Causal Graphs

Description

Infer the most likely graph of causal relations between variables from observational data

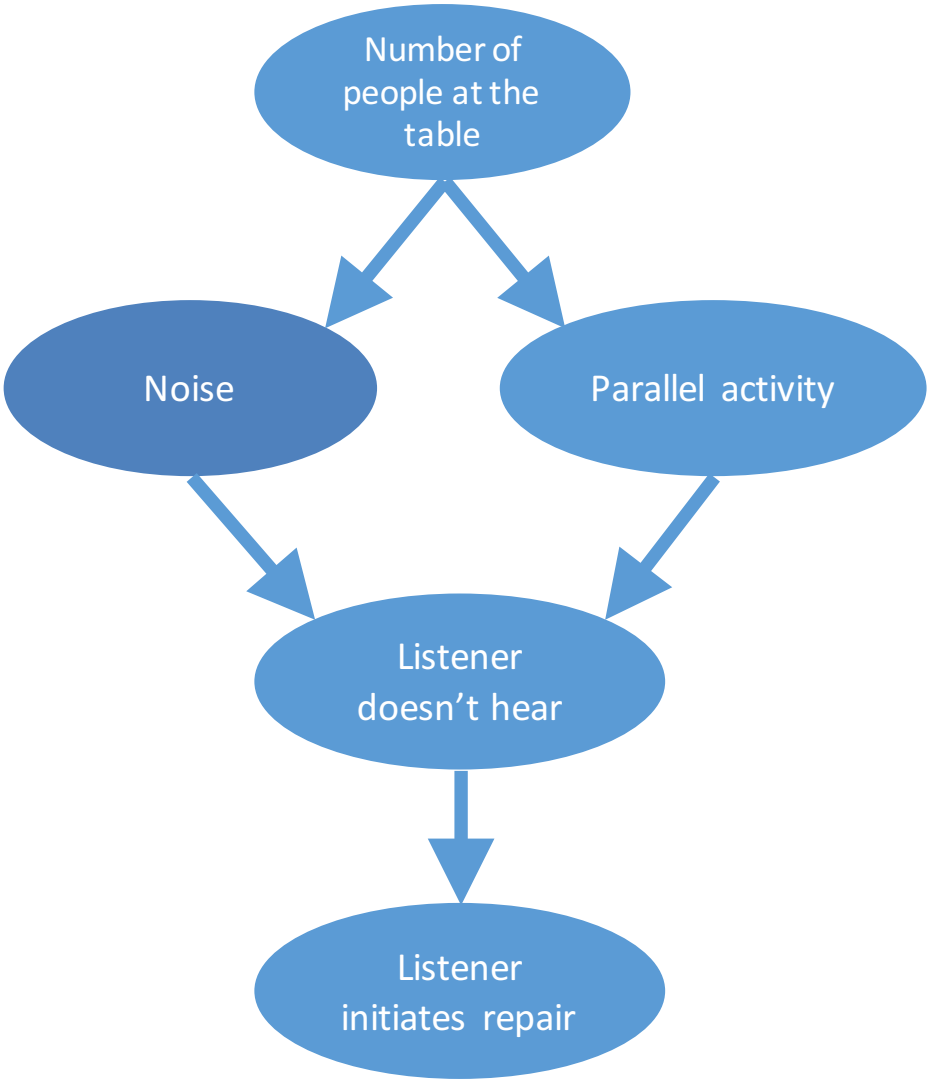
Advantages

