Decoupling Homophily and Reciprocity with Latent Space Network Models

Jiasen Yang[*] Vinayak Rao[*] Jennifer Neville[*†]

Departments of Statistics[*] and Computer Science[†]
Purdue University

August 13, 2017

PURDUE
UNIVERSITY

Dynamic Network Data

From	To	Date	Time
26e0c6c4d3b36b2594739fa30eb564b4	881938e1e49fe9d7ee0da580ae4e5946	7/7/11	12:00:02 AM
6d3d380010755cfbeded2e4ac008e8a5	6d3d380010755cfbeded2e4ac008e8a5	7/7/11	12:00:03 AM
eabf2d3d8b046cca7360e2caf1c03362	6cdca3d8290d6c9478a8c9cf6f702a0a	7/7/11	12:00:04 AM
bfd64e0e849d4ca70b719b25a34fa89a	bfd64e0e849d4ca70b719b25a34fa89a	7/7/11	12:00:09 AM
c0784dd06a9a556870d6cdf320b2042f	c0784dd06a9a556870d6cdf320b2042f	7/7/11	12:00:15 AM
589cbf6980757256743976503236bc0	43d451c9016fe24183ea83177e051cb5	7/7/11	12:00:17 AM
b311ecf15238219200d45940208b27c9	7362199de1973c3ce6fd0a8182879642	7/7/11	12:00:17 AM
8bff3584b3038174ed5064465d97578b	79de1530de5760996642b502dc41ab10	7/7/11	12:00:24 AM
6d3d380010755cfbeded2e4ac008e8a5	6d3d380010755cfbeded2e4ac008e8a5	7/7/11	12:00:46 AM
a49df0f8fd9f04ddffc35ddd75910b57	c7e640211096aa609b5c5568ded4a8c4	7/7/11	12:00:46 AM
697ac859008ce24a12343422b4188bea	697ac859008ce24a12343422b4188bea	7/7/11	12:00:49 AM
be986fcb0a18a639f2a730a2485f580d	be986fcb0a18a639f2a730a2485f580d	7/7/11	12:00:53 AM
10c1e0c9ed1a17411e784ad554db3b55	994cba4c9a4a9fbb2aedd1ec03270cb1	7/7/11	12:01:09 AM
547d16750b6ace87de48a5af95ce9d11	69c88cdaaadba8fd783e4612cc29f824	7/7/11	12:01:22 AM
6d3d380010755cfbeded2e4ac008e8a5	6d3d380010755cfbeded2e4ac008e8a5	7/7/11	12:01:27 AM
b13ec563a5e89c7a97b04662a0c43d31	b13ec563a5e89c7a97b04662a0c43d31	7/7/11	12:01:28 AM
19e74a1ab165c956d27a863854085485	2afda4ee8fc22fa741037304f43119af	7/7/11	12:02:25 AM
2acc1711a000ef1c512e8c0a00b7c122	924f4817dec2429f36b555bb0a27f059	7/7/11	12:02:26 AM
02e04607877ab6c02a0c7e93402952ca	70bbb51866b05f9b6d5a5f8ba042e2d3	7/7/11	12:02:34 AM
6808e627469fa5bd91de4cacf09d5619	6808e627469fa5bd91de4cacf09d5619	7/7/11	12:03:00 AM
ae5224db998ef857d02ccdea732f3191	66ee18050233a847ca6c81417bf0e642	7/7/11	12:03:05 AM
19e74a1ab165c956d27a863854085485	5a479e27f86d379aa5e5cae81ccaec4c	7/7/11	12:03:07 AM
e9cc27bb3cdd97b7e6c269b118169e83	e9cc27bb3cdd97b7e6c269b118169e83	7/7/11	12:03:10 AM
6a1827a939cf86b0271db68aa4f1cb43	6a1827a939cf86b0271db68aa4f1cb43	7/7/11	12:03:25 AM
0411c4c432071a78e5e081c859763408	082e77e6ed67753109ded73e2651e635	7/7/11	12:03:36 AM
feb6e82c08cd7fec916bab95e2fc4111	d1fc082e471464120e1e85ddb1da8cd1	7/7/11	12:03:46 AM
6cdca3d8290d6c9478a8c9cf6f702a0a	6cdca3d8290d6c9478a8c9cf6f702a0a	7/7/11	12:03:48 AM
332b355737cb2b7790400f2dcd4e8496	1825199e4a20c757b37df5ee0c4877e0	7/7/11	12:04:18 AM
16d35ae8ba06c11ee315a301599dcc10	96601467b1ce7c1498a36a28353ef056	7/7/11	12:04:55 AM

