Temporal models for polyadic events

Based on the work by

Zhe, Shandian, and Yishuai Dum,

"Stochastic Nonparametric Event-Tensor Decomposition." *Advances in Neural Information Processing Systems*, 2018.

Example of temporal data

On the web:

user	item	action	time
…	…	…	…
0	575	view	12/2/2017 9:50
0	1881	view	12/9/2017 18:52
0	846	basket	12/13/2017 12:28
1	1878	purchase	12/13/2017 21:29
1	576	view	12/15/2017 4:49
…	…	…	…

or

$$\textbf{Sender} \times \textbf{Reciever} \times \textbf{Event} \times \textbf{Time}$$

Task: build a model to answer questions like:

"do *user 0* and *user 1* have similar tastes"

"will *user 1* buy *product 576*?"

"*which other products user 0 might like*"

Popular approaches

Counts

- entire temporal data is disregarded
- events are considered independent
- use a (multi/non)linear function of parameters

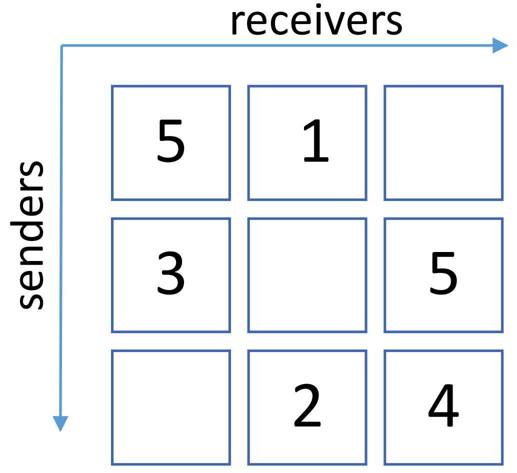
Binning

- temporal data is discretized
- events are grouped by time
- can use smoothing for time factors

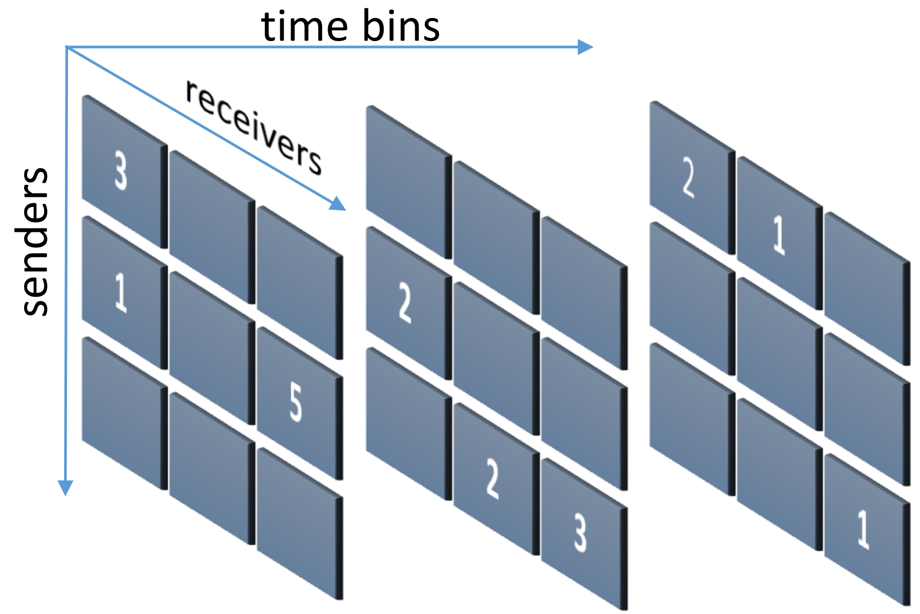
can use various distribution assumptions on observations, e.g. Gaussian, Poisson, etc.

