Dynamic Mixed Membership Stochastic Block Model for Weighted Labeled Networks

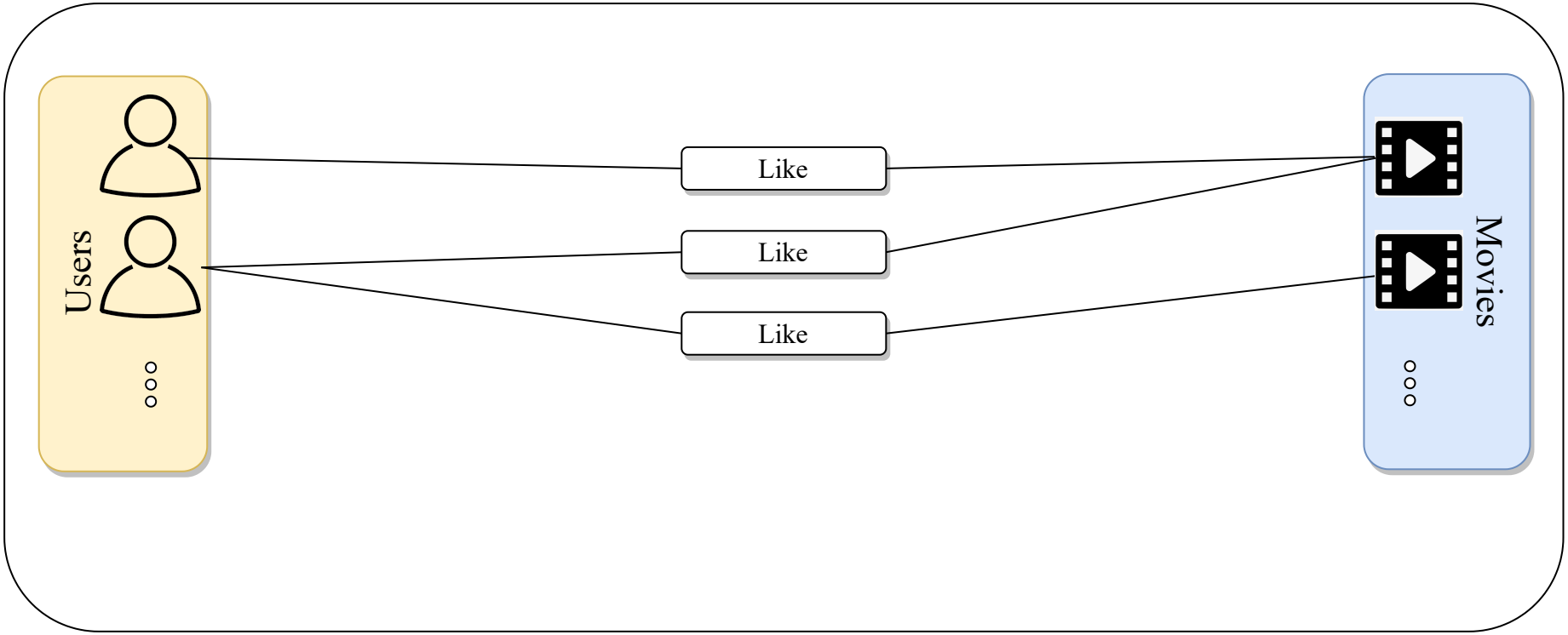
G. Poux-Médard, J. Velcin, S. Loudcher

Université de Lyon, France
Lyon 2, ERIC UR 3083

23 July 2023

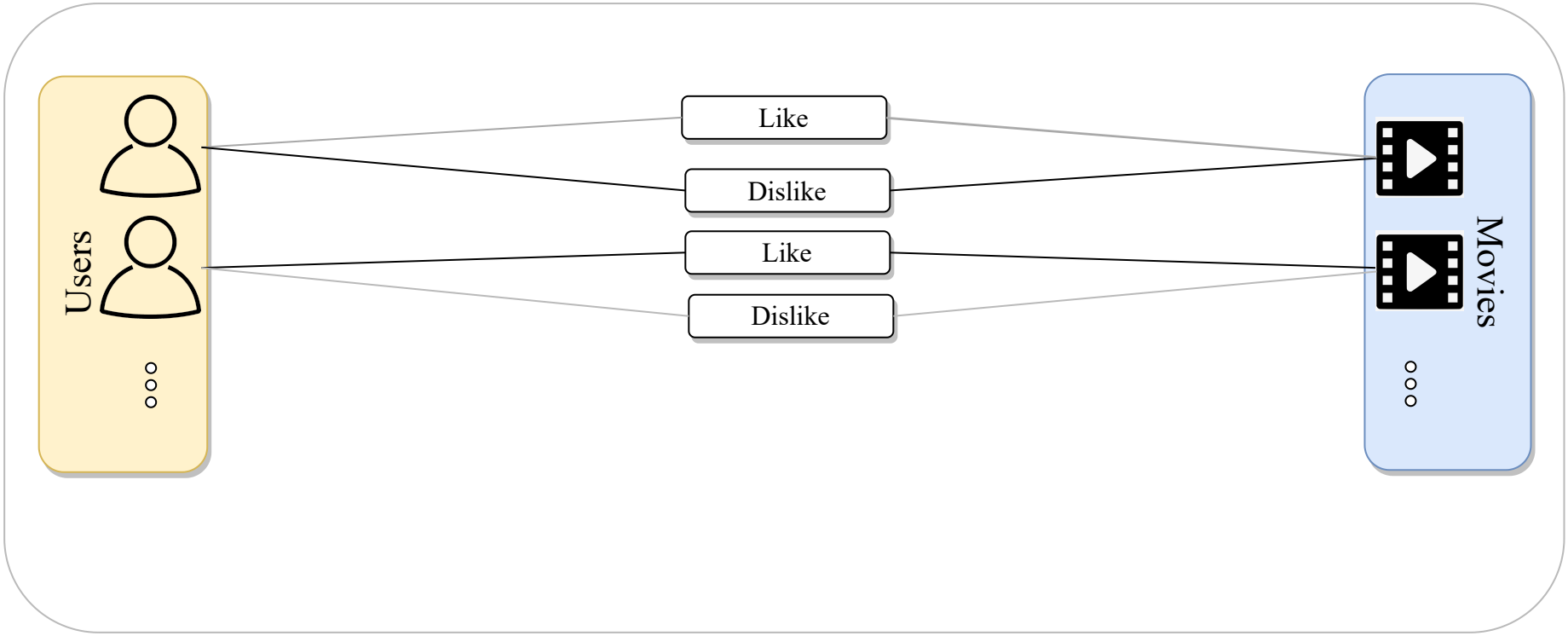
Networks

- Imagine a network where edges express users' opinion about movies
- Simple network: a link means a user liked a movie



