What data landscapes tell us about social processes:

Data-driven models from movie ratings, ideology, and party identification

Interesting data landscapes evolve when many people respond to the same question in a representative survey, an online rating website, or repeatedly in a panel survey. These macroscopic landscapes often show non-trivial shape and structure deviating from normal or uniform distributions. The evolution of the deviations is an often overlooked puzzle which solution can reveal insights on possible underlying social and individual processes. **Data-driven modeling** is the art of defining plausible social and individual mechanisms which reproduce such data landscapes, or at least stylized facts of them, with as few fitting parameters as possible.

Three puzzling landscapes and data-driven models are presented.

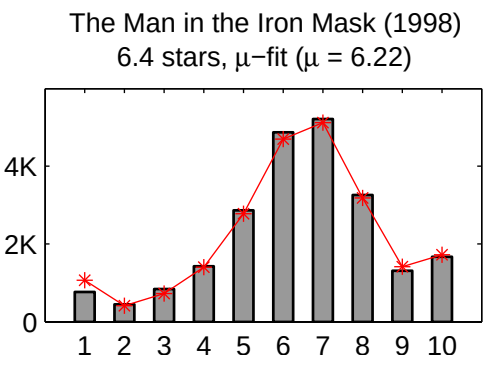
[1] Jan Lorenz. Universality of movie rating distributions. *European Physical Journal B* 71 (2009), 251–258.

[2] Thomas Metz and Jan Lorenz. Become who you are: The homing pattern in partisanship as a self-reinforcing stochastic process. *SSRN Preprint* (2013).

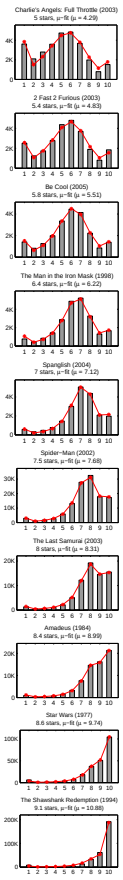
[3] Jan Lorenz. How Clustered Ideological Landscapes Emerge Through Opinion Dynamics. *ECPR General Conference Glasgow* (2014).

Movie ratings [1]

Data-generating process Users of IMDb.com rate the quality of movies they watched on a 1★—10★ scale. When users enter the website, they first see the average rating before they rate. Histograms are available. Dataset: All movies with more than 20,000 ratings were collected in 2008 (1,086 movies). Example (more at margin):

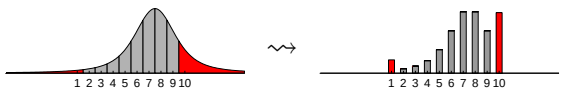


Movie rating Histograms and μ-fit

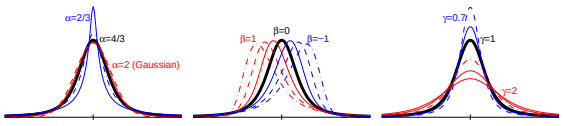


Stylized Facts Always either two or three peaks from which two lie at the extremes (1). In the region 2★—9★ the shape looks Gaussian-like (2).

Model (1) Confined discretization of a gradual perception, where individuals tend to overshoot:



(2) A rating is an average of several perceptions, thus central limit theorems suggest Levy skew α-stable distributions:

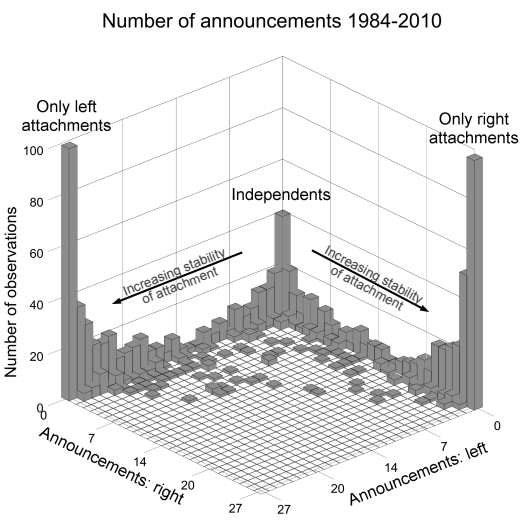


Center of location is parameterized with μ.

Conclusion The underlying distribution is not normal but fat-tailed, very typically $\alpha \approx \frac{4}{3}$. The average movie ($\mu \approx 7.5$) is most narrow (smallest γ) and not skew ($\beta = 0$). Deviation from 7.5 makes the opinion distribution broader and skew pronouncing the deviation. A customized **one-parameter fit** (μ-fit) based on these regularities models most histograms well (red lines in histograms).

