



Point process modelling for directed interaction networks

Patrick O. Perry

New York University, USA

and Patrick J. Wolfe

University College London, UK

[Received November 2010. Revised November 2012]

Summary. Network data often take the form of repeated interactions between senders and receivers tabulated over time. A primary question to ask of such data is which traits and behaviours are predictive of interaction. To answer this question, a model is introduced for treating directed interactions as a multivariate point process: a Cox multiplicative intensity model using covariates that depend on the history of the process. Consistency and asymptotic normality are proved for the resulting partial-likelihood-based estimators under suitable regularity conditions, and an efficient fitting procedure is described. Multicast interactions—those involving a single sender but multiple receivers—are treated explicitly. The resulting inferential framework is then employed to model message sending behaviour in a corporate e-mail network. The analysis gives a precise quantification of which static shared traits and dynamic network effects are predictive of message recipient selection.

Keywords: Cox proportional hazards model; Network data analysis; Partial likelihood inference; Point processes

1. Introduction

Much effort has been devoted to the statistical analysis of network data; see Jackson (2008), Goldenberg *et al.* (2009) and Kolaczyk (2009) for recent overviews. Often network observables comprise counts of interactions between individuals or groups tabulated over time. Communications networks give rise to *directed* interactions: phone calls, text messages or e-mails exchanged between a given set of individuals over a specific time period (Tyler *et al.*, 2005; Eagle and Pentland, 2006). Specific examples of repeated interactions from other types of networks include the following: Fowler's (2006) study of legislators authoring and cosponsoring bills (a collaboration network); McKenzie and Rapoport's (2007) study of families migrating between communities in Mexico (a migration network); the study of Sundaresan *et al.* (2007) of zebras congregating at locations in their habitat (an animal association network); Papachristos's (2009) study of gangs in Chicago murdering members of rival factions (an organized crime network).

In this paper, we consider partial-likelihood-based inference for general directed interaction data in the presence of covariates. We first develop asymptotic theory for the case in which interactions are strictly pairwise, and then we generalize our results to the multiple-receiver (multicast) case; we also provide efficient algorithms for partial likelihood maximization in

Address for correspondence: Patrick O. Perry, Information, Operations and Management Sciences Department, Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012, USA.
Email: pperry@stern.nyu.edu

these settings. Our main assumption on the covariates is that they be predictable, which allows them to vary with time and potentially to depend on the past.

The interaction data that we consider comprise a set of triples, with triple (t, i, j) indicating that, at time t , directed interaction $i \rightarrow j$ took place—for instance, individual i sent a message to individual j . Given such a set of triples, a primary modelling goal lies in determining which characteristics and behaviours of the senders and receivers are predictive of interaction. In this vein, three important questions stand out.

- (a) *Homophily*: is there evidence of homophily (an increased rate of interaction between similar individuals)? To what degree is a shared attribute predictive of heightened interaction?
- (b) *Network effects*: to what extent are past interaction behaviours predictive of future ones? If we observe interactions $i \rightarrow h$ and $h \rightarrow j$, are we more likely to see the interaction $i \rightarrow j$?
- (c) *Multiplicity*: how should multiple-receiver interactions of the type $i \rightarrow \{j_1, j_2, \dots, j_L\}$ be modelled? What are the implications of treating these as L separate pairwise interactions?

The issues of homophily, network effects and their interactions arise frequently in the networks literature; see, for example, McPherson *et al.* (2001), Butts (2008), Aral *et al.* (2009), Snijders *et al.* (2010) and references contained therein. Multiplicity has largely been ignored in this context, however, with notable exceptions including Lunagómez *et al.* (2009) for graphical models, and Shafiei and Chipman (2010) for network modelling.

In the remainder of this paper, we provide a modelling framework and computationally efficient partial likelihood inference procedures to facilitate analysis of these questions. We employ a Cox proportional intensity model incorporating both static and history-dependent covariates to address the first of these two questions, and a parametric bootstrap to address the third. Section 2 presents our point process model for directed pairwise interactions, along with the resultant inference procedures. Section 3 establishes consistency and asymptotic normality of the corresponding maximum partial likelihood estimator, and Section 4 extends our framework to the case of multiple-receiver interactions. Section 5 employs this framework to model message sending behaviour in a corporate e-mail network. Section 6 evaluates the strength of homophily and network effects in explaining these data, and Section 7 concludes the main body of the paper. Appendices A–C contain respectively implementation details and technical results from Sections 3 and 4. The on-line supplementary material provides comparative analyses based on related network models in the literature.

2. A point process model and partial likelihood inference

Every interaction process can be encoded by a multivariate counting measure. For sender i , receiver j and positive time t , define

$$N_t(i, j) = \#\{\text{directed interactions } i \rightarrow j \text{ in time interval } [0, t]\}.$$

For technical reasons, assume that $N_0(i, j) = 0$ and that $N_t(i, j)$ is adapted to a stochastic basis of σ -algebras $\{\mathcal{F}_t\}_{t \geq 0}$ satisfying the usual conditions. Then, $N_t(i, j)$ is a local submartingale, so, by the Doob–Meyer decomposition, there is a predictable increasing process $\Lambda_t(i, j)$, null at zero, such that $N_t(i, j) - \Lambda_t(i, j)$ is an \mathcal{F}_t -local martingale. Under mild conditions—the most important of which is that no two interactions happen simultaneously—there is a predictable continuous process $\lambda_t(i, j)$ such that $\Lambda_t(i, j) = \int_0^t \lambda_s(i, j) ds$. (In practical applications, simultaneous events exist and are an annoyance; Efron (1977) handled simultaneity through an *ad hoc* adjustment, whereas Broström (2002) added a discrete component to Λ .) The process λ is known

as the stochastic intensity of N . Heuristically,

$$\lambda_r(i, j) dt = \mathbb{P}\{\text{interaction } i \rightarrow j \text{ occurs in time interval } [t, t + dt]\}.$$

We shall model N through λ by using a version of the Cox (1972) proportional intensity model.

Let \mathcal{I} be a set of senders and \mathcal{J} be a (not necessarily disjoint) set of receivers. For each sender i , let $\bar{\lambda}_r(i)$ be a non-negative predictable process called the baseline intensity of sender i ; let $\mathcal{J}_r(i)$ be a predictable finite subset of \mathcal{J} called the receiver set of sender i . For each sender–receiver pair (i, j) , let $x_r(i, j)$ be a predictable locally bounded vector of covariates in \mathbb{R}^p . Let β_0 be an unknown vector of coefficients in \mathbb{R}^p . For the remainder of this section, assume that each interaction has a single receiver.

Given a multivariate counting process N on $\mathbb{R}_+ \times \mathcal{I} \times \mathcal{J}$, we model its stochastic intensity as

$$\lambda_r(i, j) = \bar{\lambda}_r(i) \exp\{\beta_0^\top x_r(i, j)\} \mathbf{1}\{j \in \mathcal{J}_r(i)\}. \tag{1}$$

This model posits that sender i in \mathcal{I} interacts with receiver j in $\mathcal{J}_r(i)$ at a baseline rate $\bar{\lambda}_r(i)$ modulated up or down according to the pair’s covariate vector, $x_r(i, j)$. As Efron (1977) noted, the specific parametric form for the multiplier $\exp\{\beta_0^\top x_r(i, j)\}$ is not central to the theoretical analysis, but this choice is amenable to computation and gives the parameter vector β_0 a straightforward interpretation. Butts (2008) and Vu *et al.* (2011a, b) used variants of this model to analyse repeated directed actions within social settings.

The form of model (1) is deceptively simple but remains sufficiently flexible to be useful in practice. The model allows for homophily and group level effects via inclusion of covariates of the form ‘ $\mathbf{1}\{i \text{ and } j \text{ belong to the same group}\}$ ’, where ‘group’ is some observable trait like ethnicity, gender or age group. Its real strength, though, is that $x_r(i, j)$ is allowed to be *any* predictable process; in particular $x_r(i, j)$ can depend on the history of interactions. To model reciprocation and transitivity in the interactions (with $\mathcal{I} = \mathcal{J}$), for example, choose appropriate values for Δ_k and include relevant covariates in $x_r(i, j)$:

$$\mathbf{1}\{\text{interaction } j \rightarrow i \text{ occurred in } [t - \Delta_k, t)\}$$

and

$$\mathbf{1}\{\text{for some } h, \text{ interactions } i \rightarrow h \text{ and } h \rightarrow j \text{ occurred in } [t - \Delta_k, t)\}.$$

Any process that is measurable with respect to the predictable σ -algebra is a valid covariate; this excludes only covariates depending on the future or the immediate present. In Section 5.2 we detail specific covariates that are suitable for measuring homophily and network effects.

Also note that, despite presuming \mathcal{I} and \mathcal{J} to be fixed, our analysis allows senders and receivers to enter and leave the study during the observation period. The effective number of senders at time t is the set of i such that $\bar{\lambda}_r(i) \neq 0$, which potentially varies with time. Likewise, the effective number of receivers is controlled through $\mathcal{J}_r(i)$.

Following Cox (1975), we treat the baseline rate $\bar{\lambda}_r(i)$ as a nuisance parameter and estimate the coefficient vector β_0 by using a partial likelihood. Specifically, let $(t_1, i_1, j_1), \dots, (t_n, i_n, j_n)$ be the sequence of observed interactions. The inference procedure is motivated by decomposing the full likelihood L as

$$\begin{aligned} L(t_1, i_1, j_1, t_2, i_2, j_2, \dots, t_n, i_n, j_n) &= L(t_1, i_1) L(j_1 | t_1, i_1) L(t_2, i_2 | t_1, i_1, j_1) L(j_2 | t_2, i_2, t_1, i_1, j_1) \\ &\quad \dots L(t_n, i_n | t_{n-1}, i_{n-1}, j_{n-1}, \dots, t_1, i_1, j_1) L(j_n | t_n, i_n, t_{n-1}, i_{n-1}, \dots, t_1, i_1, j_1) \\ &= \{L(t_1, i_1) L(t_2, i_2 | t_1, i_1, j_1) \dots L(t_n, i_n | t_{n-1}, i_{n-1}, j_{n-1}, \dots, t_1, i_1, j_1)\} \\ &\quad \times \{L(j_1 | t_1, i_1) L(j_2 | t_2, i_2, t_1, i_1, j_1) \dots L(j_n | t_n, i_n, t_{n-1}, i_{n-1}, \dots, t_1, i_1, j_1)\}; \end{aligned}$$

the term comprised of the product of conditional likelihoods of j_1, \dots, j_n is dubbed a partial likelihood. In continuous time, the log-partial-likelihood at time t , evaluated at β , is

$$\log\{\text{PL}_t(\beta)\} = \sum_{t_m \leq t} \left(\beta^T x_{t_m}(i_m, j_m) - \log \left[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta^T x_{t_m}(i_m, j)\} \right] \right). \tag{2}$$

In Section 3, we prove under suitable regularity conditions that the maximizer of $\log\{\text{PL}_t(\cdot)\}$ is a consistent estimator of β_0 as t increases.

The function $\log\{\text{PL}_t(\cdot)\}$ is concave and so can be maximized via Newton’s method or a gradient-based optimization approach (Nocedal and Wright, 2006). These methods require one or both of the first two derivatives of $\log\{\text{PL}_t(\cdot)\}$, which can be expressed in terms of weighted means and covariances of the covariates. The weights are

$$w_t(\beta, i, j) = \exp\{\beta^T x_t(i, j)\} \mathbf{1}\{j \in \mathcal{J}_t(i)\}, \tag{3a}$$

$$W_t(\beta, i) = \sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j). \tag{3b}$$

The inner sum in $\log\{\text{PL}_t(\beta)\}$ is $W_{t_m}(\beta, i_m)$. The function $\log\{W_t(\cdot, i)\}$ has gradient $E_t(\cdot, i)$ and Hessian $V_t(\cdot, i)$, given by

$$E_t(\beta, i) = \frac{1}{W_t(\beta, i)} \sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j) x_t(i, j), \tag{4a}$$

$$V_t(\beta, i) = \frac{1}{W_t(\beta, i)} \sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j) (x_t(i, j) - E_t(\beta, i))^{\otimes 2}, \tag{4b}$$

where $a^{\otimes 2} = a \otimes a = aa^T$. Consequently, the gradient and negative Hessian of $\log\{\text{PL}_t(\cdot)\}$ are

$$U_t(\beta) = \nabla[\log\{\text{PL}_t(\beta)\}] = \sum_{t_m \leq t} x_{t_m}(i_m, j_m) - E_{t_m}(\beta, i_m), \tag{5a}$$

$$I_t(\beta) = -\nabla^2[\log\{\text{PL}_t(\beta)\}] = \sum_{t_m \leq t} V_{t_m}(\beta, i_m). \tag{5b}$$

We call $U_t(\beta_0)$ the unnormalized score and $I_t(\beta_0)$ the observed information matrix.

Note the dependence of these terms on time varying covariates, which precludes the use of sufficient statistics and introduces additional complexity in maximizing $\log\{\text{PL}_t(\cdot)\}$. For most large interaction data sets, existing computational routines for handling Cox models (e.g. the function `coxph` from the `survival` package for R (Therneau and Lumley, 2009)) will not suffice. In Appendix A, we describe a customized method for maximizing $\log\{\text{PL}_t(\cdot)\}$ that exploits sparsity in $x_t(i, j)$.

3. Consistency of maximum partial likelihood inference

Under the model of Section 2, the maximum partial likelihood estimator (MPLE) is a natural estimate of β_0 ; the inverse Hessian of $\log\{\text{PL}_t(\cdot)\}$ evaluated at the MPLE is a natural estimate of its covariance matrix. We now give conditions under which these estimators are consistent.

In the sampling regime where observation time t is fixed and the number of senders $|\mathcal{I}|$ increases, Andersen and Gill’s (1982) consistency proof for the Cox proportional hazards model in survival analysis extends to cover model (1). This setting is natural in the context of clinical trial data, where \mathcal{I} corresponds to the set of patients under study, but does not meet the requirements that are typical of interaction data. For most interaction data we cannot control \mathcal{I} and \mathcal{J} , and

the only way to collect more data is to increase the observation time. Cox (1972, 1975) outlined a proof for general MPLE consistency that applies to our sampling regime, but his argument is heuristic; Wong’s (1986) treatment is more rigorous but does not cover continuous or time varying covariates. The general interaction data sampling regime warrants a new consistency proof.

Our proof of consistency relies on rescaling time to make the interaction times uniform. For this, define marginal processes $N_t(i) = \sum_{j \in \mathcal{J}} N_t(i, j)$ and $N_t = \sum_{i \in \mathcal{I}} N_t(i)$; also note that $t_n = \sup\{t : N_t < n\}$ is a stopping time and let \mathcal{F}_{t_n} be the σ -algebra of events before t_n . The main idea of the proof is to change time from the original scale to a scale on which $t_n - t_{n'}$ is proportional to $n - n'$.

