Sequence Analysis: A Toolbox
I: Properties of Sequences

Cees H. Elzinga
VU University Amsterdam
Faculty of Social Sciences

August 6, 2014

Structure

- Sequence Data & Sequence questions
- Single Sequence Properties
  - simple: length, distinct states, pattern
  - complex: entropy, complexity
- Applications to Family Formation and Labor Market Entry

Sequences

In social science, sequences arise when we study

- life courses
- family formation trajectories
- mobility histories
- school/work/employment careers
- ....
- ....

Sequences

In social science, sequences arise when we study

- life courses
- family formation trajectories
- mobility histories
- school/work/employment careers
- ....
- **time series of encoded behavior**

Encoding Examples

Code	Behavior/Status	Code	Behavior/Status
E	Employed	S	Single
U	Unemployed	U	Unmarried cohabitation
T	Vocational Training	M	Married
S	School	SC	Single with child(-ren)
F	Further education	UC	Cohabitation with child(-ren)
H	Higher education	MC	Married with child(-ren)

- the set of codes is called an "alphabet"
- two alphabets: school/work careers and family formation

VU UNIVERSITY AMSTERDAM

Alphabets

- the individual codes imply mutually exclusive behaviors or states
- small alphabet:
  - “poor”, “scarce” descriptions
  - finding “structure” less difficult
  - more identical patterns
- big alphabet:
  - “rich”, “detailed” descriptions
  - finding “structure” more difficult
  - less distinct patterns
- choosing an alphabet is determined by theory and/or purpose of the analysis

Sequence Data

Come in different formats:

- XT-format ("run-length encoding"):

$$\overbrace{95677\,0.842\,20\,2\,0\,4\,4}^{\text{covariates}}\ \overbrace{S/25\;M/11\;MC/85\;SC/15\;UC/2\;MC/6}^{\text{states \& durations=spells}}$$

- XX-format:

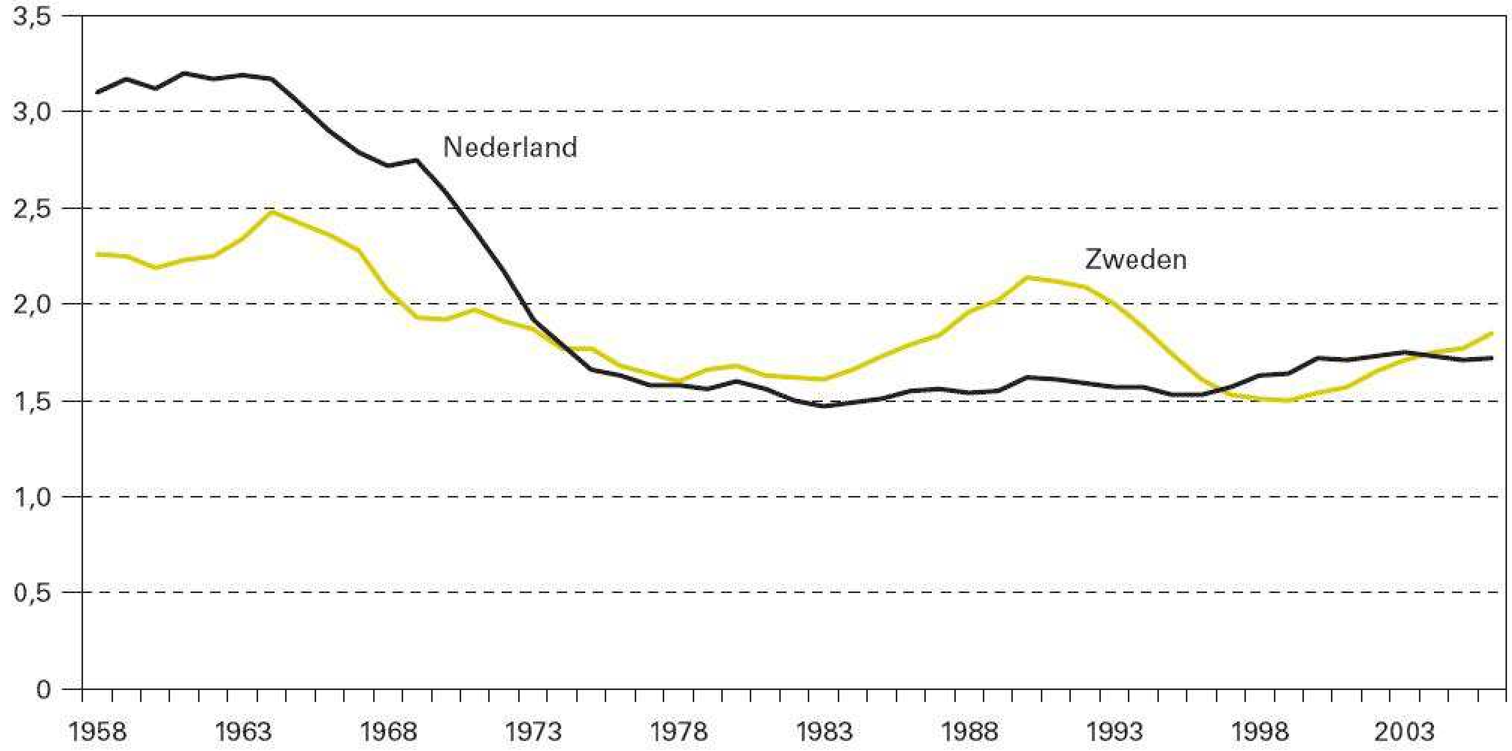
$$\overbrace{95677\ 0.842\ 20\ 2\ 0\ 4\ 4}^{\text{covariates}}\ \overbrace{\underbrace{SS\ldots S}_{25}\underbrace{M\ldots M}_{11}\underbrace{MC\ldots MC}_{85}\ldots}^{\text{states}}$$

