StatScale Seminars

Mutually exciting point process graphs for modelling dynamic networks

Imperial College London

Francesco Sanna Passino, Nick Heard
Department of Mathematics, Imperial College London
f.sannapassino@imperial.ac.uk

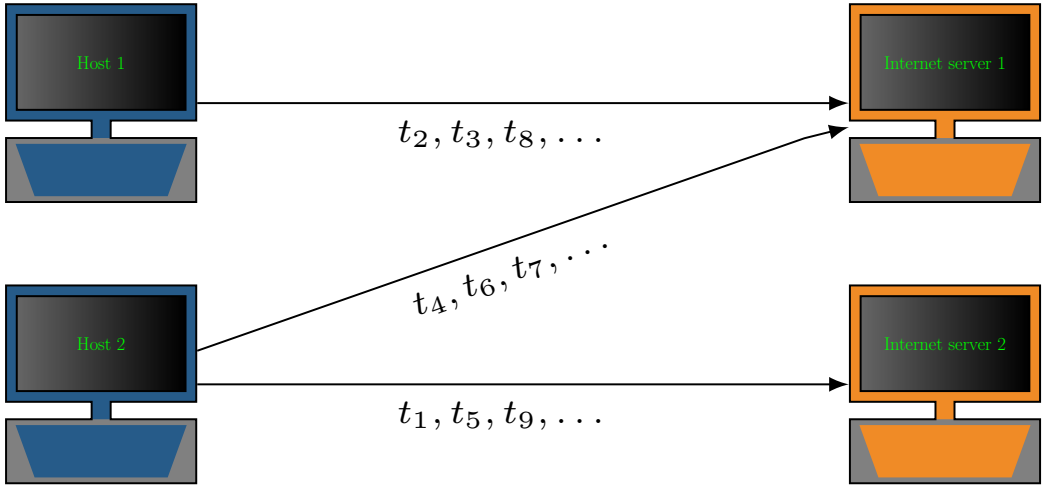
5th November, 2021

Dynamic graphs as point processes with dyadic marks

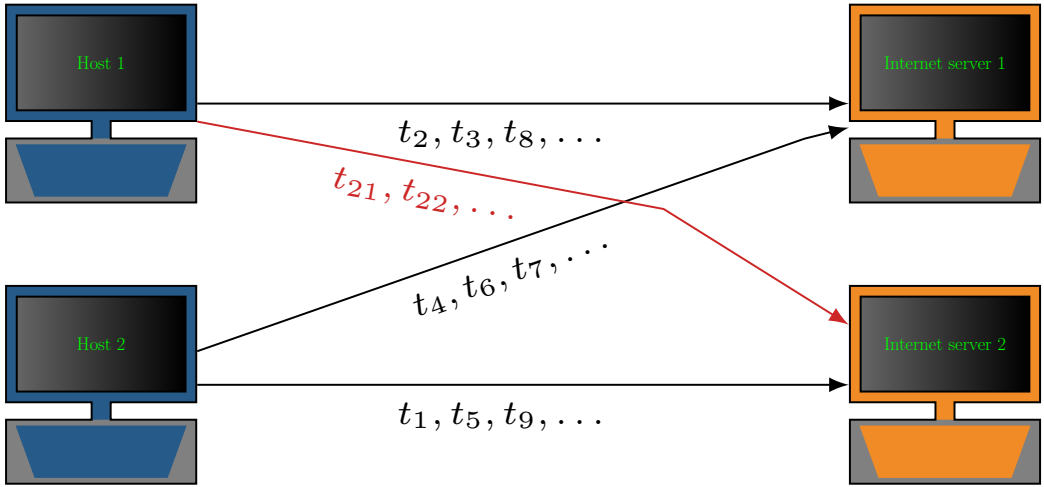
- Event data from dynamic networks are observed as triplets $(t_1, x_1, y_1), \ldots, (t_m, x_m, y_m)$, where $0 \leq t_1 \leq t_2 \leq \ldots$ are event times and the dyadic marks (x_k, y_k) denote the source and destination nodes, each belonging to a set of nodes $V = \{1, \ldots, n\}$ of size n.
- The sequence of graph edges $(x_1, y_1), \ldots, (x_m, y_m)$ induces a directed *network adjacency matrix* $\mathbf{A} = \{A_{ij}\} \in \{0, 1\}^{n \times n}$ where $A_{ij} = 1$ if node i connected to node j at least once during the entire observation period, and $A_{ij} = 0$ otherwise.

$$
\begin{matrix}(1.3, 1, 2)\\ (1.7, 3, 1)\\ (1.9, 1, 2)\\ (2.2, 3, 4)\\ (2.7, 3, 2)\\ \vdots\end{matrix}
\Rightarrow
\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array}
\Rightarrow \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}
$$

Motivation: new links in cyber-security



Motivation: new links in cyber-security



Background

- **Objective:** propose a model which can calculate *anomaly scores* for *unobserved marks*.
- **Motivation:** computer network attacks tend to form previously unobserved connections.
- *Related literature in cyber-security:* Price-Williams and Heard, 2020, demonstrate that *self-exciting processes* have an excellent performance for modelling individual edges.
- *Related methodology:* new link prediction in networks. Latent position models (LPMs, Hoff, Raftery, and Handcock, 2002) postulate that the probability of a link is a function of *node-specific* latent features:

$$\mathbb{P}(A_{ij} = 1 \mid \boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j),\ \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d,$$

  where $\kappa(\cdot)$ is a kernel function. Conditional on the latent positions, LPMs naturally admit calculations of link probabilities for unobserved links.
- **Model** proposed in this work: **mutually exciting process** on **each edge**, parametrised only by **node-specific parameters**.

Mutually exciting graphs (MEGs)

- **MEGs** are defined by a time-varying matrix of non-negative intensity functions $\boldsymbol{\lambda}(t) = \{\lambda_{ij}(t)\}$.
- Each entry is the intensity of the counting process $N_{ij}(t) = \sum_{k=1}^{m} \mathbb{1}_{[0,t]\times\{i\}\times\{j\}}(t_k, x_k, y_k)$ of events occurring on the edge (i, j): $\lambda_{ij}(t) = \lim_{\delta\to 0} \mathbb{E}[N_{ij}(t+\delta) - N_{ij}(t)|\mathcal{H}_t]$.
- For generality, it is assumed that for each edge (i, j) there exists a changepoint $\tau_{ij} \geq 0$ after which the edge becomes observable. In the simplest case, $\tau_{ij} = 0$ for all i, j.
- Each entry of $\lambda_{ij}(t)$ is represented as an **additive model** with three components:
  - The first, denoted $\alpha_i(t)$, characterises the process of arrival times involving i as **source** node;
  - The second, $\beta_j(t)$, corresponds to arrivals for which j is the **destination** node;
  - The third, $\gamma_{ij}(t)$, is an **interaction** term, also be parameterised by **node-specific parameters**.

$$\lambda_{ij}(t) = \alpha_i(t) + \beta_j(t) + \gamma_{ij}(t), \quad t \geq \tau_{ij}. \tag{1}$$

- The intensity function resembles the link function used in *additive and multiplicative effect network* models for network adjacency matrices (Hoff, 2021).