3.1. Assumptions

Let \mathcal{B} be a neighbourhood of β_0 . For a vector a , let $\|a\|$ denote its Euclidean norm; for a matrix, A , let $\|A\|$ denote its spectral norm, equal to the largest eigenvalue of $(A^T A)^{1/2}$. We require the following assumptions.

Assumption 1. The covariates are uniformly square integrable, i.e.

$$\mathbb{E}[\sup_{t,i,j} \|x_t(i, j)\|^2]$$

is bounded.

Assumption 2. The integrated covariance function is well behaved. When $\beta \in \mathcal{B}$ and $\alpha \in [0, 1]$, as $n \rightarrow \infty$, then with respect to the covariance function $\Sigma_\alpha(\beta)$ we have that

$$\frac{1}{n} \sum_{i \in \mathcal{I}} \int_0^{t_{[n\alpha]}} V_s(\beta, i) W_s(\beta, i) \bar{\lambda}_s(i) ds \xrightarrow{\mathbb{P}} \Sigma_\alpha(\beta).$$

Assumption 3. The interaction arrival times are finite. For each n ,

$$\mathbb{P}\{t_n < \infty\} = 1.$$

Assumption 4. The variance function is equicontinuous. More precisely,

$$\{V_{t_n}(\cdot, i) : n \geq 1, i \in \mathcal{I}\}$$

is an equicontinuous family of functions.

These technical assumptions are similar to those of Andersen and Gill (1982), who investigated specific settings in which their assumptions hold. Note that, when $\|x_t(i, j)\|$ is bounded and assumption 3 is in force, the remaining assumptions follow.

3.2. Main results

Assumptions 1–4 imply that the MPLE is consistent and asymptotically Gaussian, as shown by the following two theorems.

Theorem 1. Let N be a multivariate counting process with stochastic intensity as given in equation (1), with true parameter vector β_0 . Let t_n be the sequence of interaction times, and set $U_t(\beta)$ and $I_t(\beta)$ to be the gradient and negative Hessian of the log-partial-likelihood function as given respectively in equations (5a) and (5b). If assumptions 1 and 2 hold, then, as $n \rightarrow \infty$,

- (a) $n^{-1/2} U_{t_{[n\alpha]}}(\beta_0)$ converges weakly to a Gaussian process on $[0, 1]$ with covariance function $\Sigma_\alpha(\beta_0)$ and
- (b) if assumptions 3 and 4 also hold, then, for any consistent estimator $\hat{\beta}_n$ of β_0 , we have that

$$\sup_{\alpha \in [0,1]} \left\| \frac{1}{n} I_{t_{[an]}}(\hat{\beta}_n) - \Sigma_\alpha(\beta_0) \right\| \xrightarrow{P} 0.$$

We do not actually require convergence of the whole sample path, but it turns out to be just as much effort to prove as convergence of the end point. Consistency is a direct consequence of theorem 1.

Theorem 2. Let N be a multivariate counting process with stochastic intensity as given in equation (1), with true parameter vector β_0 . Let the log-partial-likelihood $\log\{\text{PL}_t(\cdot)\}$ be as defined in equation (2). Let t_n be the sequence of interaction times.

Assume for β in a neighbourhood of β_0 that $-(1/n)\nabla^2[\log\{\text{PL}_{t_n}(\beta)\}] \rightarrow^P \Sigma_1(\beta)$, where $\Sigma_1(\cdot)$ is locally Lipschitz and with smallest eigenvalue bounded away from zero. If $\hat{\beta}_n$ maximizes $\log\{\text{PL}_{t_n}(\cdot)\}$ and conclusion (a) of theorem 1 holds, then the following assumptions are true as $n \rightarrow \infty$:

- (a) $\hat{\beta}_n$ is a consistent estimator of β_0 ;
- (b) $(\hat{\beta}_n - \beta_0)\sqrt{n}$ converges weakly to a mean 0 Gaussian random variable with covariance $\Sigma_1(\beta_0)^{-1}$.

We prove theorems 1 and 2 in Appendix B.

4. Multicast interactions

In Sections 2 and 3, we have assumed that each interaction involves a single sender and a single receiver. The model and corresponding asymptotic theory are sufficient to cover strictly pairwise directed interactions (e.g. phone calls), but they do not describe interactions that can involve multiple receivers (e.g. e-mail messages). We call an interaction involving a single sender and possibly multiple receivers a multicast interaction.

In practice, multicast interactions are typically treated in an *ad hoc* manner via duplication—for example, interaction $i \rightarrow \{j_1, j_2, j_3\}$ becomes recorded as three separate pairwise interactions $i \rightarrow j_1, i \rightarrow j_2$ and $i \rightarrow j_3$ —giving rise to approximate likelihood and inference. In this section we explore the implications of using this approximate likelihood in the multicast setting. In particular we show it to be closely related to an extension of our model for directed pairwise interactions, and that the bias introduced by such an approximation can be quantified and in certain cases corrected.

For this, we introduce an extension of the model to the multicast setting. Let $\mathcal{I}, \mathcal{J}, \mathcal{J}_t(i), x_t(i, j)$ and β_0 be as in Section 2. For each sender i and positive integer L , let $\bar{\lambda}_t(i; L)$ be a non-negative predictable process called the baseline L -receiver intensity of sender i . Let $(t_1, i_1, J_1), \dots, (t_n, i_n, J_n)$ be the sequence of observed multicast interactions, with tuple (t, i, J) indicating that, at time t , sender i interacted with receiver set J . For a set J , let $|J|$ denote its cardinality.

Consider a model for multicast interactions where the rate of interaction between sender i and receiver set J is

$$\lambda_t(i, J) = \bar{\lambda}_t(i; |J|) \exp\left\{ \sum_{j \in J} \beta_0^T x_t(i, j) \right\} \prod_{j \in J} \mathbf{1}\{j \in \mathcal{J}_t(i)\}. \tag{6}$$

The log-partial-likelihood at time t , evaluated at β , is

$$\log\{\text{PL}_t(\beta)\} = \sum_{t_m \leq t} \left(\sum_{j \in J_m} \beta^T x_{t_m}(i_m, j) - \log \left[\sum_{\substack{J \subseteq \mathcal{J}_{t_m}(i_m) \\ |J|=|J_m|}} \exp\left\{ \sum_{j \in J} \beta^T x_{t_m}(i_m, j) \right\} \right] \right). \tag{7}$$

Suppose, instead of using the multicast model, that we use duplication to obtain pairwise interactions from the original multicast data. If we use model (1) for the pairwise data and ignore ties in the interaction times, we obtain an approximate partial likelihood:

$$\log\{\widetilde{\text{PL}}_t(\beta)\} = \sum_{t_m \leq t} \left(\sum_{j \in J_m} \beta^T x_{t_m}(i_m, j) - |J_m| \log \left[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta^T x_{t_m}(i_m, j)\} \right] \right). \quad (8)$$

We claim that $\log\{\widetilde{\text{PL}}_t(\beta)\}$ approximates $\log\{\text{PL}_t(\beta)\}$. Heuristically, replacing the sum over all sets of size $|J_m|$ in equation (7) with a sum over all multisets of size $|J_m|$ (i.e. allowing duplicate elements from $\mathcal{J}_{t_m}(i_m)$), observe that

$$\begin{aligned} \log \left[\sum_{\substack{J \subseteq \mathcal{J}_{t_m}(i_m) \\ |J|=|J_m|}} \exp \left\{ \sum_{j \in J} \beta^T x_{t_m}(i_m, j) \right\} \right] &\approx \log \left(\left[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta^T x_{t_m}(i_m, j)\} \right]^{|J_m|} \right) \\ &= |J_m| \log \left[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta^T x_{t_m}(i_m, j)\} \right]. \end{aligned}$$

In this sense, $\log\{\text{PL}_t(\beta)\} \approx \log\{\widetilde{\text{PL}}_t(\beta)\}$. Section 4.1 makes this statement more precise, and Section 4.2 analyses the bias that was introduced by maximizing $\log\{\widetilde{\text{PL}}_t(\beta)\}$ in lieu of $\log\{\text{PL}_t(\beta)\}$.

4.1. Approximation error

Define the receiver set growth sequence

$$G_n = \sum_{t_m \leq t_n} \frac{\mathbf{1}\{|J_m| > 1\}}{|\mathcal{J}_{t_m}(i_m)|}. \quad (9)$$

This sequence plays a critical role in bounding the error that was introduced by replacing $\log(\text{PL})$ with $\log(\widetilde{\text{PL}})$. When $|\mathcal{J}_{t_m}(i_m)|$ is constant G_n has linear growth but, when $|\mathcal{J}_{t_m}(i_m)|$ increases, G_n often has sublinear growth. For example, the Cauchy–Schwartz inequality gives

$$G_n \leq \sqrt{n} \left[\sum_{t_m \leq t_n} \frac{\mathbf{1}\{|J_m| > 1\}}{|\mathcal{J}_{t_m}(i_m)|^2} \right]^{1/2},$$

so, if $|\mathcal{J}_{t_m}(i_m)|/\sqrt{m} \rightarrow \infty$, then $G_n = \mathcal{O}(\sqrt{n})$. Theorem 3 (which is proved in Appendix C) bounds the approximation error in terms of G_n .

Theorem 3. Let (t_m, i_m, J_m) be a sequence of observations from a multivariate point process with intensity as given in equation (6). Assume that $\sup_t \|x_t(i, j)\|$ and $\sup_m \|J_m\|$ are bounded in probability. If $\log(\text{PL})$ and $\log(\widetilde{\text{PL}})$ are as defined in equations (7) and (8), and G_n is as defined in equation (9), then, for β in a neighbourhood of β_0 ,

$$\|\nabla[\log\{\text{PL}_{t_n}(\beta)\}] - \nabla[\log\{\widetilde{\text{PL}}_{t_n}(\beta)\}]\| = \mathcal{O}_P(G_n),$$

and

$$\|\nabla^2[\log\{\text{PL}_{t_n}(\beta)\}] - \nabla^2[\log\{\widetilde{\text{PL}}_{t_n}(\beta)\}]\| = \mathcal{O}_P(G_n).$$

4.2. Bias correction from the approximate partial likelihood

When we use *ad hoc* duplication, we are performing approximate inference under the multicast

model (6). In practice, even if we explicitly want to use the multicast model, computing the partial likelihood of equation (7) involves an intractable combinatorial sum, so we may resort to using the approximation instead. Maximizing $\log\{\widetilde{\text{PL}}_t(\cdot)\}$ instead of $\log\{\text{PL}_t(\cdot)\}$ introduces bias in the estimate of β_0 . Theorem 4 (which is proved in Appendix C) bounds the bias.

Theorem 4. Under the set-up of theorem 3, let $\hat{\beta}_n$ maximize $\log\{\text{PL}_{t_n}(\cdot)\}$ and let $\tilde{\beta}_n$ maximize $\log\{\widetilde{\text{PL}}_{t_n}(\cdot)\}$. Suppose for all n that the Hessian $(1/n)\nabla^2[\log\{\text{PL}_{t_n}(\cdot)\}]$ is uniformly locally Lipschitz and with smallest eigenvalue bounded away from zero in a neighbourhood of $\hat{\beta}_n$. If $G_n/n \rightarrow^P 0$, then

$$\|\tilde{\beta}_n - \hat{\beta}_n\| = \mathcal{O}_P(G_n/n).$$

That $\hat{\beta}_n$ is a consistent estimator of β_0 follows directly from the theory in Section 3, since the multicast case can be considered as a special case of the single-receiver case: consider the product $\mathcal{I} \times \mathbb{N}_+$ as the sender set, and the power set $\mathcal{P}(\mathcal{J})$ as the receiver set. For sender (i, L) , the process $\tilde{\lambda}(i; L)$ is then the baseline send intensity, and $\{J \subseteq \mathcal{J}_t(i) : |J| = L\}$ is the receiver set; for sender–receiver pair $((i, L), J)$, vector $\Sigma_{j \in J} x_t(i, j)$ is the covariate vector. Consistency of the MPLE now follows from theorem 2.

Suppose that the true MPLE, $\hat{\beta}_n$, is a root- n -consistent estimate of β_0 . (Theorem 2 gives sufficient conditions.) Theorem 4 says that, if $|\mathcal{J}_{t_m}(i_m)|$ grows sufficiently fast to make G_n smaller than $\mathcal{O}_P(\sqrt{n})$, then the approximate MPLE, $\tilde{\beta}_n$, is also root n consistent. Moreover, if $(\hat{\beta}_n - \beta_0)\sqrt{n}$ is asymptotically Gaussian, then $(\tilde{\beta}_n - \beta_0)\sqrt{n}$ is asymptotically Gaussian with the same covariance matrix but possibly a different mean. Under enough regularity, $-(1/n)[\nabla^2 \log\{\widetilde{\text{PL}}_{t_n}(\tilde{\beta}_n)\}]$ consistently estimates the limiting covariance of $(\tilde{\beta}_n - \beta_0)\sqrt{n}$. To obtain the mean, we use a parametric bootstrap as follows.

Assume that the conditions of theorem 4 hold. The residual $\tilde{\beta}_n - \beta_0$ depends continuously on β_0 and the covariate process $x_t(i, j)$. Since $\hat{\beta}_n$ is a consistent estimator of β_0 , we can estimate the bias in $\tilde{\beta}_n$ via a parametric bootstrap. We generate a bootstrap replicate data set $\{(t_m, i_m, J_m^{(r)})\}$ by drawing $J_m^{(r)}$, a random subset of $\mathcal{J}_{t_m}(i_m)$ with size $|J_m|$ whose elements are drawn proportional to $w_{t_m}(\hat{\beta}_n, i_m, \cdot)$. We then obtain a bootstrap approximate MPLE, $\tilde{\beta}_n^{(r)}$, by maximizing $\text{PL}_{t_m}^{(r)}$, where

$$\log\{\widetilde{\text{PL}}_t^{(r)}(\beta)\} = \sum_{t_m \leq t} \left(\sum_{j \in J_m^{(r)}} \beta^T x_{t_m}(i_m, j) - |J_m^{(r)}| \log \left[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\beta^T x_{t_m}(i_m, j)\} \right] \right).$$

Note that $x_t(i, j)$ is determined from the original data set, not the bootstrap data set. For each bootstrap replicate, we obtain a residual $\tilde{\beta}_n^{(r)} - \tilde{\beta}_n$. With R bootstrap replicates, we estimate the bias by

$$\widehat{\text{bias}} = \frac{1}{R} \sum_{r=1}^R \tilde{\beta}_n^{(r)} - \tilde{\beta}_n.$$

We adjust for estimator bias by replacing $\tilde{\beta}_n$ with $\tilde{\beta}_n - \widehat{\text{bias}}$.

