# Dynamic Egocentric Models for Citation Networks

**Duy Q. Vu** * DQV100@STAT.PSU.EDU
**Arthur U. Asuncion** † ASUNCION@ICS.UCI.EDU
**David R. Hunter** * DHUNTER@STAT.PSU.EDU
**Padhraic Smyth** † SMYTH@ICS.UCI.EDU
* Department of Statistics, Pennsylvania State University, University Park, PA, 16802, USA
† Department of Computer Science, University of California, Irvine, CA, 92697, USA

## Abstract

The analysis of the formation and evolution of networks over time is of fundamental importance to social science, biology, and many other fields. While longitudinal network data sets are increasingly being recorded at the granularity of individual time-stamped events, most studies only focus on collapsed cross-sectional snapshots of the network. In this paper, we introduce a dynamic egocentric framework that models continuous-time network data using multivariate counting processes. For inference, an efficient partial likelihood approach is used, allowing our methods to scale to large networks. We apply our techniques to various citation networks and demonstrate the predictive power and interpretability of the learned statistical models.

## 1. Introduction

Network analysis is of increasing interest to researchers and practitioners due to the emergence of large-scale social networks, biological and protein interaction data, citation graphs, and networks in many other fields. Since most of these networks are dynamic and evolve over time, there is increasing motivation to develop longitudinal network models, i.e., models for networks over time. Researchers have largely focused to date on analyzing discrete "snapshot" or collapsed panel data (e.g., Hanneke & Xing, 2006; Fu et al., 2009; Wyatt et al., 2010). While continuous-time models have been fitted on small networks (Wasserman, 1980; Snijders, 2005), the development and fitting of dynamic statistical models for large-scale data sets at a fine temporal granularity is still relatively unexplored.

Appearing in *Proceedings of the $28^{th}$ International Conference on Machine Learning*, Bellevue, WA, USA, 2011. 

This paper introduces a general dynamic network modeling framework that can model time-stamped event data. Our approach extends work on survival and event history analysis (Andersen et al., 1993; Butts, 2008) to large-scale network modeling and uses multivariate counting processes to model network dynamics. A benefit of this statistical framework is that it can handle arbitrary network and nodal statistics; for example, we empirically show how incorporating textual information with network statistics improves predictive performance.

In this paper, we focus on citation network analysis, which is an area of interest to machine learning and bibliometrics. Specifically, we consider processes where nodes create edges over time to nodes that joined the network earlier, and we take a restricted "egocentric" perspective that only models nodal processes for efficiency. We do not explicitly discuss more general processes (involving edge-based dynamics) in this paper, but our framework can be generalized. Moreover, we develop an efficient inference scheme that optimizes a partial likelihood that ignores the precise event times and only considers the event-to-event ordering, though we also discuss how a baseline rate of citations per unit time may be estimated if event timing is of interest.

The specific contributions of this paper are as follows:

1. We propose a statistical egocentric modeling framework for fine-grained longitudinal network data that allows for arbitrary network and nodal statistics.
2. An efficient inference scheme based on partial likelihood optimization is presented, allowing this approach to scale to large data sets.
3. We provide an empirical analysis of the predictive power and interpretability of the learned egocentric models on several real-world citation networks.

In the following sections, we introduce the egocentric network modeling framework and detail the inference algorithms. Then we empirically analyze several citation net-

works using this framework. Finally, we review related work and conclude with future research directions.

## 2. The Egocentric Network Model

In this section, we formulate the general egocentric modeling framework and provide examples of network statistics that can be specifically applied to citation network data.

### 2.1. General framework

We demonstrate below how models for counting processes (Andersen & Gill, 1982) may be used in the context of network analysis, where we need to account for the interdependence among individual processes on nodes or edges. Our egocentric framework aims to model a dynamically evolving network by placing a counting process $N_i(t)$ on node $i$, $i = 1, \ldots, n$, where $N_i(t)$ counts the number of "events" (defined based on the context) involving the $i$th node. Recurrent events have been studied extensively in the statistical literature on survival and event history analysis (e.g., Andersen et al., 1993). The basic idea of this framework is to model how the current network history influences its future development. Combining individual counting processes of all nodes results in a multivariate counting process $\mathbf{N}(t) = (N_1(t), \ldots, N_n(t))$. This counting process is genuinely multivariate – it makes no assumption about the independence of individual nodal counting processes.

This modeling framework is quite general in applicability; we apply it here to the context of citation networks. In a citation network, new papers join the network over time. At their arrival times, these papers cite others that have already been in the network. Since new papers make citations to other papers only once in their lifetimes, the main dynamic development of this network is the number of citations that papers receive over time; thus, we take $N_i(t)$ to equal the cumulative number of citations to paper $i$ at time $t$.

