CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

How fast do we forget our past social interactions? Understanding memory retention with parametric decays in relational event models

Giuseppe Arena[1,*] ![ORCID], Joris Mulder[1,2] and Roger Th. A.J. Leenders[2,3]

[1]Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands, [2]Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands, and [3]Department of Organization Studies, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands
*Corresponding author. Email: G.Arena@tilburguniversity.edu

Guest Editors (Special Issue on Relational Event Models): Carter Butts, Alessandro Lomi, Tom Snijders, Christoph Stadtfeld

Abstract

In relational event networks, endogenous statistics are used to summarize the past activity between actors. Typically, it is assumed that past events have equal weight on the social interaction rate in the (near) future regardless of the time that has transpired since observing them. Generally, it is unrealistic to assume that recently past events affect the current event rate to an equal degree as long-past events. Alternatively one may consider using a prespecified decay function with a prespecified rate of decay. A problem then is that the chosen decay function could be misspecified yielding biased results and incorrect conclusions. In this paper, we introduce three parametric weight decay functions (exponential, linear, and one-step) that can be embedded in a relational event model. A statistical method is presented to decide which memory decay function and memory parameter best fit the observed sequence of events. We present simulation studies that show the presence of bias in the estimates of effects of the statistics whenever the decay, as well as the memory parameter, are not properly estimated, and the ability to test different memory models against each other using the Bayes factor. Finally, we apply the methodology to two empirical case studies.

Keywords: relational event model; social network analysis; network dynamics; memory decay; memory retention process; profile likelihood; Bayes factor

1. Introduction

In relational event networks, the past relational event history between the actors can have an enormous impact on future relational events (Butts, 2008). Research has shown that the past can generally be well-summarized using so-called endogenous statistics to model the events to be observed. These endogenous statistics typically quantify the activity between actors in the past. For example, the endogenous statistic Inertia of actor i towards j for event m is generally computed as the total volume of past events from i to j until the previous event, that is,

$$\text{inertia}(i, j, t_m) = \sum_{e' \in E_{t_{m-1}}} \mathbb{I}(s(e') = i, r(e') = j), \qquad (1)$$

where $E_{t_{m-1}}$ denotes the event history until event $m - 1$, $s(e')$ is the sender of event e', and $r(e')$ is the receiver of event e'. Thereby the assumption is that the (logarithm of the) relational event rate between two actors depends proportionally on the number of past events between these actors.

Other examples of endogenous statistics include reciprocity, transitivity, or even more complex higher-order dynamic patterns.

By defining the endogenous statistics as the total number of past events the assumption is that all past events equally contribute to the event rate of the dyads in the subsequent period. This however may not be likely in real-life social networks. As an example let us consider a group of friends who send text messages to each other. At some point, let us assume that James sent about 24 messages to Keira, and that Vicky also sent about 24 messages to Roberto (since the observational period). Out of the 24 messages sent by James to Keira, let us assume that 22 were send more than one month ago, and 2 messages were send two weeks ago. Out of the 24 messages sent by Vicky to Roberto, all 24 messages were send in the last 5 days. Now the question is whether it is more likely that James will send a message to Keira next, or that Vicky will send a message to Roberto next? Under the assumption that sending messages is mainly driven by inertia, and inertia is computed as the total volume of past messages between actors (which is equal for both dyads), there would be an equal probability that for James to send a message to Keira, as for Vicky to send a message to Roberto. Given the fact that Vicky has been much more active to send message to Roberto in the recent past however it may be much more plausible that Vicky will send a message to Robert next than James to Keira. This would imply that recently past relational events have a stronger impact on what happens next than long-past events. Thus one could say that recently past social events are fresh in the memory of actors while long past events may not.

Alternatively it can be assumed that the weight of past events decays according to an exponential function of the transpired time of the past event (Brandes et al., 2009). Indeed, it is likely that over time actors differently weigh past events according to their time recency. The time recency of an event is defined as the time transpired at the present time point since its occurrence. This measure increases over time after the event happens and can be a crucial information in understanding whether and how the weights of events decay as time goes by and their time recency decreases. This would provide value insight to learn how the past affects the future in relational event networks. As proposed in Brandes et al.(2009), the formula for inertia is the following:

$$\text{exponential-inertia}(i, j, t_m) = \sum_{\substack{e' \in E_{t_{m-1}}: \\ s(e')=i \wedge r(e')=j}} \frac{\ln(2)}{\theta_{\text{half-life}}} \exp\left\{ -(t_m - t_{e'}) \frac{\ln(2)}{\theta_{\text{half-life}}} \right\} \tag{2}$$

where the weight of each event $(s(e'), r(e')) = (i, j)$ in the current history of events $(E_{t_{m-1}})$ follows an exponential decay governed by the half-life parameter $\theta_{\text{half-life}}$, which is assumed to be known. The transpired time of the event e', measured as $t_m - t_{e'}$, is updated at each time point. Because it increases over time, the transpired time decreases the event weight over time. The speed of such decrease depends on the value of the half-life parameter ($\theta_{\text{half-life}}$) that describes the waiting time before the weight of the past event (i, j) halves. Thus, the larger the $\theta_{\text{half-life}}$, the slower will be the decrease in weight and, in turn, long-passed events will keep having a high contribution in the calculation of the statistic, reflecting a long-lasting memory of actors. When a researcher changes the parameters governing the decay, the model statistics (such as the value of inertia) change with it and, in turn, their effect on the event rate changes as well. For this reason, the use of such pre-specified half-life parameters should be used with care. This is even more the case because the weight of past events may even decrease with a different shape than with an exponential shape in real life networks. The change of effects due to a change in the memory parameter was already explored by Brandenberger (2018), where the author shows the different estimated model effects resulting from pre-specifying different half-life values.

Another approach to modeling weight decay in relational event data was proposed by Perry & Wolfe (2013), where the past history of events at each time point is divided according to a set of

$K + 1$ increasing time widths $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_K)$ and endogenous statistics are calculated within each of the K resulting intervals. For instance, inertia for the k-th interval is calculated as

$$\text{interval-inertia}(i, j, t_m, k) = \sum_{\substack{e' \in E_{t_{m-1}}: \\ (t_m - t_{e'}) \in (\gamma_{k-1}, \gamma_k)}} \mathbb{I}(s(e') = i, r(e') = j) \quad \text{for} \quad k = 1, \ldots, K \tag{3}$$

Therefore, the effect of inertia in each interval is estimated and finally described by the vector of effects $\boldsymbol{\beta}_{\text{inertia}} = (\beta_{\text{inertia}_1}, \ldots, \beta_{\text{inertia}_K})$. No assumptions are made about the steps (which may either decrease, increase, etc.) and the estimation can be done relatively easy using existing software. Note that there is a clear relation between this stepwise approach and the above weighted approach according to the following equation

$$\begin{bmatrix} \beta_{\text{inertia}_1} \\ \vdots \\ \beta_{\text{inertia}_K} \end{bmatrix} = \begin{bmatrix} \beta_{\text{inertia}} w_1 \\ \vdots \\ \beta_{\text{inertia}} w_K \end{bmatrix} \tag{4}$$

where on the right side: $(w_1, \ldots, w_K)$ is the step-wise function of the weights of the events which is based on the widths $(\gamma_0, \ldots, \gamma_K)$ and assumes that events belonging to the same interval have the same weight, β_{inertia} is the effect of the network dynamic on the event rate. For an extended description of the relation in (4) see Appendix A1. Also the step-wise approach has potential limitations. First, in some applications it may not be natural to assume that the relative importance of a past event is relatively high and one second later (say) its importance drops considerably which is the case in such step-wise models. Second, it is generally unclear how the intervals should be chosen such that the memory (decay) in the data is accurately captured [see also Arena et al. (2022) for a related discussion]. Finally note that by considering many different intervals for all endogenous effects, the number of unknown parameters can unduly blow up resulting in a tremendous increase of our uncertainty about the model parameters.

In this paper, an alternative methodology is proposed to better learn about past events affecting future events. We assume a continuous, parameterized decay function which can either be exponential, linear, or step-wise. Each of these functions has a single memory parameter that is optimized using the observed data. A Bayesian test is proposed to determine which decay function (exponential, linear, or step-wise) fits the data best. Thereby, the methodology builds on previous approaches by (i) allowing the weight of past events to decrease in continuous time [as in Brandes et al. (2009)] but at the same time estimate the rate of the decay from the data, and (ii) finding the best fitting shape of the weight decay [as in Perry & Wolfe (2013)] without overparameterizing the model.

Related to the current work, the time sensitivity in relational event modeling has also been discussed in various other studies. The effect of time recency of past interactions was discussed by Tranmer et al. (2015), and a weekend effect was investigated by Amati et al. (2019) in a network of health care organizations, in which authors show the different network mechanisms that can be observed between week days and weekends. In another work, Bianchi & Lomi (2022) study short-term and long-term effects in network dynamics and provide examples on a high-frequency network (financial markets) as well as on a low-frequency network (patient-sharing relations among health care organizations). Furthermore, methods for estimating time-varying networks effects were proposed by Mulder & Leenders (2019), Meijerink-Bosman et al. (2022), and Meijerink-Bosman et al. (2022) using moving window approaches, and by Fritz et al. (2021) using B-splines.

Furthermore, some work has been done on external decays and on the presence of right-censoring. Stadtfeld & Geyer-Schulz (2011) discussed the problem of using external decay functions in a discrete state space and examined the use of exponential decays combined with an arbitrary threshold on the decay. They observed that despite external decays as well as events

of other types might affect the network of events under study, if the Markov process transition rates are defined over very short time spans, the impact of such factors would only be marginal. In another work, Stadtfeld & Block (2017) discussed about the hurdle of right-censoring in relational event networks and proposed a discrete time window approach that overcomes the issues generated from right-censored events.

The remainder of this paper is organized as follows: in Section 2, we introduce parametric memory decay functions and define three potential decays: (i) the one-step decay, (ii) the exponential decay, and (iii) the linear decay. Then, in Section 3, we look into the methodological consequences of treating the memory parameters as parameters to be estimated from the data, introduce the use of the profile log-likelihood in relational event models (REMs), and finally propose some possible optimization methods which aim to find the maximum likelihood estimate for the memory parameter. In Sections 4 and 5, we show the results on simulated relational event histories as well as on two real case studies.

2. Parametric functions for modeling memory decay

Recently occurred events generally have a larger impact on the next relational event that will occur in a social network than long-past events. To model this we define a weighting function, which is denoted by $w(\gamma_e(t), \theta)$ where

1. $\gamma_e(t) = t - t_e$ is the transpired time of event e at time t with $t > t_e$;
2. θ is a memory parameter with support $S(\theta) \in \mathbb{R}$, which determine the resulting shape of the decay;
3. the outcome of the weight is a non-negative real number, that is $w(\gamma_e(t), \theta) \in \mathbb{R}_0^+$.

The weight of a past event can reflect to what degree a past event is remembered, and thus, the weighting function can be viewed as an operationalization of the memory decay of actors about past events. For this reason, we shall use the term weighting function and memory decay function interchangeably in this paper.

The above weighting function is then used for computing the endogenous statistics which summarize the past event history at time t. For example, inertia, which is normally computed as the total count (volume) of past events between two actors (i, j), is now computed as a weighted count of past events weighted according to the chosen weighting function with memory parameter θ, that is,

$$\text{weighted-inertia}(i, j, t_m, \theta) = \sum_{e' \in E_{t_{m-1}}} \mathbb{I}(s(e') = i, r(e') = j) w(\gamma_{e'}(t_{m-1}), \theta) \tag{5}$$

Note that the transpired time is computed from the time of the previous event t_{m-1}, which is when the waiting time starts for observing the mth event. During the waiting time, the weight is assumed to stay constant so that the assumption of constant hazards between events is not violated.

In contrast to previous approaches we assume the memory parameter to be unknown. Regarding the memory function, many possible shapes could be considered. To keep the model computationally feasible however, three parametric functions of the memory decay are considered in this paper: a one-step decay function, an exponential decay function, and a linear decay function.

One-step decay

The one-step decay is defined as

$$w_{\text{step}}(\gamma_e(t), \theta_{\text{max}}) = \begin{cases} 1 & \text{if } \gamma_e(t) \leq \theta_{\text{max}} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$
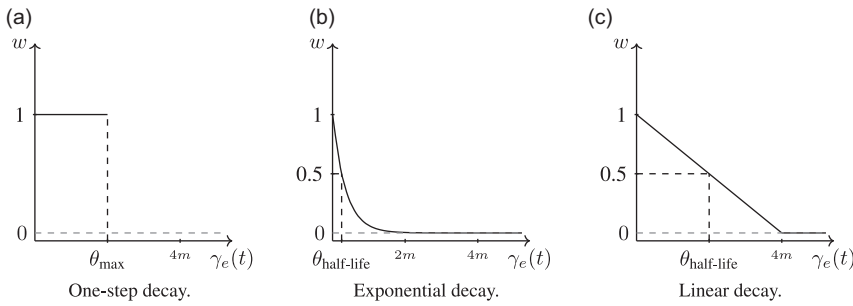
Figure 1. Three examples of memory decay: (a) One-step decay where $\theta_{\mathrm{max}} \approx 2$ months and the height of the step is fixed to 1; (b) exponential decay where $\theta_{\mathrm{half\text{-}life}} \approx 7$ days; (c) linear decay where $\theta_{\mathrm{half\text{-}life}} \approx 2$ months.

where $\theta_{\mathrm{max}} \in (0, +\infty)$ is the memory parameter and the decay describes a one-step function. Thus, events will contribute to the statistic only if their transpired time is less than the threshold θ_{max}. Moreover, the model is simplified to the case where the weight is unitary. Note that the interpretation of a coefficient of an endogenous statistic that is computed using a one-step memory function is similar to the interpretation of coefficients of count statistics which ignore memory decay. The only difference is that in the step-wise model only the events with transpired time that do not exceed the threshold value θ_{max} contribute to the rate with a value equal to the corresponding coefficient of the parameter. An example of the shape of the one-step decay is shown in Figure 1(a) where $\theta_{\mathrm{max}} \approx 2$ months. In this case, weighted inertia between actors i and j would be equal to the total number of past events between i and j in the last 2 months.