- Point process models
- Statistical network models

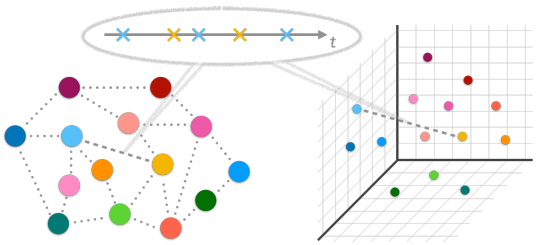
Observed data: $\{(u, v, \mathcal{H}_{uv})\}_{u,v \in V}$

- $V = \{1, \ldots, n\}$: a fixed set of vertices throughout time period $[0, T)$.
- $\mathcal{H}_{uv} = \{t_i^{uv}\}_{i=1}^{n_{uv}}$: set of all time-points at which u sent v a message.
- $n_{uv} \geq 0$: total number of messages from u to v.

Observed data: $\{(u, v, \mathcal{H}_{uv})\}_{u,v \in V}$

- $V = \{1, \ldots, n\}$: a fixed set of vertices throughout time period $[0, T)$.
- $\mathcal{H}_{uv} = \{t_i^{uv}\}_{i=1}^{n_{uv}}$: set of all time-points at which u sent v a message.
- $n_{uv} \geq 0$: total number of messages from u to v.

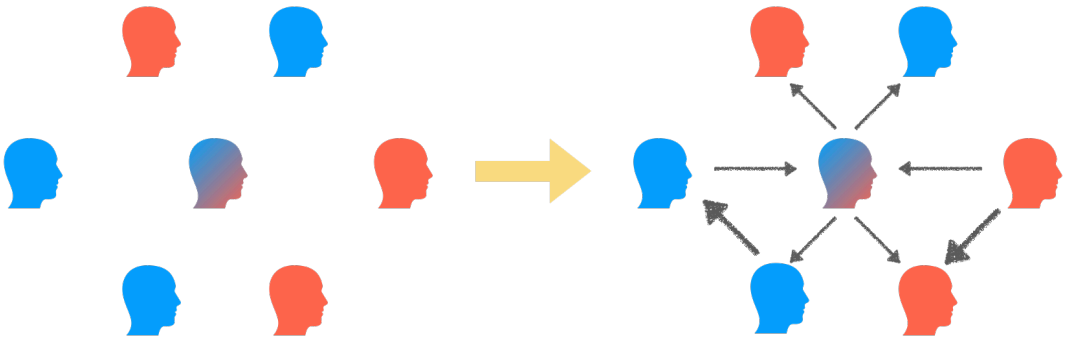
This work: Latent space point process models of dynamic networks:



- Model $\{t_i^{uv}\}_{i=1}^{n_{uv}}$ as realizations of a point process $N_{uv}(t)$, $t \in [0, T)$.
- Entire network $\Rightarrow n^2 - n$ processes (no self-loops). *Not independent!*
- Dependencies via latent space model.
- Network embedding & link prediction.

Homophily

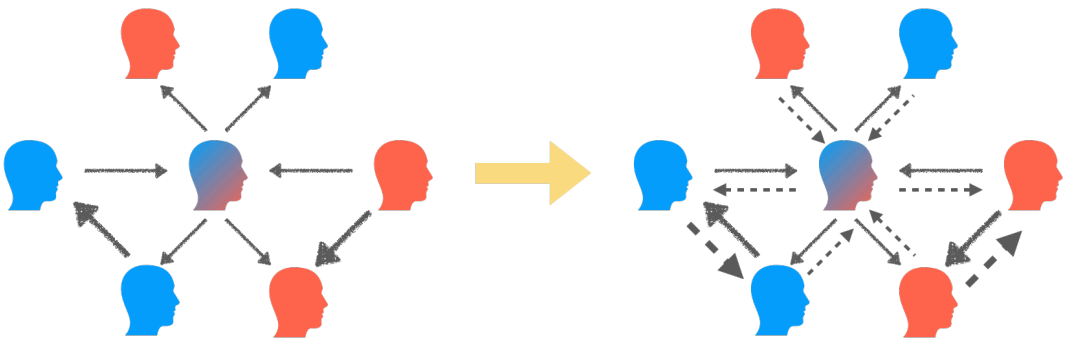
Communication occurs at a higher rate between individuals with similar features.



