The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2584-13

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

A Genome as a Toolbox: CRP and Mean-Field

Marco Cosentino Lagomarsino
Université Pierre et Marie Curie
Paris

A Genome as a Toolbox: CRP and Mean-Field

June 3nd 2014 (afternoon)

Spring School, Trieste

Marco Cosentino Lagomarsino

Génophysique / Genomic Physics Group

CNRS “Microorganism Genomics” UMR7238 Laboratory
Université Pierre et Marie Curie, Paris

1) Mean-field calculations with the CRP

Duplication / Innovation Model

i. **Duplication of an existing domain**

ii. **Innovation, genesis or transfer of a domain**

Duplication / Innovation Model

i. **Duplication of an existing domain**

ii. **Innovation, genesis or transfer of a domain**

Duplication / Innovation Model

i. **Duplication of an existing domain**

ii. **Innovation, genesis or transfer of a domain**

Duplication / Innovation Model

i. **Duplication of an existing domain**

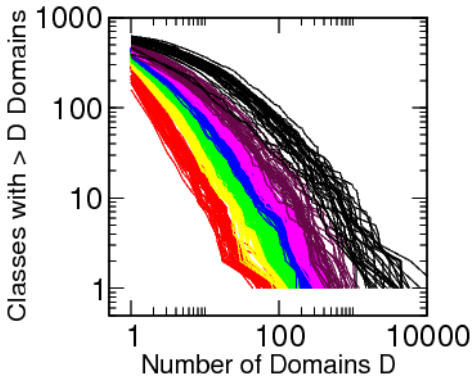
ii. **Innovation, genesis or transfer of a domain**

Duplication / Innovation Model

i. **Duplication of an existing domain**

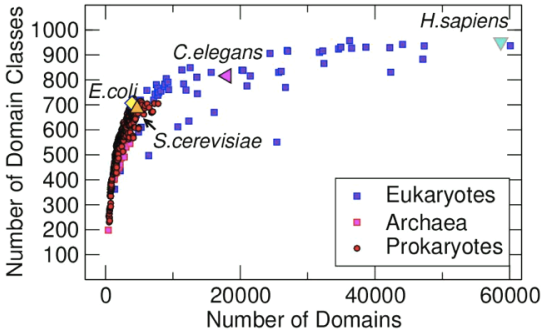
ii. **Innovation, genesis or transfer of a domain**

Duplication / Innovation Model



i. **Duplication of an existing domain**

ii. **Innovation, genesis or transfer of a domain**



Requirements

A

$$p_O + p_N = 1$$

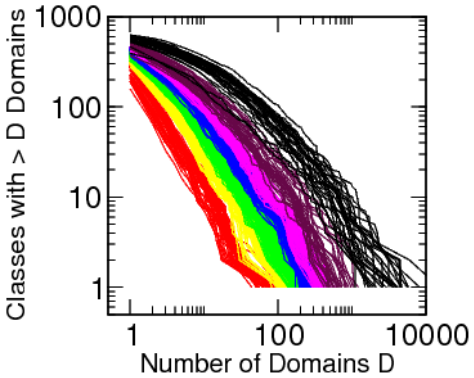
$$p_O = \sum_{i \in \text{classes}} p_O^i$$

normalizations

B

$$p_O^i \sim \frac{n_i}{n}$$

(uniform = preferential attachment)