Questions about Sequences I

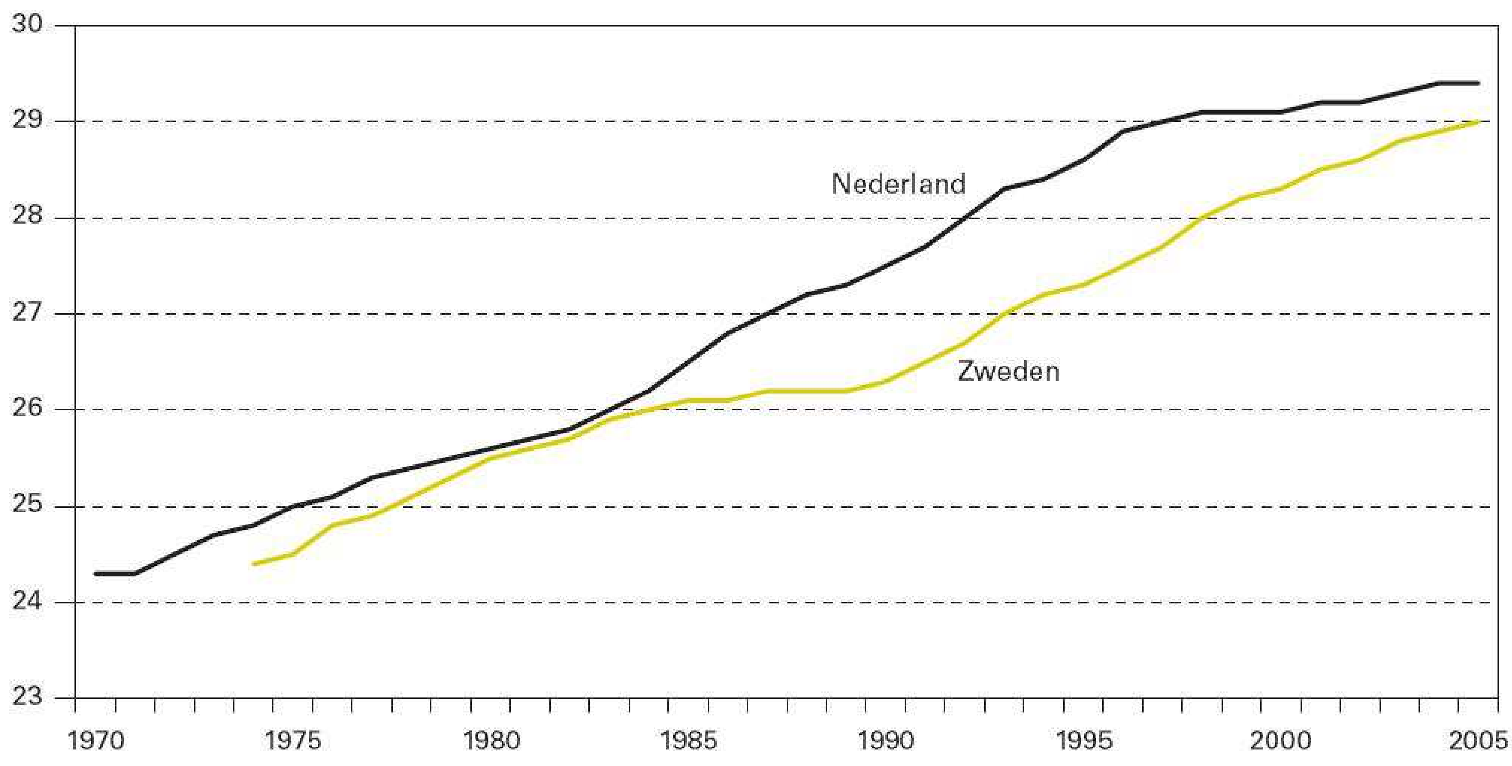
Either pertain to a particular type of event: e.g. entering parenthood

- occurrence (reproduce?)
- ordering (before marriage?)
- timing (at what age?)
- frequency (how often?)

