Serialized Interacting Mixed-Membership Stochastic Block Model

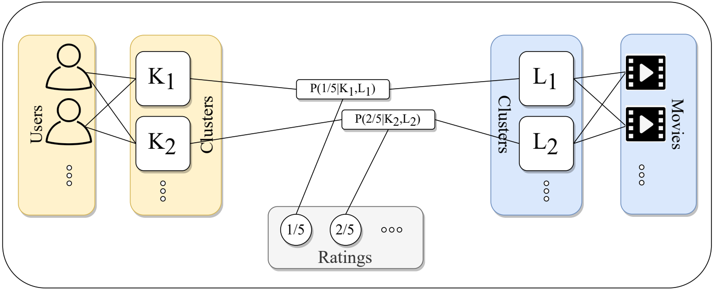
G.Poux-Médard, J. Velcin, S. Loudcher

Université de Lyon, France
Lyon 2, ERIC UR 3083

30/11/2022

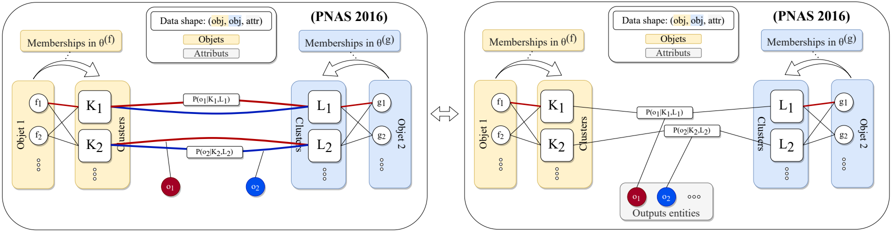
Introduction

- Classic problem: how users rate movies?
  - → Assume regularities (clusters)
- Historically: Matrix Factorization in 2009 (Y. Koren)
  - Works for linear ratings and triplets (user, movie, rating) only
- Recent developments: Mixed Membership Stochastic Block Models



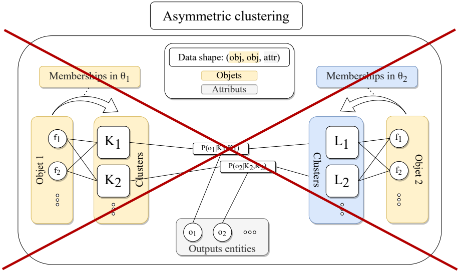
Labeled edges

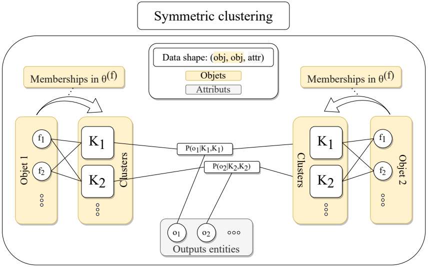
- What if ratings **do not scale linearly** (👍❤️😢😡👏...)
  - → Consider labeled edges
- (Godoy-Lorite *et al.*, 2016, PNAS)
  - Edges represent a distribution of labels (instead of a scalar weight)
  - Group memberships instantiate a PDF over labels



Interactions

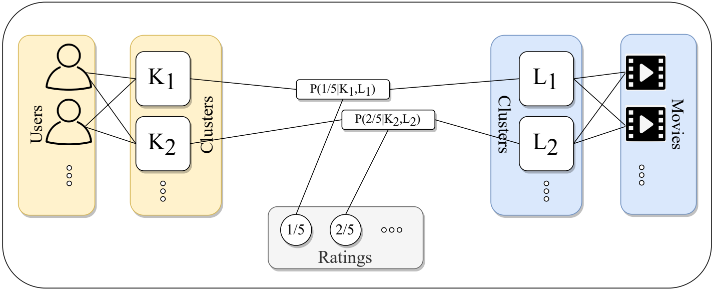
- What if the user watched **several movies** in a row?
  - $\rightarrow$ Consider information interaction
- (Poux-Médard *et al.*, 2019, RecSys)
  - Layers can share membership matrices
  - Symmetric interaction between items of the same type
    - $\rightarrow$ P(rating|movie1,movie2)=P(rating|movie2,movie1)





What if...