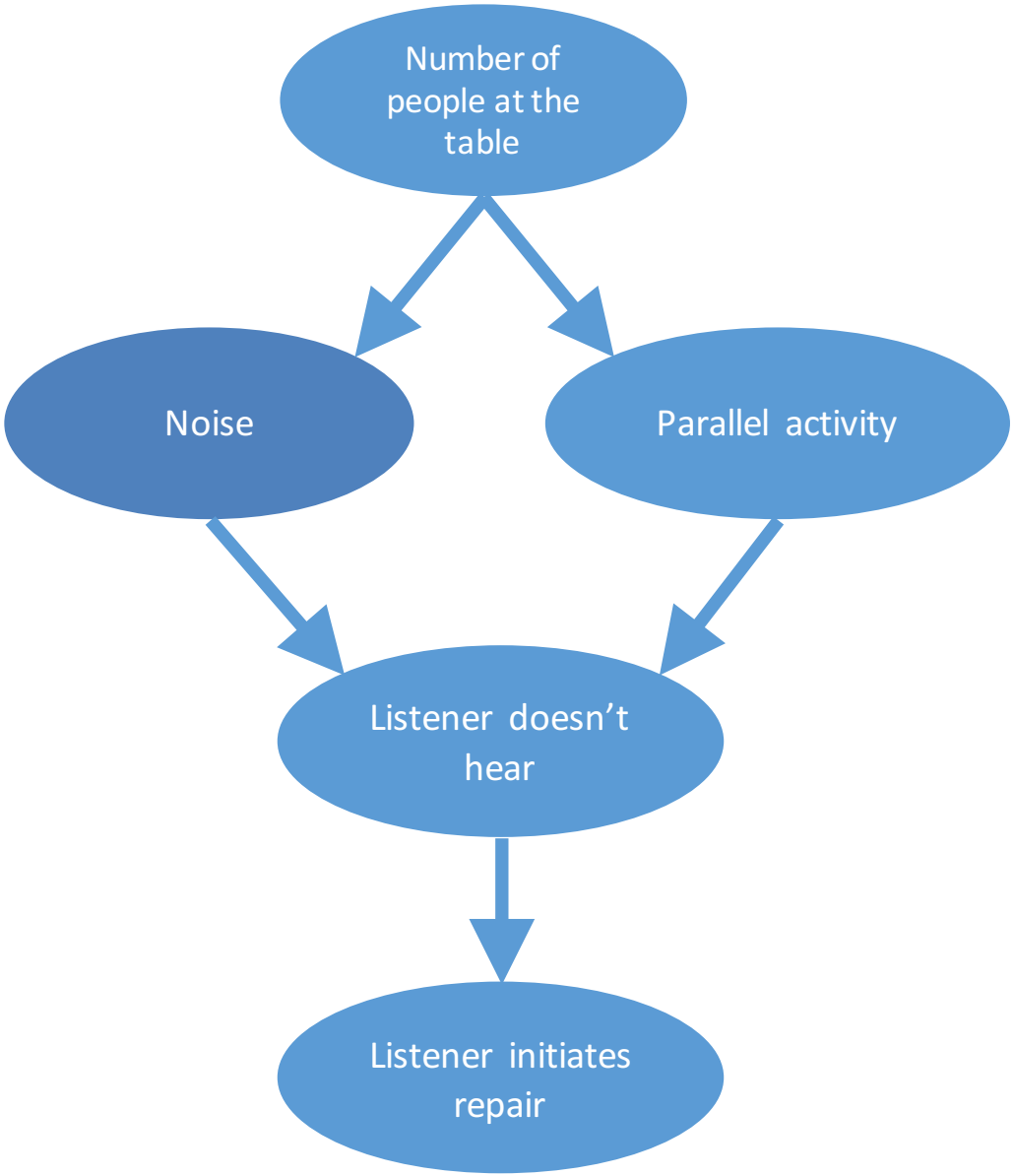
Handles large numbers of variables

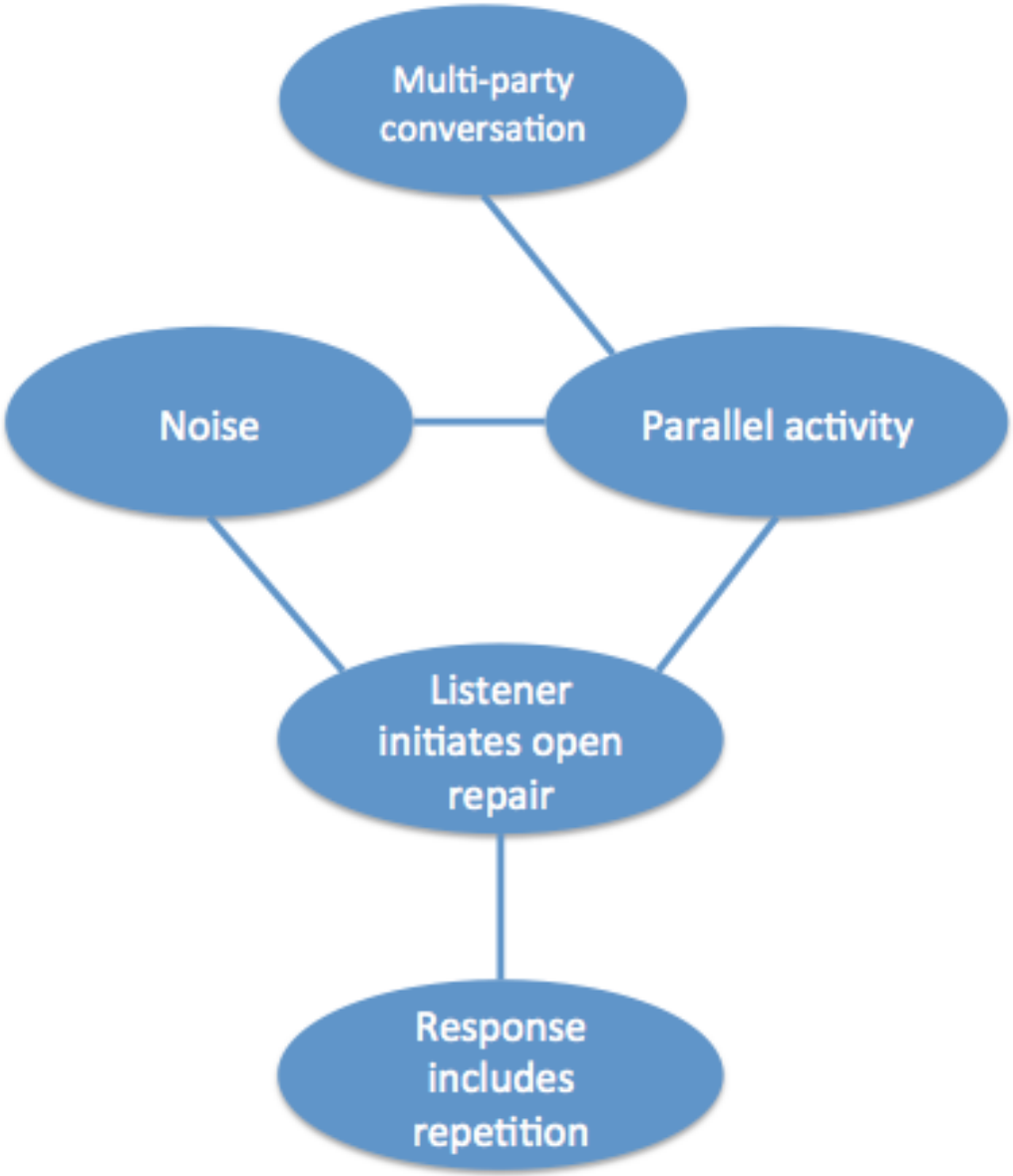
Can handle complex relationships

Easy to interpret

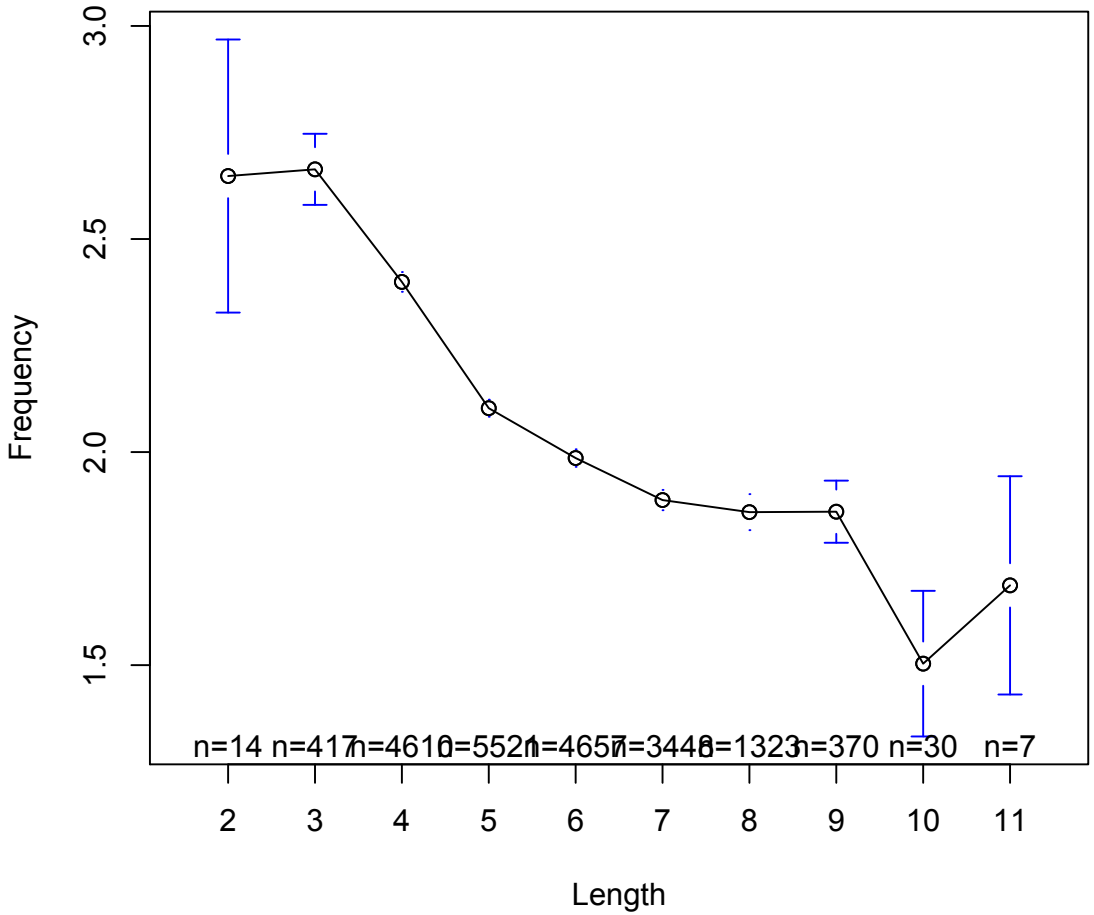