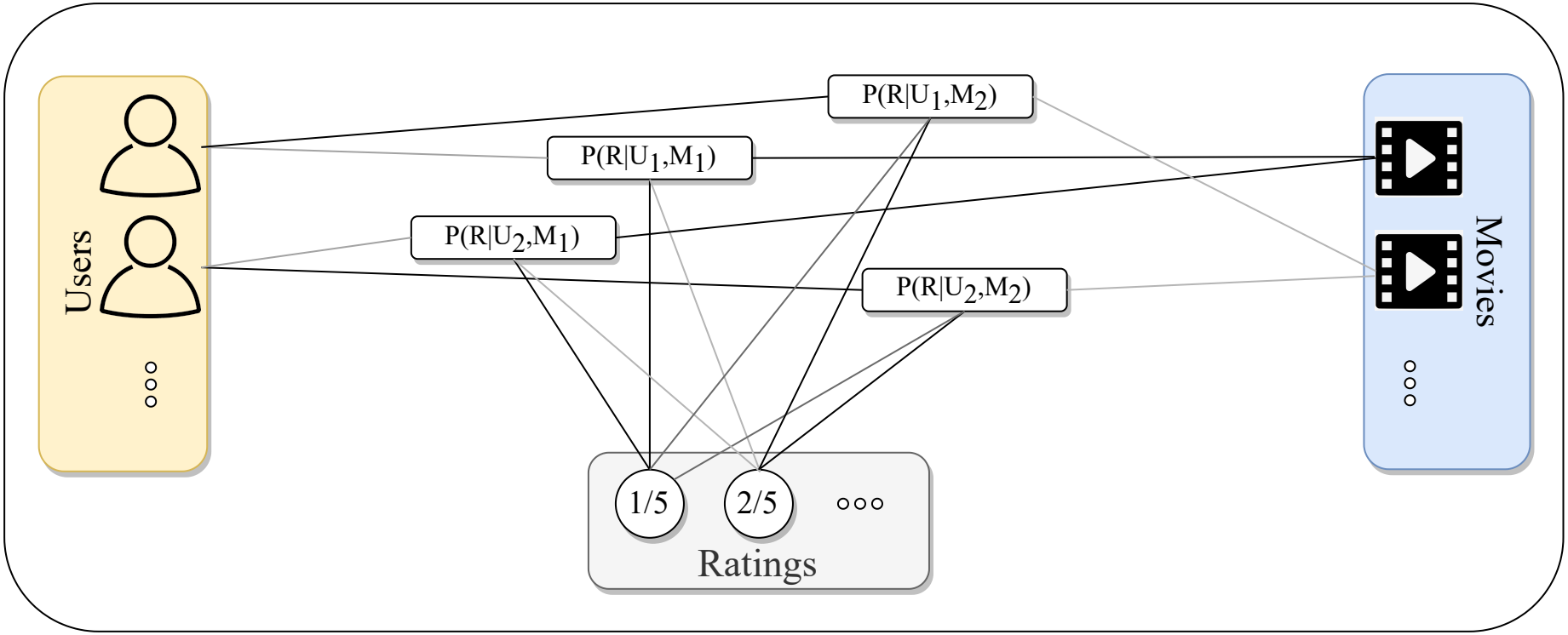
Weighted networks

- But tastes are seldom so clearly defined. Opinions can be mixed.
- Weighted network: valued links say how much a users like movies.



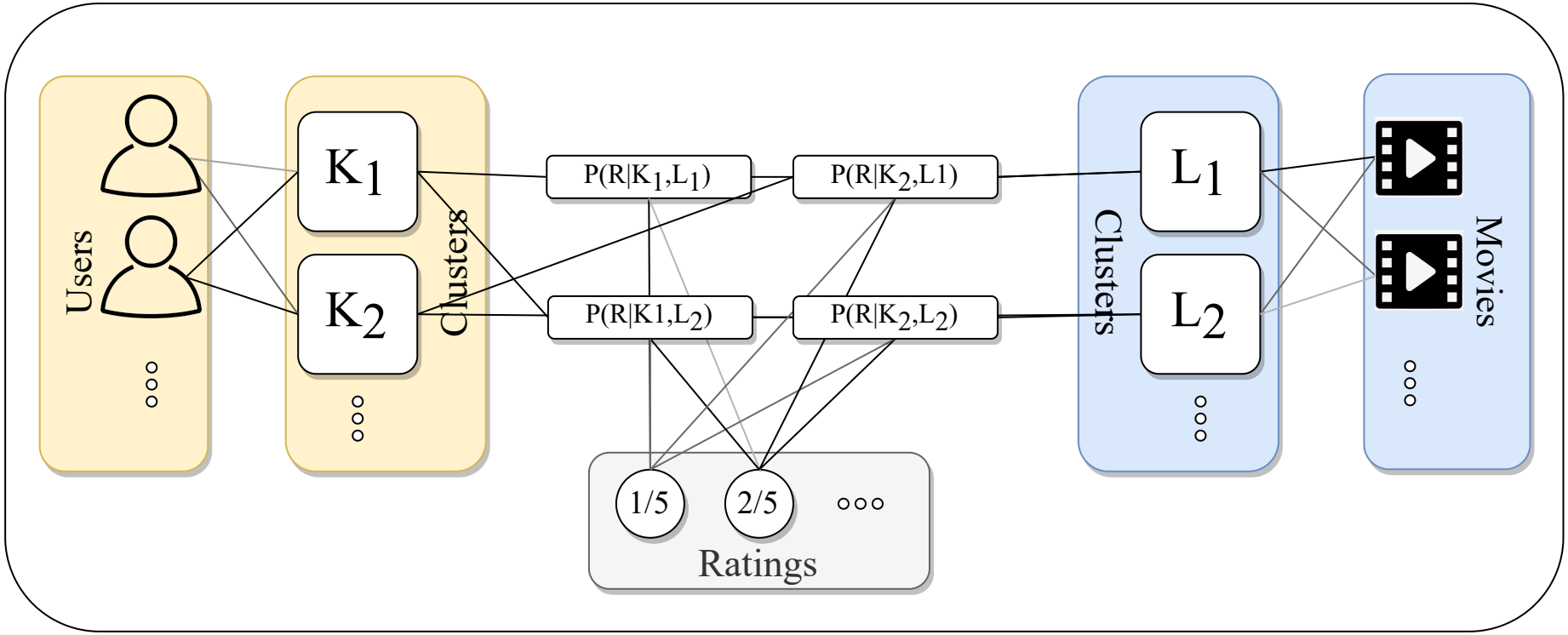
Weighted labeled networks

- What about more than two opinions? Or reactions (👍❤️😢😡👏)?
- Weighted labeled network: valued links express various opinions.



