph and star graph Expand



Results for lattices

Lattices: a problem

- Complete and star graph results heavily used p_m (the probability of adding a new invader, given there are currently m invaders).
- **However, p_m as a concept isn't well defined for lattices –** *configuration* might matter!

Lattices: geometric simplification

Lattices: geometric simplification 2

In many cases, it is possible to make an analogy to **first-passage percolation**. So these clusters are described by shape theorems, stating that these have a simple convex (but non-ball) limit shapes.

Lattices: surface area to volume

- In a d dimensional lattice, a simple convex shape of volume V has a surface area proportional to V^{η}, where $\eta = 1 - 1/d$.

Lattices: surface area to volume

- In a d dimensional lattice, a simple convex shape of volume V has a surface area proportional to V^{η}, where $\eta = 1 - 1/d$.
- The probability of adding a new invader $\propto$ to the probability of selecting a node on the boundary of the cluster of invader nodes.

Lattices: surface area to volume

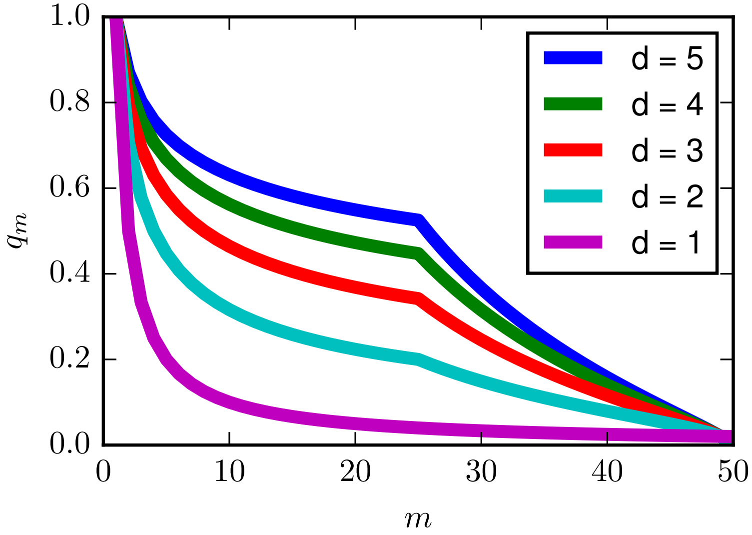
Given $\eta = 1 - 1/d$,

$$\begin{aligned} p_m &\propto \frac{1}{m} \cdot \text{Surface area of of the invader cluster} \\ &\propto q_m := \frac{\min(m, N-m)^\eta}{m}. \end{aligned}$$

Therefore,

$$T \approx \sum_{m=1}^{N-1} \mathrm{Geo}(q_m).$$

Lower dimensions are flatter



Lattices: low dimensions first

Because low dimensions have similar q_m, then

$$T \approx \sum_{m=1}^{N-1} \text{Geo}(q_m)$$

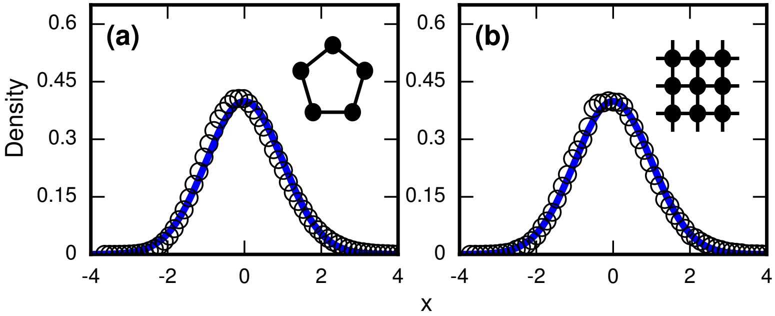
looks like a sum of similar variables.

Lattices: low dimensions first

Therefore, we can apply the

Lindeberg-Feller Central Limit Theorem

(or just use a Corollary).


(a)
(b)
Density
0.6
0.45
0.3
0.15
0
-4
-2
0
2
4
x

Lattices: no closed-form for high dimensions

Lattices: skew results

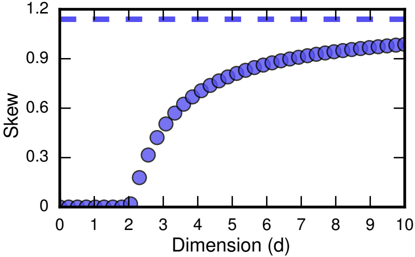
- We don't need the full distribution, just the skew.
- Getting skews is (somewhat) easy.

Lattice: $d \geq 3$ Expand

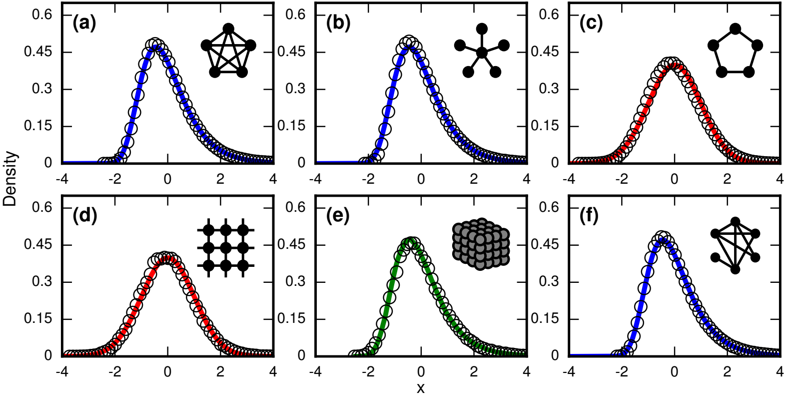
Theorem

Letting $\eta = 1 - 1/d$, the asymptotic skew of the takeover times for a $d > 2$ dimensional lattice is given by

$$Skew(d) = \frac{2\zeta(3\eta)}{\zeta(2\eta)^{3/2}}, \quad \textit{where } \zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}.$$



Summary: infinite r Expand



Complete graph, $r = 1$

Complete graph, $r = 1$

By symmetry, $p_m^+ = p_m^- = \frac{m(N-m)}{N(N-1)}$. Therefore,

$$X_n := \text{The population level after } n \text{ changes} = \sum_{i=1}^{n} x_i,$$

where $x \in \{-1, +1\}$, each with probability $1/2$.

Conditioned random walk

- Fact: We can only record an incubation period if someone *actually* gets sick.
- Therefore, we need to condition on the population X_n hitting N before ever hitting 0.

Conditioned random walk

The important stopping time is the first hitting time of 0 or N, so use

$$S = \min\{S_0, S_N\}, \text{ where } S_m = \min\{n | X_n = m\}.$$

To find the appropriate moments of the conditioned fixation times, set up the moments

$$\mu_i := E(S^i | X_S = N).$$

Conditioned random walk

From here, it is just a matter of finding the right martingales and applying the Optional Stopping Theorem.

Proposition

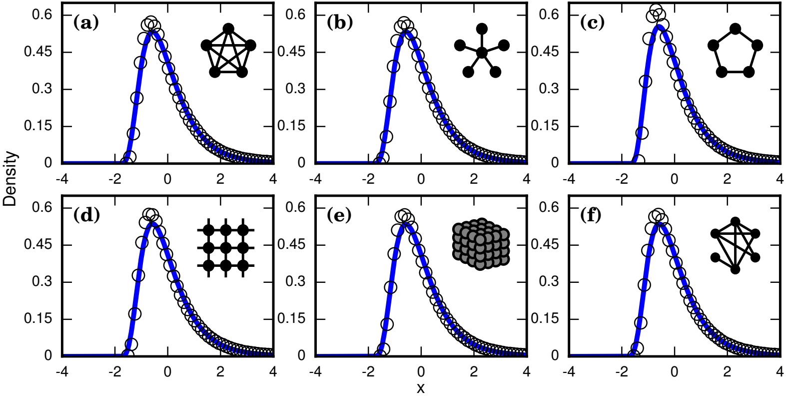
The exit time of an unbiased random walk which starts at 1 and hits N before hitting 0 has a high level of skew (≈ 1.807).

Conditioned random walk

Proposition

Conditioning on the success of the invaders induces a high level of skew.

Summary: $r = 1$ Expand



Realism?

Realism?