Children/Woman



Mother's Age at 1^{st} Birth



Questions about Sequences II

Or pertain to sequences "as a whole":

- to properties of single sequences
  - simple properties: # distinct states
  - compound properties: complexity, type
- to properties of pairs of sequences
  - (dis-)similarity/distance (e.g. careers)
  - correlation (careers & family formation)
- to properties of sets of sequences
  - homogeneity
  - typology, classification, partition

Sequence Methods

Sequence Methods

- use sequences as unit of analysis
- to study "sequence issues":
  - complexity, type
  - similarity, distance, correlation
  - homogeneity
  - typology, classification

Example-issues for Sequence Methods

- (de-)standardization of the life course
- pathways to unemployment
- intergenerational transfer of family formation patterns
- types and determinants of criminal careers
-

Single-Sequence Properties

- length (# of spells)
- # distinct spells
- total duration
- satisfies a part. pattern?
-

Data: Family and Fertility Survey

- collected between 1988 - 1999: not very recent
- UN-ECE controlled in 25 countries: 19 selected
  - 42140 records selected
- 4 birth cohorts
  - born between '45-'49, '50-'54, '55-'59, '60-'64
  - women only
- Family Formation Histories: 6 distinct statuses

		with ≥ 1 Child	
Single	S		SC
Married	M		MC
Unmarried coh.	U		UC

- covering ages 18-30 years: 144 monthly statusses
  - record: S/68 U/2 M/53 MC/21

VU UNIVERSITY AMSTERDAM

Something happened in France!

	'45-'49	'60 -'64
average # dist. states	2.87	3.11
average length	2.96	3.42
% S M MC	48.4	14.5
% S U M MC	5.7	11.6
% Miscellaneous	18.8	38.5
% dist. trajectories	13.3	18.7

. . . like in most other countries

Internal Variation

$$x = S/144$$

$$y = S/42 \;\; M/12 \;\; MC/90$$

$$z = S/42 \;\; U/24 \;\; S/14 \;\; M/6 \;\; MC/24 \;\; SC/34$$

x has less variation than y has less variation than z

QUANTIFY ?

VU UNIVERSITY AMSTERDAM

Entropy: Unpredictability

$$x = S/144$$

$$y = S/42 \;\; M/12 \;\; MC/90$$

$$z = S/42 \;\; U/24 \;\; S/14 \;\; M/6 \;\; MC/24 \;\; SC/34$$

Guessing a state from x is easier than guessing from y is easier than guessing from z

Variation as Unpredictability

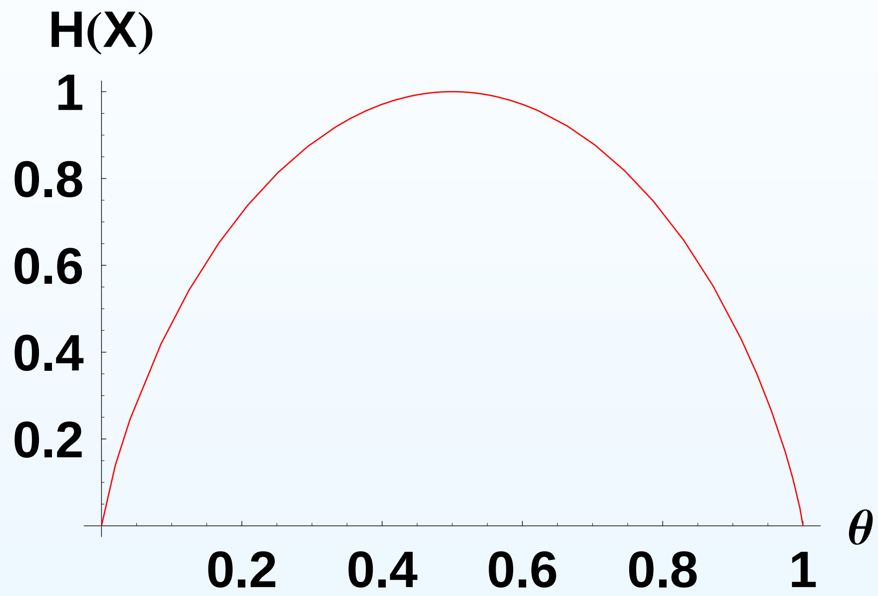
Entropy Formal

- consider an alphabet $\mathcal{A}$
  - e.g. $\mathcal{A} = \{S, U, M, SC, UC, MC\}$ or $\mathcal{A} = \{0, 1\}$
- and X a random variable over the alphabet $\mathcal{A}$
  - $p(x) = Prob(X = x)$ with $x \in \mathcal{A}$
- Entropy $H(X)$ is a property of X:
  - $H(X) = -\sum_{x \in \mathcal{A}} p(x) \log_2 p(x)$
  - convention: $0 \log 0 = 0$