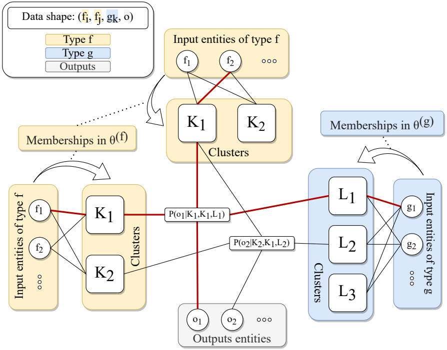
- What if ratings **do not scale linearly** (👍❤️😢😡👏...)
  → Consider labeled edges (Godoy-Lorite *et al.* 2016)
- What if the user watched **several movies** in a row?
  → Consider information interaction (Poux-Médard *et al.* 2021)
- What if we want to consider **additional information** (time)?
  → Consider additional layers of information
- What if we want **all of this** at once?
  → Serialized interacting MMSBM (this presentation)



SIMSBM - Mathematically

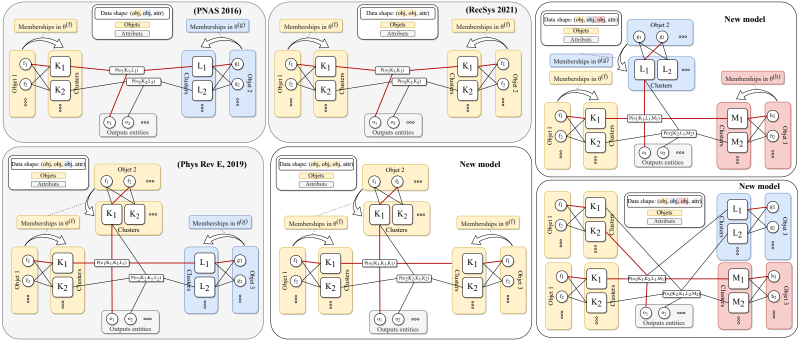
$$\mathcal{L} = \prod_{(\mathbf{f},o)\in R^{\circ}} \left(\sum_{\mathbf{k}} p_{\mathbf{k}}(o) \prod_{n} \theta^{(a(f_n))}_{f_n,k_n} \right)$$

- $\theta^{(a)}$: membership matrix for type x
- p: block-interaction tensor
- $\mathbf{k}$: a permutation of index clusters
- $\mathbf{f}$: collection of input items
- o: output item (or edge label)



SIMSBM - Graphically

- Generalises (PNAS 2016), (Phys.Rev.E, 2019) and (RecSys 2021)
  - → Any number of layers and interactions
  - → Unified framework for model selection

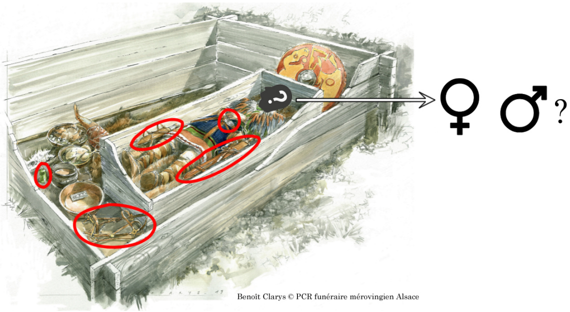


Optimization and complexity

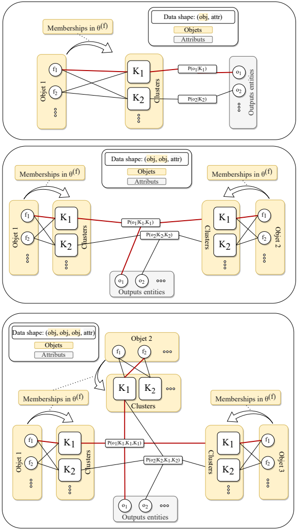
$$\mathcal{L} = \prod_{(\mathbf{f},o)\in R^\circ} \left(\sum_{\mathbf{k}} p_{\mathbf{k}}(o) \prod_{n} \theta_{f_n,k_n}^{(a(f_n))} \right)$$

- EM algorithm $\rightarrow$ Convergence to local optimum
- Scales **linearly** with the number of unique data entries $(\mathbf{f}, o)$

Applications - Archaeology

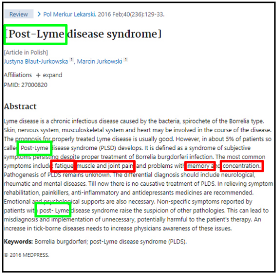


- 2000 items
- 2 genders
- 26 000 n-plets



Applications - Pubmed and Spotify

Symptoms ▶ Diseases?