sender	receiver	time
0	575	12/2/2017 9:50
0	1881	12/9/2017 18:52
0	846	12/13/2017 12:28
1	1878	12/13/2017 21:29
1	576	12/15/2017 4:49

simple count of events



count of events within time range



Preliminaries on probabilistic approach

link function

likelihood of a random variable: $x_i \sim p(x_i \,|\, \theta_i) \quad \text{where} \quad \ell(\theta_i) = m_i$

distribution parameters

parameters of our model

maximum likelihood estimation:

$$\max_{\mathcal{M}} \; L(\mathcal{M}; \mathcal{X}) \equiv \prod_{i \in \Omega} p(x_i \,|\, \theta_i)$$

set of all observations

minimization problem:

$$\min F(\mathcal{M}; \mathcal{X}) \equiv \sum_{i \in \Omega} f(x_i, m_i)$$

$$f(x, m) \equiv -\log p(x \,|\, \ell^{-1}(m))$$

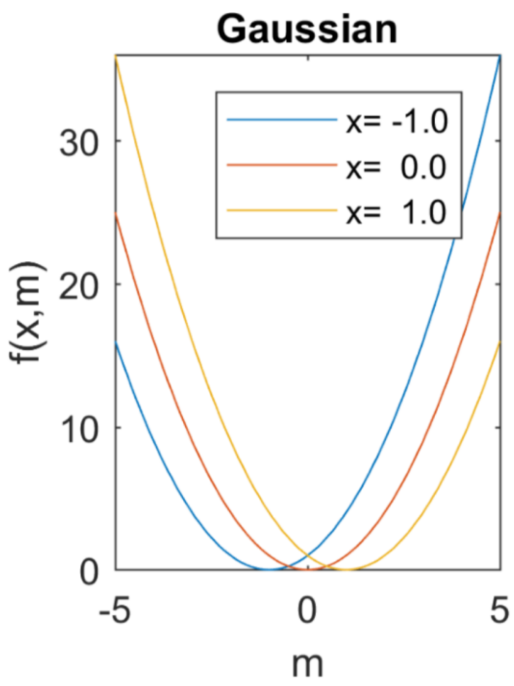
Gaussian distribution case

$$x_i = m_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim \mathcal{N}(0, \sigma)$$

$$x_i \sim \mathcal{N}(\mu_i, \sigma) \qquad \ell(\mu) = \mu \qquad \mu_i = m_i$$

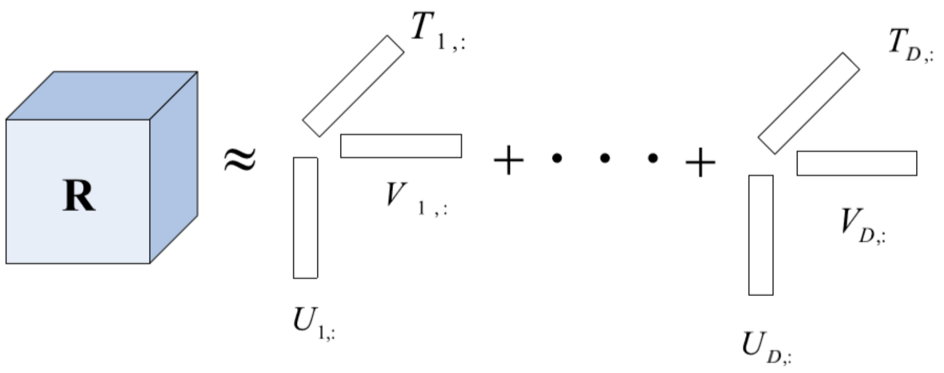
$$p(x \,|\, \mu, \sigma) = e^{-(x-\mu)^2 \,/\, 2\sigma^2} \,/\, \sqrt{2\pi\sigma^2}$$

$$\boxed{f(x, m) = (x - m)^2 \,/\, \cancel{(2\sigma^2)} + \tfrac{1}{2}\cancel{\log(2\pi\sigma^2)}}$$



Hong, David, Tamara G. Kolda, and Jed A. Duersch. "Generalized Canonical Polyadic Tensor Decomposition." *arXiv preprint arXiv:1808.07452* (2018).

