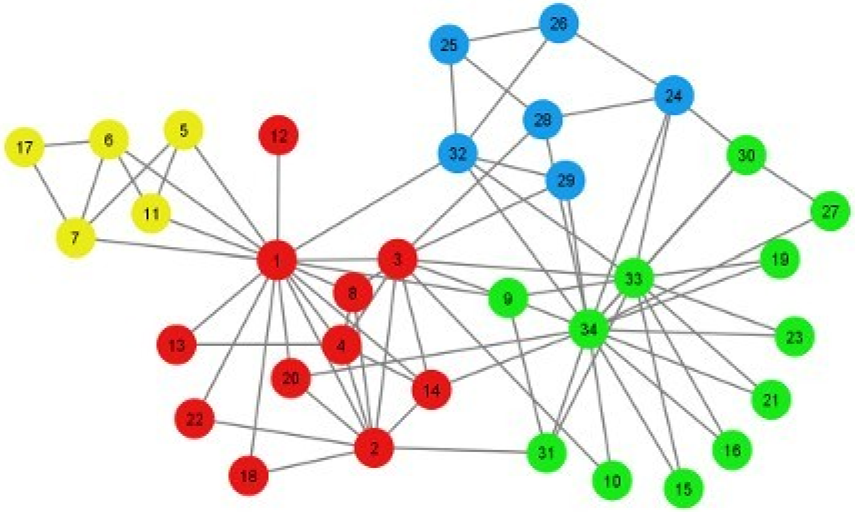
Community detection in networks with unobserved edges

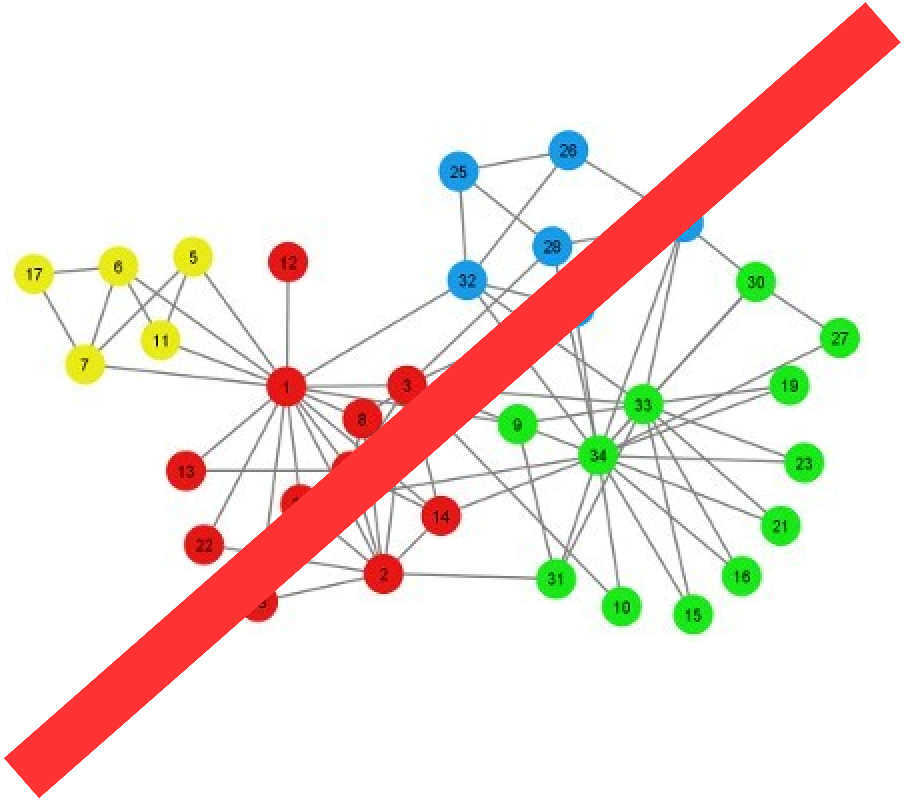
Leto Peel
Université catholique de Louvain
@PiratePeel

Community detection



Aim: partition the network according similarity of link structure

Community detection



Aim: partition the network according similarity of link structure

But we observe signals on nodes and no links!