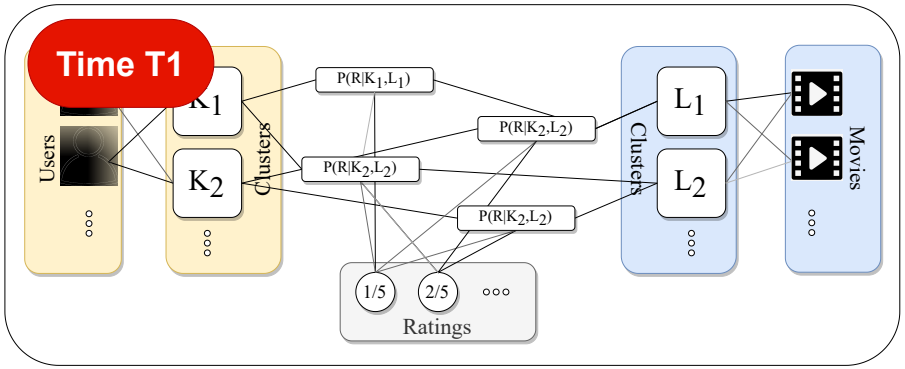
Block membership assumption

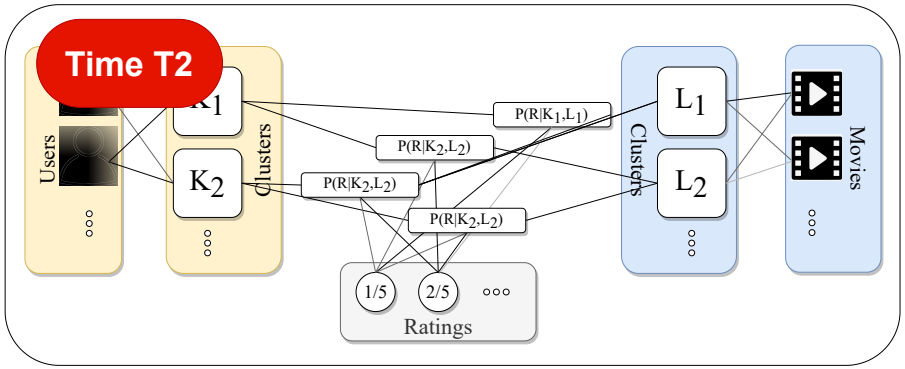
- Too many links for direct inference, but some items behave similarly.
- Blocks assumption: some users or movies can be grouped.

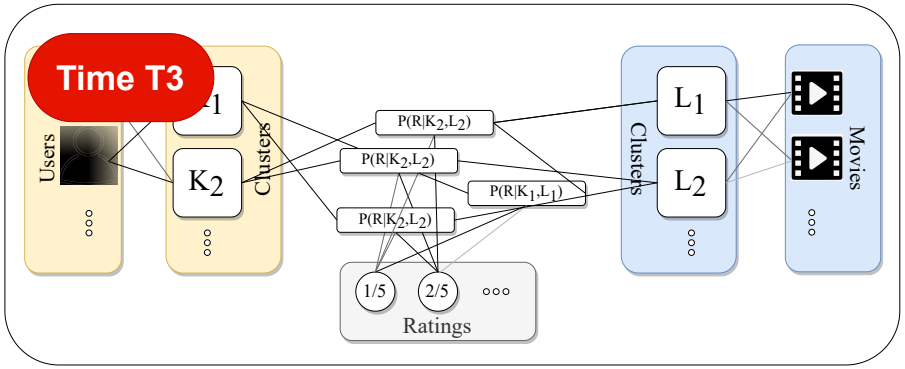


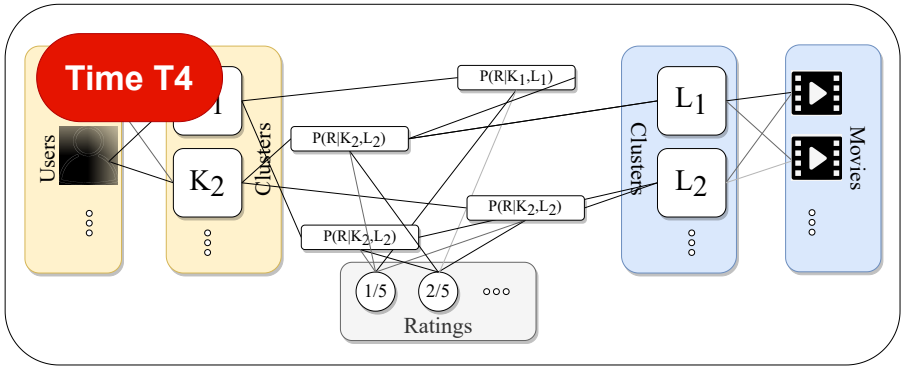
Dynamic weighted labeled networks

- Finally, tastes can evolve over time. User-movies networks seldom remain static.
- We need a blockmodel to infer weighted, labeled and dynamic networks.









Existing works

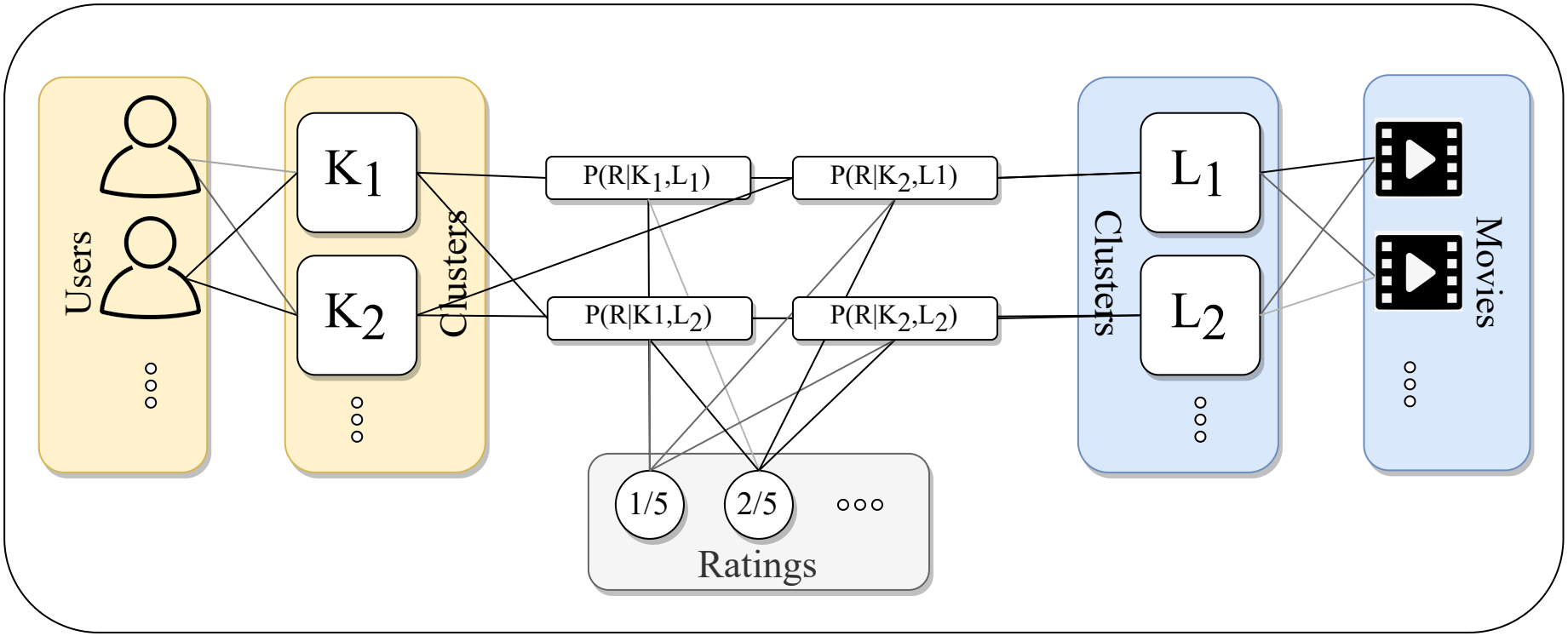
- Existing SBM-related works on:
  - Unweighted unlabeled dynamic networks
    - Refs discussed in the article: [16, 17, 21, 28, 29, 30]
  - Weighted unlabeled dynamic networks
    - Refs discussed in the article: [7, 11, 27]
  - Weighted labeled static networks
    - Refs discussed in the article: [9, 18, 19, 20, 22]