- Sartwell measured **Dispersion Factors**, the standard deviations of the logs of the data, across many diseases.

Realism?

- Sartwell measured **Dispersion Factors**, the standard deviations of the logs of the data, across many diseases.
- He measured dispersion factors between **1.1 and 1.5** for real-world diseases.

Realism?

- Sartwell measured **Dispersion Factors**, the standard deviations of the logs of the data, across many diseases.
- He measured dispersion factors between **1.1 and 1.5** for real-world diseases.
- For our high fitness simulations, we measured factors between **1.1 and 1.4**.

Realism?

- Sartwell measured **Dispersion Factors**, the standard deviations of the logs of the data, across many diseases.
- He measured dispersion factors between **1.1 and 1.5** for real-world diseases.
- For our high fitness simulations, we measured factors between **1.1 and 1.4**.
- For neutrally fit invaders, we measured factors between **1.6 and 1.7**.

The Incubation Period

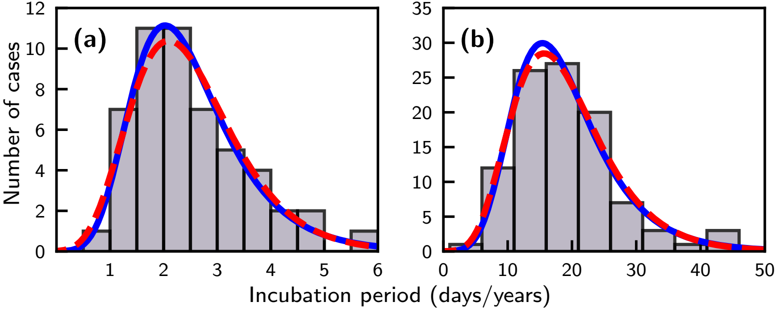


Figure: (a) Data from an outbreak of food-borne streptococcal sore throat, reported in 1950 (Sartwell, 1950). (b) Occupation-induced bladder tumors (Goldblatt, 1949).

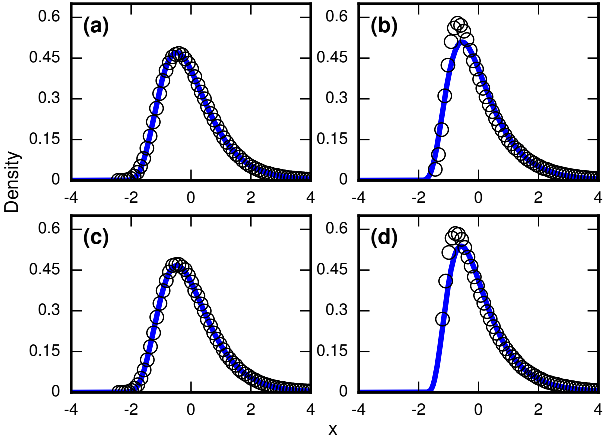


Figure: Complete graph with $r = 10$ under various forms of heterogeneity. (a) heterogeneity in invaders. (b) Heterogeneity of host sensitivity. (c) Heterogeneity of initial dosage. (d) Heterogeneity of all three.

Main points

Important Facts:

Main points

Important Facts:

- When the invader fitness is high, dynamics are dominated by **The Coupon Collector's Problem**, leading to right skewed distributions.

Main points

Important Facts:

- When the invader fitness is high, dynamics are dominated by **The Coupon Collector's Problem**, leading to right skewed distributions.
- There is a **Critical Dimension** in the infinite fitness case, with higher dimensional topologies leading to more skewed distributions.

Main points

Important Facts:

- When the invader fitness is high, dynamics are dominated by **The Coupon Collector's Problem**, leading to right skewed distributions.
- There is a **Critical Dimension** in the infinite fitness case, with higher dimensional topologies leading to more skewed distributions.
- When invader fitness is low, dynamics are dominated by a **Conditioned Random Walk**, leading to right skewed distributions.

Main points

Important Facts:

- When the invader fitness is high, dynamics are dominated by **The Coupon Collector's Problem**, leading to right skewed distributions.
- There is a **Critical Dimension** in the infinite fitness case, with higher dimensional topologies leading to more skewed distributions.
- When invader fitness is low, dynamics are dominated by a **Conditioned Random Walk**, leading to right skewed distributions.
- While population-level heterogeneity can be tuned to cause right-skewed distributions, **Such Heterogeneity Isn't Necessary** for these distributions, and just accentuate the fundamental mechanisms we already observe.

Main points

And Most Importantly...

- While lognormal-like distributions can be justified in any number of ways, **Evolutionary Network Dynamics** is one of the only phenomena common to the diverse range of diseases shown to obey Sartwell's law.

What's next?

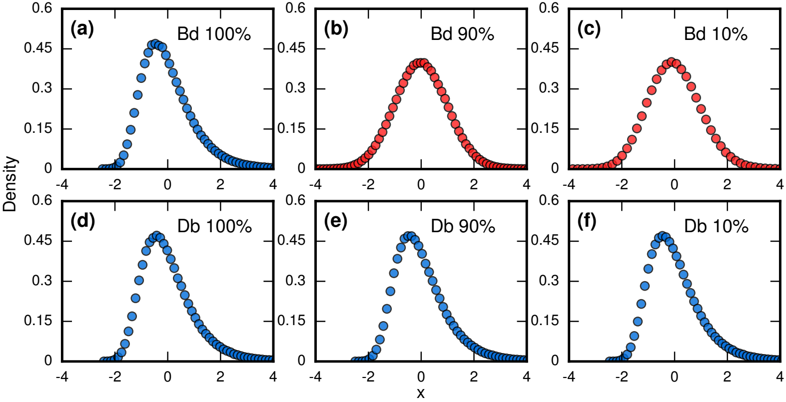


Figure: Top row: Birth-death. Bottom row: Death-birth.

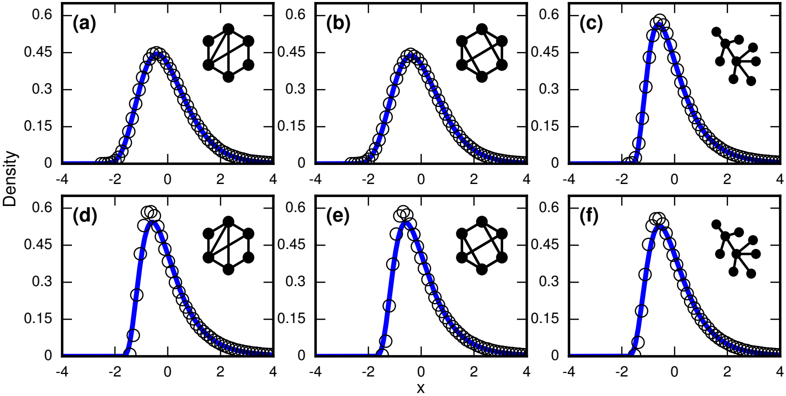


Figure: Top row: $r = \infty$. Bottom row: $r = 1$.

Variable population Expand

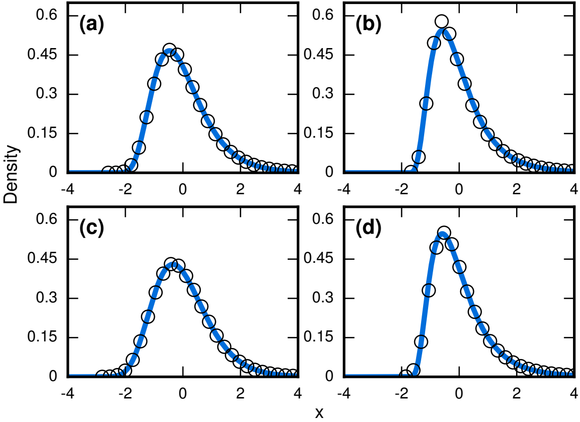


Figure: Complete graph with $r = 10$. (a) Constant total population. (b) Growing resident population. (c) Shrinking resident population. (d) Randomly varying resident population.

