Relational hyperevent models for polyadic interaction networks

Jürgen Lerner[1,2] and Alessandro Lomi[3]

[1]University of Konstanz, Konstanz, Germany
[2]RWTH Aachen, Aachen, Germany
[3]University of Italian Switzerland, Lugano, Switzerland

Address for correspondence: Jürgen Lerner, University of Konstanz, Box 67, 78457 Konstanz, Germany.
Email: juergen.lerner@uni-konstanz.de

Abstract

Polyadic, or 'multicast' social interaction networks arise when one sender addresses multiple receivers simultaneously. Available relational event models are not well suited to the analysis of polyadic interaction networks because they specify event rates for sets of receivers as functions of dyadic covariates associated with the sender and one receiver at a time. Relational hyperevent models (RHEM) address this problem by specifying event rates as functions of hyperedge covariates associated with the sender and the entire set of receivers. We illustrate the empirical value of RHEM in a comparative reanalysis of the canonical *Enron* email data set.

Keywords: hypergraphs, multicast communication, point process models, relational event models, social network analysis, technology-mediated communication

1 Introduction

Data generated by technology-mediated communication typically take the form of sequences of time stamped interaction events involving two or more actors simultaneously. This kind of one-to-many (or 'multicast') interaction is common. For instance, email users can send messages to any number of receivers (Perry & Wolfe, 2013; Zhou et al., 2007). In this paper, we call a social interaction process *polyadic* when it is characterised by events in which one 'sender' may address multiple 'receivers' simultaneously.

Data produced by time-stamped polyadic interaction processes are not unique to social network data produced by technology-mediated interaction (e.g., email messaging) or social media (e.g., twitter). Polyadic interaction data are encountered frequently in empirical research across the social sciences. Examples include scientific papers citing several references (Radicchi et al., 2012), courts judgments citing multiple legal precedents (Fowler et al., 2007), patents approved by the patent office citing multiple prior patents (Kuhn et al., 2020; Verspagen, 2007), and infected persons transmitting a virus to several others simultaneously through group contact (Colizza et al., 2007; Hâncean et al., 2021). Face-to-face conversations where one speaker addresses multiple alters simultaneously also illustrate the empirical extension of polyadic interaction processes (Gibson, 2005). Polyadic interaction may be directed or undirected. Examples of undirected polyadic interaction networks include meetings attended by multiple participants (Freeman, 2003; Lerner et al., 2021), co-authors jointly publishing a paper (Newman, 2004), coordination in task-oriented teams (Ahmadpoor & Jones, 2019; Guimera et al., 2005; Leenders et al., 2016), groups of countries agreeing to sign a multilateral treaty (Hollway & Koskinen, 2016; Simmons & Hopkins, 2005), and class action lawsuits where the plaintiff is a group of people simultaneously bringing a suit to one defendant (Bronsteen & Fiss, 2002).

Despite their obvious heterogeneity, these empirical examples share two defining features. First, interaction among the agents takes the form of time stamped relational events—rather than relational states such as 'being a friend of', 'seeking advice from', or 'regularly communicate with'. Second, the stream of observed events is not generated only by dyadic interaction, but involves multiple network nodes interacting simultaneously at single points in time. In recent years, the availability of social interaction data sharing these features has increased significantly due to the diffusion of computer-mediated communication and collaborative technologies, the availability of large-scale databases, and the diffusion of automated data collection technologies (Lazer et al., 2009). However, the availability of statistical models capable of analysing these data by leveraging their constitutive features has remained limited.

Relational event models (REMs) (Brandes et al., 2009; Butts, 2008; Lerner et al., 2013; Perry & Wolfe, 2013) provide the most promising framework for the analysis of sequences of social interaction events observed in continuous time (Bianchi & Lomi, 2022; Vu et al., 2017). Typical REM for networks of time stamped interaction events specify point processes whose intensities are functions of dyadic covariates. For instance, the intensity of events from sender i to receiver j might depend on the age or gender of i and j, on their age difference, on the frequency of previous events from i to j or from j to i, or on previous events to or from common third actors. Extant research recognises the problem posed by polyadic interaction and multicast communication. The solution that is typically offered involves specifying intensities for events in which a sender i sends a message to a set of receivers $J = \{j_1, \ldots, j_k\}$ (Perry & Wolfe, 2013). However, these intensities are still modelled as functions of dyadic covariates $x(i, j)$, $j \in J$, considering the sender and only one receiver at a time. In this way, available models for relational events assume that the multiple dyads simultaneously produced by a single polyadic interaction are either independent (Perry & Wolfe, 2013), or pertain to fictional 'collective' receivers (Butts, 2008), such as the whole 'team' as a receiver of broadcast messages. While empirically useful, and occasionally justified, neither of these solutions is fully satisfactory.

Recently proposed relational hyperevent models (RHEMs) (Lerner et al., 2021, 2019) generalise REMs by specifying the rate of interaction events from sender i to receiver set J as a function of *hyperedge covariates $x(i, J)$*—being a function of the sender and the entire set of receivers—that cannot necessarily be decomposed into dyadic covariates. As a concrete example, the average number of past interactions that pairs of actors $\{j, j'\} \subseteq J$ have jointly received from i can be expressed as a hyperedge covariate $x(i, J)$—but not as a sum of dyadic covariates $x(i, j)$, $j \in J$. Generalizing edges in graphs that connect exactly two nodes, hyperedges in a hypergraph can connect any number of nodes (Berge, 1989). A relational hyperevent is a time stamped event indexed by a hyperedge (Lerner et al., 2021).

While the conceptual transition from REM to RHEM may be intuitive, its analytical and empirical implications need more complete and rigorous articulation. In this paper, our goal is to improve our current understanding by specifying appropriate models for polyadic social interaction processes and then evaluating empirical differences with respect to comparable dyadic specifications.

A model specified strictly in terms of dyadic covariates might be unable to capture higher-order dependencies typically present in network data produced by polyadic interaction. In turn, this may yield misleading estimates of network effects, or result in lower model fit. In this paper, we take the view that higher-order dependencies—when present—should not merely be considered as an inconvenient feature of the data to be, in the best cases, controlled away. Rather, we argue that such higher-order effects provide a unique opportunity for improving our understanding of the structure and dynamics of social interaction processes, and for developing and testing innovative theories of social interaction.

In this paper, we intend to demonstrate how differences between REM and RHEM may be directly relevant for empirical studies analysing polyadic social interaction processes. We define and discuss practically relevant network effects that can be expressed in RHEMs, but not in dyadic specifications of the event rate afforded by currently available REMs. In the empirical part of the paper, we use the canonical Enron email data (Zhou et al., 2007) to illustrate how to test for polyadic dependencies in empirical data. Moreover, we assess and compare improvements in model fit implied by adding various, dyadic and hyperedge, covariates.

RHEM defined and discussed in this paper directly build on the work by Perry and Wolfe (2013), as detailed below. A number of other alternative approaches have been proposed to adapt REM to events with several receivers (or several senders). In the original statement of the model, Butts (2008) proposes the creation of 'virtual' nodes to represent sets of receivers (or senders). This approach can be appropriate for specifically chosen subsets (such as the whole team as a receiver of broadcast messages), but representing all possible subsets by virtual nodes becomes quickly unfeasible. Kim et al. (2018) propose the hyperedge event model for multicast events. Their model specifies dyadic intensities, associated with one sender and one receiver, as a function of dyadic covariates. These intensities then stochastically determine the sender of the next event and subsequently the receiver set of an interaction by that sender. Their framework does not allow for hyperedge covariates as in RHEM. Mulder and Hoff (2021) define a latent variable model for multicast interaction. However, in their model the mean suitability score of a receiver for messages initiated by a given sender is still a dyadic function, dependent on the sender and one receiver at a time.

Relational hyperevent models for undirected hyperevents (e.g., meeting events) have been proposed in Lerner et al. (2021). Earlier, RHEM have been mentioned by Lerner et al. (2019) who also defined RHEM for directed hyperevents—but did not analyse directed RHEM in any of their empirical examples. Directed RHEM have been applied to modelling contact elicitation networks by infected persons in Hâncean et al. (2021).

We demonstrate how the computation of hyperedge covariates for RHEMs may be performed with the open source software eventnet. The analysis of the Enron email data with RHEM is explained in a step-by-step tutorial linked from https://github.com/juergenlerner/eventnet/wiki/, from the data preprocessing and the computation of covariates over to model estimation in R. Thus, the analysis reported in the empirical part of this paper is fully reproducible. The software may be adopted in—and adapted to future empirical studies involving polyadic social interaction processes.

2 RHEM and hyperedge covariates for multicast social interaction

2.1 Background on REM based on dyadic covariates

We start by recalling the point process models for directed interaction networks, following closely the argument and the notation proposed by Perry and Wolfe (2013). We start our discussion from their model for the more general case of relational events defined over receiver sets of arbitrary size. As observed in their paper, the theoretical results on the consistency of the maximum partial-likelihood inference also apply to the case of polyadic (multicast) events. It may be worth repeating that the difference between RHEM and the model from Perry and Wolfe (2013) does not lie in the basic modelling framework, but in generalising dyadic covariates to hyperedge covariates.

