# Latent Behavioral Structure in Online Communities

## Case Study in `/r/science`

Steven Morse
William & Mary, Data Science
DATA 691, Graph Learning, Final Project

April 29, 2025

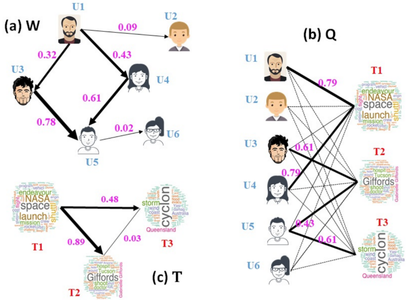
# Overview

## Problem statement

In social forums, observable behavior is often limited to who you talk to (interaction) and what you talk about (topics). Can we learn the underlying relational structure of the community given these observables? Is the structure predictive of behavior over time?

# Related work

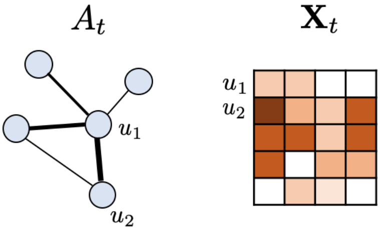
- Latent variable approaches based on features [Hoff et al., 2002, Airoldi et al., 2008]
- ... based on timing [Linderman and Adams, 2014]
- Topic modeling [Chang and Blei, 2009]
- Dynamic community-topic modeling [Zhang et al., 2019, Li et al., 2012]
- Dynamic networks [Pareja et al., 2020, Sankar et al., 2020]
- Graph autoencoders [Kipf and Welling, 2016, Simonovsky and Komodakis, 2018]
- Joint graph + feature learning [Lerique et al., 2020]




[Choudhari et al., 2021]

# Data overview and preparation

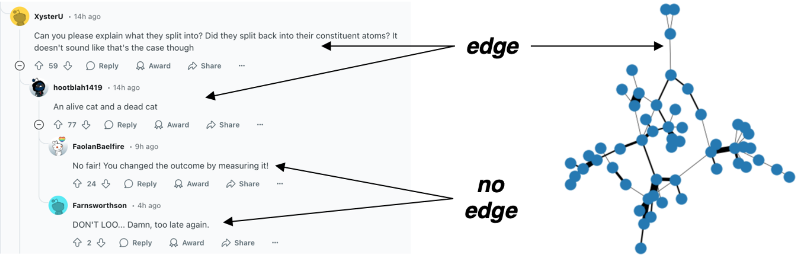
- Reddit: $>$ 12.7 billion comments over 17 years (2007-2022) [Baumgartner et al., 2020]
- We focus on 2007-2011 in the `/r/science` subreddit to enforce the idea of community.
- We engineer two inputs to our model:
  - Co-reply interaction graph
  - User topic participation signature



# User-to-user co-reply interaction

## Interaction graph

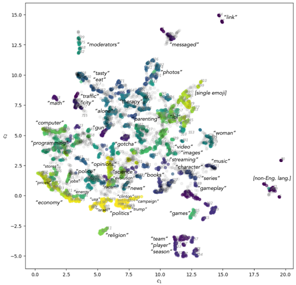
Define $G^{(\tau)} = (V, E)^{(\tau)}$ with $e_{uv} = m$ such that $m$ is a count of occurrences where $u$ replies to $v$ (or $v$ to $u$) within two comments and within two months, during $\tau$.



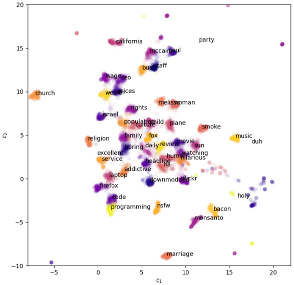
## Contextualized semantic clustering

Given an embedder $\mathcal{E} : \mathcal{S}_\ell \to \mathbb{R}^d$ and clustering $\mathcal{C} : \mathbb{R}^d \to L$,

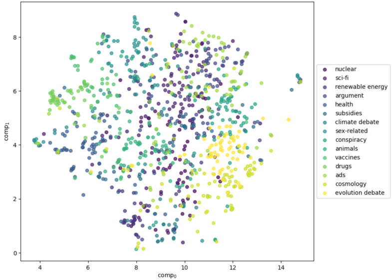
$$\mathcal{D} = \{(s_i, u_i, t_i)\}_i \quad \longrightarrow \quad \mathcal{D}_L = \{(\mathbf{e}_i, u_i, t_i, l_i)\}_i$$




2007-2022




July 2008 (All)




July 2008 (`/r/science`)

# User participation over topics

## User topic signature

Define a user $u$'s participation in $k$ topics at time $\tau$ as the signature:

$$\mathbf{x}_u^{(\tau)} = (x_k^{(\tau)}) \quad \text{with} \quad \hat{x}_k^{(\tau)} = |\mathcal{D}_L : u_i = u, t_i \in \tau, l_i = k|, \quad x_k^{(\tau)} = \frac{\hat{x}_k^{(\tau)}}{||\hat{\mathbf{x}}_u^{(\tau)}||}$$

...