Since the counting process is nondecreasing in time, it may be considered a *submartingale*; i.e., it satisfies

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

This is a standard way to model time-to-event data, though we will not delve deeply into specifics here; we refer the interested reader to the textbooks of Aalen et al. (2008) and Andersen et al. (1993). The general idea is that a submartingale may be (uniquely) decomposed as

$$\mathbf{N}(t) = \int_0^t \boldsymbol{\lambda}(s)\, ds + \mathbf{M}(t), \tag{1}$$

where the first term on the right is the "signal" at time $t$ and the second term, called a continuous-time martingale, is random "noise". Our attention will focus on modeling $\boldsymbol{\lambda}(t)$, the so-called intensity function.

To model $\boldsymbol{\lambda}(t)$, we shall denote the entire past of the network up to but not including time $t$ by $\mathbf{H}_{t^-}$ and assume that the intensity process for node $i$ is given by

$$\lambda_i(t|\mathbf{H}_{t^-}) = Y_i(t)\alpha_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t)\right), \tag{2}$$

where the "at risk" indicator function $Y_i(t)$ is defined according to the context. For citation networks, we take $Y_i(t) = I(t > t_i^{\text{arr}})$ to be 1 if the current time $t$ is greater than the arrival time $t_i^{\text{arr}}$ of node (paper) $i$. In Eq (2), $\alpha_0(t)$ represents the baseline hazard function, $\boldsymbol{\beta}$ is the vector of coefficients to estimate, and $\mathbf{s}_i(t) = (s_{i1}(t), \ldots, s_{ip}(t))$ is a vector of $p$ statistics for paper $i$ constructed based on $\mathbf{H}_{t^-}$. These statistics can be time-invariant or time-dependent; they are discussed further in Section 2.2.

The term "egocentric" in this context signifies that the counting process $N_i(t)$ is ascribed to nodes. An alternative "relational" framework, which instead defines counting processes $N_{(i,j)}(t)$ on node pairs $(i, j)$, is not appropriate for citation networks because, in the language of statistical survival analysis (e.g., Aalen et al., 2008), relationship $(i, j)$ is at risk of an event (citation) only at a single instant in time. Nonetheless, there are contexts in which both an egocentric approach and a relational approach would be appropriate, and further discussion of general time-varying network modeling ideas is given by Butts (2008) and Brandes et al. (2009). For the remainder of this paper, we drop the general language of nodes and edges, referring instead to the specific case of papers and citations.

### 2.2. Statistics from network history

As described in Section 2.1, our modeling framework can handle arbitrary statistics from the network's history. Here we detail the statistics that will be used in our experiments.

Let $y_{ij}(t)$ denote the value of the directed edge from $i$ to $j$ at time $t$. In other words, $y_{ij}(t)$ equals 1 if both $i$ and $j$ have joined the network by time $t$ and $i$ cites $j$ (we assume that the cited paper $j$ joins the network before the citing paper $i$). For each cited paper $j$ already in the network, we consider three preferential attachment (PA) statistics, three triangle statistics (Figure 1), and two out-path statistics computed based on the network history $\mathbf{H}_{t^-}$ before time $t$:

1. First-order PA: $s_{j1}(t) = \sum_{i=1}^{N} y_{ij}(t)$.
2. Second-order PA: $s_{j2}(t) = \sum_{i \neq k} y_{ki}(t) y_{ij}(t)$.
3. Recency-based first-order PA (where $T_w$ is a specified time window): $s_{j3}(t) = \sum_{i=1}^{N} y_{ij}(t) I(t - t_i^{\text{arr}} < T_w)$.
4. "Seller" statistic: $s_{j4}(t) = \sum_{i \neq k} y_{ki}(t) y_{ij}(t) y_{kj}(t)$.
5. "Broker" statistic: $s_{j5}(t) = \sum_{i \neq k} y_{kj}(t) y_{ji}(t) y_{ki}(t)$.
6. "Buyer" statistic: $s_{j6}(t) = \sum_{i \neq k} y_{jk}(t) y_{ki}(t) y_{ji}(t)$.
7. First-order out-degree (OD): $s_{j7}(t) = \sum_{i=1}^{N} y_{ji}(t)$.
8. Second-order OD: $s_{j8}(t) = \sum_{i \neq k} y_{jk}(t) y_{ki}(t)$.

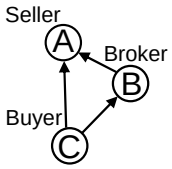



*Figure 1.* The only valid triangle configuration in citation networks and three statistics defined on it. $B$ joins the network after $A$ and cites $A$. Later, $C$ joins the network and cites $A$ and $B$.

Statistics $s_{j1}(t), \ldots, s_{j5}(t)$ are time-dependent while $s_{j6}(t), \ldots, s_{j8}(t)$ are time-invariant in that their values are unchanged after paper $j$ joins the network.