A conversation in a crowded restaurant



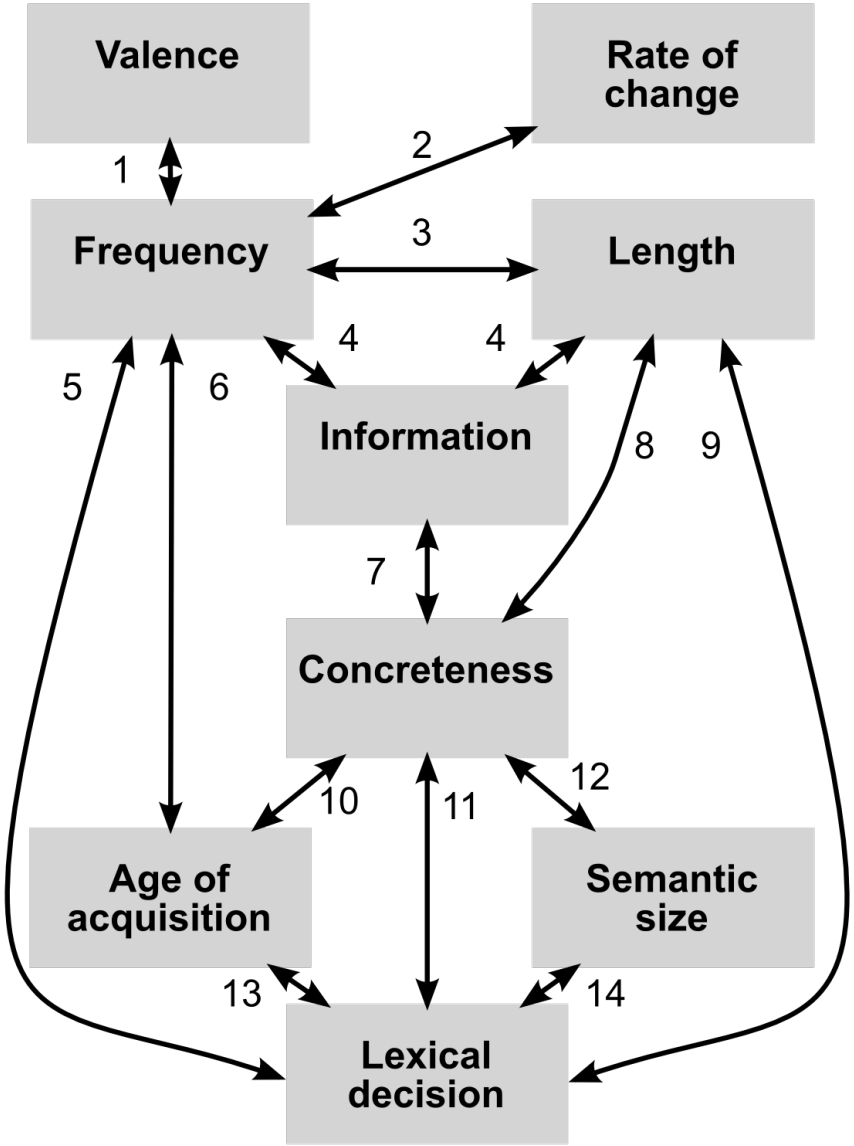

Number of people at the table
Noise
Parallel activity
Listener doesn't hear
Listener initiates repair