Selected References

- B Ottino-Löffler, J G Scott, and S H Strogatz, Evolutionary Dynamics of Incubation Periods, `http://www.biorxiv.org/content/early/2017/05/30/144139`.
- B Ottino-Löffler, J G Scott, and S H Strogatz, Takeover Times for a Simple Model of Network Infection, Physical Review E, 96m 012313 (2017).
- W A Sawyer, Ninety-Three Persons Infected by a Typhoid Carrier at a Public Dinner, Journal of the American Medical Association, 63(18) (1914).
- P E Sartwell, The Distribution of Incubation Periods of Infectious Disease, Am. J. Hyg. 51, 310-318 (1950).
- M W Goldblatt, Vesical tumours induced by chemical compounds, Occupational and Environmental Medicine 6:6581 (1949).
- L Ryan, Meet the best 9-year-old Pokemon card player in Minnesota, Star Tribune, March 9 (2015).
- P A P Moran, The Effect of Selection in a Haploid Genetic Population, Proc. Camb. Phil. Soc., 54(8) (1958).
- E Lieverman, C Hauert, and M A Nowak, Evolutionary Dynamics on Graphs, Nature, 433(7023) (2005).
- P Erdős and A Rényi, On a Classical Problem of Probability Theory, Publ. Math. Inst. Hung. Acad. Sci 6 (1961).
- L E Baum and P Billingsley, Asymptotic Distributions for the Coupon Collector's Problem, The Annals of Mathematical Statistics, 36(6) (1965).
- A Auffinger, M Damron, and J Hanson, 50 years of First Passage Percolation, `https://arxiv.org/abs/1511.03262` (2016).
- Durrett R. 1991. Probability: Theory and Examples. Belmont: Brooks/Cole.

This research was supported by a Sloan Fellowship to Bertrand Ottino-Löffler, in the Center for Applied Mathematics in Cornell, as well as by NSF grants DMS-1513179 and CCF-1522054.

Questions?

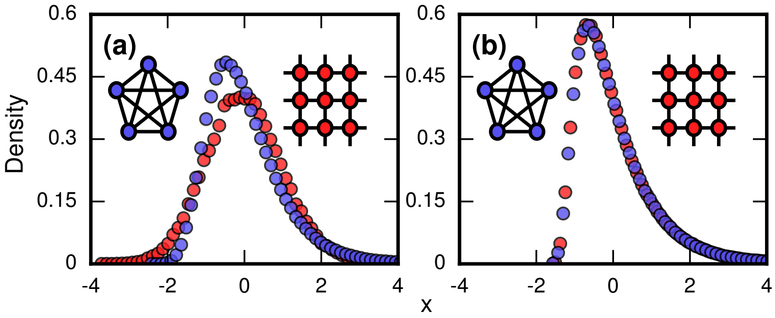
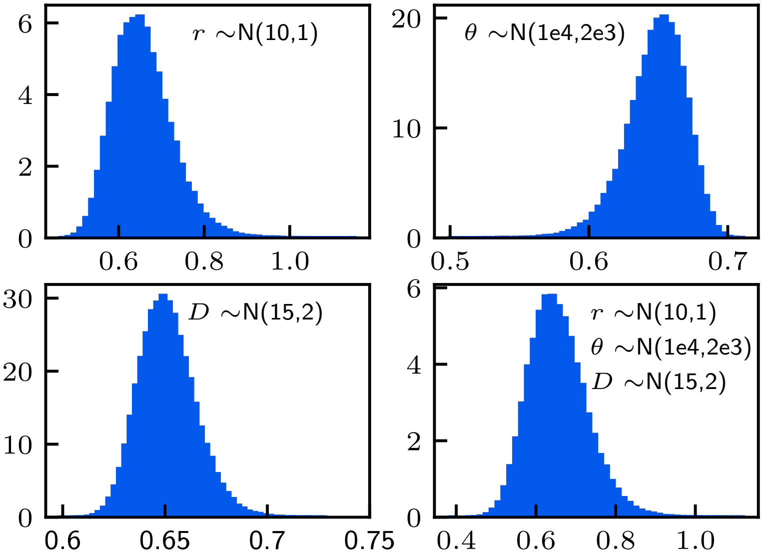


Figure: (a) $r = \infty$. (b) $r = 1$.

Personal: people.cam.cornell.edu/∼bjo34/
Twitter: @OttinoLoffler

Appendix: Additional info

A common misconception [Return] [Expand]

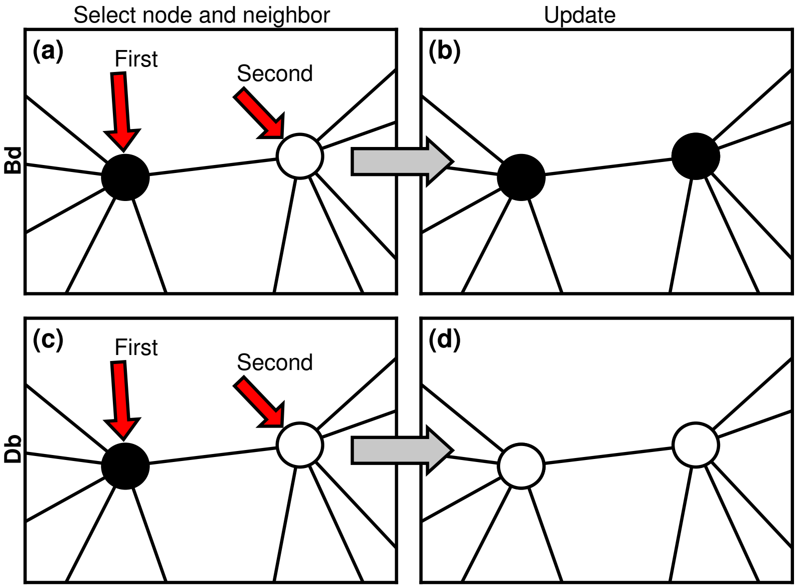


Takeover Time Return

Definition

The **Fixation (or Takeover) Time** of a network evolutionary process is the time between the appearance of a single invader and 100% of the resident nodes being replaced by invaders. (The initial population of invaders and the final takeover threshold can both be adjusted, if desired.)

The Moran model family (I) Return



The Moran model family (II) Return

Step order	Birth first	Death first
Fitness-step first	Bd	Db
Fitness-step second	bD	dB
Both fitness-steps	BD	DB

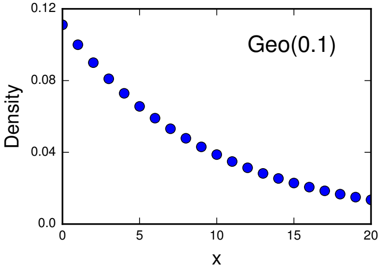
Geometric random variables Return R. Coupon

Definition

Define Geo(p) to be a geometric random variable with distribution

$$P(\text{Geo}(p) = k) = (1-p)^{k-1}p$$

for $k = 1, 2, \ldots$ and $0 < p \leq 1$.



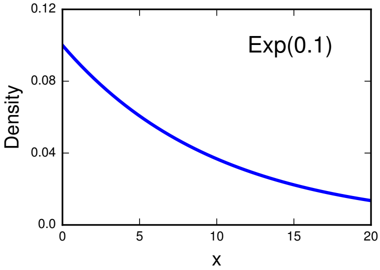
Exponential random variables Return

Definition

Define $\mathcal{E}(p)$ be an exponential random variable with density

$$pe^{-px}dx$$

for $x \geq 0$.



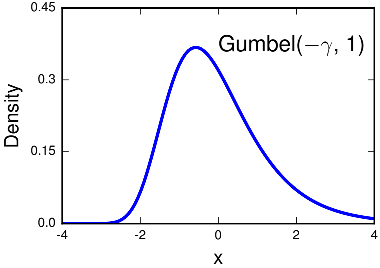
Gumbel random variables R. Main R. Coupon

Definition

Define Gumbel(α, β) be a Gumbel random variable with density

$$\beta^{-1} e^{-(x-\alpha)/\beta} \exp\left(-e^{-(x-\alpha)/\beta}\right) dx$$

for all x.