In the context of modeling the counting process of citations for a given paper $j$, the coefficients $\beta_1, \ldots, \beta_8$ corresponding to the eight statistics above should be interpreted as the strength (and direction) of the corresponding effect in the presence of all other effects. For instance, $\beta_1$ measures the "rich get richer" effect, $\beta_2$ measures the effect due to being cited by well-cited papers, and $\beta_3$ measures a type of recency effect, i.e., a temporary elevation of citation intensity following a number of recent citations. All of these effects should be interpreted as though the other effects that are included in the model have already been accounted for.

In addition to the network effects described above, we can take the heterogeneity of the nodes into account by exploiting the information contained in the abstracts of papers (when these abstracts are available). For this purpose, we use Latent Dirichlet Allocation (LDA), also known as topic modeling (Blei et al., 2003), as follows. After converting each abstract into a bag-of-words representation, an LDA model is learned on the training set, and topic proportions $\boldsymbol{\theta}$ as defined by Blei et al. (2003) are generated for each training node. The learned LDA model is also used to estimate topic proportions $\boldsymbol{\theta}$ for each node in the test set through a standard fold-in procedure. With the learned topic proportions, we construct a vector of similarity statistics for each paper $j$ at each arrival time $t_i^{\text{arr}}$ of arriving paper $i$:

$$\mathbf{s}_j^{\text{LDA}}(t_i^{\text{arr}}) = \boldsymbol{\theta}_i \circ \boldsymbol{\theta}_j, \tag{3}$$

where $\circ$ denotes the element-wise product of two vectors. Our approach here is similar to that taken by Chang & Blei (2009). We are free to choose the number of topics; for the arXiv-TH data set with abstracts, we utilize 50 topics. Note that each topic-specific similarity value has a corresponding $\beta$ coefficient.

## 3. Inference

Efficient inference for these models can be achieved using a partial likelihood approach and other computational techniques, which we discuss in the following subsections.

### 3.1. Full versus partial likelihood

There are two different inference approaches for the egocentric network model that differ in the choice of whether to specify a parametric form for the baseline hazard $\alpha_0(t)$ in Eq (2). A fully parametric approach, specifying $\alpha_0(t, \boldsymbol{\gamma})$ as an exponential, Weibull, or piecewise constant distribution, might be useful if applications of interest are time-related, such as predicting the number of citations to a paper in some future time period. Using this approach, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ can be estimated by maximizing the full log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{e=1}^{m} \Bigg[ \log \alpha_0(t_e, \boldsymbol{\gamma}) + \boldsymbol{\beta}^\top \mathbf{s}_{i_e}(t_e) - \sum_{i=1}^{n} \int_{t_{e-1}}^{t_e} Y_i(u) \alpha_0(u, \boldsymbol{\gamma}) \exp\left\{\boldsymbol{\beta}^\top \mathbf{s}_i(u)\right\} du \Bigg] \tag{4}$$

(Aalen et al., 2008), where $m$ is the number of citation events and $i_e$ and $t_e$ denote the paper cited and time of the $e$th event. For the purposes of estimating and then validating $\hat{\beta}$, we may split the citation events into a training set and a test set, respectively, in which case $m$ is the number of events in the training set. The parametric approach of Eq (4) has two disadvantages: First, the specified form may be incorrect, leading to biased estimates; second, the integral above increases computational complexity, particularly if some elements of $\mathbf{s}_i(t)$ depend continuously on $t$.

In other applications, such as network effect inference or citation recommendation for new papers, we are interested in estimating only $\boldsymbol{\beta}$, not $\alpha_0(t)$, so an approach that leaves $\alpha_0(t)$ unspecified may be preferable. This is the approach we take in this paper, considering $\alpha_0(t)$ to be essentially a nuisance parameter. The $\boldsymbol{\beta}$ parameters may then be estimated by maximizing the partial likelihood of Cox (1972):

$$L(\boldsymbol{\beta}) = \prod_{e=1}^{m} \frac{\exp\left(\boldsymbol{\beta}^\top \mathbf{s}_{i_e}(t_e)\right)}{\sum_{i=1}^{n} Y_i(t_e) \exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t_e)\right)}. \tag{5}$$

Large-sample results such as consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$ estimated based on the partial likelihood are derived in Andersen & Gill (1982); furthermore, it is worth mentioning that the partial likelihood is a special case of a broader class of composite likelihoods (Varin & Vidoni, 2005). Besides avoiding problems resulting from the misspecification of $\alpha_0(t)$, this partial likelihood approach only requires the availability of time-dependent statistics at event times. When the number of nodes whose statistics are updated between two event times is small, as is the case in our egocentric citation network model, we can exploit this fact computationally by updating the denominator of Eq (5) only when needed, from one event to the next.