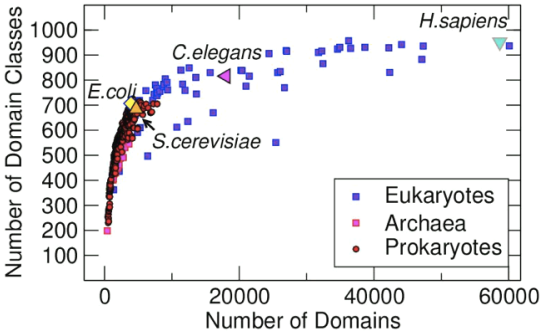


C

p_N *NOT constant in F, n*

$$p_N \sim \frac{1}{n}$$

$$p_N \sim \frac{F}{n}$$

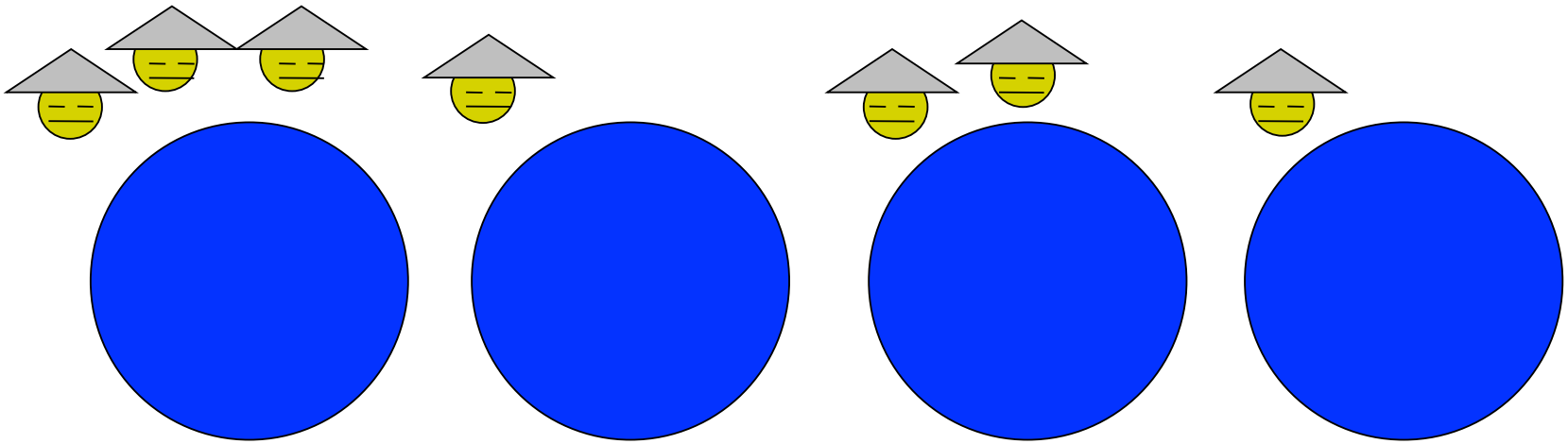


Simplest Case

$$p_O = \frac{n - f\alpha}{n + \theta} \qquad p_N = \frac{\theta + f\alpha}{n + \theta}$$

$$\theta > -\alpha$$
$$0 \leq \alpha \leq 1$$

Chinese Restaurant Process



Exercise: write a simulation of this process

$$p_O = \frac{n-f\alpha}{n+\theta} \qquad p_N = \frac{\theta+f\alpha}{n+\theta} \qquad \begin{array}{l}\theta > -\alpha \\ 0 \le \alpha \le 1\end{array}$$

Plot some realizations of
f(n) (#families)
f(j,n) (#families with j members)

Mean-field for families

$$p_O = \frac{n - f\alpha}{n + \theta} \qquad p_N = \frac{\theta + f\alpha}{n + \theta}$$

$$\theta > -\alpha$$
$$0 \leq \alpha \leq 1$$

$$p_O^{(i)} = \frac{n_i - \alpha}{n + \theta}$$

$$\frac{d\langle n_i \rangle}{dn} = p_O^{(i)}(\langle n_i \rangle)$$

$$\frac{d\langle f \rangle}{dn} = p_N$$

Mean-field for families

$$p_O = \frac{n - f\alpha}{n + \theta} \qquad p_N = \frac{\theta + f\alpha}{n + \theta} \qquad \begin{array}{l} \theta > -\alpha \\ 0 \leq \alpha \leq 1 \end{array}$$

$$p_O^{(i)} = \frac{n_i - \alpha}{n + \theta}$$

$$\langle n_i \rangle \sim \langle n_{0,i} \rangle \frac{n}{n_0}$$

$\alpha \neq 0$

$$F(n) = \frac{1}{\alpha} \left[(\alpha + \theta) \left(\frac{n+\theta}{\theta} \right)^{\alpha} - \theta \right] \sim n^{\alpha}$$

$\alpha = 0$

$$F(n) = \theta \log \left(\frac{n+\theta}{\theta} \right) \sim \theta \log \frac{n}{\theta}$$

Estimating the histogram /1

$$\langle n_i \rangle \sim \langle n_{0,i} \rangle \frac{n}{n_0}$$

Hence (drop the <>) $j > n_i$ if $\quad n_0 > n^* \sim \frac{n}{j}$

The cumulative histogram at size *n* is estimated by the ratio #(families born before *n**) / # (families)

$$P(n_i > j) \simeq \frac{F(n^*)}{F(n)} \sim \frac{1}{j^\alpha} \quad \alpha \neq 0$$

$$\sim 1 - \log(j) \quad \alpha = 0$$

Estimating the histogram /2

$$p_O = \frac{n - f\alpha}{n+\theta} \qquad p_N = \frac{\theta + f\alpha}{n+\theta} \qquad \begin{matrix} \theta > -\alpha \\ 0 \le \alpha \le 1 \end{matrix}$$