Lindeberg-Feller Central Limit Theorem Return R. Normal

Theorem

Suppose the random variables $Y_{m,N}$ *are such that* $E[Y_{m,N}] = 0$ *and* $\sum_m E[Y_{m,N}^2] = 1$ *for all* N*, and also*

$$\lim_{N \to \infty} \sum_{m=1}^{N} E[Y_{m,N}^2; |Y_{m,N}| > \epsilon] = 0 \qquad (2)$$

for all $\epsilon > 0$*. Then*

$$\sum_{m=1}^{N} Y_{m,N} \xrightarrow{d} Normal(0, 1) \qquad (3)$$

Lemma

If we have random variables X_i with variances σ_i^2 and skews κ_i, then their sum has a skewness of

$$Skew\left(\sum_i X_i\right) = \frac{\sum_i \kappa_i \sigma_i^3}{\left(\sum_i \sigma_i^2\right)^{3/2}}. \tag{4}$$

Optional stopping theorem Return R. RW

Theorem

Given a martingale M_n and a stopping time S, then

$$E[M_S] = E[M_0]$$

Agreement of geometric and exponential variables I

Return R. Coupon R. Star

Proposition

Given probabilities $p_m := p_m(M)$ *and* $L := L(M)$ *divergent, if*

$$\lim_{M\to\infty} \sum_{m=1}^{M} \frac{1}{p_m L^2} = 0,$$

then

$$\frac{1}{L}\left(\sum_{m=1}^{M} Geo(p_m) - 1/p_m\right) \sim \frac{1}{L}\left(\sum_{m=1}^{M} \mathcal{E}(p_m) - 1\right).$$

The symbol "$\sim$" means the ratio of characteristic functions goes to 1 as N gets large.

Poof of prop 1 (I) Return

This is proven by finding the characteristic functions for both sides, and showing that the ratio of these functions goes to 1 as M gets large.

Poof of prop 1 (II) Return

Our variables:

$$T_G = \sum_{m=1}^{M} \text{Geo}(p_m)$$

$$T_E = \sum_{m=1}^{M} \mathcal{E}(p_m)$$

$$\mu = \sum_{m=1}^{M} 1/p_m$$

Poof of prop 1 (III) Return

Want to show

$$\frac{T_G - \mu}{L} \sim \frac{T_E - \mu}{L}.$$

Poof of prop 1 (IV) Return

Our characteristic functions:

$$\phi_G = E\left[e^{it\frac{T_G-\mu}{L}}\right] = \prod_{m=1}^{M} \frac{p_m \exp\left[(it/L)\left(1-1/p_m\right)\right]}{1-\left(1-p_m\right)\exp\left(it/L\right)}$$

$$\phi_E = E\left[e^{it\frac{T_E-\mu}{L}}\right] = \prod_{m=1}^{M} \frac{\exp\left[-it/(p_m L)\right]}{1-it/(p_m L)}$$

Poof of prop 1 (V) Return

Fix t and take a ratio:

$$\phi_E/\phi_G = \prod_{m=1}^{M} \frac{\exp(-it/L) - (1 - p_m)}{p_m\,[1 - it/(p_m L)]}.$$

Poof of prop 1 (VI) Return

L gets large, so there's some vanishing $R_1 := R_2(M)$ such that

$$\exp(-it/L) = 1 + (-it/L) + R_1 t^2/L^2.$$

So then we have

$$\phi_E/\phi_G = \prod_{m=1}^{M} \left(1 + \frac{t^2}{p_m L^2} \frac{R_1}{1 - it/(p_m L)}\right).$$

Poof of prop 1 (VII) Return

$$\phi_E/\phi_G = \prod_{m=1}^{M}\left(1 + \frac{t^2}{p_m L^2}\frac{R_1}{1 - it/(p_m L)}\right)$$

Notice

- $|1 - it/(p_m L)| \geq 1$.
- The sum of $1/(p_m L^2)$ goes to 0.
- Therefore, each individual $p_m L^2$ gets large for all m.
- So, the second term is small.
- Therefore it can be rewritten exactly as an appropriate exponential.

Poof of prop 1 (VIII) Return

So

$$\phi_E/\phi_G = \prod_{m=1}^{M} \exp\left[R_2 t^2/(p_m L)\right]$$
$$= \exp\left[t^2 R_2 \sum_{m=1}^{M} \frac{1}{p_m L^2}\right] \to 1,$$

where the final limit comes from our assumption on $\sum_{m=1}^{M} \frac{1}{p_m L^2}$. ✓

Agreement of geometric and exponential variables II

Return | R. 1D | R. 2D | R. 3D

Proposition

Use the setup of the previous prop, and define $\sigma_G^2 = Var(T_G)$ *and* $\sigma_E^2 = Var(T_E)$. *If*

$$\lim_{M\to\infty} \frac{\sum_{m=1}^M p_m^{-1}}{\sum_{m=1}^M p_m^{-2}} = 0,$$

then

$$\frac{T_G - \mu}{\sigma_G} \sim \frac{T_E - \mu}{\sigma_E}.$$

Poof of prop 2 (I) Return

First, we check if

$$\frac{T_G - \mu}{\sigma_G} \sim \frac{T_E - \mu}{\sigma_G}.$$

follows from "Agreement of geometric and exponential variables I."

Poof of prop 2 (II) Return

Notice that

$$\begin{aligned}\lim_{M\to\infty}\sum_{m=1}^{M}\frac{1}{p_m\sigma_G^2} &= \lim_{M\to\infty}\frac{\sum_{m=1}^{M}p_m^{-1}}{\sum_{m=1}^{M}p_m^{-2}-p_m^{-1}}\\ &= \lim_{M\to\infty}\frac{\sum_{m=1}^{M}p_m^{-1}/\sum_{m=1}^{M}p_m^{-2}}{1-\sum_{m=1}^{M}p_m^{-1}/\sum_{m=1}^{M}p_m^{-2}}\\ &= \frac{0}{1-0}=0\end{aligned}$$

by hypothesis, so the condition of "Agreement of geometric and exponential variables I" is met.

Poof of prop 2 (III) Return

Therefore,

$$\frac{T_G - \mu}{\sigma_G} \sim \frac{T_E - \mu}{\sigma_G} = \frac{\sigma_E}{\sigma_G}\frac{T_E - \mu}{\sigma_E}.$$

Poof of prop 2 (IV) Return

However,

$$\lim_{M\to\infty} \frac{\sigma_E^2}{\sigma_G^2} = \lim_{M\to\infty} \frac{\sum_{m=1}^{M} p_m^{-2}}{\sum_{m=1}^{M} p_m^{-2} - p_m^{-1}}$$

$$= \lim_{M\to\infty} \frac{1}{1 - \sum_{m=1}^{M} p_m^{-1} / \sum_{m=1}^{M} p_m^{-2}}$$

$$= 1.$$

Poof of prop 2 (V) Return

Therefore,

$$\frac{T_G - \mu}{\sigma_G} \sim \frac{T_E - \mu}{\sigma_E},$$

so the proposition is proven. ✓

Proposition

Let $T = \sum_{m=1}^{M} \mathcal{E}(p_m)$, *define* $\sigma^2 = Var(T) = \sum_{m}^{M} p_m^{-2}$, *and let* $\lim_{M\to\infty} p_m\sigma = \infty$. *If*

$$\lim_{M\to\infty} \sum_{m=1}^{M} \exp\left(-\epsilon p_m \sigma\right) = 0,$$

then

$$\frac{T-\mu}{\sigma} \xrightarrow{d} Normal(0,1).$$

Poof of condition for normality (I) Return

To prove this, we apply the Lindeberg-Feller central limit theorem [12] to the random variables

$$Y_{m,M} := \frac{\mathcal{E}(p_m) - 1/p_m}{\sigma},$$

since

$$\sum_m (Y_m, M) = (T - \mu)/\sigma.$$

Poof of condition for normality (II) Return

To apply this, we need to check three conditions.

Poof of condition for normality (III) Return

First condition holds!

$$E[Y_{m,M}] = 0$$

Poof of condition for normality (IV) Return

Second condition holds!

$$\sum_m E[Y^2_{m,M}] = 1$$

Poof of condition for normality (V) Return

The third condition is more difficult ...

$$\forall \epsilon \lim_{M \to \infty} \text{Lind}_M := \lim_{M \to \infty} \sum_{m=1}^{M} E[Y_{m,M}^2; |Y_{m,M}| > \epsilon] = 0?$$

Poof of condition for normality (VI) Return

Notice that $Y_{m,M} < -\epsilon$ implies

$$\mathcal{E}(p_m) < p_m^{-1} - \epsilon\sigma_E^2 = p_m^{-1}(1 - \epsilon p_m \sigma).$$

By hypothesis, the right hand side will eventually be less than 0, meaning that eventually $Y_{m,M} < -\epsilon$ will be impossible.