Latent space models of (static) networks: $p_{uv} \propto e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2}$, $\mathbf{z}_u, \mathbf{z}_v \in \mathbb{R}^d$. (Hoff *et al.*, 2002)

Reciprocity

Individuals tend to reciprocate communications from other individuals.

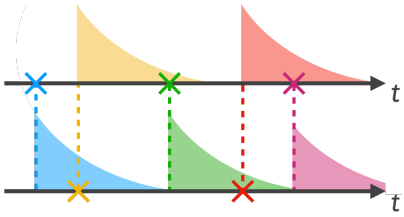


Modeling reciprocating relationships with Hawkes processes. (Blundell *et al.*, 2012)

Model the pair of processes $N_{uv}(t)$ and $N_{vu}(t)$ as a bivariate Hawkes process:

$$\lambda_{uv}(t|\mathcal{H}_{uv}, \mathcal{H}_{vu}) = \gamma_{uv} + \sum_{k:\, t_k^{vu} < t} \phi_{uv}(t - t_k^{vu})$$

Event history: $\mathcal{H}_{uv}(t) = \{t_k^{uv} : t_k^{uv} < t\}$.



(Hawkes, 1971)

Weighted combination of basis kernels:

$$\underbrace{\phi_{uv}(\cdot)}_{\text{Triggering kernel}} = \sum_{b=1}^{B} \xi_b^{uv} \phi_b(\cdot)$$

e.g., $\phi_b(t) = e^{-t/\tau}$, $e^{-t/\tau} \sin^2\left(\frac{\pi t}{\tau}\right)$.

Hawkes Process (HP) Model

$$\lambda_{uv}(t) = \gamma + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, \phi_b(t - t_k^{vu})$$

$$N_{uv}(\cdot) \sim \text{HawkesProcess}(\lambda_{uv}(\cdot))$$

(*cf.* Blundell *et al.*, 2012)

Latent Space Point Process Models

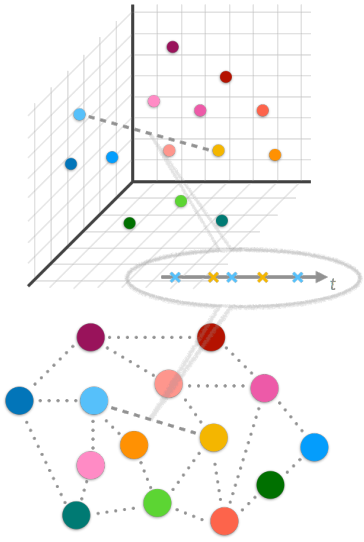
- Poisson-rate Latent Space (PLS) Model
- Hawkes Dual Latent Space (DLS) Model
- Hawkes Base-rate Latent Space (BLS) Model
- Hawkes Reciprocal Latent Space (RLS) Model

Poisson-rate Latent Space (PLS) Model

$$\mathbf{z}_v \sim \mathcal{N}(0, \sigma^2 \, \mathrm{I}_{d\times d}) \qquad \forall v \in V$$

$$\lambda_{uv}(t) = \gamma \, e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2} \qquad \forall u \neq v$$

$$N_{uv}(\cdot) \sim \text{PoissonProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$



Hawkes Dual Latent Space (DLS) Model

$$\mathbf{z}_v \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d\times d}) \qquad \forall v \in V$$

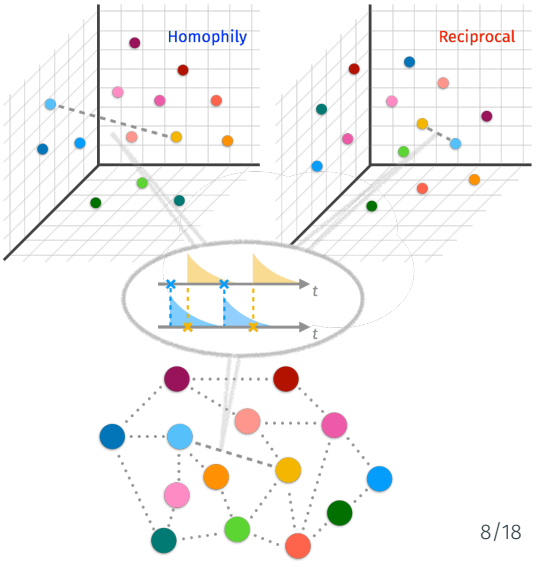
$$\boldsymbol{\mu}_v \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_{d\times d}) \qquad \forall v \in V$$