  → Detailed survey in (Lee *et al.*, 2019, Applied Network Science, 4)
- Weighted labeled dynamic networks: this presentation

Serialized interacting mixed membership SBM

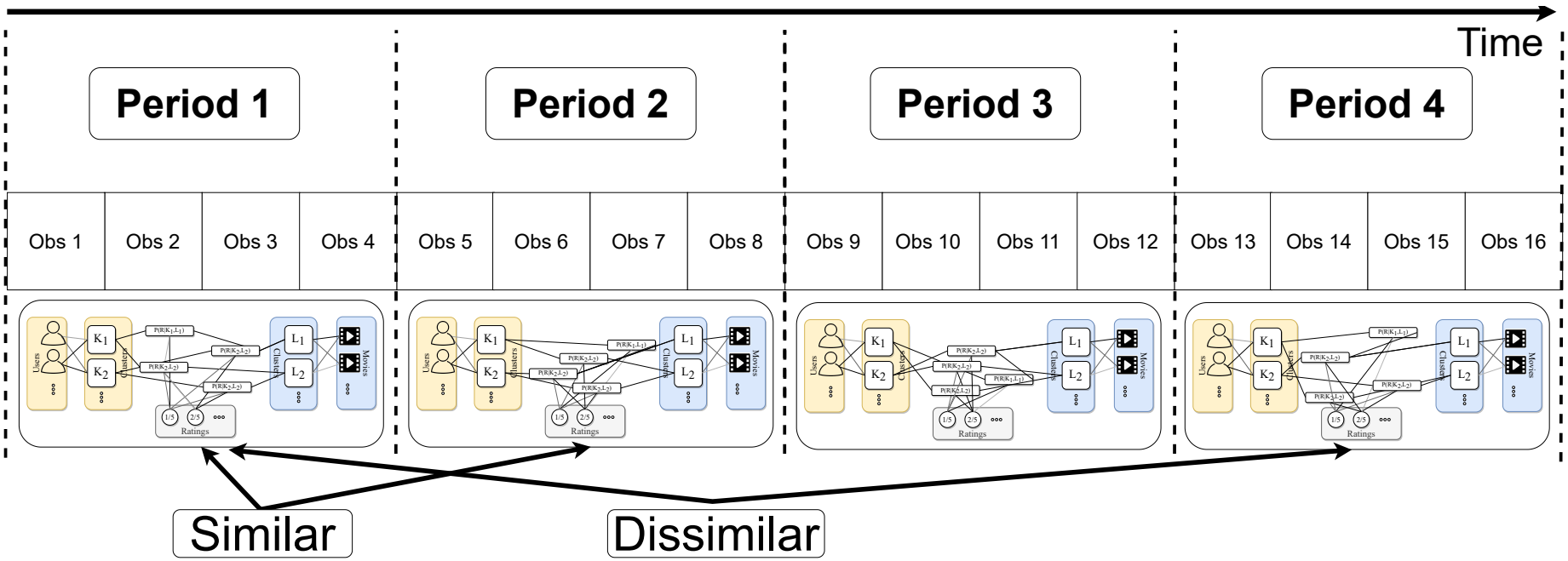
- Base model: (Poux-Médard *et al.*, 2022, ICDM)
  - Weighted labeled static networks

  → Must make it dynamic



Slicing and smooth evolution

- Slice the data and fit one model per epoch



- One assumption: models for epochs close in time should be similar

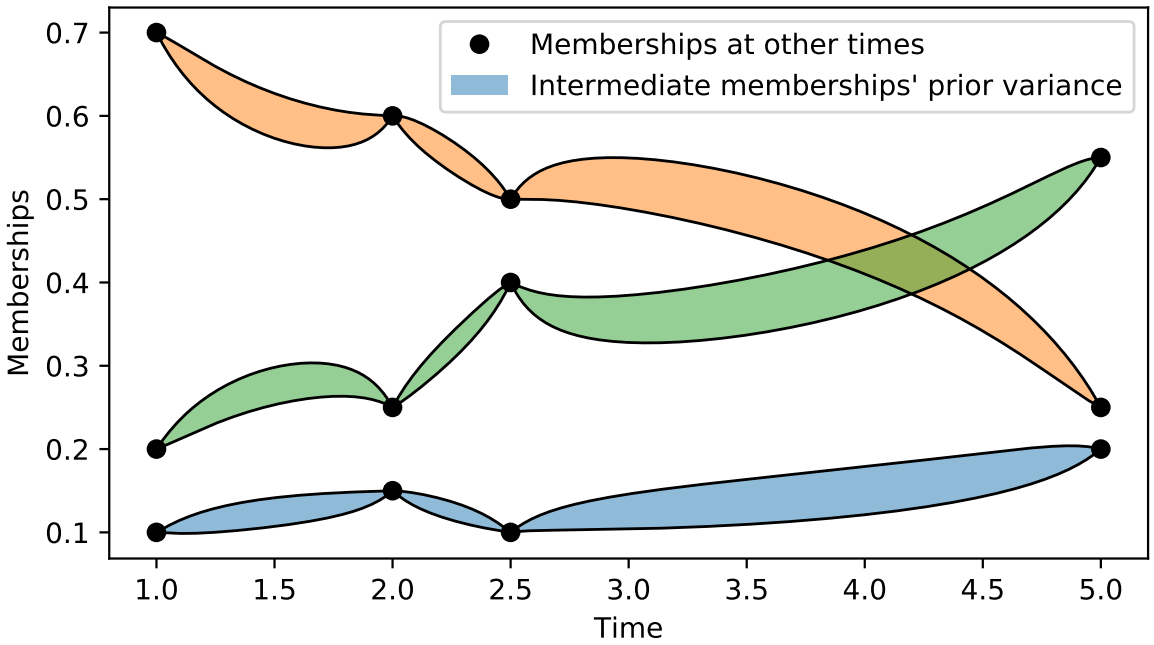
→ Smooth parameters variation

Method

- To enforce that: prior probability on each parameter

→ Distribution s.t. most probable value is the average:

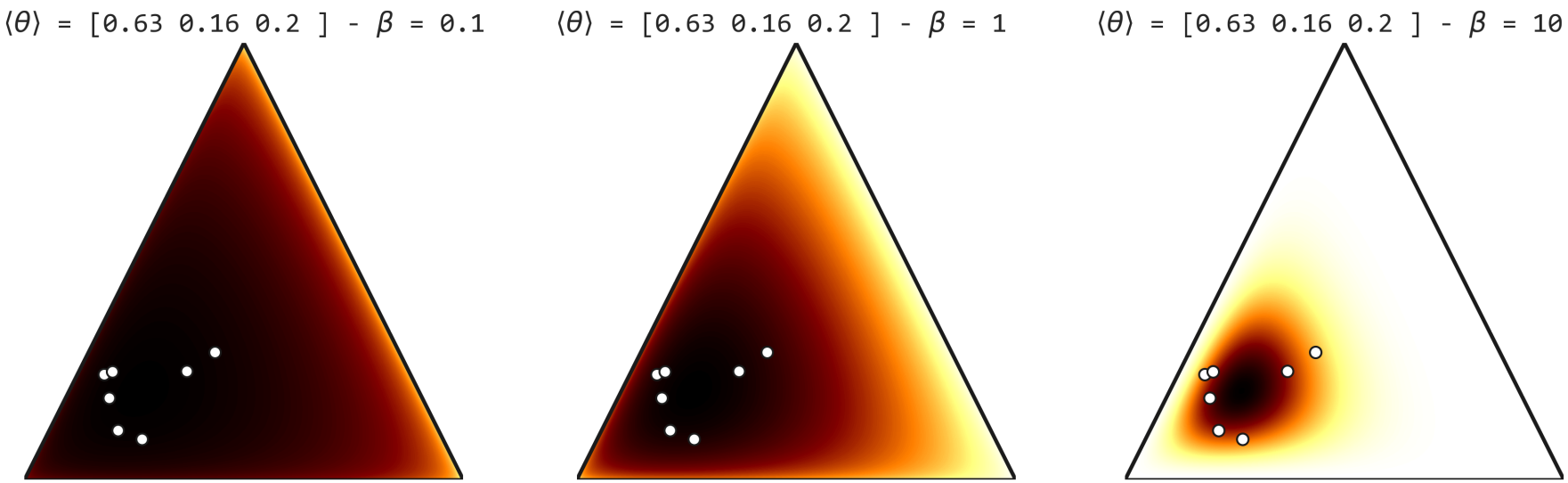
$$M[\theta(t)] = \frac{1}{T}\sum_{s \neq t}^{T} \theta(s) := < \theta(t) >$$