4.3. Simulation

We show a simulation study to verify the result of theorem 4 empirically. In the study, we have one sender, and a receiver count $|\mathcal{J}|$ ranging from 32 to 1000. Each receiver was assigned a constant covariate vector $x(j)$ whose elements were independent Bernoulli random variables with success probability $\frac{1}{2}$. The components of the true coefficient vector β were drawn independently from the standard normal distribution.

We chose sample sizes n ranging from 32 to 100000. For each receiver count $|\mathcal{J}|$, we drew

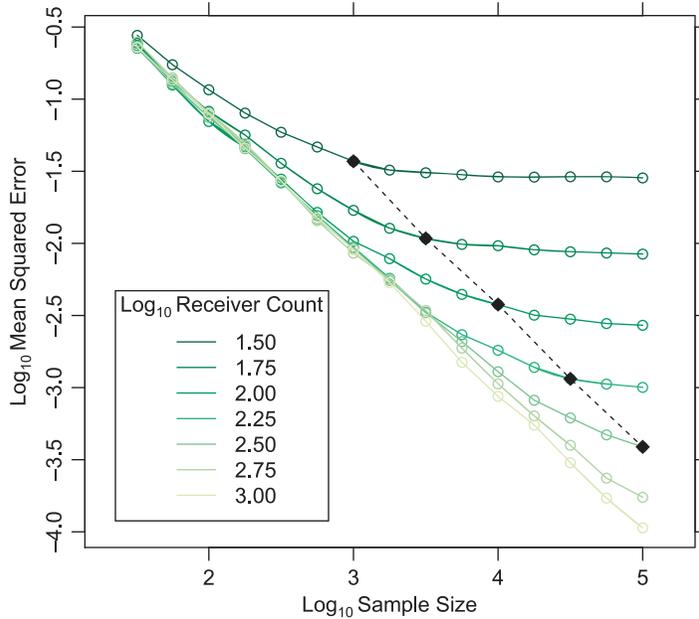


Fig. 1. Multicast coefficient estimation error with approximate MPLE: the receiver count $|\mathcal{J}|$ is equal to the square root of sample size n along the broken line

n multicast messages, with the receiver set J_m for message m determined as follows: we determined the size $|J_m|$ by drawing from a geometric distribution with success probability $p=0.4$, so that $\mathbb{P}\{|J_m|=L\}=(1-p)^{L-1}p$ for $L \geq 1$; once $|J_m|$ had been determined, we chose between all receiver sets with cardinality $|J_m|$, with $\mathbb{P}\{J_m=J\} \propto \exp\{\sum_{j \in J} \beta^T x(j)\}$. Once we had generated the message data, we computed $\tilde{\beta}$ by maximizing the approximate log-partial-likelihood analogous to equation (8). Finally, we computed $\|\beta - \tilde{\beta}\|$.

We repeated this procedure for 100 random replicates at each receiver count and sample size, and computed the mean-squared-error of $\tilde{\beta}$ by averaging the value of $\|\beta - \tilde{\beta}\|^2$ over all replicates. Fig. 1 displays the results. From the spacings of the asymptotes of the curves in Fig. 1, we can see that, if $|\mathcal{J}|$ does not grow with n , then the error $\|\beta - \tilde{\beta}\|^2$ is roughly $\mathcal{O}(|\mathcal{J}|^{-2})$ for large n ; strictly speaking, the assumptions of theorem 4 do not hold in this scenario since $G_n = \mathcal{O}_P(n/|\mathcal{J}|)$, but nevertheless theorem 4 predicts an error rate of $\mathcal{O}(|\mathcal{J}|^{-2})$. For theorem 4 to apply, we require that $|\mathcal{J}|$ grow with n . From the slope of the broken line in Fig. 1, we can see that, if $|\mathcal{J}| = \sqrt{n}$, then $\|\beta - \tilde{\beta}\|^2$ is roughly $\mathcal{O}_P(n^{-1})$; this agrees with theorem 4, since $G_n = \sqrt{n}$ in this situation.

5. Fitting the model to a corporate e-mail network

Recall from Section 1 that, given a set of interaction data triples (t, i, j) , a primary modelling goal lies in determining which characteristics and behaviours of the senders and receivers are predictive of interaction. The modelling and inference framework that was introduced above enables us to address these concerns directly, as we now demonstrate through the analysis of a corporate e-mail network consisting of a large subset of the e-mail messages sent within the Enron corporation between 1998 and 2002. These e-mail interaction data give rise to the following questions.

- (a) *Homophily*: to what extent are traits shared between individuals (gender, department or seniority) predictive of interaction behaviours?
- (b) *Network effects*: to what extent are dyadic or even triadic network effects, as characterized by past interaction behaviours, relevant to predicting future interaction behaviours?

We undertake our analysis by using the multicast proportional intensity modelling framework developed in Sections 2 and 3 above, employing both static covariates reflecting actor traits, as well as dynamic covariates capturing network effects. The bootstrap technique that was introduced in Section 4 for multicast interactions is then used to reduce bias in the estimated effects, as well as to demonstrate that our asymptotic approximations are reasonable in this data modelling regime. We conclude this section with a discussion of the goodness of fit of our model in this setting, before turning our attention in Section 6 to an evaluation of the strength of homophily and network effects in explaining these data.

5.1. Data and methods

Our example analysis uses publicly available data from the Enron e-mail corpus (Cohen, 2009), a large subset of the e-mail messages that were sent within the Enron corporation between 1998 and 2002, and made public as the result of a subpoena by the US Federal Energy Regulatory Commission during an investigation into fraudulent accounting practices. We analyse the data set that was compiled by Zhou *et al.* (2007), comprising 21635 messages sent between 156 employees between November 13th, 1998, and June 21st, 2002, along with the genders, seniorities and departments of these employees.

Approximately 30% of these messages have more than one recipient across their ‘To’, ‘CC’ and ‘BCC’ fields, with a few messages having more than 50 recipients. In the subsequent analysis, we exclude messages with more than five recipients—a subjectively chosen cut-off that avoids e-mails sent *en masse* to large groups.

We model these data by using the multicast proportional intensity model of Section 4, with $\mathcal{I} = \mathcal{J} = \{1, 2, \dots, 156\}$ and $\mathcal{J}_i(i) = \mathcal{I} \setminus \{i\}$, and with static and dynamic covariates described in the next section. We fit the model by first maximizing the approximate log-partial-likelihood $\log\{\text{PL}_t(\beta)\}$ of model (8), and then employing a parametric bootstrap to estimate and correct the resultant bias in parameter estimates. We calculate standard errors by using the corresponding asymptotic theory. In the setting of this example, the interaction count is high, so the asymptotic framework that was developed in Sections 3 and 4 is natural. The main violation of assumptions 1–4 is that our covariates (described in Section 5.2) may in principle be unbounded; nevertheless, bootstrap calculations (described in Section 5.3) show that the asymptotic approximations that we employ remain reasonable in this regime.

We wrote custom software in the C programming language to fit the model by using Newton’s method. Our implementation exploits structure in the covariates to make the computational complexity of the fitting procedure roughly linear in the number of messages and the number of actors. Appendix A describes the fitting procedure in detail. It took approximately 20 min to fit the full model by using a standard desktop computer with a 2.4-GHz processor and 4 Gbytes of random-access memory. Each bootstrap replicate took approximately 10 min to generate and fit, using the original estimate as a starting point for the fitting algorithm. Most of the complexity in the fitting procedure is due to the inclusion of triadic covariates as described below; including only dyadic covariates reduces the fitting time to approximately 1 min.

5.2. Covariates

The goal of our investigation is to assess the predictive ability of actor traits and network

effects. For this, we choose covariates that encode these traits and effects. Each covariate is encoded as a component of the time varying dyad-dependent vector $x_t(i, j)$, which is linked to the rate of interaction between sender i and receiver j via the multicast proportional intensity model (1).

5.2.1. Static covariates to measure homophily and group level effects

Consider first those actor traits that do not vary with time: the actors' genders, departments and seniorities. We encode the traits of actor i and their second-order interactions by using nine actor-dependent binary (0–1) variables, as described in Table 1.

We encode all 20 identifiable first-order interactions between the traits of sender i and receiver j as components of $x_t(i, j)$. We do this by using variates of the form $Y(j)$ and $X(i) \cdot Y(j)$, where X and Y are chosen from the list of four actor-dependent variates (L, T, J, F). We also include four receiver-specific covariates of the form $\mathbf{1} \cdot Y(j)$. We cannot identify the coefficients for covariates of the form $X(i) \cdot \mathbf{1}$; if a component of $x_t(i, j)$ is the same for all values of j , then the corresponding component of β will not be identifiable since the product of the two can be absorbed into $\bar{\lambda}_t(i)$ without changing the likelihood.

We measure homophily by way of the estimated coefficients for covariates of the form $X(i) \cdot X(j)$. For example, if the sum of the coefficients of $\mathbf{1} \cdot J(j)$ and $J(i) \cdot J(j)$ is large and positive, this tells us that junior employees exhibit homophily in their choice of message recipients.

5.2.2. Dynamic covariates to measure network effects

Static effects are useful for determining which traits are predictive of the relative rate of interaction between sender i and receiver j , but they do not shed light on network effects. Therefore, we are also interested in the predictive relevance of the dynamic network behaviours that are described in Table 2. The first two behaviours (*send* and *receive*) are 'dyadic', involving exactly two actors, whereas the last four (*2-send*, *2-receive*, *sibling* and *cosibling*) are 'triadic', involving exactly three actors.

To measure first-order dependence of message exchange behaviour on these network effects, we introduce binary indicators for all six effects as components of $x_t(i, j)$. These indicators depend on the sender i , the receiver j and the history of the process at the current time t . By the shorthand notation $1\{\text{send}\}$, we denote the indicator variable depending on sender i , receiver j and the current time t , which indicates whether i has sent j a message before time t , with the remaining notations ($1\{\text{receive}\}$, $1\{2\text{-receive}\}$, etc.) defined similarly.

To measure higher order time dependence, we introduce additional covariates of the following form. We partition the interval $[-\infty, t)$ into $K = 7$ subintervals:

Table 1. Actor-specific traits, with counts of how many of the 156 actors share each trait

Variate	Characteristic of actor i	Count
$L(i)$	Member of the Legal Department	25
$T(i)$	Member of the Trading Department	60
$J(i)$	Seniority is junior	82
$F(i)$	Gender is female	43

Table 2. Dynamic covariates to measure network effects

Behaviour	Actors	Description
send	$i \rightarrow j$	i has sent j a message in the past
receive	$i \leftarrow j$	i has received a message from j in the past
2-send	$i \rightarrow h \rightarrow j$	There is an actor h such that i has sent h a message and h has sent j a message in the past
2-receive	$i \leftarrow h \leftarrow j$	There is an actor h such that i has received a message from h , and h has received a message from j
sibling		There is an actor h such that h has sent i and j messages in the past
cosibling		There is an actor h such that h has received messages from i and j

$$[-\infty, t) = [t - \Delta_K, t - \Delta_{K-1}) \cup [t - \Delta_{K-1}, t - \Delta_{K-2}) \cup \dots \cup [t - \Delta_1, t - \Delta_0)$$

where $\infty = \Delta_K > \Delta_{K-1} > \dots > \Delta_1 > \Delta_0 = 0$ and ‘ $t - \infty$ ’ is defined to be $-\infty$. Specifically, we set $\Delta_k = 7.5 \text{ min} \times 4^k$ for $k = 1, \dots, K - 1$ so that for k in this range Δ_k takes the values 30 min, 2 h, 8 h, 32 h, 5.33 days and 21.33 days.

Define the half-open interval $I_t^{(k)} = [t - \Delta_k, t - \Delta_{k-1})$. For $k = 1, \dots, K$ we define the dyadic effects

$$\begin{aligned} \text{send}_t^{(k)}(i, j) &= \#\{i \rightarrow j \text{ in } I_t^{(k)}\}, \\ \text{receive}_t^{(k)}(i, j) &= \#\{j \rightarrow i \text{ in } I_t^{(k)}\}; \end{aligned}$$

for sender i , such that these covariates measure the number of messages sent to, and respectively received by, receiver j in time interval $I_t^{(k)}$.

The dyadic effects have been defined in the manner above to enable easy interpretation of the corresponding coefficients. To illustrate this, for $k = 1, \dots, K$, suppose that β_k is the coefficient corresponding to $\text{send}_t^{(k)}(i, j)$. If we observe the message $i \rightarrow j$ at time t , then, for future time t' in the interval $(t, t + \Delta_1]$, the rate $\lambda_{t'}(i, j)$ will be multiplied by the factor $\exp(\beta_1)$; for t' in the interval $(t + \Delta_1, t + \Delta_2]$, the rate will be multiplied by $\exp(\beta_2)$; this continues similarly, with the rate being multiplied by $\exp(\beta_k)$ whenever $t' \in (t + \Delta_{k-1}, t + \Delta_k]$ and, equivalently, when $\Delta_{k-1} < t' - t \leq \Delta_k$. Thus, the coefficients β_1, \dots, β_K measure the effect of a ‘send event’ and how this effect decays over time. We expect that β_k will decrease as k increases, but we do not enforce this constraint on the estimation procedure.

The triadic effects involve pairs of messages. For $k = 1, \dots, K$ and $l = 1, \dots, K$ we define the triadic effects

$$\begin{aligned} 2\text{-send}_t^{(k,l)}(i, j) &= \sum_{h \neq i, j} \#\{i \rightarrow h \text{ in } I_t^{(k)}\} \#\{h \rightarrow j \text{ in } I_t^{(l)}\}, \\ 2\text{-receive}_t^{(k,l)}(i, j) &= \sum_{h \neq i, j} \#\{h \rightarrow i \text{ in } I_t^{(k)}\} \#\{j \rightarrow h \text{ in } I_t^{(l)}\}, \\ \text{sibling}_t^{(k,l)}(i, j) &= \sum_{h \neq i, j} \#\{h \rightarrow i \text{ in } I_t^{(k)}\} \#\{h \rightarrow j \text{ in } I_t^{(l)}\}, \\ \text{cosibling}_t^{(k,l)}(i, j) &= \sum_{h \neq i, j} \#\{i \rightarrow h \text{ in } I_t^{(k)}\} \#\{j \rightarrow h \text{ in } I_t^{(l)}\}. \end{aligned}$$

For sender i and receiver j , the covariate $\text{2-send}_t^{(k,l)}(i, j)$ counts the pairs of messages such that, for some h distinct from i and j , message $i \rightarrow h$ occurred in interval $I_t^{(k)}$ and message $h \rightarrow j$ occurred in interval $I_t^{(l)}$; the other covariates behave similarly.