In the above discussion, we have assumed that at most one event happens at any time, i.e., only one paper is cited at a specific time. However, in reality, citations occur in groups: when joining the network, a paper simultaneously makes citations to many papers. In addition, the publication times of multiple papers may coincide, either because they are published simultaneously or because the observation times are rounded to some large unit. As long as the ratio of the number of simultaneously cited papers to the total number of papers currently in the network is small, we can deal with ties using the Breslow approximation (Klein & Moeschberger, 2003, section 8.4),

$$L(\boldsymbol{\beta}) = \prod_{e=1}^{m} \frac{\prod_{i_e \in C_e} \exp\left(\boldsymbol{\beta}^\top \mathbf{s}_{i_e}(t_e)\right)}{\left[\sum_{i=1}^{n} Y_i(t_e) \exp\left(\boldsymbol{\beta}^\top \mathbf{s}_i(t_e)\right)\right]^{|C_e|}}, \quad (6)$$

where $C_e$ is the set of papers cited at time $t_e$.

### 3.2. Computational issues

A Newton-Raphson algorithm is used to find $\hat{\boldsymbol{\beta}}$ by maximizing the logarithm of the Breslow approximation (6) and to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$ as the inverse of the Hessian matrix of the last iteration. The algorithm implements a simple step-halving procedure, halving the length of the step if necessary until $\log L(\boldsymbol{\beta})$ increases. The iterations continue until every element in $\nabla \log L(\boldsymbol{\beta})$ is smaller that $10^{-3}$ in absolute value, or until the increase in $\log L(\boldsymbol{\beta})$ is less than $10^{-100}$, whichever happens first.

The computational complexity for naively evaluating $\log L(\boldsymbol{\beta})$ in Eq (6), along with its gradient vector and Hessian matrix, is $O(p^2mn)$, where $p$ is the dimension of $\boldsymbol{\beta}$, $m$ is the number of distinct citation event times in the training set, and $n$ is the number of nodes in the training set. If we include the solving of a $p$-dimensional linear system as required for a Newton-Raphson iteration, each iteration requires $O(p^2mn + p^3)$ computations. However, for our purposes, the value of $p$ is much smaller than $m$ or $n$, so the $p^3$ term is negligible relative to $p^2mn$.

We present two ways to make inference more efficient. First, we target the factor $n$ by using a caching data structure to exploit the sparsity of nodes that are updated between two event times. Second, we apply the statistical theory of recurrent event models (Andersen et al., 1993) aimed at reducing the factor $m$ while maintaining the accuracy of learned parameters $\boldsymbol{\beta}$.

A straightforward calculation of the partial likelihood can be summarized as follows. First, a data structure of the initial network is initialized, as well as a matrix where each row $i$ is a vector of statistics for node $i$. Then for each new citation event $e$, the following steps are performed:

1. The current statistics matrix is used to calculate two terms, $\boldsymbol{\beta}^\top \mathbf{s}_{i_e}(t_e)$ and $\kappa(e)$, where
$$\kappa(e) = \sum_{i=1}^{n} Y_i(t_e) \exp\left\{\boldsymbol{\beta}^\top \mathbf{s}_i(t_e)\right\}. \quad (7)$$
2. Edge $e$ is added to the network data structure.
3. The statistics matrix is updated based on the added $e$.

To obtain the maximum likelihood estimator of $\boldsymbol{\beta}$, the above procedure will be repeated many times for different values of $\boldsymbol{\beta}$. Since steps (2) and (3) do not depend on $\boldsymbol{\beta}$, one computational improvement that can be made is to cache the time series of network updates, as follows. We step through all event times from the beginning. As each edge is added to the network, we figure out which nodes and their corresponding statistics are updated and then cache all of these updates, i.e., store those elements of the statistics matrix to be updated as well as the list of updated nodes. In our experiments on models that only include statistics based on network structure, the number of nodes affected at an event time is very small. Therefore, despite using more memory, caching saves a significant amount of computational time spent on repeating steps (2) and (3) for each new value of $\boldsymbol{\beta}$ by simply modifying the matrix of nodal statistics using the cached updates.

Specifically, the caching method works as follows. The terms of the sum $\kappa(e)$ of Eq (7) are only different from those of $\kappa(e-1)$ for those nodes whose nodal statistics are updated or who join the network within interval $[t_{e-1}, t_e)$. Letting $U_{e-1}$ be the set of nodes whose nodal statistics are updated during $[t_{e-1}, t_e)$ and $C_{e-1}$ be the set of nodes who join the network during $[t_{e-1}, t_e)$, we have

$$\begin{aligned}\kappa(e) &= \kappa(e-1) + \sum_{i \in C_{e-1}} \exp\left\{\boldsymbol{\beta}^\top \mathbf{s}_i(t_e)\right\} \\ &\quad + \sum_{i \in U_{e-1}} \left[\exp\left\{\boldsymbol{\beta}^\top \mathbf{s}_i(t_e)\right\} - \exp\left\{\boldsymbol{\beta}^\top \mathbf{s}_i(t_{e-1})\right\}\right].\end{aligned}$$

Each summation above involves a small number of terms relative to $n$. We cache values such as $\mathbf{s}_i(t_e)$ for $i \in C_{e-1}$ and $i \in U_{e-1}$ in our initial first pass through the whole data set so that these summations may be calculated quickly, and the resulting $\kappa(e)$ summed, for arbitrary values of $\boldsymbol{\beta}$.

