The threat of disinformation to epistemic security: an ~~exactly~~ solvable model

Who can you trust?

If you can't trust your ~~barber~~ president, who can you trust?

José F. Fontanari
Universidade de São Paulo
Brasil

supported by FAPESP and CNPq

Epistemic security: breakdown of trusted sources of information is one of the most pressing problems today.

truth-telling *vs*. lying

Groundwork of the Metaphysics of Morals, Kant (1785)

A world in which everyone tells the truth is possible, whereas one in which everyone lies is unthinkable - not in the sense that it would be bad, but in the sense that it cannot exist.

Batesian mimicry (1865)

Papilio polytes

mimic

Pachliopta aristolochiae

unpalatable

approach:

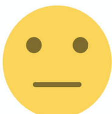
(evolutionary) game theory

+ quantitative genetics

evolutionary game-theoretic model

- **individual-environment interaction** w.p. $1 - w$ → trust

individual i



estimate ξ_i

$\xi_i \sim N(\mu, \sigma)$ → hazardousness

environment

μ

w.p. w

environmental challenge

probability that individual *i* survives the environmental challenge

$$S_i = \exp\left[-\frac{1}{2}\left(\xi_i - \mu\right)^2\right]$$

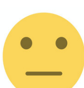
viability S

$$P(S) = \frac{1}{\sqrt{\pi\sigma^2}}\frac{S^{1/\sigma^2-1}}{\sqrt{-\ln S^{1/\sigma^2}}}$$