Motivating examples…

Identify assets whose prices vary coherently to better manage risk

Motivating examples...

Identify regions of the brain to predict the onset of psychosis and learn about the ageing of the brain

Identify assets whose prices vary coherently to better manage risk

Motivating examples...

Identify regions of the brain to predict the onset of psychosis and learn about the ageing of the brain

Identify assets whose prices vary coherently to better manage risk

Identify climate zones to better understand factors affecting our climate

Is there really a network?

Is there really a network?

We don’t have to directly observe something to believe it is true

Common practise

- Calculate pairwise correlations between signals (e.g. Pearson’s).
- Threshold (and Binarize) the matrix of correlations.
- Perform community detection on this (notional) network

Problems

- This procedure commonly invokes point-estimates at each step
  - Does not capture the uncertainty of individual links

Problems

- This procedure commonly invokes point-estimates at each step
  - Does not capture the uncertainty of individual links

- Unclear how to include missing data.

- No intrinsic/clear notion of the right number of communities.

The signals we observe from many nodes are driven by a few latent factors

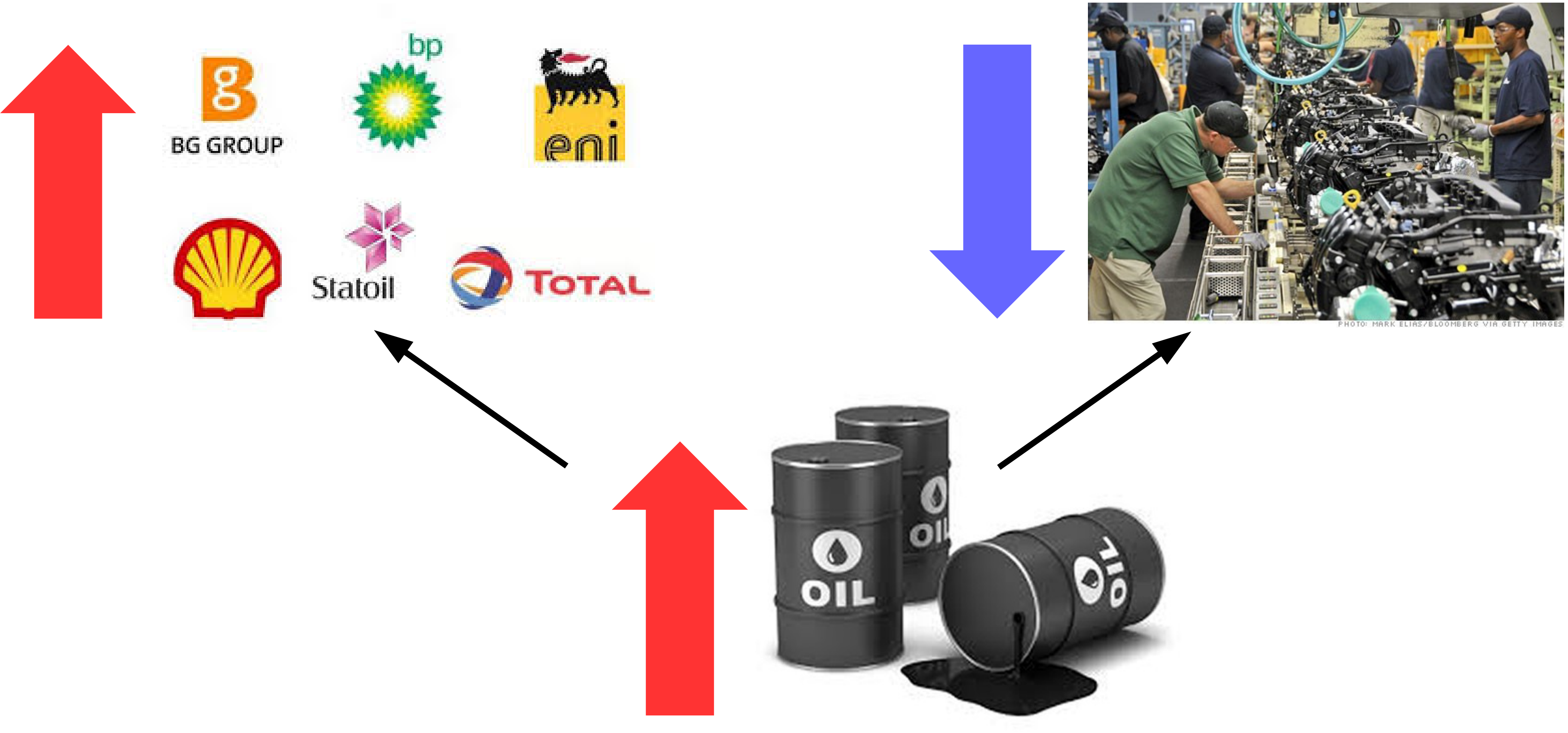
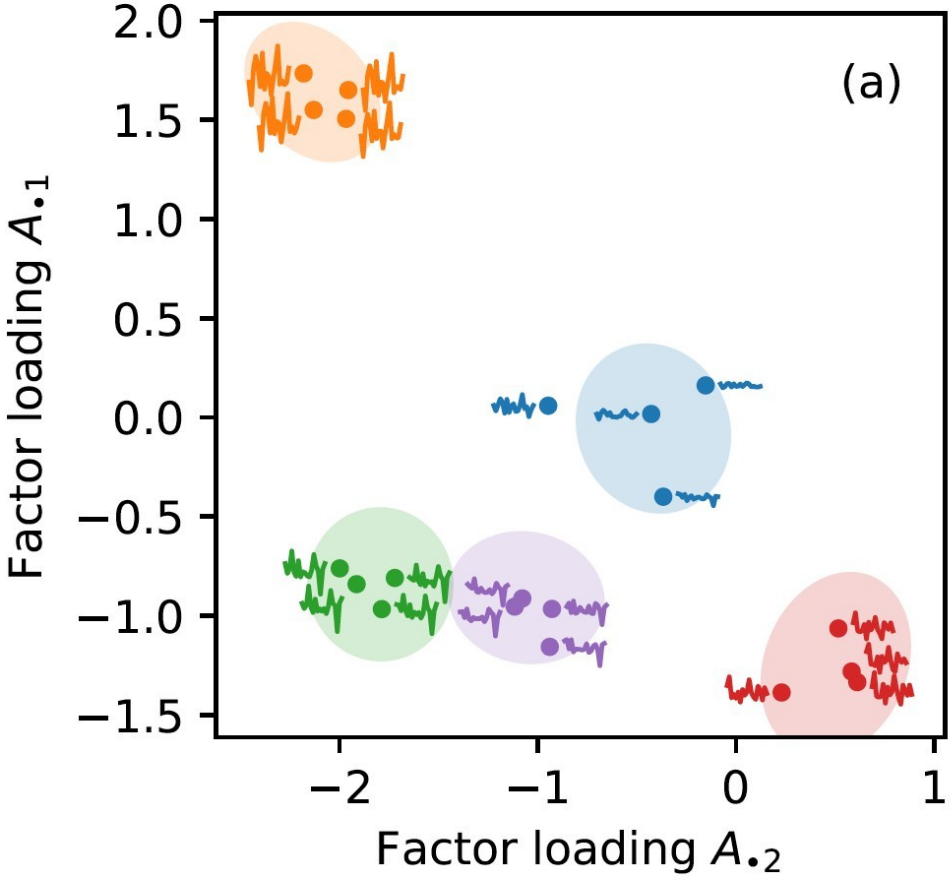


PHOTO: MARK ELIAS/BLOOMBERG VIA GETTY IMAGES

The signals we observe from many nodes are driven by a few latent factors



Notion of a community is: a group of nodes that influenced similarly by the latent factors

$$y_{ti} | A, x, \tau \sim \mathrm{Normal}\left(\sum_{q=1}^{p} x_{tq} A_{iq}, \tau_i^{-1}\right)$$

