Dynamic Egocentric Models for Citation Networks

Duy Vu
Arthur Asuncion
David Hunter
Padhraic Smyth

To appear in *Proceedings of the 28th International Conference on Machine Learning*, 2011

MURI meeting, June 3, 2011

Outline

Egocentric Counting Processes

- Goal: Model a dynamically evolving network
- Following standard recurrent event theory, place a counting process $N_i(t)$ on node i, $i = 1, \ldots, n$.
- $N_i(t)$ counts the number of "events" involving the ith node.
- Combine $N_i(t)$ gives a multivariate counting process $\mathbf{N}(t) = (N_1(t), \ldots, N_n(t))$.
- Genuinely multivariate; no assumption about the independence of $N_i(t)$.
- "Egocentric" using Carter's terminology because i are nodes, not node pairs.

Modeling of Citation Networks

- New papers join the network over time.
- At arrival, a paper cites others that are already in the network.
- Main dynamic development is the number of citations *received*.
- Thus, $N_i(t)$ equals the cumulative number of citations to paper i at time t.
- "Egocentric" means $N_i(t)$ is ascribed to nodes. Alternative "relational" framework, using $N_{(i,j)}(t)$, is not appropriate here: Relationship (i,j) is at risk of an event (citation) only at a single instant in time.
- Further discussion of general time-varying network modeling ideas given by Butts (2008) and Brandes et al (2009).

The Doob-Meyer Decomposition

Each $N_i(t)$ is nondecreasing in time, so $\mathbf{N}(t)$ may be considered a *submartingale*; i.e., it satisfies

$$E\left[\mathbf{N}(t) \,|\, \text{past up to time } s\right] \geq \mathbf{N}(s) \quad \text{for all } t > s.$$

The Doob-Meyer Decomposition

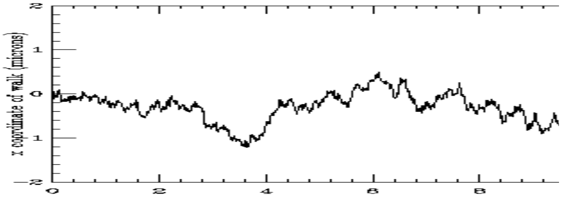
Each $N_i(t)$ is nondecreasing in time, so $\mathbf{N}(t)$ may be considered a *submartingale*; i.e., it satisfies

$$E\left[\mathbf{N}(t) \,|\, \text{past up to time } s\right] \geq \mathbf{N}(s) \quad \text{for all } t > s.$$

Any submartingale may be uniquely decomposed as

$$\mathbf{N}(t) = \int_0^t \boldsymbol{\lambda}(s)\, ds + \mathbf{M}(t) :$$

- $\boldsymbol{\lambda}(t)$ is the “signal” at time t (this *intensity function* is what we will model)
- $\mathbf{M}(t)$ is a continuous-time Martingale.



Modeling the Intensity Process

The intensity process for node i is given by

$$\lambda_i(t|\mathbf{H}_{t^-}) = Y_i(t)\alpha_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t)\right),$$

where

Modeling the Intensity Process

The intensity process for node i is given by

$$\lambda_i(t|\mathbf{H}_{t^-}) = Y_i(t)\alpha_0(t) \exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t)\right),$$

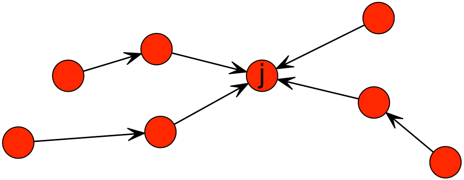
where

- $Y_i(t) = I(t > t_i^{\text{arr}})$ is the "at-risk indicator"
- $\mathbf{H}_{t^-}$ is the past of the network up to but not including time t
- $\alpha_0(t)$ is the baseline hazard function
- $\boldsymbol{\beta}$ is the vector of coefficients to estimate
- $\mathbf{s}_i(t) = (s_{i1}(t), \ldots, s_{ip}(t))$ is a p-vector of statistics for paper i

Preferential Attachment Statistics

For each cited paper j already in the network...

- First-order PA: $s_{j1}(t) = \sum_{i=1}^{N} y_{ij}(t)$. *"Rich get richer" effect*
- Second-order PA: $s_{j2}(t) = \sum_{i \neq k} y_{ki}(t) y_{ij}(t)$.
  Effect due to being cited by well-cited papers
- Recency-based first-order PA (we take $T_w = 180$ days):
  $s_{j3}(t) = \sum_{i=1}^{N} y_{ij}(t) I(t - t_i^{\mathrm{arr}} < T_w)$.
  Temporary elevation of citation intensity after recent citations

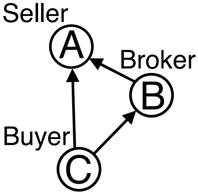


Statistics in red are time-dependent. Others are fixed once j joins the network.