Substantively a one-step decay may be appropriate in social networks where actors only have a relatively short-term memory. It may then be reasonable to assume that only the past events within this short window affect the endogenous statistics, and that the past events in this window affect the endogenous variables (approximately) equal. Computationally the one-step model is convenient as we would only need to look back until θ_{max} to compute the endogenous statistics.

The one-step function was used by Mulder & Leenders (2019) using a prespecified memory length. Mulder & Leenders (2019) also assumed that network parameters may change over time. This was achieved by estimating the network parameters within a time window which was set equal to the chosen memory length while moving the window over the observed event history. In the current paper we do not consider a model where network parameters can change over time. Instead we assume that the network parameters are homogeneous over time but, in the case of a one-step decay, we do assume that the past events affect the endogenous statistics until a certain threshold value (i.e., θ_{max}), which is assumed unknown.

Exponential decay

The functional form for the exponential decay is

$$w_{\exp}(\gamma_e(t), \theta_{\mathrm{half\text{-}life}}) = \exp\left\{ -\gamma_e(t) \frac{\ln(2)}{\theta_{\mathrm{half\text{-}life}}} \right\} \tag{7}$$

for $\gamma_e(t) \in (0, +\infty)$ where $\theta_{\mathrm{half\text{-}life}} \in (0, +\infty)$ is the memory parameter that measures the minimum elapsed time after which the event weight is halved. In this formulation, we let the weights start decaying from 1 instead of $\frac{\ln(2)}{\theta_{\mathrm{half\text{-}life}}}$ as it is defined in Brandes et al. (2009) and this will only affect the scaling of the effects $\boldsymbol{\beta}$. One of the possible shapes of an exponential decay is shown in Figure 1(b) where $\theta_{\mathrm{half\text{-}life}} \approx 7$ days.

The interpretation of a coefficient of an endogenous statistic that is computed using an exponential decay function is slightly more complicated than for a regular count statistic because the contribution of each past event to the rate (and the hazard) depends on the transpired time

since the event was observed. For example, when the coefficient of inertia is equal to 3 and the decay function in Figure 1(b) is considered with a half-life of 7 days which starts at 1, the last event that was observed has a maximal contribution to the rate (and hazard) equal to 3. Furthermore, the contribution of events that were observed approximately 7 days ago contribute with approximately 1.5, and events that were observed approximately 14 days ago contribute with approximately 0.75 to the rate and hazard.

Theoretically an exponential memory decay implies that the weight reduces to half its value in a fixed amount of time, regardless of the current weight. Furthermore the model assumes that past events are never "forgotten" as in the one-step model. Depending on the context this may be realistic. Computationally the exponential decay is somewhat demanding as it requires one to look back at the entire past history for computing endogenous statistics. However eventually the weights become negligible, and thus, can be approximated as zero.

The exponential model was proposed by Brandes et al. (2009) to model relational events between political actors (e.g., countries) during conflicts. Instead of estimating the half-life parameter from the observed data, the model was fitted using different prespecified half-life parameters. This yielded fairly consistent results in their empirical applications. It is yet unknown whether this result holds in general. This will be explored later in this paper when fitting the model using misspecified memory parameters.

Linear decay

The linear decay function is defined as

$$w_{\text{linear}}(\gamma_e(t), \theta_{\text{half-life}}) = \left(1 - \frac{1}{2\theta_{\text{half-life}}} \gamma_e(t)\right) \mathbb{I}(\gamma_e(t) \leq 2\theta_{\text{half-life}}) \tag{8}$$

for $\gamma_e(t) \in (0, +\infty)$ and with $\theta_{\text{half-life}} \in (0, +\infty)$, which quantifies the time until the weight is halved, similar as in the exponential decay in (7). Unlike the exponential decay on the other hand the weight becomes 0 after the transpired time reaches $2\theta_{\text{half-life}}$, similar as the one-step model. In this sense the linear decay model can be seen as a middle ground between the one-step decay and the exponential decay function. An example of a linear weight decay is shown in Figure 1(c) for $\theta_{\text{half-life}} = 2$ months.

The interpretation of a coefficient of an endogenous statistic that is computed using a linear decay function may be slightly more complicated than for a regular count statistic (because we need to take the transpired time of past events into account) but possibly the interpretation is easier than for statistics with the exponential decay function because linear trends are relatively easy to understand. For example, if inertia would be equal to 3 and the decay function in Figure 1(c) is considered having a half-life of 2 months, the last event that was observed has a maximal contribution to the rate (and hazard) equal to 3, and events that occurred approximately 1, 2, 3, and 4 months or more contribute to the event rate with 2.25, 1.5, 0.75, and 0.

The linear decay model may be appropriate where the contribution of past events to the endogenous statistics (and thus to the logarithm of the event rate) is an approximately linear function of the transpired time, which, at some point, becomes approximately zero. Similar as the one-step model, the model is computationally convenient because one would not need to take the entire past history into account in the computation of the endogenous statistics. It may be somewhat less realistic however that the decay is assumed to be exactly 0. To our knowledge a linear decay model has not yet been considered for relational event modeling.

Normalizing decay functions and updating statistics

All the three weight decays start at 1, decay towards zero but are not normalized. However, they can be normalized by multiplying them with a normalizing constant of $\frac{\log(2)}{\theta}$ for the exponential

decay and $\frac{1}{\theta}$ for the one-step and the linear decay. The effect of the normalization of the weights directly translates into a re-scaling of the effect β of each endogenous statistic on the event rate, without changing nor re-scaling the estimate of the memory parameter. Normalization is recommended whenever it is needed to compare effects β across different parametrizations or across network dynamics with different memory parameter but following the same parametrization.

When endogenous statistics at each time point are updated according to one of the weight decays introduced in this section, the transpired time of events in the history is updated with respect to the time point that precedes the present one. For instance, if we need to update statistics at time t_m, the history of events that we are going to consider will be $E_{t_{m-1}}$, that is the collection of events from the onset until and including the event occurred at t_{m-1} and the time we consider to compute the time transpired of each event in the history will be the time of the last event in $E_{t_{m-1}}$, that is t_{m-1}. Therefore, since statistics are assumed to be updated at the last observed time point and not during the waiting time between two subsequent events, no right-censoring has to be taken into account in our analysis and the assumption of constant hazards during waiting times is not violated.

3. The profile log-likelihood in REM

The functions to model memory decay presented in Section 2 are three examples of univariate decays that can be embedded in the likelihood function of a REM [Butts (2008)] as well as in an actor-oriented model [DyNAM; Stadtfeld & Block (2017)]. In these decay functions, the memory parameter has support in $(0, +\infty)$. Thus, with the purpose of avoiding a constrained optimization for the memory parameter, we can re-parametrize it as $\theta = \exp\{\psi\}$, where $\psi \in \mathbb{R}$ is the natural logarithm of the memory parameter θ.

We now consider a sequence E_{t_M} of M relational events occurring among N actors where the likelihood function of a REM, which depends on the memory decay parameter ψ, is written as

$$
\begin{aligned}
&\mathcal{L}\big(\boldsymbol{\beta}, \psi; E_{t_M}\big) \\
&= \prod_{m=1}^{M} \left[\lambda\big(s_{e_m}, r_{e_m}, X_{e_m}, E_{t_{m-1}}, \boldsymbol{\beta}, \psi\big) \prod_{e' \in \mathcal{R}} \exp\left\{-\lambda\big(s_{e'}, r_{e'}, X_{e'}, E_{t_{m-1}}, \boldsymbol{\beta}, \psi\big) (t_m - t_{m-1})\right\} \right]
\end{aligned}
\tag{9}
$$

where at each time point a vector of endogenous and exogenous statistics in X_{e_m} is available for every possible dyad in the risk set ($\mathcal{R}$). Although we assume a time-invariant risk set the method can straightforwardly be applied to dynamic risk sets. Parameters $\boldsymbol{\beta}$ describe the effect of the statistics on the event rate and ψ represents the logarithm of the memory parameter under a specific memory decay, which is assumed to be the same for all the endogenous statistics. In the context of maximization of the likelihood function we are interested in finding the vector of parameters $(\boldsymbol{\beta}, \psi)$ that maximizes the likelihood given the observed sequence of events which is equivalent to minimizing the negative log-likelihood:

$$
\arg\min(\boldsymbol{\beta}, \psi)\{-\ln \mathcal{L}(\boldsymbol{\beta}, \psi; E_{t_M})\}
\tag{10}
$$

The optimization problem in (10) is generally solved by calculating the derivatives of the function up to and including the second order. In the case of a REM with an unknown memory parameter, endogenous statistics are no more sufficient for the estimation of the corresponding vector of effects $\boldsymbol{\beta}$, because their value changes depending on the value of the memory parameter ψ. Thus, only the sequence of events can be referred to as sufficient statistic both for the endogenous effects in $\boldsymbol{\beta}$ and for ψ. Moreover, derivatives for the memory parameter can either increase the computational burden or fail to exist (for instance, in the one-step decay function). In light of this, we can take advantage of the negative profile log-likelihood for a given memory parameter
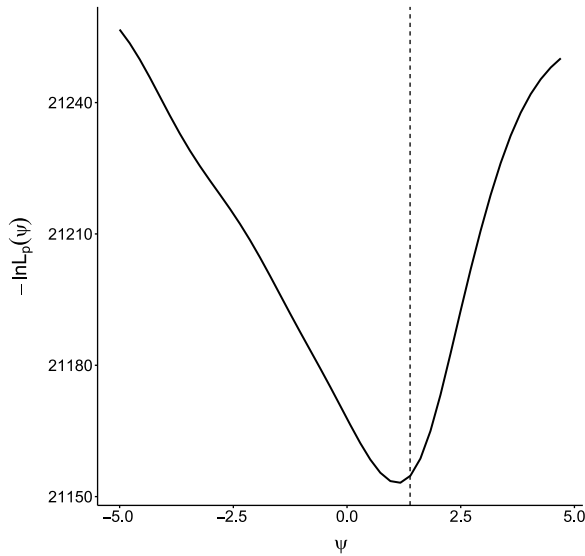
Figure 2. Negative profile log-Likelihood for the log-half-life parameter (exponential memory decay with true value $\psi =$ ln $(4) \approx 1.386$, dashed vertical line) for one randomly simulated event history.

and investigate whether the memory decay assumption is supported from the data and where the minimum potentially lies in. The profile negative log-Likelihood for ψ can be written as,

$$- \ln \mathcal{L}_{\mathrm{p}}(\psi) = \min_{\boldsymbol{\beta}} \{ - \ln \mathcal{L}(\boldsymbol{\beta}, \psi; E_{t_M}) \} \tag{11}$$

where the value of $- \ln \mathcal{L}_{\mathrm{p}}(\psi)$ is obtained as the minimum value of the negative log-likelihood where the memory parameter is fixed and the optimization is carried over $\boldsymbol{\beta}$ (as in a regular REM). Equation (11) comes down to one optimization for each fixed value of $\psi \in \mathbb{R}$. If there exists a minimum for $- \ln \mathcal{L}_{\mathrm{p}}(\psi)$, that value will correspond to the global minimum of both ψ and the optimized vector $\boldsymbol{\beta}$, thus they will be a solution for the optimization of the negative log-likelihood $- \ln \mathcal{L}(\boldsymbol{\beta}, \psi; E_{t_M})$.

An example of the negative profile log-likelihood based on one randomly simulated relational event history with an exponential memory decay is shown in Figure 2 where the function reaches its minimum close to the true value of the log-half-life parameter ($\psi = \ln (4) \approx 1.386$, indicated by the dashed vertical line). The slight deviation from the true value can be explained from random sampling (explored in more detail in the next section).

A drawback of the optimization of the negative Profile log-likelihood is that such methods do not provide a measure for the standard error of the memory parameter nor its covariances with the vector of effects $\boldsymbol{\beta}$. A way to estimate the accuracy of the estimate for the memory parameter and the related covariances consists in embedding the weight decay function in the model and in optimizing the complete log-Likelihood in (9). However, this approach requires the weight decay function to be differentiable at least twice.