Observed time series

Latent factor time series

Factor loadings

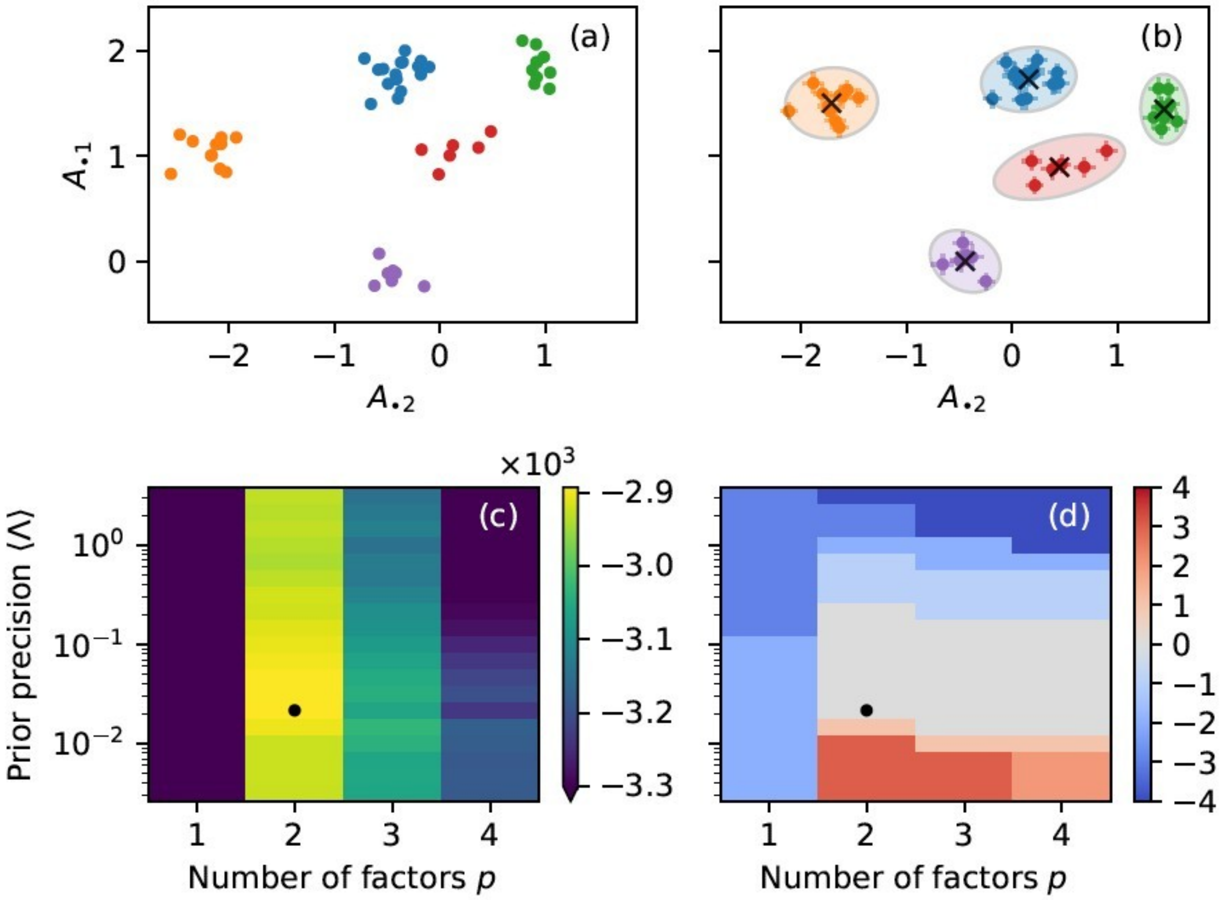
$$y_{ti}|A, x, \tau \sim \text{Normal}\left(\sum_{q=1}^{p} x_{tq}A_{iq}, \tau_i^{-1}\right)$$

$$A_i \sim \sum_{k=1}^{K} z_{ik}\text{Normal}\left(\overset{\text{Community mean}}{\mu_k}, \overset{\text{Community precision}}{\Lambda_k^{-1}}\right),$$

$$\text{where } z_{ik} = \begin{cases} 1 & \text{if } g_i = k \\ 0 & \text{otherwise} \end{cases}.$$

