Scalable statistical estimation methods for large, time-varying networks

Duy Vu[1]
Arthur Asuncion[2]
David Hunter[1]
Padhraic Smyth[3]

[1]Department of Statistics, Penn State
[2]Google Inc.
[3]Department of Computer Science, UC-Irvine
Supported by ONR MURI Award Number N00014-08-1-1015

MURI grant meeting, January 10, 2012

Outline

Counting processes for evolving networks

Egocentric Models vs. Relational Models

Egocentric Network Models

Model Structure

Application: Citation Networks

Refer to Vu et al (ICML 2011) for further details

Relational Network Models

Refer to Vu et al (NIPS 2011) for further details

See also Perry and Wolfe (2010)

Outline

Counting processes for evolving networks

Egocentric Models vs. Relational Models

Egocentric Network Models

Model Structure

Application: Citation Networks

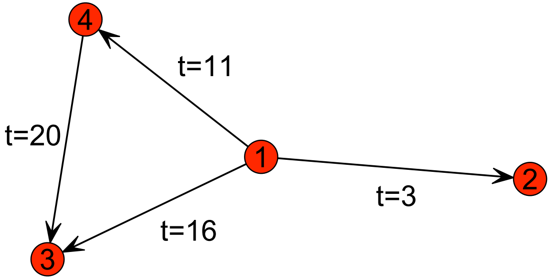
Refer to Vu et al (ICML 2011) for further details

Relational Network Models

Refer to Vu et al (NIPS 2011) for further details

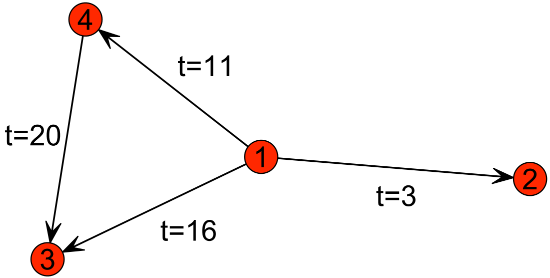
See also Perry and Wolfe (2010)

Counting Processes for networks



- Goal: Model a dynamically evolving network using counting processes.

Counting Processes for networks



- Goal: Model a dynamically evolving network using counting processes.

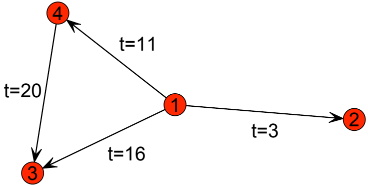
- Two possibilities (using terminology of Butts, 2008):
  - **Egocentric:** The counting process $N_i(t) =$ cumulative number of "events" involving the ith node by time t.
  - **Relational:** The counting process $N_{ij}(t) =$ cumulative number of "events" involving the (i, j)th node pair by time t.

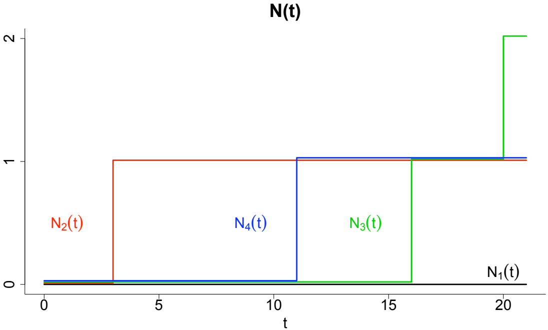
Counting Process approach: Egocentric example

- Combine the $N_i(t)$ to give a multivariate counting process

$$\mathbf{N}(t) = (N_1(t), \ldots, N_n(t)).$$

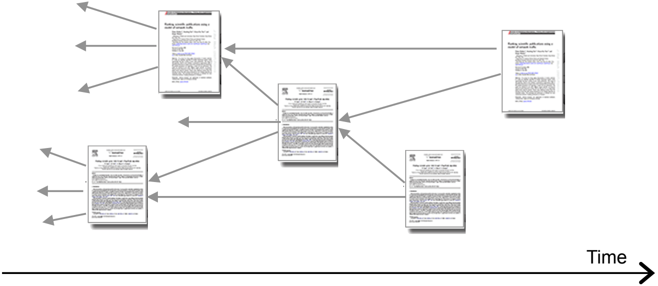
- Genuinely multivariate; no assumption about the independence of $N_i(t)$.