Entropy of a fair coin

- $H(X) = -\sum_{x \in \mathcal{A}} p(x) \log p(x)$
- $\mathcal{A} = \{Heads, Tails\}$ with $p(Heads) = \theta$ and $p(Tails) = 1 - \theta$
- $H(X) = -[\theta \log \theta + (1-\theta) \log(1-\theta)]$
- coin fair: $\theta = \frac{1}{2}$
  - $H(X) = -[\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}] = 1$

Entropy of all coins



- H is maximal when uncertainty about the outcome is maximal
- $H = 0$ when there is no surprise: $\theta = 0$ or $\theta = 1$.

Entropy of sequences

$$x = S/144$$

$$y = S/42 \;\; M/12 \;\; MC/90$$

$$z = S/42 \;\; U/24 \;\; S/14 \;\; M/6 \;\; MC/24 \;\; SC/34$$

$$H(x) = 0$$

$$H(y) = -[\frac{42}{144}\log_2\frac{42}{144} + \frac{12}{144}\log_2\frac{12}{144} + \ldots] = 0.204$$

$$H(z) = 1.353$$

Relative Entropy

- Entropy is maximal when distribution is "flat": equiprobable states
  - $p(x) = \frac{1}{k}$ for $|\mathcal{A}| = k$
- Maximal entropy increases with size of alphabet
- $H(X) = -\sum \frac{1}{k} \log_2 \frac{1}{k} = \log_2 k$
  - $0 \leq H(X) \leq \log_2 |\mathcal{A}|$
- Relative entropy: $0 \leq \frac{H(X)}{\log_2 |\mathcal{A}|} \leq 1$

Labor Market Entry Data

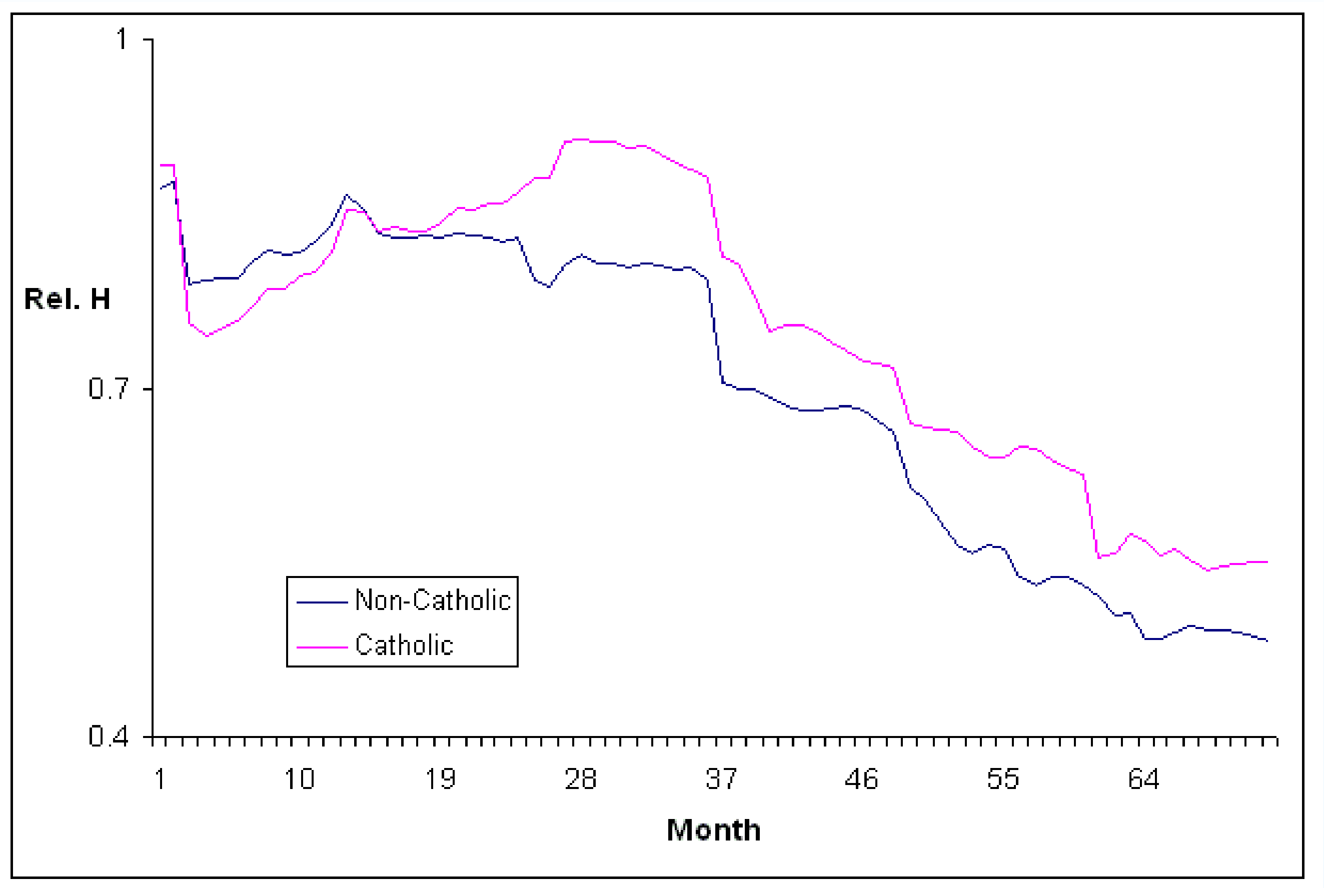
- 712 records of School Work Statuses
- from youngsters that were 16 in 1993
- i.e. at the end of compulsory education
- observed for consecutive 72 months