Even though we cannot obtain standard errors for the memory parameters in a straightforward manner, the profile log-Likelihood can be used to quantify our relative uncertainty (from a Bayesian perspective) about different models that assume different values for the memory parameter as in a model selection problem. For example, when looking at the example data that was used in Figure 2, we could think of a set of models $\mathcal{M}_1 : \psi = -5.0$, $\mathcal{M}_2 : \psi = -4.9$, $\mathcal{M}_3 : \psi = -4.8, \ldots, \mathcal{M}_{101} : \psi = 5$. The Bayesian information criterion for model $\mathcal{M}_t$ [BIC; Schwarz (1978)] is then defined by

$$\mathrm{BIC}(\mathcal{M}_t) = k \log(M) - 2 \ln \mathcal{L}(\hat{\boldsymbol{\beta}}, \psi; E_{t_M}), \tag{12}$$

where $\hat{\boldsymbol{\beta}}$ is the MLE assuming the memory parameter of length k, and ψ is given under model $\mathcal{M}_t$. Thus, this BIC is directly available using standard statistical software. Consequently the Bayes factor (BF) between one model, say, $\mathcal{M}_{t_1}$ assuming a certain value for the memory parameter against another model, say, $\mathcal{M}_{t_2}$, assuming another value (possibly assuming another memory function as well), is then given by

$$\mathrm{BF}(\mathcal{M}_{t_1}, \mathcal{M}_{t_2}) = \exp\{\mathrm{BIC}(\mathcal{M}_{t_2})/2 - \mathrm{BIC}(\mathcal{M}_{t_1})/2\}, \tag{13}$$

which quantifies the relative evidence in the data in favor of $\mathcal{M}_{t_1}$ against $\mathcal{M}_{t_2}$. Thus, via this route we can even test non-nested models having different memory functions and assuming different memory parameters.

4. Simulations: Synthetic relational event histories with memory decay

Numerical simulations were conducted to explore the bias and the change in fit observed when memory decay values and/or decay parametrization are misspecified. Furthermore, we explored the performance of the BF to test between models with different memory decays. Finally, a simulation was carried out to investigate the behavior of the estimates in the scenario were the assumption of piece-wise-constant hazard is no longer met. The simulations studies under four different populations will be referred to as Simulations 1, 2 3, and 4.

Simulation 1: Exponential memory decay

In Simulation 1, 100 relational event histories are generated, each with $M = 5,000$ events occurring among $N = 20$ actors. The log event rate for any dyad $(i, j) \in \mathcal{R}$ at time t is specified as follows:

$$\ln \lambda(i, j, t) = \beta_{\mathrm{Intercept}} + \beta_{\mathrm{Dyadic}_1} \mathrm{Dyadic}_1(i, j) +$$

$$+ \beta_{\mathrm{Dyadic}_2} \mathrm{Dyadic}_2(i, j) + \beta_{\mathrm{Inertia}} \text{weighted-Inertia}(i, j, t, \theta_{\mathrm{half\text{-}life}}) +$$

$$+ \beta_{\mathrm{Reciprocity}} \text{weighted-Reciprocity}(i, j, t, \theta_{\mathrm{half\text{-}life}}) +$$

$$\beta_{\mathrm{TClosure}} \text{weighted-TClosure}(i, j, t, \theta_{\mathrm{half\text{-}life}}) +$$

$$+ \beta_{\mathrm{ABAY}} \mathrm{ABAY}(i, j, t) \tag{14}$$

where Dyadic_1 and Dyadic_2 are two exogenous variables that are time-invariant and asymmetric (i.e. $\mathrm{Dyadic}_1(i, j) \neq \mathrm{Dyadic}_1(j, i)$). Weighted inertia, weighted reciprocity, and weighted transitivity closure [TClosure, based on the definition presented in Arena et al. (2022)] are endogenous statistics based on a weighted count using an exponential memory decay with $\theta_{\mathrm{half\text{-}life}} = 4$ (with $\psi = \ln(\theta_{\mathrm{half\text{-}life}}) \approx 1.386$). ABAY is an endogenous turn-continuing participation shift (Butts, 2008) which does not follow any memory decay. The vector of true parameters is ($\beta_{\mathrm{Intercept}} = -3.5$, $\beta_{\mathrm{Dyadic}_1} = 0.5$, $\beta_{\mathrm{Dyadic}_2} = -0.3$, $\beta_{\mathrm{Inertia}} = 0.2$, $\beta_{\mathrm{Reciprocity}} = 0.3$, $\beta_{\mathrm{TClosure}} = 0.1$, $\beta_{\mathrm{ABAY}} = 0.2$).

Simulation 2: Linear memory decay

In Simulation 2, 100 relational event histories are generated, each with $M = 5,000$ events occurring among $N = 20$ actors. The log event rate for any dyad $(i, j) \in \mathcal{R}$ at time t is specified as in (14). However, in this simulation the memory decay for weighted inertia, weighted reciprocity, and weighted transitivity closure follows a linear decay with $\theta_{\mathrm{half\text{-}life}} = 4$ (with $\psi = \ln(\theta_{\mathrm{half\text{-}life}}) \approx 1.386$). The vector of true parameters is the same as the one used in Simulation 1 except for the Intercept which is $\beta_{\mathrm{Intercept}} = -3$.

Simulation 3: One-step memory decay

The same configuration is considered as for Simulation 2 but with a one-step memory decay for weighted inertia, weighted reciprocity and weighted transitivity closure using threshold $\theta_{\max} = 4$ (with $\psi = \ln(\theta_{\max}) \approx 1.386$).

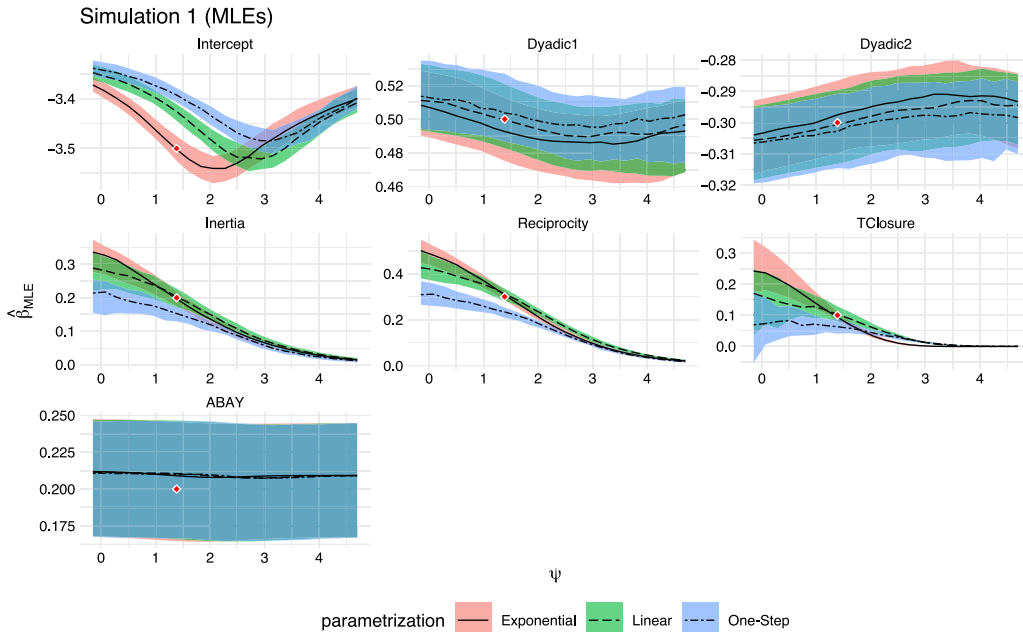*Simulation 4: Exponential memory decay and decreasing hazard*

In this simulation, 100 relational event histories are generated, each with $M = 5,000$ events occurring among $N = 20$ actors. To explore the effect of violations of the piece-wise constant hazard assumption, the waiting times are generated from a Weibull distribution where the shape parameter is assumed equal to 0.5. With such value of the shape parameter, hazards decrease over the waiting times. The scale parameter of the Weibull is still a function of the rates and the log event rate for any dyad $(i, j) \in \mathcal{R}$ at time t is specified as in (14) with the exception of the Intercept that is assumed $\beta_{\text{Intercept}} = -10$. The weight decay of the endogenous statistics follows the same exponential decay as in Simulation 1 with a half-life parameter $\theta_{\text{half-life}} = 4$ (with $\psi = \ln(\theta_{\text{half-life}}) \approx 1.386$).

4.1 Exploring bias based on a misspecified memory decay

The first purpose of the first three simulations studies is to understand whether and to what degree maximum likelihood estimates of the effects of the statistics in a REM are affected by the value of the memory parameter. This is important as memory decay is often prespecified in an ad hoc manner.

In Figures 3, 5, and 7, the trend for each estimated effect over the logarithm of the memory parameter, ψ, is shown under the three memory decays (exponential, linear and one-step). The shaded areas delimit the first and the third quartile of the distribution (based on the 100 simulations) of the estimated effect at any given value of ψ. The black lines show the trend of the median of each effect over the 100 simulations, and they have a different line type according to each parametrization. The diamond-shaped point marks the coordinates of the true memory parameter ($\ln(4) \approx 1.386$) and the true value of each specific effect. In all the simulations we see that all the endogenous variables which were assumed to follow a memory decay (Inertia, Reciprocity, and Transitivity Closure) as well as the Intercept are considerably affected by bias in the case of a misspecified memory parameter. Only if (i) the memory model assumed is the correct one and (ii) the memory parameter is around its maximum likelihood estimate, the distribution of the estimates across the simulations tend to concentrate around the true $\boldsymbol{\beta}$. As a consequence of this, it is evident that by not accounting for the memory parameter in the maximum likelihood optimization of a REM as well as not investigating different memory decays will likely lead the researcher to biased estimates.

Furthermore, Figure 4, 6, and 8 show a comparison between rescaled negative profile log-Likelihoods across 100 simulations within each simulation study (Simulation 1, 2, and 3). In each simulation study, each of the 100 simulated event sequences were optimized under each of the three parametrizations of the memory decay (Exponential, Linear, and One-Step). Therefore, for each event sequence a negative profile log-Likelihood is found for each parametrization. The set of negative profile log-Likelihoods under each parametrization and per each event sequence are then rescaled based on the global minimum across the three parametrizations and the local minimum within each parametrization, resulting in the new scale on the y-axis ($-\ln(L_{\mathrm{p}})_{\text{rescaled}}$). Each Figure shows three regions with different line types, one per each parametrization. Each region represents the (rescaled) value assumed by the 95% of the simulations in one parametrization across different values of the log-memory parameter (on the x-axis). The vertical dashed bold line marks the true value for the logarithm of the memory parameter ($\psi = \ln(4) \approx 1.386$). The results suggest that the proposed method using the profile log likelihood results in accurate estimates of

Figure 3. Simulation 1 (the true decay is exponential). Trend of the maximum likelihood estimates over the logarithm of the memory parameter, ψ, and under each of the three memory decays (exponential, linear, and one-step). The shaded regions delimit the first and the third quartile of the distribution (based on 100 simulated event sequences) of the estimated effect β over ψ. The black lines show the trend of the median of each effect across the 100 simulations, and they have a different line type according to each parametrization. The diamond-shaped point marks the coordinates of the true memory parameter $(\ln(4) \approx 1.386)$ and the true value of each specific effect.

the memory parameter in a well-specified model. Moreover, the true data generating model results in the best fit overall. Finally we see that in all three simulation studies, the three parametrizations result in roughly the same fit when towards small values of the memory parameter (negative values on the logarithmic scale) as well as towards larger values (greater than 3.0 on the logarithmic scale). This implies that in the case of a complete mismatch of the true decay parameter and the decay parameter that is used for model fitting, it does not matter which decay function would be used.

When including memory decay parameters in REMs it is also important to verify whether the assumption of proportional hazards is violated or not. In order to accomplish this, we analyzed the Schoenfeld's residuals (Schoenfeld, 1982) calculated for those endogenous statistics which are assumed to follow a weight decay (Inertia, Reciprocity and Transitivity Closure). In each of the three simulations, residuals in the 100 replicates distributed around zero and showed no trend over time, which implies that the assumption of proportional hazard was not violated.

4.2 Testing different decay functions via the Bayes factor

The second purpose of simulation studies 1, 2, and 3 is to explore the performance of the BF to test different memory models. We measured the relative evidence in favor of the true model given the simulated relational event sequence. In each of the three simulations, for every generated event sequence, we computed the BF of the model of the true weight decay function against the best model under the remaining other two decays using Equation (13) for each simulated dataset. The formulation of the BF is such that $\mathrm{BF}(\mathcal{M}_1, \mathcal{M}_2) > 1$ (<1) implies evidence for $\mathcal{M}_1(\mathcal{M}_2)$. If the BF is on the log scale the cutoff value equals 0. By investigating the distribution of the BF across
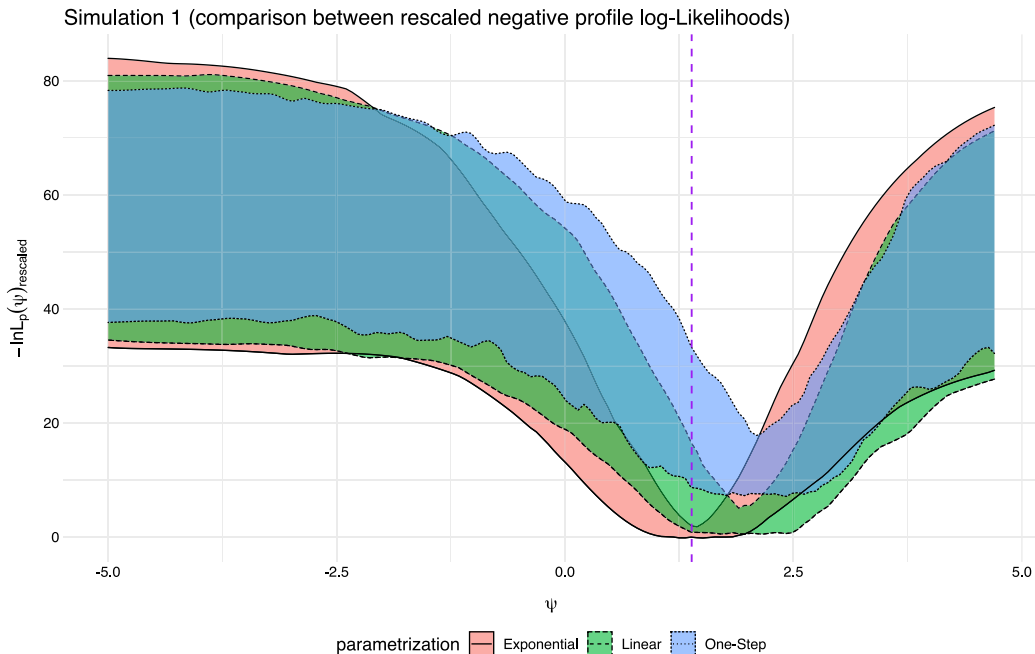
Figure 4. Simulation 1 (comparison between rescaled negative profile log-Likelihoods across simulations). The *y*-axis is the $-\ln(L_p)$ that is rescaled based on the global minimum across the three parametrizations and the local minimum within each parametrization. The figure shows three regions with different line types, one per each parametrization. Each region represents the (rescaled) value assumed by the 95% of the simulations in one parametrization across different values of the memory parameter (here on its logarithmic scale on the *x*-axis). The vertical dashed bold line marks the true value for the logarithm of the memory parameter ($\psi = \ln(4) \approx 1.386$). The three weight decays result in showing about the same evidence towards small values of the memory parameter (negative values on the logarithmic scale) as well as towards larger values (greater than 3.0 on the logarithmic scale). However, when in the neighborhood close to the true value of the memory parameter, the tree parametrizations show a diverging evidence, with the Exponential model being the lowest, which is the true parametrization used in the generation of the 100 event sequences.

the 100 sequences for all the simulations we get insights how well the BF can distinguish between different memory models. Figure 9 plots the distribution of the BFs.