Time binning



- Hours / Days / Weeks, etc.

The model: User × Item × Time span → Relevance

- Using CP approximation
- All factors follow normal distribution with zero mean

Smoothing assumption*:

$$p\big(\mathbf{T}(k,:)|\mathbf{T}(k-1,:)\big) = \mathcal{N}\big(\mathbf{T}(k,:)|\mathbf{T}(k-1,:),\sigma^2\mathbf{I}\big)$$

Results in additional regularization terms: $||\mathbf{T}(k,:) - \mathbf{T}(k-1,:)||^2$

*Xiong, Liang, et al. "Temporal collaborative filtering with bayesian probabilistic tensor factorization.“ *Proceedings of the 2010 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2010.

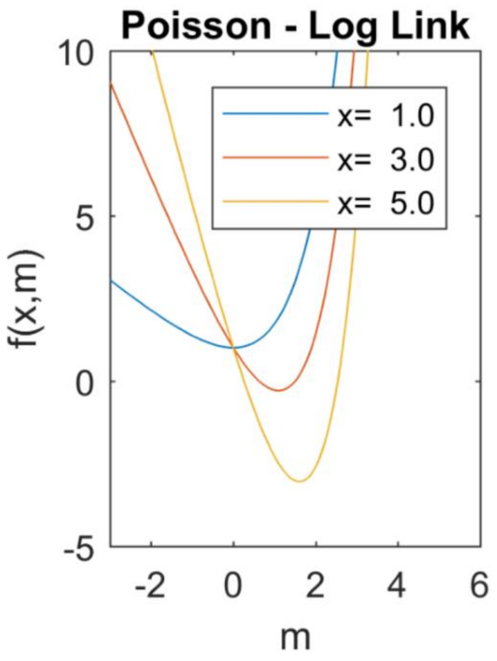
Poisson distribution case

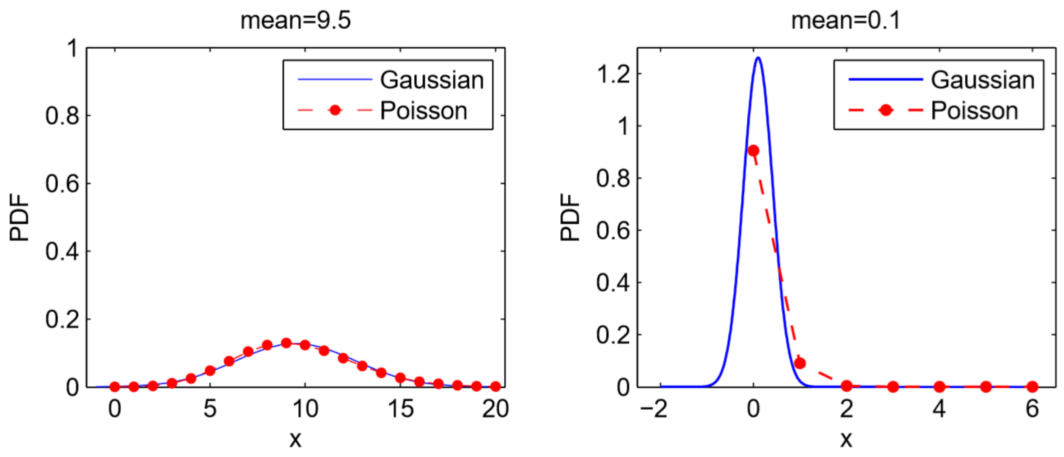
$$p(x \mid \lambda) = e^{-\lambda}\lambda^{x} \,/\, x! \quad \text{for} \quad x \in \mathbb{N}$$

$$\ell(\lambda) = \log \lambda$$

mean and variance

$$f(x, m) = e^{m} - xm \quad \text{for} \quad x \in \mathbb{N}, m \in \mathbb{R}$$





Figures credit: Hong, David, Tamara G. Kolda, and Jed A. Duersch. "Generalized Canonical Polyadic Tensor Decomposition." *arXiv preprint arXiv:1808.07452* (2018).