$$\boldsymbol{\varepsilon}_v^{(b)} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{d\times d}) \qquad \forall v \in V,\ b = 1, \ldots, B$$

$$\mathbf{x}_v^{(b)} \sim \boldsymbol{\mu}_v + \boldsymbol{\varepsilon}_v^{(b)} \qquad \forall v \in V,\ b = 1, \ldots, B$$

$$\lambda_{uv}(t) = \underbrace{\gamma\, e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2}}_{\text{Homophily base-rate}} + \underbrace{\sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \beta\, e^{-\|\mathbf{x}_u^{(b)} - \mathbf{x}_v^{(b)}\|_2^2}\, \phi_b(t - t_k^{vu})}_{\text{Reciprocal component}}$$

$$N_{uv}(\cdot) \sim \text{HawkesProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$



Base-rate Latent Space (BLS) Model

$$\mathbf{z}_v \sim \mathcal{N}(0, \sigma^2 \, \mathrm{I}_{d \times d})$$

$$\lambda_{uv}(t) = \underbrace{\gamma \, e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2}}_{\text{Homophily base-rate}} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, \phi_b(t - t_k^{vu})$$

$$N_{uv}(\cdot) \sim \text{HawkesProcess}(\lambda_{uv}(\cdot))$$

Reciprocal Latent Space (RLS) Model

$$\mathbf{x}_v \sim \mathcal{N}(0, \sigma^2 \, \mathrm{I}_{d \times d})$$

$$\lambda_{uv}(t) = \gamma + \underbrace{\sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, e^{-\|\mathbf{x}_u - \mathbf{x}_v\|_2^2} \, \phi_b(t - t_k^{vu})}_{\text{Reciprocal component}}$$

$$N_{uv}(\cdot) \sim \text{HawkesProcess}(\lambda_{uv}(\cdot))$$

(*cf.* Linderman and Adams, 2014)

Model Summary

Model	$\lambda_{uv}(t)$	#parameters
HP	$\gamma + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, \phi_b(t - t_k^{vu})$	$\mathcal{O}(B)$
PLS	$\gamma \, e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2}$	$\mathcal{O}(n \cdot d)$
BLS	$\gamma \, e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, \phi_b(t - t_k^{vu})$	$\mathcal{O}(n \cdot d)$
RLS	$\gamma + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \xi_b \, e^{-\|\mathbf{x}_u - \mathbf{x}_v\|_2^2} \, \phi_b(t - t_k^{vu})$	$\mathcal{O}(n \cdot d)$
DLS	$\gamma \, e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \beta \, e^{-\|\mathbf{x}_u^{(b)} - \mathbf{x}_v^{(b)}\|_2^2} \, \phi_b(t - t_k^{vu})$	$\mathcal{O}(n \cdot d \cdot B)$

HP Hawkes Process Model (*cf.* Blundell *et al.*, 2012)

PLS Poisson-rate Latent Space Model

RLS Reciprocal Latent Space Model (*cf.* Linderman and Adams, 2014)

BLS Base-rate Latent Space Model

DLS Dual Latent Space Model

Perform maximum a posteriori (MAP) inference for all model parameters.

Full log-likelihood of observed communications $\left\{(u, v, \{t_i^{uv}\}_{i=1}^{n_{uv}})\right\}_{u,v \in V}$:

$$\log \mathcal{L} = \sum_{\substack{u,v=1 \\ u \neq v}}^{n} \left\{ -\Lambda_{uv}(0, T) + \sum_{i=1}^{n_{uv}} \log \lambda_{uv}(t_i^{uv}) \right\}$$

and gradients available in closed form.

Cache data statistics: $\forall u, v \in V, b = 1, \ldots, B$,

$$\Delta_{b,T}^{vu} \triangleq \sum_{k=1}^{n_{vu}} [\Phi_b(T - t_k^{vu}) - \Phi_b(0)] \qquad \delta_{b,i}^{uv} \triangleq \sum_{k:\, t_k^{vu} < t_i^{uv}} \phi_b(t_i^{uv} - t_k^{vu})$$

Optimization via L-BFGS-B (Byrd et al., 1995).