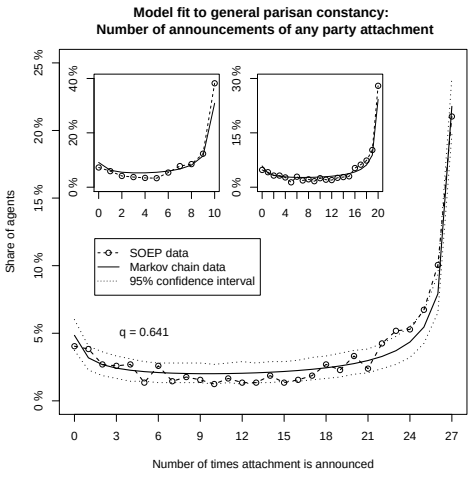
Party attachment [2]

Data-generating process Individuals are asked for party attachment ("Do you feel attached, and to which party?") from 1984 to 2010 in the German Socio-Economic Panel. Partisanship is quantified by counting the number of attachments uttered in these 27 years (N=965 West-Germans with unambiguous answers).



Stylized Facts (1) Most people name parties only left or right. (2) The distribution is U-shaped with peaks at never and always partisan.

Model Initially an attachment is uttered with probability q. At time t the probability of utterance is $\frac{q+x(t)}{t+1}$ where $x(t)$ is the number of uttered attachments up to t.

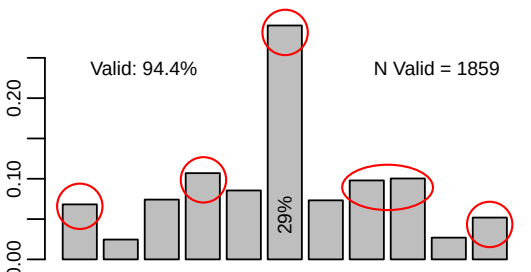


Conclusion Calibrated to $q = 0.641$ this self reinforcing stochastic process explains the distribution of partisanship almost perfect ($R^2 = 0.96$).

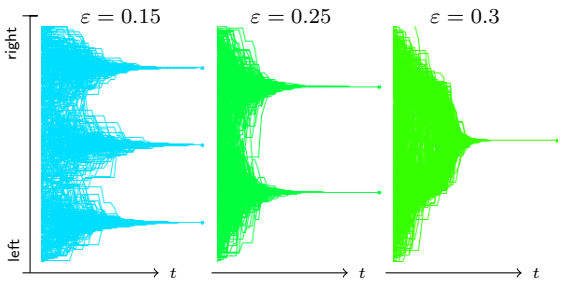
Ideological landscapes [3]

Data-generating process Representative people are asked about their self-placement on an 11 level ideological left—right axis in the European Social Survey 2002 to 2012. Between waves people discuss politics which might change their self-placement.

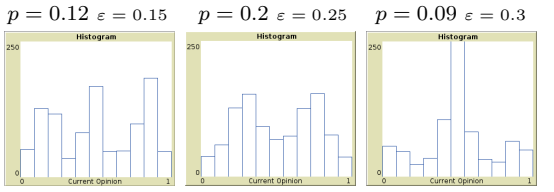
Stylized Facts Never a standard distribution; always largest peak at the center; multiple peaks ubiquitous; very often extremal peaks; often off-center peaks. Example from France 2012 (more at margin):



Model Agent-based model ($N = 1000$) with initial ideology uniform in $[0, 1]$. **Homophile adaptation**: In random pairwise encounters agents i, j adjust to $\frac{x_i+x_j}{2}$ if they are close in ideology $|x_i - x_j| < \varepsilon$.

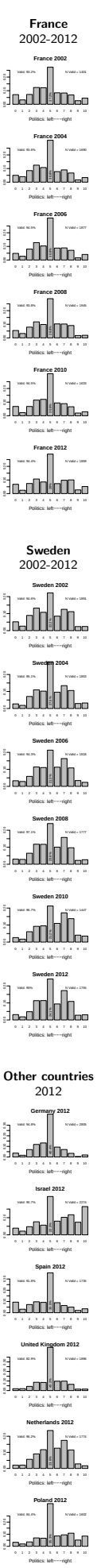


Reconsideration: With probability p an agent starts again from random opinion. Snapshots of histograms with 11 bins:



Conclusion The moderately clustered ideological landscapes can partly be explained by a model of opinion dynamics under homophile adaptation including possible reconsideration of ideology.

Some ideology landscapes



Jan Lorenz post@janlo.de, www.janlo.de
BIGSSS, **Jacobs University Bremen**. Poster presented at:
GESIS Computational Social Science Winter Symposium, Köln, Dec 2014

DFG
Project LO 2024/2-1 (Jan Lorenz)
Opinion Dynamics and Collective Decisions:
Procedures, Behavior and Systems Dynamics

BIGSSS
Bremen International Graduate School of Social Sciences

JACOBS UNIVERSITY