Poof of condition for normality (VII) Return

Therefore

$$\lim_{M\to\infty} \mathrm{Lind}_M = \lim_{M\to\infty} \sum_{m=1}^{M} E[Y_{m,M}^2; Y_{m,M} > \epsilon].$$

Poof of condition for normality (VIII) Return

So for large enough M and defining $c_m := 1 + \epsilon p_m \sigma$, we have

$$\begin{aligned}
\mathrm{Lind}_M &:= \sum_{m=1}^{M} \int_{c_m/p_m}^{\infty} \left(\frac{x - 1/p_m}{\sigma} \right)^2 e^{-p_m x} p_m dx \\
&= \sum_{m=1}^{M} \frac{1}{\sigma^2 p_m^2} \int_{c_m}^{\infty} (y-1)^2 e^{-y} dy \\
&= \sum_{m=1}^{M} \frac{1}{\sigma^2 p_m^2} e^{-c_m} (c_m^2 + 1) \\
&= \sum_{m=1}^{M} \frac{1}{\sigma^2 p_m^2} e^{-\epsilon p_m \sigma} (2 + 2\epsilon p_m \sigma + \epsilon^2 p_m^2 \sigma^2).
\end{aligned}$$

Poof of condition for normality (IX) Return

By hypothesis, $p_m\sigma$ grows without bound, so the $p_m^2\sigma^2$ term will be dominant.
Therefore, there is some constant D such that we can create the upper bound:

$$\begin{aligned}\mathrm{Lind}_M &\leq \sum_{m=1}^{M} \frac{1}{\sigma^2 p_m^2} e^{-\epsilon p_m \sigma} D\sigma^2 p_m^2 \\ &\leq D \sum_{m=1}^{M} \exp\left(-\epsilon p_m \sigma\right).\end{aligned}$$

Poof of condition for normality (X) Return

By hypothesis,

$$\lim_{M\to\infty} \sum_{m=1}^{M} \exp\left(-\epsilon p_m \sigma\right) = 0,$$

therefore,

$$\lim_{M\to\infty} \mathrm{Lind}_M \leq \lim_{M\to\infty} D \sum_{m=1}^{M} \exp\left(-\epsilon p_m \sigma\right) = 0.$$

Poof of condition for normality (XI) Return

Therefore, the Lindeberg condition holds, so

$$\sum_m (Y_{m,M}) = \frac{T-\mu}{\sigma} \xrightarrow{d} \text{Normal}(0,1). \checkmark$$

Proposition

Given exponential random variables $\mathcal{E}(p_m)$ for $m = 1, \ldots, M$, with p_m distinct, then $\sum_{m=1}^{M} \mathcal{E}(p_m)$ is distributed according to the density

$$g_M(x) = \sum_{m=1}^{M} p_m e^{-p_m x} \prod_{k=1, k \neq m}^{M} \frac{p_k}{p_k - p_m} \tag{5}$$

Proof of sum of exponentials (I) Return

This is a (mostly) simple exercise in induction.

Proof of sum of exponentials (II) Return

Basis case of $M = 1$ is trivial.

Proof of sum of exponentials (III) Return

To get the inductive step, just convolute,

$$g_{M+1}(x) = \int_0^x p_{M+1} e^{-p_{M+1}(x-y)} g_M(y) dy.$$

Proof of sum of exponentials (IV) Return

Directly compute to get

$$g_{M+1}(x) = \sum_{m=1}^{M} p_m e^{-p_m x} \prod_{k \neq m}^{M+1} \frac{p_m}{p_k - p_m}$$
$$+ \sum_{m=1}^{M} \frac{p_m p_{M+1}}{p_m - p_{M+1}} e^{-p_{M+1} x} \prod_{k \neq m}^{M} \frac{p_k}{p_k - p_m}.$$

The first term is good, the second isn't.

Proof of sum of exponentials (V) Return

The second term becomes

$$\text{Second Term} = e^{-p_{M+1}x}\left(\prod_{m=1}^{M}\frac{p_m}{p_m - p_{M+1}}\right) b(p_{M+1}),$$

where we define

$$b(z) := \sum_{m=1}^{M}\prod_{k\neq m}^{M}\frac{p_k - z}{p_k - p_m}.$$

We can interpret $b(z)$ as a polynomial of at most degree $M-1$ in z (a Lagrange polynomial, to be specific).

Proof of sum of exponentials (VI) Return

- For any $k \in \{1, \ldots, M\}$ then $b(p_k) = 0$.
- Therefore, $b(z) - 1$ is a polynomial with M distinct roots.
- $b(z) - 1$ has maximum degree of $M - 1$.

Therefore, $b(z) \equiv 1$.

Proof of sum of exponentials (VII) Return

Plugging in gives

$$g_{M+1}(x) = \sum_{m=1}^{M+1} p_m e^{-p_m x} \prod_{k=1, k\neq m}^{M+1} \frac{p_k}{p_k - p_m},$$

which is the desired result. ✓

Proposition

Let T_C be the time to complete a set of N cards by drawing one at a time, uniformly at random with replacement. The T_C is distributed according to

$$\frac{T_C - \mu}{N} \xrightarrow{d} Gumbel(-\gamma, 1),$$

where $\mu = N\log(N) + N\gamma + O(\log(N))$.

Proof of coupon collector (I) Return

If we have collected m coupons then the probability of the next coupon being a new one is

$$p_m = \frac{N - m}{N}.$$

Proof of coupon collector (II) Return

Therefore, the time it takes to get the next card is just

$$\mathrm{Geo}(p_m),$$

where this is a geometric random variable. Expand

Proof of coupon collector (III) Return

So, the full collection time is

$$T_C = \sum_{m=1}^{N-1} \text{Geo}\left(p_m\right)$$

$$= \sum_{m=1}^{N-1} \text{Geo}\left(\frac{N-m}{N}\right)$$

$$= \sum_{k=1}^{N-1} \text{Geo}\left(\frac{k}{N}\right)$$

Proof of coupon collector (IV) Return

By switching from geometric to exponential random variables Prop 1, and setting $T_E = \sum_{m=1}^{M} \mathcal{E}(p_m)$, we get

$$\frac{T_C - \mu}{N} \sim \frac{T_E - \mu}{N}.$$

Proof of coupon collector (V) Return

Because we have a sum of exponential random variables, we have

$$\frac{T_E}{N} = \sum_{m=1}^{N-1} \mathcal{E}(Np_m) = \sum_{k=1}^{N-1} \mathcal{E}(k).$$

Proof of coupon collector (VI) Return

By using induction, Expand we have that T_E/N has a distribution given by

$$g_N(x) = \sum_{k=1}^{N-1} k e^{-kx} \prod_{r=1, r\neq k}^{N-1} \frac{r}{r-k} \\ = (N-1)e^{-x}(1-e^{-x})^{N-2}$$

Proof of coupon collector (VII) Return

To find the final distribution, reintroduce the shift of μ/N and take the limit to get

$$f(x) = e^{-(x+\gamma)} e^{-e^{-(x+\gamma)}},$$

which is a special case of the Gumbel distribution. Expand

Proof of coupon collector (VIII) Return

Therefore, we have

$$\frac{T_C - \mu}{N} \xrightarrow{d} \text{Gumbel}(-\gamma, 1). \checkmark$$

Normality of 1D ring Return

Proposition

The distribution of fixation times T for a 1D ring of N nodes is

$$\frac{T-\mu}{\sigma} \xrightarrow{d} \mathit{Normal}(0,1),$$

where μ and σ^2 are the mean and variance of T.

Normality of 1D ring (I) Return

The invader population on a ring will always be a single chain with two free ends.

Normality of 1D ring (II) Return

Given m invaders, the probability of a new invader being added is

$$p_m = \frac{1}{m}$$

for $m = 1, 2, \ldots, M$, where $M := N - 1$.