The caching scheme above works well when the statistics $\mathbf{s}_i(t)$ are defined so as to be essentially local, in the sense that the appearance of a new paper in the network, with its list of network-edge-generating citations, will only affect the values of $\mathbf{s}_i(t)$ for those papers $i$ that are actually cited. However, as discussed in Section 2.2, we analyze citation networks using two separate collections of $\mathbf{s}_i(t)$ statistics, and only one of these collections can be said to possess the locality property. In the other case involving LDA-based

*Table 1.* Characteristics of citation data sets.

| | 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 |

matching statistics between pairs of papers, the values of the $s_i(t_e)$ statistics change for all papers $i$ currently in the network at the time of event $e$. Thus, caching will not be effective for the LDA statistics.

Instead, for the LDA matching statistics we will rely on the concept of non-informative censoring in survival and event history analysis (Andersen et al., 1993) to attack the factor $m$ in $O(mn)$. Assuming that the training data are defined as the network history from time 0 until time $\tau_r$, we reason as follows. Rather than using all event times in $[0, \tau_r]$ to construct the partial likelihood for estimation, we can perform inference on a training window $[\tau_l, \tau_r]$ with a small number of distinct event times without sacrificing too much statistical efficiency. The main condition to ensure this feature is that $\tau_l$ is determined independently of the event process, which certainly holds for our examples since selection of $\tau_l$ is independent of the citation process. Moreover, all observations in $[0, \tau_l)$ are still used to construct the network history $\mathbf{H}_{t^-}$ for a given $t$; the only change to the approximate partial likelihood in Eq (6) is that the limits of the product shall be from some $m_0 > 1$ to $m$, rather than 1 to $m$. We verify this strategy empirically in Section 4.3.

## 4. Experimental Analysis

We apply the egocentric framework to citation networks and analyze the predictive power and interpretability of our approach. The following sections detail the data sets, the models learned, and the experimental setup and results.

### 4.1. Data sets

Citation networks are the main focus of this paper since they are of interest to the machine learning community and are well suited to egocentric modeling. We use data from the American Physical Society (APS)[1] and arXiv[2]. Table 1 summarizes the characteristics of these data sets.

The APS data contains the citation network for articles appearing in Physical Review Letters, Physical Review, and Reviews of Modern Physics from 1893 through 2009. Most of these timestamps for the articles are grouped into months while recent papers have day-resolution; thus, the ratio between unique timestamps and papers is low.

[1] https://publish.aps.org/datasets
[2] http://snap.stanford.edu/data

The arXiv-PH data contains the citation network of arXiv high energy physics phenomenology articles spanning from January 1993 to March 2002. Timestamps are available on a daily scale. Meanwhile, arXiv-TH is a high energy physics theory data set of articles spanning from January 1993 to April 2003. Timestamps are recorded on a continuous-time scale (millisecond resolution). There are 25,004 citation event times corresponding to the number of papers that make citations when they join the network. Besides temporal network information, arXiv-TH also has paper abstracts which will be used to illustrate how paper metadata can be integrated into the egocentric model.

### 4.2. Specific models

We consider the following egocentric models:

1. A baseline preferential attachment (**PA**) egocentric model with only first-order preferential attachment statistic $s_1(t)$. Under the rank metric, this baseline is equivalent to a nonparametric growth model based on the PA mechanism (Barabasi & Albert, 1999) where papers are ranked based on the current number of citations that they have received so far.
2. A **P2PT** model that includes $s_1, s_2, s_4, \ldots, s_8$.
3. A **P2PTR180** model that includes all the statistics $s_1, \ldots, s_8$. The difference between P2PT and P2PTR180 is the inclusion of the recency first-order PA statistic $s_3(t)$ with a window of 180 days.

Since abstracts are available for arXiv-TH, the following models will also be considered:

4. An **LDA** egocentric model where LDA-based matching topic proportion statistics $\mathbf{s}_j^{\mathrm{LDA}}(t_i^{\mathrm{arr}})$ are used.
5. An **LDA+P2PTR180** egocentric model where all network statistics and LDA statistics are considered.

### 4.3. Experiments

We evaluate the predictive power of these models through rolling prediction experiments. We split each data set chronologically into three phases: a statistics-building phase, a training phase, and a test phase. The statistics-building phase is used to construct the network history and to build up the network statistics. The training phase is used to construct the partial likelihood and estimate the

*Table 2.* Number of unique citation event times in the statistics-building, training, and test phases.