Let $\mathcal{I}$ be a finite set of senders and $\mathcal{J}$ be a finite set of receivers, not necessarily disjoint from $\mathcal{I}$. We denote elements from $\mathcal{A} = \mathcal{I} \cup \mathcal{J}$ as *actors*. For a sender $i \in \mathcal{I}$ and a point in time $t > 0$, let $\mathcal{J}_t(i) \subseteq \mathcal{J}$ denote the set of actors that could potentially receive an interaction from i at t. If $\mathcal{A} = \mathcal{I} = \mathcal{J}$, then it is often the case that $\mathcal{J}_t(i) = \mathcal{I} \setminus \{i\}$, implying that a sender can send interaction events to everyone but herself—that is, loops are excluded. For a sender i and receiver j, let $x_t(i, j)$ be a p-dimensional vector of covariates and β_0 a p-dimensional vector of unknown parameters. For a sender i and a positive integer L (giving the receiver set size), the baseline intensity is denoted by $\bar{\lambda}_t(i, L)$.

Perry and Wolfe (2013) define a model for counting processes on $\mathbb{R}_+ \times \mathcal{I} \times \mathcal{P}(\mathcal{J})$ where the intensity on (i, J), with $i \in \mathcal{I}$ and $J \subseteq \mathcal{J}$, is modelled as

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

Intuitively, $\lambda_t(i, J) \Delta t$ is the expected number of events that i sends to the receiver set J in the time interval $[t, t + \Delta t]$. The intensity $\lambda_t(i, J)$ is assumed to be the baseline intensity $\bar{\lambda}_t(i, |J|)$ multiplied with the relative rate, $\exp\{\beta_0^{\mathrm{T}} \sum_{j \in J} x_t(i, j)\}$. Thus, the parameter vector β_0 controls which

covariates $\{x_t(i, j); \ j \in J\}$ make J a more or less likely receiver set for interaction events sent by i. The covariates in $x_t(i, j)$ can depend on actor characteristics, such as the age of j or the age difference of i and j, but they can also depend on the history of the process. For instance, the covariates $x_t(i, j)$ can include a count of the number of past events that i has sent to j, before t.

Importantly, the model summarised in (1) assumes that the covariate vectors $\{x_t(i, j), \ j \in J\}$ increase or decrease the rate of events from sender i to the receiver set J *independently of each other*. Equivalent to (1), we can model the intensity $\lambda_t(i, J)$ as the baseline intensity $\bar{\lambda}_t(i, |J|)$ multiplied with the relative rate $\exp\{\beta_0^{\mathrm{T}} x_t(i, J)\}$, where the covariates $x_t(i, J)$ of sender i and *receiver set J* are assumed to have the specific form

$$x_t(i, J) = \sum_{j \in J} x_t(i, j). \tag{2}$$

The equation above implies that in the original model the suitability of j as a receiver of interaction sent by i is assumed to be independent of the other receivers $j' \in J \setminus \{j\}$ of the same interaction—an assumption that is challenged in our paper. Generalizing this model, RHEM allow covariates $x_t(i, J)$ that do not necessarily decompose into a sum of dyadic covariates of the form $\sum_{j \in J} x_t(i, j)$; compare Section 2.2.

Let $(t_1, i_1, J_1), \ldots, (t_n, i_n, J_n)$ be the observed sequence of polyadic interactions where (t, i, J) indicates that at time t sender i interacts with receiver set J. The model from Perry and Wolfe (2013), given in (1), leads to the log partial likelihood at t evaluated at $\beta \in \mathbb{R}^p$:

$$\log L_t(\beta) = \sum_{t_m \leq t} \left(\beta^{\mathrm{T}} \sum_{j \in J_m} x_{t_m}(i_m, j) - \log \left[\sum_{J \in \binom{\mathcal{J}_{t_m}(im)}{|J_m|}} \exp\left\{ \beta^{\mathrm{T}} \sum_{j \in J} x_{t_m}(i_m, j) \right\} \right] \right), \tag{3}$$

where for a set X and an integer L, we write $\binom{X}{L} = \{X' \subseteq X; \ |X'| = L\}$ for the set of all subsets of size L. Perry and Wolfe (2013) prove that, under generally accepted assumptions, parameters maximising (3), i.e., the maximum partial-likelihood estimates (MPLE), are a consistent estimator of β_0.

2.2 Relational hyperevent model

Building on the model proposed by Perry and Wolfe (2013), the definition of RHEM for directed polyadic interaction can be obtained by substituting $x_t(i, J)$ for $\sum_{j \in J} x_t(i, j)$ in (1) and (3). More precisely, given the notation from Section 2.1, RHEM define a model for counting processes on $\mathbb{R}_+ \times \mathcal{I} \times \mathcal{P}(\mathcal{J})$ where the intensity on (i, J), with $i \in \mathcal{I}$ and $J \subseteq \mathcal{J}$, is modelled as

$$\lambda_t(i, J) = \bar{\lambda}_t(i, |J|) \exp\{\beta_0^{\mathrm{T}} x_t(i, J)\} \mathbf{1}\{J \subseteq \mathcal{J}_t(i)\}. \tag{4}$$

The covariates $x_t(i, J)$ do not necessarily decompose into a sum of dyadic covariates and may depend on exogenous actor-level characteristics or they can depend on the history of the process. Examples for the former include the average age of receivers in J, the average absolute age difference between the sender i and the receivers in J, or the average absolute age difference between pairs of receivers in J. Examples for the latter (i.e., history-dependent) covariates include the number of past interactions that i has sent to receiver set J:

$$\#\{\text{interaction}(t', i, J) \text{with } t' < t\},$$

the average number of past interactions that actors $j \in J$ have received from i:

$$\frac{\sum_{j \in J} \#\{\text{interaction}(t', t, J') \text{ with } t' < t \wedge j \in J'\}}{|J|},$$

or the average number of past interactions that pairs of actors $\{j, j'\} \subseteq J$ have jointly received from i:

$$\frac{\sum_{\{j,j'\}\in\binom{J}{2}} \#\{\text{interaction}(t', i, J') \text{ with } t' < t \wedge j, j' \in J'\}}{\binom{|J|}{2}}.$$

Note that the first and the third of the above examples for history-dependent covariates do not admit a decomposition of the form (2). The second of these examples for history dependent covariates admits a decomposition of the form (2) *if* we are willing to additionally allow multiplication of covariates with a constant that depends only on the receiver set size $|J|$.

Let $(t_1, i_1, J_1), \ldots, (t_n, i_n, J_n)$ be the observed sequence of polyadic interactions. The model given in (4) leads to the following log partial likelihood at t evaluated at $\beta \in \mathbb{R}^p$

$$\log L_t(\beta) = \sum_{t_m \leq t} \left(\beta^{\mathrm{T}} x_{t_m}(i_m, J_m) - \log \left[\sum_{J \in \binom{\mathcal{J}_{t_m}(i_m)}{|J_m|}} \exp\{\beta^{\mathrm{T}} x_{t_m}(i_m, J)\} \right] \right). \tag{5}$$

The arguments from Perry and Wolfe (2013) on the consistency of MPLE of their model also apply to parameters maximising (5). Note that the only difference between (3) and (5) is that we allow more general covariates.

2.3 Case control sampling

The log partial likelihood (5) can no longer be approximated by replacing the sum over all J in the risk set $\binom{\mathcal{J}_{t_m}(i_m)}{|J_m|}$ with a sum over all receivers in $\mathcal{J}_{t_m}(i_m)$, as it was the case for models specified with dyadic covariates in Perry and Wolfe (2013). Having to sum over all subsets would lead to excessive, or even intractable, computational runtime for all but the smallest $\mathcal{J}_{t_m}(i_m)$ and J_m.

Previous work on REM and RHEM for the analysis of large networks (Lerner & Lomi, 2020; Lerner et al., 2021; Vu et al., 2015) suggested to reduce excessive computational runtime by replacing the risk set with a sampled risk set obtained via case control sampling (Borgan et al., 1995; Keogh & Cox, 2014). The derivation of the partial likelihood based on sampled risk sets (7), starting from the partial likelihood using the entire risk set (5), can actually be obtained by adapting the general approach from Borgan et al. (1995) to our special case—and in the remainder of this subsection we will only briefly indicate the relevant steps. We recall that the main contribution of this paper are the hyperedge covariates and not the particular method to estimate model parameters. We point out that when applying RHEM to data with a very small number of nodes, case control sampling is not needed (Bianchi & Lomi, 2022; Lerner & Lomi, 2022) and parameters may be estimated by maximising (5). In application settings where the number of nodes is larger, we propose to estimate RHEM parameters by maximising (7).

For a given positive integer k (the number of non-events per event), let $\tilde{\mathcal{R}}_t(\mathcal{J}_t(i), J, k) \subseteq \binom{\mathcal{J}_t(i)}{|J|}$ be a set of subsets of $\mathcal{J}_t(i)$ that is sampled uniformly at random from

$$\left\{ \mathcal{R} \subseteq \binom{\mathcal{J}_t(i)}{|J|}; \ J \in \mathcal{R} \wedge |\mathcal{R}| = k+1 \right\}. \tag{6}$$

In other words, $\tilde{\mathcal{R}}_t(\mathcal{J}_t(i), J, k)$ contains the observed receiver set J (the 'case') plus k alternative receiver sets (the 'controls'), sampled without replacement uniformly and independently at random from $\{J' \subseteq \mathcal{J}_t(i); \ |J'| = |J| \wedge J' \neq J\}$.