Experimental Evaluation

Dataset Description

Enron Core network of the Enron email dataset during 01/2000–04/2002; 155 Enron executives, 9,646 email messages over 453 days.

Email Email communications within Purdue University during 07/2011—02/2012. Filtered out mailing-lists, extracted 100 largest-degree nodes; 34,438 emails over 237 days.

Facebook Facebook wall messages among Purdue students during 03/2007–03/2008. Extracted 100 largest-degree nodes; 18,865 wall messages over 385 days.

Experiment Setup

For each dataset, sort the messages according to their time-stamps.
Training set: first 70% messages. Test set: remaining 30% messages.

$B = 4$ Hawkes basis kernels: $\phi_1(t) = e^{-\frac{t}{1/24}}$, $\phi_2(t) = e^{-t}$, $\phi_3(t) = e^{-t/7}$, $\phi_4(t) = e^{-t/7} \sin^2\left(\frac{\pi t}{7}\right)$.

Randomly sample 100 time-points t_i during the test period.

For each pair of nodes, the predicted probability that at least a link will appear in the $[t_i, t_i + \delta)$ time window is $p_{uv} = 1 - \exp\{-\int_{t_i}^{t_i+\delta} \lambda_{uv}(s)\, ds\}$. Set $\delta = 14$ (days).

For each t_i, compute ROC-AUC over all node-pairs. Report mean and SD across t_i's.

Model	Enron	Email	Facebook
HP	0.750 (0.070)	0.881 (0.088)	0.931 (0.095)
PLS	0.681 (0.041)	0.843 (0.087)	0.874 (0.078)
BLS	0.738 (0.065)	0.868 (0.095)	0.927 (0.096)
RLS	0.750 (0.070)	0.881 (0.088)	0.931 (0.095)
DLS	**0.928 (0.018)**	**0.971 (0.006)**	**0.979 (0.008)**

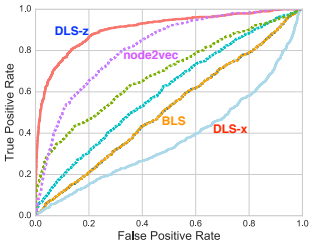
Given the learned latent feature vectors $\{z_v\}_{v \in V}$, compute the predicted probability that an edge exists in the test graph via $p_{uv} \propto e^{-\|z_u - z_v\|_2^2}$.

Compute (static) link prediction ROC-AUC for all pairs of nodes.

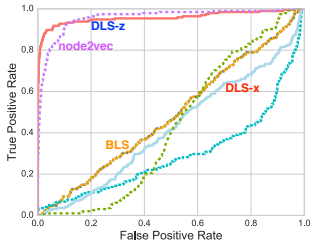
Static graph embeddings: spectral clustering and node2vec (Grover and Leskovec, 2016).

Model	Enron	Email	Facebook
PLS	0.512	0.483	0.505
BLS	0.512	0.483	0.505
RLS	0.601	0.295	0.445
DLS	**0.906**	**0.958**	0.947
Spectral	0.687	0.428	0.452
node2vec	0.829	**0.958**	**0.956**