Triangle Statistics

For each cited paper j already in the network...

- "Seller" statistic: $s_{j4}(t) = \sum_{i \neq k} y_{ki}(t) y_{ij}(t) y_{kj}(t)$.
- "Broker" statistic: $s_{j5}(t) = \sum_{i \neq k} y_{kj}(t) y_{ji}(t) y_{ki}(t)$.
- "Buyer" statistic: $s_{j6}(t) = \sum_{i \neq k} y_{jk}(t) y_{ki}(t) y_{ji}(t)$.

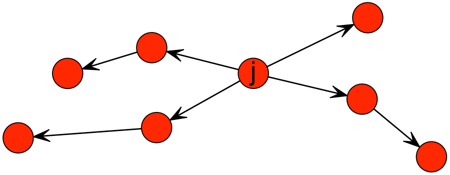


Statistics in red are time-dependent. Others are fixed once j joins the network.

Out-Path Statistics

For each cited paper j already in the network...

- First-order out-degree (OD): $s_{j7}(t) = \sum_{i=1}^{N} y_{ji}(t)$.
- Second-order OD: $s_{j8}(t) = \sum_{i \neq k} y_{jk}(t) y_{ki}(t)$.



Statistics in red are time-dependent. Others are fixed once j joins the network.

Topic Modeling Statistics

Additional statistics, using abstract text if available, as follows:

- An LDA model (Blei et al, 2003) is learned on the training set.
- Topic proportions $\boldsymbol{\theta}$ generated for each training node.
- LDA model also used to estimate topic proportions $\boldsymbol{\theta}$ for each node in the test set.
- We construct a vector of similarity statistics:

$$\mathbf{s}_j^{\mathrm{LDA}}(t_i^{\mathrm{arr}}) = \boldsymbol{\theta}_i \circ \boldsymbol{\theta}_j,$$

  where $\circ$ denotes the element-wise product of two vectors.
- We use 50 topics; each s_j component has a corresponding β.

Partial Likelihood

Recall: The intensity process for node i is

$$\lambda_i(t|\mathbf{H}_{t^-}) = Y_i(t)\alpha_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t)\right).$$

If $\alpha_0(t) \equiv \alpha_0(t, \gamma)$, we may use the "local Poisson-ness" of the multivariate counting process to obtain (and maximize) a likelihood function (details omitted).

Partial Likelihood

Recall: The intensity process for node i is

$$\lambda_i(t|\mathbf{H}_{t^-}) = Y_i(t)\alpha_0(t) \exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t)\right).$$

If $\alpha_0(t) \equiv \alpha_0(t, \boldsymbol{\gamma})$, we may use the "local Poisson-ness" of the multivariate counting process to obtain (and maximize) a likelihood function (details omitted).

However, we treat α_0 as a nuisance parameter and take a partial likelihood approach as in Cox (1972): Maximize

$$L(\boldsymbol{\beta}) = \prod_{e=1}^{m} \frac{\exp\left(\boldsymbol{\beta}^\top \mathbf{s}_{i_e}(t_e)\right)}{\sum_{i=1}^{n} Y_i(t_e) \exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t_e)\right)} = \prod_{e=1}^{m} \frac{\exp\left(\boldsymbol{\beta}^\top \mathbf{s}_{i_e}(t_e)\right)}{\kappa(t_e)}$$

Trick: Write $\kappa(t_e) = \kappa(t_{e-1}) + \Delta\kappa(t_e)$, then optimize $\Delta\kappa(t_e)$ calculation.

Data Sets We Analyzed

Three citation network datasets from the physics literature:

1. **APS:** Articles in *Physical Review Letters*, *Physical Review*, and *Reviews of Modern Physics* from 1893 through 2009. Timestamps are monthly for older, daily for more recent.
2. **arXiv-PH:** arXiv high-energy physics phenomenology articles from Jan. 1993 to Mar. 2002. Timestamps are daily.
3. **arXiv-TH:** High-energy physics theory articles spanning from January 1993 to April 2003. Timestamps are continuous-time (millisecond resolution). Also includes text of paper abstracts.

	Papers	Citations	Unique Times
APS	463,348	4,708,819	5,134
arXiv-PH	38,557	345,603	3,209
arXiv-TH	29,557	352,807	25,004

Three Phases

1. **Statistics-building phase:** Construct network history and build up network statistics.
2. **Training phase:** Construct partial likelihood and estimate model coefficients.
3. **Test phase:** Evaluate predictive capability of the learned model.

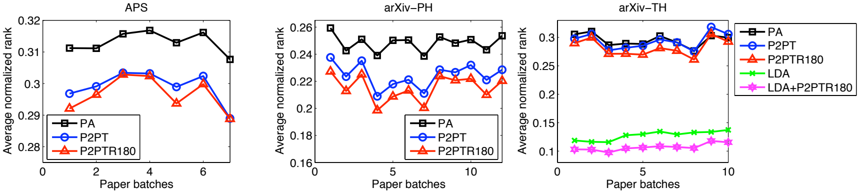
Statistics-building is ongoing even through the training and test phases. The phases are split along citation event times.