As with the dyadic effects, the triadic effects are designed so that their coefficients have a straightforward interpretation. However, since triadic effects involve pairs of messages, the interpretation is a little more involved. We illustrate with the $\text{2-send}_t^{(k,l)}(i, j)$ covariate having coefficient $\beta_{k,l}$ for $k = 1, \dots, K$ and $l = 1, \dots, K$. Take i and j to be two actors. Suppose that at time t we observe the message $h \rightarrow j$. At this point, we look through the history of the process for all messages of the form $i \rightarrow h$; when paired with the original $h \rightarrow j$ message, each of these defines a ‘2-send event’. The other 2-send events are defined as follows: if at time s we observe the message $i \rightarrow h$, then we enumerate all observed messages $h \rightarrow j$ in the history of the process; when each of these is paired with the original $i \rightarrow h$ event it constitutes a 2-send event. A pair (s, t) can be associated with each 2-send event, where s is the time of the $i \rightarrow h$ message and t is the time of the $h \rightarrow j$ message. At time t' after s and t , the existence of the 2-send event causes the sending rate $\lambda_{t'}(i, j)$ to be multiplied by the factor $\exp(\beta_{k,l})$, where $t' \in (s + \Delta_{k-1}, s + \Delta_k]$ and $t' \in (t + \Delta_{l-1}, t + \Delta_l]$. We expect $\beta_{k,l}$ to decrease as k and l increase, though again we do not enforce this constraint in the fitting procedure.

As previously noted, Butts (2008) used a variant of the proportional intensity model to capture interaction behaviour in social settings. As such, a correspondence can be drawn between certain of the covariates in Butts (2008) and those outlined above. If we set $K = 1$, then the send_t covariate is equivalent to an unnormalized version of Butts’s persistence covariate, and the sum $\text{send}_t + \text{receive}_t$ becomes an unnormalized version of Butts’s preferential attachment covariate. For the triadic effects, Butts’s OTP, ITP, ISP and OSP covariates are analogous to the 2-send, 2-receive, sibling and cosibling covariates, although the exact definitions differ slightly. (For example, $\text{OTP}_t(i, j)$ is defined as $\sum_h \min[\#\{i \rightarrow h \text{ in } (-\infty, t)\}, \#\{h \rightarrow j \text{ in } (-\infty, t)\}]$.) The versions of these covariates that we have introduced above, however, are designed to enable a more precise quantification of the time dependence of network effects, as well as a more straightforward interpretation of the corresponding coefficients. In related models, Vu *et al.* (2011a, b) used similar covariates, except that they did not partition $[-\infty, t)$ into subintervals.

5.3. Bootstrap bias correction

Given the model specification, data and covariates outlined above, we can estimate the parameter vector β_0 under the approximate log-partial-likelihood (8). Recall that the results of Section 4 bound the bias resulting from this approximate MPLE procedure as a function of the growth rate of the recipient set \mathcal{J} over time. Here, treating the set \mathcal{J} of 156 Enron employees as constant, the resultant bias is of order $1/|\mathcal{J}|$ —and, since $|\mathcal{J}| = 156$ is of the order of the square root of the number 21365 of messages in the data set, we can correct this bias by using the parametric bootstrap outlined at the end of Section 4.

Fig. 2 summarizes the corresponding bootstrap residuals (from 500 replicates) for each component of the estimated parameter vector β_0 ; we can see from Fig. 2 that treating messages with multiple recipients as multiple single-recipient messages introduces bias of the order of the standard error for most of the coefficients. There is a pronounced negative bias in coefficient estimates for the dyadic effects, which is representative of a more general phenomenon. Sparsity in the components of $x_t(i, j)$ (when considered as a function of j), when combined with high values of the corresponding entries β , leads to negative bias in the coefficient estimates when there are messages with multiple recipients. The approximation in equation (7) is worst when, for some j^* , weight $w_{t_m}(i_m, j^*)$ far exceeds all other values of $w_{t_m}(i_m, j)$, so that $w_{t_m}(i_m, j^*) \approx W_{t_m}(i_m)$;

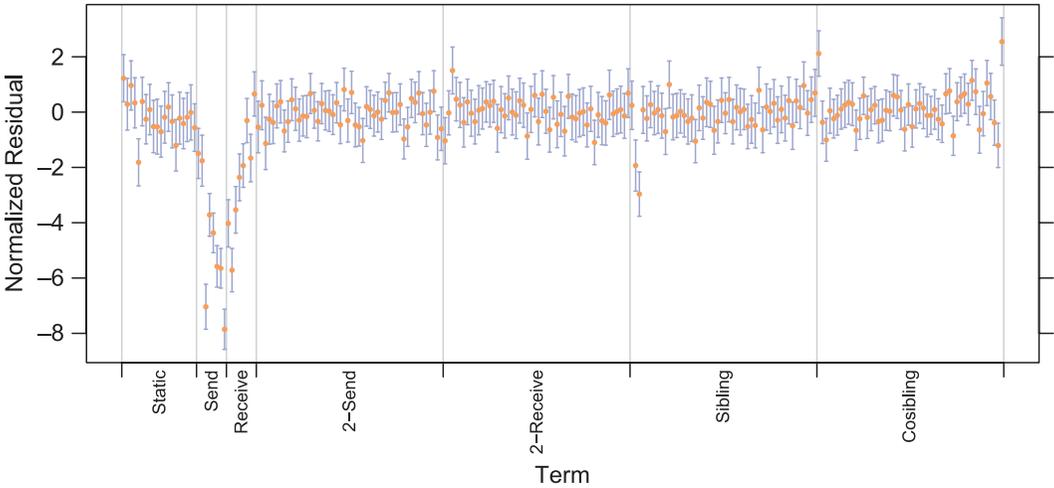


Fig. 2. Enron bootstrap residuals—summary of bootstrap residuals for estimated coefficients by using the Enron data set, normalized by estimated standard errors (the coefficients are grouped by model term): ●, means; |, ±1 standard deviation

when $|J_m|$ is large, the maximum of \widetilde{PL} will avoid this situation by shrinking β where $x_{t_m}(i_m, j)$ is sparse. The dyadic covariates are particularly sparse, so the estimates for their coefficients are particularly vulnerable to this bias.

Besides correcting for bias, the bootstrap simulations give us confidence that the asymptotic approximations are reasonable. The simulated standard errors are very close to those predicted by the theory, despite the norm $\|x_t(i, j)\|_2$ being potentially unbounded, contrary to the assumptions of theorem 1.

5.4. Goodness of fit

Table 3 details an *ad hoc* analysis of deviance for the fitted model, showing how the approximate deviance (twice the approximate log-partial-likelihood) behaves as we add consecutive terms to the model. Group level (static) effects account for 15% of the null deviance and network effects

Table 3. *Ad hoc* analysis of deviance for the Enron model†

<i>Term</i>	<i>Degrees of freedom</i>	<i>Deviance</i>	<i>Residual degrees of freedom</i>	<i>Residual deviance</i>
Null			32261	325412
Static	20	50365	32241	275047
Send	8	107942	32233	167105
Receive	8	5919	32225	161186
Sibling	50	3601	32175	157585
2-send	50	516	32125	157069
Cosibling	50	1641	32075	155428
2-receive	50	158	32025	155270

†Residual deviance is defined as twice the approximate negative log-partial-likelihood from equation (8). The ‘static’ term contains the group level effects, and the other terms contain the network effects.

account for 37%. The most dramatic decrease in the residual deviance comes from introducing the ‘send’ terms into the model; with only 8 degrees of freedom, they can account for 33% of the null deviance. The full model accounts for 52% of the null deviance.

The residual deviance for the full model is approximately 4.8 times the residual degrees of freedom, and so an *ad hoc* adjustment for this overdispersion is to multiply the calculated standard errors by $\sqrt{4.8} \approx 2.2$.

Note, however, that the residual deviance by itself is not adequate as a goodness-of-fit measure, as it depends only on the estimated coefficients (see section 4.4.5 of McCullagh and Nelder (1989) for discussion of a related problem for logistic regression with sparse data). To shed more light on how well the model fits these data, we use a normalized version of the martingale residual from Therneau *et al.* (1990), which we call a Pearson residual. Specifically, given $\hat{\beta}$, we define

$$\hat{N}_t(i, j) = \sum_{i_m \leq t} \frac{w_{i_m}(\hat{\beta}, i, j)}{W_{i_m}(\hat{\beta}, i)} \mathbf{1}\{i_m = i\},$$

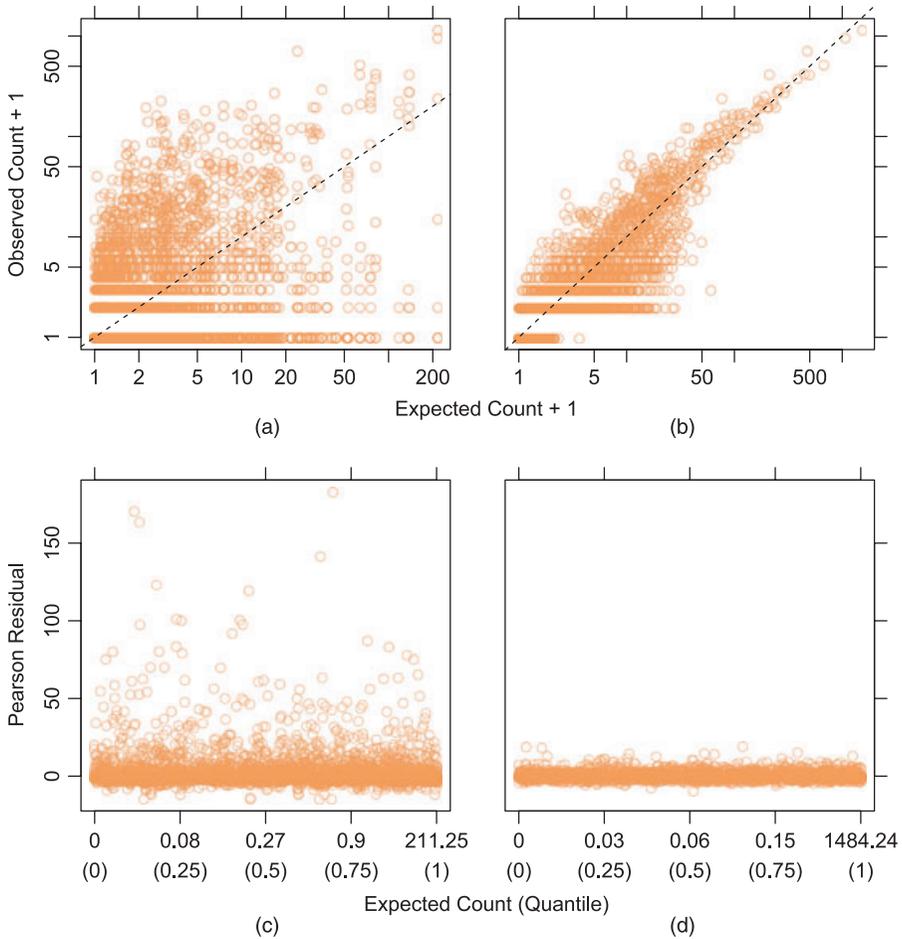


Fig. 3. Godness-of-fit plots for (a) static and (b) static and dynamic observed count $\hat{N}_{\infty}(i, j)$ plotted against expected count $\hat{N}_{\infty}(i, j)$, and (c) static and (d) static and dynamic Pearson residual $\{\hat{N}_{\infty}(i, j) - \hat{N}_{\infty}(i, j)\} / \hat{N}_{\infty}(i, j)^{1/2}$ versus expected count

which is the expected number of $i \rightarrow j$ events given the estimated model, with $\int \bar{\lambda}_t(i) dt$ estimated by the Breslow (1974) estimate $\int W_t(\hat{\beta}, i)^{-1} \sum_j dN_{i,j}(t)$. The martingale residual analogous to that of Therneau *et al.* (1990) is then defined as $N_t(i, j) - \hat{N}_t(i, j)$; we normalize this quantity by an estimate of its standard deviation to obtain a ‘Pearson’ residual:

$$\{N_t(i, j) - \hat{N}_t(i, j)\} / \hat{N}_t(i, j)^{1/2}.$$

Fig. 3(a) shows a plot of $N_\infty(i, j)$ versus $\hat{N}_\infty(i, j)$ for two different models. In the ‘static’ model, we include only the static covariates, whereas, in the full (‘static and dynamic’) model, we also include all six types of network covariates. The fit for the static model is poor. For instance, it repeatedly predicts up to 200 $i \rightarrow j$ events where we observed only one or two; likewise, the model predicts one or fewer events where we observed up to 20. For the full model, which includes the dynamic covariates to account for network effects, the fit is much better, with the relationship between observed and expected interaction counts being roughly linear.

Fig. 3(b) shows the Pearson residuals. For the full model, more than 95% are less than 1.21 in absolute value, and the maximum absolute residual is 18.7. In contrast, the 95% quantile for the absolute residuals in the static model is at 3.5, and the maximum absolute residual is 182.7. The sum of squares of the residuals (X^2) is 17281 in the full model; that for the static model is over 34 times higher (596253). We do not know what a ‘reasonable’ value for X^2 is; an *ad hoc* degrees-of-freedom calculation suggests that for the full number this should be roughly equal to $23944 = 156 \times 155 - (20 + 2 \times 8 + 4 \times 50)$, which suggests that the full model is too aggressive. The bootstrap simulations confirm this, with 17055 being 5.6 standard deviations below the mean value X^2 for the bootstrap replicates.

For a more parsimonious model, we might drop most of the triadic effects. Indeed, the model which only uses dyadic effects has an X^2 -value of 21094. However, at this stage we desire a model with the lowest possible bias, and also wish to acquire estimates for all of the network effects.

6. Evaluating the strength of homophily and network effects

Given the model fitting procedure and results that were described above, we may now evaluate the strength of homophily and network effects in predicting the interaction behaviour that is observed in our data.

6.1. Assessing evidence for homophily in the Enron data

The analyses of Section 5 have established that our multicast proportional intensity model with chosen covariates is reasonably accurate in describing message recipient selection, conditional on the sender and the history of the process. Thus, we are justified in using the estimated coefficients from the model to assess the predictive ability of the corresponding covariates.