Main effects

- Define the source and destination counting processes as $N_i(t) = \sum_{k=1}^{m} \mathbb{1}_{[0,t]\times\{i\}}(t_k, x_k)$ and $N'_j(t) = \sum_{k=1}^{m} \mathbb{1}_{[0,t]\times\{j\}}(t_k, y_k)$.
- Let $\ell_{i1}, \ell_{i2}, \dots$ denote the event indices $\{k : x_k = i\}$ such that i appears as source node, and $\ell'_{j1}, \ell'_{j2}, \dots$ denote the event indices $\{k : y_k = j\}$ for which j is the destination node.
- To allow self-excitation of both source and destination nodes, the latent functions $\alpha_i(t)$ and $\beta_j(t)$ are assigned the following form:

$$\alpha_i(t) = \alpha_i + \sum_{k>N_i(t)-r}^{N_i(t)} \omega_i(t - t_{\ell_{ik}}), \quad \beta_j(t) = \beta_j + \sum_{k>N'_j(t)-r}^{N'_j(t)} \omega'_j(t - t_{\ell'_{jk}}),$$

  where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n), \boldsymbol{\beta} = (\beta_1, \dots, \beta_n) \in \mathbb{R}^n_+$ are node-specific baseline intensity levels, and ω_i, ω'_i are node-specific, non-increasing excitation functions from $\mathbb{R}_+$ to $\mathbb{R}_+$.
- For simplicity, the excitation functions assume the following **scaled exponential form**, for non-negative parameters $\boldsymbol{\mu}_i, \boldsymbol{\mu}'_j, \boldsymbol{\phi}_i, \boldsymbol{\phi}'_j \in \mathbb{R}^n_+$:

$$\omega_i(t) = \mu_i \exp\{-(\mu_i + \phi_i)t\}, \qquad \omega'_j(t) = \mu'_j \exp\{-(\mu'_j + \phi'_j)t\}.$$

Interactions

- Let $\ell_{ij1}, \ell_{ij2}, \ldots$ be the indices $\{k : x_k = i, y_k = j\}$ of events observed on the edge (i, j).
- The interaction term $\gamma_{ij}(t)$ in (1) assumes a similar form to the main effects, but with a background rate obtained as the **inner product** between node-specific baseline parameter vectors $\boldsymbol{\gamma}_i, \boldsymbol{\gamma}'_j \in \mathbb{R}^d_+$:

$$\gamma_{ij}(t) = \boldsymbol{\gamma}_i^\intercal \boldsymbol{\gamma}'_j + \sum_{k > N_{ij}(t) - r}^{N_{ij}(t)} \omega_{ij}(t - t_{\ell_{ijk}}),$$

- The inner product baseline is inspired by random dot product graphs (RDPGs; see, for example, Athreya et al., 2018) for link probabilities.
- For simplicity, the excitation function $\omega_{ij}(t)$ is expressed as a **sum of scaled exponentials**, parameterised by four node-specific, non-negative latent d-vectors $\boldsymbol{\nu}, \boldsymbol{\nu}'_j, \boldsymbol{\theta}_i, \boldsymbol{\theta}'_j \in \mathbb{R}^d_+$:

$$\omega_{ij}(t) = \sum_{\ell=1}^{d} \nu_{i\ell}\nu'_{j\ell} \exp\{-(\theta_{i\ell} + \nu_{i\ell})(\theta'_{j\ell} + \nu'_{j\ell})t\}.$$

MEGs: an example

- The integer parameter r expresses *how many events* are taken into account in the intensity function. There are three limiting cases:
  - $r = 0$: Poisson process (the process is independent of previous events);
  - $r = 1$: Markov process (dependence only on the distance to the last event);
  - $r \to \infty$: Hawkes process (dependence on the entire history of the process).

(a) Main effect $\alpha_i(t)$ (b) Main effect $\beta_j(t)$ (c) Interaction $\gamma_{ij}(t)$

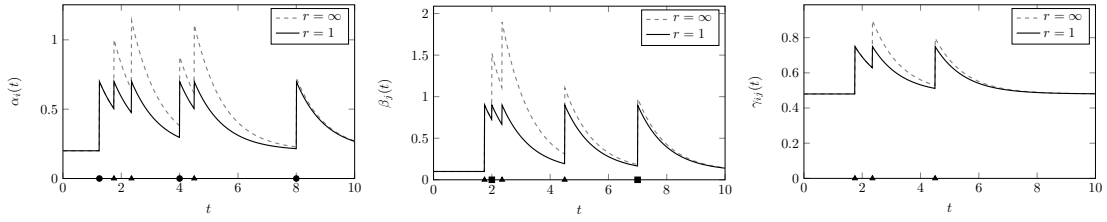


Figure 1. Cartoon of a 1-dimensional MEG model with $\alpha_i = 0.2$, $\mu_i = 0.5$, $\phi_i = 0.5$, $\beta_j = 0.1$, $\mu'_j = 0.8$, $\phi'_j = 0.2$, $\gamma_i = 0.8$, $\nu_i = 0.9$, $\theta_i = 1.1$, $\gamma'_j = 0.6$, $\nu'_j = 0.3$, $\theta'_j = 0.2$. Events with source node i and destination node j are denoted by triangles; other events with source node i are are denoted with circles, and other events with destination node j are denoted by squares.

MEGs: the resulting edge intensity function

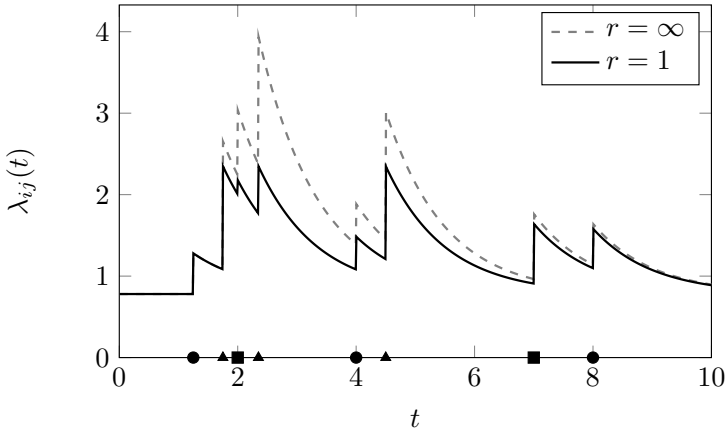


Figure 2. Cartoon of a 1-dimensional MEG model with $\alpha_i = 0.2$, $\mu_i = 0.5$, $\phi_i = 0.5$, $\beta_j = 0.1$, $\mu'_j = 0.8$, $\phi'_j = 0.2$, $\gamma_i = 0.8$, $\nu_i = 0.9$, $\theta_i = 1.1$, $\gamma'_j = 0.6$, $\nu'_j = 0.3$, $\theta'_j = 0.2$. Events with source node i and destination node j are denoted by triangles; other events with source node i are are denoted with circles, and other events with destination node j are denoted by squares.