Adapting the general approach of Borgan et al. (1995) to our special case, we obtain a counting process on $\mathbb{R}_+ \times \mathcal{I} \times \mathcal{P}(\mathcal{J}) \times \mathcal{P}[\mathcal{P}(\mathcal{J})]$, where the tuple $(t, i, J, \mathcal{R})$ indicates that at time t, sender i interacts with receiver set J and $\mathcal{R}$ is the set that has been sampled from (6). If $\pi_t(\mathcal{R}|i, J)$ denotes the

conditional probability that $\mathcal{R}$ is the sampled risk set at t, given that an event on (i, J) occurs at t, then this counting process has intensity

$$\lambda_t(i, J, \mathcal{R}) = \bar{\lambda}_t(i, |J|) \exp\{\beta_0^{\mathrm{T}} x_t(i, J)\} \mathbf{1}\{J \subseteq \mathcal{J}_t(i)\} \pi_t(\mathcal{R}|i, J).$$

Uniform sampling from (6) is a special case of nested case-control sampling (Borgan et al., 1995), in particular it is $\pi_t(\mathcal{R}|i, J) = \pi_t(\mathcal{R}|i, J')$ for all $J, J' \in \mathcal{R}$, and by Borgan et al. (1995) we can estimate parameters $\hat{\beta}$ by maximising the log partial-likelihood function

$$\log \tilde{L}_t(\beta) = \sum_{t_m \leq t} \left(\beta^{\mathrm{T}} x_{t_m}(i_m, J_m) - \log \left[\sum_{J \in \tilde{\mathcal{R}}_{t_m}(\mathcal{J}_{t_m}(i_m), J_m, k)} \exp\{\beta^{\mathrm{T}} x_{t_m}(i_m, J)\} \right] \right). \tag{7}$$

Borgan et al. (1995) prove conditions under which the MPLE is a consistent estimator. Estimating parameters based on the partial likelihood (7) uses only information about the selection of receiver sets J_m out of all *sampled* sets of receivers $J \in \tilde{\mathcal{R}}_{t_m}(\mathcal{J}_{t_m}(i_m), J_m, k)$. On the other hand, it does not consider information contained in the event time t_m, sender i_m, and receiver set size $|J_m|$, which are absorbed in the baseline rate $\bar{\lambda}_t(i, |J|)$. Therefore, including covariates that are only functions of event times, senders, or size of the receiver sets would lead to non-identifiable parameters.

2.4 Hyperedge covariates
Since hyperedge covariates $x_t(i, J)$ are defined on the entire set of receivers, a large number of structurally different covariates is possible. While our list is certainly far from exhaustive, we define in the following a collection of hyperedge covariates that are practically relevant in the empirical analysis of multicast interaction networks. We define two types of covariates: 'attribute effects', dependent on actor-level attributes and 'network effects', dependent on the history of the process, that is, on sequences of previously observed events.

2.4.1 Covariates dependent on actor-level attributes
Suppose that available data include information on one or several numeric actor-level attributes $z : \mathcal{A} \to \mathbb{R}$. Then, in general, RHEM covariates $x_t(i, J)$ based on such actor ('node-specific') characteristics may be obtained either by functions of the values in the receiver set such as mean$\{z(j); \ j \in J\}$, or by functions of the receivers' values in relation to the sender's value $z(i)$, such as mean$\{|z(j) - z(i)|; \ j \in J\}$. Covariates that are only functions of the sender i would lead to a non-identifiable parameter, since their effect would be absorbed by the baseline rate $\bar{\lambda}_t(i, |J|)$.

Concrete examples of covariates are discussed below, where we also discuss the possibility of defining covariates based on categorical attributes. In this article, we consider only time-invariant attributes. However, the covariates discussed below extend directly to time-varying actor-level attributes by considering values at the given time t.

Receiver set average of attribute z captures a first-order effect of an attribute on receiving interactions and is defined to be

$$\text{rec\_avg\_}z_t(i, J) = \sum_{j \in J} z(j)/|J|.$$

For example, in the empirical analysis we apply this definition to four binary attributes *female*, *senior*, *trade*, and *legal*. The receiver set average, thus, gives the proportion of receivers that are females, seniors, members of the 'trade' department, or members of the 'legal' department, respectively. The definition also applies to non-binary numeric attributes.

Sender–receiver heterophily. We assess heterophily between senders and receivers by averaging the absolute difference between the attribute value of the sender i and the values of the receivers $j \in J$:

$$\text{send\_rec\_diff\_}z_t(i, J) = \sum_{j \in J} |z(i) - z(j)|/|J|.$$

For example, for the attribute *female* we obtain the fraction of receivers having the opposite gender than the sender. If senders have a tendency to interact with similar receivers, then a higher value of sender–receiver homophily would make an interaction on the hyperedge (i, J) less likely. Thus, in the presence of homophily we would expect a negative parameter—a positive parameter would point to heterophily.

In the case of a categorical attribute z, such as the attribute 'department', we assess sender–receiver heterophily via the fraction of receivers that have a different value than the sender:

$$\text{send\_rec\_diff}\_z_t(i, J) = \sum_{j \in J} \mathbf{1}\{z(i) \neq z(j)\}/|J|.$$

Receiver set heterophily. We assess heterophily in the receiver set by averaging the absolute difference of the attribute value between receiver pairs:

$$\text{rec\_set\_diff}\_z_t(i, J) = \sum_{\{j,j'\} \in \binom{J}{2}} \frac{|z(j') - z(j)|}{\binom{|J|}{2}}.$$

If senders have a tendency to interact with homogeneous receiver sets (irrespective of their own attribute value), then a higher value of the covariate capturing receiver set homophily would make an interaction less likely. Thus, we would expect a negative parameter in the presence of homophily; a positive parameter would point to heterophily.

For a categorical attribute z, we assess receiver set heterophily via the fraction of receiver pairs that have a different attribute value:

$$\text{rec\_set\_diff}\_z_t(i, J) = \sum_{\{j,j'\} \in \binom{J}{2}} \frac{\mathbf{1}\{z(j') \neq z(j)\}}{\binom{|J|}{2}}.$$

Note that receiver set heterophily is structurally different from sender–receiver heterophily (Snijders & Lomi, 2019). Theoretically it might be the case that senders have neither a preference, nor a reluctance to send messages to, say, receivers in their own department. Yet, in the same data it is possible that typical receiver sets are mostly composed of members from one department. In such a hypothetical scenario, we would not find evidence for sender–receiver homophily but we would find evidence for receiver set homophily.

2.4.2 Network effects

In addition to covariates based on actor-level attributes, we define several covariates expressing dependence of the intensity $\lambda_t(i, J)$ on the history of the process. Given an observed sequence of events generated by polyadic interaction $E = \{(t_1, i_1, J_1), \ldots, (t_n, i_n, J_n)\}$, the value of these covariates at time t is computed as a function of earlier events $E_{<t} = \{(t_m, i_m, J_m) \in E; \ t_m < t\}$.

Similar to previous work on REM and RHEM (Amati et al., 2019; Brandes et al., 2009; Lerner et al., 2013, 2021), we let the effect of past events decay over time. More precisely, for a given half-life period $T_{1/2} > 0$, the weight of a past event (t_m, i_m, J_m) at current time $t > t_m$ is defined to be $w(t - t_m) = \exp\left(-(t - t_m)(\log 2/T_{1/2})\right)$. While alternatives exist to exponential decay of past events, Schecter and Quintane (2021) found it to be generally adequate. The objective of our paper is to clarify the benefit of hyperedge covariates in comparison to dyadic covariates. We consider issues related to the effect of elapsed time of past events as orthogonal to the objective of the current paper. Hyperedge covariates and dyadic covariates considered in the empirical part of this paper are defined with the same decay mechanism and the same half-life period.

Exact repetition. An empirically plausible effect would capture relational 'inertia', or the tendency of actors to continue to do what they did in the past. In fact, in many cases, it may be important to

verify the presence of more complex mechanisms over and above the simple repetition of past be-haviour. In polyadic interaction networks, this inertial behavioural tendency would lead to future events that repeat the sender and the entire receiver set of past events. Such an effect can be cap-tured by the covariate:

$$\text{exact\_repetition}_t(i, J) = \sum_{(t_m, i_m, J_m) \in E_{<t}} w(t - t_m) \cdot \mathbf{1}(i_m = i \wedge J_m = J).$$

Thus, we sum the current weight over all previous events that have the same sender and exactly the same receiver set. Exact repetition captures effects in which the same sender repeatedly addresses the same receivers, for instance, email communication with fixed receiver lists ('mailing lists').
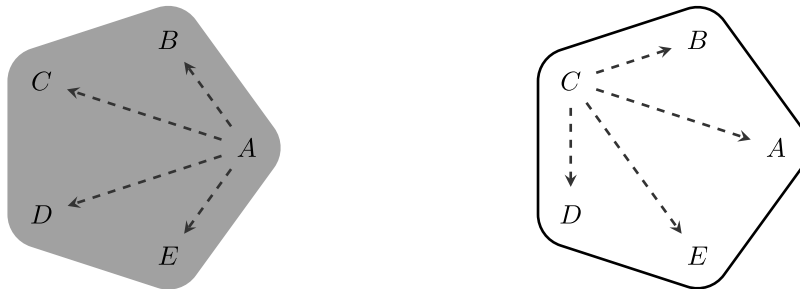
Unordered repetition. Besides exact repetition, there can be situations of interaction within a stable group of actors with turn-taking among the senders (Gibson, 2005). Such an effect can be captured by the covariate

$$\text{unordered\_repetition}_t(i, J) = \sum_{(t_m, i_m, J_m) \in E_{<t}} w(t - t_m) \cdot \mathbf{1}(\{i_m\} \cup J_m = \{i\} \cup J).$$