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

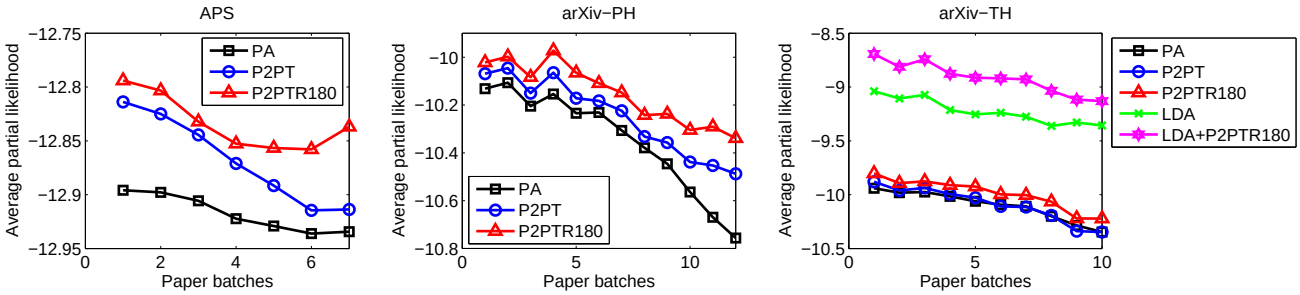



*Figure 2.* Likelihoods of test paper citations (shown as averages over paper batches) for each data set.

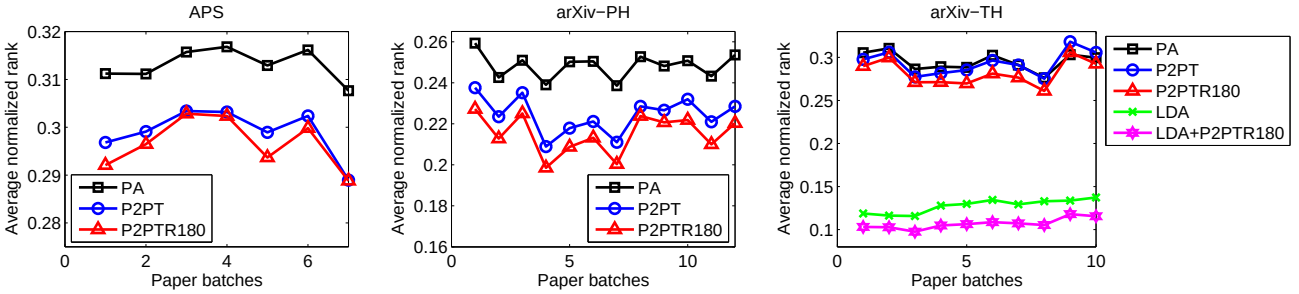



*Figure 3.* Ranks of test paper citations (shown as averages over paper batches) for each data set.

model coefficients. The test phase is used to evaluate the predictive capability of the learned model. Note that statistics-building is ongoing even through the training and test phases. The phases are split along citation event times. The sizes of each phase are presented in the Table 2.

The statistics-building phase is relatively long to mitigate truncation effects at the beginning of network formation as well as the effect of severely grouped event times, which biases parameter estimates. However, these training and test windows still cover a substantial period of time (e.g. for APS, the last 200 unique times covers 2.5 years). Note that performance is relatively invariant to the size of the training windows. Using windows of size 2000 and 5000 for arXiv-TH, we achieved essentially the same results.

The evaluation metrics that we consider are held-out partial likelihoods and held-out normalized ranks. In the same fashion as described in Section 3.1, a held-out partial (log)likelihood is computed for each paper in the test set by taking the average of the partial likelihoods for each citation event. We compute a "rank" for each citation event by sorting the likelihoods of each possible citation in decreasing order and determining the position of each true citation in that sorted list. We normalize this rank by dividing by the number of possible citations, and the paper's rank is the average of the normalized ranks of each observed citation. A lower rank indicates better predictive performance.

In Figure 2, the held-out likelihoods of the PA, P2PT, and P2PTR180 models are shown. To avoid clutter, we show the average held-out partial likelihoods of batches of papers in the test phase, chronologically ordered. The batch sizes are 3000 for APS, 500 for PH, and 500 for TH. These results show that the P2PT and P2PTR180 egocentric models generally outperform the PA baseline. Furthermore, for arXiv-TH, we include the LDA and LDA+P2PTR180 models and find that adding the LDA statistics into the egocentric framework significantly boosts performance.