Multi-party conversation
Noise
Parallel activity
Listener initiates open repair
Response includes repetition

Frequency and Length

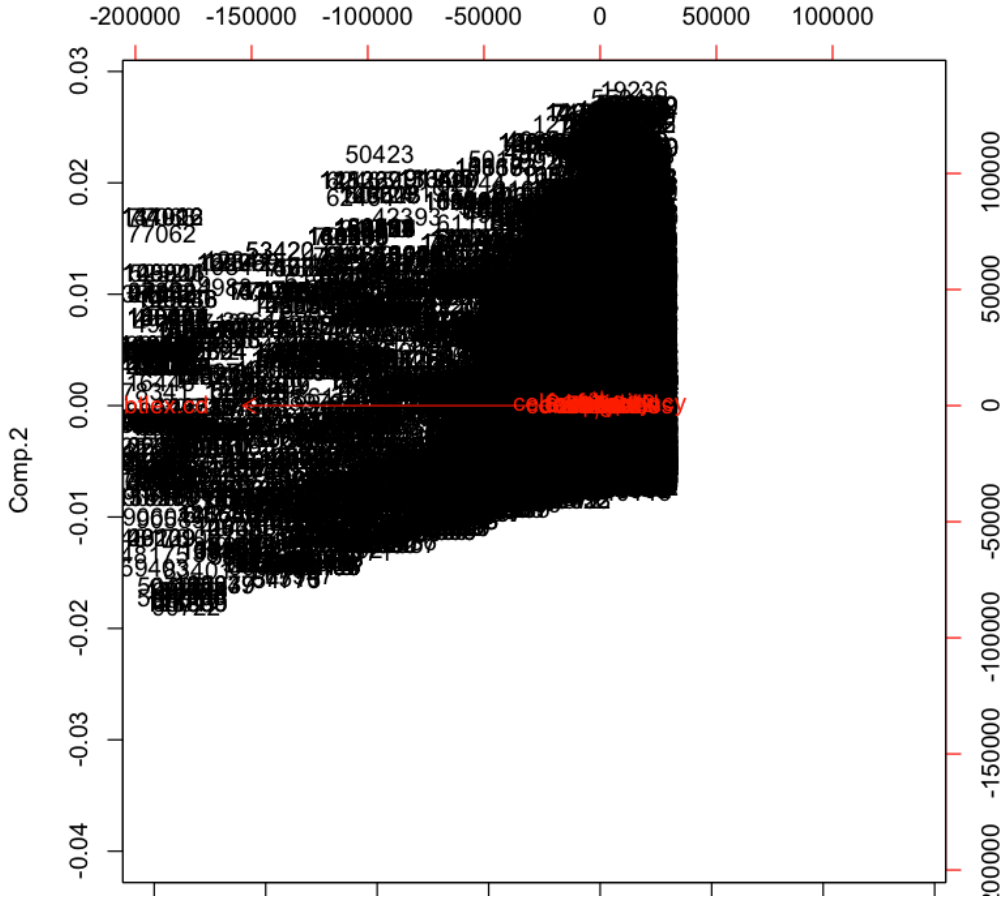


Properties of words



1. Boucher & Osgood (1969)
2. Pagel et al. (2007)
3. Zipf (1936)
4. Piantadosi et al. (2011)
5. Balota et al., 2004
6. Kuperman et al., 2013
7,8. Piantadosi et al. (2011b)
9. Hudson & Bergman, 1985
10. Reilly & Jacobs, 2007
11,12. Yao et al. (2013)
13. Walker & Hulme (1999)
14. Sereno et al. (2009)

Solution: Factor analysis?



Factor Loading rankings:

- Contextual Diversity
- Orthgraphic neighbourhood
- Familiarity
- Surprisal

Problems:

No causal analysis

Massive non-linearity

Difficult to interpret

Solution: Causal Graph Analysis

From observational data, find the most likely directed, acyclic graph of causal connections

Assumptions:

- Closed system
- Acyclic graph: No co-evolution

Inferring causal graphs

PC algorithm (Sprites et al., 2000; Kalisch et al., 2012)

Start with fully unconnected graph

For each pair of variables:

Try to find evidence that the variables are independent:

no correlation,

or correlation is explained by a set of other variables

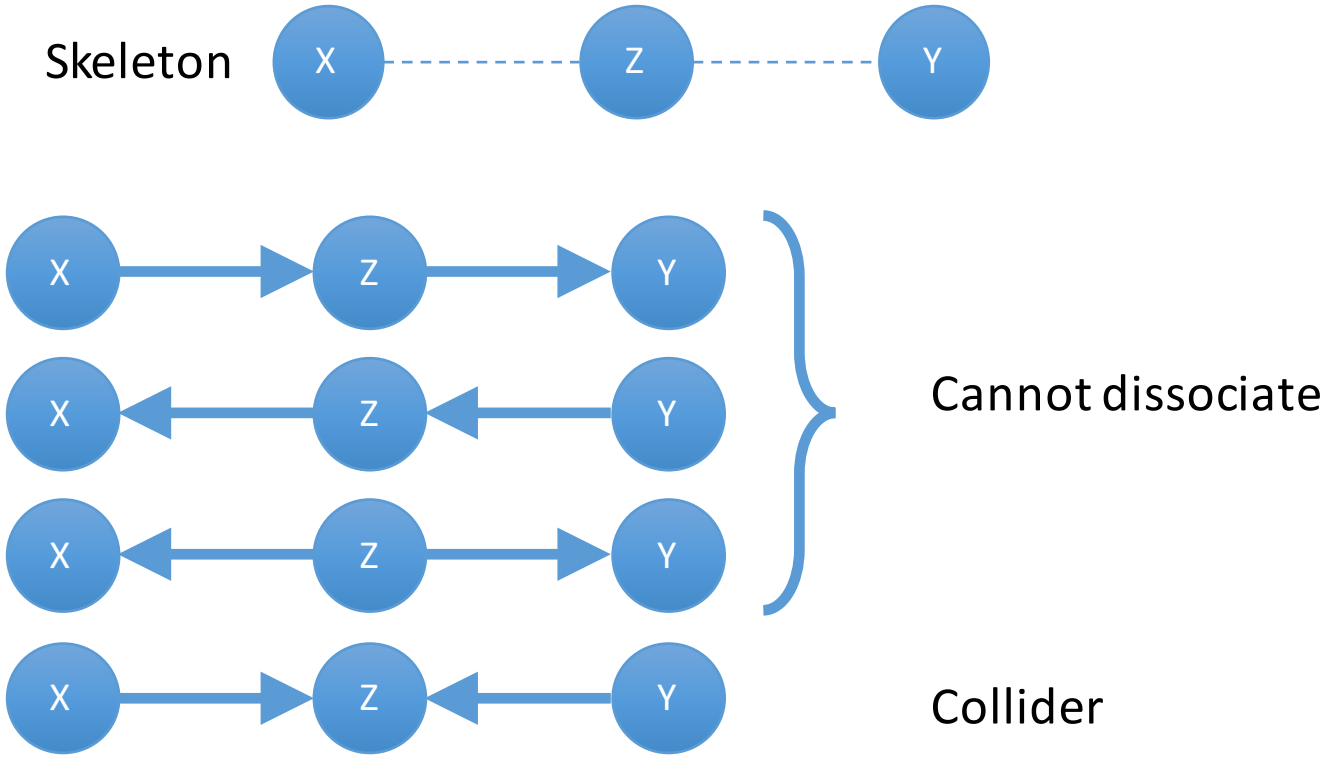
Any statistical test can be used (e.g. conditional independence)

If variables are not independent, add an edge.

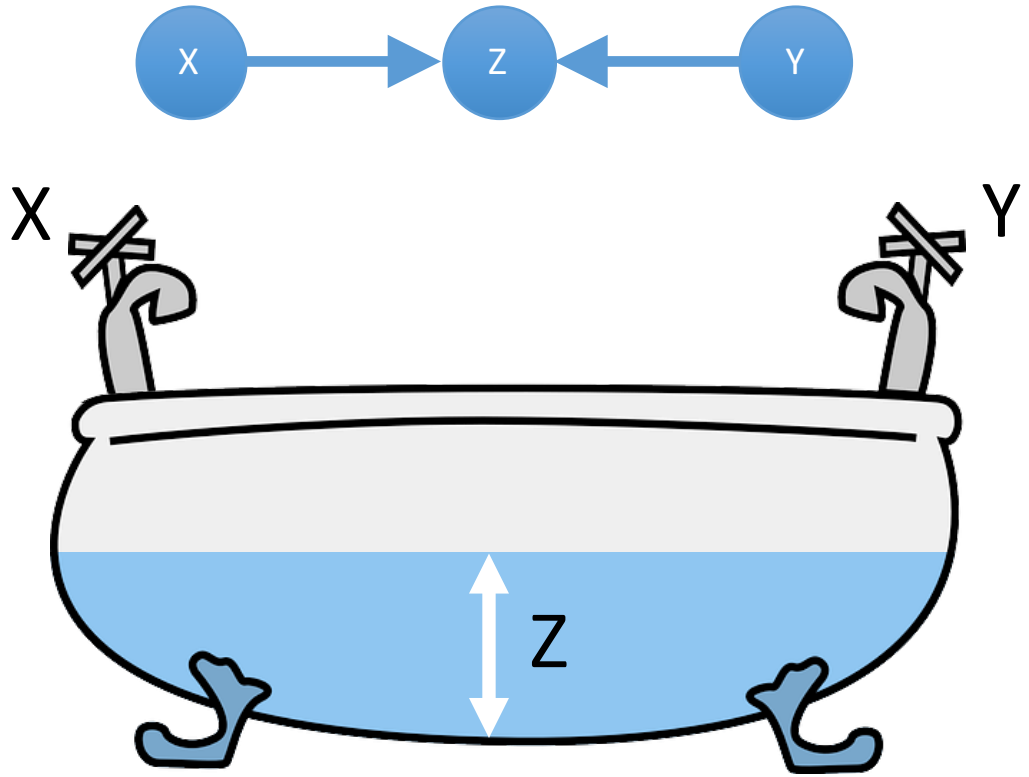
The PC algorithm is an efficient way of performing only the tests which need to be done.