Log-likelihood function of MEG models

- For a sequence of observed events $\mathcal{H}_T = \{(x_1, y_1, t_1), \dots, (x_m, y_m, t_m)\}$, the log-likelihood is:

$$\log L(\mathcal{H}_T; \mathbf{\Psi}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \sum_{k=1}^{n_{ij}} \log \lambda_{ij}(t_{\ell_{ijk}}) - \int_{\tau_{ij}}^{T} \lambda_{ij}(t) \mathrm{d}t \right\}, \tag{2}$$

  where n_{ij} is the number of events observed on the edge (i, j).
- Double summations: for $r = \infty$, calculating the log-likelihood on each edge is $\mathcal{O}(n_{ij}^2 + n_i^2 + n_j'^2)$.
- Assume sequences of arrival times $t_{i1} < \cdots < t_{iN_i(T)}$ involving i as source node, and $t'_{j1} < \cdots < t'_{jN'_j(T)}$ such that j is the destination of the connection.
- Assume that a subset of $n_{ij} \leq \min\{N_i(T), N'_j(T)\}$ events is observed on (i, j), and denote the indices of such events as $u_{ij,1}, \dots, u_{ij,n_{ij}}$ and $u'_{ij,1}, \dots, u'_{ij,n_{ij}}$ for each sequence. Then:

$$\log \lambda_{ij}(t_{\ell_{ijk}}) = \log \left\{ \alpha_i + \mu_i \sum_{h=1}^{u_{ij,k}-1} e^{-(\mu_i+\phi_i)(t_{\ell_{ijk}} - t_{ih})} + \beta_j + \mu'_j \sum_{h=1}^{u'_{ij,k}-1} e^{-(\mu'_j+\phi'_j)(t_{\ell_{ijk}} - t'_{jh})} \right.$$
$$\left. + \boldsymbol{\gamma}_i^{\mathsf{T}} \boldsymbol{\gamma}'_j + \sum_{q=1}^{d} \nu_{iq} \nu'_{jq} \sum_{h=1}^{k-1} e^{-(\nu_{iq}+\theta_{iq})(\theta'_{jq}+\nu'_{jq})(t_{\ell_{ijk}} - t_{\ell_{ijh}})} \right\}. \tag{3}$$

A recursive form of the log-likelihood

- Using a technique similar to the method proposed in Ogata, 1978, it is possible to calculate (3) in linear time using a **recursive formulation of the inner summations**.
- For $k \in \{1, 2, \dots, n_{ij}\}$, define $\psi_{ij}(k)$, $\psi'_{ij}(k)$ and $\tilde{\psi}_{ijq}(k)$ as follows:

$$\psi_{ij}(k) = \sum_{h=1}^{u_{ij,k}-1} e^{-(\mu_i+\phi_i)(t_{\ell_{ijk}}-t_{ih})}, \qquad \psi'_{ij}(k) = \sum_{h=1}^{u'_{ij,k}-1} e^{-(\mu'_j+\phi'_j)(t_{\ell_{ijk}}-t'_{jh})},$$

$$\tilde{\psi}_{ijq}(k) = \sum_{h=1}^{k-1} e^{-(\nu_{iq}+\theta_{iq})(\nu'_{jq}+\theta'_{jq})(t_{\ell_{ijk}}-t_{\ell_{ijh}})}, \; q = 1, \dots, d. \tag{4}$$

- Using (3) and (4), the first term of the log-likelihood (2) becomes:

$$\sum_{k=1}^{n_{ij}} \log \lambda_{ij}(t_{\ell_k}) = \sum_{k=1}^{n_{ij}} \log \left\{ \alpha_i + \beta_j + \gamma_i \gamma'_j + \mu_i \psi_{ij}(k) + \mu'_j \psi'_{ij}(k) + \nu_i \nu'_j \tilde{\psi}_{ij}(k) \right\}.$$

A recursive form of the log-likelihood

- The recursive structure of the log-likelihood stems from the following proposition:

Proposition

The terms $\psi_{ij}(k), \psi'_{ij}(k)$ and $\tilde{\psi}_{ij}(k)$ can be written recursively as:

$$\psi_{ij}(k) = e^{-(\mu_i+\phi_i)(t_{\ell_{ijk}}-t_{\ell_{ij,k-1}})}\left[1+\psi_{ij}(k-1)\right] + \sum_{h=u_{ij,k-1}+1}^{u_{ij,k}-1} e^{-(\mu_i+\phi_i)(t_{\ell_{ijk}}-t_{ih})},$$

$$\psi'_{ij}(k) = e^{-(\mu'_j+\phi'_j)(t'_{\ell'_k}-t'_{\ell'_{k-1}})}\left[1+\psi'_{ij}(k-1)\right] + \sum_{h=u'_{ij,k-1}+1}^{u'_{ij,k}-1} e^{-(\mu'_j+\phi'_j)(t_{\ell_{ijk}}-t'_{jh})},$$

$$\tilde{\psi}_{ijq}(k) = e^{-(\nu_{iq}+\theta_{iq})(\nu'_{jq}+\theta'_{jq})(t_{\ell_{ijk}}-t_{\ell_{ij,k-1}})}\left[1+\tilde{\psi}_{ijq}(k-1)\right].$$

- Importantly, the recursive structure also extends to the gradient $\boldsymbol{g} = \frac{\partial}{\partial \boldsymbol{\Psi}} \log L(\mathcal{H}_T; \boldsymbol{\Psi})$.

Inference via the EM algorithm

- An EM algorithm can be implemented using an extension of the procedure of Fox et al., 2016.
- Reparametrisation of the scaled exponential excitation functions:
  - the decay rates $\mu_i + \phi_i, \mu'_j + \phi'_j, \nu_{iq} + \theta_{iq}$ and $\nu'_{jq} + \theta'_{jq}$ are rewritten as $\tilde{\phi}_i, \tilde{\phi}'_j, \tilde{\theta}_{iq}$ and $\tilde{\theta}'_{jq}$,
  - the jumps are expressed as products between $\tilde{\phi}_i, \tilde{\phi}'_j, \tilde{\theta}_{iq}$ and $\tilde{\theta}'_{jq}$ and

$$\tilde{\mu}_i = \frac{\mu_i}{\mu_i + \phi_i}, \qquad \tilde{\mu}'_j = \frac{\mu'_j}{\mu'_j + \phi'_j}, \qquad \tilde{\nu}_{iq} = \frac{\nu_{iq}}{\nu_{iq} + \theta_{iq}}, \qquad \tilde{\nu}'_{jq} = \frac{\nu'_{jq}}{\nu'_{jq} + \theta'_{jq}},$$

  where such parameters lie in $[0, 1]$.
- Consider arrival times $t_{i1} < \cdots < t_{iN_i(T)}$ involving i as source node, and $t'_{j1} < \cdots < t'_{jN'_j(T)}$ such that j is the destination. Similarly, let $t_{ij1} < \cdots < t_{ijN_{ij}(T)}$ denote the events on (i, j).
- For $r = \infty$, the conditional intensity function (1) for an edge, for $t \geq \tau_{ij}$, is:

$$\lambda_{ij}(t) = \alpha_i + \sum_{k=1}^{N_i(t)} \omega_i(t - t_{ik}) + \beta_j + \sum_{k=1}^{N'_j(t)} \omega'_j(t - t'_{jk}) + \sum_{q=1}^{d} \gamma_{iq}\gamma'_{jq} + \sum_{k=1}^{N_{ij}(t)} \sum_{q=1}^{d} \omega_{ijq}(t - t_{ijk}), \tag{5}$$

where $\omega_{ij}(\cdot)$ has been expressed as a sum of d functions $\omega_{ijq}(t) = \tilde{\nu}_{iq}\tilde{\theta}_{iq}\tilde{\nu}'_{jq}\tilde{\theta}'_{jq} \exp\{-\tilde{\theta}_{iq}\tilde{\theta}'_{jq}t\}$.