- **individual-individual interaction (copying)**

individual j

individual i



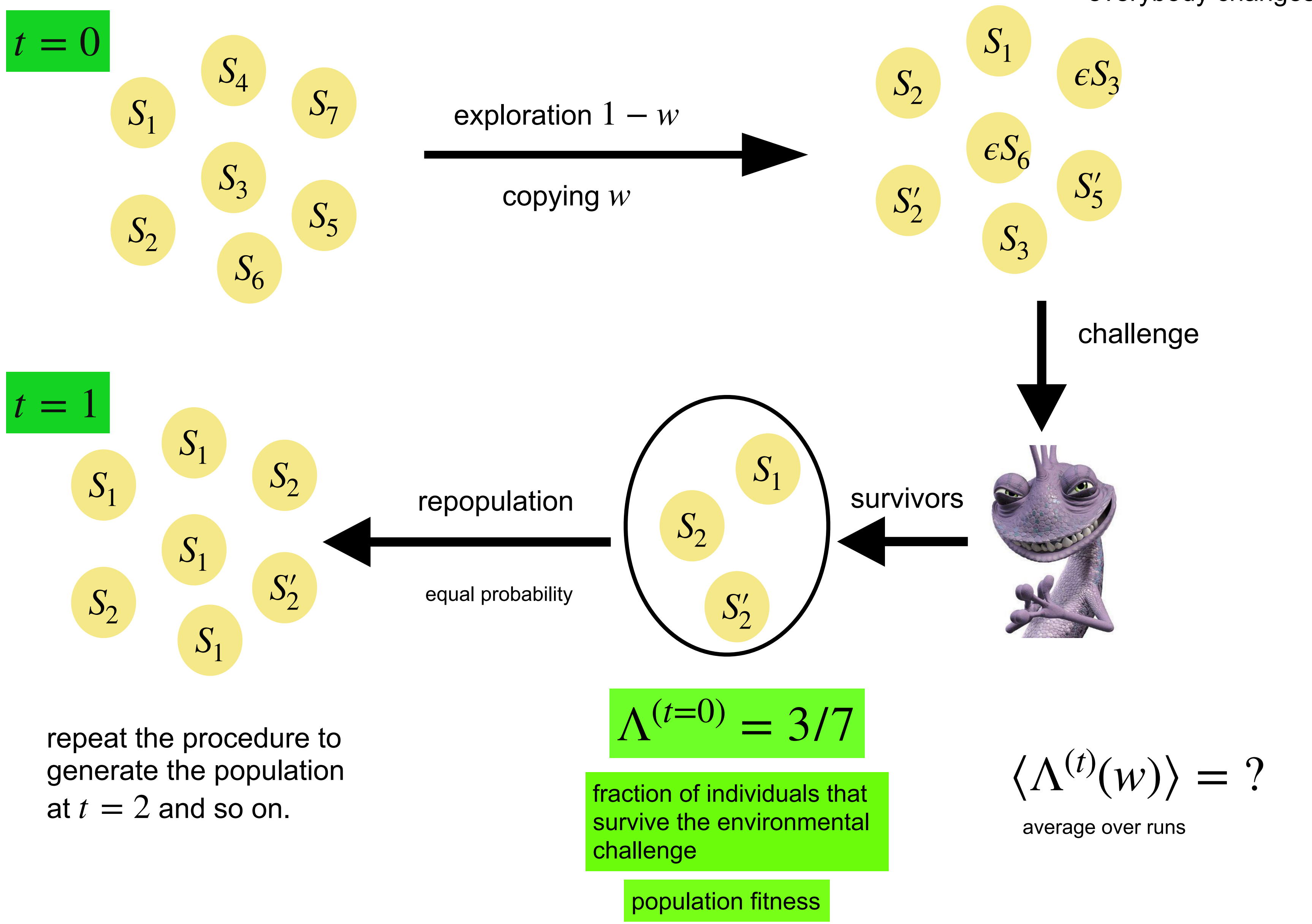
deceitfulness ↓ cost ↓

ξ_j w.p. γ → $S_i = \epsilon S_j$ $\epsilon \sim \text{Uniform}(1 - \eta, 1)$

ξ_j w.p. $1 - \gamma$ → $S_i = S_j$

- **population dynamics**

population size N=7 fixed



- **intuition for $\gamma = 0$**

A survivor at generation *t* with viability *S* has a well-defined lineage back to the generation τ when the viability value *S* first appeared. $\tau = 0, 1, \ldots, t$

τ — probability new viability S appears (mutant), survives the challenge and passes to generation $\tau + 1$

$$(1 - w) \times S \times \frac{1}{\Lambda^{(\tau)}}$$

$\tau + 1$ — probability the ancestor is copied and the copyist survives the challenge and passes to generation $\tau + 2$

$$w \times S \times \frac{1}{\Lambda^{(\tau+1)}}$$

$\tau + 2$ — probability the ancestor is copied and the copyist survives the challenge and passes to generation $\tau + 3$

$$w \times S \times \frac{1}{\Lambda^{(\tau+2)}}$$

.
.
.

t — probability the ancestor is copied and the last copyist survives the challenge:

$$w \times S$$

probability that an individual at generation t survives the challenge by copying an individual who has copied and individual at $t - 1$, who has copied an individual at $t - 2$, etc… who has copied an individual who explored the environment at generation τ:

$$\frac{(1 - w) w^t S^{t+1}}{\Lambda^{(\tau)} \Lambda^{(\tau+1)} \ldots \Lambda^{(t-1)}}$$

https://doi.org/10.48550/arXiv.2205.07969

- **analytical solution for $N \to \infty$**

$\mathbb{E}_S(S^n)$ probability S survives n challenges

$$\langle \Lambda^{(0)}(w) \rangle = (1-w)\mathbb{E}_S(S) + w b_1 \mathbb{E}_S(S)$$

$$\langle \Lambda^{(1)}(w) \rangle = (1-w)\left[\mathbb{E}_S(S) + \frac{w}{\langle \Lambda^{(0)} \rangle} b_1 \mathbb{E}_S(S^2)\right] + \frac{w^2}{\langle \Lambda^{(0)} \rangle} b_1 b_2 \mathbb{E}_S(S^2)$$

$$\langle \Lambda^{(2)}(w) \rangle = (1-w)\left[\mathbb{E}_S(S) + \frac{w}{\langle \Lambda^{(1)} \rangle} b_1 \mathbb{E}_S(S^2) + \frac{w^2}{\langle \Lambda^{(1)} \rangle \langle \Lambda^{(0)} \rangle} b_1 b_2 \mathbb{E}_S(S^3)\right] + \frac{w^3}{\langle \Lambda^{(1)} \rangle \langle \Lambda^{(0)} \rangle} b_1 b_2 b_3 \mathbb{E}_S(S^3)$$

$$\langle \Lambda^{(3)}(w) \rangle = (1-w)\Big[\underbrace{\mathbb{E}_S(S)}_{\tau=3} + \underbrace{\frac{w}{\langle \Lambda^{(2)} \rangle} b_1 \mathbb{E}_S(S^2)}_{\tau=2} + \underbrace{\frac{w^2}{\langle \Lambda^{(2)} \rangle \langle \Lambda^{(1)} \rangle} b_1 b_2 \mathbb{E}_S(S^3)}_{\tau=1} + \underbrace{\frac{w^3}{\langle \Lambda^{(2)} \rangle \langle \Lambda^{(1)} \rangle \langle \Lambda^{(0)} \rangle} b_1 b_2 b_3 \mathbb{E}_S(S^4)}_{\tau=0}\Big] + \frac{w^4}{\langle \Lambda^{(2)} \rangle \langle \Lambda^{(1)} \rangle \langle \Lambda^{(0)} \rangle} b_1 b_2 b_3 b_4 \mathbb{E}_S(S^4)$$