Expression of the prior

- In our case: parameters drawn from a Dirichlet distribution
  - White dots: temporal neighbours $\theta(s \neq t)$
  - Colors: prior probability for parameter $\theta(t)$
  - β: hyperparameter to control the variance
- Most frequent value (mode) $\mathbb{M}[Dir(\vec{x}|\vec{\alpha})|_i] \propto \vec{\alpha}|_i - 1$

$\rightarrow$ $P(\theta(\vec{t})) = Dir(\theta(\vec{t})|1 + < \theta(\vec{t}) >)$



Making static models dynamic

- Base model: SIMSBM (Poux-Médard, 2022, ICDM)
  - Generalizes static MMSBM (Godoy-Lorite, 2016, PNAS), (Tarres-Deulofeu, 2019, Phys.Rev.X) and (Poux-Médard, 2021, RecSys)
- Plug-in to make these models dynamic:

(Godoy-Lorite, 2016) + SDSBM

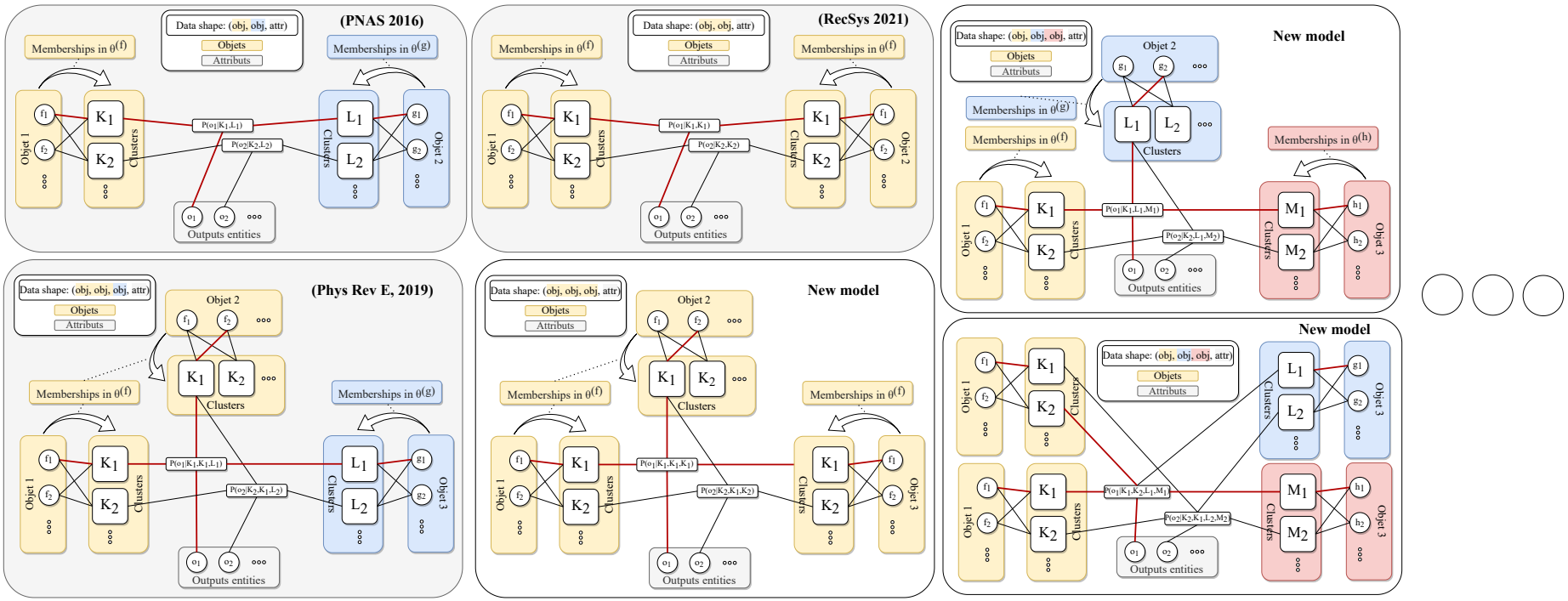
$$\begin{cases} \theta^{(t)}_{i,k} = \frac{\sum_l \sum_{(o,j)\in\partial(i,t)} \omega^{(t)}_{i,j,o}(k,l)+\beta\langle\theta^{(t)}_{i,k}\rangle}{N_{i,t}+\beta} \\ \eta^{(t)}_{j,l} = \frac{\sum_k \sum_{(o,i)\in\partial(j,t)} \omega^{(t)}_{i,j,o}(k,l)+\beta\langle\eta^{(t)}_{j,l}\rangle}{N_{j,t}+\beta} \\ p^{(t)}_{k,l}(o) = \frac{\sum_{(i,j,t)\in\partial o} \omega^{(t)}_{i,j,o}(k,l)+\beta\langle p^{(t)}_{k,l}(o)\rangle}{\sum_{(i,j,o,t)\in R^\circ} \omega^{(t)}_{i,j,o}(k,l)+\beta} \end{cases}$$

(Tarrès-Deulofeu, 2019) + SDSBM

$$\begin{cases} \theta^{(t)}_{h,k} = \frac{\sum_{l,m} \sum_{(o,i,j)\in\partial(h,t)} \omega^{(t)}_{h,i,j,o}(k,l,m)+\beta\langle\theta^{(t)}_{h,k}\rangle}{N_{h,t}+\beta} \\ \eta^{(t)}_{i,l} = \frac{\sum_{k,m} \sum_{(o,h,j)\in\partial(i,t)} \omega^{(t)}_{h,i,j,o}(k,l,m)+\beta\langle\eta^{(t)}_{i,l}\rangle}{N_{i,t}+\beta} \\ p^{(t)}_{k,l,m}(o) = \frac{\sum_{(h,i,j,t)\in\partial o} \omega^{(t)}_{h,i,j,o}(k,l,m)+\beta\langle p^{(t)}_{k,l,m}(o)\rangle}{\sum_{(h,i,j,o,t)\in R^\circ} \omega^{(t)}_{h,i,j,o}(k,l,m)+\beta} \end{cases}$$