Inference via the EM algorithm

- Conditional on $\mathcal{H}_t$, the subsequent event on the edge (i, j) could be interpreted as the **offspring** of one of the $2 + d + \min\{r, N_i(t)\} + \min\{r, N'_j(t)\} + d\min\{r, N_{ij}(t)\}$ components of the intensity (5), each corresponding to a non-homogeneous Poisson process in (t, ∞).
- $\lambda_{ij}(t)$ is written as a superimposition of conditional intensities of different processes, where the event allocations are missing data, giving a **branching structure** to the event hierarchy.
- For missing data problems, the traditional approach in statistics is to deploy the Expectation-Maximisation algorithm (EM, Dempster, Laird, and Rubin, 1977).
- Strategy: introduce latent binary variables to reconstruct the branching structure.
  - For events generated from the background rates α_i, β_j and $\gamma_{iq}\gamma'_{jq}$, $q = 1, \ldots, d$ (also known as *immigrant events*), the corresponding latent variables are denoted by the letter b. For example:

$$b_{ij\ell}^{(\alpha)} = \begin{cases} 1 & \text{if } t_{ij\ell} \text{ is a background event obtained from the Poisson process with rate } \alpha_i, \\ 0 & \text{otherwise.} \end{cases}$$

  - For the events that are not generated from the background rates, the corresponding latent variables are denoted with the letter z. For example:

$$z_{ijk\ell q}^{(\gamma)} = \begin{cases} 1 & \text{if } t_{ij\ell} \text{ is offspring on the } k\text{-th event on } (i, j), \text{ generated from } \omega_{ijq}(\cdot), \\ 0 & \text{otherwise.} \end{cases}$$

Inference via the EM algorithm

- If the branching structure is included, the **complete data** log-likelihood can be obtained:

(6)

$$\log L(\mathcal{H}_T; \tilde{\mathbf{\Psi}}, \mathbf{B}, \mathbf{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{\ell=1}^{n_{ij}}\left[b_{ij\ell}^{(\alpha)}\log(\alpha_i) + \sum_{k>N_i(t_{ij\ell})-r}^{N_i(t_{ij\ell})} z_{ij\ell k}^{(\alpha)}[\log(\tilde{\mu}_i\tilde{\phi}_i) - \tilde{\phi}_i(t_{ij\ell}-t_{ik})]\right.\right.$$

$$+ b_{ij\ell}^{(\beta)}\log(\beta_i) + \sum_{k>N_j'(t_{ij\ell})-r}^{N_j'(t_{ij\ell})} z_{ij\ell k}^{(\beta)}[\log(\tilde{\mu}_j'\tilde{\phi}_j') - \tilde{\phi}_j'(t_{ij\ell}-t_{jk}')] + \sum_{q=1}^{d}\left(b_{ij\ell q}^{(\gamma)}[\log(\gamma_{iq}) + \log(\gamma_{jq}')]\right.$$

$$\left.\left. + \sum_{k>N_{ij}(t_{ij\ell})-r}^{N_{ij}(t_{ij\ell})} z_{ij\ell kq}^{(\gamma)}[\log(\tilde{\nu}_{iq}\tilde{\theta}_{iq}) + \log(\tilde{\nu}_{jq}'\tilde{\theta}_{jq}') - \tilde{\theta}_{iq}\tilde{\theta}_{jq}'(t_{ij\ell}-t_{ijk})]\right)\right] - \int_{\tau_{ij}}^{T}\lambda_{ij}(t)\mathrm{d}t\right\}.$$

Inference via the EM algorithm – E-step

- The E-step of the EM algorithm consists in calculating $\mathbb{E}_{\mathbf{B},\mathbf{Z}|\mathcal{H}_T,\boldsymbol{\Psi}^*}\{\log L(\mathcal{H}_T;\tilde{\boldsymbol{\Psi}},\mathbf{B},\mathbf{Z})\}$.
- From (6), this reduces to calculating the *responsibilities*:

$$\xi_{\cdot}^{(\cdot)} = \mathbb{P}_{\mathbf{B},\mathbf{Z}|\mathcal{H}_T,\tilde{\boldsymbol{\Psi}}^*}\left\{b_{\cdot}^{(\cdot)} = 1 \mid \tilde{\boldsymbol{\Psi}}^*\right\}, \qquad \zeta_{\cdot}^{(\cdot)} = \mathbb{P}_{\mathbf{B},\mathbf{Z}|\mathcal{H}_T,\tilde{\boldsymbol{\Psi}}^*}\left\{z_{\cdot}^{(\cdot)} = 1 \mid \tilde{\boldsymbol{\Psi}}^*\right\},$$

- Such probabilities are simply represented by the relative contributions of different components to the conditional intensity (5). For $r = \infty$:

$$\begin{aligned}
\xi_{ij\ell}^{(\alpha)} &\propto \alpha_i, & \zeta_{ij\ell k}^{(\alpha)} &\propto \tilde{\mu}_i\tilde{\phi}_i \exp\{-\tilde{\phi}_i(t_{ij\ell} - t_{ik})\}\mathbb{1}_{(t_{ik},\infty)}(t_{ij\ell}),\\
\xi_{ij\ell}^{(\beta)} &\propto \beta_j, & \zeta_{ij\ell k}^{(\beta)} &\propto \tilde{\mu}'_j\tilde{\phi}'_j \exp\{-\tilde{\phi}'_j(t_{ij\ell} - t'_{jk})\}\mathbb{1}_{(t'_{jk},\infty)}(t_{ij\ell}),\\
\xi_{ij\ell q}^{(\gamma)} &\propto \gamma_{iq}\gamma'_{jq}, & \zeta_{ij\ell kq}^{(\gamma)} &\propto \tilde{\nu}_{iq}\tilde{\theta}_{iq}\tilde{\nu}'_{jq}\tilde{\theta}'_{jq} \exp\{-\tilde{\theta}_{iq}\tilde{\theta}'_{jq}(t_{ij\ell} - t_{ijk})\}\mathbb{1}_{(t_{ijk},\infty)}(t_{ij\ell}),
\end{aligned}$$

with normalising constant $\lambda_{ij}(t_{ij\ell})$, calculated using parameter values $\tilde{\boldsymbol{\Psi}}^*$.

Inference via the EM algorithm – M-step

- At the M-step, the expectation $\mathbb{E}_{\mathbf{B},\mathbf{Z}|\mathcal{H}_T,\tilde{\mathbf{\Psi}}^*}\{\log L(\mathcal{H}_T;\tilde{\mathbf{\Psi}},\mathbf{B},\mathbf{Z})\}$ calculated at the E-step is maximised with respect to $\tilde{\mathbf{\Psi}}$, and updated parameter estimates are obtained.
- For most of the parameters in the MEG model with scaled exponential excitation function, the maxima are analytically available:

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{n}\sum_{\ell=1}^{n_{ij}}\xi_{ij\ell}^{(\alpha)}}{\sum_{j=1}^{n}(T-\min\{T,\tau_{ij}\})}, \qquad \hat{\tilde{\mu}}_i = \frac{\sum_{j=1}^{n}\sum_{\ell=1}^{n_{ij}}\sum_{k=1}^{N_i(t_{ij\ell})}\zeta_{ij\ell k}^{(\alpha)}}{\sum_{j=1}^{n}\sum_{k=1}^{n_i}[e^{-\tilde{\phi}_i\min\{T,\max\{\tau_{ij}-t_{ik},0\}\}}-e^{-\tilde{\phi}_i(T-t_{ik})}]},$$

$$\hat{\gamma}_{iq} = \frac{\sum_{j=1}^{n}\sum_{\ell=1}^{n_{ij}}\xi_{ij\ell q}^{(\gamma)}}{\sum_{j=1}^{n}\gamma'_{jq}(T-\min\{T,\tau_{ij}\})}, \qquad \hat{\tilde{\nu}}_{iq} = \frac{\sum_{j=1}^{n}\sum_{\ell=1}^{n_{ij}}\sum_{k=1}^{N_{ij}(t_{ij\ell})}\zeta_{ij\ell kq}^{(\gamma)}}{\sum_{j=1}^{n}\tilde{\nu}'_{jq}\sum_{k=1}^{n_{ij}}[1-e^{-\tilde{\theta}_{iq}\tilde{\theta}'_{jq}(T-t_{ijk})}]},$$

and similarly for $\hat{\beta}_j$, $\hat{\tilde{\mu}}'_j$, $\hat{\gamma}'_{jq}$ and $\hat{\tilde{\nu}}'_{jq}$.