In each of the three simulations, the distributions of the ln (BF) concentrates on positive values (>0), with at least the 95% of the generated sequences supporting the true memory decay and values of the BF (on its logarithmic scale) show a somewhat strong evidence as well. The worst performance was observed in the case of Simulation 1 (exponential decay) when it was compared with a linear decay, where the BF pointed towards the linear model in 20% of the simulated data sets. This result shows that the linear decay can potentially mimic an underlying exponential decay. Of course note that because the BF is consistent, the error probability would go to 0 when increasing the sample size of the simulated event history and the evidence for the true model would go to infinity.

4.3 Exploring estimation errors due to a misspecified hazard function

In Simulation 4, we explore the potential estimation error of the proposed REM with an exponential memory decay while the data were generated using Weibull waiting times (which violate the piece-wise-constant hazard assumption). Figure 10 shows the distribution of the endogenous and exogenous REM parameters $\boldsymbol{\beta}$ based on the optimized memory parameter ψ for each generated dataset. The Figure shows that the distribution of the maximum likelihood estimate of each
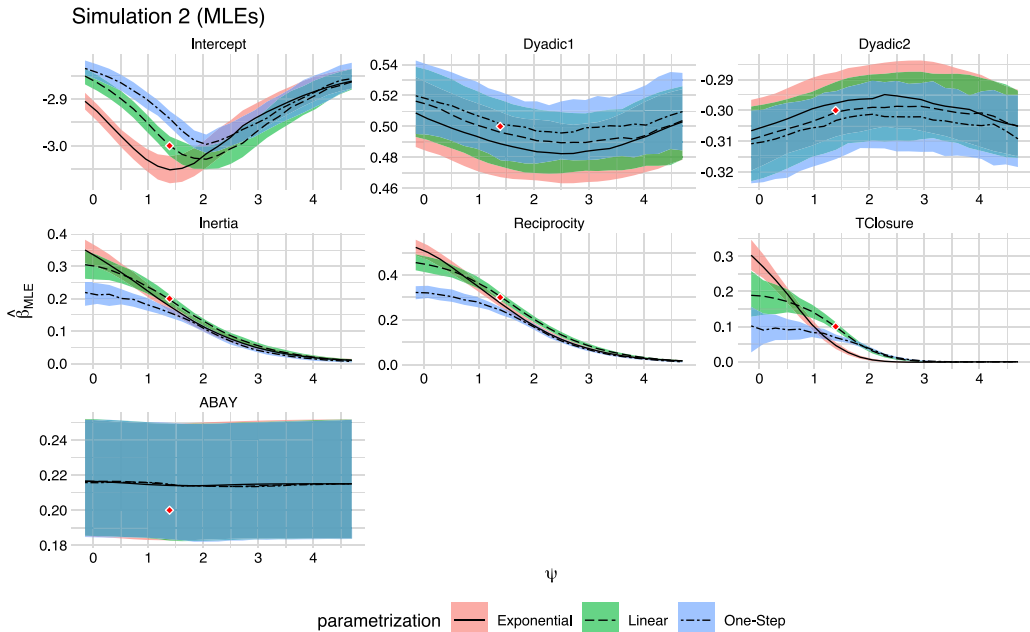
Figure 5. Simulation 2 (the true decay is linear). Trend of the maximum likelihood estimates over the logarithm of the memory parameter, ψ, and under each of the three memory decays (exponential, linear, and one-step). The shaded regions delimit the first and the third quartile of the distribution (based on 100 simulated event sequences) of the estimated effect β over ψ. The black lines show the trend of the median of each effect across the 100 simulations, and they have a different line type according to each parametrization. The diamond-shaped point marks the coordinates of the true memory parameter ($\ln(4) \approx 1.386$) and the true value of each specific effect.

effect is generally centered around its true value which suggests practically no clear estimation error, except for the intercept which roughly ranges between -4.95 and -4.70 while the true value was -10.0. This suggests that the decreasing hazard under the data generating model is mainly picked up by the intercept of the fitted model while leaving the other model parameters generally unchanged on average. This suggests that the estimated network parameters are safe to interpret in the case of a misspecified model due to a decreasing hazard underlying the data.

5. Investigate memory decay in empirical relational event networks

In this section we apply the methodology to the following two real-life case studies:

- A network of Indian socio-political actors sending demands to one another;
- A network of students sending text messages among each other.

The goal was to learn which memory decay function best describes the weight decrease of past events when modeling future events using endogenous network statistics. We also illustrate the impact of misspecified memory parameters on the network coefficients. Moreover, the fit and predictive performance of the best fitting memory decay model was compared with a model which ignores memory decay to study the importance of memory decay in empirical relational event networks. Finally we provide insights about the computational costs of the approach for relational events with different numbers of actors and different numbers of events.

Simulation 2 (comparison between rescaled negative profile log-Likelihoods)
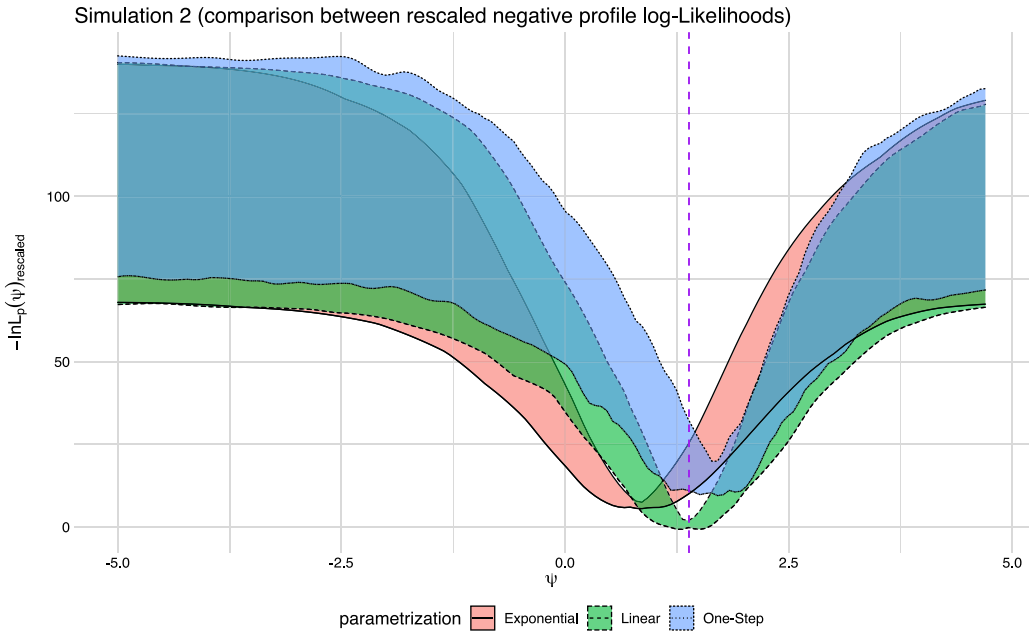
Figure 6. Simulation 2 (comparison between rescaled negative profile log-Likelihoods across simulations). The *y*-axis is the $-\ln(L_p)$ that is rescaled based on the global minimum across the three parametrizations and the local minimum within each parametrization. The figure shows three regions with different line types, one per each parametrization. Each region represents the (rescaled) value assumed by the 95% of the simulations in one parametrization across different values of the memory parameter (here on its logarithmic scale on the *x*-axis). The vertical dashed bold line marks the true value for the logarithm of the memory parameter ($\psi = \ln(4) \approx 1.386$). The three weight decays result in showing about the same evidence towards small values of the memory parameter (negative values on the logarithmic scale) as well as towards larger values (greater than 3.0 on the logarithmic scale). However, when in the neighborhood close to the true value of the memory parameter, the tree parametrizations show a diverging evidence, with the Linear model being the lowest, which is the true parametrization used in the generation of the 100 event sequences.

5.1 Demands among Indian socio-political actors

The Indian data were retrieved from the Integrated Crisis Early Warning System (Boschee et al., 2015). This database is available in the Harvard Dataverse repository and it collects relational events that describe interactions (found in news articles) occurring between socio-political actors all over the world. We focus our analysis on the sequence of requests that were recorded within the Indian territory. In the original data, such requests were further classified in humanitarian, military or economic ones but we avoid such distinction in our analysis.

The relational event sequence consists of $M = 7{,}567$ demands recorded between June 2012 and April 2020 and sent among the 10 most active actor types: citizens, government, police, member of the Judiciary, India, Indian National Congress Party, Bharatiya Janata Party, ministry, education sector, and "other authorities." The time variable is recorded at a daily level, therefore events that co-occurred are considered evenly spaced throughout a day and the memory parameter is measured in days. The logarithm of the rate (λ) for the demand sent by actor i to actor j at time t is modeled as

$$\ln \lambda(i, j, t) = \beta_{\text{Intercept}} + \beta_{\text{Inertia}} \text{weighted-Inertia}(i, j, t, \psi)$$
$$+ \beta_{\text{Reciprocity}} \text{weighted-Reciprocity}(i, j, t, \psi) + \beta_{\text{TClosure}} \text{weighted-TClosure}(i, j, t, \psi)$$
$$+ \beta_{\text{ABAY}} \text{ABAY}(i, j, t) \tag{15}$$
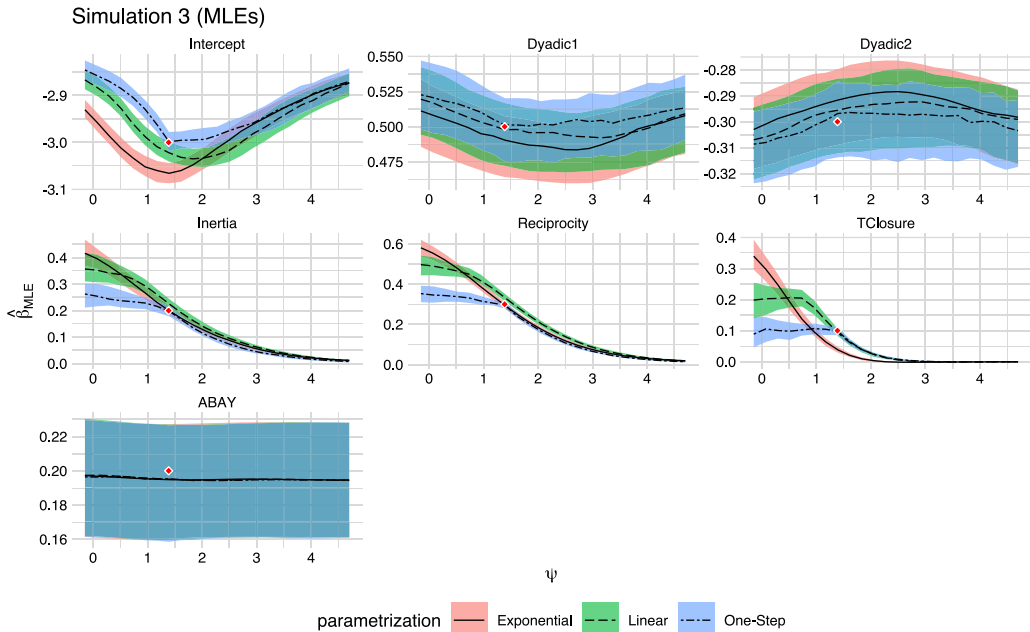
Figure 7. Simulation 3 (the true decay is one-step). Trend of the maximum likelihood estimates over the logarithm of the memory parameter, ψ, and under each of the three memory decays (exponential, linear, and one-step). The shaded regions delimit the first and the third quartile of the distribution (based on 100 simulated event sequences) of the estimated effect β over ψ. The black lines show the trend of the median of each effect across the 100 simulations, and they have a different line type according to each parametrization. The diamond-shaped point marks the coordinates of the true memory parameter ($\ln(4) \approx 1.386$) and the true value of each specific effect.

where weighted-Inertia, weighted-Reciprocity and weighted-Transitivity Closure (TClosure) are assumed to follow a weight decay governed by ψ, the logarithm of the memory parameter.