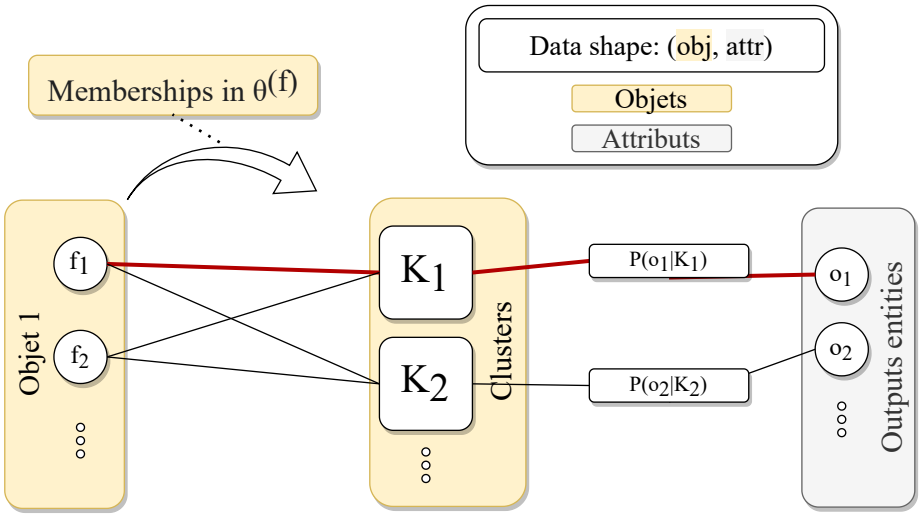
A note on SIMSBM

- SIMSBM allows to model a variety of situations:
  - Networks such as User-movie-ratings, Symptom-symptom-disease, Word-word-answer, User-producer-actor, and many other.
- SDSBM allows to make any of these dynamic
  - Simply add one term in the updates equations
  - Computational cost only scales with the number of epochs



Synthetic data

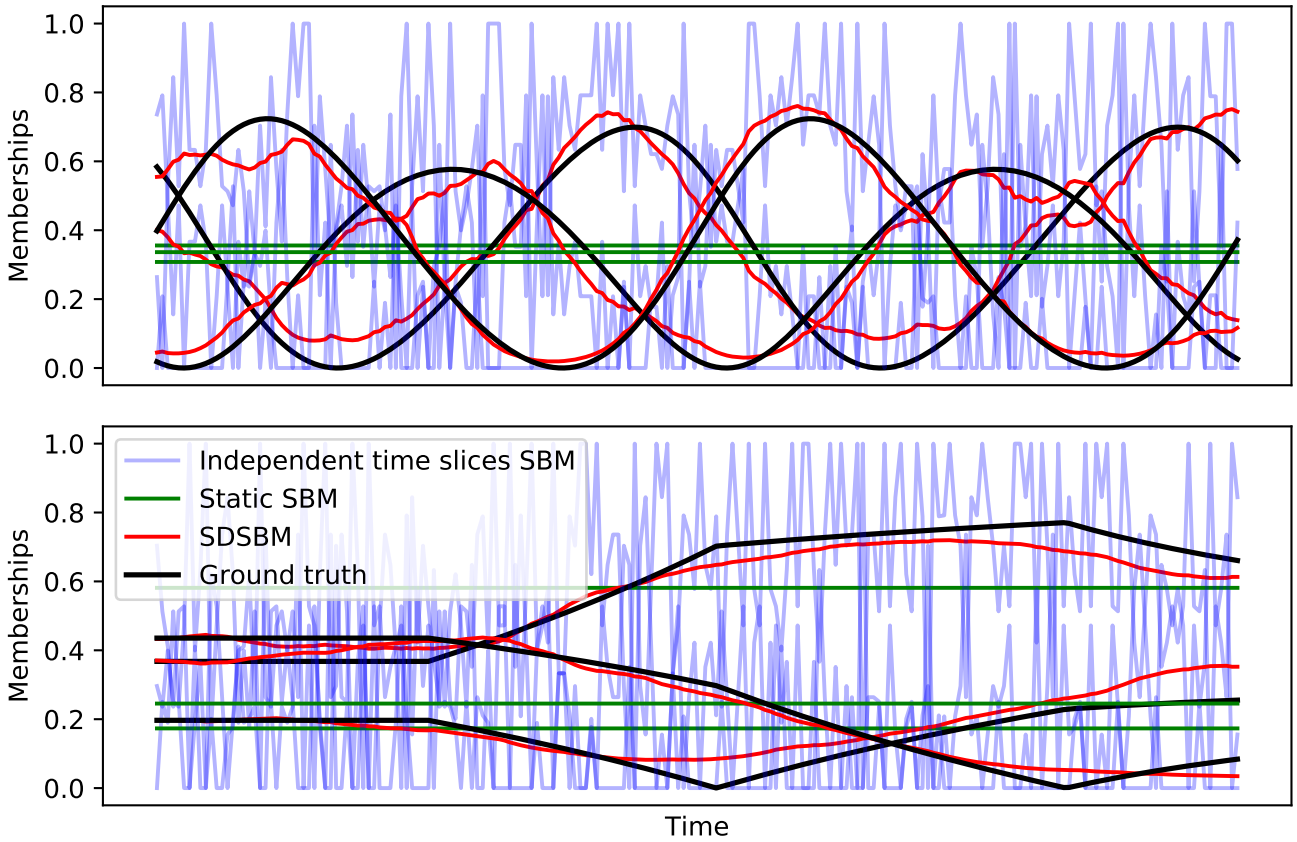
- Simplest model for demonstration purposes: SIMSBM(1)



- 100 items sharing 3 clusters choose between 3 outputs
- 1000 observations (item, output) sliced in 100 epochs
- 5-folds cross validation
- Eval. in prediction (AUCROC) and error on the params. (RMSE)
- Comparison to
  - MMSBM (Godoy-Lorite, 2016): the base, static model
  - T-MBM (Tarrès-Deulofeu, 2019): independent epoch