- For the remaining parameters, a solution is not available, but recursive equations can be obtained:

$$\tilde{\theta}_{iq} = \frac{\sum_{j=1}^{n}\sum_{\ell=1}^{n_{ij}}\sum_{k=1}^{N(t_{ij\ell})}\zeta_{ij\ell kq}^{(\gamma)}}{\sum_{j=1}^{n}\sum_{\ell=1}^{n_{ij}}\{\sum_{k=1}^{N_{ij}(t_{ij\ell})}\zeta_{ij\ell kq}^{(\gamma)}\tilde{\theta}'_{jq}(t_{ij\ell}-t_{ijk})+\tilde{\nu}_{iq}\tilde{\nu}'_{jq}\tilde{\theta}'_{jq}(T-t_{ij\ell})e^{-\tilde{\theta}_{iq}\tilde{\theta}'_{jq}(T-t_{ij\ell})}\}},$$

Similar equations are available for $\tilde{\phi}_i$, $\tilde{\phi}'_j$ and $\tilde{\theta}'_{jq}$.

Inference via gradient ascent

- The EM algorithm guarantees convergence to a local maximum.
- Issue: it is **not scalable** to large networks or to a large numbers of events, since it requires to define $n_{ij}[2 + d + N_i(T) + N'_j(T) + dn_{ij}]$ additional latent variables for each edge.
- Possible solution: use gradient-based algorithms for optimisation, using the recursive form of the log-likelihood for evaluation in linear time on each edge $\Rightarrow$ **Adam** (Kingma and Ba, 2015).

Algorithm: *Adam* gradient ascent algorithm for optimisation of the log-likelihood (2)

Input: step size $\eta \in \mathbb{R}_+$, decay rates $\rho_1, \rho_2 \in (0, 1)$, smoothing $\varepsilon \in \mathbb{R}_+$, initial values $\boldsymbol{\Psi}_0$.
Output: model parameters $\boldsymbol{\Psi}$ corresponding to a local maximum of $\log L(\mathcal{H}_T; \boldsymbol{\Psi})$.

Initialise estimates of the first and second moment of the gradient: $\boldsymbol{m}_0 = \mathbf{0}, \boldsymbol{v}_0 = \mathbf{0}$,
for $k = 1, 2, \ldots$ **do**
  calculate gradient $\boldsymbol{g}_k = \frac{\partial}{\partial \boldsymbol{\Psi}} \log L(\mathcal{H}_T; \boldsymbol{\Psi})\big|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}_{k-1}}$, evaluated at $\boldsymbol{\Psi}_{k-1}$,
  update EWMA estimate of first moment: $\boldsymbol{m}_k = \rho_1 \boldsymbol{m}_{k-1} + (1 - \rho_1)(\boldsymbol{g}_k \times \boldsymbol{\Psi}_{k-1})$,
  update second moment: $\boldsymbol{v}_k = \rho_2 \boldsymbol{v}_{k-1} + (1 - \rho_2)[(\boldsymbol{g}_k \times \boldsymbol{\Psi}_{k-1}) \times (\boldsymbol{g}_k \times \boldsymbol{\Psi}_{k-1})]$,
  update parameters: $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}_{k-1} \times \exp\left\{\eta \boldsymbol{m}_t \big/ (1 - \rho_1^k)\left(\sqrt{\boldsymbol{v}_t/(1 - \rho_2^k)} + \varepsilon\right)\right\}$,
until *convergence in* $\log L(\mathcal{H}_T; \boldsymbol{\Psi})$.

Simulation and assessment of the goodness-of-fit

- **Simulation: thinning.** In order to validate the inferential procedure, it is necessary to simulate data from the MEG model (1), which can be interpreted as an extended multivariate Hawkes process where some of the parameters are shared across the individual processes. Therefore, simulating MEG models is possible under the framework described in Ogata, 1981, and follows the standard technique of simulation via *thinning*.
- **Goodness-of-fit: distribution of the p-values of out-of-sample events.** Given arrival times $t_1, \ldots, t_{n_{ij}}$ on the edge (i, j), the upper tail p-values are:

$$p_{ijk} = \exp\left\{ - \int_{t_{k-1}}^{t_k} \lambda_{ij}(s) \mathrm{d}s \right\}, k = 1, \ldots, n_{ij}.$$

Under the null hypothesis of correct specification of the conditional intensity $\lambda_{ij}(t)$, the p-values are uniformly distributed.

MEG on a fully connected graph - Main effects only

- 100 simulations of 3,000 events on a fully connected MEG with $n = 2$:
  - $\lambda_{ij}(t) = \alpha_i(t) + \beta_j(t), r = \infty,$
  - $\boldsymbol{\alpha} = [0.01, 0.05], \boldsymbol{\beta} = [0.07, 0.03], \boldsymbol{\mu} = [0.2, 0.15], \boldsymbol{\mu}' = [0.1, 0.25], \boldsymbol{\phi} = [0.8, 0.85], \boldsymbol{\phi}' = [0.9, 0.75].$

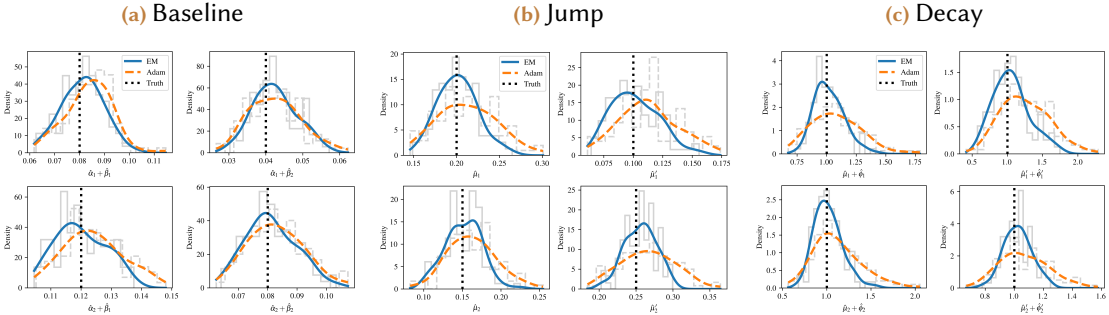


Figure 3. Histograms (with corresponding kernel density estimates) of parameter estimates obtained using EM and Adam from 100 simulations of 3,000 events on a fully connected MEG with $n = 2$, $\lambda_{ij}(t) = \alpha_i(t) + \beta_j(t)$, $r = \infty$, $\boldsymbol{\alpha} = [0.01, 0.05]$, $\boldsymbol{\beta} = [0.07, 0.03]$, $\boldsymbol{\mu} = [0.2, 0.15]$, $\boldsymbol{\mu}' = [0.1, 0.25]$, $\boldsymbol{\phi} = [0.8, 0.85]$, $\boldsymbol{\phi}' = [0.9, 0.75]$.