Results in a ‘skeleton’ graph

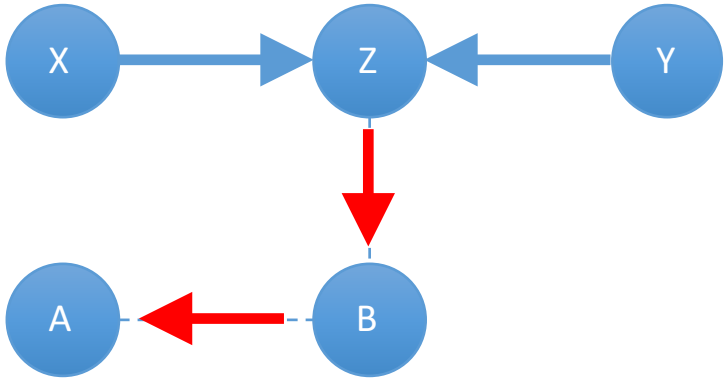
Orienting the edges



Colliders



Orienting the edges

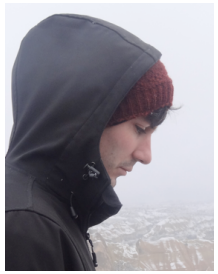


Case study: Properties of words

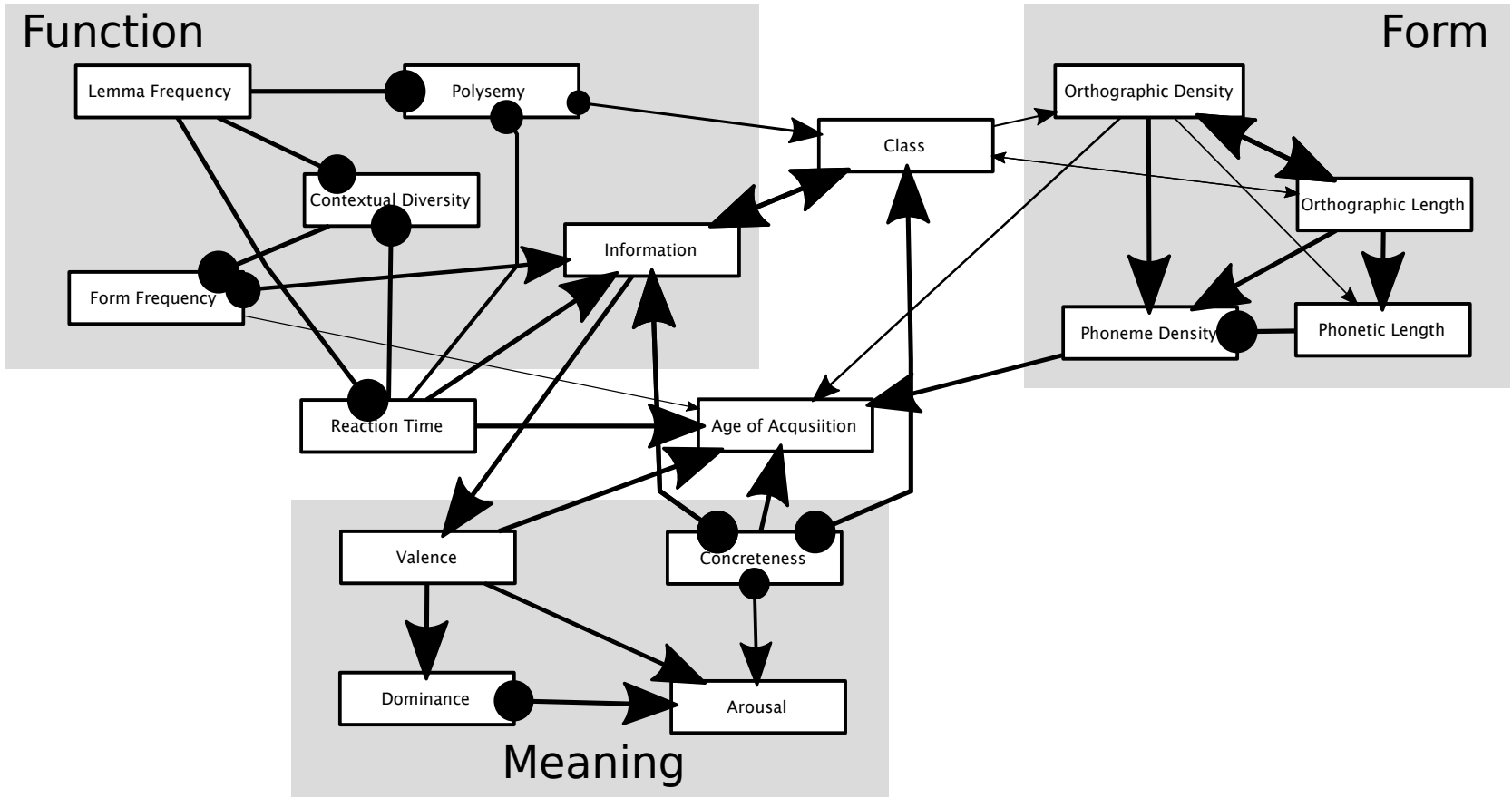
- Length (CELEX)
- Frequency (form and lemma) (CELEX)
- Part of Speech (CELEX)
- Age of Acquisition (Kuperman et al. 2013)
- Valence, Arousal, Dominance (Warriner et al. 2013)
- Concreteness (Brysbaert et al., in press)
- Lexical decision time (Keuleers et al., 2012)
- Orthographic neighbourhood (OLD20)
- Phonological neighbourhood (IPhOD)
- Contextual diversity (Subtlex)
- Familiarity (Wordnet Polysemy)
- Surprisal (information) (Piantadosi et al.)