We first investigated the three weight decays presented in Section 2 (exponential, linear, and one-step decay) by optimizing the negative log-Likelihood for each model (a plot of the $-\ln L_p(\psi)$ is shown in Figure 11). Finally we chose the best fitting model among the three models, that is the one with the lowest BIC. In Table 1 the BIC's of the three best models are reported along with the BIC of a model in which no memory decay was specified. In the same table, the log-BF is calculated by considering as reference the model with the lowest BIC among the three, that is the exponential one. We see that there is convincing evidence in the data in favor of the exponential decay model as the data were approximately $\exp(10.97)$, $\exp(58.27)$, and $\exp(1952.88)$ times more likely under the exponential model than under the linear decay model, the one-step model, and the model without memory, respectively. Thus, we choose to continue the analysis with the exponential decay model. In Figure 12 the trend of the MLEs is plotted over the logarithm of the memory parameter (ψ) for the exponential decay model. Again, we see a considerable impact of the choice of the memory parameter which suggests that choosing the memory parameter in an ad hoc manner is not advised. For example, we see that transitivity closure can vary from approximate 0 to more than 1 within the considered range of the memory parameter.

The estimated half-life of the exponential memory decay in this network is $\exp(\hat{\psi}) \approx 64$ days. Thus the weight of past requests tends to halve after about 2 months. This case study has been already shown in Arena et al. (2022) where a semi-parametric strategy was applied to model memory decay by means of an ensemble of many step-wise decay models. In that analysis, which does not make parametric assumptions about memory decay function (and is therefore computationally much more expensive), the shape of the decay also followed an approximate

Simulation 3 (comparison between rescaled negative profile log-Likelihoods)
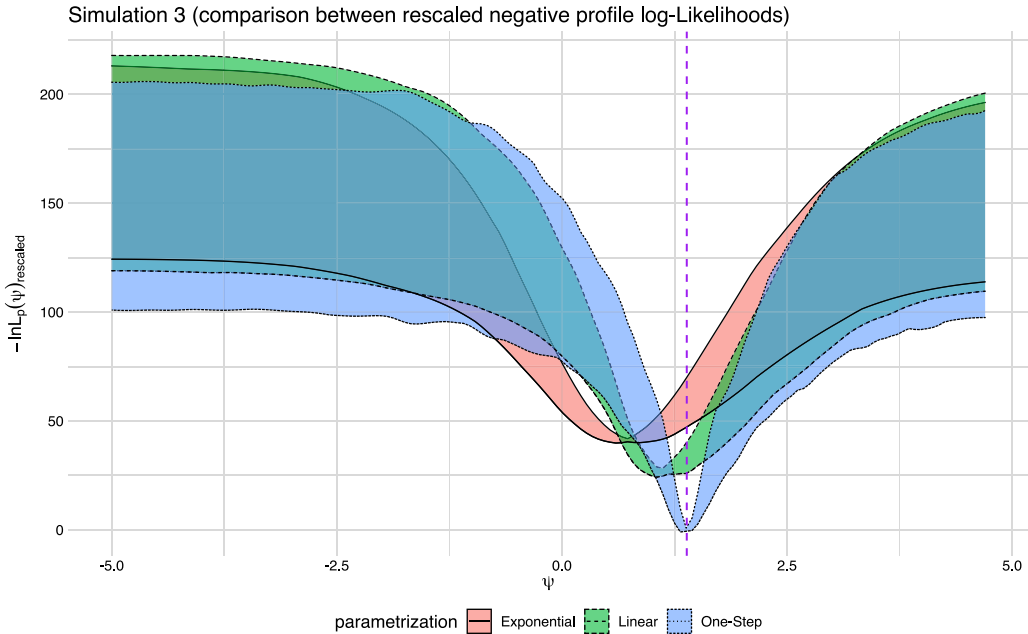
Figure 8. Simulation 3 (comparison between rescaled negative profile log-Likelihoods across simulations). The *y*-axis is the $-\ln(L_p)$ that is rescaled based on the global minimum across the three parametrizations and the local minimum within each parametrization. The figure shows three regions with different line types, one per each parametrization. Each region represents the (rescaled) value assumed by the 95% of the simulations in one parametrization across different values of the memory parameter (here on its logarithmic scale on the *x*-axis). The vertical dashed bold line marks the true value for the logarithm of the memory parameter ($\psi = \ln(4) \approx 1.386$). The three weight decays result in showing about the same evidence towards small values of the memory parameter (negative values on the logarithmic scale) as well as towards larger values (greater than 3.0 on the logarithmic scale). However, when in the neighborhood close to the true value of the memory parameter, the tree parametrizations show a diverging evidence, with the One-Step model being the lowest, which is the true parametrization used in the generation of the 100 event sequences.

exponential shape, which is the same as we find here using the parametric approaches presented in this paper.

In Table 2, the estimates of the effects $\boldsymbol{\beta}$ at the optimized memory parameter are reported. When interpreting these coefficients it is important to note that the memory function was normalized such that the surface underneath the line equals 1. Given the half-life parameter of 64 days, this implies that the function in Figure 1(b) would be multiplied with $\ln(2)/64$. Thus, given the estimated inertia effect of 6.9, the contribution of the last observed event to the last observed dyad is equal to a factor of $\exp(6.9 \times (\ln(2)/64)) \approx 1.08$, that is, an increase of 8% (which is the maximal contribution), and if an event was observed for this dyad approximately 64 days ago, this would have resulted in a contribution to the rate with a factor of $\exp(6.9 \times (\ln(2)/64) \times 0.5) \approx 1.04$, that is, an increase of 4%.

To illustrate the importance of modeling memory decay, we evaluated the predictive performance of the best fitting REM with an exponential decay function and compared it with a REM which ignores memory decay by giving all past events an equal weight.

The plot in Figure 13 shows the ROC curves of both models which clearly shows the superior performance of the model with an exponential memory decay function.

5.2 Text messages among students

The sms data consist of a sequence of text messages sent among a group of university students (freshmen) during a period of four weeks. The original event sequence is part of the interaction
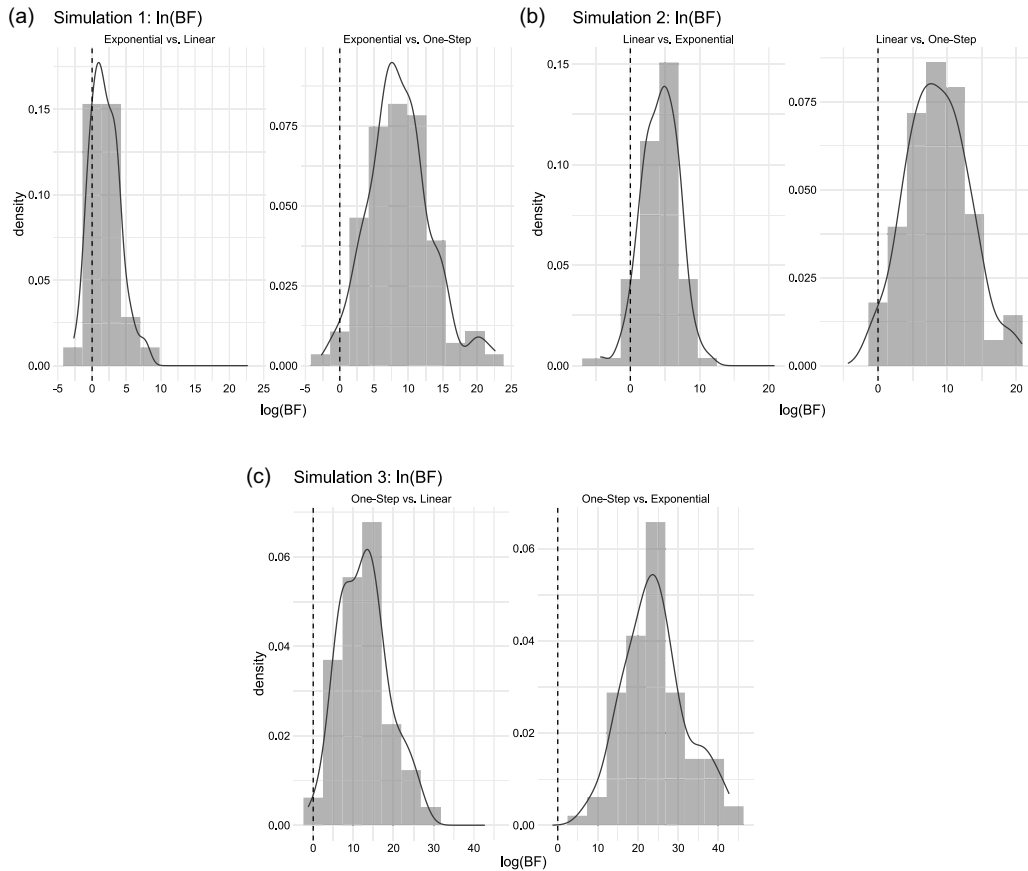
Figure 9. Distribution of ln (BF) (where ln (BF) > 0 translates to as evidence in favor of the true model) : (a) in Simulation 1 the true weight decay is exponential, thus the two Bayes factors are BF($M_{Exponential}$, M_{Linear}) and BF($M_{Exponential}$, $M_{One\text{-}Step}$) and the number of simulations where ln (BF) > 0 is respectively 80 and 98 out of 100; (b) in Simulation 2 the true weight decay is linear, thus the two Bayes factors are BF(M_{Linear}, $M_{Exponential}$) and BF(M_{Linear}, $M_{One\text{-}Step}$) and the number of simulations where ln (BF) > 0 is respectively 95 and 98 out of 100; (c) in Simulation 3 the true weight decay is one-step, thus the two Bayes factors are BF($M_{One\text{-}Step}$, M_{Linear}) and BF($M_{One\text{-}Step}$, $M_{Exponential}$) and the number of simulations where ln (BF) > 0 is respectively 99 and 100 out of 100.

data collected in the Copenhagen Networks Study (Sapiezynski et al., 2019) and it consists of 568 students and 24,333 events (number of text messages).

We ran the same analysis on three sub-sequences of events (increasing in both number of actors and number of events) so as to have a better understanding of the computational complexity of the methodology presented in this paper as well as to explore the method for networks of different sizes, both in terms of the number of actors and the number of events. For the selection of the three sub-sequences: (i) we ran a clustering algorithm that works on the optimization of a modularity score (Clauset et al., 2004; Csardi & Nepusz, 2006), (ii) we sorted the clusters based on the length of the event sequences, from the longest to the shortest, and (iii) we considered three sub-networks where the first was based on the first cluster of actors, the second was based on the first two clusters and, finally, the third was based on the first eight clusters. Each sub-sequence of events also includes the interactions between actors belonging to a different cluster. In Table 3, we show the size of each network in terms of number of students (# Actors) and text messages sent (# Events).
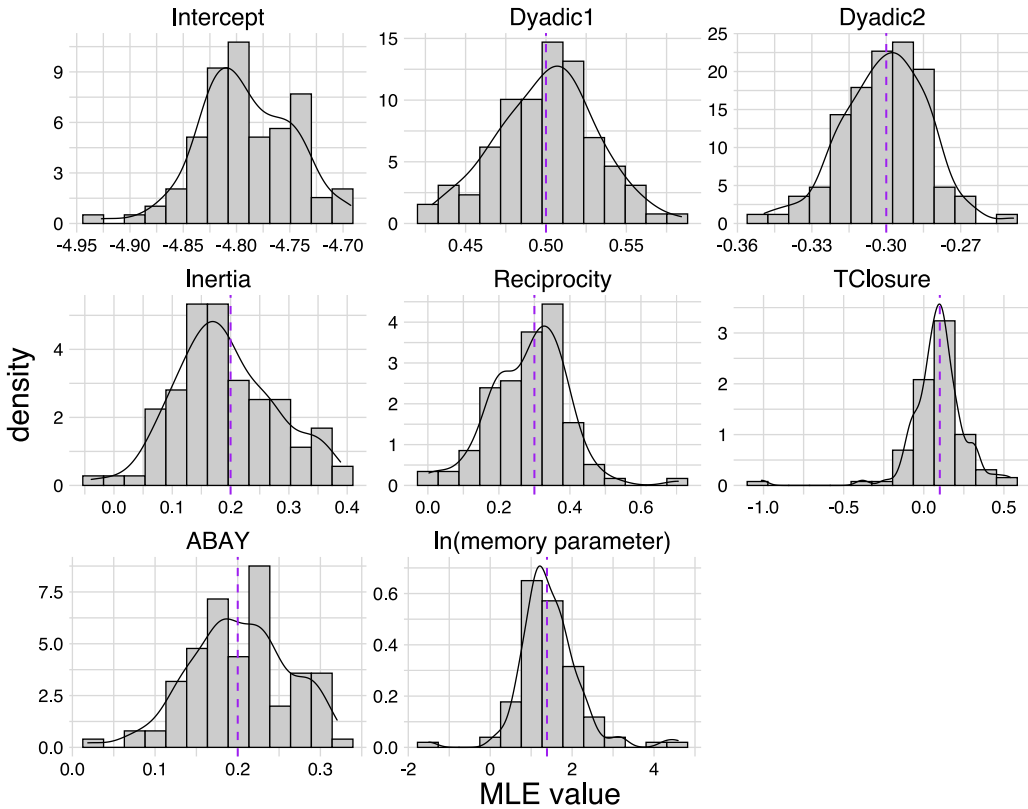
Figure 10. Simulation 4 (the true waiting time in the 100 simulated event sequences is distributed as a Weibull with shape parameter equal to 0.5). Distribution of the maximum likelihood estimates of the effects of the statistics as well as the memory parameter in a REM (with piece-wise constant hazard assumption). Each vertical dashed line corresponds to the true value of each effect.