Review > Pol Merkur Lekarski. 2016 Feb;40(236):129-33.

[Post-Lyme disease syndrome]

[Article in Polish]
Justyna Błaut-Jurkowska [1], Marcin Jurkowski [1]

Affiliations + expand
PMID: 27000820

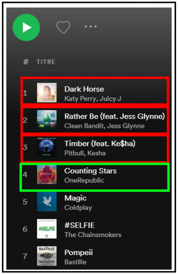
Abstract

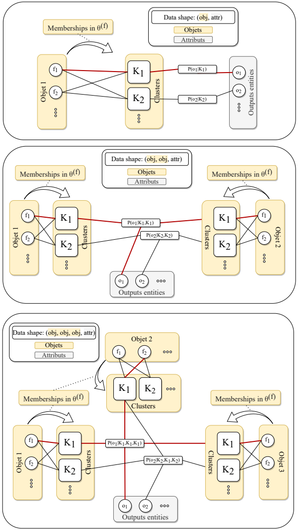
Lyme disease is a chronic infectious disease caused by the bacteria, spirochete of the Borrelia type. Skin, nervous system, musculoskeletal system and heart may be involved in the course of the disease. The prognosis for properly treated Lyme disease is usually good. However, in about 5% of patients so called Post-Lyme disease syndrome (PLSD) develops. It is defined as a syndrome of subjective symptoms persisting despite proper treatment of Borrelia burgdorferi infection. The most common symptoms include fatigue, muscle and joint pain and problems with memory and concentration. Pathogenesis of PLDS remains unknown. The differential diagnosis should include neurological, rheumatic and mental diseases. Till now there is no causative treatment of PLDS. In relieving symptom rehabilitation, painkillers, anti-inflammatory and antidepressants medicines are recommended. Emotional and psychological supports are also necessary. Non-specific symptoms reported by patients with post- Lyme disease syndrome raise the suspicion of other pathologies. This can lead to misdiagnosis and implementation of unnecessary, potentially harmful to the patient's therapy. An increase in tick-borne diseases needs to increase physicians awareness of these issues.

Keywords: Borrelia burgdorferi; post-Lyme disease syndrome (PLDS).

© 2016 MEDPRESS.

Songs ▶ Next song?

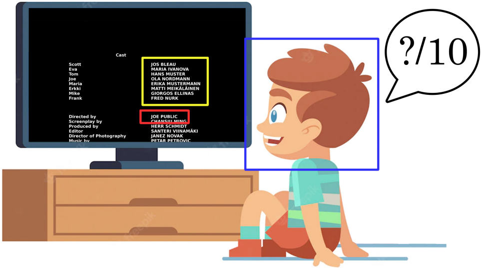


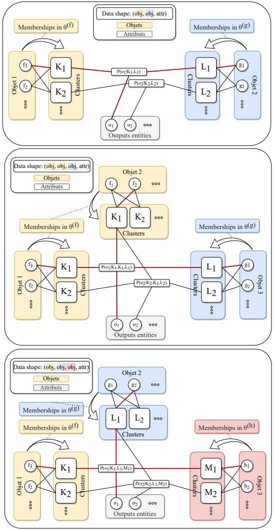


- 322 symptoms
- 4442 diseases
- 2 000 000 n-plets

- 2028 artists
- 2028 artists
- 50 000 n-plets

Applications - Imdb





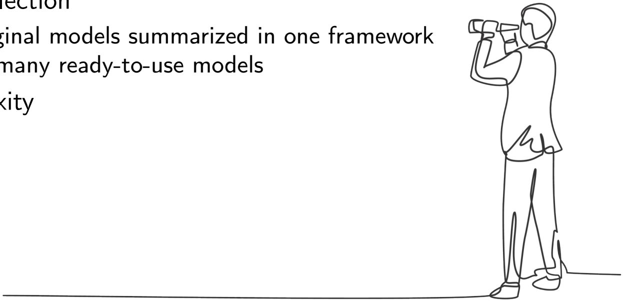
- 2502 users
- 809 actors
- 255 directors
- 10 ratings
- 1 000 000 n-plets