Many scalable techniques:

- APR-CP (Kolda)
- KL-based optimization

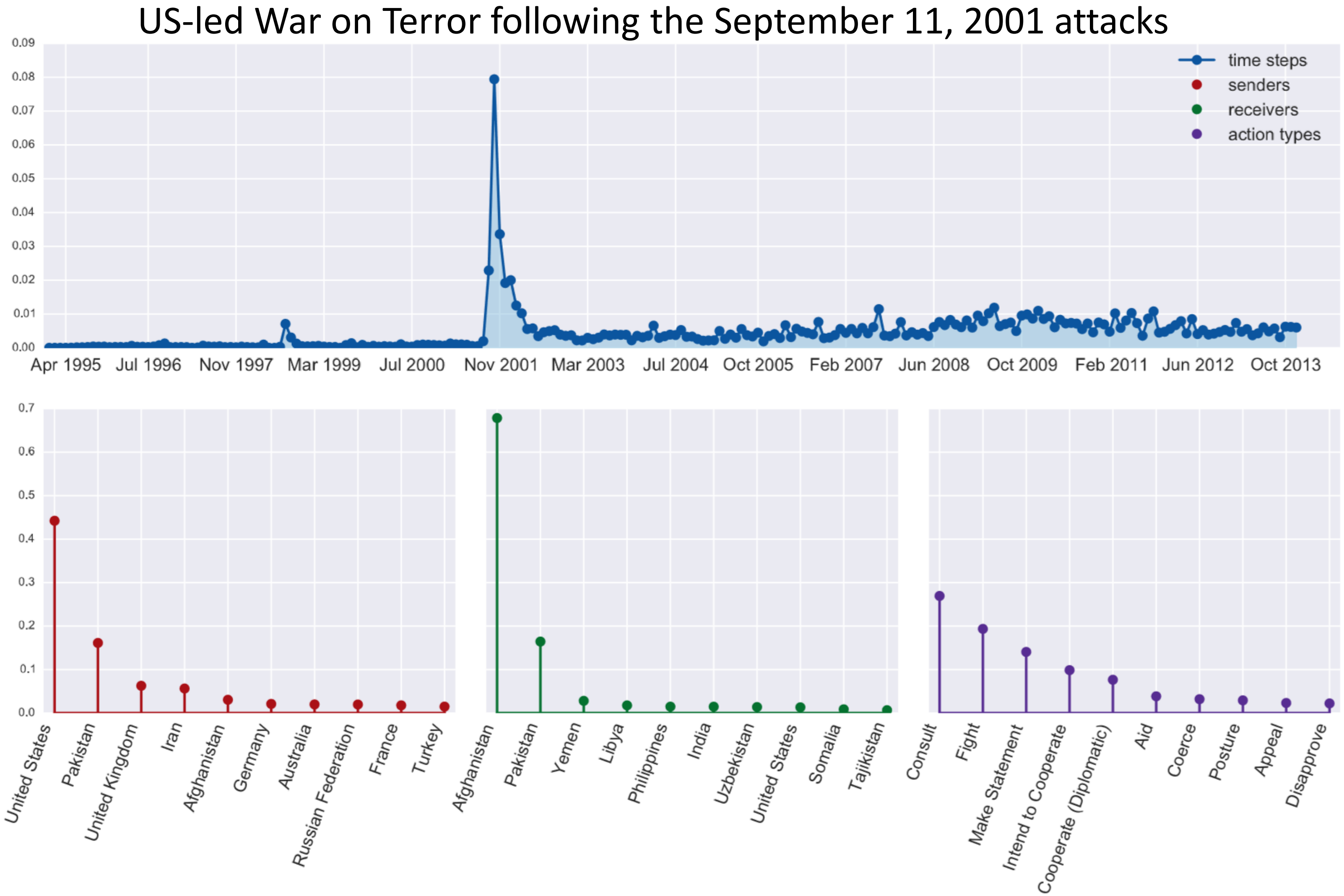
Other distributions

Table 1: Statistically-motivated loss functions. Parameters in blue are assumed to be constant. Numerical adjustments are indicated in red.