In contrast to (ordered) exact repetition, the unordered repetition effect allows that different actors take on the role of sender—as long as the union of the sender with the receiver set remains constant. A typical example of this behaviour is email communication using a 'reply-to-all' functionality: a re-ceiver of a previous message sends a message to the previous sender and to all other receivers. See Figure 1 for an illustration of unordered repetition. In contrast to unordered repetition, exact repe-tition would require that the future event in Figure 1 is exactly on the hyperedge $(A, \{B, C, D, E\})$, that is, the same sender A sends another interaction to the same receiver set $\{B, C, D, E\}$.
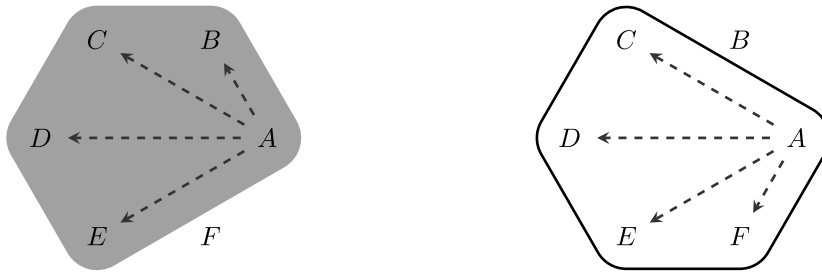
Partial receiver set repetition. The two (exact and unordered) repetition covariates defined above still give an incomplete picture of stability in multicast interaction events, since they require that sets of actors involved in past and current events have to be identical. A possible event that mostly, but not exactly, repeats the receiver set of a past event—for instance, if some new receivers are added and/or if some previous receivers are removed—is treated identically to a possible event with a completely disjoint receiver set. To quantify partial repetition, we define a parametric fam-ily of covariates capturing to what extent subsets of a possible receiver set have jointly received past events.

To shorten notation, we define the *hyperedge indegree* of a set of receivers $J' \subseteq \mathcal{J}$ by considering past events that have been jointly received by all members of J'—possibly together with varying

Figure 1. Stylised example illustrating unordered repetition. *Left*: history of a past event $e_1 = (t_1, A, \{B, C, D, E\})$ displayed as a grey-shaded area; dashed lines connect the sender to the receivers. *Right*: a candidate hyperedge $h = (C, \{A, B, D, E\})$ for a future hyperevent. The past event e_1 increases the value of unordered repetition on h at time $t > t_1$. In communication networks, unordered repetition could point to turn-taking among a stable set of conversation participants.

Figure 2. Stylised example illustrating (sender-specific) partial receiver set repetition. *Left*: history of a past event $e_1 = (t_1, A, \{B, C, D, E\})$. *Right*: a candidate hyperedge $h = (A, \{C, D, E, F\})$ for a future hyperevent. Among the four receivers in h, three have individually received the past event e_1. Among the six pairs of receivers in h, three have jointly received the past event e_1. Among the four triples of receivers in h, one has jointly received the past event e_1. The past event e_1 increases the value of sender-specific partial receiver set repetition of order $p = 1, 2, 3$ on h at $t > t_1$. If the sender of h was another actor G, instead of A, then the past event would still increase the value of partial receiver set repetition, but it would not increase the value of the sender-specific variant.

other receivers outside of J':

$$\text{hy\_deg}_t^{(in)}(J') = \sum_{(t_m, i_m, J_m) \in E_{<t}} w(t - t_m) \cdot \mathbf{1}(J' \subseteq J_m).$$

Partial receiver set repetition (or subset repetition among the receivers) is parametrised by a positive integer p, giving the cardinality of the subsets that repeatedly receive joint messages:

$$\text{rec\_sub\_rep}_t^{(p)}(i, J) = \sum_{J' \in \binom{J}{p}} \frac{\text{hy\_deg}_t^{(in)}(J')}{\binom{|J|}{p}}.$$

For $p = 1$ we obtain the average indegree of individual receivers $j \in J$ by considering past interactions received by j, downweighted by the elapsed time. For $p = 2$ we consider past interactions that have been jointly received by pairs $\{j, j'\} \subseteq J$, and so on.

Partial receiver set repetition—and the related sender-specific partial receiver set repetition, defined below—are illustrated in Figure 2. In the notation of that figure, if we ignore the decay over time and if e_1 is the entire history, we get for a point in time $t > t_1$ the following values:

$$\text{rec\_sub\_rep}_t^{(1)}(h) = 3/4$$

$$\text{rec\_sub\_rep}_t^{(2)}(h) = 3/6$$

$$\text{rec\_sub\_rep}_t^{(3)}(h) = 1/4.$$

Partial receiver set repetition of order $p > 3$ is zero in this example.

Sender-specific partial receiver set repetition. Partial receiver set repetition defined above does not consider whether past interactions jointly received by $J' \subseteq J$ originated from the same sender i. Thus, these covariates capture partial repetition of receivers by any sender. To consider only partial receiver set repetition by the same sender, we first define the *sender-specific hyperedge degree* by:

$$\text{hy\_deg}_t(i, J') = \sum_{(t_m, i_m, J_m) \in E_{<t}} w(t - t_m) \cdot \mathbf{1}(i = i_m \wedge J' \subseteq J_m).$$

Sender-specific partial receiver set repetition (or sender-specific subset repetition in the receiver set) is parametrised by a positive integer p, giving the cardinality of the subset that repeatedly

receives joint messages from the given sender i.

$$\text{send\_rec\_sub\_rep}_t^{(p)}(i, J) = \sum_{J' \in \binom{J}{p}} \frac{\text{hy\_deg}_t(i, J')}{\binom{|J|}{p}}.$$

For $p = 1$, we obtain the average weight of past interactions that individual actors $j \in J$ received from the given sender i, where the average is taken over all those receivers. For $p = 2$, we consider past interactions that have been sent by i and that have been jointly received by pairs $\{j, j'\} \subseteq J$, and so on.

In the example illustrated in Figure 2, we obtain the same values for sender-specific partial receiver set repetition as for partial receiver set repetition, since the candidate hyperedge for a future event $h = (A, \{C, D, E, F\})$ has the same sender as the past event e_1. However, another candidate hyperedge $h' = (G, \{C, D, E, F\})$ would have the same value as h in the partial receiver set repetition covariates but would get the value zero in the sender-specific variants.

(Sender-specific) partial receiver set repetition of order $p \geq 2$ may induce a clustering in the set of actors, revealing subsets of actors that are likely to jointly receive the same interactions. Partial receiver set repetition implies a 'global' clustering applying to all interactions, irrespective of their sender. The sender-specific variants allow for different clusterings of receivers that vary with the sender.
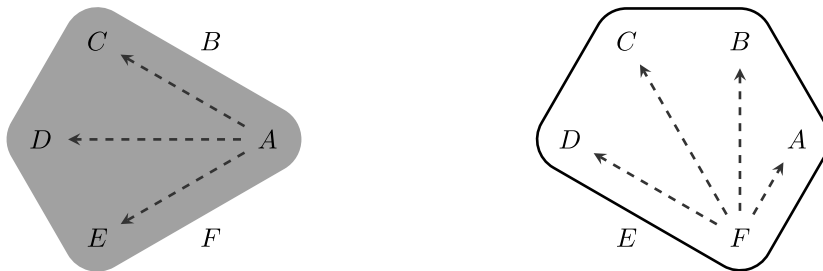
Past interaction among receivers. Yet another network effect in polyadic social interaction network arises if actors send interactions to a sender of a past interaction together with a subset of the receivers of that past interaction. This is, for instance, a frequent pattern in scientific citation networks where a paper P cites another paper P' and some of the references of P'. This pattern can be captured by the following family of covariates, parametrised by a positive integer p, giving the number of the repeated receivers of the previous event:

$$\text{interact\_rec}_t^{(p)}(i, J) = \sum_{j \in J, J' \in \binom{J \setminus \{j\}}{p}} \frac{\text{hy\_deg}_t(j, J')}{|J| \cdot \binom{|J|-1}{p}}.$$

Interaction among receivers is illustrated in Figure 3. In the notation of that figure, if we ignore the decay over time and if e_1 is the entire history, we get for a point in time $t > t_1$ the following values:

$$\text{interact\_rec}_t^{(1)}(h) = 2/(4 \cdot 3)$$

$$\text{interact\_rec}_t^{(2)}(h) = 1/(4 \cdot 3).$$

Figure 3. Stylised example illustrating the covariate 'interaction among receivers'. *Left*: history of a past event $e_1 = (t_1, A, \{C, D, E\})$. *Right*: a candidate hyperedge $h = (F, \{A, B, C, D\})$ for a future hyperevent. The sender F of h sends an interaction to the sender A of the past event e_1 and to two of its receivers (C and D). The past event e_1 increases the value of the covariate interaction among receivers on h at $t > t_1$ for $p = 1, 2$. For $p > 2$ that covariate is zero since there are no three previous receivers that receive an interaction together with the previous sender.

Interaction among receivers of order $p > 2$ is zero in this example.