Applications - Quantitative results

			F1	P@1	AUCROC	AUCPR	RankAvgPrec	CovErrNorm
MrBanks	Ply, Sit (3), Gen, Age	SIMSBM(1,1,1,1)	0.7124(2)	0.6549(3)	0.7071(2)	0.7141(3)	0.8274(1)	0.1726(1)
		SIMSBM(1,2,1,1)	0.7107(2)	0.6696(5)	0.7120(4)	0.7158(5)	0.8348(3)	0.1652(3)
		SIMSBM(1,3,1,1)	**0.7348(2)**	**0.7172(5)**	**0.7610(4)**	**0.7646(4)**	**0.8586(3)**	**0.1414(3)**
		TF	0.6795	0.6037	0.4702	0.4967	0.8019	0.1981
		NMF	0.7178	0.6976	0.7232	0.7182	0.8409	0.1591
		KNN	0.7023	0.6648	0.6859	0.6623	0.8324	0.1676
		NB	0.6867	0.6382	0.6323	0.6250	0.8191	0.1809
Spotify	Artists (3)	SIMSBM(1)	0.1741(4)	0.2155(7)	**0.7908(6)**	0.1603(3)	0.3827(4)	0.0786(3)
		SIMSBM(2)	0.3156(5)	**0.3348(4)**	0.7661(5)	0.2545(3)	**0.4528(3)**	0.0938(6)
		SIMSBM(3)	**0.3243(4)**	0.3209(3)	0.7384(6)	**0.2613(3)**	0.4366(3)	0.1079(7)
		TF	0.0262	0.0042	0.4805	0.0159	0.0962	0.1550
		NMF	0.0371	0.0658	0.5650	0.0403	0.1762	0.2557
		KNN	0.3201	0.3009	0.7079	0.2400	0.3941	0.5212
		NB	0.0463	0.0846	0.7005	0.0576	0.2264	**0.0763**
PubMed	Symptoms (3)	SIMSBM(1)	0.2915(2)	0.5576(4)	0.7475(1)	0.2658(1)	0.4641(1)	0.2033(1)
		SIMSBM(2)	0.3127(1)	0.5704(1)	0.7613(1)	0.2840(1)	0.4838(1)	0.1991(1)
		SIMSBM(3)	**0.3219(1)**	**0.5790(1)**	**0.7666(1)**	**0.2895(1)**	**0.4937(1)**	**0.1983(1)**
		TF	0.1607	0.1003	0.5605	0.1777	0.1370	0.5118
		NMF	0.1606	0.0293	0.5368	0.2158	0.2321	0.2959
		KNN	0.2414	0.3251	0.6154	0.2324	0.2891	0.7730
		NB	0.2600	0.1618	0.7054	0.2389	0.2036	0.3058
Imdb	Usr, Dir, Cast	SIMSBM(1,1,1)	**0.3896(1)**	**0.3437(2)**	**0.7593(1)**	**0.3293(2)**	**0.5705(1)**	**0.1654(1)**
		TF	0.2547	0.2238	0.5039	0.1513	0.4549	0.2636
		NMF	0.1127	0.0483	0.5005	0.1529	0.1406	0.8319
		KNN	0.2596	0.1890	0.5501	0.1681	0.3268	0.5248
		NB	0.2558	0.2373	0.5362	0.1617	0.4632	0.2571

Conclusion

- In summary:
  - Numerous real-world applications
    - Recommender systems (Spotify, Imdb, ...)
    - Social sciences (archaeology, behaviour analysis, ...)
    - Miscellaneous (medical diagnosis, retweet prediction, ...)
  - Easy model selection
    - Three original models summarized in one framework
    - Infinitely many ready-to-use models
  - Linear complexity



Thanks for your attention!

Webpage: `https://gaelpouxmedard.github.io/`

Code and data: `https://github.com/GaelPouxMedard/SIMSBM/`