Distribution	Link function	Loss function	Constraints
$\mathcal{N}(\mu, \sigma)$	$m = \mu$	$(x-m)^2$	$\boxed{x, m \in \mathbb{R}}$
$\text{Gamma}(k, \sigma)$	$m = k\sigma$	$x/(m+\epsilon) + \log(m+\epsilon)$	$x > 0, m \geq 0$
$\text{Rayleigh}(\theta)$	$m = \sqrt{\pi/2}\,\theta$	$2\log(m+\epsilon) + (\pi/4)(x/(m+\epsilon))^2$	$x > 0, m \geq 0$
$\text{Poisson}(\lambda)$	$m = \lambda$	$m - x\log(m+\epsilon)$	$x \in \mathbb{N}, m \geq 0$
	$m = \log\lambda$	$e^m - xm$	$\boxed{x \in \mathbb{N},\ m \in \mathbb{R}}$
$\text{Bernoulli}(\rho)$	$m = \rho\,/\,(1-\rho)$	$\log(m+1) - x\log(m+\epsilon)$	$x \in \{0, 1\}, m \geq 0$
	$m = \log(\rho\,/\,(1-\rho))$	$\log(1+e^m) - xm$	$x \in \{0, 1\},\ m \in \mathbb{R}$
$\text{NegBinom}(r, \rho)$	$m = \rho\,/\,(1-\rho)$	$(r+x)\log(1+m) - x\log(m+\epsilon)$	$x \in \mathbb{N},\ m \geq 0$

Source: Hong, David, Tamara G. Kolda, and Jed A. Duersch. "Generalized Canonical Polyadic Tensor Decomposition." *arXiv preprint arXiv:1808.07452* (2018).

International relations analysis



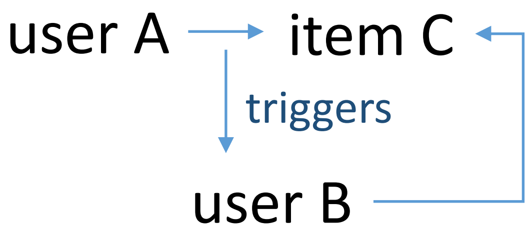
4D tensor data:
County × County × Action × Time

entities with the highest values of latent factors are displayed

Schein, Aaron, et al. "Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts."
Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

Problems

- Loosing information (due to aggregation and binning)
- Not able to catch local causal/triggering effects



need a state space transition mechanism

event-tensor abstraction

Hawkes process

base rate

$$\lambda(t) = \lambda_0 + \sum_{t_i<t} h(t - t_i)$$

triggering/excitation function

Joint probability of a sequence $\{t_1, \dots t_n\}$

$$\boxed{p(\{t_1, \dots t_n\}) = e^{-\int_0^T \lambda(t)} \prod_{j=1}^{n} \lambda(t_j)}$$

The model

Work by Zhe, Shandian, and Yishuai Dum,
"Stochastic Nonparametric Event-Tensor Decomposition." *NeurIPS*, 2018.

index of event

sequence of all observations: $$S = [(s_1, \mathbf{i}_1), \ldots, (s_N, \mathbf{i}_N)]$$

event timestamp

some nonlinear function

$$\boxed{\lambda_{\mathbf{i}}(t) = \lambda_{\mathbf{i}}^0 + \sum_{s_n < t} h_{\mathbf{i}_n \to \mathbf{i}}(t - s_n)}$$

$$\lambda_{\mathbf{i}}^0 = e^{f(\mathbf{x}_{\mathbf{i}})}$$

$$\mathbf{x}_{\mathbf{i}} = [\underbrace{\mathbf{U}^{(1)}(i_1, :), \ldots, \mathbf{U}^{(K)}(i_K, :)}]$$