Status	Code	Status	Code
Employed	E	Unemployed	U
Training	T	School	S
Further Education	F	Higher education	H

(McVicar, D. & M. Anyadike-Danes, 2002)

42 0.21 1 1 1 0 0 0 0 0 1 0 0 1 E/11 U/5 T/1 E/4 T/27 U/18 E/2 U/2

Entropy over Time



Entropy and Event-Order

$$x = U/36 \;\; E/36$$
$$y = U/18 \;\; E/18 \;\; U/18 \;\; E/18$$

$$H(x) = -2[\frac{36}{72}\log_2\frac{36}{72}] = 1 = H(y)$$

- Entropy is sensitive to event-frequency, not to event-order!
- "Variation" in y exceeds the "variation" in x
- What do we mean by "variation"?

Quantifying through Subsequences

Toy-sequence: $abacb$

Subsequences: generate by taking away characters

Take	Subsequences
5 (all)	λ (empty)
4	a, b , c
3	ab, ac, bb, $\ldots$
2	aac, bcb, $\ldots$
1	$abac$, $abcb$, $\ldots$
0 (none)	$abacb$

$\phi(x)$: # distinct subsequences of x

$$\phi(abacb) = 26,\ 1 \leq \phi(x_1 \ldots x_n) \leq 2^n$$

Quantifying Complexity I

- family formation became more complex/turbulent
  - more states (single, cohabitated, (divorced))
  - repetition of states (union dissolution)
- The more distinct subsequences, the more turbulent
- $\phi(abcd) = 16 > \phi(abca) = 15 > \phi(abac) = 14$
- $1 \leq \phi(x^\ell) \leq 2^\ell$ (big numbers: impractical!)
- $0 \leq C(x) = \log_2(\phi(x)) \leq \ell$

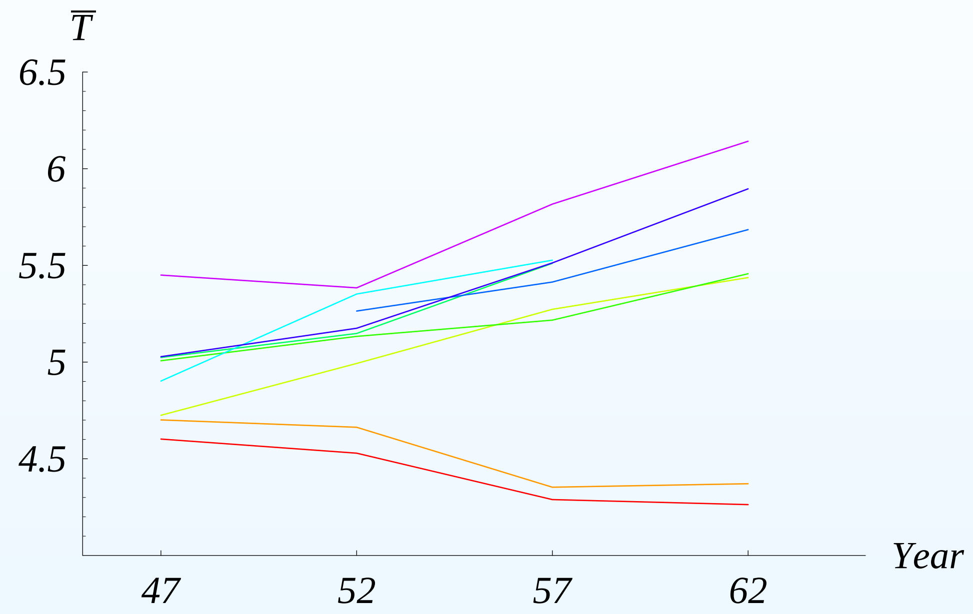
Quantifying Complexity II

- complexity decreases with variance of durations
- $C(a/99 \;\; b/1) < C(a/50 \;\; b/50)$

$$0 \leq C(x) = \log_2 \left(\phi(x) \cdot \frac{s^2_{t,max}(x) + 1}{s^2_t(x) + 1} \right)$$