Our first task is to gauge the predictive strength of homophily. For this, Table 4 shows the estimated group level coefficients for our model. Notably, homophily is evident for almost all main effects (department, seniority and gender): the estimated coefficients of $L(j)$, $T(j)$ and $J(j)$ are all negative, whereas the sum of the estimated coefficients of $F(j)$ and $F(i) \cdot F(j)$ is positive. Negative homophily is evidenced in that the sum of the coefficients for $L(j)$ and $L(i) \cdot L(j)$ is negative. The coefficient of $F(j)$ and the sum of the coefficients for $T(j)$ and $T(i) \cdot T(j)$, and $J(j)$ and $J(i) \cdot J(j)$ are not significant.

Taking gender as an example, the way that the homophily effect manifests is as follows: if i is a female sending a message at time t , and person j is identical to person j' except for

Table 4. Estimated coefficients and standard errors for group level covariates of the form $X(i) \cdot Y(j)$, where i is the sender, j is the receiver and $X(i)$ and $Y(j)$ are given in the row and column headings†

Sender	Coefficients for the following receivers:			
	<i>L</i>	<i>T</i>	<i>J</i>	<i>F</i>
1	-0.91† (0.04)	-0.36† (0.04)	-0.34† (0.04)	0.04 (0.03)
<i>L</i>	0.63† (0.05)	0.28† (0.05)	0.22† (0.04)	0.15† (0.04)
<i>T</i>	0.32† (0.07)	0.43† (0.05)	0.27† (0.05)	-0.07 (0.05)
<i>J</i>	0.06 (0.05)	0.28† (0.04)	0.37† (0.03)	-0.13† (0.03)
<i>F</i>	0.59† (0.05)	-0.21† (0.05)	-0.09 (0.04)	0.15† (0.03)

†Significant (via Wald test) at level 10^{-3} .

gender, then i is more likely to send to the similarly gendered individual. The relative rate is $\exp(0.04 + 0.15) \approx 1.2$. The characterization for other types of homophily is similar.

Conspicuously, the only example of negative homophily is when the sender i is in the Legal Department. In this case, if person j is identical to person j' except for department, then i is more likely to send to an individual in a different department. The relative rates for the three departments are $\exp(0.63 - 0.91) \approx 0.76$ for the Legal Department, $\exp(0.28 - 0.36) \approx 0.92$ for the Trading Department and $\exp(0) = 1$ for any other department.

Were we interested only in homophily, we might be tempted to forgo the proportional intensity model (1), and instead to perform a contingency table analysis. The on-line supplementary material explores this approach in detail. The major shortcoming of the contingency table approach is that it assumes that the messages are independent, which leads to bias in the parameter estimates.

6.2. Evaluating the importance of network effects

In Section 6.1 we established that homophily was predictive of sending behaviour, even after accounting for network effects. We now investigate the characteristics of these network effects and establish which of these effects are of greatest importance.

To begin our analysis, Table 5 shows the estimated coefficients for the network indicator effects, giving a crude picture of the predictive importance of each network effect. The estimated coefficients are all positive, indicating that network effects strengthen the ties between individuals. The estimated coefficient for 1{send} is over three times larger than the other coefficients, agreeing with the general notion that one is most likely to do today the things that one did yesterday. The next tier of indicator effects comprises 1{sibling}, 1{sibling} and 1{2-send}, whose estimated coefficients range from 0.67 to 1.06. Two triadic effects, 1{2-receive} and 1{cosibling}, are not significantly predictive of sending behaviour.

The estimated coefficients for the recency-dependent covariates, which are shown in Figs 4 and 5, give a more complete picture of network effects. Firstly, we can see that dyadic effects persist for over 3 weeks from the time that a message has been sent. The decay of the estimated

Table 5. Estimated coefficients for network indicator effects†

Variate	Coefficient
1{send}	3.26 (0.03)
1{receive}	0.97 (0.02)
1{2-send}	0.67 (0.05)
1{2-receive}	0.01 (0.04)
1{sibling}	1.06 (0.05)
1{cosibling}	0.09 (0.04)

†Standard errors are given in parentheses.

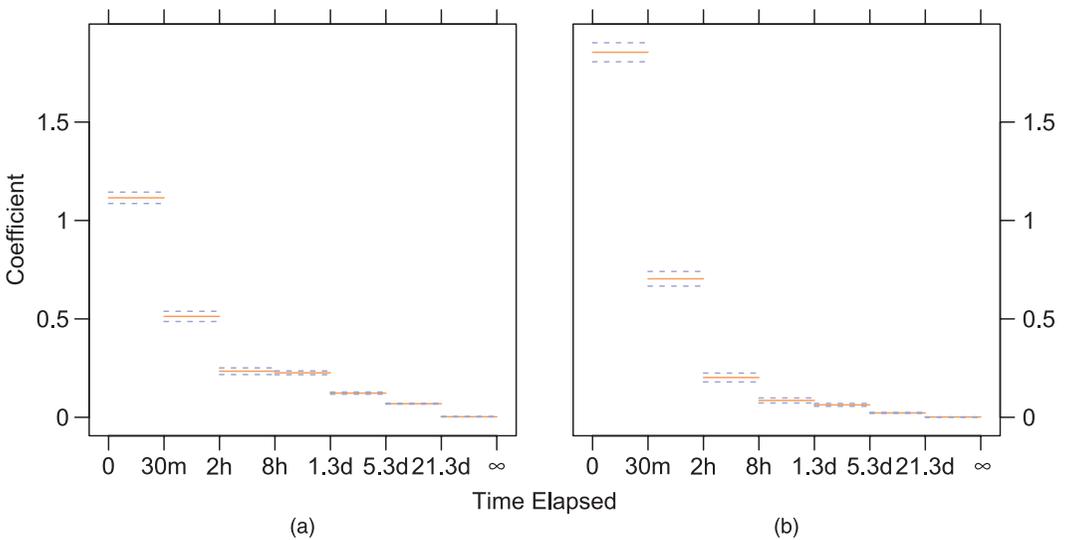


Fig. 4. Estimated coefficients for dyadic effects, with standard errors: (a) send; (b) receive

coefficients is roughly exponential in the time elapsed, corresponding to a superexponential decay in the relative sending rate. For 30 min after i sends a message to j , our estimated model predicts that the rate at which i sends to j will be multiplied by $\exp(1.11) \approx 3.05$, and the rate at which j sends to i will be multiplied by $\exp(1.85) \approx 6.39$; then, between 30 min and 2 h, the rates will be multiplied by $\exp(0.51) \approx 1.67$ and $\exp(0.70) \approx 2.02$ respectively; this proceeds similarly until after 21.3 days, when the rates will be multiplied by $\exp(0.003) \approx 1.002$ and $\exp(0.002) \approx 1.002$.

Comparing the coefficients for $\text{send}_t^{(k)}$ with those of $\text{receive}_t^{(k)}$ we see that the latter are higher for $k \leq 2$, whereas the former are higher for $k > 2$. The corresponding intuition is that, if A is

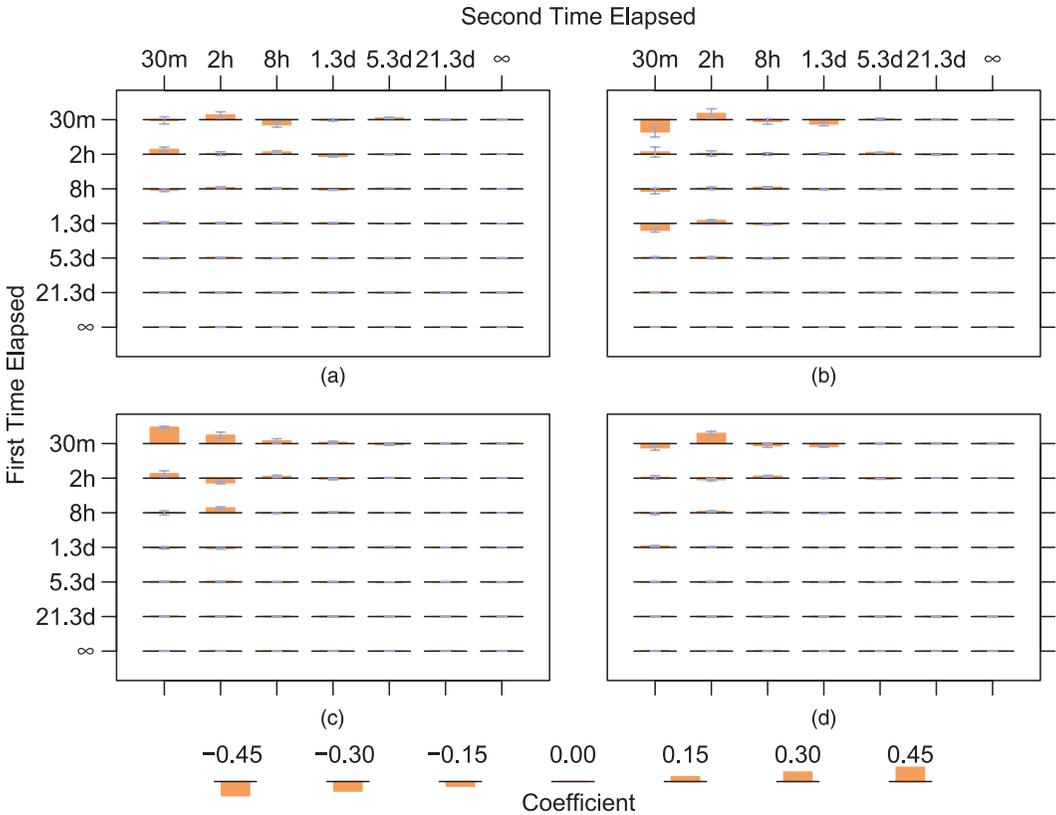


Fig. 5. Estimated coefficients for triadic effects, with standard errors: (a) 2-send; (b) 2-receive; (c) sibling; (d) cosibling

sending a message up to 2 h after receiving a message from B, then A is likely to respond to B but, after that, A is more likely to send to an individual whom A e-mailed at the time of receiving B’s original message (provided that B and this other individual are identical in all other respects). The time window during which reciprocation is more important than past habit is less than 8 h.

From Fig. 5, we can see that the triadic effects are in general less pronounced and are much more short lived than the dyadic effects. About 86% of the estimated coefficients are within 3 standard errors of 0; even those that are significantly non-zero mostly lie between -0.05 and 0.05 . The exceptions are the coefficients for $sibling_t^{(1,1)}$ (0.51), $sibling_t^{(2,2)}$ (-0.14), $sibling_t^{(3,2)}$ (0.15), $cosibling_t^{(1,2)}$ (0.32), $2\text{-receive}_t^{(4,1)}$ (-0.21) and $2\text{-receive}_t^{(4,2)}$ (0.09). We may interpret these coefficients as follows.

- (a) *sibling*: if B sent A and C messages in the last 30 min or between 2 and 8 h ago, then A and C are more likely to send messages to each other; however, if B sent A and C messages between 30 min and 2 h ago, then A and C are less likely to send messages to each other.
- (b) *cosibling*: if A sent a message to B in the last 30 min, and C sent a message to B between 30 min and 2 h ago, then A will send to B at a higher rate.
- (c) *2-receive*: if A sent a message to B in the last 30 min, and B sent a message to C between 8 h and 32 h ago, then C will send to A at a lower rate; if, however, the message from A to B was sent between 30 min and 2 h ago, then C will send to A at a higher rate.

Given the emphasis on transitivity in the networks literature, it may at first seem disconcerting that most of the estimated coefficients for the time-dependent triadic effects are found to be insignificant in this analysis. However, one must bear in mind that, except for messages sent to them directly, individuals are likely to have no knowledge of their colleagues' e-mail activities, and therefore there is no reason why this activity should directly affect sending behaviours. Any predictive power that the triadic effects have, then, must be due to correlation with exogenous factors. In this light, it is not surprising that the triadic effects are small and have small time horizons.

The results above provide a detailed view of the ways in which network effects can manifest themselves in data. The on-line supplementary material contains comparative analyses based on an actor-oriented model and an exponential random-graph model. (See Snijders *et al.* (2010) and Anderson *et al.* (1999) respectively for detailed surveys.) These analyses further bolster our confidence in the results of this section.

7. Conclusion

Our analysis of the Enron corpus in Sections 5 and 6 has demonstrated the ways in which static and dynamic effects manifest themselves in e-mail communication networks, and we expect similar conclusions to hold broadly for other types of directed interaction data. Relative to alternatives such as contingency table analyses, actor-oriented network models and exponential random-graph models, an advantage of our approach lies in its ability to model the given data directly, rather than in an aggregated form. We can adjust for network effects to obtain more reliable estimates of homophily, and by using continuous time information we obtain precise quantification on the time-dependent behaviour of the network effects.

In this work, our focus has been on the coefficient vector β . We have used partial likelihood for its estimation, enabling us to treat each sender-specific baseline intensity $\lambda_t(i)$ as a nuisance parameter. Were we to use the model for prediction, we would need to estimate baseline intensities; this could be done by using a Nelson–Aalen estimator as in Andersen *et al.* (1993).

The foundation of our work is Cox's (1972) proportional intensity model and partial likelihood theory, tools which he first introduced 40 years ago and which have been significantly developed since then (Cox, 1975; Fleming and Harrington, 1991; Andersen *et al.* 1993; Martinussen and Scheike, 2006; Cook and Lawless, 2007). These tools are used extensively in the context of survival analysis but require further development for use in modelling interaction data. In this vein, we have extended the associated theory in two directions: first, we have provided results that are asymptotic in time rather than in the size of the population under study; second, we have shown that treating multicast interactions via duplication leads to bias in the parameter estimates (which can in turn be corrected in certain regimes).

We find that the proportional intensity model with time varying covariates is particularly useful for modelling repeated directed interactions. The model is simple, flexible and well established, and it facilitates investigation into which traits and behaviours are predictive of interaction.

Acknowledgements

We thank Joe Blitzstein, Susan Holmes, Art Owen and Andrew Thomas for helpful remarks and encouragement. We benefited from many helpful comments by the journal's reviewers, who brought Butts (2008) to our attention and pushed us to expand the data analysis section. The work was supported in part by the Army Research Office under 'Presidential early career award for scientists and engineers' W911NF-09-1-0555, the Army Research Office under 'Multidisciplinary university research initiative' award 58153-MA-MUR and the Royal Society under a Wolfson Research Merit Award.

Appendix A: Implementation

To compute the MPLE, we use Newton's method as described in Boyd and Vandenberghe (2004). This requires an efficient algorithm for computing the gradient and Hessian of the log-partial-likelihood. For simplicity, we describe the case of strictly pairwise interactions with no ties in the interaction times. We use the notation from Section 2, with model (1) and partial likelihood (2). Recall that $x_t(i, j)$ is in \mathbb{R}^p . Assume that $|\mathcal{I}| = I$ and $|\mathcal{J}| = J$.