Egocentric Example: 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: Number of citations *received*.

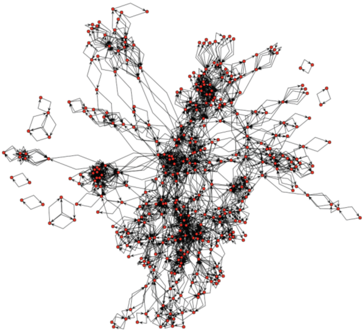


- $N_i(t)$: Number of citations to paper i by time t.
- "At-risk" indicator $R_i(t)$: Equal to $I\{t_i^{\mathrm{arr}} < t\}$.

Relational Example: Modeling a network of contacts

- Metafilter: Community weblog for sharing links and discussing content among its users.
- Pattern of contacts: Dynamically evolving network
- Links are *non-recurrent*; i.e., $N_{ij}(t)$ is either 0 or 1.
- "At-risk" indicator $R_{ij}(t) = I\{\max(t_i^{\text{arr}}, t_j^{\text{arr}}) < t < t_{e_{ij}}\}$.

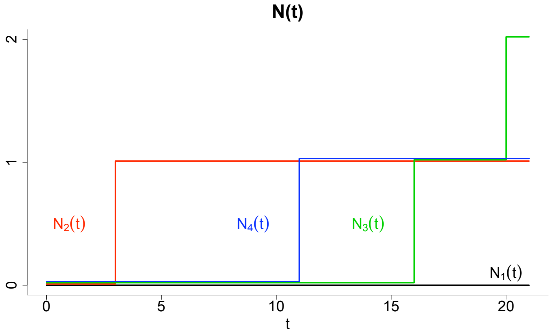
contacter	contactee	date
1	14155	2004-06-15 12:00:00.000
1	2238	2004-06-15 12:00:00.000
1	14275	2004-06-15 12:00:00.000
...		
13099	7683	2004-06-17 16:31:51.040
15231	14752	2004-06-17 16:31:51.040
...		
45087	7610	2007-10-31 12:23:15.683
16719	61	2007-10-31 13:28:38.670
48758	1	2007-10-31 13:47:16.843



Submartingales: Egocentric Case

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

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

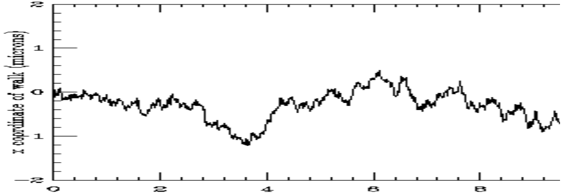


Theory: The Doob-Meyer Decomposition

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, called the *intensity function*
- $\mathbf{M}(t)$ is the "noise," a continuous-time Martingale.



- We will model each $\lambda_i(t)$ or $\lambda_{ij}(t)$.

Outline

Counting processes for evolving networks
- Egocentric Models vs. Relational Models

Egocentric Network Models
- Model Structure
- Application: Citation Networks
  - *Refer to Vu et al (ICML 2011) for further details*

Relational Network Models
  - *Refer to Vu et al (NIPS 2011) for further details*
  - *See also Perry and Wolfe (2010)*

Modeling the Intensity Process, Part I: Egocentric Case

The intensity process for node i is given by

- Cox Proportional Hazard Model, fixed coefficients:

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

- Aalen additive model, time-varying coefficients:

$$\lambda_i(t|\mathbf{H}_{t^-}) = R_i(t)\big(\beta_0(t) + \boldsymbol{\beta}(t)^\top \mathbf{s}_i(t)\big),$$

where

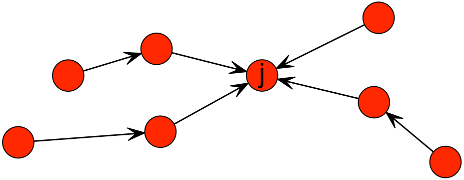
  - $R_i(t) = I(t > t_i^{\text{arr}})$ is the "at-risk indicator"
  - $\mathbf{H}_{t^-}$ is the past of the network up to but not including time t
  - $\alpha_0(t)$ or $\beta_0(t)$ is the baseline hazard function
  - $\boldsymbol{\beta}$ is the vector of coefficients to estimate
  - $\mathbf{s}_i(t) = (s_{i1}(t), \ldots, s_{ip}(t))$ is a p-vector of statistics for paper i

Let us consider the citation network examples. . .

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



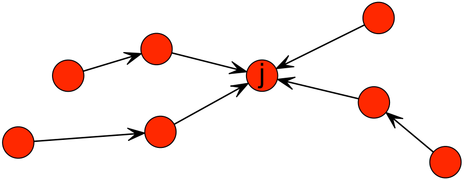
Statistics in red are time-dependent. Others are fixed once j joins the network.

NB: $\mathbf{y}(t^-)$ *is the network just prior to time* t.

Recency PA Statistic

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

- 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



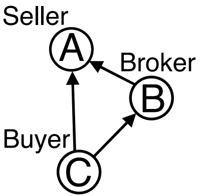
Statistics in red are time-dependent. Others are fixed once j joins the network.

NB: $\mathbf{y}(t^-)$ *is the network just prior to time* t.

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: $s_{j6}(t) = \sum_{i \neq k} y_{jk}(t) y_{ki}(t) y_{ji}(t^-)$.



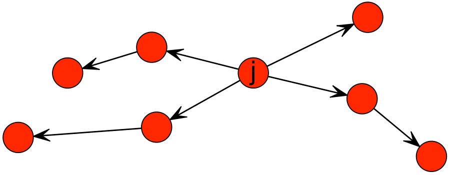
Statistics in red are time-dependent. Others are fixed once j joins the network.

NB: $\mathbf{y}(t^-)$ is the network just prior to time t.

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.

NB: $\mathbf{y}(t^-)$ is the network just prior to time t.

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 $\mathbf{s}_j$ component has a corresponding β.

Partial Likelihood (how to fit the Cox PH Model)

Recall: The intensity process for node i is

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

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).

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

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

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

Least Squares (How to fit the Aalen Additive Model)

Recall: The intensity process for node i is

$$\lambda_i(t|\mathbf{H}_{t^-}) = R_i(t)\big(\beta_0(t) + \boldsymbol{\beta}(t)^\top \mathbf{s}_i(t)\big).$$

- We do inference not for the β_k but rather for their time-integrals

$$B_k(t) = \int_0^t \beta_k(s)ds. \tag{1}$$

- Then

$$\hat{\mathbf{B}}(t) = \sum_{t_e \le t} J(t_e)\big[\mathbf{W}(t_e)^\top \mathbf{W}(t_e)\big]^{-1} \mathbf{W}(t_e)^\top \Delta\mathbf{N}(t_e), \tag{2}$$

  where
  - $\mathbf{W}(t)$ is $N(N-1)\times p$ with (i,j)th row $R_{ij}(t)\mathbf{s}(i,j,t)^\top$;
  - $J(t)$ is the indicator that $\mathbf{W}(t)$ has full column rank.

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

Why Such Long Building Phases?