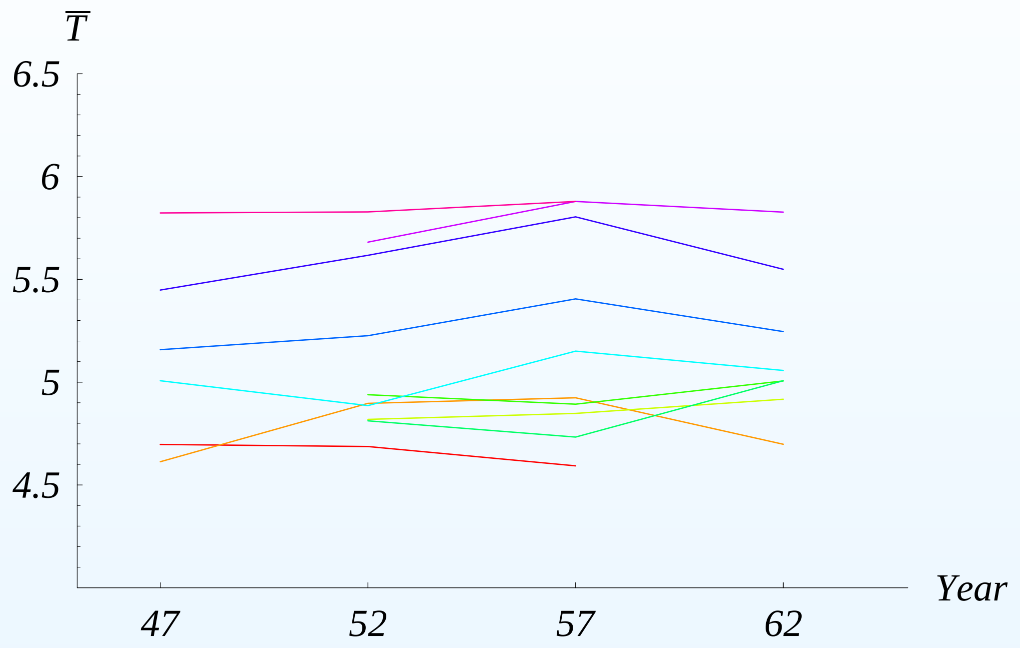
$$s^2_{t,max} = \frac{(n-1)(n - \sum_i t_x(i))^2}{n^2}$$

$\overline{C}$ changes in some countries ...