Suppose that $(t_1, i_1, j_1), \dots, (t_n, i_n, j_n)$ is the sequence of observed interactions. Set $n(i) = \#\{i_m : i_m = i\}$. The partial likelihood factors into a product of terms, one for each sender:

$$\text{PL}_t(\beta) = \prod_{i \in \mathcal{I}} \text{PL}_t(\beta, i), \quad \text{PL}_t(\beta, i) = \prod_{\substack{i_m \leq t, \\ i_m = i}} \frac{w_{i_m}(\beta, i, j_m)}{W_{i_m}(\beta, i)}.$$

This factorization allows us to compute $\log\{\text{PL}_t(\beta)\}$ and its derivatives by computing the sender-specific terms in parallel and then adding them together.

The gradient and Hessian of the sender-specific log-partial-likelihood are respectively

$$\nabla[\log\{\text{PL}_t(\beta, i)\}] = \sum_{\substack{i_m \leq t, \\ i_m = i}} x_{i_m}(i, j_m) - \sum_{\substack{i_m \leq t, \\ i_m = i}} E_{i_m}(\beta, i), \quad (10a)$$

$$-\nabla^2[\log\{\text{PL}_t(\beta, i)\}] = \sum_{\substack{i_m \leq t, \\ i_m = i}} V_{i_m}(\beta, i), \quad (10b)$$

where $E_t(\beta, i)$ and $V_t(\beta, i)$ are as defined in equations (5a) and (5b). When $x_t(i, j)$ is constant over time, sufficient statistics for β imply that these formulae simplify. Otherwise, computing the first two derivatives of $\log\{\text{PL}_{t_n}(\beta)\}$ necessitates iterating over all messages, potentially requiring time $\mathcal{O}(nJp^2)$. For small to medium sized data sets, this is manageable, but for large network data sets it can become prohibitive. In what follows we show how to exploit sparsity to reduce the computation time drastically.

A.1. Initial values

We shall need to compute $W_0(\beta, i)$, $w_0(\beta, i, j)$, $E_0(\beta, i)$ and $V_0(\beta, i)$ for all values of i and j . In the worst case, doing so will take $\mathcal{O}(IJP^2)$. However, often the senders belong to a small number, $\bar{I} \ll I$, of groups such that if, i and i' are in the same group, then the corresponding values of W_0 , π_0 , E_0 and V_0 are the same, reducing the total complexity to $\mathcal{O}(\bar{I}JP^2)$. The remaining complexity estimates assume that the initial values have all been precomputed.

A.2. Exploiting sparsity

We first decompose x into its static (non-time-varying) and dynamic parts as follows:

$$x_t(i, j) = x_0(i, j) + \Delta x_t(i, j). \quad (11)$$

Typically, we can quickly compute the dynamic part $\Delta x_t(i, j)$ at each observed message time by incrementally updating it. Further, $\Delta x_t(i, j)$ is 0 for most (i, j) pairs—often $\Delta x_t(i, j)$ is 0 unless i and j have a common acquaintance or they have interacted in the past. For convenience, set $\mathcal{J}_0(i) = \mathcal{J}$. Let

$$\bar{\mathcal{J}}(i) = \{j \in \mathcal{J} : j \in \mathcal{J}_t(i) \text{ and } \Delta x_t(i, j) \neq 0 \text{ for some } t\} \cup \{j \in \mathcal{J} : j \notin \mathcal{J}_t(i) \text{ for some } t\}.$$

For fixed t and i , assume that computing $\Delta x_t(i, j)$ for all values of j takes amortized time $\mathcal{O}(d\bar{J})$.

Since $\mathcal{J}_0(i) = \mathcal{J}$, we have that

$$\begin{aligned} w_t(\beta, i, j) &= w_0(\beta, i, j) \exp\{\beta^T \Delta x_t(i, j)\} \mathbf{1}\{j \in \mathcal{J}_t(i)\} \\ &= w_0(\beta, i, j) + \Delta w_t(i, j), \\ W_t(\beta, i) &= W_0(\beta, i) + \sum_{j \in \bar{\mathcal{J}}(i)} \Delta w_t(i, j), \end{aligned}$$

where

$$\Delta w_t(i, j) = w_0(\beta, i, j)[\exp\{\beta^T \Delta x_t(i, j)\} \mathbf{1}\{j \in \mathcal{J}_t(i)\} - 1];$$

here we have used that $\Delta w_t(i, j)$ is 0 unless $j \in \bar{\mathcal{J}}(i)$. Write

$$\pi_t(\beta, i, j) = w_t(\beta, i, j) / W_t(\beta, i);$$

then, defining

$$\begin{aligned} \gamma_t(i) &= W_0(\beta, i) / W_t(\beta, i), \\ \Delta\pi_t(\beta, i, j) &= \Delta w_t(\beta, i, j) / W_t(\beta, i), \end{aligned}$$

we can express $\pi_t(\beta, i, j)$ as

$$\pi_t(\beta, i, j) = \gamma_t(i) \pi_0(\beta, i, j) + \Delta\pi_t(\beta, i, j).$$

Moreover, given the initial values $W_0(\beta, i)$ and $w_0(\beta, i, j)$, we can efficiently keep track of $\gamma_t(i)$ and $\Delta\pi_t(\beta, i, j)$: for any i and t , it takes amortized time $\mathcal{O}(\bar{J}dp)$ to evaluate $\gamma_t(i)$ and all values of $\Delta\pi_t(i, j)$ as j varies.

A.3. Computing the gradient

In evaluating the gradient of the log-partial-likelihood as given by equation (10a), the sum $\sum_m x_{im}(i, j_m)$ can be computed in time $\mathcal{O}(np)$, whereas the computationally expensive term is $\sum_m E_{im}(\beta, i_m)$. In what follows we show how to exploit sparsity in x to reduce the associated computational overhead.

To simplify the notation, we suppress the dependence of all quantities on β and i . Consider π_t and $\Delta\pi_t$ to be vectors of length J , and write

$$\pi_t = \gamma_t \pi_0 + \Delta\pi_t.$$

Also, let $X_t = X_t(i)$ and $\Delta X_t = \Delta X_t(i)$ be the $J \times p$ matrices whose j th rows are $x_t(i, j)$ and $\Delta x_t(i, j)$ respectively, so that

$$X_t = X_0 + \Delta X_t.$$

Using these expressions, we obtain

$$E_t = X_t^T \pi_t = \gamma_t E_0 + X_0^T \Delta\pi_t + \Delta X_t^T \pi_t,$$

and thus

$$\sum_{i_m=i}^m E_{im} = \left(\sum_{i_m=i}^m \gamma_{i_m} \right) E_0 + X_0^T \left(\sum_{i_m=i}^m \Delta\pi_{i_m} \right) + \sum_{i_m=i}^m \Delta X_{i_m}^T \pi_{i_m}.$$

Taking advantage of the sparsity in ΔX_t and $\Delta\pi_t$, computing the three sums on the right-hand side takes time $\mathcal{O}\{n(i)\bar{J}dp\}$. Once the sums are known, the multiplication $(\sum \gamma_{i_m})E_0$ takes time $\mathcal{O}(p)$, and the multiplication $X_0^T(\sum \Delta\pi_{i_m})$ takes time $\mathcal{O}\{\bar{J}p\}$. Thus, we can compute $\sum_{i_m=i}^m E_{im}$ in time $\mathcal{O}\{n(i)\bar{J}dp\}$. Computing these terms separately for each i and then summing over all i to obtain the total gradient requires time $\mathcal{O}(n\bar{J}dp + Ip)$.

A.4. Computing the Hessian

Computing the Hessian according to equation (10b) proceeds similarly to the case of the gradient. We need to compute the sum $\sum_m V_{im}(\beta, i_m)$ efficiently; whereas a naive computation requires time $\mathcal{O}(nJp^2)$, this can be significantly improved by exploiting sparsity in $x_t(i, j)$.

For this, define $\Pi_t(\beta, i)$ to be the $J \times J$ diagonal matrix with $(\Pi_t(\beta, i))_{jj} = \pi_t(\beta, i, j)$, and set $\Delta\Pi_t(\beta, i) = \Pi_t(\beta, i) - \Pi_0(\beta, i)$. Suppressing the dependence on β and i , we have

$$\begin{aligned} V_t &= X_t^T (\Pi_t - \pi_t \pi_t^T) X_t \\ &= X_0^T (\Pi_t - \pi_t \pi_t^T) X_0 + \Delta X_t^T (\Pi_t - \pi_t \pi_t^T) X_0 + X_0^T (\Pi_t - \pi_t \pi_t^T) \Delta X_t + \Delta X_t^T (\Pi_t - \pi_t \pi_t^T) \Delta X_t. \end{aligned}$$

The first of these terms reduces to

$$X_0^T (\Pi_t - \pi_t \pi_t^T) X_0 = \gamma_t V_0 + \gamma_t (1 - \gamma_t) E_0 E_0^T - E_0 (\gamma_t \Delta\pi_t)^T X_0^T - X_0 (\gamma_t \Delta\pi_t) E_0^T + X_0^T (\Delta\Pi_t - \Delta\pi_t \Delta\pi_t^T) X_0,$$

and the second can be expressed as

$$\Delta X_t^\top (\Pi_t - \pi_t \pi_t^\top) X_0 = (\gamma_t \Delta X_t \pi_t) E_0^\top + \Delta X_t^\top (\Pi_t + \pi_t \pi_t^\top) X_0.$$

The third term is the transpose of the second; the fourth does not simplify.

To compute the sum $\sum_{i_m=i}^m V_{i_m}$, we only accumulate sums of terms that change with time: γ_t , $\Delta \pi_t$, $\gamma_t(1 - \gamma_t)$, $\gamma_t \Delta \pi_t$, $\Delta \pi_t \Delta \pi_t^\top$, $\gamma_t \Delta X_t \pi_t$, $\Delta X_t^\top (\Pi_t + \pi_t \pi_t^\top)$ and $\Delta X_t^\top (\Pi_t - \pi_t \pi_t^\top) \Delta X_t$. Doing so takes time $\mathcal{O}(\bar{J} d p^2)$ for each time increment. As with the gradient computation, we compute the sums separately for each i and then sum over all i , so that the total computation time is $\mathcal{O}(n \bar{J} d p^2 + I p^2)$.

A.5. Total computation time

To perform one Newton step in maximization of the log-partial-likelihood (2), we must first compute the gradient and Hessian of the log-partial-likelihood at the current value of β , and then compute the inverse of the Hessian and its product with the gradient. Once we have the Hessian, computing its inverse takes time $\mathcal{O}(p^3)$. Typically, it takes $\mathcal{O}(1)$ Newton steps to compute the maximum of a convex function (the constant is often below 30). The key factors in determining the computation time by using the factors laid out above are \bar{I} , \bar{J} and d .

- (a) The value of \bar{I} depends on the structure of $x_0(i, j)$. Specifically, \bar{I} is equal to the number of distinct values of the matrix $X_0(i)$ as i varies. For the Enron data, we have that $\bar{I} = 12$: each sender belongs to one of 12 groups determined by group (L, T or O), seniority (J or S) and gender (F or M), and so the matrix $X_0(i)$ depends only on the group of i .
- (b) The value of \bar{J} depends on the sparsity of $x_t(i, j)$. If $x_t(i, j)$ includes only dyadic network effects, then \bar{J} will typically be of size $\mathcal{O}(1)$ or $\mathcal{O}(J^\alpha)$ for a fractional value α ; when we add triadic effects, this size will typically grow to at most $\mathcal{O}(J^{2\alpha})$.
- (c) The value of d depends on further structure in $x_t(i, j)$. In our implementation, $d = \mathcal{O}(1)$ for dyadic effects and $d = \mathcal{O}(J)$ for triadic effects.

The total computational cost per Newton step is thus $\mathcal{O}(\bar{I} \bar{J} p^2 + n \bar{J} d p^2 + I p^2 + p^3)$, with the significance of this expression being that it is nearly linear in I, J and n . Thus, the algorithm scales naturally to large data sets.

Appendix B: Results from Section 3

B.1. Proof of theorem 1

Observe that the process $N_t(i, j)$ has compensator $\Lambda_t(i, j) = \int_0^t \lambda_s(i, j) ds$; similarly, processes $N_t(i)$ and N_t have compensators $\Lambda_t(i) = \sum_{j \in \mathcal{J}} \Lambda_t(i, j)$ and $\Lambda_t = \sum_{i \in \mathcal{I}} \Lambda_t(i)$. Correspondingly, define local martingales $M_t(i, j) = N_t(i, j) - \Lambda_t(i, j)$, $M_t(i) = N_t(i) - \Lambda_t(i)$ and $M_t = N_t - \Lambda_t$; also define

$$H_t(i, j) = x_t(i, j) - E_t(\beta_0, i),$$

where $E_t(\beta, i)$ is as defined in equation (4a).

As observed by Andersen and Gill (1982), the score function $U_t(\cdot)$ evaluated at β_0 has a simple representation in terms of these processes:

$$\begin{aligned} U_t(\beta_0) &= \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \int_0^t H_s(i, j) dN_s(i, j) \\ &= \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \int_0^t H_s(i, j) dM_s(i, j), \end{aligned}$$

since $\sum_{j \in \mathcal{J}} \int_0^t H_s(i, j) d\Lambda_s(i, j) = 0$. Since, by assumption 1, x is uniformly bounded, H is as well. Each term in the sum above is thus locally square integrable, with predictable covariation

$$\begin{aligned} \left\langle \int H_s(i, j) dM_s(i, j), \int H_s(i', j') dM_s(i', j') \right\rangle_t &= \int_0^t H_s(i, j) \otimes H_s(i', j') d\langle M(i, j), M(i', j') \rangle_s \\ &= \int_0^t (H_s(i, j))^{\otimes 2} d\Lambda_s(i, j) \mathbf{1}\{i=i', j=j'\} \end{aligned}$$

(Fleming and Harrington (1991), theorem 2.4.3). There is a sequence of stopping times localizing all $M(i, j)$ simultaneously, so $U(\beta_0)$ is locally square integrable with predictable variation

$$\begin{aligned} \langle U(\beta_0) \rangle_t &= \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \int_0^t (H_s(i, j))^{\otimes 2} d\Lambda_s(i, j) \\ &= \sum_{i \in \mathcal{I}} \int_0^t V_s(\beta_0, i) d\Lambda_s(i). \end{aligned} \tag{12}$$

Now we rescale time. For each positive n define a discretized time-scaled version of the score that is right continuous with limits from the left. The process is defined for times α in $[0, 1]$; between times in $[k/n, (k + 1)/n)$, it takes the value U_k , i.e.

$$\tilde{U}_\alpha^{(n)}(\beta) = U_{\lfloor n\alpha \rfloor}(\beta). \tag{13}$$

To prove part (a), lemma 1 shows that $\tilde{U}_\alpha^{(n)}(\beta_0)$ is a square integrable martingale adapted to $\tilde{\mathcal{F}}_\alpha^{(n)} = \mathcal{F}_{\lfloor n\alpha \rfloor}$, the σ -algebra of events before $t_{\lfloor n\alpha \rfloor}$. Since it depends only on values at jump times, the quadratic variation of $\tilde{U}^{(n)}(\beta_0)$ at time α is equal to the quadratic variation of $U(\beta_0)$ at time $t_{\lfloor n\alpha \rfloor}$. Therefore, since quadratic and predictable variation have the same limit when it exists (Rebolledo (1980), proposition 1), assumption 2 implies that $\langle (1/\sqrt{n})\tilde{U}^{(n)}(\beta_0) \rangle_\alpha \rightarrow^P \Sigma_\alpha(\beta_0)$. Lemma 2 in turn verifies that $(1/\sqrt{n})\tilde{U}^{(n)}(\beta_0)$ satisfies a Lindeberg condition necessary for the application of Rebolledo’s (1980) martingale central limit theorem. Thus the process converges in distribution to a Gaussian process with covariance function $\Sigma_\alpha(\beta_0)$ as claimed.