Reciprocation and out-in popularity. Actors often have the tendency to reply to interactions they received in the past. Reciprocation is captured by considering interactions that the sender i of the current interaction has received from actors $j \in J$:

$$\text{recip}_t(i, J) = \sum_{j \in J} \text{hy\_deg}_t(j, \{i\})/|J|.$$

The effect *out-in popularity*, arises when senders of past interactions receive future interactions from any actor in the network—not necessarily from the receivers of the past interactions. We define the *outdegree* of an actor $i' \in \mathcal{A}$ by considering past events that have been send by i'. (Note that in contrast to the hyperedge indegree, the outdegree cannot be defined for a set of more than one actor, since every interaction has only one sender.)

$$\deg_t^{(out)}(i') = \sum_{(t_m, i_m, J_m) \in E_{<t}} w(t - t_m) \cdot \mathbf{1}(i' = i_m).$$

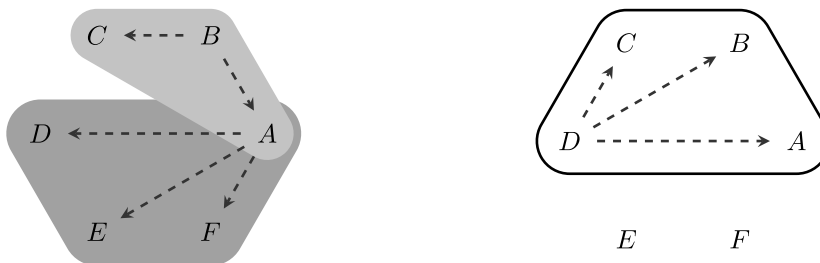Out-in popularity is defined by

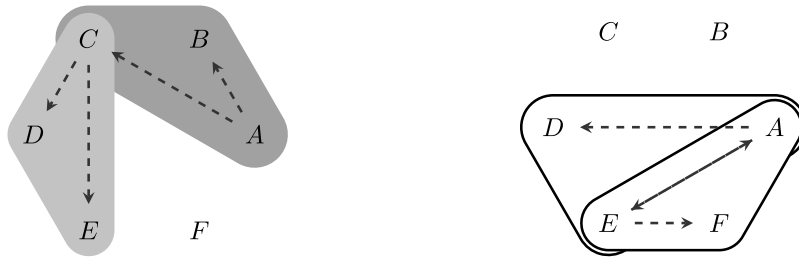$$\text{out\_in\_pop}_t(i, J) = \sum_{j \in J} \deg_t^{(out)}(j)/|J|.$$

Reciprocation and out-in popularity are illustrated in Figure 4. In the notation of that figure, if we ignore the decay over time and if e_1, e_2 is the entire history, we get for a point in time $t > t_1, t_2$ the following values:

$$\text{recip}_t(h) = 1/3$$
$$\text{out\_in\_pop}_t(h) = 2/3.$$

Triadic effects. Interactions might further depend on past interactions that the sender and the receivers had with common third actors. By varying the directions of the past interactions to or from the sender and the receivers we obtain four variants of triadic closure, denoted by transitive closure, cyclic closure, incoming balance, and outgoing balance. In the summations below, a iterates

Figure 4. Stylised example illustrating reciprocation and out-in popularity. *Left*: history of two past events $e_1 = (t_1, A, \{D, E, F\})$ and $e_2 = (t_2, B, \{A, C\})$. *Right*: a candidate hyperedge $h = (D, \{A, B, C\})$ for a future hyperevent. The past event e_1 increases the value of the reciprocation covariate on h at $t > t_1$, since e_1 has been sent by A, a receiver of h, among others to D, the sender of h. The past event e_2 does not increase reciprocation on h—but it increases out-in popularity on h, since e_2 has been send by B, a receiver of h.

Figure 5. Stylised example illustrating transitive closure and cyclic closure. *Left*: history of two past events $e_1 = (t_1, A, \{B, C\})$ and $e_2 = (t_2, C, \{D, E\})$. *Right*: two candidate hyperedges $h = (A, \{D, E\})$ and $h' = (E, \{A, F\})$ for future hyperevents. The past events e_1, e_2 increase the value of transitive closure on h at $t > t_1$, t_2, since h transitively closes two paths: from A over C to D and from A over C to E. The past events e_1, e_2 increase the value of cyclic closure on h' at $t > t_1$, t_2, since h' closes a cycle from A to C to E to A.

Figure 6. Stylised example illustrating incoming balance. *Left*: history of two past events $e_1 = (t_1, C, \{A, B\})$ and $e_2 = (t_2, C, \{D, E\})$. *Right*: a candidate hyperedge $h = (A, \{D, E\})$ for a future hyperevent. The past events e_1, e_2 increase the value of incoming balance on h at $t > t_1$, t_2, since the sender of h, A received a past event from C and two receivers of h, D, and E also received a past event from the same sender C.

over all actors $\mathcal{A} \setminus \{i, j\}$.

$$\text{transitive\_closure}_t(i, J) = \sum_{j \in J, a \neq i,j} \frac{\min\{\text{hy\_deg}_t(i, \{a\}), \text{hy\_deg}_t(a, \{j\})\}}{|J|}$$

$$\text{cyclic\_closure}_t(i, J) = \sum_{j \in J, a \neq i,j} \frac{\min\{\text{hy\_deg}_t(a, \{i\}), \text{hy\_deg}_t(j, \{a\})\}}{|J|}$$

$$\text{in\_balance}_t(i, J) = \sum_{j \in J, a \neq i,j} \frac{\min\{\text{hy\_deg}_t(a, \{i\}), \text{hy\_deg}_t(a, \{j\})\}}{|J|}$$

$$\text{out\_balance}_t(i, J) = \sum_{j \in J, a \neq i,j} \frac{\min\{\text{hy\_deg}_t(i, \{a\}), \text{hy\_deg}_t(j, \{a\})\}}{|J|}.$$

Transitive closure and cyclic closure are illustrated in Figure 5. In the notation of that figure, if we ignore the decay over time and if e_1, e_2 is the entire history, for a point in time $t > t_1$, t_2 we obtain the following values:

$$\text{transitive\_closure}_t(h) = 2/2$$
$$\text{cyclic\_closure}_t(h) = 1/2.$$

Incoming balance is illustrated in Figure 6. In the notation of that figure, if we ignore the decay over time and if e_1, e_2 is the entire history, for a point in time $t > t_1$, t_2 we obtain the following

Figure 7. Stylised example illustrating the outgoing balance covariate. *Left*: history of two past events $e_1 = (t_1, A, \{B, C\})$ and $e_2 = (t_2, E, \{D, C\})$. *Right*: a candidate hyperedge $h = (A, \{D, E\})$ for a future hyperevent. The past events e_1, e_2 increase the value of outgoing balance at $t > t_1, t_2$, since the sender of h, A has sent a past event to C and one receiver of h, E has also sent a past event to C.

value:

$$in\_balance_t(h) = 2/2.$$

Outgoing balance is illustrated in Figure 7. In the notation of that figure, if we ignore the decay over time and if e_1, e_2 is the entire history, for a point in time $t > t_1, t_2$ we obtain the following value:

$$out\_balance_t(h) = 1/2.$$

2.4.3 Dyadic covariates vs. hyperedge covariates

As mentioned, we denote a hyperedge covariate $x_t(i, J)$ as a *dyadic covariate* if it admits a decomposition of the form $x_t(i, J) = C(|J|) \cdot \sum_{j \in J} x_t(i, j)$, compare (2), where $C(|J|)$ is a multiplicative constant depending only on the size of the receiver set J. Here we discuss, which of the hyperedge covariates defined above are dyadic covariates.

Among the covariates based on actor-level attributes, the receiver set average and sender-receiver heterophily are dyadic covariates. In contrast, receiver set heterophily is not a dyadic covariate since the terms in its definition jointly consider pairs of receivers.

Neither exact repetition nor unordered repetition are dyadic covariates since the terms in their definition jointly consider a potentially unbounded number of receivers. Partial receiver set repetition of order p is a dyadic covariate for $p = 1$ but not for any $p > 1$. The same holds for sender-specific partial receiver set repetition of order p. Interaction among receivers of order p is not a dyadic covariate for any $p \geq 1$. Note that the definition of this covariate considers subsets of the receiver set J containing p previous receivers and one previous sender. Reciprocation, out-in popularity, and all four triadic closure covariates are dyadic. For the latter, note that in the definition of these covariates we sum over single receivers $j \in J$, rather than over larger subsets of J.

3 Empirical analysis

3.1 Empirical setting and data

We demonstrate the empirical value of the model in an analysis of the Enron email data set—a collection of corporate emails exchanged by employees of Enron Corporation that was made public after the company filed for bankruptcy in December 2001. For additional information on the history of the data, we refer interested readers to Zhou et al. (2007). Prior empirical analyses of the email corpus may be found in Diesner et al. (2005) and in Keila and Skillicorn (2005). To facilitate comparability, we analyse the subset of the data cleaned and processed by Zhou et al. (2007) that has also been used in the empirical example reported in Perry and Wolfe (2013) and that is available at https://github.com/patperry/interaction-proc/tree/master/data/enron. Analysis with the eventnet software requires conversion of these data into a different format. The conversion steps are explained in https://github.com/juergenlerner/eventnet/tree/master/data/enron where

Table 1. Number of emails (bottom row) with given number of receivers (top row)

$|J|$	1	2	3	4	5	6	7	8	9	10	>10
Frequency:	14,985	2,962	1,435	873	711	180	176	61	24	29	199

the converted data are also available for download. As mentioned before, the entire analysis reported in this paper is detailed in a step-by-step tutorial; see Section 4.6.

In the empirical section that follows we refer to this subset as the 'Enron data' or just the 'data'. The data comprises 21,635 emails (treated as hyperevents) among 156 Enron employees. Additionally, we use the actor attributes, *gender* (female = 1, male = 0), *seniority* (senior = 1, junior = 0), and the categorical attribute *department* (taking the values 'Legal', 'Trading', and 'Other'). None of the actor attributes changes over time and there are no missing values.