Classes with

1 *2* *3* *members etc.*

$F(1,n) \longrightarrow F(2,n) \longrightarrow F(3,n) \longrightarrow \ldots$

$p_O^{(1)}$ $p_O^{(2)}$ $p_O^{(3)}$

p_N

$$p_O^{(i)} = \frac{n_i - \alpha}{n+\theta}$$

Hierarchy of "rate" Equations

(same "master equation" procedure as zero-range process, Barabasi-Albert model, etc.)

$$\begin{cases} \partial_n F(n) = & \frac{\alpha F(n) + \theta}{n+\theta} \\ \partial_n F(1,n) = & \frac{\alpha F(n) + \theta}{n+\theta} - (1-\alpha)\frac{F(1,n)}{n+\theta} \\ \partial_n F(2,n) = & (1-\alpha)\frac{F(1,n)}{n+\theta} - (2-\alpha)\frac{F(2,n)}{n+\theta} \\ \cdots & \cdots \end{cases}$$

Estimating the histogram /2

Ansatz (Large n)

$$F(j,n) \approx \chi_j F(n)$$

$$\begin{cases} \partial_n F(n) = & \dfrac{\alpha F(n)}{n} \\ \alpha\chi_1 = & \alpha - (1-\alpha)\chi_1 \\ \alpha\chi_2 = & (1-\alpha)\chi_1 - (2-\alpha)\chi_2 \\ \cdots & \\ \alpha\chi_j = & (j-1-\alpha)\chi_{j-1} - (j-\alpha)\chi_j \end{cases}$$

Estimating the histogram /2

Ansatz $F(j,n) \approx \chi_j F(n)$

$$\begin{cases} \chi_1 = & \alpha \\ 2\chi_2 = & (1-\alpha)\chi_1 \\ \cdots & \\ j\chi_j = & (j-1-\alpha)\chi_{j-1} \end{cases}$$

$$\chi_j = \prod_{l=1}^{j-1} (l-\alpha)\frac{1}{\Gamma(j+1)}\alpha = \frac{\alpha}{\Gamma(1-\alpha)}(j-1)^{(1-\alpha)}\frac{\Gamma(j-1)}{\Gamma(j+1)}$$

$$P(j) = \chi_j = \alpha\frac{1}{\Gamma(1-\alpha)}\left[\frac{1}{j}\right]^{1+\alpha}$$

“Qian-Gerstein” process

Qian et al, JMB, 2001

$$p_O = (1 - r) \qquad p_N = r$$

$$p_O^{(i)} = (1 - r)\frac{n_i}{n}$$

“Qian-Gerstein” process

Qian et al, JMB, 2001

Problems:

Gives $F(n) \sim n$

Gives exponent > 2 for $F(j,n)$

Fit parameters by genome (no common trend detected)

2) General facts about the CRP

There is a lot of math (probability) literature about the CRP

Paradigm of “exchangeable” distribution

$$P(n_1, n_2, \ldots, n_f) = P(n_{\pi(1)}, n_{\pi(2)}, \ldots, n_{\pi(f)})$$

NB: independence implies exchangeability but not viceversa

(Pitman, st Flour 2006, for the hard-boiled)

Example: Polya urn

Urn with W_0 white balls and B_0 black ones. Iteratively,

1) Draw a ball

2) Place the ball back with *a* balls of the same color

Xi = 1 BLACK Xi = 0 WHITE

$$\mathbf{P}(1,1,0,1) = = \frac{B_0}{B_0+W_0} \times \frac{B_0+a}{B_0+W_0+a} \times \frac{W_0}{B_0+W_0+2a} \times \frac{B_0+2a}{B_0+W_0+3a}$$

$$\mathbf{P}(1,0,1,1) = = \frac{B_0}{B_0+W_0} \times \frac{W_0}{B_0+W_0+a} \times \frac{B_0+a}{B_0+W_0+2a} \times \frac{B_0+2a}{B_0+W_0+3a}$$

But the sequence $\{X_i, i \geq 1\}$ is not iid

De Finetti's theorem

Echangeable RVs are conditionally independent

For Polya (binary variables),
Mixture of Bernoulli with a "hidden variable"

$$\mathbf{P}(X_1 = x_1, ..., X_n = x_n) = \int_0^1 p^{S_n}(1-p)^{n-S_n} dF(p)$$

$$S_n = \sum_i x_i$$

… in this case the density turns out to be

$$\text{Beta}\left(\frac{B_0}{B_0 + W_0}, \frac{W_0}{B_0 + W_0}\right)$$

De Finetti's theorem

More in general

$$p(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} p(x_i|\theta) p(\theta) d\theta$$

For the CRP the basic distribution can be multinomial and the mixing one can be computed