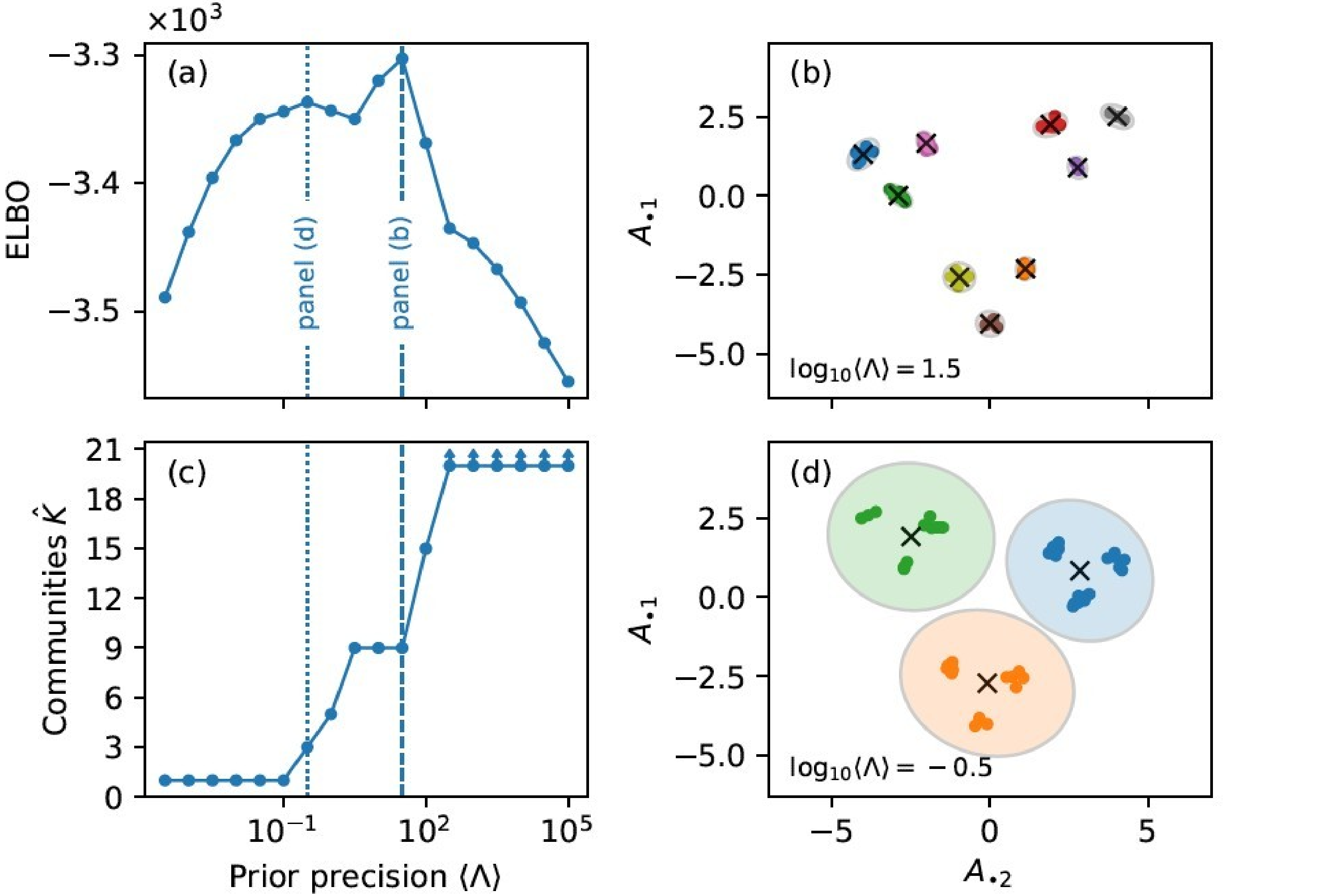
Generated

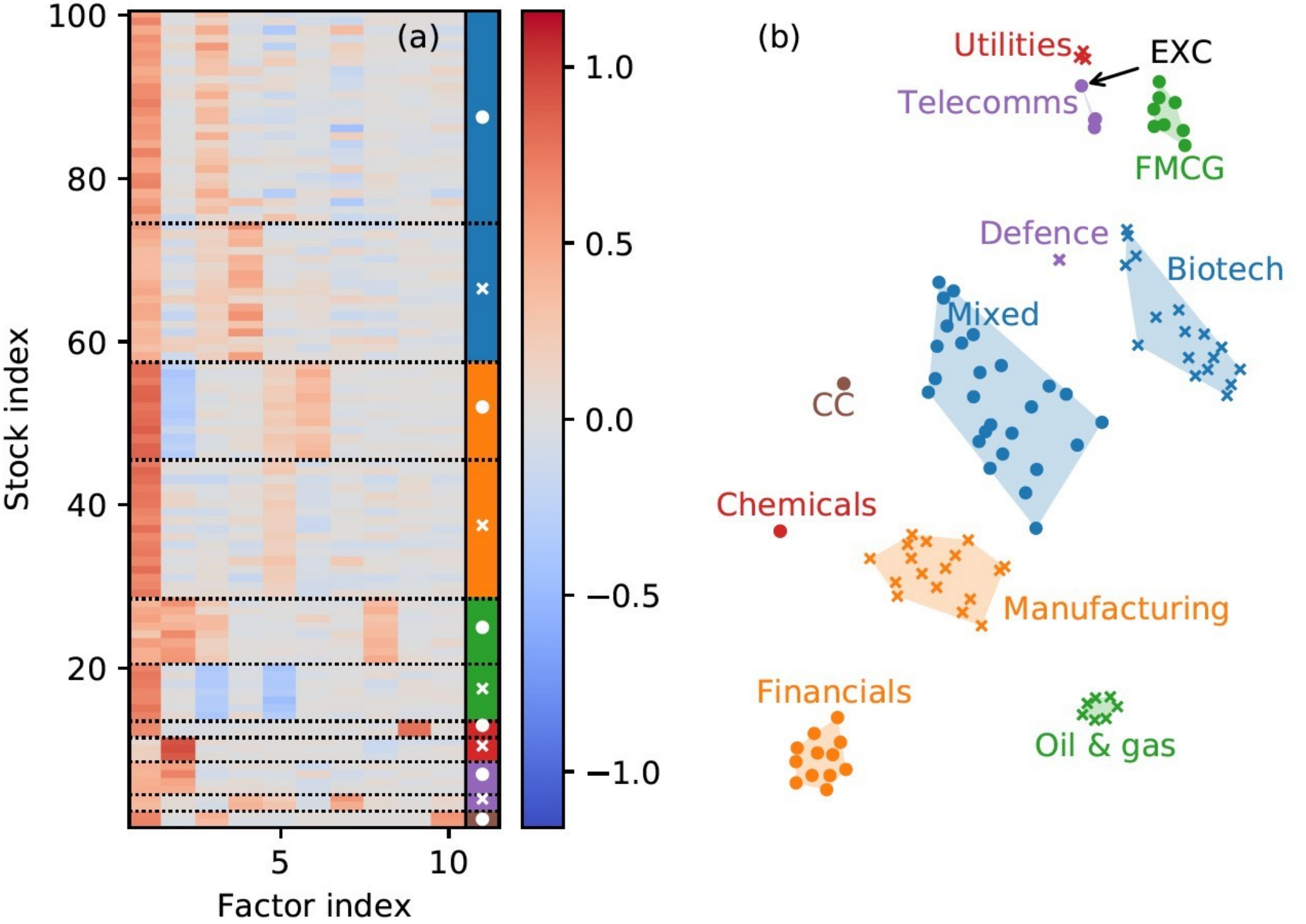
Inferred



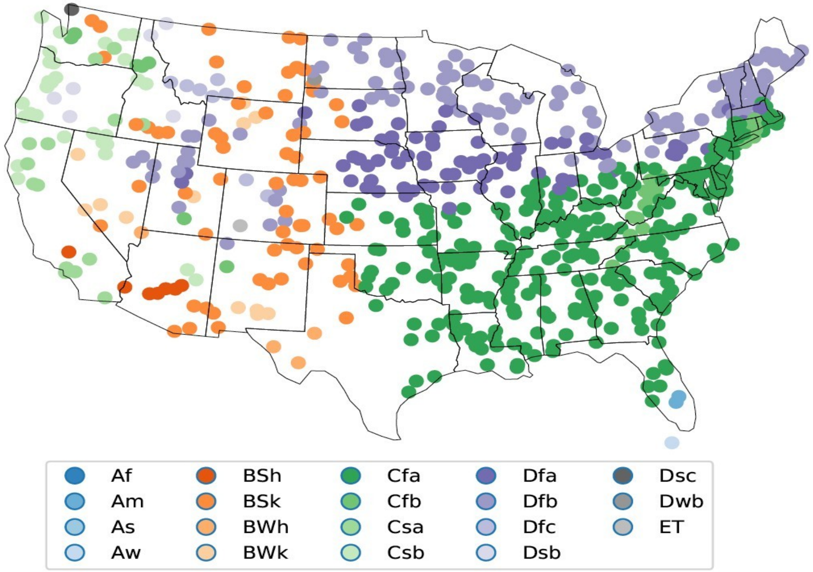
Lower bound on the marginal likelihood (ELBO)

Difference between $K_{generated}$ and $K_{inferred}$