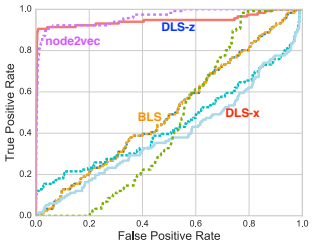
Homophily and Reciprocal Latent Spaces



Enron

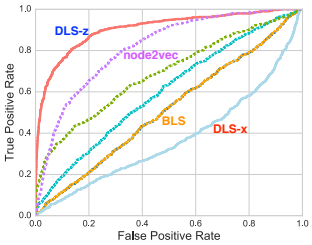


Email

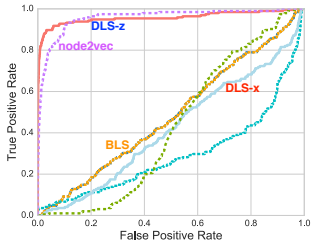


Facebook

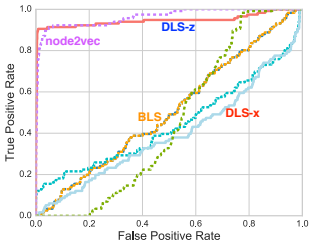
Homophily and Reciprocal Latent Spaces



Enron



Email



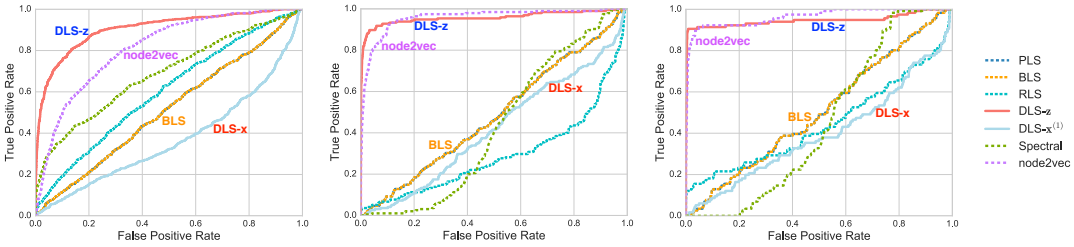
Facebook

(DLS) $$\lambda_{uv}(t) = \gamma\, e^{-\|\mathbf{z}_u - \mathbf{z}_v\|_2^2} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \beta\, e^{-\|\mathbf{x}_u^{(b)} - \mathbf{x}_v^{(b)}\|_2^2}\, \phi_b(t - t_k^{vu})$$

The homophily latent space $\{\mathbf{z}_v\}_{v \in V}$ outperforms reciprocal latent spaces $\{\{\mathbf{x}_v^{(b)}\}_{v \in V}\}_{b=1}^{B}$.

However, BLS ("DLS - $\{\{\mathbf{x}_v^{(b)}\}_{v \in V}\}_{b=1}^{B}$") also performs poorly in link prediction.

Homophily and Reciprocal Latent Spaces



Enron Email Facebook

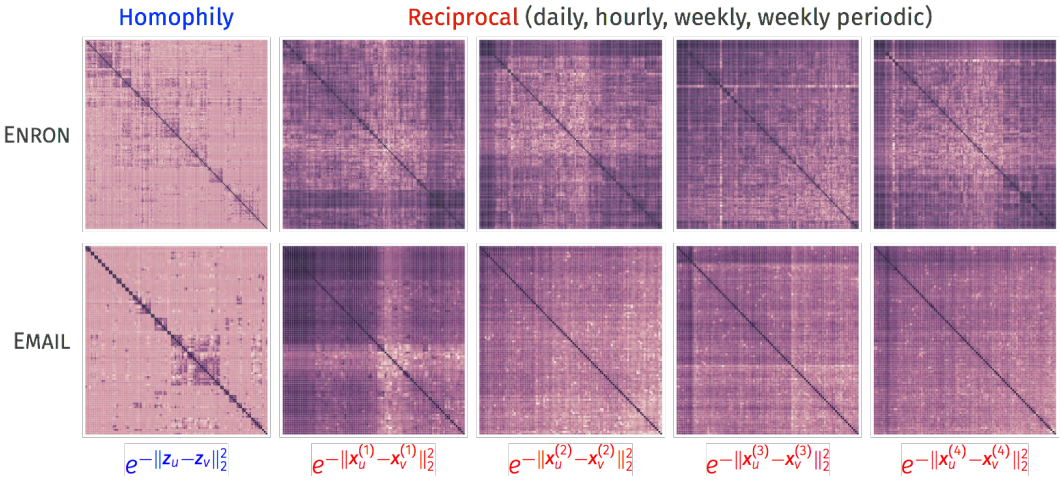
(DLS) $$\lambda_{uv}(t) = \gamma\, e^{-\|z_u - z_v\|_2^2} + \sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \beta\, e^{-\|x_u^{(b)} - x_v^{(b)}\|_2^2}\, \phi_b(t - t_k^{vu})$$

The homophily latent space $\{z_v\}_{v \in V}$ outperforms reciprocal latent spaces $\{\{x_v^{(b)}\}_{v \in V}\}_{b=1}^{B}$.

However, BLS ("DLS - $\{\{x_v^{(b)}\}_{v \in V}\}_{b=1}^{B}$") also performs poorly in link prediction.