Note: each class in the CRP behaves like a Polya Urn

Links of CRP with…

Ewens sampling formula

Neutral theory biodiversity

Stick-breaking process

Bayesian clustering

…

CRP limit theorems for number of families *F(n)*

For $\alpha = 0$ mean and variance of *F(n)* scale as $\theta \log \frac{n}{\theta}$

For $\alpha > 0$ $s = \frac{f(n)}{n^{\alpha}}$ asymptotically follows a

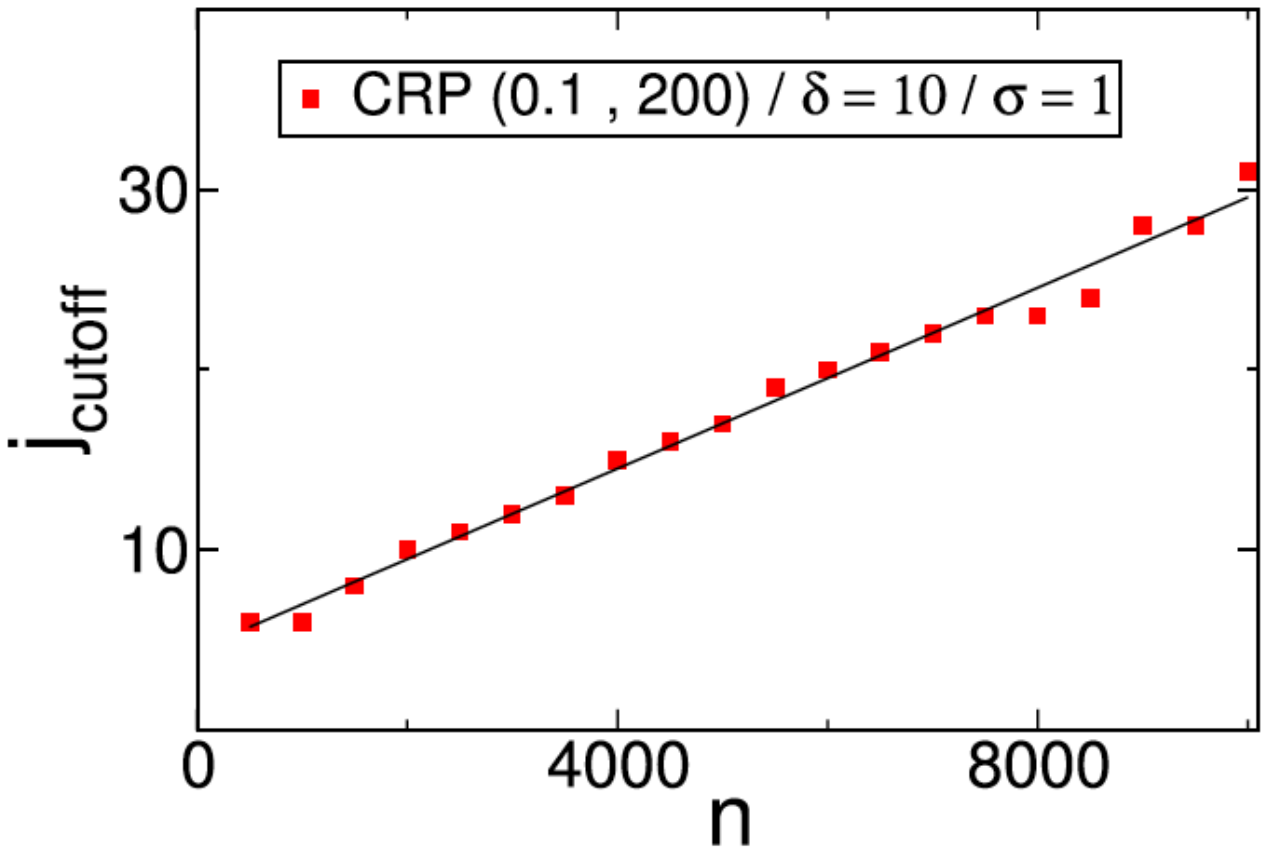
finite-variance distribution

=> No self-averaging

3) Finite-size effects

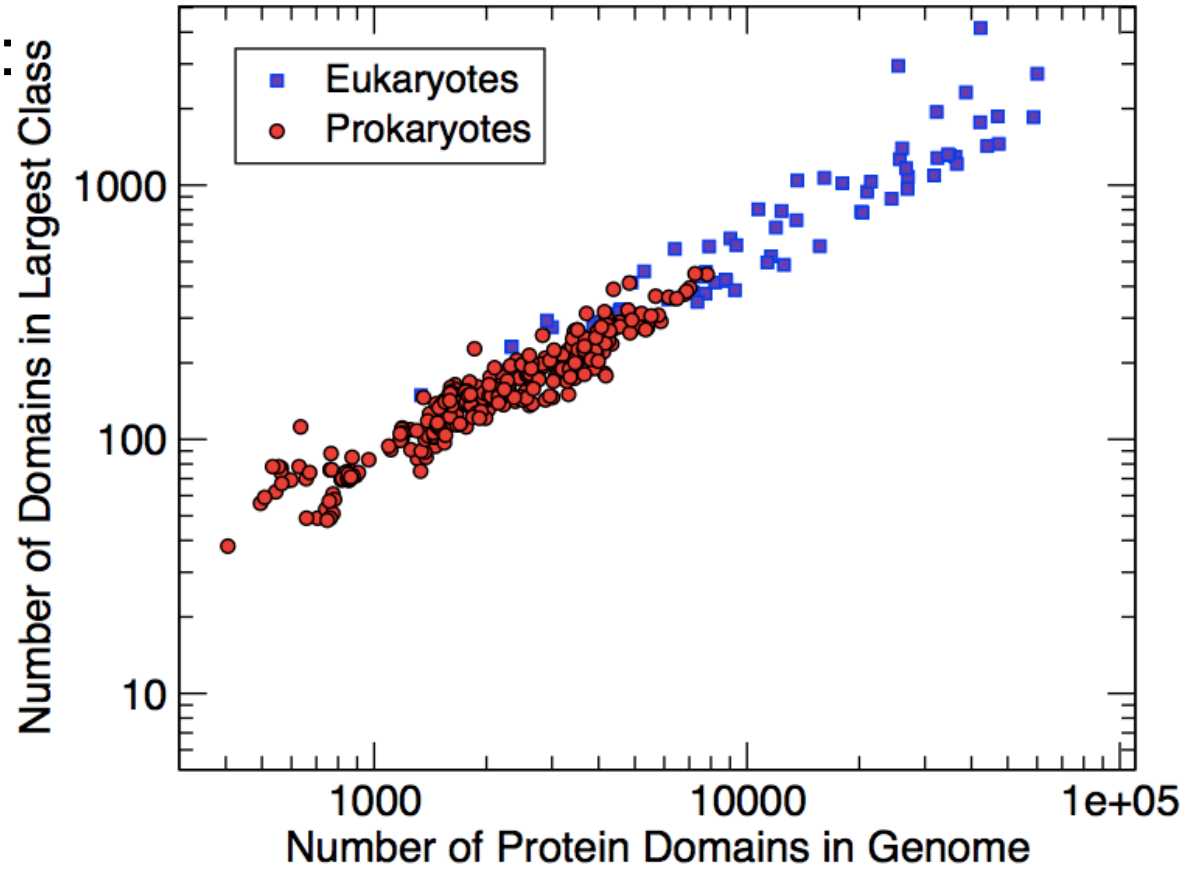
The cutoff of *F(j,n)* scales *linearly* with size

CRP



The cutoff of *F(j,n)* scales *linearly* with size

Genomes:

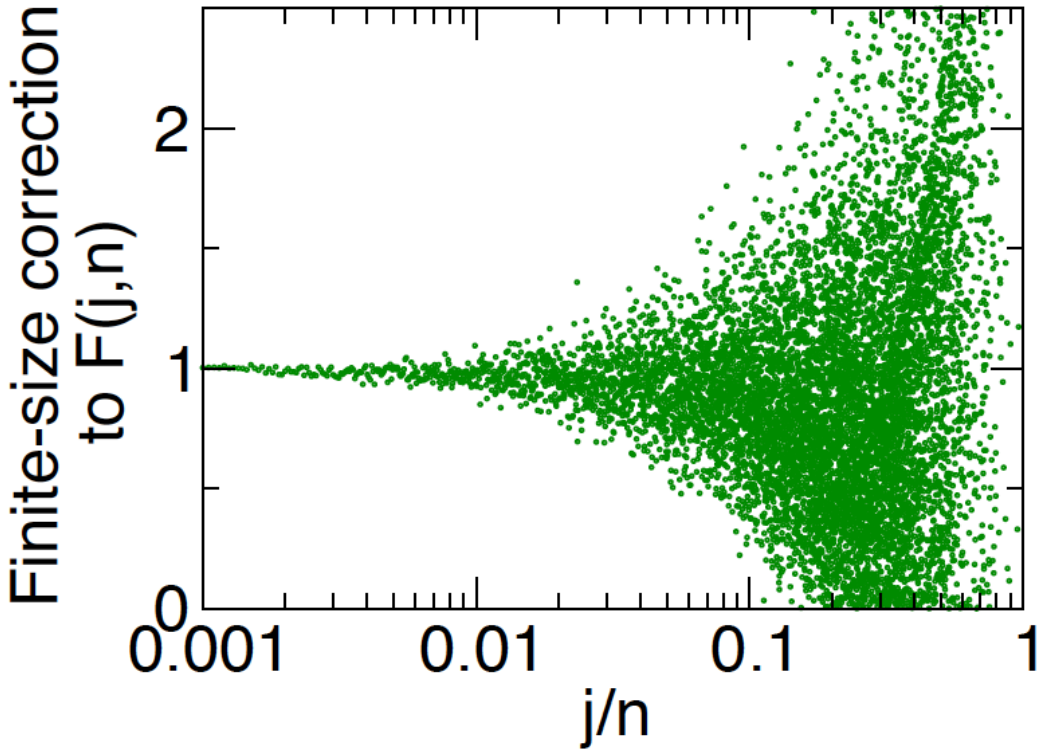


Not true for the “Qian-Gerstein” process

The CRP Has Anomalous Finite-size Effects

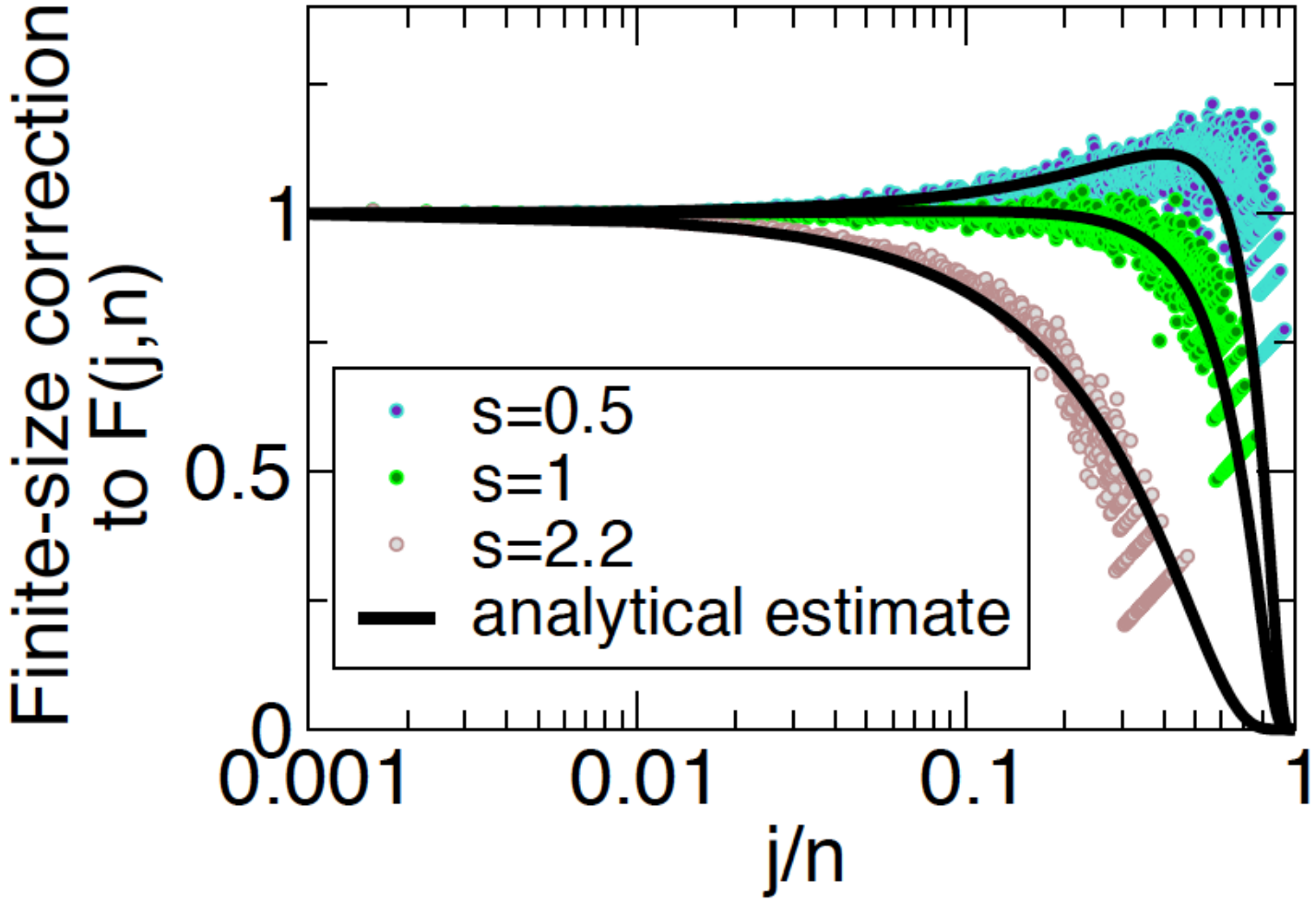
For $\alpha>0$ $F(n)/n^{\alpha}$ converges to a probability distribution
This corresponds to non-selfaveraging: *StDev[F(n)]/F(n)* diverges