MEG on a fully connected graph - Increasing number of events

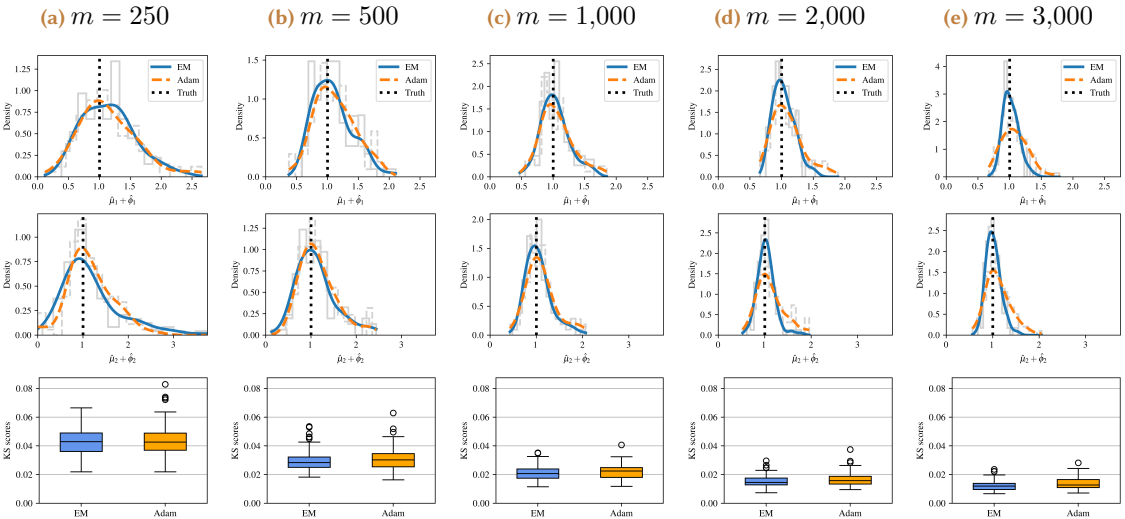


Figure 4. Histograms (with corresponding kernel density estimates) of estimates for $\boldsymbol{\mu} + \boldsymbol{\phi}$ and boxplots of KS scores obtained using EM and Adam from 100 simulations from the same model as Figure 3, with $m \in \{250, 500, 1{,}000, 2{,}000, 3{,}000\}$ events.

MEG on a fully connected graph - Interactions only

- 100 simulations of 3,000 events on a fully connected MEG with $n = 2$:
  - $\lambda_{ij}(t) = \gamma_{ij}(t)$, $r = \infty$,
  - $\boldsymbol{\gamma} = [0.1, 0.5]$, $\boldsymbol{\gamma}' = [0.1, 0.3]$, $\boldsymbol{\nu} = [0.6, 0.4]$, $\boldsymbol{\nu}' = [0.5, 0.25]$, $\boldsymbol{\theta} = [0.4, 0.6]$, $\boldsymbol{\theta}' = [0.5, 0.75]$.

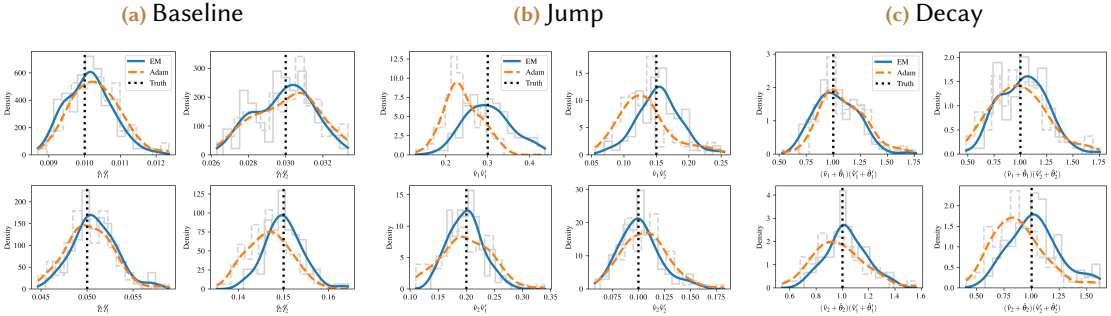


Figure 5. Histograms (with corresponding kernel density estimates) of parameter estimates and boxplots of KS scores obtained using EM and Adam from 100 simulations of 3,000 events on a fully connected MEG with $n = 2$, $\lambda_{ij}(t) = \gamma_{ij}(t)$, $r = \infty$, $\boldsymbol{\gamma} = [0.1, 0.5]$, $\boldsymbol{\gamma}' = [0.1, 0.3]$, $\boldsymbol{\nu} = [0.6, 0.4]$, $\boldsymbol{\nu}' = [0.5, 0.25]$, $\boldsymbol{\theta} = [0.4, 0.6]$, $\boldsymbol{\theta}' = [0.5, 0.75]$.

Results: Enron e-mail network

- The Enron e-mail network collection is a record of e-mails exchanged between the employees of Enron Corporation before its bankruptcy.
- These data have already been demonstrated to be well-modelled as self-exciting point processes by Fox et al., 2016.
- $34{,}427$ distinct triplets (x_k, y_k, t_k), corresponding to messages exchanged between $n = 184$ employees between November 1998 and June 2002, forming a total of $3{,}007$ edges.
- Some of the emails are sent to multiple receivers, and only $18{,}031$ unique event times are observed, implying that on average each e-mail is sent to approximately 1.90 nodes.
- Because an e-mail can have multiple recipients, and because the event times are recorded to the nearest second, the likelihood must be adapted with the arrivals modelled by an analogous discrete time process.
- The model is trained on $30{,}704$ e-mails, and tested on the remaining $3{,}723$ e-mails.
- In the training set, $2{,}720$ edges are observed, and 811 in the test set, of which 287 are *not* observed in the training period.

Results: Enron e-mail data, $\tau_{ij} = t_{\ell_{ij1}}$

Table 1. Training and test KS scores on the Enron e-mail network for different configurations of the MEG model.