***March 2007:***
13 (cosmology): Human understanding of time cannot be complete as long as …
7 (sex-related): The sudden reduction of pain is a powerful simulacrum of pleasure.
1 (sci-fi): Time Cub(beaten by an angry horde)
9 (animals): The title make a good quote, but the implications of a city as an organism ….
0 (nuclear): It's like there's an anti-progressive monster inside California struggling to get out. :-(
11 (drugs): Uhhh....for me yes. I want the drugs *now* :-(
3 (argument): There's an important point about politics that needs to be made. Americans are …
0 (nuclear): That wasn't my point at all - in fact most Western countries *including* France …
5 (subsidies): I'm not an anti-oil nut, but drilling in our own territory is indeed more .... palatable.

...

$\hat{\mathbf{x}}_u^{(\tau)}$

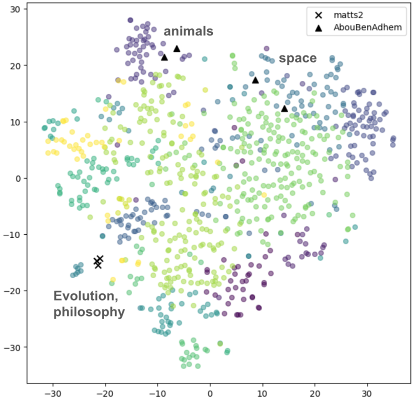
| | author | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | total_posts |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 485 | Prysorra | 4 | 1 | 2 | 1 | 4 | 1 | 3 | 1 | 2 | 11 | 5 | 9 | 4 | 0 | 3 | 51 |

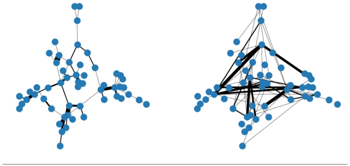
$\mathbf{x}_u^{(\tau)}$

[0.23 0.06 0.11 0.06 0.23 0.06 0.17 0.06 0.11 0.63 0.29 0.52 0.23 0.00 0.17]

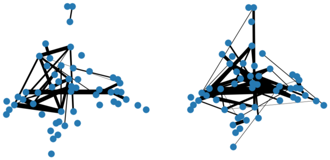
# Observations

- Users' signatures and their co-reply behaviors evolve over time.
- Topic signatures and co-reply interaction are only weakly correlated.



Interaction graphs $(t_0, t_1)$



Feature similarities $(t_0, t_1)$

# (Temporal) Dual-decoder Graph Autoencoder (TDGAE)

- Learn a latent graph $Z$ of behavioral homophily
- Encoder:

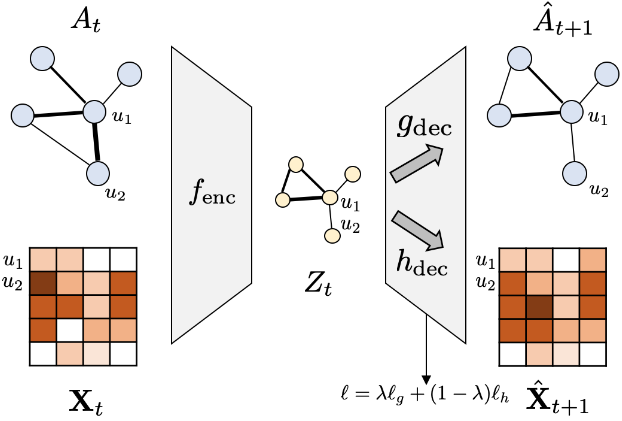
$$\mathbf{h}^{(1)} = \mathrm{GCN}(\mathbf{x}, E)$$
$$\mathbf{h}^{(2)} = \mathrm{GCN}(\mathbf{h}^{(1)}, E)$$
$$\mathbf{z}_{uv} = \mathrm{MLP}([\mathbf{h}_u^{(2)}, \mathbf{h}_v^{(2)}])$$

- (Graph) Decoder:

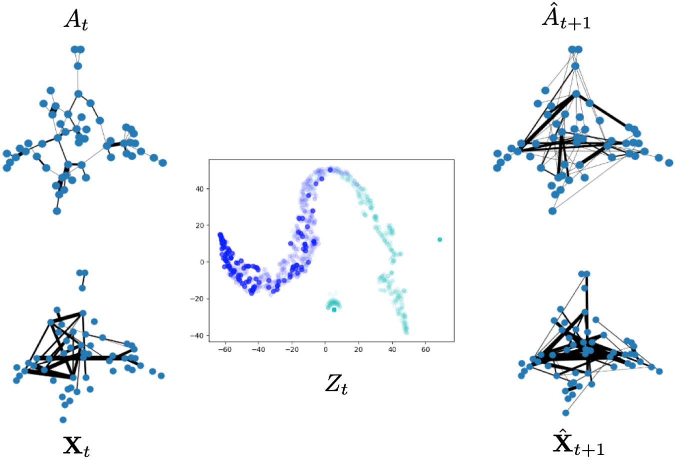
$$\mathbf{v} = \mathrm{MLP}(\mathbf{z})$$
$$p(e_{uv}) = \sigma(\mathbf{v})$$

- (Feature) Decoder:

$$\mathbf{a} = \mathrm{Agg}(\mathbf{z})$$
$$\mathbf{x} = \mathrm{MLP}(\mathbf{a})$$

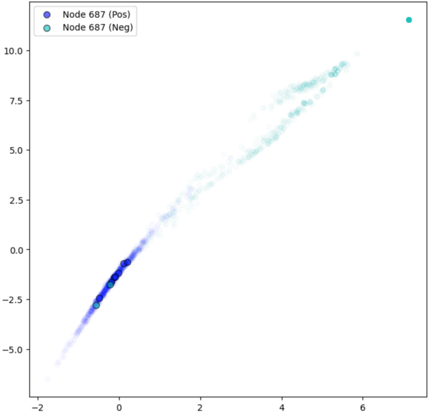


# Example




2007: Jan-Jun ($t$), Jul-Dec ($t + 1$)

# Example




With latent dimension $d = 2$, single node highlighted

# Performance

- "Naive" model:
  - Edges: $e_{uv}^{(\tau)} = e_{uv}^{(\tau+1)}$
  - Features: $\mathbf{x}_u^{(\tau)} = \mathbf{x}_u^{(\tau+1)}$
- LR (features-only) is linear regression $f : \mathbb{R}^d \to \mathbb{R}^d$
- GCN: single-layer $\mathrm{GCN}(A_\tau, \mathbf{X}_\tau)$

| | Edges (AUC) | Features (MSE) |
|---|---|---|
| Naive | 0.508 | 0.0685 |
| LR | - | **0.0410** |
| GCN | 0.894 | 0.0433 |
| DGAE | **0.903** | 0.0445 |

# Next steps

- https://github.com/stmorse/sgg
- More data; non-social network data
- Deeper networks, harder baselines
- Re-examine model architecture (e.g. $h(Z_t)$ or $h(g(Z_t))$ or something else)
- Latent space dimension tuning

# References I


Airoldi, E. M., Blei, D., Fienberg, S., and Xing, E. (2008).
Mixed membership stochastic blockmodels.
*Advances in neural information processing systems*, 21.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020).
The pushshift reddit dataset.
In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Chang, J. and Blei, D. (2009).
Relational topic models for document networks.
In *Artificial intelligence and statistics*, pages 81–88. PMLR.

Choudhari, J., Bedathur, S., Bhattacharya, I., and Dasgupta, A. (2021).
Analyzing topic transitions in text-based social cascades using dual-network hawkes process.
In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 305–319. Springer.

# References II

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002).
Latent space approaches to social network analysis.
*Journal of the american Statistical association*, 97(460):1090–1098.

Kipf, T. N. and Welling, M. (2016).
Variational graph auto-encoders.
*arXiv preprint arXiv:1611.07308*.

Lerique, S., Abitbol, J. L., and Karsai, M. (2020).
Joint embedding of structure and features via graph convolutional networks.
*Applied Network Science*, 5:1–24.

Li, D., Ding, Y., Shuai, X., Bollen, J., Tang, J., Chen, S., Zhu, J., and Rocha, G. (2012).
Adding community and dynamic to topic models.
*Journal of Informetrics*, 6(2):237–253.

# References III


Linderman, S. and Adams, R. (2014).
Discovering latent network structure in point process data.
In *International conference on machine learning*, pages 1413–1421. PMLR.

Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., Schardl, T., and Leiserson, C. (2020).
Evolvegcn: Evolving graph convolutional networks for dynamic graphs.
In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5363–5370.

Sankar, A., Wu, Y., Gou, L., Zhang, W., and Yang, H. (2020).
Dysat: Deep neural representation learning on dynamic graphs via self-attention networks.
In *Proceedings of the 13th international conference on web search and data mining*, pages 519–527.

Simonovsky, M. and Komodakis, N. (2018).
Graphvae: Towards generation of small graphs using variational autoencoders.
In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, pages 412–422. Springer.

# References IV


Zhang, Y., Wu, B., Ning, N., Song, C., and Lv, J. (2019).
Dynamic topical community detection in social network: A generative model approach.
*IEEE Access*, 7:74528–74541.

# Questions?