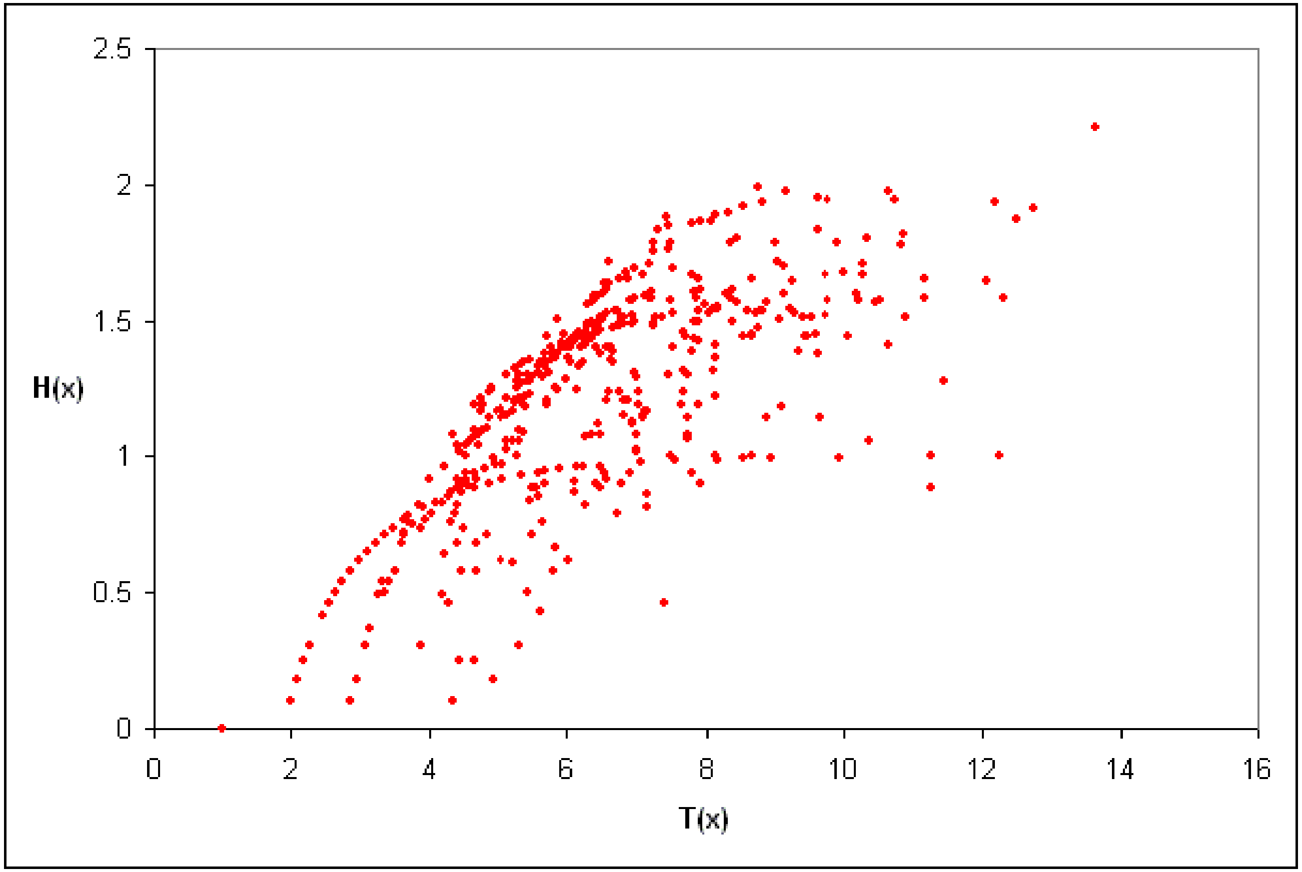
- In Portugal and Italy $\overline{C}$ decreases
- In Can, Aus, Nor, Fin, USA, Fra and NZ, $\overline{C}$ increases

But not in most countries...



- In Pol, Esp, Hun, Czech, Slo, Lit, Lat, Est, Neth and Swe, $\overline{C}$ does not change
- but remains constant at very different levels: Pol vs Swe

Entropy & Complexity



- Complexity and Entropy for Labor Market Entry Data

Nuts & Bolts for Weirdos

- the i^{th} *prefix* $x^i = x_1 \dots x_i$: the first i states
- count $\phi(\lambda = x^0) = 1$, $\phi(x^1) = 2$, $\phi(x^2)$, $\phi(x^3)$, etc
- if x_n is "new": double the number found for $\phi(x^{n-1})$
- if x_n is not new: double and compensate

$$\phi(x^n) = \begin{cases} 2\phi(x^{n-1}) & \text{if} \quad x_n \text{ is new} \\ 2\phi(x^{n-1}) - \phi(x^{\ell-1}) & \text{if} \quad x_n \text{ is not new} \end{cases}$$