The finite-size correction to *F(j,n)*,
F(j,n)/F(j,∞) is realization-dependent



The CRP Has Anomalous Finite-size Effects

The finite-size correction to *F(j,n)*
related to the realization-specific scaling of *F(n)*



Comparison

CRP

The cutoff scales linearly

Density diverges

Realization-dependent “bump”

Zero-range process ~ Qian-Gerstein

The cutoff scales sublinearly

Density may diverge or not, tunable in ZRP

High density gives condensation

4) Role of gene loss

Adding Uniform Domain Loss Does Not Affect the Scaling

$$p_O = (1-\delta)\frac{n - f\alpha}{n+\theta}$$

i. **Duplication of an existing domain**

ii. **Innovation, genesis or transfer of a domain**

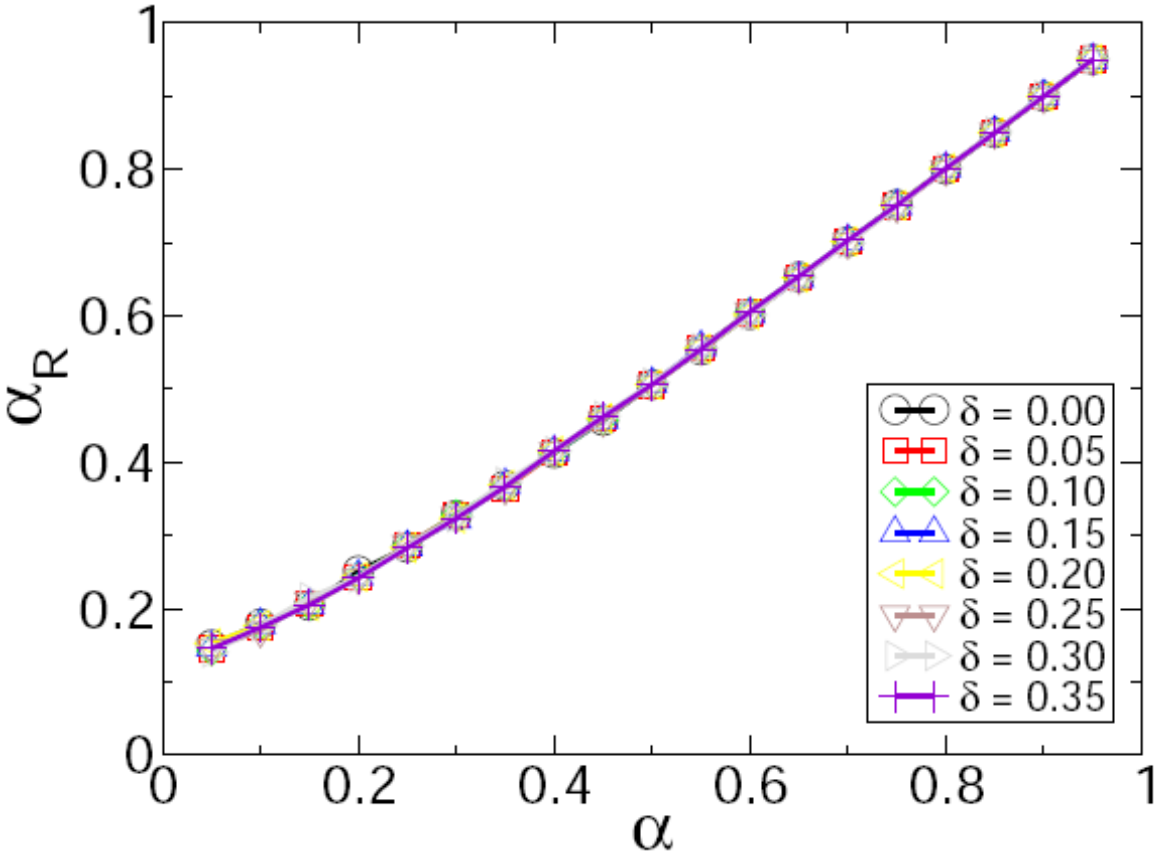
$$p_N = (1-\delta)\frac{\theta + f\alpha}{n+\theta}$$