Normality of 1D ring (III) Return

The total fixation time is therefore

$$T = \sum_{m=1}^{M} \mathrm{Geo}(p_m) = \sum_{m=1}^{M} \mathrm{Geo}(1/m)$$

Normality of 1D ring (IV) Return

- Use proposition to switch to exponential variables. Expand
- Use proposition to apply CLT. Expand

Normality of 1D ring (V) Return

Therefore,

$$\frac{T-\mu}{\sigma} \xrightarrow{d} \text{Normal}(0,1),$$

as desired. ✓

Gumbel for star network Return

Proposition

The distribution of fixation times T for a star network of N spokes is

$$\frac{T - E(T)}{N^2} \xrightarrow{d} Gumbel(-\gamma, 1).$$

Gumbel for star network (I) Return

WLOG, the hub will always be the position of the first invader.

Gumbel for star network (II) Return

Given this, the probability of adding a new invader is

$$\begin{aligned}&(\text{probability of choosing the hub})\\ &\times(\text{probability of then replacing a resident spoke})\end{aligned}$$

Gumbel for star network (III) Return

Given m spokes are invaders, the probability of adding a new invader is

$$p_m = \frac{1}{m+1} \cdot \frac{N-m}{N},$$

for $m = 0, 1, \ldots, N-1$.

Gumbel for star network (IV) Return

By using Prop 1 and replacing L with N^2, we get

$$\begin{aligned}
\frac{T - E(T)}{N^2} &= \sum_{m=0}^{N-1} \frac{\text{Geo}(p_m) - 1/p_m}{N^2} \\
&\sim \sum_{m=0}^{N-1} \frac{\mathcal{E}(p_m) - 1/p_m}{N^2} \\
&=: \frac{T_S - E(T)}{N^2}
\end{aligned}$$

Gumbel for star network (V) Return

To complete this proof, we are going to compare the time to complete a star network to the time it takes to complete a coupon collection, which is defined as

$$T_C := \sum_{k=1}^{N} \mathcal{E}\left(\frac{m}{N}\right)$$

Take the characteristic functions of both
Our characteristic functions:

$$\phi_S = E\left[e^{it\frac{T_S-\mu}{N^2}}\right] = \prod_{k=1}^{N} \frac{\exp - it/(N^2 p_k)}{1 - it/(N^2 p_k)}$$

$$\phi_C = E\left[e^{it\frac{T_C-\mu}{N}}\right] = \prod_{k=1}^{N} \frac{\exp - it/k}{1 - it/k}$$

Gumbel for star network (VII) Return

Taking the ratio gives

$$\frac{\phi_C}{\phi_S} = \prod_{k=1}^{N} \exp\left[\frac{-it}{N}\left(1-\frac{1}{k}\right) + \log\left(1+\frac{it}{N}\frac{k-1}{k-it}\right)\right].$$

Gumbel for star network (VIII) Return

Via Taylor expansion at large N,

$$\frac{\phi_C}{\phi_S} = \exp\left[\frac{it}{N}\sum_{k=1}^{N}\frac{it(k-1)}{k(k-it)} + \frac{t^2}{2}\sum_{k=1}^{N}\frac{R_m}{N^2}\left(\frac{k-1}{k-it}\right)^2\right].$$

Gumbel for star network (IX) Return

$$\frac{\phi_C}{\phi_S} = \exp\left[\frac{it}{N}\sum_{k=1}^{N}\frac{it(k-1)}{k(k-it)} + \frac{t^2}{2}\sum_{k=1}^{N}\frac{R_m}{N^2}\left(\frac{k-1}{k-it}\right)^2\right]$$

- The first sum is bounded by $O(\log(N))$.
- The residual function R_m similarly is too small compared to N^2.

Gumbel for star network (X) Return

Therefore,

$$\frac{\phi_C}{\phi_S} \xrightarrow{N\to\infty} 1,$$

and so

$$\frac{T_S - \mu}{N^2} \sim \frac{T_C - \mu}{N}.$$

Gumbel for star network (XI) Return

By applying The Coupon Collector we get

$$\frac{T - E(T)}{N^2} \xrightarrow{d} \text{Gumbel}(-\gamma, 1),$$

as desired. ✓

2D as normal Return

Proposition

Given

$$q_m := \frac{\min(m, N-m)^{1/2}}{m}$$

and

$$T := \sum_{m=1}^{N-1} Geo(q_m),$$

then given $\mu := E[T]$ *and* $\sigma^2 := Var(T)$, *then*

$$\frac{T-\mu}{\sigma} \xrightarrow{d} Normal(0,1).$$

2D as normal (I) Return

The first step is to use Prop 2 to switch from geometric to exponential random variables, so

$$\frac{T-\mu}{\sigma} \sim \frac{T_E-\mu}{\mathrm{Var}(T_E)^{1/2}},$$

where $T_E = \sum_{m=1}^{N-1} \mathcal{E}(q_m)$.

2D as normal (II) Return

Next, we split the sum into two halves

$$T_E = T_a + T_b := \sum_{m=1}^{N/2-1} \mathcal{E}(q_m) + \sum_{m=N/2}^{N-1} \mathcal{E}(q_m),$$

N even WLOG.

2D as normal (III) Return

We are going to be applying the Normality Condition to both of T_a and T_b.

2D as normal (IV) Return

This involves satisfying two conditions:

- $q_m^2 \mathrm{Var}(T_x) \xrightarrow{N\to\infty} \infty$
- $\sum_m \exp\left(-\epsilon q_m \sqrt{\mathrm{Var}(T_x)}\right) \xrightarrow{N\to\infty} 0$

2D as normal (V) Return

First, we are doing the first half T_a,

$$\text{Var}(T_a) = \sum_{m=1}^{N/2-1} q_m^{-2} = \sum_{m=1}^{N/2-1} m = \frac{(N/2-1)^2 + (N/2-1)}{2}.$$

2D as normal (VI) Return

Therefore,

$$q_m^2 \mathrm{Var}(T_a) = \frac{1}{m}\frac{(N/2-1)^2+(N/2-1)}{2}$$
$$\geq \frac{N}{16} \to \infty$$

for large N, so the first condition is satisfied.

2D as normal (VII) Return

To account for the second condition, we use the asymptotic inequalities $q_m^2 \mathrm{Var}(T_a) > N/8$ and $q_m > 1/\sqrt{N}$. So

$$\sum_{m=1}^{N/2-1} \exp\left(-\epsilon q_m \sqrt{\mathrm{Var}(T_a)}\right) \leq N \exp\left(-\epsilon \sqrt{N}/8\right) \to 0. \checkmark$$

So T_a is distributed according to a normal.

2D as normal (VIII) Return

The second half is similar. First condition:

$$\begin{aligned}\mathrm{Var}(T_b) &= \sum_{k=1}^{N/2} q_{N-k}^{-2} = \sum_{k=1}^{N/2} \left(\frac{N-m}{\sqrt{k}}\right)^2 \\ &= N^2 \left(\sum_{k=1}^{N/2} \frac{1}{k} - \frac{2}{N} \sum_{k=1}^{N/2} 1 + \frac{1}{N^2} \sum_{k=1}^{N/2} k\right) \\ &\geq \frac{N^2}{4} \log(N).\end{aligned}$$

2D as normal (IX) Return

Therefore

$$q_k^2 \mathrm{Var}(T_b) \geq \frac{k}{(N-k)^2} \frac{N^2}{4} \log(N) \geq \frac{1}{4} \log N \xrightarrow{N \to \infty} \infty,$$

which satisfies the first condition.

2D as normal (X) Return

For the second condition:

$$\sum_{k=1}^{N/2} \exp\left(-\epsilon q_k \sqrt{\mathrm{Var}(T_b)}\right) \leq \sum_{k=1}^{N} \exp\left(-\frac{\epsilon}{2}\frac{\sqrt{k}}{N-k} N\sqrt{\log(N)}\right)$$
$$\leq \sum_{k=1}^{N} \exp\left(-\frac{\epsilon}{2}\sqrt{k\log(N)}\right)$$
$$\leq \int_0^\infty \exp\left(-\frac{\epsilon}{2}\sqrt{x\log(N)}\right) dx$$
$$= \frac{8}{\epsilon^2 \log(N)} \xrightarrow{N\to\infty} 0. \checkmark$$

2D as normal (XI) Return

Therefore, both T_a and T_b are normals. The sum of two normals is a normal, so

$$\frac{T - \mu}{\sigma} \xrightarrow{d} \text{Normal}(0, 1),$$

as desired. ✓

Skew for $d \geq 3$ Return

Proposition

Given $\eta = 1 - 1/d$ *and*

$$q_m = \frac{\min(m, N-m)^\eta}{m}$$

and

$$T := \sum_{m=1}^{N-1} Geo(q_m),$$

then, letting ζ *be the usual Riemann zeta function,*

$$Skew(T) \to \frac{2\zeta(3\eta)}{\zeta(2\eta)^{3/2}}.$$

Skew for $d \geq 3$ (I) Return

The first step is to use Prop 2 to switch from geometric to exponential random variables, so

$$\frac{T-\mu}{\sigma} \sim \frac{T_E-\mu}{\mathrm{Var}(T_E)^{1/2}},$$

where $T_E = \sum_{m=1}^{N-1} \mathcal{E}(q_m)$.