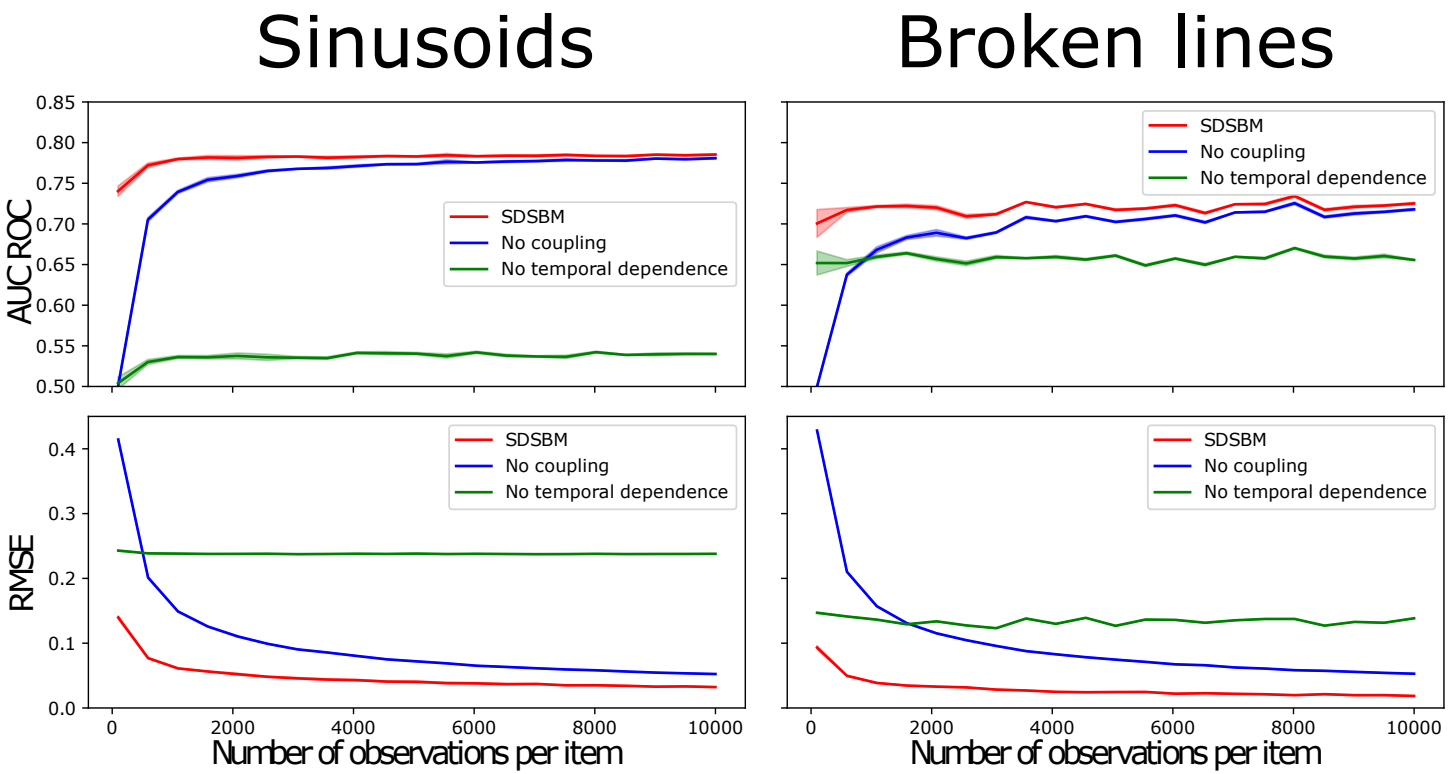
Qualitative results

- Two dynamic patterns: sinusoidal variations and random broken lines
- Reminder: our only hypothesis is that parameters vary smoothly



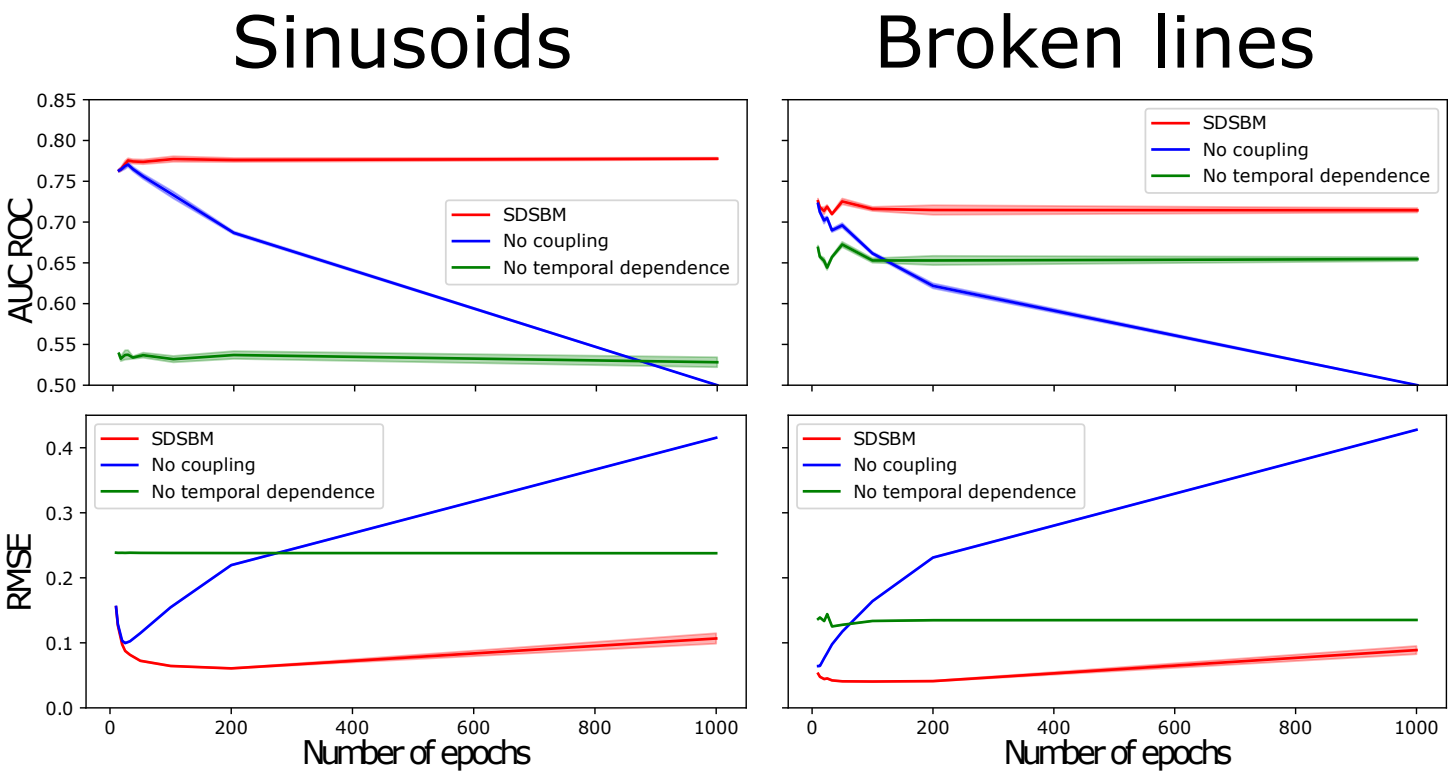
Synthetic data – Scarcity of observations

- Number of observations per item from 100 (1 obs/epoch) to 10.000 (100 obs/epoch)
- SDSBM works even with very few data



Synthetic data – Robustness against slicing

- Number of epochs from 10 to 1000, with 1k observations per item
- SDSBM is robust against the number of epochs



Real-world data – Datasets

- Three datasets from (Kumar *et al.*, 2019, KDD):
  - **Reddit**: 10k users interact with 1k subreddits over a month
    - Predict which subreddit a user interacts with at time t
  - **Wikipedia**: 8k users edit 1k pages over a month
    - Predict which page a user edits at time t
  - **LastFM**: 1k users listen to 1k songs over a month
    - Predict which song a user listens to at time t
- One dataset gathered from the Clauss-Slaby DB: the **Epigraphy** dataset
  - 117k ancient roman engravings from 100BC to 500AD
  - 18 status (slave, soldier, senator, etc.)
  - 62 regions (Latium, Hispania, Gallia Narbonensis, Syria, etc.)
  - Predict the region where a status appeared at time t