ℓ: last occurrence of character x_n in x^{n-1}

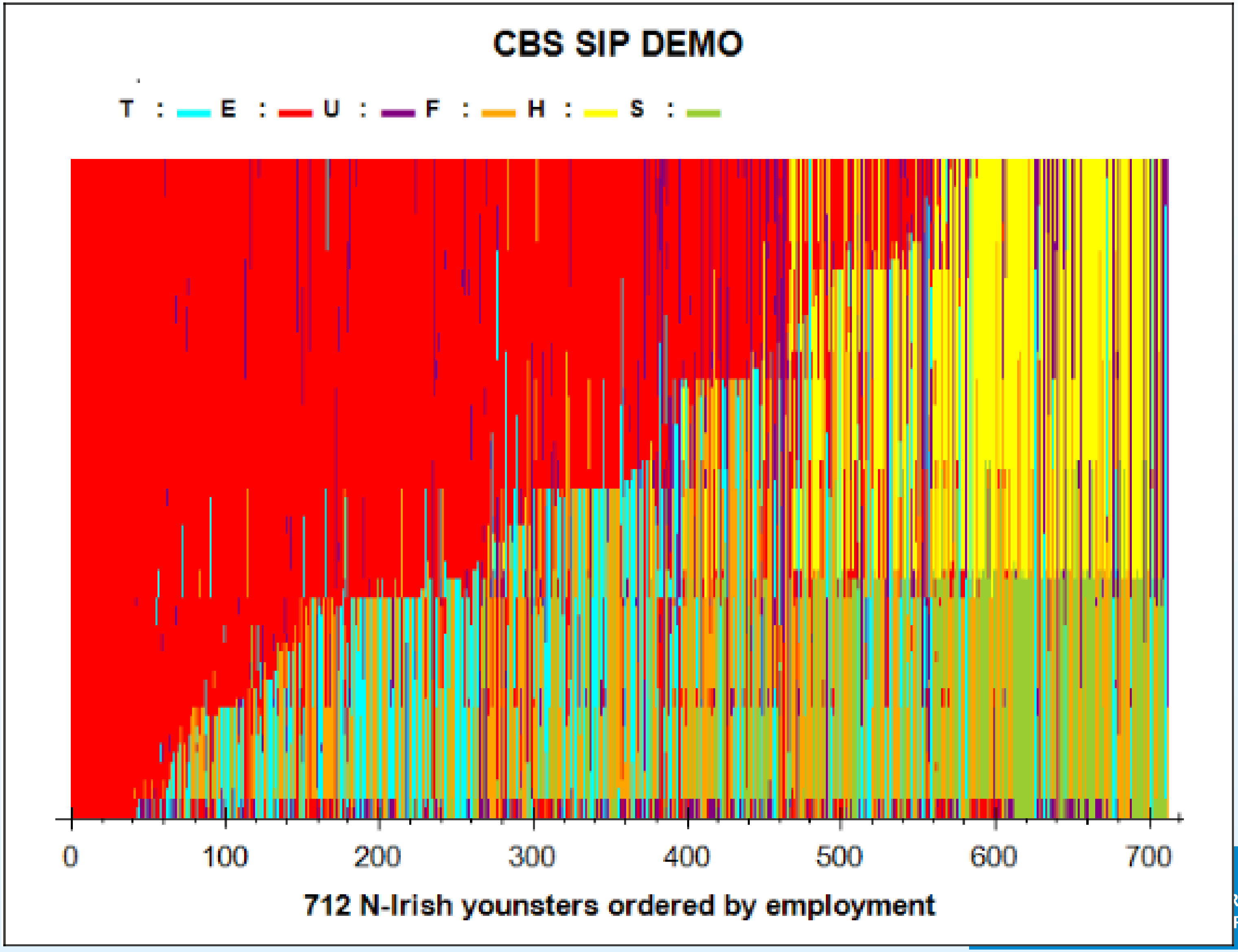
	λ	a	b	a	c	b
ℓ	0	0	0	1	0	2
ϕ	1	2	4	7	14	26

VU UNIVERSITY AMSTERDAM

Visual Rendering: Sequence Index Plots I

- each **state** is assigned a **fixed** color/shading
- for each sequence
  - for each state, create a bar of appropriate color,
  - of a length, proportional to duration of the state
  - and stack the state-bars in the right order
- order the stacked bars according to some criterion
  - e.g. time spend in a particular state
  - loading on some MDS- or PC-dimension

Sequence Index Plots: Example



Further Reading

- Brzinsky-Fay,C., U. Kohler & M. Luniak "Sequence Analysis with Stata", The Stata Journal (2006) 4, pp. 435-460.
- Gabadinho, A., G. Ritschard, N.S. Müller and M. Studer (2011), Analyzing and visualizing state sequences in R with TraMineR, Journal of Statistical Software, 40(4), pp. 1-37. Freely available from http://www.jstatsoft.org/v40/i04/.
- Piccarreta, R. and O. Lior "Exploring Sequences: a Graphical Tool based on Multidimensional Scaling", Journal of the Royal Statistical Society A, 173(1)(2010) 165-184.
- Bürgin, Reto and Gilbert Ritschard. 2014.A Decorated Parallel Coordinate Plot for Categorical Longitudinal Data. The American Statistician 68:98-103.
- Fasang, Anette and Tim Futing Liao (2014) Visualizing Sequences in the Social Sciences: Relative Frequency Sequence Plots, Sociological Methods & Research (Online First).
- Gabadinho, A., G. Ritschard, M. Studer and N.S. Müller (2011), Extracting and Rendering Representative Sequences, In Fred, A., Dietz, J.L.G., Liu, K. & Filipe, J. (eds) Knowledge Discovery, Knowledge Engineering and Knowledge Management., Series: Communications in Computer and Information Science (CCIS), Vol. 128, Berlin: Springer-Verlag. pp. 94-106. doi: 10.1007/978-3-642-19032-2.
- Rohwer, G. and U. Pötter, TDA User Manual (tman.pdf), 2005. (http://www.stat.ruhr-uni-bochum.de/tman.html)
- Elzinga, C.H. CHESA 3.1 User manual (2009) (http://home.fsw.vu.nl/ch.elzinga/)

VU UNIVERSITY AMSTERDAM