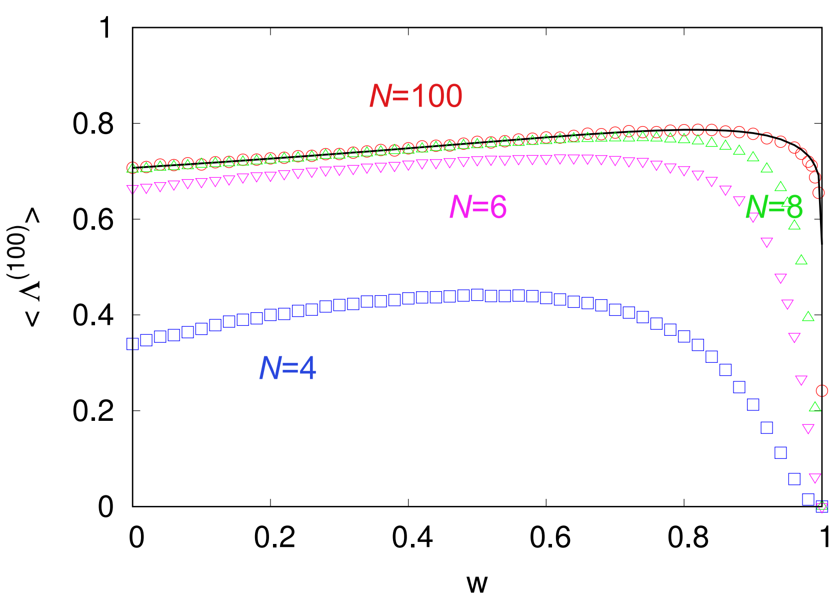
$$b_\tau = 1 - \gamma + \gamma \mathbb{E}_\epsilon(\epsilon^\tau)$$

- **analytical solution for $N \to \infty$ (continuation)**

$$\langle \Lambda^{(t)}(w) \rangle = (1-w) \sum_{\tau=0}^{t} a_{\tau,t} \mathbb{E}_S(S^{\tau+1}) w^{\tau} + a_{t+1,t} \mathbb{E}_S(S^{t+1}) w^{t+1}$$

$$a_{0,t} = 1 \qquad a_{\tau,t} = \frac{b_\tau}{\langle \Lambda^{(t-\tau)} \rangle} a_{\tau-1,t} \qquad \langle \Lambda^{(-1)} \rangle \equiv 1$$

mean population fitness at t=100



$\gamma = 0.5,\ \eta = 0.1,\ \sigma^2 = 1$

theoretical predictions fit the simulation data perfectly for large *N*.

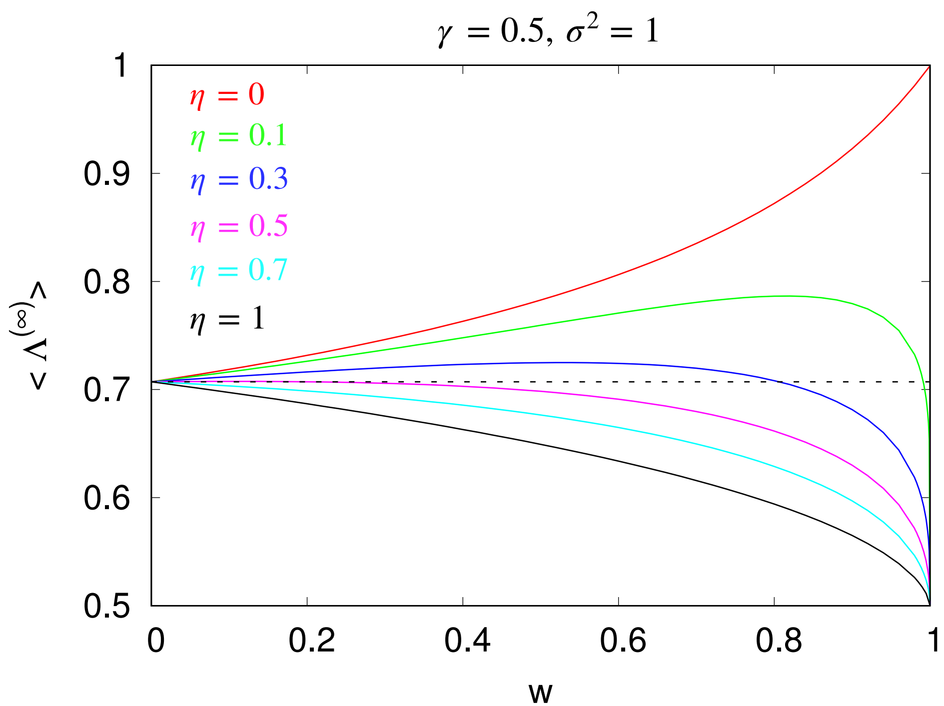
- **equilibrium analysis ($t \to \infty$)**