Figure 3 shows the held-out ranks. As with the likelihood plots, we report the average rank over batches of papers in the test phase. Note that random guessing yields a normalized rank of 0.5, and so all of the models are performing substantially better than random. Moreover, these ranks demonstrate that the egocentric models with more network statistics typically outperform the PA baseline. The ranks also confirm that adding LDA statistics increases predictive performance. In Figure 3, LDA+P2PTR180 gives a 16% relative rank improvement over LDA alone, suggesting that a mixture of network and nodal statistics is helpful. A nonparametric paired Wilcoxon test on the ranks of each test paper obtained by LDA and LDA+P2PTR180 yields a p-value of $6 \times 10^{-12}$, suggesting that the difference in ranks is statistically significant.

Figure 4 shows the recall as a function of cut-point, for arXiv-TH. Recall (accumulated over all test events) is defined as the percentage of the true citations that are found in the sorted likelihood list from positions 1 to $K$, where $K$ is the cut-point. As with the held-out likelihoods and ranks,

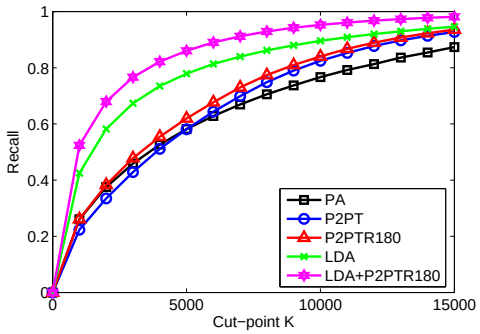



*Figure 4.* Recall of top-$K$ recommendation list on arXiv-TH.

we find that P2PT and P2PTR180 outperform the baseline PA, and the inclusion of the LDA statistics improves recall.

Another strength of the egocentric model is its interpretability. We interpret the coefficients of the LDA+P2PTR180 model that are estimated on the arXiv-TH data (see Table 3). Note that we only interpret the sign of each coefficient in the presence of all other effects. For example, the positive value of the $\beta_1$ coefficient indicates the following: for two papers which have all other identical statistics, the one with a higher preferential attachment statistic is more likely to be cited in the future. In addition, the recency PA coefficient $\beta_3$ is also positive, which suggests that a paper with a higher number of citations within 6 months is more likely to be cited than a similar paper with less recent citations. In other words, there is evidence of a recency citation effect, or a temporary elevation of citation intensity following a number of recent citations.

We also found an interesting citation pattern related to network statistics on triangles. The negative value for the $\beta_4$ coefficient suggests that, for two papers $A$ and $D$ in Figure 5(a) (with all other statistics identical), if paper $A$ has a higher seller statistic, it is less likely to be cited in the future. Intuitively, this makes sense, since the lack of this triangle allows $A$ to have more diverse citation pathways in the future. Similarly, the buyer coefficient $\beta_6$ is also negative, suggesting that for two papers $C$ and $E$ in Figure 5(b), the paper $C$ with a higher buyer statistic is less likely to be cited. In other words, there is evidence of a diversity effect; a paper with diverse citing and cited patterns is more likely to be cited in the future.

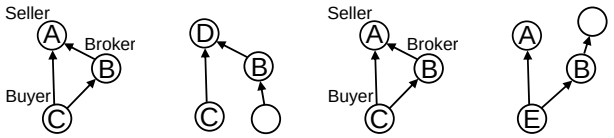
## 5. Related Work

The analysis of citation networks has a lengthy history within the bibliometric community, finding such analysis to be useful in uncovering historical scientific trends (Price, 1965), discovering author interactions (Börner et al., 2004), and determining the impact factors of journals (Garfield, 1972). Within the machine learning community, there has been a focus towards automated citation recommendation. For example, Strohman et al. (2007) and He et al. (2011) develop approaches for recommending paper citations using a variety of engineered features. Probabilistic models that incorporate textual content and other metadata have also been investigated (Dietz et al., 2007; Nallapati et al., 2008). One such model is the Relational Topic Model (Chang & Blei, 2009), which incorporates text by defining the probability of a link between two nodes to be proportional to a similarity score between their topic proportions. A key difference between these methods and our approach is that these methods generally do not consider the temporal aspect of the network. Moreover, while our paper has not primarily focused on crafting specific features for citation modeling, our egocentric approach can incorporate such features within a statistical framework.

More generally, there has been increasing interest in modeling longitudinal network data, as surveyed by Goldenberg et al. (2009). Dynamic exponential random graph models have been used to model snapshot data (Hanneke & Xing, 2006; Wyatt et al., 2010). There also exist dynamic models which operate in latent space in order to group similar nodes together (Sarkar & Moore, 2005; Fu et al., 2009). These models differ from our egocentric approach in that they operate on snapshot data and make Markovian assumptions, while our approach operates on fine-grained event data and conditions on the en-

*Table 3.* Estimated coefficients for network statistics in the LDA+P2PTR180 model. All of these estimates are statistically significant at the level $\alpha = 0.0001$.