To prove part (b), recalling that $M_t(i) = N_t(i) - \Lambda_t(i)$, combine equations (5b) and (12) to obtain the relationship

$$\sum_i \int_0^{t_{\lfloor n\alpha \rfloor}} V_s(\beta_0, i) dM_s(i) = I_{t_{\lfloor n\alpha \rfloor}}(\beta_0) - \langle \tilde{U}^{(n)}(\beta_0) \rangle_\alpha. \tag{14}$$

When $\alpha \in [0, 1]$, a repeated application of the triangle inequality to

$$\left\| \frac{1}{n} I_{t_{\lfloor n\alpha \rfloor}}(\hat{\beta}_n) - \frac{1}{n} \{I_{t_{\lfloor n\alpha \rfloor}}(\beta_0) - I_{t_{\lfloor n\alpha \rfloor}}(\beta_0)\} - \Sigma_\alpha(\beta_0) \right\|$$

using relationship (14) yields

$$\begin{aligned} \left\| \frac{1}{n} I_{t_{\lfloor n\alpha \rfloor}}(\hat{\beta}_n) - \Sigma_\alpha(\beta_0) \right\| &\leq \left\| \frac{1}{n} \sum_i \int_0^{t_{\lfloor n\alpha \rfloor}} \{V_s(\hat{\beta}_n, i) - V_s(\beta_0, i)\} dN_s(i) \right\| \\ &\quad + \left\| \frac{1}{n} \sum_i \int_0^{t_{\lfloor n\alpha \rfloor}} V_s(\beta_0, i) dM_s(i) \right\| + \left\| \frac{1}{n} \sum_i \int_0^{t_{\lfloor n\alpha \rfloor}} V_s(\beta_0, i) d\Lambda_s(i) - \Sigma_\alpha(\beta_0) \right\|. \end{aligned}$$

We show that all three terms converge to 0 in probability. The first term above is uniformly bounded by $\sup_{n', i} \|V_{n'}(\hat{\beta}_n, i) - V_{n'}(\beta_0, i)\|$, which converges to 0 since $\hat{\beta}_n \rightarrow^P \beta_0$ by hypothesis of the theorem and $\{V_{n'}(\cdot, i)\}$ is an equicontinuous family by assumption 4. Lemma 3 proves, as a consequence of assumption 3 and Lenglart’s (1977) inequality, that the second term converges to 0 uniformly in α . The third term converges to 0 by assumption 2, thereby concluding the proof.

B.2. Supporting lemmas for theorem 1

Lemma 1. Using the notation of theorem 1, under assumption 1 the process $\tilde{U}_\alpha^{(n)}(\beta_0)$ from equation (13) is a square integrable martingale adapted to $\tilde{\mathcal{F}}_\alpha^{(n)} = \mathcal{F}_{\lfloor n\alpha \rfloor}$.

Proof. The conditional expectation property holds provided that $\mathbb{E}[U_{t_n}(\beta_0) | \mathcal{F}_{t_{n-1}}] = U_{t_{n-1}}(\beta_0)$. Define $K = \sup_{t, i, j} \|x_t(i, j)\|$. Note that $\|H_t(i, j)\| \leq 2K$. Thus,

$$\begin{aligned} \|U_{t \wedge t_n}(\beta_0)\| &\leq 2K(N_{t \wedge t_n} + \Lambda_{t \wedge t_n}), \\ \mathbb{E}[\sup_t \|U_{t \wedge t_n}(\beta_0)\|^2] &\leq 8 \mathbb{E}[K^2]^{1/2} (\mathbb{E}[N_{t_n}^2] + \mathbb{E}[\Lambda_{t_n}^2])^{1/2}. \end{aligned}$$

By assumption 1, $\mathbb{E}[K^2]$ is finite and, by construction, N_{t_n} is bounded. Since $N_{t \wedge t_n}$ is a counting process, $\mathbb{E}[\Lambda_{t_n}^2]$ is finite, also (this follows from results in section 2.3 of Fleming and Harrington (1991)). Thus, $U_{t \wedge t_n}(\beta_0)$ is uniformly integrable. The optional sampling theorem now applies to give the conditional expectation property of $\tilde{U}^{(n)}(\beta_0)$. For square integrability, note that $\sup_{1 \leq m \leq n} \mathbb{E}\|U_{t_m}\|^2 \leq \mathbb{E}[\sup_t \|U_{t \wedge t_n}(\beta_0)\|^2]$.

Lemma 2. Using the notation of theorem 1, under assumption 1, the Lindeberg condition for Rebolledo’s (1980) central limit theorem is satisfied: for any positive ε ,

$$\frac{1}{n} \sum_{i,j} \int_0^{t_n} \|H_s(i, j)\|^2 \mathbf{1}\{\|H_s(i, j)\| > \sqrt{n\varepsilon}\} d\Lambda_s(i, j) \xrightarrow{P} 0.$$

Proof. With $K = \sup_{t,i,j} \|x_t(i, j)\|$ as above, the integral is bounded by $4K^2 \mathbf{1}\{n^{-1/2}K > \varepsilon/2\} \Lambda_{t_n}/n$. Since $\mathbb{E}[K^2] < \infty$ by assumption 1, the first term converges to 0 in probability. Since $\mathbb{E}[\Lambda_{t_n}] = \mathbb{E}[N_{t_n}] = n$, the product of the two also converges to 0 in probability. Thus, the Lindeberg condition is satisfied.

Lemma 3. Using the notation of theorem 1, under assumptions 1 and 3 we have that

$$\left\| \frac{1}{n} \sum_i \int_0^{t_{[an]}} V_s(\beta_0, i) dM_s(i) \right\| \xrightarrow{P} 0$$

uniformly in α .

Proof. Lengart’s (1977) inequality and assumption 3 imply that, for any positive ρ and δ ,

$$\mathbb{P}\left\{ \sup_{t \in [0, t_n]} \left\| \frac{1}{n} \sum_i \int_0^t V_s(\beta_0, i) dM_s(i) \right\| \geq \rho \right\} \leq \frac{\delta}{\rho^2} + \mathbb{P}\left\{ \frac{1}{n^2} \sum_i \int_0^{t_n} \|V_s(\beta_0, i)\|^2 d\Lambda_s(i) \geq \delta \right\}$$

(see Fleming and Harrington (1991), corollary 3.4.1) for a related proof). As in the proof of lemma 1, set $K = \sup_{t,i,j} \|x_t(i, j)\|$. The sum is bounded by $(16K^4/n)\Lambda_{t_n}/n$. Since $n^{-1/2}K^2 \xrightarrow{P} 0$ by assumption 1 and $\mathbb{E}[\Lambda_{t_n}] = n$, the right-hand side of the inequality converges to δ/ρ^2 . Since δ is arbitrary, the right-hand side must converge to 0.

B.3. Proof of theorem 2

We follow Haberman’s (1977) approach to proving consistency, which relies on Kantorovich’s (1952) analysis of Newton’s method. Tapia (1971) has given an elementary proof of the Kantorovich theorem. We state a weak form of the result as a lemma.

Lemma 4 (Kantorovich theorem). Let $P(x) = 0$ be a general system of non-linear equations, where P is a map between two Banach spaces. Let $P'(x)$ denote the Jacobian (Fréchet differential) of P at x , assumed to exist in D_0 , a convex open neighbourhood of x_0 . Assume that

- (a) $\|P'(x_0)^{-1}\| \leq B$,
- (b) $\|P'(x_0)^{-1}P(x_0)\| \leq \eta$,
- (c) $\|P'(x) - P'(y)\| \leq K\|x - y\|$, for all x and y in D_0 ,

with $h = BK\eta \leq \frac{1}{2}$.

Let $\Omega_* = \{x : \|x - x_0\| \leq 2\eta\}$. If $\Omega_* \subset D_0$, then the Newton iterates, $x_{k+1} = x_k - P'(x_k)^{-1}P(x_k)$, are well defined, remain in Ω_* and converge to x^* in Ω_* such that $P(x^*) = 0$. In addition,

$$\|x^* - x_k\| \leq \frac{\eta}{h} \frac{(2h)^{2^k}}{2^k}, \quad k = 0, 1, 2, \dots$$

B.3.1. Proof of theorem 2

Set $U_t(\cdot)$ and $I_t(\cdot)$ to be the gradient and negative Hessian of the log-partial-likelihood, as defined in equations (5a) and (5b). Since $I_t(\beta)$ is a sum of rank 1 matrices with positive weights, it is positive semidefinite, and $\log\{PL_t(\cdot)\}$ is a concave function. By the assumption that the smallest eigenvalue of $\Sigma_1(\cdot)$ is bounded away from 0 in a neighbourhood of β_0 , for n sufficiently large, if $\log\{PL_t(\cdot)\}$ has a local maximum in that neighbourhood then it must be the unique global maximum.

We find the local maximum by applying Newton’s method to the gradient of $(1/n)\log\{PL_{t_n}(\cdot)\}$ taking β_0 as the initial iterate. Define

$$Z_n = - \left\{ \frac{1}{n} I_{t_n}(\beta_0) \right\}^{-1} \frac{1}{n} U_{t_n}(\beta_0).$$

The first Newton iterate, $\beta_{n,1}$, is equal to $\beta_0 - Z_n$. Part (b) of theorem 1 and the assumptions of theorem 2 imply that $\{(1/n)I_{t_n}(\beta_0)\}^{-1}$ exists for n sufficiently large, so that Z_n is well defined. Moreover, part (a) of theorem 1 and Slutsky’s theorem imply that $Z_n \xrightarrow{P} 0$ and $\sqrt{n}Z_n \xrightarrow{d} \mathcal{N}\{0, \Sigma_1(\beta_0)^{-1}\}$.

Now we may apply Kantorovich’s theorem to bound $\|\hat{\beta}_n - \beta_0\|$ and $\|\hat{\beta}_n - \beta_{n,1}\|$ as follows. By assumption, there is a neighbourhood of β_0 , say D_0 , and finite K and B , such that $\|(1/n)I_n(\beta) - (1/n)I_n(\beta')\| \leq K\|\beta - \beta'\|$ and $\|(1/n)I_n(\beta_0)^{-1}\| \leq B$ for $\beta, \beta' \in D_0$. Define $\eta_n = \|Z_n\|$ and $h_n = BK\eta_n$, noting that h_n and η_n are size $\mathcal{O}_P(n^{-1/2})$. Thus, for n sufficiently large,

- (a) $\|\hat{\beta}_n - \beta_0\| \leq 2\eta_n \rightarrow^P 0$,
- (b) $\|\hat{\beta}_n - (\beta_0 - Z_n)\| \sqrt{n} \leq 2\sqrt{n}\eta_n h_n \rightarrow^P 0$.

Thus, $\hat{\beta}_n \rightarrow^P \beta_0$, and $(\hat{\beta}_n - \beta_0)\sqrt{n}$ and $Z_n\sqrt{n}$ converge weakly to the same limit.

Appendix C: Results from Section 4

C.1. Proof of theorem 3

When $J \subseteq \mathcal{J}_t(i)$, set $X_t(i, J) = \sum_{j \in J} x_t(i, j)$ and $w_t(\beta, i, J) = \exp\{\beta^T X_t(i, J)\}$. As a slight abuse of notation, when j is an element of $\mathcal{J}_t(i)$, take ‘ $w_t(\beta, i, j)$ ’ to mean $w_t(\beta, i, \{j\})$. Define weights

$$W_t(\beta, i; L) = \sum_{\substack{J \subseteq \mathcal{J}_t(i), \\ |J|=L}} w_t(\beta, i, J),$$

$$\tilde{W}_t(\beta, i; L) = \left\{ \sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j) \right\}^L,$$

and note that the approximation error in $\log\{\tilde{\text{PL}}_t(\beta)\}$ comes from replacing W with \tilde{W} . The gradients of the weights are

$$E_t(\beta, i; L) = \nabla[\log\{W_t(\beta, i; L)\}] = \frac{1}{W_t(\beta, i; L)} \sum_{\substack{J \subseteq \mathcal{J}_t(i), \\ |J|=L}} w_t(i, J) X_t(i, J),$$

$$\tilde{E}_t(\beta, i; L) = \nabla[\log\{\tilde{W}_t(\beta, i; L)\}] = L \frac{\sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j) x_t(i, j)}{\sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j)}.$$

The second is the expectation of $\sum_{i=1}^L x_t(i, j_i)$ when j_1, \dots, j_L are drawn independently and identically from $\mathcal{J}_t(i)$ with weights $w_t(\beta, i, \cdot)$; the first is the same expectation, conditional on the event that j_1, \dots, j_L are all unique. Let $\mathbb{P}_{t,\beta,i;L}$ and $\tilde{\mathbb{P}}_{t,\beta,i;L}$ denote the two probability laws for j_1, \dots, j_L , and let $\tilde{\mathbb{E}}_{t,\beta,i;L}$ and $\mathbb{E}_{t,\beta,i;L}$ denote expectations with respect to them, so that $E_t(\beta, i; L) = \mathbb{E}_{t,\beta,i;L}[\sum_{i=1}^L x_t(i, j_i)]$ and $\tilde{E}_t(\beta, i; L) = \tilde{\mathbb{E}}_{t,\beta,i;L}[\sum_{i=1}^L x_t(i, j_i)]$.