trust-always pure strategy ($w = 1$)

$$\langle \Lambda^{(\infty)}(1) \rangle = \begin{cases} 1-\gamma \text{ se } \eta > 0 \\ 1 \text{ se } \eta = 0 \end{cases}$$

trust-no-one pure strategy ($w = 0$)

$$\langle \Lambda^{(\infty)}(0) \rangle = \frac{1}{\sqrt{1+\sigma^2}}$$



What matters is the value of $w = \tilde{w}$ that maximizes the fraction of individuals that survive the environmental challenge.

$\tilde{w} = 0$ for $\eta > 4 - 2\sqrt{3} \approx 0.536$

transition point determined by the condition $\frac{d < \Lambda^{(\infty)} >}{dw}\Big|_{w=0} = 0$:

$$\eta_c^0 = \frac{2}{\gamma}\left(1 - \frac{\sqrt{1+2\sigma^2}}{1+\sigma^2}\right)$$

what's η?

$S_i = \epsilon S_j$

$\epsilon \sim \text{Uniform}(1-\eta, 1)$

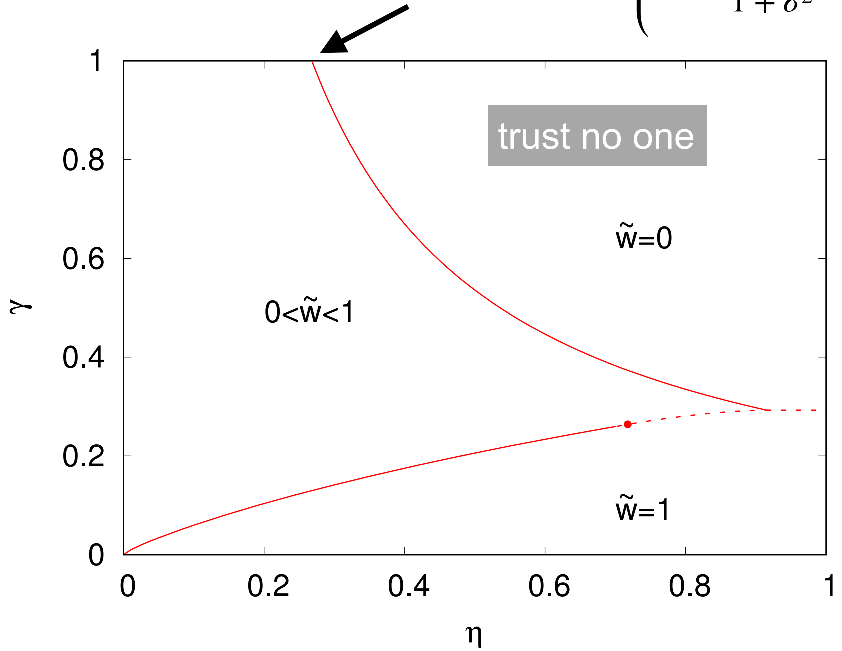
cost of believing false information

$\langle \Lambda^{(\infty)}(w) \rangle$ can be seen as minus the free-energy in a Landau-Ginsburg framework

- **phase diagram**

$\sigma^2 = 1$

$$\eta_c^0(\gamma = 1) = 2\left(1 - \frac{\sqrt{1+2\sigma^2}}{1+\sigma^2}\right)$$



trust-no-one regime disappears if $\eta_c^0(\gamma = 1) > 1$, i.e.,

$$\sigma^2 > 3 + 2\sqrt{3} \approx 6.46$$

- **lessons**

  - Increase of the hazardousness of the environment σ^2 favors trust. interesting

  - Increase of cost η of believing corrupted information favors the trust-always regime ($\tilde{w} = 1$). not obvious

  - Increase of deceitfulness γ and of cost η of believing corrupted information favors trust-no-one regime ($\tilde{w} = 0$). obvious

Who can we trust?

if the environment is harsh, trust any survivor.

Zahavi’s honest signalling principle

Thanks for the attention!