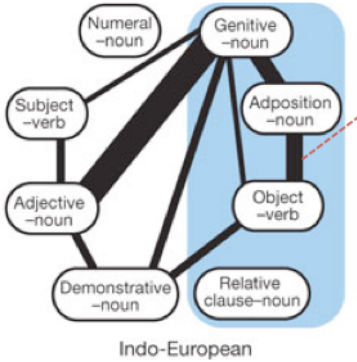
10,000 English words

Marloes Maathuis



Damián Blasi

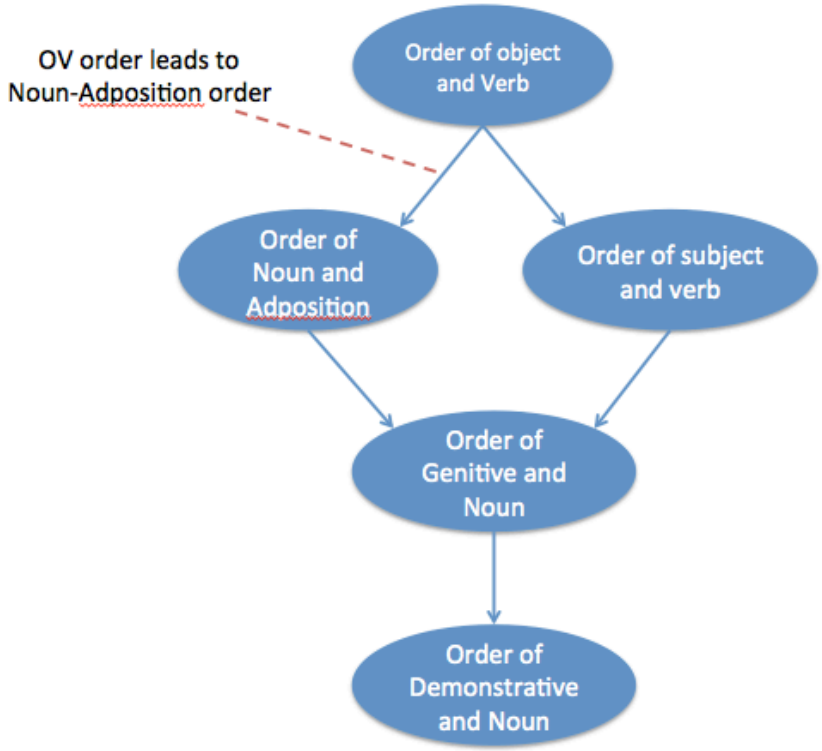

Function
Form
Meaning
Lemma Frequency
Polysemy
Contextual Diversity
Information
Form Frequency
Reaction Time
Class
Age of Acqusiition
Orthographic Density
Orthographic Length
Phoneme Density
Phonetic Length
Valence
Concreteness
Dominance
Arousal