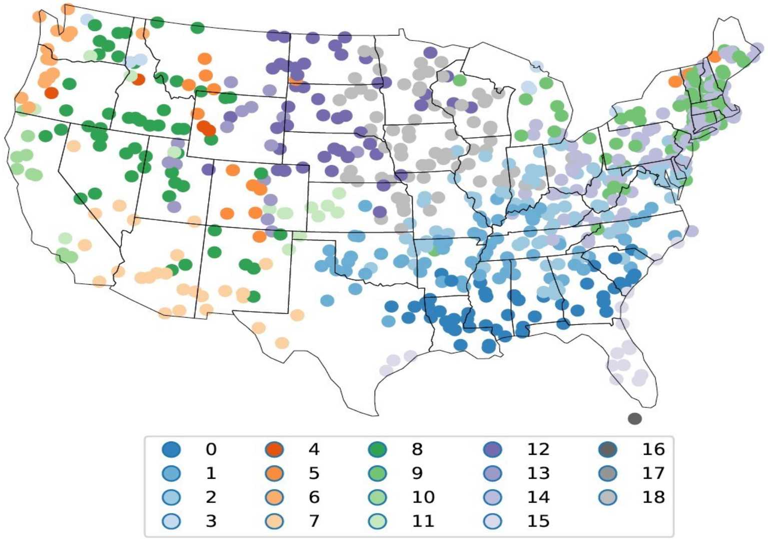

×10^3
−3.3
−3.4
−3.5
ELBO
(a)
panel (d)
panel (b)
(b)
2.5
0.0
−2.5
−5.0
A_{•1}
log_{10}⟨Λ⟩ = 1.5
(c)
21
18
15
12
9
6
3
0
Communities K̂
10^{−1}
10^2
10^5
Prior precision ⟨Λ⟩
(d)
2.5
0.0
−2.5
−5.0
A_{•1}
log_{10}⟨Λ⟩ = −0.5
−5
0
5
A_{•2}


(a)
Stock index
100
80
60
40
20
5
10
Factor index
1.0
0.5
0.0
−0.5
−1.0
(b)
Utilities
EXC
Telecomms
FMCG
Defence
Biotech
Mixed
CC
Chemicals
Manufacturing
Financials
Oil & gas