KS scores (train & test)			Main effects $\alpha_i(\cdot)$ and $\beta_j(\cdot) \downarrow$			
$\tau_{ij} \downarrow$	Interactions $\gamma_{ij}(\cdot) \downarrow$		Absent	Poisson ($r = 0$)	Markov ($r = 1$)	Hawkes ($r = \infty$)
$\tau_{ij} = t_{\ell_{ij1}}$ (MLE)	Absent		– –	0.4530 0.4133	0.3678 0.3484	0.4443 0.3586
	Poisson ($r = 0$)	$d = 1$	0.4252 0.4221	0.3946 0.4179	0.3434 0.3574	0.4255 0.3560
		$d = 5$	0.3490 0.3851	0.3498 0.3953	0.3165 0.3677	0.3491 0.3613
		$d = 10$	0.3339 0.3763	0.3347 0.3688	0.3112 0.3470	0.3376 0.3575
	Markov ($r = 1$)	$d = 1$	0.1662 0.2029	0.1491 0.1945	0.1305 0.1777	0.1702 0.1874
		$d = 5$	0.0916 0.1875	0.0910 0.1684	**0.0885 0.1628**	0.0916 0.1746
		$d = 10$	**0.0885** 0.1743	**0.0885** 0.1848	**0.0885** 0.1696	**0.0885** 0.1743
	Hawkes ($r = \infty$)	$d = 1$	0.2640 0.2755	0.2825 0.2887	0.2538 0.2637	0.2599 0.2871
		$d = 5$	0.2304 0.2904	0.2284 0.2760	0.2271 0.2774	0.2420 0.2981
		$d = 10$	0.2461 0.2923	0.2521 0.2865	0.2413 0.3091	0.2498 0.3129

- The MLE approach has a drawback: the p-values for the first observation on each edge are *always* 1. This implies that the KS scores are bounded below by $2720/30704 \approx 0.0885$ for the training set and $287/3723 \approx 0.0770$ for the test set.

Results: Enron e-mail data, $\tau_{ij} = 0$

Table 2. Training and test KS scores on the Enron e-mail network for different configurations of the MEG model.

KS scores (train & test)			Main effects $\alpha_i(\cdot)$ and $\beta_j(\cdot) \downarrow$			
$\tau_{ij} \downarrow$	Interactions $\gamma_{ij}(\cdot) \downarrow$		Absent	Poisson ($r = 0$)	Markov ($r = 1$)	Hawkes ($r = \infty$)
$\tau_{ij} = 0$	Absent		– –	0.4530 0.4133	0.3678 0.3484	0.4443 0.3586
	Poisson ($r = 0$)	$d = 1$	0.7039 0.7926	0.6627 0.7753	0.6543 0.7148	0.7059 0.6050
		$d = 5$	0.5623 0.7059	0.5646 0.7206	0.5748 0.7008	0.7060 0.6053
		$d = 10$	0.5354 0.6853	0.5332 0.6739	0.5725 0.6952	0.7060 0.6059
	Markov ($r = 1$)	$d = 1$	0.3135 0.3324	0.3004 0.3326	0.3262 0.3240	0.2027 0.1999
		$d = 5$	0.0760 0.1664	0.0825 0.1584	0.0855 0.1782	0.0495 **0.0924**
		$d = 10$	0.0775 0.1649	0.0793 0.1546	0.0816 0.1606	**0.0402** 0.0971
	Hawkes ($r = \infty$)	$d = 1$	0.2871 0.2486	0.2333 0.2449	0.2485 0.2379	0.1749 0.1991
		$d = 5$	0.1939 0.2167	0.1885 0.2246	0.2010 0.2137	0.1467 0.1994
		$d = 10$	0.2029 0.2395	0.2158 0.2470	0.2207 0.2339	0.1606 0.1943

- Comparison to alternative node-based models:
  - Poisson processes $\lambda_i(t) = \alpha_i$. KS score: 0.4088;
  - Hawkes processes $\lambda_i(t) = \alpha_i + \sum_{k=1}^{N_i(t)} \mu_i \exp\{-(\mu_i + \phi_i)(t - t_{ik})\}$. KS score: 0.2499;
  - Mutually exciting process with intensity $\lambda_i(t) = \alpha_i + \sum_{k=1}^{N_i'(t)} \mu_i \exp\{-(\mu_i + \phi_i)(t - t'_{ik})\}$ (Fox et al., 2016). KS score: 0.2806. It could be inferred that users tend to respond to multiple e-mails in sessions, and not necessarily immediately after an individual e-mail is received.

Results: Enron e-mail data, $\tau_{ij} = 1/A_{ij}$

Table 3. Training and test KS scores on the Enron e-mail network for different configurations of the MEG model.

KS scores (train & test)			Main effects $\alpha_i(\cdot)$ and $\beta_j(\cdot) \downarrow$			
$\tau_{ij} \downarrow$	Interactions $\gamma_{ij}(\cdot) \downarrow$		Absent	Poisson ($r = 0$)	Markov ($r = 1$)	Hawkes ($r = \infty$)
$\tau_{ij} = \begin{cases} 0, & A_{ij} = 1 \\ \infty, & A_{ij} = 0 \end{cases}$	Absent		– –	0.4530 0.4133	0.3678 0.3484	0.4443 0.3586
	Poisson ($r = 0$)	$d = 1$	0.5158 0.6038	0.4812 0.5864	0.3742 0.3602	0.4197 0.2808
		$d = 5$	0.4269 0.5516	0.4309 0.5641	0.3553 0.3598	0.3938 0.2803
		$d = 10$	0.4035 0.5413	0.4084 0.5565	0.3430 0.3537	0.3659 0.2810
	Markov ($r = 1$)	$d = 1$	0.1950 0.2115	0.1600 0.2017	0.1504 0.1422	0.1309 0.1445
		$d = 5$	0.0709 0.1222	0.0746 0.1008	0.0696 0.0917	**0.0152** 0.0848
		$d = 10$	0.0619 0.1029	0.0627 0.1079	0.0634 0.0836	0.0213 **0.0800**
	Hawkes ($r = \infty$)	$d = 1$	0.1870 0.2084	0.1816 0.2049	0.1783 0.1747	0.1719 0.1879
		$d = 5$	0.1377 0.1805	0.1374 0.1840	0.1391 0.1642	0.1553 0.2154
		$d = 10$	0.1556 0.2023	0.1588 0.2046	0.1546 0.1863	0.1640 0.2082

- The best performance (KS score 0.0152) is achieved when a Markov process is used for the interaction, with $d = 5$ or $d = 10$, combined with a Hawkes process for the main effects.
- In general, setting τ_{ij} using the adjacency matrix seems to outperform competing strategies for estimation of τ_{ij} in terms of KS scores.
- The interaction term should be included in the model.

Results: ICL NetFlow data

- Many enterprises routinely collect network flow (NetFlow) data, representing summaries of connections between internet protocol (IP) addresses.
- The data consists of $1{,}951{,}067$ arrival times (in milliseconds), recorded in three weeks.
  - Sources: $n_1 = 173$ clients hosted within the Department of Mathematics at ICL;
  - Destinations: $n_2 = 6{,}083$ internet servers connecting on ports 80 and 443;
  - $156{,}186$ unique edges in total.
- Only edges such that the percentage of arrival times observed between 7am and 12am is larger than 99%, corresponding to the college opening hours, were considered.
- The MEG model is trained on the first two weeks of data, corresponding to $1{,}299{,}372$ events, and tested on $651{,}695$ events observed in the final week.
- The number of unique edges in the training set is $115{,}600$, and $70{,}408$ in the test set.
- Only $29{,}822$ edges are observed in both time windows, which implies that $40{,}586$ new edges are observed in the test set.

Results: ICL NetFlow data, KS scores

Table 4. Kolmogorov-Smirnov scores on the ICL NetFlow data for different configurations of the MEG model.