In all the three selected relational event sequences, the time variable is available as timestamp which is converted to hours transpired since the beginning of the observation time. Thus the memory parameter will be measured in hours as well. In addition to the event sequence we also know the gender of the students and whether they are friends on Facebook or not. With these two information we specified two dyadic variables: (1) SameGender which assumes the value 1 if the two actors interacting have the same gender, 0 otherwise; (2) FBfriends that assumes the value 1 where the sender and the receiver of the text message are friends on Facebook, 0 otherwise.

We specify the same model for the three sub-networks; thus, the logarithm of the rate (λ) for a text message sent by actor i to actor j at time t is modeled as

$$
\begin{aligned}
\ln \lambda(i, j, t) = {} & \beta_{\text{Intercept}} + \beta_{\text{SameGender}} \text{SameGender}(i, j) + \beta_{\text{FBfriends}} \text{FBfriends}(i, j) \\
& + \beta_{\text{Inertia}} \text{weighted-Inertia}(i, j, t, \psi) + \beta_{\text{Reciprocity}} \text{weighted-Reciprocity}(i, j, t, \psi) \\
& + \beta_{\text{ABAY}} \text{ABAY}(i, j, t)
\end{aligned}
\tag{16}
$$

Also in this application weighted-Inertia and weighted-Reciprocity are assumed to follow a weight decay governed by ψ and the three parametrizations were examined, in the same fashion as with the Indian data.

In Figure 14, we see that for all three networks the negative profile log-Likelihood for the exponential model is lowest, suggesting the best fit for an exponential decay. This is also confirmed

Table 1. (**Indian data**) BIC of the best model (where the memory parameter is optimized) under each of the three memory decays (exponential, linear, and one-step) and for the model w/o memory. The lowest BIC is the one of the exponential model (56,753.55), and the two log-Bayes-factor are calculated based on the following model comparisons: BF($\mathcal{M}_{\text{Exponential}}, \mathcal{M}_{\text{Linear}}$), BF($\mathcal{M}_{\text{Exponential}}, \mathcal{M}_{\text{One-Step}}$), and BF($\mathcal{M}_{\text{Exponential}}, \mathcal{M}_{\text{w/o memory}}$).

	Decay			
	Exponential	Linear	One-Step	w/o memory
BIC	56,753.55	56,775.5	56,870.08	60,659.31
ln (BF)	-	10.97	58.27	1952.88

Table 2. (**Indian data**) Maximum likelihood estimates for the exponential decay. The estimate of the logarithm of the memory parameter is 4.156, that is an half-life of exp(4.156) $\approx$ 64 days. Estimates of effects β are all significant.
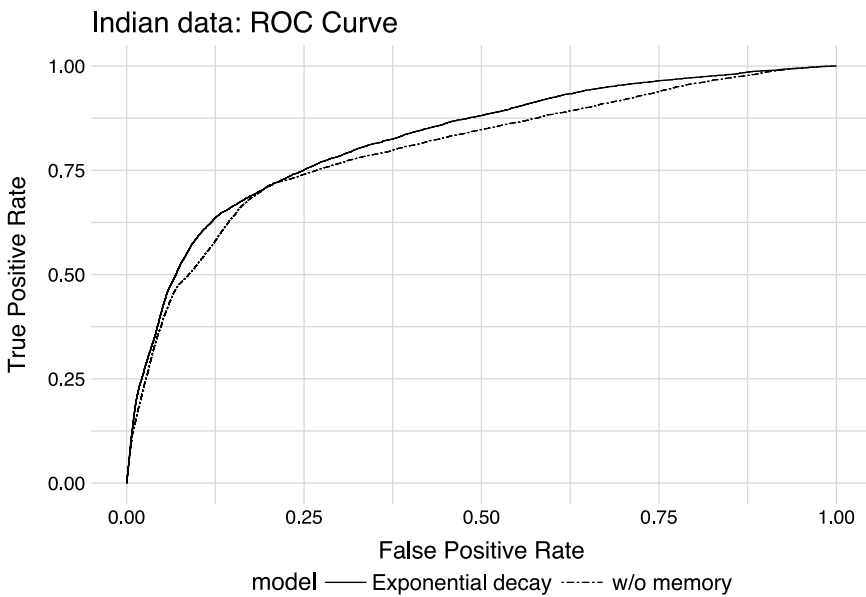
	Exponential decay					
	$\hat{\psi}$	Intercept	Inertia	Reciprocity	TClosure	ABAY
$\hat{\beta}$	4.156	−4.373	6.912	2.888	0.146	0.229
$se(\hat{\beta})$	-	0.018	0.074	0.111	0.005	0.033
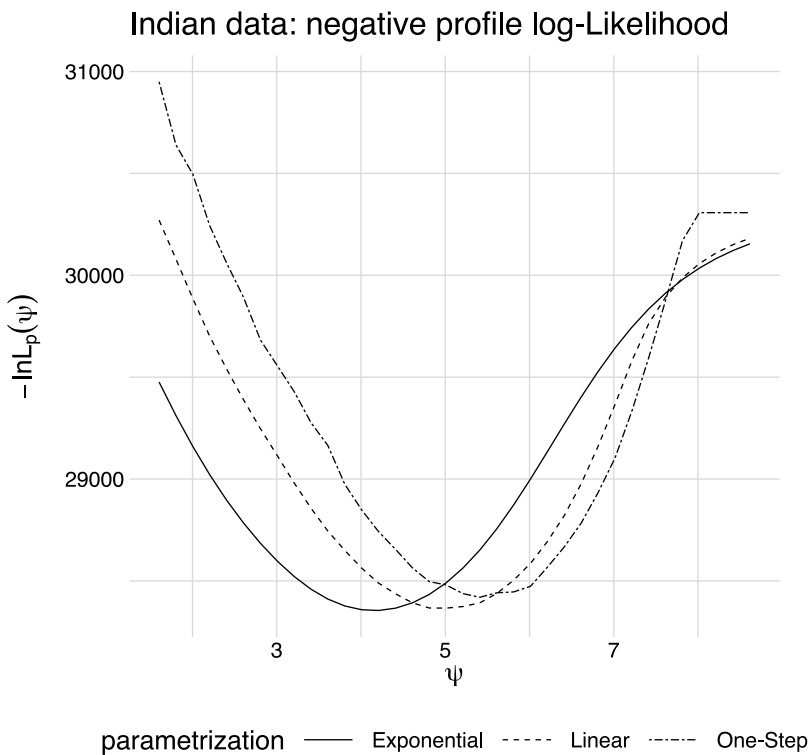
Figure 11. (**Indian data**) negative profile log-Likelihood ($-\ln L_p(\psi)$) under each of the three memory decays (exponential, linear, and one-step).

Indian data: trend of MLEs when the weight decay is exponential
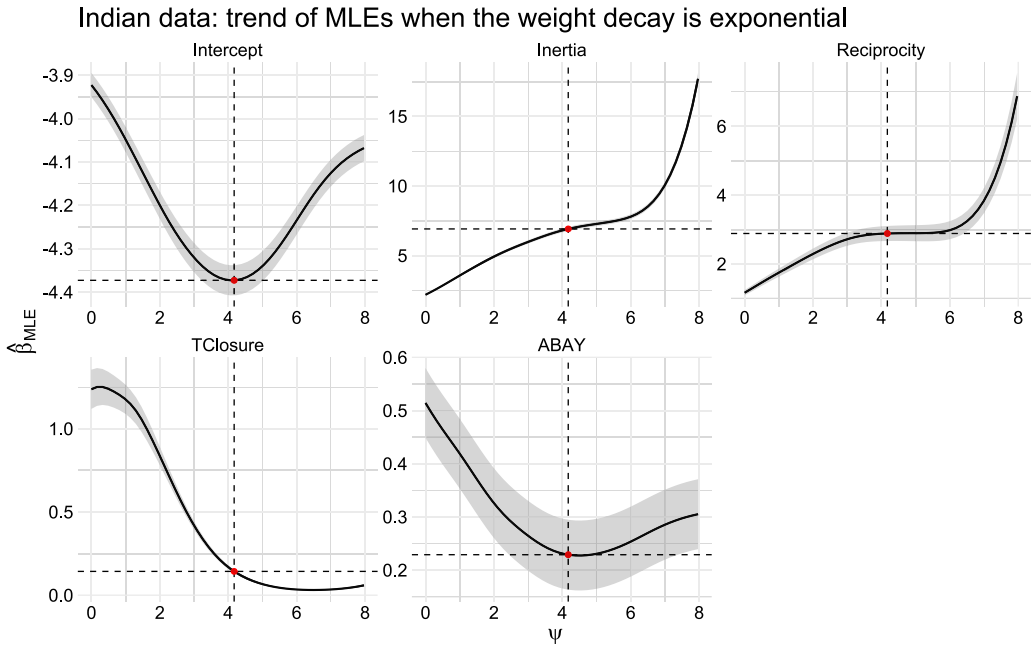
Figure 12. (**Indian data**) trend of the maximum likelihood estimates (MLEs) for the exponential decay over ψ (logarithm of the memory parameter). The dashed black lines in each plot mark the estimate for the log-memory-parameter $\hat{\psi}_{MLE}$ (vertical lines) and the estimates of the effects $\boldsymbol{\beta}$ (horizontal lines) at the corresponding $\hat{\psi}_{MLE}$. The shaded regions are the 95% confidence intervals for the effects $\boldsymbol{\beta}$ estimated at any value of ψ.


Figure 13. (**Indian data**) ROC curve of model with exponential memory decay and model without memory.

by comparing the three optima, thus by the BIC's and the BFs shown in Table 4. For the exponential model, the trend of the estimates $\boldsymbol{\beta}$ over ψ for the model with eight clusters are shown in Figure 15, where the dot marks the maximum likelihood for each effect at the $\hat{\psi}_{MLE}$. The trends

Table 3. (**Sms data**) Dimensions of the three sub-networks used in the example

| Network of sms	# Actors	# Events
1 cluster	23	3,678
2 clusters	53	7,311
8 clusters	199	13,943
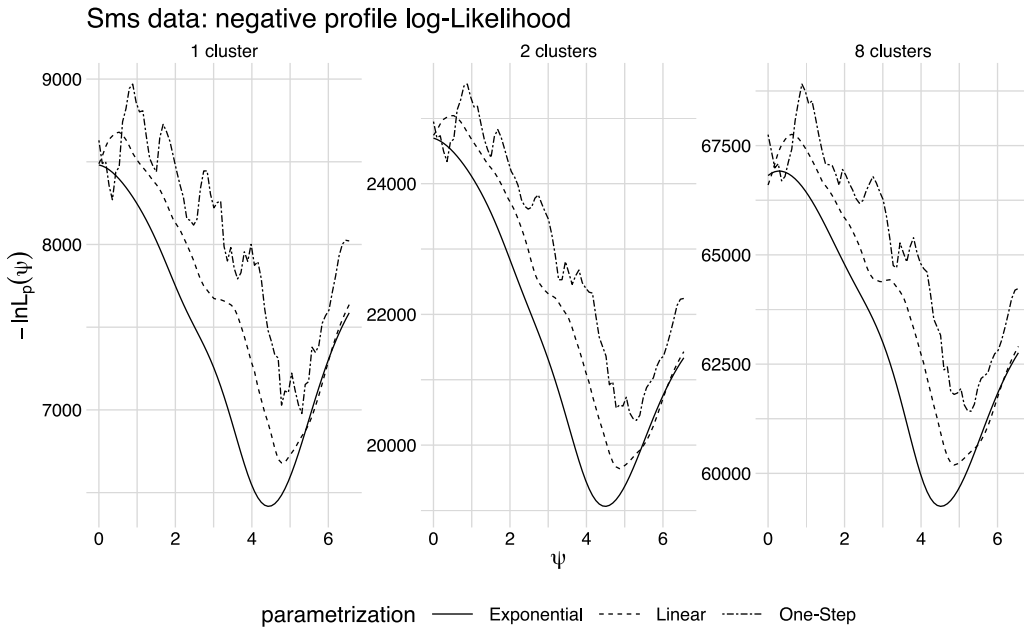
Figure 14. (**Sms data**) negative profile log-Likelihood $(-\ln L_p(\psi))$ under each of the three memory decays (exponential, linear, and one-step) and for each sub-network (one cluster, two clusters, and eight clusters).

of the other two networks (1 cluster and 2 clusters) are shown in Appendix A2. The optimal half-life ranges between approximately 84 and 92 h, which implies that text messages become half as important to predict future observations after a little bit more than 3.5 days.

The maximum likelihood estimates for the exponential model regarding the three networks are shown in Table 5 using normalized decay functions, which should be taken into account when interpreting the endogenous effects. For example, for the network based on eight clusters, inertia was estimated to be equal to 1.423, which implies that the rate of the last observed dyad is multiplied with $\exp(1.423 \times (\ln(2)/\exp(4.519))) \approx 1.011$, which implies an increase of about 1.1%, and if an event was observed, say, $\exp(4.519) \approx 92$ hours ago, this would have resulted in an increase of about 0.5% of the rate. Furthermore, both the variables SameGender and FBfriends show clear effects on the event rate. A negative effect for SameGender suggests that the text messages are more likely to be exchanged between students of a different gender. Indeed, the parameter $\hat{\beta}_{\text{SameGender}} = -0.714$ suggests that the hazard (sms) rate for a dyad where both actors have the same gender will be $(\exp(-0.714) - 1) \approx -51\%$ lower than the rate in which the two actors have different gender (holding all the other statistics constant). The effect for the variable FBfriends shows that the sequence of sms is strongly represented by students that are also friends on Facebook, since such variable results to have a large positive effect on the sms rate.