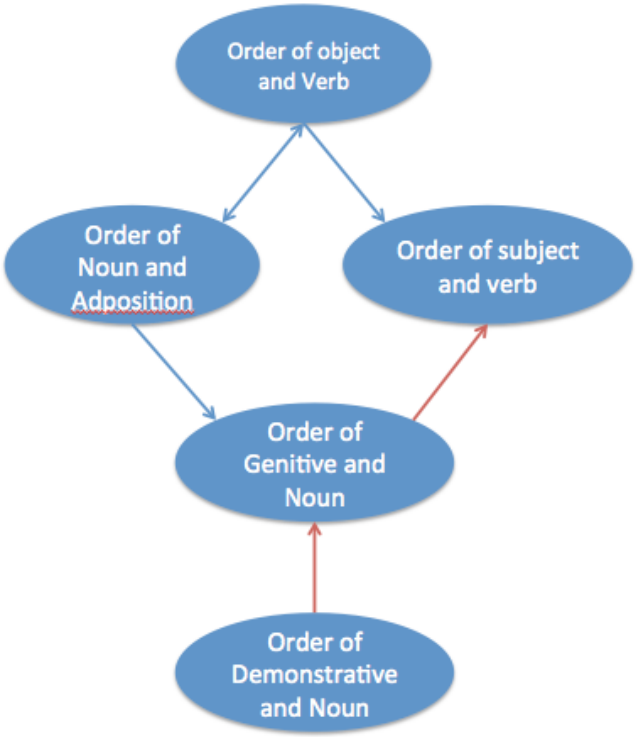


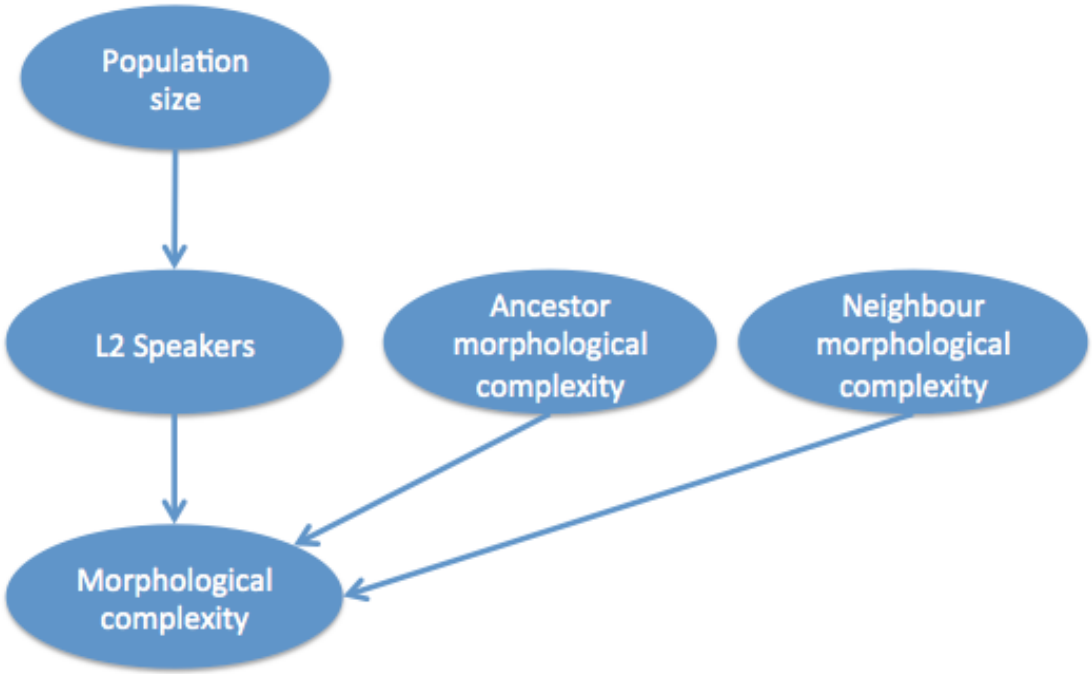
Indo-European

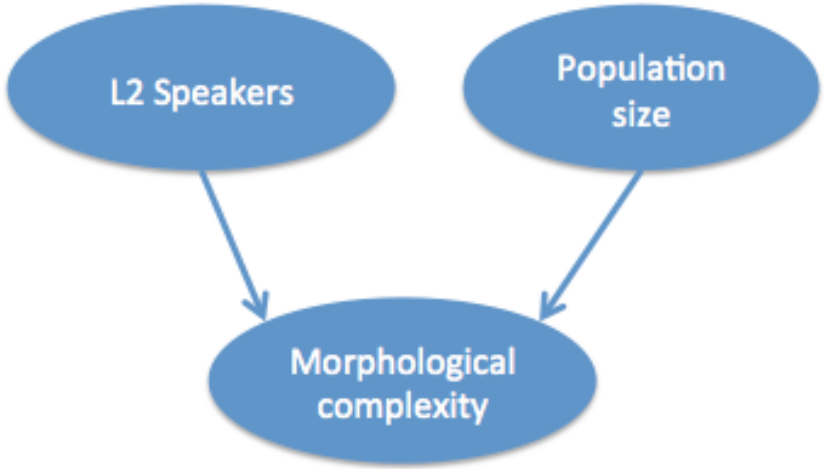
OV order leads to Noun-Adposition order

	OV	VO
N-Ad	472	41
Ad-N	14	454






Population size
L2 Speakers
Ancestor morphological complexity
Neighbour morphological complexity
Morphological complexity


L2 Speakers
Population size
Morphological complexity

Application to diversity

Demographics

Population size

Population density

Country area

Attitudes

Uncertainty Avoidance

Long term/short term orientation

Masculinity

Collectivism

Power distance

Climate

Water Availability

Frost days

Mean growing season

Mean temperature

Biodiversity

State

GDP

Gini

Pathogen prevalence

Km roads

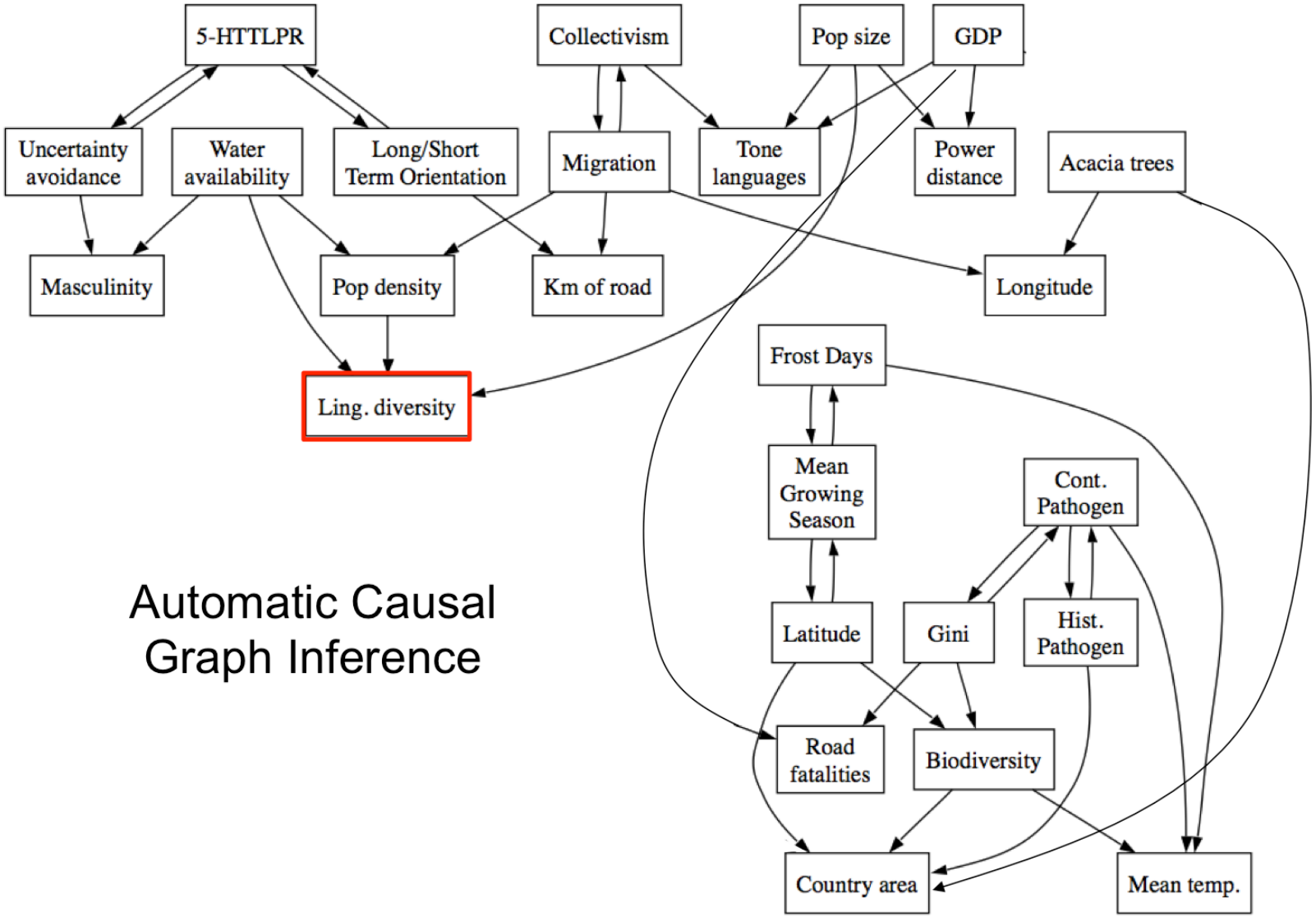
Random

Road fatalities

Migration

Acacia trees

Prevalence of 5-HTTLPR gene



Automatic Causal Graph Inference