The observed emails have exactly one sender and between one and 57 receivers. About 30.7% of the emails have more than one receiver and the average number of receivers is 1.77. The receiver set size distribution is further detailed in Table 1. We analyse all of these emails, that is, we do not discard emails with many receivers.

The time values of the emails in the data correspond to seconds—but time resolution is by the minute since all given time values are divisible by 60. There are 20,994 emails (about 97%) that have a unique time stamp, there are 305 time points at which exactly two different emails have been sent, nine time points are shared by three emails, and there is one time point at which four different emails have been sent. We order simultaneous emails arbitrarily. Given the high level of time resolution and the relatively low number of simultaneous hyperevents (i.e., tied event times), we believe that this decision is unlikely to affect our results in any meaningful way. However, we note that established ways to deal with tied event times exist (Breslow, 1974; Efron, 1977; Hertz-Picciotto & Rockhill, 1997; Kalbfleisch & Prentice, 1973) and are, for instance, implemented in the R package survival.

3.2 Model specification and selection

The core objectives of our analysis are to understand and illustrate higher order effects in polyadic interaction networks. We estimate two types of models, a conventional REM using only dyadic covariates ('dyadic model') and RHEM ('polyadic model').

The dyadic model includes the covariates receiver set average for the actor-level attributes *gender*, *seniority*, *trading*, and *legal* and sender–receiver heterophily for *gender*, *seniority*, and for the categorical attribute *department*. The dyadic model also includes the network effects partial receiver set repetition and sender-specific partial receiver set repetition of order $p = 1$, reciprocation, out-in popularity, and all four triadic closure effects.

The RHEM includes all covariates of the dyadic model and in addition the covariates receiver set heterophily for *gender*, *seniority*, and for the categorical attribute *department*. Moreover, the RHEM includes exact repetition, unordered repetition, as well as (sender-specific) partial receiver set repetition and interaction among receivers for varying values of p that have been determined in a preliminary analysis (see below). Additionally, we specify and estimate two further variants of the dyadic REM and two of the polyadic RHEM. These restricted models include covariates based on actor-level attributes only, and network effects only, respectively.

Model estimation proceeds in two steps. In the first step, we apply the eventnet software (Lerner & Lomi, 2020; Lerner et al., 2021), available at https://github.com/juergenlerner/eventnet, to sample k non-event hyperedges associated with each observed hyperevent and to compute (a superset of) all covariates of observed events and sampled non-events. We set the number of non-events per event to $k = 100$. We initially compute (sender-specific) partial receiver set repetition and interaction among receivers of order $p = 1, \ldots, 10$. All network effects are defined with the half-life period $T_{1/2}$ set to one week. (While the decision about the half life is essentially arbitrary, and unrelated to the objectives of this paper, it can be expected that one week somewhat levels out day-of-the-week effects, which are likely to be prominent in a corporate communication network.) In the second step, we estimate parameters of models specified by varying lists of covariates with the coxph function in the R package survival (Therneau, 2015; Therneau &

Grambsch, 2013). To assess the variation of estimates over different samples, we repeat the sampling of risk sets 100 times, recomputing the covariate values each time. This gives us, for each actual choice of the covariate vector x_t, 100 different log partial-likelihood functions (7) and 100 potentially different estimated parameter vectors $\hat{\beta}$. Since covariates modelling network effects are skewed, we transform them by $x \mapsto \sqrt{x}$, but we do not standardise covariates.

To determine reasonable values of p for the covariates (sender-specific) partial receiver set repetition and interaction among receivers, we incrementally add these covariates for increasing values of p and monitor robustness of parameter estimation over different samples, as well as model fit assessed by information criteria (AIC and BIC). We find that for a model specified with partial receiver set repetition up to order $p = 4$ and sender-specific partial receiver set repetition and interaction among receivers up to order $p = 3$, estimation converges for all 100 samples and model fit measured in AIC and BIC consistently increases for growing order of these effects. In the following, we denote this model with 28 covariates as 'the RHEM' (see the rightmost model in Table 3).

Among the covariates included in the RHEM, two pairs have correlation exceeding 0.9. These are the pairs exact repetition and unordered repetition, with a correlation of 0.93, and partial receiver set repetition and sender-specific partial receiver set repetition of order $p = 3$, with a correlation equal to 0.98. To test how findings are affected by such high correlations, we additionally fit a reduced RHEM without unordered repetition and without sender-specific partial receiver set repetition of order $p = 3$. This reduced RHEM has a lower model fit than the RHEM including all effects (AIC = 73,286 for the reduced model and AIC = 72,919 for the largest RHEM). Most parameters that are included in both models are not affected qualitatively in the sense that they keep their signs and significance levels. The only exceptions are (1) out-in popularity becomes significantly positive in the reduced model (and is non-significant in the largest RHEM) and (2) the parameter of exact repetition switches its sign from significantly positive to significantly negative when we additionally include unordered repetition. We argue that this sign switch does not point to lack of robustness in our findings—but rather to a relevant effect in multicast

Table 2. Estimated parameters of dyadic models

	att (dyad)	net (dyad)	att+net (dyad)
rec_avg_female	0.24 (0.02)***		0.26 (0.02)***
rec_avg_senior	0.71 (0.02)***		0.45 (0.02)***
rec_avg_legal	1.22 (0.02)***		0.15 (0.03)***
rec_avg_trading	−0.61 (0.03)***		−0.22 (0.03)***
send_rec_diff_female	−0.54 (0.02)***		−0.24 (0.02)***
send_rec_diff_senior	−1.04 (0.02)***		−0.49 (0.02)***
send_rec_diff_dept	−2.10 (0.02)***		−0.88 (0.02)***
rec_sub_rep_1		0.11 (0.01)***	0.04 (0.01)**
send_rec_sub_rep_1		2.70 (0.02)***	2.54 (0.02)***
reciprocation		1.04 (0.02)***	1.00 (0.02)***
out_in_pop		−0.00 (0.01)	0.01 (0.01)
in_balance		0.11 (0.01)***	0.10 (0.01)***
out_balance		−0.02 (0.01)*	−0.04 (0.01)***
transitive_closure		0.12 (0.01)***	0.11 (0.01)***
cyclic_closure		−0.09 (0.01)***	−0.11 (0.01)***
AIC	164,871.80	87,983.89	85,158.09
Num. events	21,635	21,635	21,635
Num. obs.	2,185,135	2,185,135	2,185,135

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Table 3. Estimated parameters of RHEM

	att (rhem)	net (rhem)	att+net (rhem)
rec_avg_female	0.28 (0.02)***		0.21 (0.02)***
rec_avg_senior	0.68 (0.02)***		0.32 (0.02)***
rec_avg_legal	1.23 (0.02)***		0.13 (0.03)***
rec_avg_trading	−0.49 (0.03)***		−0.11 (0.03)***
send_rec_diff_female	−0.50 (0.02)***		−0.19 (0.02)***
send_rec_diff_senior	−0.94 (0.02)***		−0.42 (0.02)***
send_rec_diff_dept	−1.88 (0.02)***		−0.73 (0.02)***
rec_set_diff_female	−0.43 (0.04)***		−0.18 (0.07)**
rec_set_diff_senior	−1.42 (0.04)***		−0.65 (0.07)***
rec_set_diff_dept	−2.22 (0.04)***		−0.99 (0.07)***
exact_repetition		−0.49 (0.06)***	−0.37 (0.06)***
unordered_repetition		1.04 (0.05)***	0.98 (0.05)***
rec_sub_rep_1		0.07 (0.01)***	0.02 (0.01)
rec_sub_rep_2		0.65 (0.06)***	0.47 (0.06)***
rec_sub_rep_3		2.18 (0.27)***	1.93 (0.26)***
rec_sub_rep_4		6.73 (0.80)***	6.12 (0.79)***
send_rec_sub_rep_1		1.99 (0.04)***	1.80 (0.04)***
send_rec_sub_rep_2		5.08 (0.21)***	4.83 (0.21)***
send_rec_sub_rep_3		5.15 (1.01)***	4.03 (0.92)***
reciprocation		0.65 (0.03)***	0.63 (0.03)***
out_in_pop		−0.01 (0.01)	0.01 (0.01)
interact_rec_1		2.89 (0.09)***	2.67 (0.09)***
interact_rec_2		7.30 (0.88)***	6.93 (0.87)***
interact_rec_3		34.90 (5.15)***	32.74 (5.21)***
in_balance		0.08 (0.01)***	0.08 (0.01)***
out_balance		−0.09 (0.01)***	−0.10 (0.01)***
transitive_closure		0.06 (0.01)***	0.05 (0.01)***
cyclic_closure		−0.04 (0.01)**	−0.05 (0.01)***
AIC	159,300.98	75,196.57	72,919.66
Num. events	21,635	21,635	21,635
Num. obs.	2,185,135	2,185,135	2,185,135

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

communication networks that can be well explained (see the discussion related with exact repetition and unordered repetition below).

4 Results and discussion

4.1 Discussion of effects

Estimated parameters and standard errors of the three models specified with dyadic covariates are reported in Table 2, those of RHEM including dyadic and higher-order effects are reported in Table 3. In the following, we discuss findings to highlight representative effects, or types of effects, across all models in which the respective covariates are included. Our goal is not to discuss the findings exhaustively, but rather to offer guidance how to interpret effects and illustrate the

additional results that RHEM make possible, and the new questions they afford. Even though we fitted models to 100 different samples, the reported parameters have been estimated from one arbitrarily selected single sample. Indeed, we believe that this situation is more representative for empirical studies than results derived from repeated sampling. We report summary statistics of the distribution of parameters over 100 samples in Appendix A. We emphasise that dyadic REM and RHEM are estimated on the same sample; the only difference is the effects included in the empirical model specifications.