Table 4. (**Sms data**) Per each sub-network (one cluster, two clusters, eight clusters) the BIC of the best model (where the memory parameter is optimized) under each of the three memory decays (exponential, linear, and one-step) and for the model w/o memory. In all the sub-networks, The lowest BIC is the one of the exponential model, and the two log-Bayes-factor are calculated based on the following model comparisons: BF($\mathcal{M}_{\text{Exponential}}, \mathcal{M}_{\text{Linear}}$), BF($\mathcal{M}_{\text{Exponential}}, \mathcal{M}_{\text{One-Step}}$) and BF($\mathcal{M}_{\text{Exponential}}, \mathcal{M}_{\text{w/o memory}}$).

		Decay			
		Exponential	Linear	One-Step	w/o memory
1 cluster	BIC	12,882.49	13,408.32	14,004.85	16,092.17
	ln (BF)	-	262.92	561.18	1,604.84
2 clusters	BIC	38,178.28	39,332.40	40,810.29	44,540.81
	ln (BF)	-	577.06	1,316.01	3,181.26
8 clusters	BIC	118,548.10	120,438.30	122,899.70	128,520.70
	ln (BF)	-	945.12	2,175.85	4,986.32

Sms data (8 clusters): trend of MLEs when the weight decay is exponential
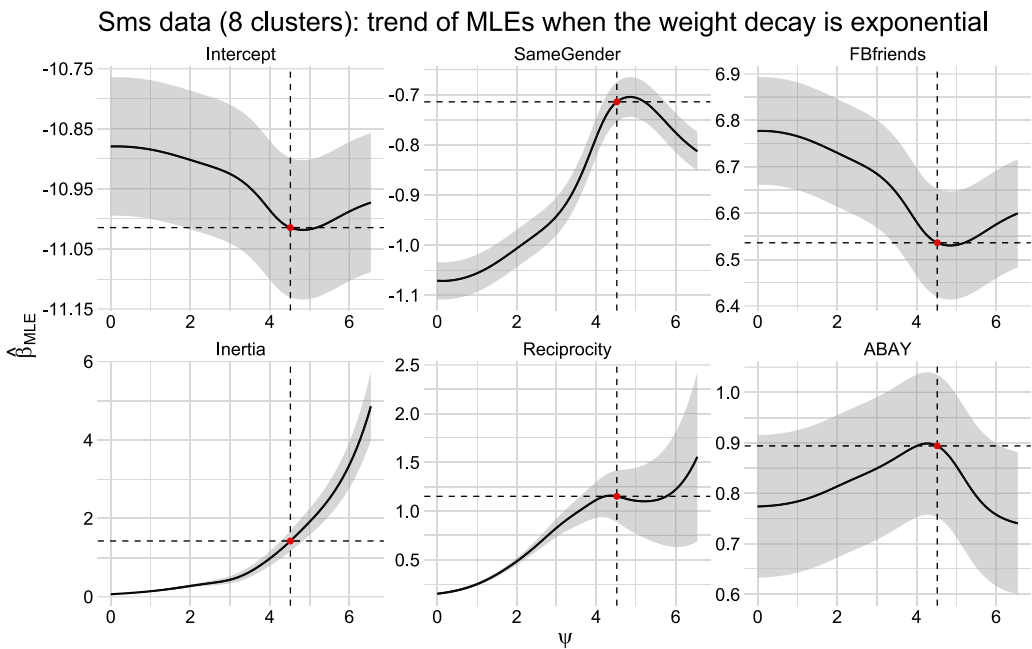
Figure 15. Sms data (eight clusters): trend of the maximum likelihood estimates (MLEs) for the exponential decay over ψ (logarithm of the memory parameter). The dashed black lines in each plot mark the estimate for the log-memory-parameter $\hat{\psi}_{\text{MLE}}$ (vertical lines) and the estimates of the effects $\boldsymbol{\beta}$ (horizontal lines) at the corresponding $\hat{\psi}_{\text{MLE}}$. The shaded regions are the confidence intervals at 0.95 for the effects $\boldsymbol{\beta}$ estimated at any value of ψ. For the Transitivity Closure (TClosure) estimates are plotted for an interval of ψ to make the trend much more readable.

To get more insights about the predictive performance of the best fitting exponential decay model in comparison to a model which ignores memory decay, we checked the ROC curves. These are displayed in Figure 16. Again we see that there is an improvement in predictive performance of the exponential decay model over the model without memory decay. The improvement is relatively small for the data based on 8 clusters.

Table 5. (**Sms data**) Maximum likelihood estimates for the exponential decay in each of the three sub-networks (1 cluster, 2 clusters, 8 clusters). The estimate of the logarithm of the memory parameter ($\exp(\hat{\psi})$) ranges approximately between 84 and 92 h in the three networks. Estimates of effects β are overall significant.

					Exponential decay			
		$\hat{\psi}$	Intercept	SameGender	FBfriends	Inertia	Reciprocity	ABAY
1 cluster	$\hat{\beta}$	4.438	−8.706	−1.357	5.745	0.662	0.984	−0.665
	$se(\hat{\beta})$	-	0.214	0.056	0.215	0.175	0.175	0.139
2 clusters	$\hat{\beta}$	4.491	−9.214	−1.392	5.691	1.228	0.852	−0.407
	$se(\hat{\beta})$	-	0.109	0.036	0.110	0.134	0.134	0.110
8 clusters	$\hat{\beta}$	4.519	−11.015	−0.714	6.536	1.423	1.152	0.894
	$se(\hat{\beta})$	-	0.059	0.020	0.059	0.135	0.135	0.072
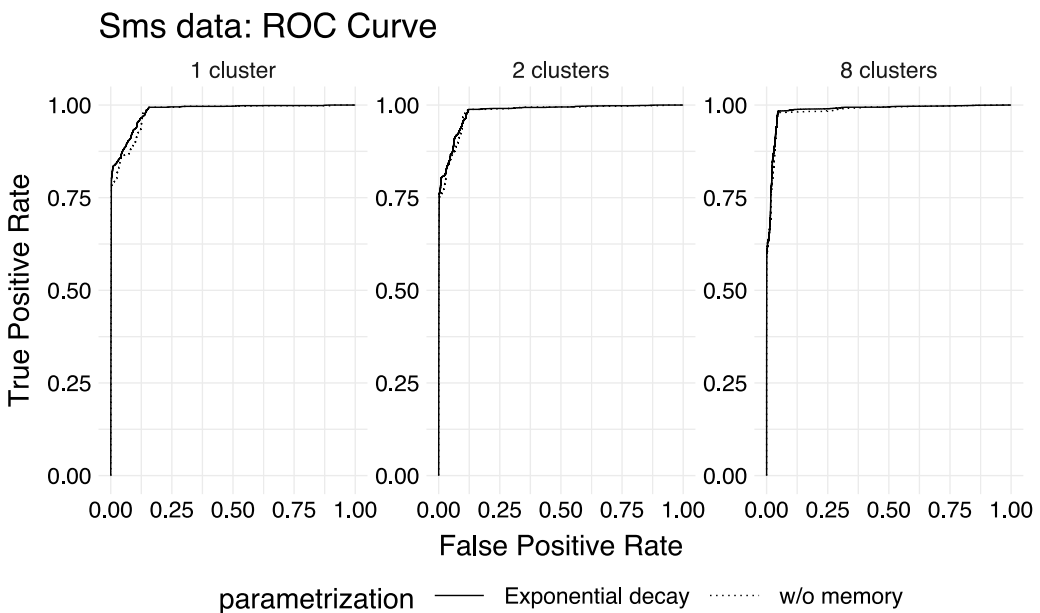
Figure 16. (**Sms data**): ROC curve of model with exponential memory decay and model without memory for the three sub-networks (one cluster, two clusters, and eight clusters).

The final objective of this study was to provide insights about the computational time of the methodology by considering the needed time for one iteration in the estimation stage. We ran one iteration for the optimization of the parameter β by fixing the log-memory parameter ψ to the maximum likelihood estimate from the sub-sequence with 8 clusters. We considered the three sub-sequences and for each sub-sequence, we run the estimation of the effects β over an increasing sequence length of 1,800, 3,600, 7,200 and, 13,000 events. Per each sequence length, we run the estimation 100 times. In Figure 17 the median (over 100 repetitions) running time for one iteration in the optimization stage is shown across both sub-networks (one, two, and eight clusters) and sequence lengths. The model that is estimated in the optimization stage is the same one introduced in the data example in Section 5.2. The time is reported in seconds on the y-axis. The sequence length on the x-axis is the number of events for the specific sub-network. The first

Sms data: median running time of one iteration in the optimization stage
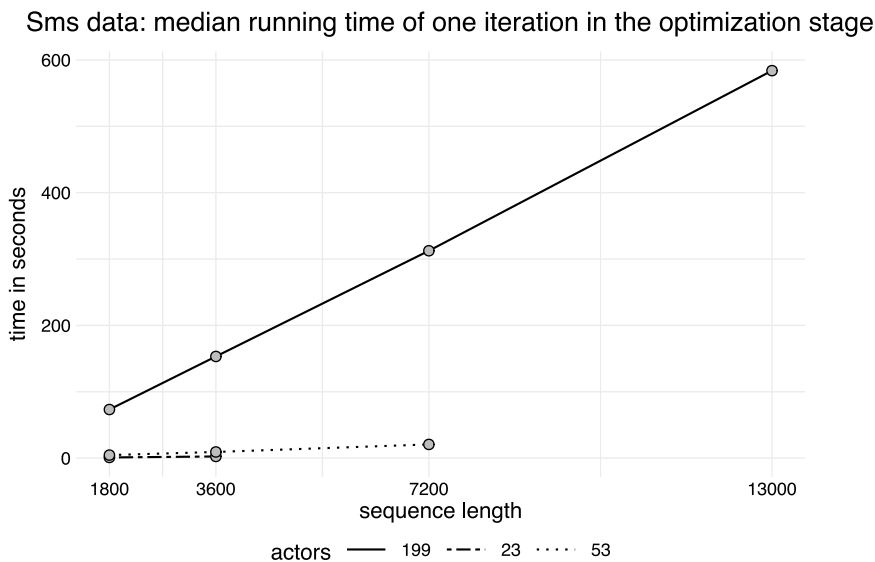
Figure 17. (**Sms data**): Median running time of one iteration in the optimization stage. The model that is estimated in the optimization stage is the same one introduced in the data example in Section 5.2. The time is reported in seconds (on the *y*-axis), the sequence length is the number of events considered (on the *x*-axis). The first and the second sub-sequence (networks with one and two clusters) do not show running times for larger lengths because they reach their maximum length (see Table 3).

and the second sub-sequence (networks with one and two clusters) do not show running times for lengths larger than their maximum length (see Table 3). We observe that the median running time of one iteration follows a linear increasing trend with a slope that becomes larger for the networks with a higher number of actors. This shows that the computation is feasible for small networks of 23 actors to fairly large networks of 199 actors.

6. Discussion

In the literature on relational event networks, weight decay functions have been used to capture the decreasing importance of past events to compute endogenous network statistics as a function of the transpired time. To achieve this, a parametric function can be chosen to model the decay of the weight of past events together with a chosen memory parameter that describes the speed of the decay or memory length. In previous studies both the decay function and the memory parameter governing have often been pre-specified in a fairly ad hoc manner. As an alternative, the method presented in this paper allows one to find the best fitting decay function and memory parameter using the observed sequence of relational events by inspecting several parametric decay functions.

The simulation studies and empirical applications in this paper showed that a misspecification of the shape of the memory decay and/or a misspecification of the memory parameter lead to biased estimates of the effects of (endogenous) statistics, and consequently this may result in incorrect conclusions about the temporal interaction behavior in the network. This was shown by visualizing the trends of the estimated effects as a function of the memory parameter. For this reason, it is not recommended to use memory functions or memory parameters that are arbitrarily specified. Instead we recommend to optimize the decay using the observed data as our studies revealed that such biases are generally avoided in that case. Hence, we recommend network researchers who are interested to learn how the past affects the future in relational event networks to estimate the memory decay in the endogenous statistics using the proposed methodology.

Of course, this paper only considered three possible parametric functions to capture memory decay, all using one memory parameter. Many more single-parameter functions could be considered. Moreover, the methodology could be extended to functions with two or more unknown parameters (e.g., smoothed-one step decay, negative power decay, hyperbolic-like decay which also allows for a long-term memory plateau). We leave this for future research. Decays depending on a multiple parameters of course add complexity both to the optimization stage and to the interpretation of the parameters. For this reason, the use of univariate memory models may be preferred as a first step to study memory decay in empirical relational event networks.

Another important direction for future research is to improve the estimation of the memory parameter that allows a quantification of its uncertainty, and how this transcends to the network parameters. Both classical likelihood methods as well as full Bayesian approaches could be considered for this purpose. Furthermore, even though our simulation revealed that the model is fairly robust against violations of the piece-wise constant hazard assumption, the potential impact of such misspecifications would be useful to explore in more depth in future research.

Finally we note that the code for the processing of the original data along with the application of the methodology presented in this paper are available in the OSF repository with identifier DOI: 10.17605/OSF.IO/FD9QX (also reachable at https://doi.org/10.17605/OSF.IO/FD9QX). The software developed to run the method discussed in this work will be available in the R package `bremory` which will become available in the coming months.