Real-world data – Numerical results

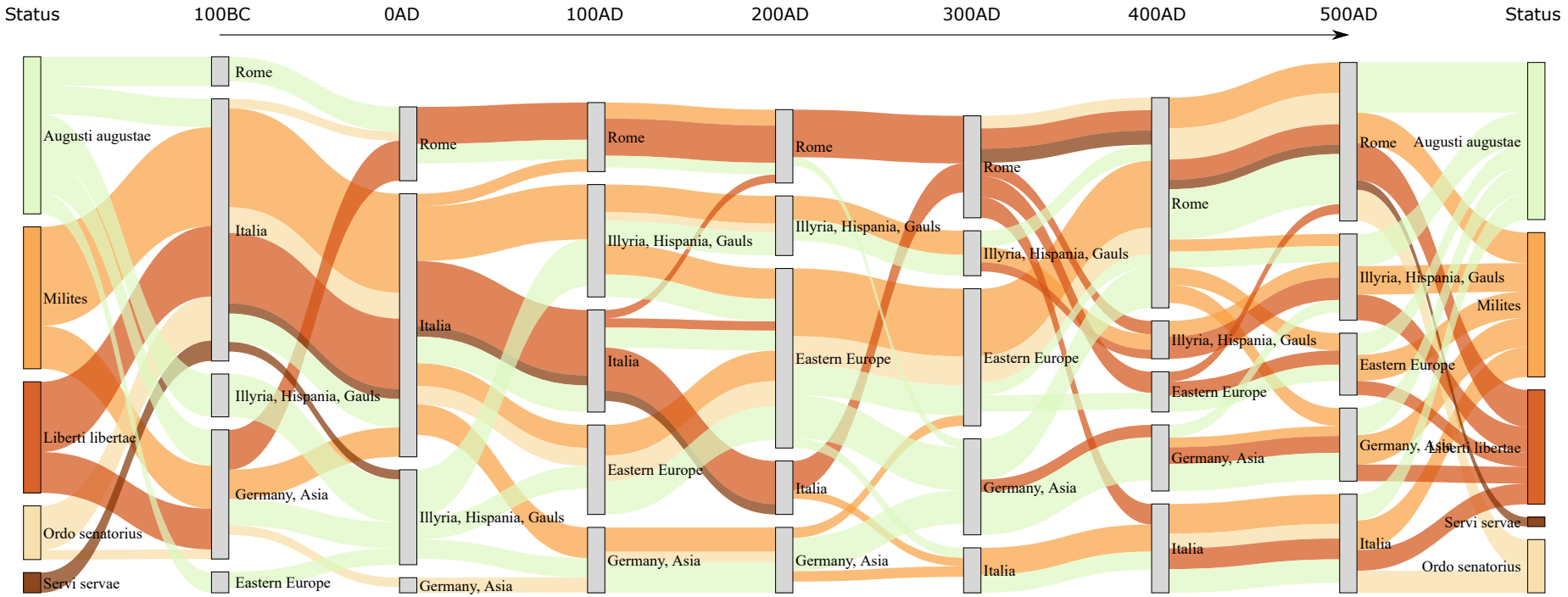
- SDSBM outperforms static and non-coupling baselines

		ROC	AP	NCE
Epi	SDSBM	**0.9025(11)**	**0.3700(17)**	**0.1151(11)**
	NC	0.8420(22)	0.3435(36)	0.1582(19)
	SIMSBM(1)	0.8597(12)	0.2141(16)	0.1451(13)
Lastfm	SDSBM	**0.8942(8)**	**0.0168(1)**	**0.1284(11)**
	NC	0.8393(5)	0.0157(2)	0.1785(7)
	SIMSBM(1)	0.8647(5)	0.0115(2)	0.1493(4)
Wiki	SDSBM	**0.9759(2)**	**0.0659(9)**	**0.0459(3)**
	NC	0.9092(7)	0.0608(10)	0.1195(8)
	SIMSBM(1)	0.9576(7)	0.0622(4)	0.0565(8)
Reddit	SDSBM	**0.9803(3)**	**0.4295(54)**	**0.0312(3)**
	NC	0.8508(5)	0.3598(17)	0.1846(7)
	SIMSBM(1)	0.9798(2)	**0.4269(40)**	0.0322(3)

Real-world data – Epigraphy dataset

- Output for the Epigraphy dataset
- Some historical facts that we retrieve:
  - Italy demilitarization around 100AD
  - Spread of libertii through Europe
  - 3rd century crisis and military reinforcement of the capital

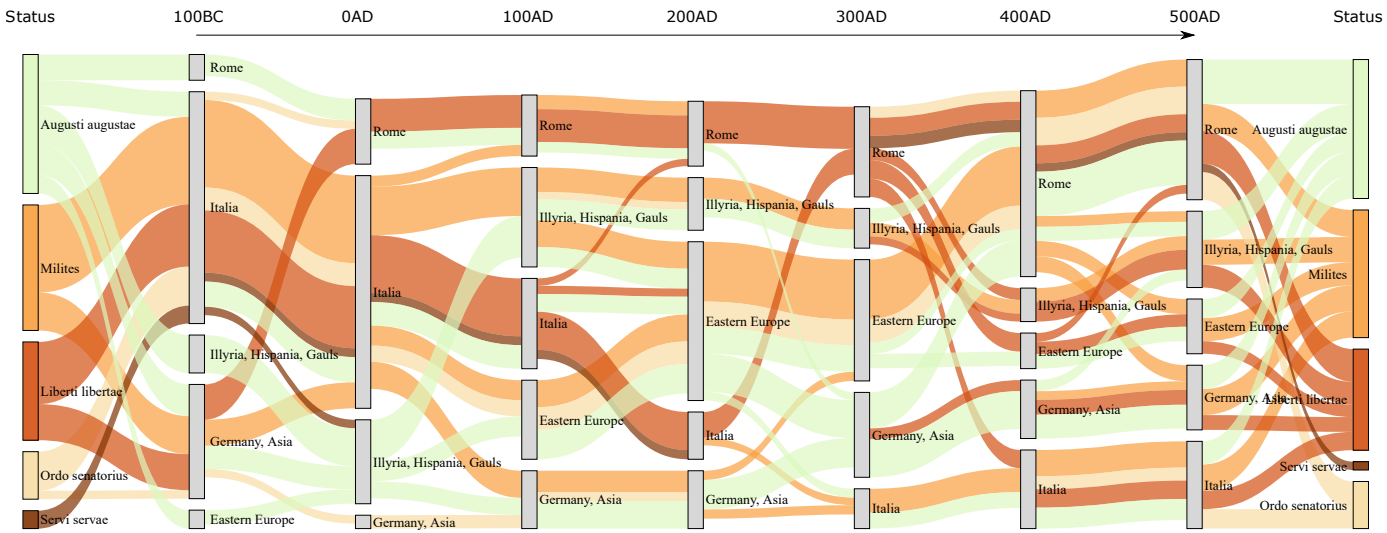
→ A dedicated study would be needed to go beyond mere hints



Conclusion and perspectives

- Conclusions:
  - Method to model weighted, labeled and dynamic networks
  - A single hypothesis: parameters vary smoothly
  - Plug-in for existing models
  - Works when data is scarce $\rightarrow$ interest for humanities
- Perspectives:
  - Get rid of the slicing to infer continuous parameters
    - Expressing the Dirichlet prior as a CRP?
    - Kernel methods?
  - Infer the hyperparameter β
  - Infer the averaging kernel

Thanks for your attention!



More here

- Webpage: `https://gaelpouxmedard.github.io/`
- Code and data: `https://github.com/GaelPouxMedard/SIMSBM/`