Effects of actor-level attributes

The attribute distribution in receiver sets is captured by the four covariates measuring receiver set averages (rec_avg_z). All four models including attribute effects agree that receiver sets have an over-representation of females, seniors, and members of the legal department and an under-representation of members of the trading department (recall that the base, or reference, department is 'other'). Parameters of covariates assessing the difference in attribute values between the sender and the receivers consistently reveal that senders tend to write messages to receivers of the same gender, seniority, and department. On top of this 'sender–receiver homophily', the two RHEM containing attribute effects reveal that receiver sets tend to be homogeneous with respect to gender, seniority, and department (negative parameters of the 'receiver set difference' covariates). We observe that without any exception effect sizes of covariates based on attributes become smaller when we additionally control for network effects. Moreover, typically—but not without exception—sizes of attribute effects are stronger in the dyadic models than in the RHEM. Intuitively, these results reveal a possible tendency to over-estimate the effects of actor-specific attributes in models that do not control for network effects, or that do not control for higher-order effects, respectively.

Exact repetition and unordered repetition

In the RHEM, the parameter associated with unordered repetition is significantly positive and that of exact repetition is significantly negative. These two effects have to be interpreted jointly—as we illustrate building on the example from Figure 1. Assume that at t_1 actor A sends an email to $\{B, C, D, E\}$ and consider two alternative hyperedges $h = (C, \{A, B, D, E\})$ and $h' = (A, \{B, C, D, E\})$ at a later point in time $t > t_1$. The past email $e_1 = (t_1, A, \{B, C, D, E\})$ increases the value of exact repetition and it increases the value of unordered repetition on h' at t. The joint effect of these two covariates is positive (since unordered repetition has a larger parameter than the absolute value of the negative repetition parameter). Thus, the past event $e_1 = (t_1, A, \{B, C, D, E\})$ makes a repeated email from A to $\{B, C, D, E\}$ at $t > t_1$ more likely. However, an event at t on hyperedge h, that is, sent from C to $\{A, B, D, E\}$ would be even more likely than an email from A to $\{B, C, D, E\}$. This is because the past email $e_1 = (t_1, A, \{B, C, D, E\})$ increases the value of unordered repetition on h at t (having a positive effect) but it does not increase the value of repetition (which would have had a negative effect).

Thus, while conversations within fixed lists of actors (e.g., A, B, C, D, E in the example from Figure 1) are overrepresented, it is more likely that a future email within the same fixed set of actors will have a sender different from that of the preceding email. This points to a form of 'turn-taking' (Gibson, 2005) within fixed conversation groups. We note that it would not be possible to express these effects exclusively with dyadic covariates.

Exact repetition and unordered repetition express a form of behavioural inertia, indicating that actors tend to repeat what they did in the past.

(Sender-specific) partial receiver set repetition and interaction among receivers

In the RHEM almost all covariates expressing subset repetition in the receiver set (with order $p = 1, 2, 3, 4$), all covariates for sender-specific subset repetition in the receiver set (with order $p = 1, 2, 3$), and interaction among receivers have a significantly positive effect. (Sender-specific) receiver subset repetition of order one are dyadic effects and thus are also included in the dyadic model, where they are associated with positive parameters.

All three types of effects can be interpreted in defining a similarity (or suitability) measure on sets of actors based on past interaction, which in turn tends to increase the rate of future interactions

send by actor i to receiver set J. Specifically, partial receiver set repetition considers J as a suitable receiver set to the extent that subsets $J' \subseteq J$ have already received common interactions; sender-specific partial receiver set repetition considers the hyperedge (i, J) as suitable for the next interaction to the extent that subsets $J' \subseteq J$ have already received common interactions from the sender i; the interaction among receivers effect considers J as a suitable receiver set to the extent that subsets $J' \subseteq J$ have already received common interactions from a sender $i' \in J \setminus J'$. In all three types of effects, the order of the effect corresponds to the size of the subsets J'.

The findings demonstrate that repetition from the same sender to the same receivers (possibly within a larger and varying set of yet other receivers) is not a purely dyadic phenomenon. Rather, if two (or three or four) actors have already jointly received an email, then they are more likely to do so again. This points to a clustering of the set of all possible receivers into subsets that are likely to receive the same emails. We find evidence for a 'global' clustering (that is, one that applies to the average sender) and for a sender-specific clustering (so that one sender can structure the space of receivers in a different way than another sender).

Reciprocation and out-in popularity

A tendency to reciprocate interaction is found in all four models containing this effect. Thus, receivers of past interaction have the tendency to send interaction to the sender of that past interaction. Beyond that effect, all four models agree that there is no evidence for out-in popularity. This latter effect tests whether those actors who have sent out more interaction in the past are likely to receive more future interaction—from anyone in the network, that is, not necessarily from the receivers of the past interaction.

Triadic effects

The triadic effects can be included in the dyadic model and in the RHEM. All four models agree that there is a tendency for transitive closure but a tendency against cyclic closure. Together these two findings are compatible with the 'hierarchical' interpretation that messages tend to go from higher status to lower status actors (Lerner & Lomi, 2017). Note that these findings are equally compatible with the interpretation that messages tend to go 'upward' in the hierarchy. We also find that two actors who have received (possibly different) messages from the same third actor (incoming balance) are more likely to interact themselves but two actors who have both sent messages to the same third actor (outgoing balance) are less likely to interact themselves.

4.2 Contribution of individual covariates to the log-likelihood

In Table 4, we list improvements in log likelihood of the 28 models specified with exactly one covariate over the null model (i.e., the model with no covariates). We also report the contribution of individual covariates in the full model by measuring improvements of the full model (i.e., the 'RHEM' with 28 covariates) over the 28 models obtained by dropping exactly one covariate. These latter measurements indicate how much each individual covariate contributes over the 27 others.

We find that sender-specific partial receiver set repetition of order one (which is a dyadic covariate) is the strongest individual effect, closely followed by unordered repetition and exact repetition (both of which are non-dyadic hyperedge covariates). The covariates based on actor-level attributes imply much lower contributions to the log likelihood than the network effects. Looking at the contributions of single effects on top of all other 27 covariates (rightmost column) we find—with exception of the strongest effect—a remarkably different order. However, the second strongest effect, sender-specific partial receiver set repetition of order two, is again a non-dyadic hyperedge effect. We conclude that the contributions of some of the higher-order effects are roughly comparable to those of the strongest dyadic effects.

4.3 Evidence of higher-order effects

Overall, we find strong evidence of the presence of higher-order effects in the data. Covariates that go beyond purely dyadic effects have been found to be significantly predictive of future event distributions. Moreover, as shown above, the contributions of some higher-order effects to the log likelihood have similar magnitudes as the strongest dyadic covariates.

Table 4. Improvement in log-likelihood resulting from individual covariates

	Over null model	In full model
send_rec_sub_rep_1	53,649.62	1,247.12
unordered_repetition	47,906.19	173.16
exact_repetition	44,603.05	19.00
reciprocation	36,047.57	169.54
send_rec_sub_rep_2	22,539.50	530.31
transitive_closure	20,468.45	10.09
out_balance	18,562.00	57.88
rec_sub_rep_2	18,444.34	25.76
cyclic_closure	17,300.53	6.30
in_balance	16,599.53	43.17
rec_sub_rep_1	16,274.49	1.46
interact_rec_1	15,086.75	433.19
rec_sub_rep_3	13,251.15	24.83
send_rec_sub_rep_3	13,089.58	13.23
send_rec_diff_dept	11,048.89	524.04
interact_rec_2	10,794.70	46.04
out_in_pop	10,294.25	0.07
rec_sub_rep_4	8,382.27	34.92
interact_rec_3	6,614.91	28.90
rec_avg_legal	5,599.23	7.55
rec_set_diff_dept	2,430.60	103.30
send_rec_diff_senior	2,250.83	183.04
rec_avg_trading	1,955.37	8.20
rec_set_diff_senior	1,071.38	43.45
rec_avg_senior	1,001.24	92.09
send_rec_diff_female	779.91	33.37
rec_avg_female	461.45	38.04
rec_set_diff_female	70.88	3.37

As outlined above, we contend that such interdependence between different receivers should not be considered as an annoyance to be controlled away—but rather as an opportunity to develop and test additional theories about the structure of polyadic interaction networks. Some of the higher-order effects are highly relevant for empirical studies. Homophily in the receiver set, exact repetition and unordered repetition, and (sender-specific) partial receiver-set repetition of order $p \geq 2$ all have relevant implications for the structure and dynamics of polyadic interaction networks.

4.4 Qualitative differences between RHEM and dyadic models

In most cases, parameters of covariates that have been included in dyadic models and in RHEM are estimated to have the same sign and significance levels. However, in most cases the absolute values of these parameters are considerably smaller in the RHEM than in the dyadic models. This may suggest that failure to control for higher-order effects could lead to an overestimation of effect sizes. In our empirical analysis, the dyadic models suggested that partial receiver set repetition of order one is significantly positive, while the largest RHEM found the same effect to be not significant.

4.5 Model fit of RHEM and dyadic model

Regarding model fit, measured with information criteria (AIC), we find two patterns. First, network effects improve the model fit much more than attribute effects and the joint models including attribute and network effects have the best fit in the family of dyadic models and in RHEMs. Second, RHEM fit the data better than the corresponding dyadic models. We find that the RHEM purely specified with network effects 'net (rhem)' already has a better fit than the full dyadic model 'att+net (dyad)'.