CP factors $\mathcal{U} = \{\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}\}$

$$\boxed{h_{\mathbf{i}_n \to \mathbf{i}}(t - s_n) = k(\mathbf{x}_{\mathbf{i}_n}, \mathbf{x}_{\mathbf{i}}) h_0(t - s_n)}$$

kernel function

$$h_0(t - s_n) = \mathbb{1}(s_n \in A_t) \beta e^{-\frac{1}{\tau}(t - s_n)}$$

collection of preceding events

$$A_t = \{s_j | s_j \in P_t(C_{\max}), t - \Delta_{\max} \leq s_j \leq t\}$$

Likelihood estimate

events for entry i

$$p(\{m_{\mathbf{i}}, f_{\mathbf{i}}\}|\mathcal{U}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, c(\mathbf{X}, \mathbf{X})) \prod_{\mathbf{i}} e^{-\int_0^T \lambda_{\mathbf{i}}(t)\mathrm{d}t} \prod_{j=1}^{n_{\mathbf{i}}} \lambda_{\mathbf{i}}(s_{\mathbf{i}}^n)$$

$p(\mathbf{f}|\mathcal{U})$ follows a Gaussian process (for estimating $\mathbf{f}$)

- variables are highly entangled within nonlinear terms
- leads to intractable objective

Tricks to compute:

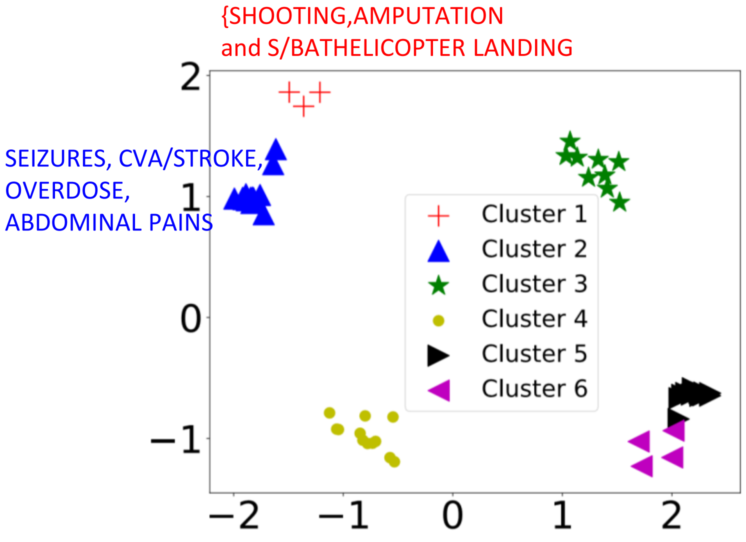
- Hawkes process as the union of Poisson processes (Poisson super-position theorem)
- Add low-parametric latent cause variable for each event with variational posterior $q(\mathbf{z})$
- Use SOTA sparse variational GP framework
- Randomly partition both the events and the tensor entries into mini-batches $\{N_k\}$ and $\{M_l\}$
- Solved by modified EM algorithm

Final objective:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{g})}\big(\log \frac{p(\mathbf{g})}{q(\mathbf{g})}\big) + \sum_k \frac{|N_k|}{N} \sum_{j \in N_k} \phi_{s_j, \bar{A}_{s_j}} \frac{N}{|N_k|} + \sum_k \sum_l \frac{|N_k|}{N} \frac{|M_l|}{M} \sum_{j \in N_k} \sum_{\mathbf{i} \in M_l} \psi_{s_j, \mathbf{i}, \mathbf{i}_j} \frac{N}{|N_k|} \frac{M}{|M_l|}$$