References

Amati, V., Lomi, A., & Mascia, D. (2019). Some days are better than others: Examining time-specific variation in the structuring of interorganizational relations. *Social Networks*, *57*, 18–33. doi: 10.1016/j.socnet.2018.10.001.

Arena, G., Mulder, J., & Leenders, R. T. A. (2022). A Bayesian semi-parametric approach for modeling memory decay in dynamic social networks. *Sociological Methods & Research*, doi: 10.1177/00491241221113875.

Bianchi, F., & Lomi, A. (2022). From ties to events in the analysis of interorganizational exchange relations. *Organizational Research Methods*, doi: 10.1177/10944281211058469.

Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., & Ward, M. (2015). *ICEWS coded event data*. Harvard Dataverse, doi: 10.7910/DVN/28075,

Brandenberger, L. (2018). Trading favors—Examining the temporal dynamics of reciprocity in congressional collaborations using relational event models. *Social Networks*, *54*, 238–253. doi: 10.1016/j.socnet.2018.02.001.

Brandes, U., Lerner, J., & Snijders, T. A. (2009). Networks evolving step by step: statistical analysis of dyadic event data. Paper presented at Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009 (pp. 200–205). New York City, NY: IEEE. doi: 10.1109/ASONAM.2009.28.

Butts, C. (2008). A relational event model for social action. *Sociological Methodology*, *38*(1), 155–120. doi: 10.2307/20451153.

Clauset, A., Newman, M. E. J., Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111. doi: 10.1103/PhysRevE.70.066111 Retrieved from.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems, 1695, https://igraph.org.

Fritz, C., Thurner, P. W., & Kauermann, G. (2021). Separable and semiparametric network-based counting processes applied to the international combat aircraft trades. *Network Science*, *9*(3), 291–311. doi: 10.1017/nws.2021.9.

Meijerink-Bosman, M., Back, M., Geukes, K., Leenders, R., & Mulder, J. (2022). Discovering trends of social interaction behavior over time: An introduction to relational event modeling. *Behavior Research Methods*, doi: 10.3758/s13428-022-01821-8.

Meijerink-Bosman, M., Leenders, R., & Mulder, J. (2022). Dynamic relational event modeling: Testing, exploring, and applying. *PLoS One*, *17*(8), 1–23. doi: 10.1371/journal.pone.0272309 Retrieved from.

Mulder, J., & Leenders, R. T. A. (2019). Modeling the evolution of interaction behavior in social networks: A dynamic relational event approach for real-time analysis. *SChaos, Solitons and Fractals*, *119*, 73–85. doi: 10.1016/j.chaos.2018.11.027.

Perry, P. O., & Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *75*(5), 821–849. doi: 10.1111/rssb.12013.

Sapiezynski, P., Stopczynski, A., Lassen, D. D., & Lehmann, S. (2019). Interaction data from the Copenhagen networks study. *Scientific Data*, 6(1), 1–10. doi: 10.1038/s41597-019-0325-x.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1), 239–241, http://www.jstor.org/stable/2335876

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. doi: 10.1214/aos/1176344136.

Stadtfeld, C., & Block, P. (2017). Interactions, actors, and time: Dynamic network actor models for relational events. *Sociological Science*, 4(14), 318–352. doi: 10.15195/v4.a14.doi.

Stadtfeld, C., & Geyer-Schulz, A. (2011). Analyzing event stream dynamics in two-mode networks: An exploratory analysis of private communication in a question and answer community. *Social Networks*, 33(4), 258–272. doi: 10.1016/j.socnet.2011.07.004.

Tranmer, M., Marcum, C. S., Morton, F. B., Croft, D. P., & de Kort, S. R. (2015). Using the relational event model (REM) to investigate the temporal dynamics of animal social networks. *Animal Behaviour*, 101, 99–105. doi: 10.1016/j.anbehav.2014.12.005.

A. Appendix

A.1. Weights following a step-wise function

In Perry & Wolfe (2013) endogenous statistics are calculated based on a set of intervals of the transpired time of past events. Consider a vector of $K + 1$ increasing widths $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_K)$ width $\gamma_0 < \ldots < \gamma_K$ and a network dynamic like inertia. After we calculate inertia in the intervals at all the time points, the estimated effects define a step-wise function for the effect of the specific network dynamic at any time point in the event sequence. The step-wise effect function is described by $\boldsymbol{\beta}_{\text{inertia}} = (\beta_{\text{inertia}_1}, \ldots, \beta_{\text{inertia}_K})$

$$\beta_{\text{inertia}_1} \sum_{\substack{e' \in E_{t_{m-1}}: \\ (t_m - t_{e'}) \in (\gamma_0, \gamma 1]}} \mathbb{I}(s(e') = i, r(e') = j) + \ldots + \beta_{\text{inertia}_K} \sum_{\substack{e' \in E_{t_{m-1}}: \\ (t_m - t_{e'}) \in (\gamma_{K-1}, \gamma K]}} \mathbb{I}(s(e') = i, r(e') = j) \tag{A1}$$

where per each interval in $k = 1, \ldots, K$ the effect β_{inertia_k} multiplies by the value of the inertia computed in the kth interval at time t_m. The (A1) can be rewritten in a way similar to (2) where the weight decay this time follows a step-wise function defined on the vector of increasing widths $\boldsymbol{\gamma}$. Indeed, by considering the same vector of widths and associating a vector of K weights $\boldsymbol{w} = (w_1, \ldots, w_K)$ to each interval the step-wise weight decay becomes

$$w(\gamma) = \begin{cases} w_1 & \text{if } \gamma \in (\gamma_0, \gamma_1] \\ \vdots \\ w_K & \text{if } \gamma \in (\gamma_{K-1}, \gamma_K] \\ 0 & \text{otherwise} \end{cases} \tag{A2}$$

Hence, the statistic can be written as a weighted sum as in (2) but in this case following a step-wise decay for the weights,

$$\beta_{\text{inertia}} \text{inertia}(i, j, t_m) = \beta_{\text{inertia}} \sum_{e' \in E_{t_{m-1}}} \mathbb{I}(s(e') = i, r(e') = j) w(t_m - t_{e'}) \tag{A3}$$

where $w(t_m - t_{e'})$ follows the step-wise function in (A2). Considering that weights are the same within each interval, the sum in (A3) can be re-arranged as follows

$$\beta_{\text{inertia}}\text{inertia}(i, j, t_m)$$

$$= \beta_{\text{inertia}} \left[\sum_{\substack{e' \in E_{t_{m-1}}: \\ (t_m - t_{e'}) \in (\gamma_0, \gamma 1]}} \mathbb{I}(s(e')=i, r(e')=j)w_1 + \ldots + \sum_{\substack{e' \in E_{t_{m-1}}: \\ (t_m - t_{e'}) \in (\gamma_{K-1}, \gamma_K]}} \mathbb{I}(s(e')=i, r(e')=j)w_K \right] =$$

$$= \beta_{\text{inertia}}w_1 \sum_{\substack{e' \in E_{t_{m-1}}: \\ (t_m - t_{e'}) \in (\gamma_0, \gamma 1]}} \mathbb{I}(s(e') = i, r(e') = j) + \ldots + \beta_{\text{inertia}}w_K \sum_{\substack{e' \in E_{t_{m-1}}: \\ (t_m - t_{e'}) \in (\gamma_{K-1}, \gamma K]}} \mathbb{I}(s(e') = i, r(e') = j) \quad \text{(A4)}$$

The (A4) is exactly the same formula in (A1) with the only difference that here the step-wise function of weights is explicitly written. Therefore, the equivalence between the two vectors of effects can be written as follows

$$\begin{bmatrix} \beta_{\text{inertia}_1} \\ \vdots \\ \beta_{\text{inertia}_K} \end{bmatrix} = \begin{bmatrix} \beta_{\text{inertia}}w_1 \\ \vdots \\ \beta_{\text{inertia}}w_K \end{bmatrix} \quad \text{(A5)}$$

The same idea of a changing weight of events according to their time recency is also here but it is proposed from a different perspective. In the specific case of a step-wise function, the number of steps and their widths are the parameters describing the function.

A.2. Sms data (sub-networks with 1 cluster and 2 clusters): Trend of MLEs when the weight decay is exponential
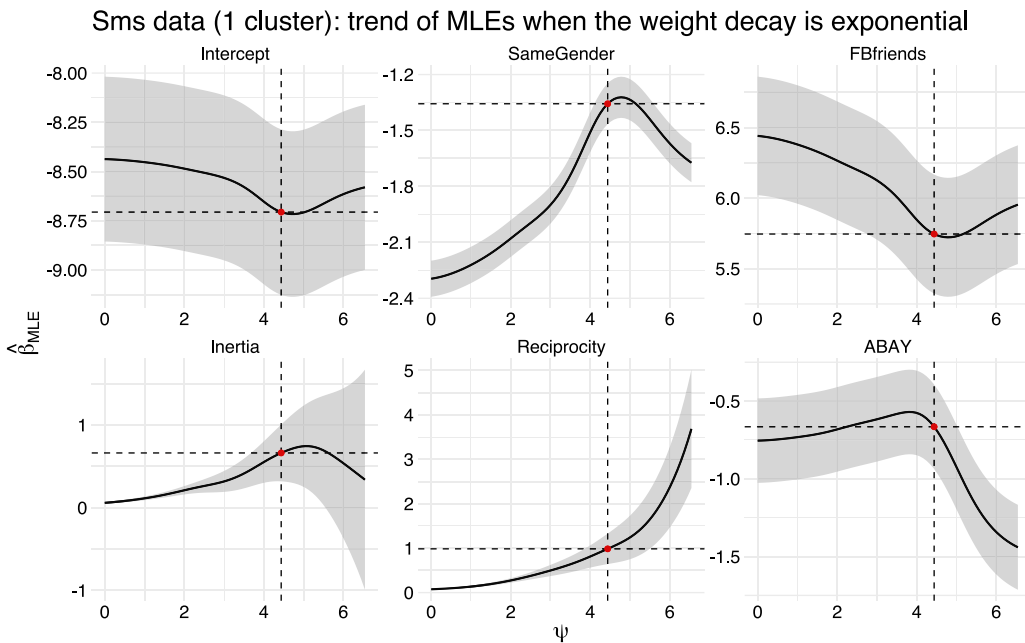
Figure A1. Sms data (1 cluster): trend of the maximum likelihood estimates (MLEs) for the exponential decay over ψ (logarithm of the memory parameter). The dashed black lines in each plot mark the estimate for the log-memory-parameter $\hat{\psi}_{\text{MLE}}$ (vertical lines) and the estimates of the effects β (horizontal lines) at the corresponding $\hat{\psi}_{\text{MLE}}$. The shaded regions are the confidence intervals at 0.95 for the effects β estimated at any value of ψ.
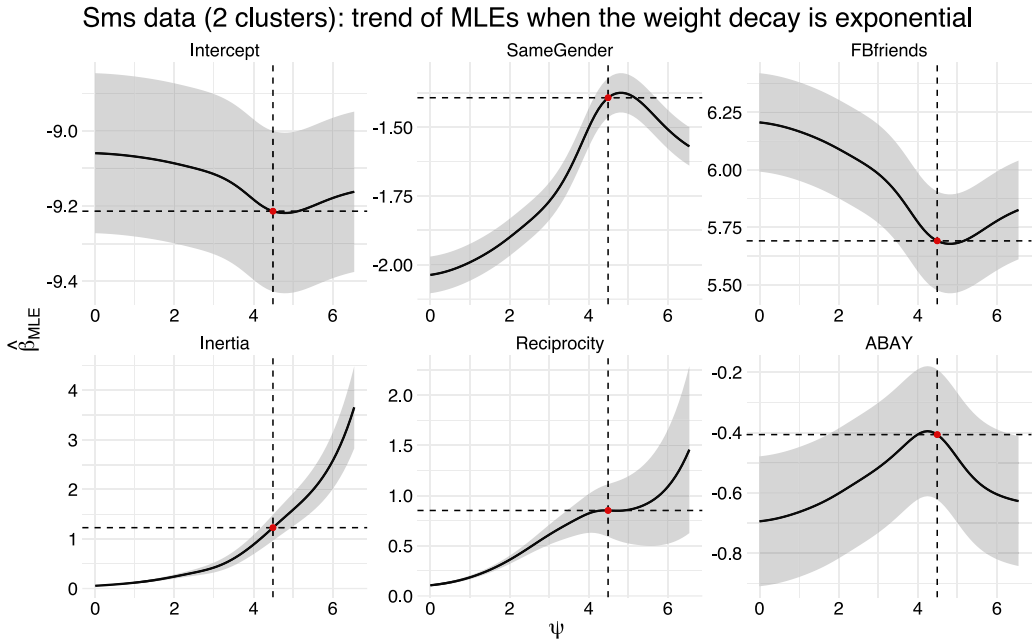
Figure A2. **Sms data (2 clusters)**: trend of the maximum likelihood estimates (MLEs) for the exponential decay over ψ (logarithm of the memory parameter). The dashed black lines in each plot mark the estimate for the log-memory-parameter $\hat{\psi}_{MLE}$ (vertical lines) and the estimates of the effects β (horizontal lines) at the corresponding $\hat{\psi}_{MLE}$. The shaded regions are the confidence intervals at 0.95 for the effects β estimated at any value of ψ.

Cite this article: Arena G., Mulder J. and Leenders R. Th. A. J. (2023). How fast do we forget our past social interactions? Understanding memory retention with parametric decays in relational event models. *Network Science* **11**, 267–294. https://doi.org/10.1017/nws.2023.5