The bound on $\nabla[\log\{\text{PL}_m(\beta)\}] - \nabla[\log\{\tilde{\text{PL}}_m(\beta)\}]$ derives from a bound on $E_t(\beta, i; L) - \tilde{E}_t(\beta, i; L)$. Write

$$E_t(\beta, i; L) - \tilde{E}_t(\beta, i; L) = \mathbb{E}_{t,\beta,i;L} \left[\sum_{i=1}^L x_t(i, j_i) \right] - \tilde{\mathbb{E}}_{t,\beta,i;L} \left[\sum_{i=1}^L x_t(i, j_i) \right].$$

We define probability law $\mathbb{P}_{t,\beta,i;L}^*$ and associated random variables j_1, \dots, j_L and $\tilde{j}_1, \dots, \tilde{j}_L$, such that marginally j_1, \dots, j_L are distributed according to $\mathbb{P}_{t,\beta,i;L}$ and $\tilde{j}_1, \dots, \tilde{j}_L$ are distributed according to $\tilde{\mathbb{P}}_{t,\beta,i;L}$, but the variables are coupled to have non-trivial chance of agreeing. Then,

$$\|E_t(\beta, i; L) - \tilde{E}_t(\beta, i; L)\| = \left\| \mathbb{E}_{t,\beta,i;L}^* \left[\sum_{i=1}^L x_t(i, j_i) - \sum_{i=1}^L x_t(i, \tilde{j}_i) \right] \right\|$$

$$\leq 2L \left[\sup_{j \in \mathcal{J}_t(i)} \|x_t(i, j)\| \right] \mathbb{P}_{t,\beta,i;L}^* \{(j_1, \dots, j_L) \neq (\tilde{j}_1, \dots, \tilde{j}_L)\}.$$

The coupling is as follows.

- (a) Draw $(\tilde{j}_1, \dots, \tilde{j}_L)$ according to $\tilde{\mathbb{P}}_{t,\beta,i;L}$.
- (b) If $(\tilde{j}_1, \dots, \tilde{j}_L)$ are all unique, set $(j_1, \dots, j_L) = (\tilde{j}_1, \dots, \tilde{j}_L)$; otherwise draw (j_1, \dots, j_L) independently according to $\mathbb{P}_{t,\beta,i;L}$.

With $K = \sup_{j \in \mathcal{J}_t(i)} \|x_t(i, j)\|$, lemma 5 shows that

$$\mathbb{P}_{t, \beta, i; L}^* \{(j_1, \dots, j_L) \neq (\tilde{j}_1, \dots, \tilde{j}_L)\} \leq \binom{L}{2} \frac{\exp(4K\|\beta\|)}{|\mathcal{J}_t(i)|}.$$

The resulting bound on $\|\nabla[\log\{\mathbf{PL}_t(\beta)\}] - \nabla[\log\{\widetilde{\mathbf{PL}}_t(\beta)\}]\|$ now follows by expressing

$$\nabla[\log\{\widetilde{\mathbf{PL}}_t(\beta)\}] - \nabla[\log\{\mathbf{PL}_t(\beta)\}] = \sum_{i_m \leq t} E_{i_m}(\beta, i_m; |J_m|) - \tilde{E}_{i_m}(\beta, i_m; |J_m|).$$

Using

$$\|E_t(\beta, i; L) - \tilde{E}_t(\beta, i; L)\| \leq KL^2(L-1) \frac{\exp(4K\|\beta\|)}{|\mathcal{J}_t(i)|},$$

we obtain

$$\|\nabla[\log\{\widetilde{\mathbf{PL}}_t(\beta)\}] - \nabla[\log\{\mathbf{PL}_t(\beta)\}]\| \leq K \exp(4K\|\beta\|) \sum_{i_m \leq t} \frac{|J_m|^2(|J_m| - 1)}{|\mathcal{J}_m(i_m)|}.$$

We obtain the final bound for the gradients by replacing the numerators of the summands with $\sup_m |J_m|$.

Using the same methods, lemma 6 derives the bound on the difference in Hessians.

C.2. Supporting lemmas for theorem 2

Lemma 5. Using the notation and assumptions of theorem 3,

$$\mathbb{P}_{t, \beta, i; L}^* \{(j_1, \dots, j_L) \neq (\tilde{j}_1, \dots, \tilde{j}_L)\} \leq \binom{L}{2} \frac{\exp(4K\|\beta\|)}{|\mathcal{J}_t(i)|},$$

where $K = \sup_t \|x_t(i, j)\|$.

Proof. The left-hand side is bounded by the probability that the samples $\tilde{j}_1, \dots, \tilde{j}_L$ are all unique, which can be bounded by

$$\sum_{k < l} \mathbb{P}_{t, \beta, i; L}^* \{\tilde{j}_k = \tilde{j}_l\} = \binom{L}{2} \sum_{j \in \mathcal{J}_t(i)} \left\{ \frac{w_t(\beta, i, j)}{\sum_{j' \in \mathcal{J}_t(i)} w_t(\beta, i, j')} \right\}^2.$$

Note that $\exp(-K\|\beta\|) \leq w_t(\beta, i, j) \leq \exp(K\|\beta\|)$, so

$$\sum_{j \in \mathcal{J}_t(i)} \left\{ \frac{w_t(\beta, i, j)}{\sum_{j' \in \mathcal{J}_t(i)} w_t(\beta, i, j')} \right\}^2 \leq \frac{\exp(4K\|\beta\|)}{|\mathcal{J}_t(i)|}.$$

Lemma 6. Using the notation and assumptions of theorem 3,

$$\|\nabla^2[\log\{\widetilde{\mathbf{PL}}_t(\beta)\}] - \nabla^2[\log\{\mathbf{PL}_t(\beta)\}]\| \leq 2K^2 \exp(4K\|\beta\|) \sum_{i_m \leq t} \frac{|J_m|^3(|J_m| - 1)}{|\mathcal{J}_m(i_m)|}.$$

Proof. The argument is similar to the bound on the difference in gradients in the proof of theorem 3. The Hessians of the weights are

$$\begin{aligned} V_t(\beta, i; L) &= \nabla^2[\log\{W_t(\beta, i; L)\}] = \frac{1}{W_t(\beta, i; L)} \sum_{\substack{J \subseteq \mathcal{J}_t(i), \\ |J|=L}} w_t(\beta, i, J) (X_t(i, J) - E_t(\beta, i; L))^{\otimes 2}, \\ \tilde{V}_t(\beta, i; L) &= \nabla^2[\log\{\tilde{W}_t(\beta, i; L)\}] = L \frac{\sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j) (x_t(i, j) - (1/L)\tilde{E}_t(\beta, i; L))^{\otimes 2}}{\sum_{j \in \mathcal{J}_t(i)} w_t(\beta, i, j)}. \end{aligned}$$

The first is the covariance matrix of $\sum_{j=1}^L x_t(i, j)$ under $\mathbb{P}_{t, \beta, i; L}$; the second is the covariance matrix of the same quantity under $\mathbb{P}_{t, \beta, i; L}^*$. The result follows in the same manner as in the proof of theorem 3. The relevant intermediate bound is

$$\|V_i(\beta, i; L) - \tilde{V}_i(\beta, i; L)\| \leq 2K^2L^3(L-1) \frac{\exp(4K\|\beta\|)}{|\mathcal{I}_i(i)|}.$$

C.3. Proof of theorem 4

We know that Newton's method applied to $(1/n)\log\{\widetilde{\text{PL}}_{t_n}(\cdot)\}$ converges to $\tilde{\beta}_n$ after sufficiently many iterations. We employ $\hat{\beta}_n$ as the initial iterate and use the Kantorovich theorem (lemma 4) to bound $\|\tilde{\beta}_n - \hat{\beta}_n\|$.

In the notation of lemma 4, $P(\cdot)$ is the gradient of $(1/n)\log\{\widetilde{\text{PL}}_{t_n}(\cdot)\}$ and $P'(\cdot)$ is its Hessian. The conditions of theorem 4 imply that assumptions (a) and (c) hold uniformly in n for some finite B and K . Set

$$\eta_n = \left\| \left(\nabla^2 \left[\frac{1}{n} \log\{\widetilde{\text{PL}}_{t_n}(\hat{\beta}_n)\} \right] \right)^{-1} \nabla \left[\frac{1}{n} \log\{\widetilde{\text{PL}}_{t_n}(\hat{\beta}_n)\} \right] \right\|$$

and set $h_n = BK\eta_n$. Since $\nabla[\log\{\text{PL}_{t_n}(\hat{\beta}_n)\}] = 0$, theorem 3 and the boundedness of the inverse Hessian imply that $\eta_n = \mathcal{O}_P(G_n/n)$. Therefore, for n sufficiently large,

$$\|\tilde{\beta}_n - \hat{\beta}_n\| \leq \frac{\eta_n}{h} \frac{(2h)^{2^0}}{2^0} = 2\eta_n = \mathcal{O}_P\left(\frac{G_n}{n}\right).$$

References

- Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical Models based on Counting Processes*. New York: Springer.
- Andersen, P. K. and Gill, R. D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1100–1120.
- Anderson, C. J., Wasserman, S. and Crouch, B. (1999) A p^* primer: logit models for social networks. *Soc. Netw.*, **21**, 37–66.
- Aral, S., Muchnik, L. and Sundararajan, A. (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natn. Acad. Sci. USA*, **106**, 21544–21549.
- Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization*. Cambridge: Cambridge University Press.
- Breslow, N. E. (1974) Covariance analysis of censored survival data. *Biometrics*, **30**, 89–99.
- Broström, G. (2002) Cox regression; ties without tears. *Commun. Statist. Theor. Meth.*, **31**, 285–297.
- Butts, C. T. (2008) A relational event framework for social action. *Sociol. Methodol.*, **38**, 155–200.
- Cohen, W. W. (2009) Enron email dataset. (Available from <http://www.cs.cmu.edu/~enron/>.)
- Cook, R. J. and Lawless J. F. (2007) *The Statistical Analysis of Recurrent Events*. Berlin: Springer.
- Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187–220.
- Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.
- Eagle, N. and Pentland, A. S. (2006) Reality mining: sensing complex social systems. *Persnl Ubiquit. Comput.*, **10**, 255–268.
- Efron, B. (1977) The efficiency of Cox's likelihood function for censored data. *J. Am. Statist. Ass.*, **72**, 557–565.
- Fleming, T. R. and Harrington, D. P. (1991) *Counting Processes and Survival Analysis*. New York: Wiley.
- Fowler, J. H. (2006) Connecting the Congress: a study of cosponsorship networks. *Polit. Anal.*, **14**, 456–487.
- Goldenberg, A., Zheng, A. X., Fienberg, S. E. and Airoldi, E. M. (2009) A survey of statistical network models. *Found. Trends Mach. Learn.*, **2**, 129–233.
- Haberman, S. J. (1977) Maximum likelihood estimates in exponential response models. *Ann. Statist.*, **5**, 815–841.
- Jackson, M. O. (2008) *Social and Economic Networks*. Princeton: Princeton University Press.
- Kantorovich, L. V. (1952) Functional analysis and applied mathematics. (Engl. transl. C. D. Benster). *Report 1509*. National Bureau of Standards, Gaithersburg.
- Kolaczyk, E. D. (2009) *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.
- Lengart, E. (1977) Relation de domination entre deux processus. *Ann. Inst. H. Poincaré*, **13**, 171–179.
- Lunagómez, S., Mukherjee, S. and Wolpert, R. L. (2009) Geometric representations of hypergraphs for prior specification and posterior sampling. *Technical Report 2009-01*. Duke University, Durham.
- Martinussen, T. and Scheike, T. H. (2006) *Dynamic Regression Models for Survival Data*. New York: Springer.
- McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
- Mckenzie, D. and Rapoport, H. (2007) Network effects and the dynamics of migration and inequality: theory and evidence from Mexico. *J. Devlpmnt Econ.*, **84**, 1–24.
- McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001) Birds of a feather: homophily in social networks. *A. Rev. Sociol.*, **27**, 415–444.
- Nocedal, J. and Wright, S. J. (2006) *Numerical Optimization*, 2nd edn. New York: Springer.

- Papachristos, A. V. (2009) Murder by structure: dominance relations and the social structure of gang homicide. *Am. J. Sociol.*, **115**, 74–128.
- Rebolledo, R. (1980) Central limit theorems for local martingales. *Probab. Theor. Reltd Flds*, **51**, 269–286.
- Shafiei, M. and Chipman, H. (2010) Mixed membership stochastic block-models for transactional networks. In *Proc. 10th Int. Conf. Data Mining*, pp. 1019–1024. Piscataway: Institute of Electrical and Electronics Engineers Press.
- Snijders, T. A. B., Van de Bunt, G. V. and Steglich, C. E. G. (2010) Introduction to stochastic actor-based models for network dynamics. *Soc. Netwks*, **32**, 44–60.
- Sundaresan, S. R., Fischhoff, I. R., Dushoff, J. and Rubenstein, D. I. (2007) Network metrics reveal differences in social organization between two fission-fusion species, Grevy's zebra and onager. *Oecologia*, **151**, 140–149.
- Tapia, R. A. (1971) The Kantorovich theorem for Newton's method. *Am. Math. Monthly*, **78**, 389–392.
- Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990) Martingale-based residuals for survival models. *Biometrika*, **77**, 147–160.
- Therneau, T. and Lumley, T. (2009) survival: Survival analysis, including penalised likelihood. *R Package Version 2.35-8*. (Available from <http://CRAN.R-project.org/package=survival>.)
- Tyler, J. R., Wilkinson, D. M. and Huberman, B. A. (2005) E-mail as spectroscopy: automated discovery of community structure within organizations. *Inform. Soc.*, **21**, 143–153.
- Vu, D. Q., Asuncion, A., Hunter, D. and Smyth, P. (2011a) Continuous-time regression models for longitudinal networks. *Adv. Neurl Inform. Process. Syst.*, **24**, 2492–2500.
- Vu, D. Q., Asuncion, A. U., Hunter, D. R. and Smyth, P. (2011b) Dynamic egocentric models for citation networks. In *Proc. 28th Int. Conf. Machine Learning*, pp. 857–864. New York: Association for Computing Machinery.
- Wong, W. H. (1986) Theory of partial likelihood. *Ann. Statist.*, **14**, 88–123.
- Zhou, Y., Goldberg, M., Magdon-Ismail, M. and Wallace, W. A. (2007) Strategies for cleaning organizational emails with an application to Enron email dataset. In *Proc. 5th A. Conf. North American Association for Computational Social Organization Science*. Pittsburgh: North American Association for Computational Social and Organizational Science.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Point process modelling for directed interaction networks: Supplementary material'.