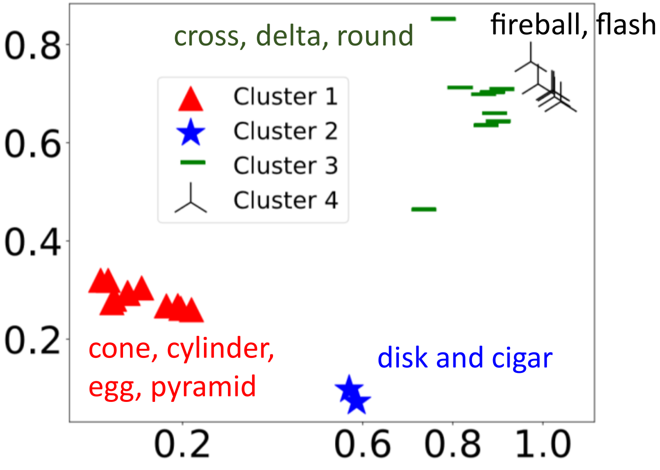
Results

k-means + BIC clustering over the latent space



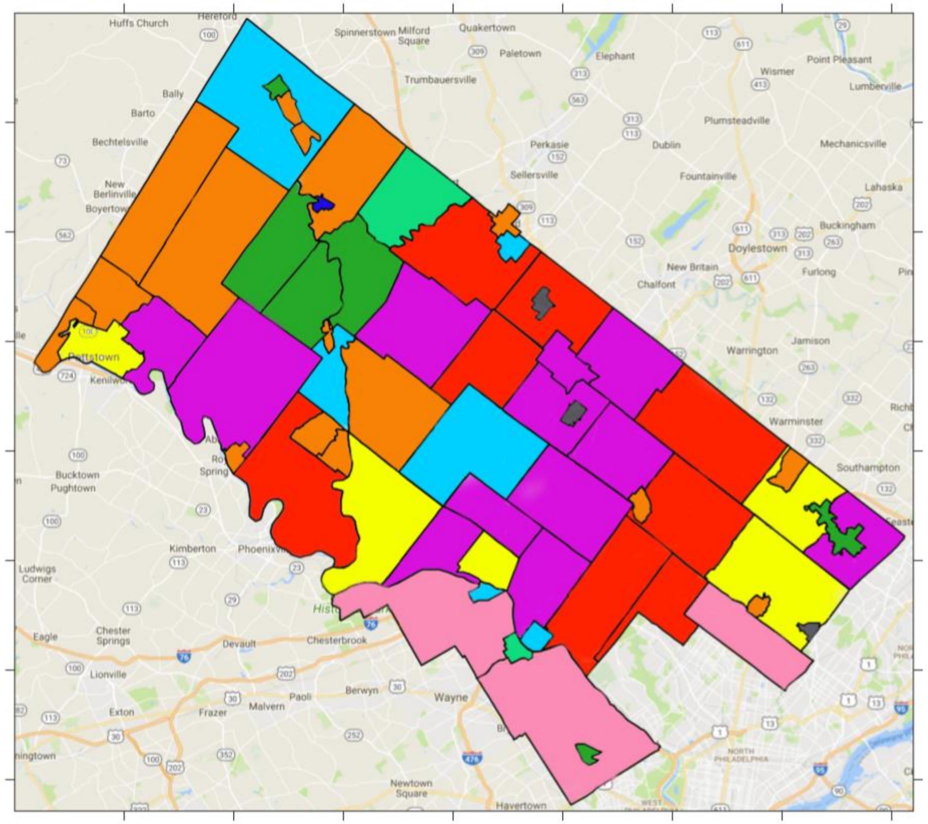
(a) EMS titles

(EMS title, township) data



(b) UFO shapes

(UFO shape, city) data



(c) Townships

Figure 3: Structures reflected from the latent factors learned by our model on *911* on *UFO*. In (c), the clusters of townships are shown in the actual map.

Questions?