Skew for $d \geq 3$ (II) Return

We will now split up T_E into front and back halves, so

$$T := T_a + T_b := \sum_{m=1}^{N/2-1} \mathcal{E}(q_m) + \sum_{m=N/2}^{N-1} \mathcal{E}(q_m).$$

Skew for $d \geq 3$ (III) Return

To find which part has the larger contribution, we calculate the variance of each half. So first there's

$$\mathrm{Var}(T_a) = \sum_{m=1}^{N/2-1} q_m^{-2} = \sum_{m=1}^{N/2-1} m^{2/d} \leq \int_0^{N/2} x^{2/d} dx \leq N^{2/d+1}.$$

Skew for $d \geq 3$ (IV) Return

Then, since $\eta \geq 2/3$, the second half has

$$\begin{aligned}
\mathrm{Var}(T_b) &= \sum_{k=1}^{N/2} q_{N-k}^{-2} = \sum_{k=1}^{N/2} \frac{(N-k)^2}{k^{2\eta}} \\
&= N^2 \left(\sum_{k=1}^{N/2} \frac{1}{k^{2\eta}} - \frac{2}{N} \sum_{k=1}^{N/2} k^{1-2\eta} + \frac{1}{N^2} \sum_{k=1}^{N/2} k^{2/d} \right) \\
&\to N^2 \zeta(2\eta)
\end{aligned}$$

Skew for $d \geq 3$ (V) Return

To find the skew of T, we use the Skew Lemma to find

$$\begin{aligned} \text{Skew}(T) &= \frac{\text{Skew}(T_a)\text{Var}(T_a)^{3/2} + \text{Skew}(T_b)\text{Var}(T_b)^{3/2}}{(\text{Var}(T_a) + \text{Var}(T_b))^{3/2}} \\ &= \frac{\text{Skew}(T_a)(\text{Var}(T_a)/\text{Var}(T_b))^{3/2} + \text{Skew}(T_b)}{(1 + \text{Var}(T_a)/\text{Var}(T_b))^{3/2}} \\ &\to \text{Skew}(T_b). \end{aligned}$$

since $2 > 2/d + 1$ for $d \geq 3$, and so the variance from T_b dominates that from T_a.

Skew for $d \geq 3$ (VI) Return

(Sidebar: it is also possible to use the Normal Condition to show T_a goes to a normal, and therefore contributes no skew.)

Skew for $d \geq 3$ (VII) Return

Reuse the lemma to find

$$\text{Skew}(T_b) = \frac{\sum_{k=1}^{N/2} 2q_{N-k}^{-3}}{\left(\sum_{k=1}^{N/2} q_{N-k}^{-2}\right)^{3/2}}.$$

Skew for $d \geq 3$ (VIII) Return

The denominator just limits to $N^3\zeta(2\eta)^{3/2}$, whereas the numerator goes to

$$\sum_{k=1}^{N/2} 2q_{N-k}^{-3} = 2\sum_{k=1}^{N/2} \frac{1}{k^{3\eta}}(N-k)^3$$

$$= 2N^3\left(\sum_{k=1}^{N/2}\frac{1}{k^{3\eta}} - \frac{3}{N}\sum_{k=1}^{N/2}k^{1-3\eta} + \frac{3}{N^2}\sum_{k=1}^{N/2}k^{2-3\eta} + \frac{1}{N^3}\sum_{k=1}^{N/2}k^{3/d}\right)$$

$$\to 2N^3\zeta(3\eta)$$

Skew for $d \geq 3$ (IX) Return

Therefore,

$$\text{Skew}(T) = \frac{2\zeta(3\eta)}{\zeta(2\eta)^{3/2}}$$

as desired. ✓

Random walk skew Return

Proposition

The time for a simple random walk starting at 1 to hit N before hitting 0 has an asymptotic skew of ≈ 1.807.

Random walk skew (I) Return

Let's represent a random walk of n steps by

$$X_n = 1 + \sum_{i=1}^{n} x_i$$

where $x_i \in \{-1, +1\}$, each with probability $1/2$.

Random walk skew (II) Return

To account for the hitting time, define the stopping time

$$S_m = \min\{n | X_n = m\},$$

so the first time the random walk hits m.

Random walk skew (III) Return

To include the absorbing states at 0 and N, the walk's stopping time is

$$S = \min(S_0, S_N).$$

Random walk skew (IV) Return

To find the conditioned skew, define the conditioned moments by

$$\mu_i := E(S^i | X_S = N),$$

for $i = 1, 2, 3$.

Random walk skew (V) Return

To find these, we are going to use *martingale theory*. A martingale is a sequence of random variables M_n and filter of sigma fields $\mathcal{F}_N$ such that, among other things,

$$E(M_{n+1}|\mathcal{F}_n) = M_n.$$

Random walk skew (VI) Return

For this $\mathcal{F}_n$ will be the sigma field consisting of all information from the first n steps of the random walk. Therefore,

$$E(x_{n+1}|\mathcal{F}_n) = 0$$
$$E(x_{n+1}^2|\mathcal{F}_n) = 1.$$

Random walk skew (VII) Return

The first martingale we will define is

$$M_n^{(1)} := X_n^3 - 3nX_n.$$

Random walk skew (VIII) Return

Checking the condition gives

$$\begin{aligned} E(M^{(1)}_{n+1}|\mathcal{F}_n) =& E\left(X^3_{n+1} - 3(n+1)X_{n+1}|\mathcal{F}_n\right) \\ =& E\left((X_n + x_{n+1})^3 - 3(n+1)(X_n + x_{n+1})|\mathcal{F}_n\right) \\ =& E\left(X^3_n + 3x_{n+1}X^2_n + 3x^2_{n+1}X_n \right.\\ & \left. -3(n+1)X_n - 3(n+1)x_{n+1}|\mathcal{F}_n\right) \\ =& X^3_n - 3nX_n \\ =& M^{(1)}_n. \end{aligned}$$

So $M^{(1)}_n$ well-approximates its future, meaning that it is a proper martingale.

Random walk skew (IX) Return

Therefore, we can cite Optional Stopping to get

$$E\left(M_0^{(1)}\right) = E\left(M_S^{(1)}\right).$$

Random walk skew (X) Return

The left-hand side is easy, since we start at $n = 0$ and $X_0 = 1$,

$$E\left(M_0^{(1)}\right) = 1^3 - 3 \cdot 0 \cdot 1 = 1.$$

Random walk skew (XI) Return

Because stopping occurs either at $X_S = 0$ or $X_S = N$, the right-hand side becomes

$$\begin{aligned}
E\left(M_S^{(1)}\right) &= P(X_S = N)E(M_S^{(1)}|X_S = N) + P(X_S = 0)E(M_S^{(1)}|X_S = 0) \\
&= \frac{1}{N}E\left(X_n^3 - 3nX_n|X_S = N\right) + \frac{N-1}{N}E\left(X_n^3 - 3nX_n|X_S = 0\right) \\
&= \frac{1}{N}\left(N^3 - 3\mu_1 N\right) + \frac{N-1}{N}E\left(0^3 - 3E(S|X_S = 0)0\right) \\
&= N^2 - 3\mu_1.
\end{aligned}$$

Random walk skew (XII) Return

Solve to get

$$\mu_1 = \frac{N^2 - 1}{3}.$$

Random walk skew (XIII) Return

Getting the next two moments is the same basic calculation. First, check the martingale property for both of

$$M_n^{(2)} = X_n^5 - 10nX_n^3 + (15n^2 + 10n)X_n$$
$$M_n^{(3)} = X_n^7 - 21nX_n^5 + (105n^2 + 70n)X_n^3 - (105n^3 + 210n^2 + 112n)X_n.$$

Random walk skew (XIV) Return

Use optional stopping to get conditions and reveal the next two moments, which are given by

$$\mu_2 = \frac{7N^4 - 20N^2 + 13}{45}$$
$$\mu_3 = \frac{31N^6 - 147N^4 + 189N^2 - 73}{315}.$$

Random walk skew (XV) Return

Therefore!

$$\text{Skew} = \frac{\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3}{\left(\mu_2 - \mu_1^2\right)^{3/2}} \xrightarrow{N\to\infty} \frac{8}{7}\left(\frac{5}{2}\right)^{1/2} \approx 1.807.$$

So conditioning a random walk introduces large skew. ✓

A common misconception Return

Figure: Histograms of incubation period distribution given a basic exponential growth model with random parameters. Each plot represents 10^5 simulations.