US cities climate data



Koppen climate zones

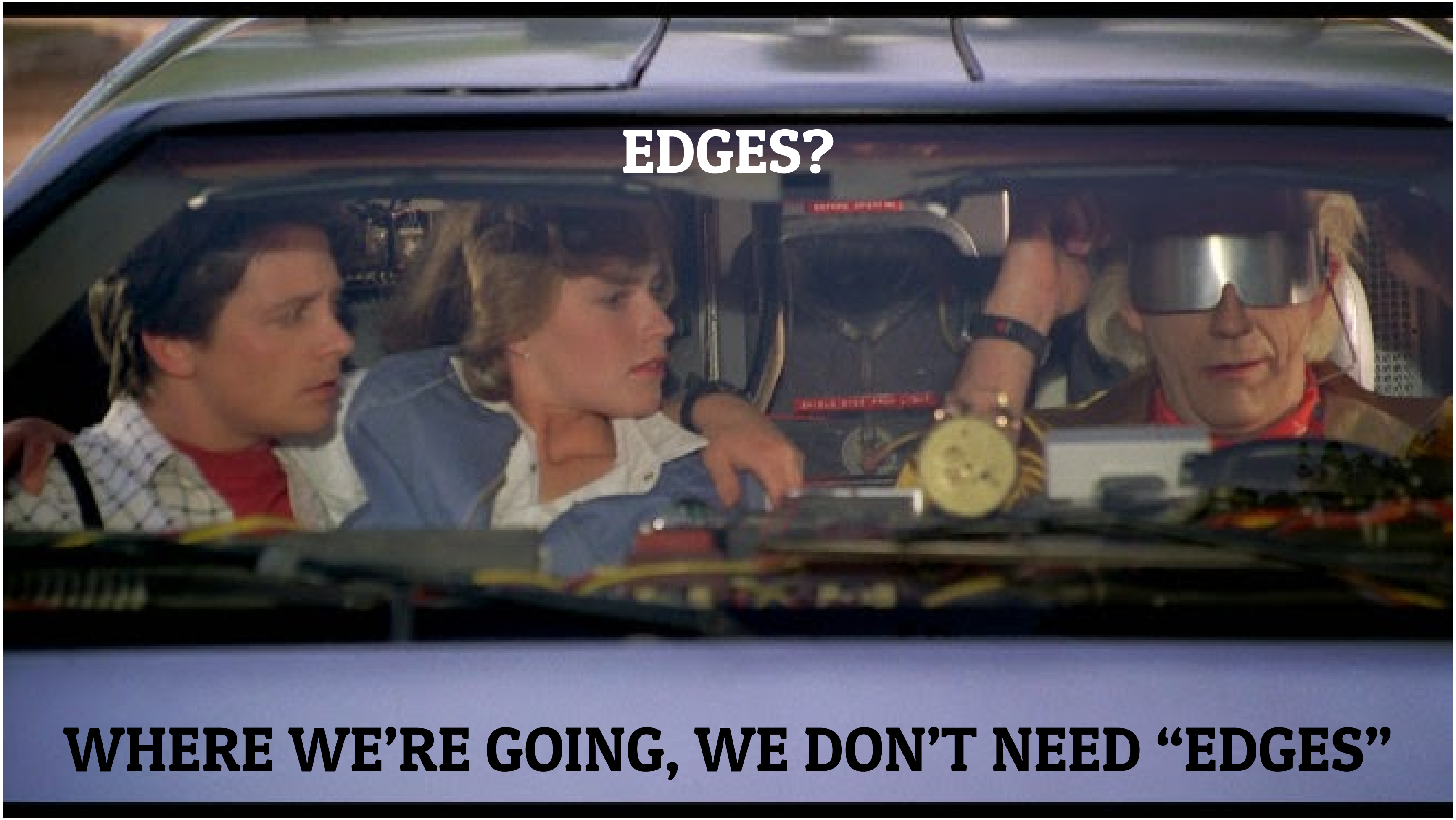


inferred climate zones

What happened to the network?

- Since we skip explicit interpretation of A our inference framework is basically a Bayesian (time-series) clustering.

- One can re-interpret AA^T as a network, or interpret distances between time-series in the latent-space as links in a network, but this is optional.


EDGES?
WHERE WE’RE GOING, WE DON’T NEED “EDGES”

In collaboration with…

Till
Hoffmann

Nick
Jones

Renaud
Lambiotte

Contact:

✉ leto.peel@uclouvain.be

@PiratePeel