$$p_L = \delta$$

iii. **Loss of an existing domain**

Adding Uniform Domain Loss Does Not Affect the Scaling

Observed Exponent



Value of the α parameter

A Model with Weighted Loss has Interesting Scaling Behavior

$$p_O = (1-\delta)\frac{n - f\alpha}{n+\theta}$$

i. **Duplication of an existing domain**

$$p_N = \frac{\theta + f\alpha}{n+\theta}$$

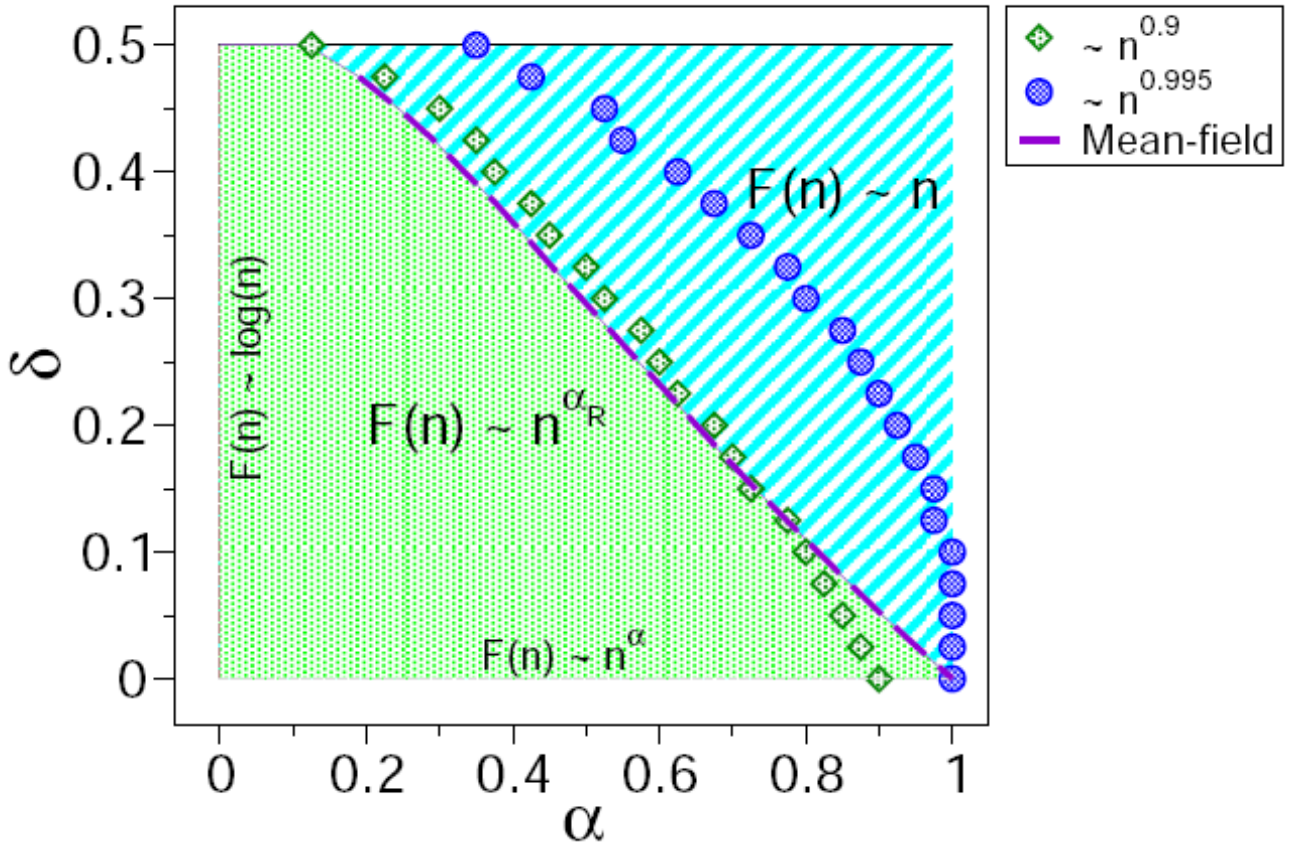
ii. **Innovation, genesis or transfer of a domain**

$$p_L = \delta\frac{n - f\alpha}{n+\theta}$$

iii. **Loss of an existing domain**

A Model with Weighted Loss has Interesting Scaling Behavior

"phase diagram" of *F(n)* scaling



Conclusions

- One can do a lot of simple estimates with this process
- Math literature gives all you need (and more!) for the basic CRP (but not the variants)
- Finite-size behavior interesting for statistical physics
- Gene loss affects the qualitative behavior only weakly