Explanation: The reciprocal latent spaces achieve a denoising effect that "explains away" communications primarily driven by reciprocity.

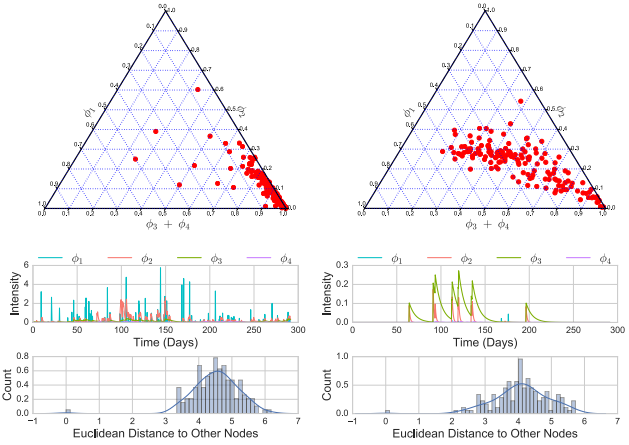
Visualization of Inferred Node-Similarity Matrices



For each pair of nodes u and v, compute relative similarity in the b-th kernel via

$$p_{uv}^{(b)} \triangleq \frac{e^{-\|x_u^{(b)} - x_v^{(b)}\|_2^2}}{\sum_{h=1}^{B} e^{-\|x_u^{(h)} - x_v^{(h)}\|_2^2}}$$

Embed pairs of nodes onto a probability simplex where each pair $u, v \in V$ is represented by a point $(p_{uv}^{(1)}, \ldots, p_{uv}^{(B)})^\top$.



(e) Enron Node #108 (f) Enron Node #92

Concluding Remarks

Model dynamic network data with latent space point process models.

Crucial to account for heterogeneity across both homophily and reciprocity:

- In dynamic link prediction, incorporating the homophily latent space produces a significant gain across all three real-world datasets.
- In static link prediction, the reciprocal latent spaces greatly improve the quality of the estimated homophily latent space.

Our finding is related to that of Rudolph et al. (2016): Modeling each observation conditioned on its context improves the quality of the learned embeddings.

Questions?

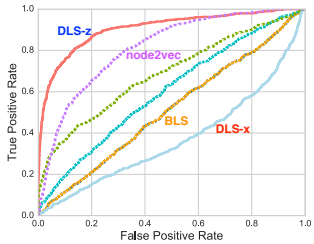
Thank you!

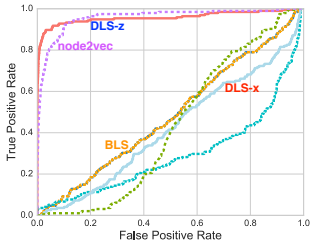
Predictive Log-Likelihood

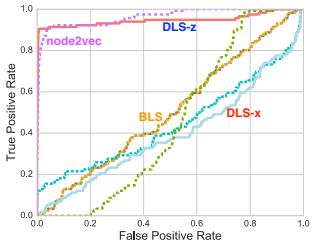
Learn each model on training set; compute predictive log-likelihood on test set.

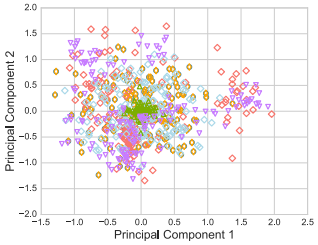
Model	Enron	Email	Facebook
HP	**-16226.155**	-2129.940	-7871.895
PLS	-37803.978	-112684.130	-66742.379
BLS	-21779.686	-9850.932	-12119.869
RLS	-16565.449	-2113.254	-7867.870
DLS	-16422.946	**185.264**	**-6421.609**

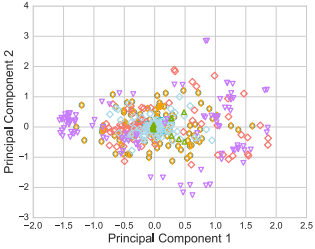
Network Embedding

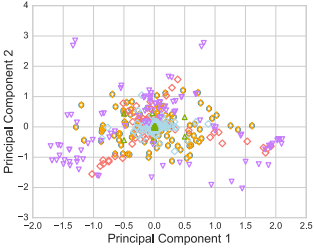


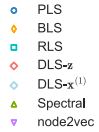












Enron

Email

Facebook