Number of unique citation event times in the three phases:

	Building	Training	Test
APS	4,934	100	100
arXiv-PH	2,209	500	500
arXiv-TH	19,004	1000	5000

Average Normalized Ranks

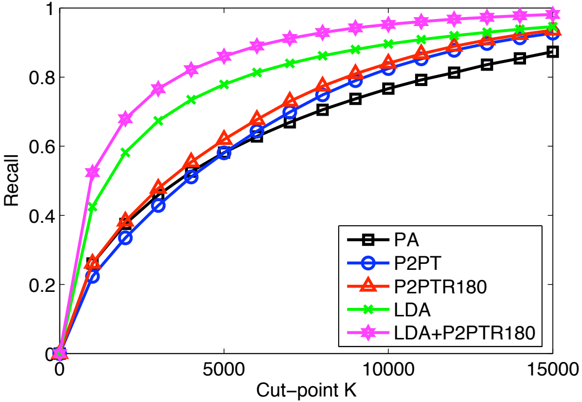
- Compute "rank" for each true citation among sorted likelihoods of each possible citation.
- Normalize by dividing by the number of possible citations.
- Average of the normalized ranks of each observed citation.
- Lower rank indicates better predictive performance.



- Batch sizes are 3000, 500, 500, respectively.
- **PA**: pref. attach only ($s_1(t)$); **P2PT**: $s_1, \ldots, s_8$ except s_3;
- **P2PTR180**: $s_1, \ldots, s_8$; **LDA**: LDA stats only

Recall Performance

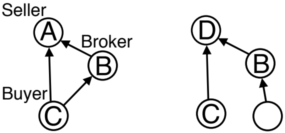
Recall: Proportion of true citations among largest K likelihoods.



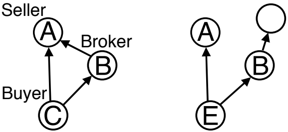
- **PA**: pref. attach only $(s_1(t))$; **P2PT**: $s_1, \ldots, s_8$ except s_3;
- **P2PTR180**: $s_1, \ldots, s_8$; **LDA**: LDA stats only

Coefficient Estimates for LDA + P2PTR180 Model

Statistics	Coefficients (β)
s_1 (PA)	0.01362
s_2 (2^{nd} PA)	0.00012
s_3 (PA-180)	0.02052
s_4 (Seller)	-0.00126
s_5 (Broker)	-0.00066
s_6 (Buyer)	-0.00387
s_7 (1^{st} OD)	0.00090
s_8 (2^{nd} OD)	0.02052



Diverse seller effect:
D more likely cited than *A*.



Diverse buyer effect:
E more likely cited than *C*.

References

Blei, D.M., Ng, A.Y., and Jordan, M.I.
Latent Dirichlet allocation.
Journal of Machine Learning Research, 3:993–1022, 2003.

Brandes, U., Lerner, J., and Snijders, T.A.B.
Networks evolving step by step: Statistical analysis of dyadic event data.
In *Advances in Social Network Analysis and Mining*, pp. 200–205. IEEE, 2009.

Butts, C.T.
A relational event framework for social action.
Sociological Methodology, 38(1):155–200, 2008.

Cox, D. R.
Regression models and life-tables.
Journal of the Royal Statistical Society, Series B, 34:187–220, 1972.

Why Such Long Building Phases?

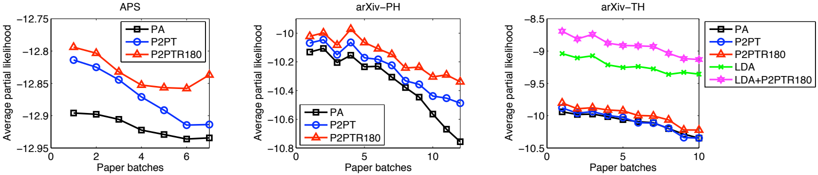
- The lengthy building phase mitigates truncation effects at the beginning of network formation and effects of severely grouped event times
- Training and test windows still cover a substantial period of time (e.g. 2.5 years for APS)
- Performance is relatively invariant to the size of the training windows. We achieved essentially the same results using windows of size 2000 and 5000 for arXiv-TH.

Number of unique citation event times in the three phases:

	Building	Training	Test
APS	4,934	100	100
arXiv-PH	2,209	500	500
arXiv-TH	19,004	1000	5000

Average Partial Loglikelihood

- Compute average of the partial likelihoods for each citation event.



- Batch sizes are 3000, 500, 500, respectively.
- **PA**: pref. attach only $(s_1(t))$; **P2PT**: $s_1, \ldots, s_8$ except s_3;
- **P2PTR180**: $s_1, \ldots, s_8$; **LDA**: LDA stats only