KS scores (train & test)		Main effects $\alpha_i(\cdot)$ and $\beta_j(\cdot) \downarrow$			
Interactions $\gamma_{ij}(\cdot) \downarrow$		Absent	Poisson ($r = 0$)	Markov ($r = 1$)	Hawkes ($r = \infty$)
Absent		– –	0.7351 0.7148	0.6678 0.6489	0.7312 0.6950
Poisson ($r = 0$)	$d = 1$	0.7328 0.7157	0.7325 0.7150	0.6672 0.6480	0.7316 0.6960
	$d = 5$	0.7295 0.7167	0.7313 0.7123	0.6673 0.6487	0.7275 0.6967
	$d = 10$	0.7260 0.7174	0.7289 0.7140	0.6680 0.6493	0.7270 0.6969
Markov ($r = 1$)	$d = 1$	0.2194 0.1723	0.2242 0.1657	0.2038 0.1440	0.1645 0.1281
	$d = 5$	0.1024 0.1080	0.0896 0.0805	**0.0728 0.0738**	0.1041 0.0899
	$d = 10$	0.0843 0.0764	0.0871 0.0761	0.0850 0.0843	0.1100 0.0883
Hawkes ($r = \infty$)	$d = 1$	0.1080 0.0802	0.0747 0.1182	0.1082 0.0794	0.0884 0.1262
	$d = 5$	0.1576 0.1819	0.1532 0.2126	0.1677 0.2143	0.2307 0.2383
	$d = 10$	0.1584 0.1935	0.1546 0.2112	0.1619 0.2206	0.2388 0.2503

- $\tau_{ij} = 0$ if $A_{ij} = 1$, $\tau_{ij} = \infty$ if $A_{ij} = 0$.

Results: ICL NetFlow data, Q-Q plots

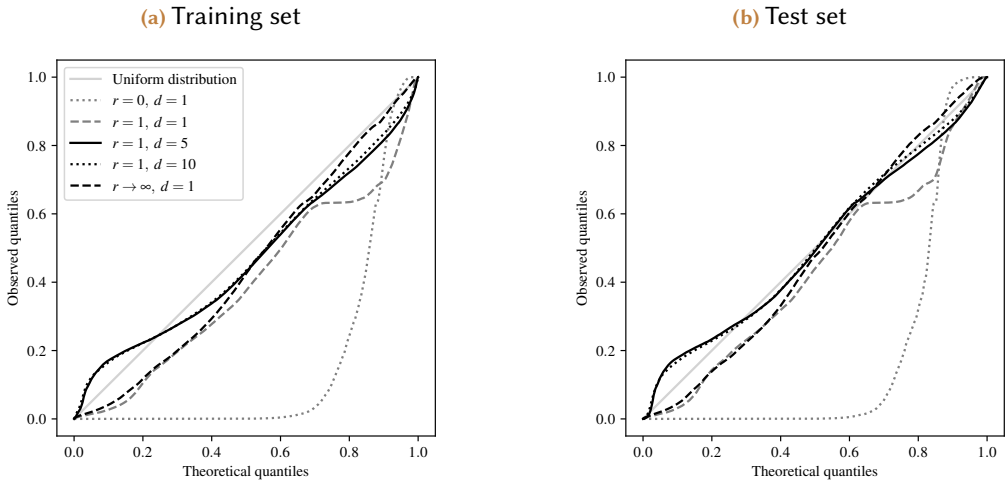


Figure 6. Q-Q plots for the training and test p-values obtained from different MEG models, with main effects $\alpha_i(t)$ and $\beta_j(t)$ with $r = 1$, and different parameters for the interaction term $\gamma_{ij}(t)$, specified in the legend.

Results: ICL NetFlow data, KS scores

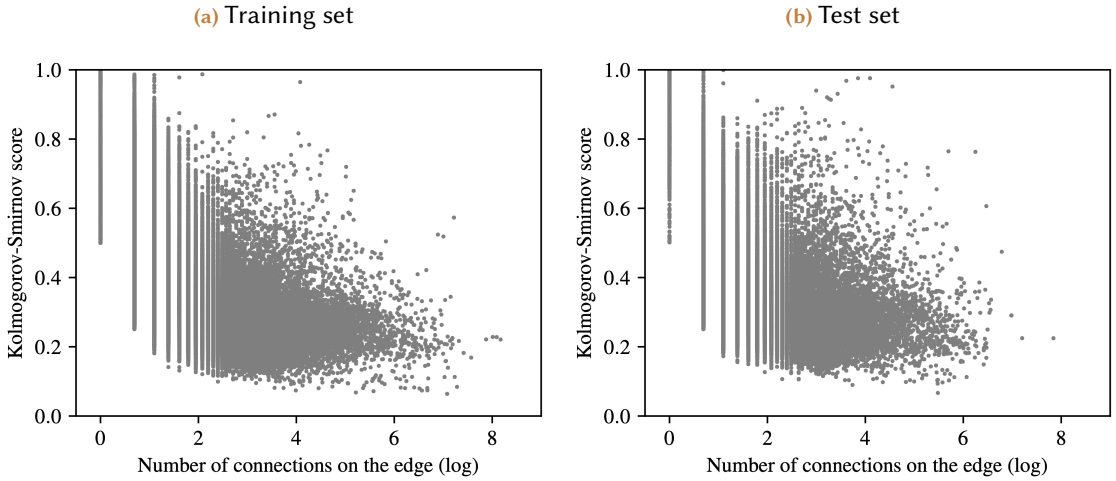


Figure 7. Scatterplot of the Kolmogorov-Smirnov scores, calculated for each edge, versus the logarithm of the total number of connections on the edge, for the best performing model in Table 4.

Conclusion

- Mutually-exciting graphs (MEG), network-wide models for point processes with dyadic marks, have been proposed.
- Edge-specific intensities are obtained only via node-specific parameters, which is useful for large and sparse graphs.
- MEG is able to predict events observed on *new* edges.
- MEG greatly outperforms results previously obtained in the literature on the Enron e-mail network.
- More details in Sanna Passino and Heard, 2021. Scan the QR code to get the arχiv preprint – **[v2]** coming soon!
- *python* code on GitHub: `fraspass/meg`.

References I

Athreya, A. et al. (2018). "Statistical Inference on Random Dot Product Graphs: a Survey". In: *Journal of Machine Learning Research* 18.226, pp. 1–92.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B* 39.1, pp. 1–22.

Fox, E.W. et al. (2016). "Modeling E-mail Networks and Inferring Leadership Using Self-Exciting Point Processes". In: *Journal of the American Statistical Association* 111.514, pp. 564–584.

Hoff, P. (2021). "Additive and Multiplicative Effects Network Models". In: *Statistical Science* 36.1, pp. 34–50.

Hoff, P. D, A. E. Raftery, and M. S. Handcock (2002). "Latent space approaches to social network analysis". In: *Journal of the American Statistical Association* 97.460, pp. 1090–1098.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: a Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR*. Ed. by Yoshua Bengio and Yann LeCun. San Diego, CA, USA.

Ogata, Y. (1978). "The asymptotic behaviour of maximum likelihood estimators for stationary point processes". In: *Annals of the Institute of Statistical Mathematics* 30.2, pp. 243–261.

References II

- Ogata, Y. (1981). "On Lewis' simulation method for point processes". In: *IEEE Transactions on Information Theory* 27.1, pp. 23–31.
- Price-Williams, M. and N. A. Heard (2020). "Nonparametric Self-exciting Models for Computer Network Traffic". In: *Statistics and Computing* 30.2, pp. 209–220.
- Sanna Passino, F. and N. A. Heard (2021). "Mutually exciting point process graphs for modelling dynamic networks". In: *arXiv e-prints*. arXiv: 2102.06527 [cs.SI].