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




(a) Diverse Seller Effect (b) Diverse Buyer Effect

*Figure 5.* Preferential citation patterns on triangles. In 5(a), paper $D$ with a diverse cited pattern is more likely to attract future citations than paper $A$. In 5(b), paper $E$ with a diverse citing pattern is more likely to be cited than paper $C$.

tire network history. Meanwhile, continuous-time Markov processes have been applied to the analysis of longitudinal network data (Snijders, 2005). However, these studies have typically focused on very small networks; moreover, they also make Markovian assumptions. Huang & Lin (2009) treat each edge between two nodes as a time series and uses an autoregressive integrated moving average model for link prediction; however, this approach is not suitable for citation networks since a citation link only appears once and remains fixed once it appears.

The work closest to our own is that of Butts (2008), who develops a relational event framework based on survival analysis theory. We can view our approach as extending this work to large-scale longitudinal network modeling, where we develop a general egocentric model that is applicable to citation networks.

## 6. Discussion and Conclusion

We have introduced a statistical egocentric framework for modeling longitudinal network data and have developed an efficient inference approach for such models. Empirical analysis on real-world citation network data suggests that the egocentric approach has utility in terms of both prediction and interpretability.

Though our approach only exploits the ordering of events, one may explicitly consider the timing of events by estimating the baseline hazard $\alpha_0(t)$ of Eq (2). For instance, after computing $\hat{\boldsymbol{\beta}}$ for LDA+P2PTR180 on arXiv-TH, we were able to estimate the cumulative baseline hazard $A_0(\tau_l, \tau_r) = \int_{\tau_l}^{\tau_r} \alpha_0(s)ds$ via the so-called Breslow estimator (Aalen et al., 2008, section 4.1.2) and found that it had roughly constant slope. This constant slope (despite no assumption of a parametric form) is consistent with a constant baseline hazard, suggesting that a parametric approach might avoid the disadvantages mentioned in Section 3.1. Thus, one may use Eq (2) in conjunction with the estimated $\alpha_0(t)$ to construct an intensity process for each individual node, if time-specific predictions are desired.

There are many additional avenues for future work. In the space of citation networks, one can incorporate advanced statistics based on authors, journals, and other metadata; furthermore, the joint learning of LDA and network model parameters can potentially yield improved performance. While we have focused on the egocentric perspective, counting processes may be placed on graph edges as well; thus, our framework can be generalized to other types of networks, such as biological and social networks.

### Acknowledgments

This work is supported by ONR under the MURI program, Award Number N00014-08-1-1015.

## References

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

Andersen, P.K. and Gill, R.D. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. *Statistical Models Based on Counting Processes*. Springer, 1993.

Barabasi, A.L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

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

Börner, K., Maru, J.T., and Goldstone, R.L. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5266, 2004.

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.

Chang, J. and Blei, D.M. Relational topic models for document networks. In *AI and Statistics*, pp. 81–88, 2009.

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

Dietz, L., Bickel, S., and Scheffer, T. Unsupervised prediction of citation influences. In *International Conference on Machine Learning*, pp. 233–240. ACM, 2007.

Fu, W., Song, L., and Xing, E. P. Dynamic mixed membership blockmodel for evolving networks. In *International Conference on Machine Learning*, pp. 329–336. ACM, 2009.

Garfield, E. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.

Goldenberg, A., Zheng, A.X., Fienberg, S.E., and Airoldi, E.M. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2009.

Hanneke, S. and Xing, E. P. Discrete temporal models of social networks. In *Conference on Statistical Network Analysis*, 2006.

He, Q., Kifer, D., Pei, J., Mitra, P., and Giles, L. Citation recommendation without author supervision. In *International Conference on Web Search and Data Mining*, pp. 755–764, 2011.

Huang, Z. and Lin, D.K.J. The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2):286–303, 2009.

Klein, J.P. and Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2003.

Nallapati, R.M., Ahmed, A., Xing, E.P., and Cohen, W.W. Joint latent topic models for text and citations. In *SIGKDD Conference*, pp. 542–550. ACM, 2008.

Price, D. J. Networks of scientific papers. *Science*, 149:510, 1965.

Sarkar, P. and Moore, A. Dynamic social network analysis using latent space models. *SIGKDD Explorations*, 7(2):31–40, 2005.

Snijders, T.A.B. Models for longitudinal network data. *Models and Methods in Social Network Analysis*, pp. 215–247, 2005.

Strohman, T., Croft, W. B., and Jensen, D. Recommending citations for academic papers. In *SIGIR*, pp. 705–706, 2007.

Varin, C. and Vidoni, P. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519, 2005.

Wasserman, S. Analyzing social networks as stochastic processes. *Journal of the Amer. Stat. Association*, 75(370):280–294, 1980.

Wyatt, D., Choudhury, T., and Bilmes, J. Discovering long range properties of social networks with multi-valued time-inhomogeneous models. In *AAAI Conference*, 2010.