Figure: (a) In the Birth-death (Bd) update rule, a node anywhere in the network is selected at random, with probability proportional to its fitness, and one of its neighbors is selected at random, uniformly. (b) The neighbor takes on the type of the first node. In biological terms, one can interpret this rule in two ways: either the first node transforms the second; or it gives birth to an identical offspring that replaces the second. (c) In the Death-birth (Db) update rule, a node is selected at random to die, with probability inversely proportional to its fitness, and one of its neighbors is selected at random, uniformly, to give birth to one offspring. (d) The first node is replaced by the offspring of the second.

Complete graph and star graph Return

Figure: Simulated distributions of invader fixation times. Starting from a single invader at a random node, the state of the network was updated by Birth-death dynamics on both a complete graph and a two-dimensional (2D) lattice. Results for the Death-birth update rule (not shown) are identical. All distributions are normalized to have zero mean and unit variance. (a) Infinitely fit invader. For invader fitness $r \to \infty$, the distribution is right-skewed for a complete graph (blue symbols). It approaches a Gumbel distribution as $N \to \infty$, where N is the number of nodes in the network. In contrast, for a 2D lattice (red symbols) the incubation periods are normally distributed. Simulations used 10^6 repetitions on a complete graph of $N = 150$ nodes, and 10^5 repetitions for a 2D lattice of $N = 30^2$ nodes. (b) Neutrally fit invader. Distributions of incubation periods are shown for invader fitness $r = 1$, using 10^6 repetitions on a complete graph of $N = 50$ nodes (blue symbols), and 10^5 repetitions for a 2D lattice of $N = 7^2$ nodes (red symbols).

Figure: Simulated, fitted, and normalized distributions of incubation periods for Birth-death dynamics on a complete graph of $N = 500$ nodes. Unless stated otherwise, each simulation used an invader fitness of $r = 10$, measured times till complete takeover ($f = 1$), and started from an initial dose of 1 invader. Runs where the dosage was not smaller than the truncation point were rejected. The blue curves indicate noncentral lognormals fitted via the method of moments. (a) Heterogeneous fitness of invader. Every run used a different r selected from a Gamma distribution with a shape parameter of 10. (b) Heterogeneity of host response. Instead of waiting until all N residents had been replaced by invaders, every run used a different truncation point uniformly selected from $\{2, 3, \ldots N\}$. (c) Heterogeneity of dosage. Every run had a different starting population drawn from a Poisson of mean 10 and a shift of 1. (d) Heterogeneity of invader fitness, host response, and dosage. Every run used an r drawn from Gamma(10), a truncation point f drawn from Uniform(0,1), and a dosage drawn from Poisson(10)+1.

Complete graph and star graph Return

Figure: Distributions of invader fixation times, normalized to have zero mean and unit variance, are shown for infinite-r Birth-death dynamics on various networks. Open circles show simulation results. Curves show analytical predictions. Insets show schematics of networks. (a) The distribution of fixation times for a complete graph on $N = 150$ nodes, for 10^6 runs. (b) The distribution of fixation times for a star graph with $N = 75$ spokes, for 10^6 runs.

Lattices: $d = 1$ and $d = 2$ Return

Figure: Distributions of invader fixation times, normalized to have zero mean and unit variance, are shown for infinite-r Birth-death dynamics on various networks. Open circles show simulation results. Curves show analytical predictions: blue curves are Gumbels, red are normals, and green is an intermediate distribution. Insets show schematics of networks. (a) The distribution of fixation times for a 1D ring on $N = 75$ nodes, for 10^6 runs. (b) The distribution of fixation times for a 2D lattice of $N = 60 \times 60$ nodes, for 10^5 runs.

Summary: infinite r Return

Figure: Distributions of invader fixation times, normalized to have zero mean and unit variance, are shown for infinite-r Birth-death dynamics on various networks. Open circles show simulation results. Curves show analytical predictions: blue curves are Gumbels, red are normals, and green is an intermediate distribution. Insets show schematics of networks. (a) The distribution of fixation times for a complete graph on $N = 150$ nodes, for 10^6 runs. (b) The distribution of fixation times for a star graph with $N = 75$ spokes, for 10^6 runs. (c) The distribution of fixation times for a 1D ring on $N = 75$ nodes, for 10^6 runs. (d) The distribution of fixation times for a 2D lattice of $N = 60 \times 60$ nodes, for 10^5 runs. (e) The distribution of fixation times for a 3D lattice of $N = 11^3$ nodes, for 10^5 runs. The predicted distribution is the result of an approximating sum of exponential random variables under 10^6 repetitions. (f) The distribution of fixation times for an Erdős-Rényi random graph on $N = 115$ nodes with an edge probability of $\rho = 0.5$.

Summary: $r = 1$ Return

Figure: Simulated and fitted distributions of invader fixation times are shown for Birth-death dynamics on various networks. All distributions were normalized to have mean zero and unit variance. The curves indicate noncentral lognormals fitted via the method of moments. (a) Complete graph on $N = 50$ nodes, for 10^6 runs. (b) Star graph with $N = 25$ spokes, for 10^6 runs. (c) One-dimensional ring on $N = 50$ nodes, for 10^6 runs. (d) Two-dimensional lattice on $N = 7 \times 7$ nodes, for 10^6 runs. (e) Three-dimensional lattice on $N = 4^3$ nodes for 10^6 runs. (f) Erdős-Rényi random graph on $N = 25$ nodes with an edge probability of $\rho = 0.5$.

Figure: Distribution of incubation periods for both Birth-death (Bd) and Death-birth (Db) dynamics, using a complete graph of $N = 5000$ nodes, and a infinite invader fitness. Incubation periods are now defined as times needed for invaders to take over a fraction f of the whole network. All distributions are normalized to have zero mean and unit variance. Data points are color-coded according to the nature of the distribution: blue indicates a Gumbel distribution, and red indicates a normal distribution.

Figure: Simulated and fitted distributions of invader fixation times for Birth-death dynamics on small-world, scale-free, and k-regular networks. All distributions were normalized to have mean zero and unit variance. The curves indicate non-central lognormals fitted to the first three moments of the data. All distributions are the result of 10^6 simulations. The figures in the top row ((a), (b), (c)) used invader fitness $r = \infty$, whereas the figures in the bottom row ((d), (e), (f)) used neutral fitness $r = 1$. (a) Newman-Watts-Strogatz small-world ring network with shortcut probability of $\rho = 0.25$ on $N = 75$. (b) Random 3-regular graph on $N = 100$ nodes. (c) Barabasi-Albert scale-free network with a minimum degree of 3 and $N = 100$ nodes. (d) Newman-Watts-Strogatz small-world ring network with shortcut probability of $\rho = 0.25$ on $N = 25$ nodes. (e) Random 3-regular graph on $N = 22$ nodes. (f) Barabasi-Albert scale-free network with a minimum degree of 3 and $N = 22$ nodes.

Variable population Return

Figure: Simulated, fitted, and normalized distributions of incubation periods for Birth-death dynamics on a complete graph that initially has $N = 500$ nodes. Invader fitness is set at $r = 10$. The blue curves indicate noncentral lognormals fitted via the method of moments. (a) Constant total population. (b) Growing population. At every time step, there is a constant $1/N$ chance that a new resident node will appear. The new node is adjacent to all preexisting nodes. (c) Shrinking population. At every time step, there is a constant $1/N$ chance that a random resident node will be removed. (d) Randomly varying population. At every time step, a resident node is either added or removed from the population, both events occurring with probability $1/2$.