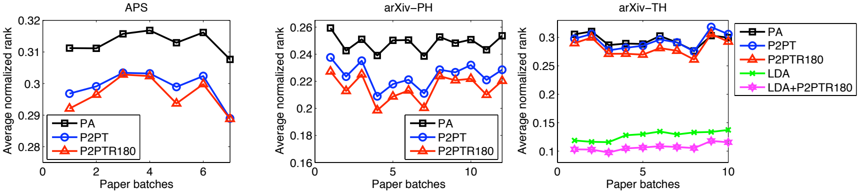
- The lengthy building phase mitigates truncation effects at the beginning of network formation and effects of severely grouped event times
- Training and test windows still cover a substantial period of time (e.g. 2.5 years for APS)
- Performance is relatively invariant to the size of the training windows. 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 Normalized Ranks

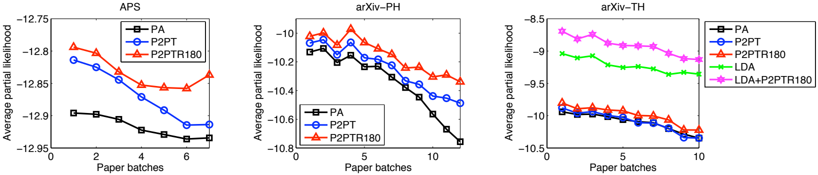
- Compute "rank" for each true citation among sorted likelihoods of each possible citation.
- Normalize by dividing by the number of possible citations.
- Average of the normalized ranks of each observed citation.
- Lower rank indicates better predictive performance.



- 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

Average Partial Loglikelihood

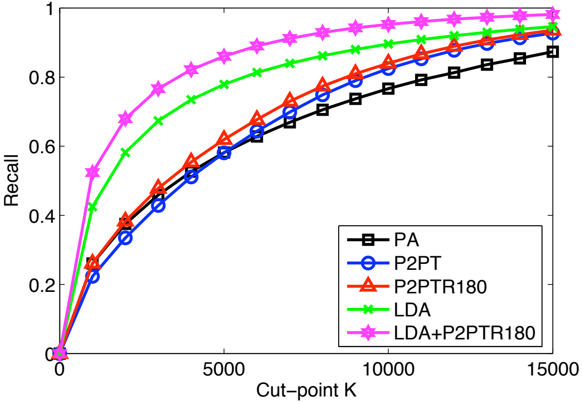
- Compute average of the partial likelihoods for each citation event.



- 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

Recall Performance

Recall: Proportion of true citations among largest K likelihoods.

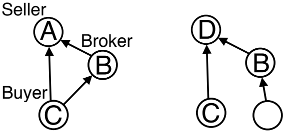


- **PA**: pref. attach only $(s_1(t))$; **P2PT**: $s_1, \ldots, s_8$ except s_3;
- **P2PTR180**: $s_1, \ldots, s_8$; **LDA**: LDA stats only

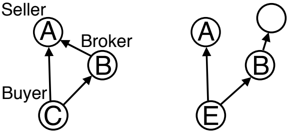
Coefficient Estimates for LDA + P2PTR180 Model

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

All coefficient estimates are significant at the 0.0001 level.



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



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

Outline

Counting processes for evolving networks

Egocentric Models vs. Relational Models

Egocentric Network Models

Model Structure

Application: Citation Networks

Refer to Vu et al (ICML 2011) for further details

Relational Network Models

Refer to Vu et al (NIPS 2011) for further details

See also Perry and Wolfe (2010)

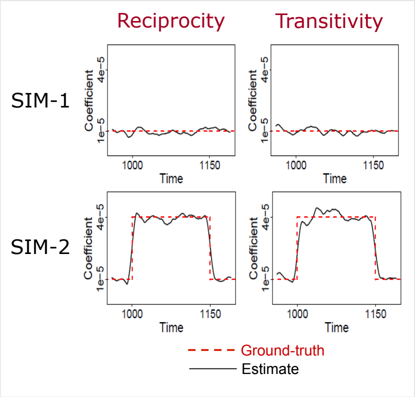
Network Data Sets

- Simulated data (SIM-1, SIM-2)
- Real networks:
  - Irvine: an online social network at UC Irvine (4/2004 to 10/2004).
  - MetaFilter: a community weblog contact network (8/2007 to 2/2011).

	Nodes	Edges	Stats-Building Phase	Training Phase	Test Phase
Irvine	1,899	20,296	7,073	7,646	5,507
MetaFilter	51,362	76,791	60,376	8,763	7,620

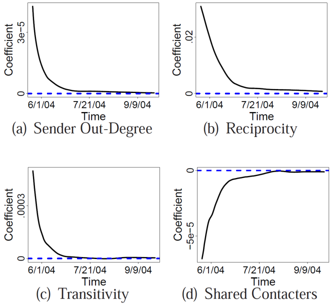
Recovering Time-Varying Coefficients

- Simulated data from ground-truth coefficients:
  - **SIM-1**: Constant coefficients for reciprocity, transitivity.
  - **SIM-2**: Varying coefficients for reciprocity, transitivity.
- Learned time-varying coefficients of Aalen model on simulated data.

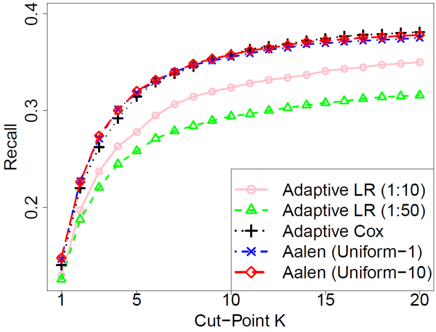


Irvine Data Set

- Aalen coefficients suggest two distinct phases of network evolution, consistent with an independent analysis [Panzarasa et al, 2009].
- On prediction experiments, Aalen/Cox outperforms logistic regression.

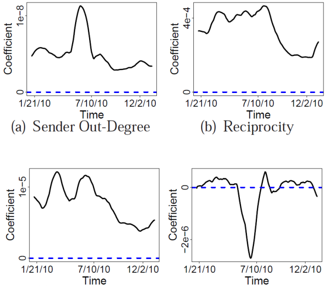


(a) Sender Out-Degree (b) Reciprocity (c) Transitivity (d) Shared Contacters

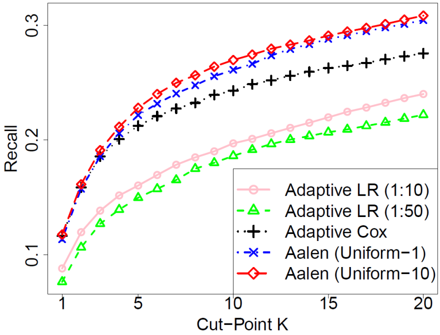


Metafilter Data Set

- Network effects continuously change over time.
- Time-varying Aalen model outperforms Cox model.



(a) Sender Out-Degree (b) Reciprocity (c) Transitivity (d) Shared Contacters



Cited References

Aalen, O. O., Borgan, O., and Gjessing, H. K.
Survival and Event History Analysis: A Process Point of View
Springer, 2008.

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

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.

Perry, P. O. and Wolfe, P. J.
Point process modeling for directed interaction networks
arXiv:1011.1703v1 [stat.ME] 8 Nov 2010

Salathé, M. and Khandelwal, S.
Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control
PLoS Computational Biology, 7(10): e1002199. doi:10.1371/journal.pcbi.1002199, 2011.

Vu, D. Q., Asuncion, A. U., Hunter, D. R., and Smyth, P.
Dynamic Egocentric Models for Citation Networks,
Proceedings of the 28th International Conference on Machine Learning (ICML 2011), 857–864, 2011.

Vu, D. Q., Asuncion, A. U., Hunter, D. R., and Smyth, P.
Continuous-Time Regression Models for Longitudinal Networks
Advances in Neural Information Processing Systems 24 (NIPS 2011), to appear.