4.6 Replicability

The analysis is explained in a tutorial linked from https://github.com/juergenlerner/eventnet/wiki/, from the data preprocessing and the computation of covariates over to model estimation in R. Thus, the analysis reported in this paper is fully replicable and the tutorial points out some model variations that might be relevant in various application scenarios. The software may be freely adopted in—and adapted to future empirical studies involving polyadic social interaction processes.

5 Conclusion

The objectives of our paper are to introduce, analyse, and illustrate hyperedge covariates for polyadic interaction networks and to compare RHEM, including higher-order effects, with REM specified exclusively in terms of dyadic covariates. In our illustrative empirical analysis of the Enron email data, we find consistent evidence for the significance and importance of higher-order effects captured by hyperedge covariates that jointly consider receiver sets of size larger than one. Failure to control for such higher-order dependencies can also affect estimated parameters of dyadic covariates, although, in our analysis, this change seems to affect mostly parameter sizes (and occasionally also their significance levels), but in most cases does not switch the direction of effects (i.e., parameter signs).

The results we reported suggest that researchers interested in analysing polyadic interaction networks should not restrict their models to dyadic specifications. From a high-level perspective, the step from models for multicast networks specified via dyadic covariates (Perry & Wolfe, 2013) to RHEM is relatively intuitive. However, hyperedge covariates provide a much richer set of possible effects and considerably increase the fit of models for the given empirical data.

These results strengthen our view that higher-order effects should not be considered merely as an annoyance that has to be controlled for—but rather as an opportunity to develop and test additional theories about the structure and dynamics of social interaction and communication networks. Some of the effects that we have documented could not have been discovered or tested with available models purely specified via dyadic covariates. Concretely, this includes findings on unordered repetition (pointing to turn-taking within a fixed list of conversation participants), partial receiver set repetition (related with a clustering of the actors into groups that are likely receivers of a joint message), and sender-specific partial receiver set repetition (pointing to patterns in which different senders cluster the actors into potentially different groups).

We point out that average hyperedge sizes in the Enron email data are still rather small: 70% of the emails have exactly one receiver and the average number of receivers is 1.77. In other empirical settings, for instance, coauthorship networks (Newman, 2004), hyperevents are typically much larger. For example, in the coauthorship networks considered in Lerner et al. (2019), the average number of authors per paper is close to eight and some events have size up to 100. As another example, in the meeting events extracted from contact diaries analysed in Lerner et al. (2021), the average number of participants per meeting is 4.4. With larger hyperevents, it is typically possible to include subset repetition covariates of higher order.

Our primary goal in this paper was to establish hyperedge covariates and compare RHEM to dyadic models, rather than elaborate new parameter estimation techniques. To deal with the exponential size of the risk set, we proposed case-control sampling—which is a readily available and well-established technique in the analysis of rare events (Borgan et al., 1995; Keogh & Cox, 2014). Hyperevents could indeed be considered as rare events since a randomly selected subset of actors is very unlikely to experience even one common event. However, case-control sampling, in which we uniformly sample from the risk set, revealed some limitations, mostly due to the fact that some

higher-order covariates are very sparse in the risk set. In turn, most sampled non-events did not provide any information on those covariates which increased the necessary sample size. A more efficient sampling scheme could be stratified sampling, where, for instance, one could include as strata non-events that are non-zero in some subset repetition covariates. Another possibility would be Markov-chain Monte Carlo (MCMC) methods which do not sample uniformly from the (stratified or unstratified) risk set but sample proportionally to the conditional event probability of subsets. The additional cost of MCMC methods, however, is that sampling is no longer independent of model parameters.

We point out that, even if we do not include subset repetition covariates of order five or higher, our models do not preclude hyperevents of that size. The reason for this is that subset repetition covariates are nested within each other: whenever a past event implies a non-zero value in subset repetition of order p, then it necessarily implies non-zero values in subset repetition of lower order $p' < p$; compare the examples given in Section 2.4.2.

Besides improved parameter estimation techniques for RHEM, we consider the further development of other RHEM covariates as a promising direction for future work. As we have discussed, the collection of hyperedge covariates proposed in this paper is far from exhausting the

Table A1. Summary statistics (quantiles at probabilities 0.0, 0.025, 0.5, 0.975, 1.0) of RHEM parameters over 100 samples

	0%	2.5%	50%	97.5%	100%
rec_avg_female	0.183	0.196	0.212	0.230	0.236
rec_avg_senior	0.322	0.327	0.339	0.354	0.358
rec_avg_legal	0.105	0.111	0.131	0.153	0.165
rec_avg_trading	−0.124	−0.122	−0.108	−0.092	−0.089
send_rec_diff_female	−0.203	−0.200	−0.185	−0.172	−0.166
send_rec_diff_senior	−0.440	−0.433	−0.420	−0.409	−0.407
send_rec_diff_dept	−0.753	−0.751	−0.737	−0.724	−0.717
rec_set_diff_female	−0.286	−0.259	−0.203	−0.133	−0.091
rec_set_diff_senior	−0.754	−0.743	−0.687	−0.640	−0.613
rec_set_diff_dept	−1.072	−1.056	−0.987	−0.929	−0.903
exact_repetition	−0.517	−0.490	−0.428	−0.377	−0.361
unoredered_repetition	0.962	0.970	1.019	1.076	1.087
rec_sub_rep_1	0.020	0.022	0.028	0.036	0.038
rec_sub_rep_2	0.321	0.374	0.468	0.564	0.580
rec_sub_rep_3	1.258	1.358	1.868	2.347	2.588
rec_sub_rep_4	3.047	3.270	4.943	7.006	8.248
send_rec_sub_rep_1	1.765	1.775	1.807	1.841	1.850
send_rec_sub_rep_2	4.078	4.502	5.173	6.083	6.269
send_rec_sub_rep_3	3.615	4.133	8.764	15.288	16.344
reciprocation	0.555	0.576	0.607	0.640	0.652
out_in_pop	−0.002	0.001	0.010	0.018	0.019
interact_rec_1	2.365	2.401	2.599	2.798	2.866
interact_rec_2	4.325	5.322	7.851	10.518	10.962
interact_rec_3	3.922	10.185	22.745	33.437	36.689
in_balance	0.051	0.057	0.068	0.076	0.078
out_balance	−0.125	−0.123	−0.111	−0.098	−0.093
transitive_closure	0.045	0.048	0.057	0.073	0.073
cyclic_closure	−0.061	−0.057	−0.042	−0.025	−0.022

Table A2. Summary statistics (quantiles at probabilities 0.0, 0.025, 0.5, 0.975, 1.0) of dyadic REM parameters over 100 samples

	0%	2.5%	50%	97.5%	100%
rec_avg_female	0.224	0.237	0.254	0.275	0.280
rec_avg_senior	0.448	0.457	0.468	0.485	0.489
rec_avg_legal	0.119	0.129	0.151	0.174	0.182
rec_avg_trading	−0.240	−0.237	−0.221	−0.202	−0.197
send_rec_diff_female	−0.252	−0.251	−0.234	−0.220	−0.215
send_rec_diff_senior	−0.512	−0.504	−0.491	−0.479	−0.476
send_rec_diff_dept	−0.909	−0.905	−0.892	−0.878	−0.867
rec_sub_rep_1	0.036	0.037	0.044	0.052	0.054
send_rec_sub_rep_1	2.505	2.508	2.530	2.547	2.555
reciprocation	0.976	0.985	1.007	1.034	1.042
out_in_pop	0.008	0.011	0.018	0.026	0.028
in_balance	0.081	0.088	0.098	0.105	0.110
out_balance	−0.059	−0.059	−0.048	−0.037	−0.032
transitive_closure	0.102	0.105	0.116	0.127	0.134
cyclic_closure	−0.126	−0.122	−0.107	−0.094	−0.086

possibilities. Developing further effects that are relevant for given application scenarios will further increase the empirical extension and generality of RHEM.

Conflict of interests: The authors declare no conflict of interest.

Data availability

The empirical data analysed in this work are linked from https://github.com/juergenlerner/eventnet/wiki/Directed-RHEM-for-multicast-interaction-(tutorial), along with pointers to the used software, scripts, and a step-by-step tutorial.

Appendix. Variation across samples

We repeated parameter estimation on 100 different samples of non-events associated with each observed event. In Table A1, we report summary statistics of RHEM parameters over these 100 samples and in Table A2 we report summaries of parameters of the dyadic REM. We can observe that the distributions of parameters capturing (sender-specific) subset repetition and interaction among receivers have a dispersion that is increasing with the order of these covariates. This can be explained by the fact that these covariates are increasingly sparse among the non-events, that is, most randomly sampled non-events assume the value zero on these higher-order statistics. As indicated in the conclusion, this sparsity implies the need for a large sample size when parameters are estimated via case-control sampling—but could possibly be reduced by other parameter estimation techniques (for instance, via stratified sampling or MCMC methods). However, even the distributions of these parameters consistently do not cross zero. Parameter distributions of most other effects show less relative variability.

References

Ahmadpoor M., & Jones B. F. (2019). Decoding team and individual impact in science and invention. *Proceedings of the National Academy of Sciences*, 116(28), 201812341. https://doi.org/10.1073/pnas.1812341116

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. https://doi.org/10.1016/j.socnet.2018.10.001

Berge C. (1989). *Hypergraphs: Combinatorics of finite sets*. North-Holland.

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

Borgan Ø., Goldstein L., & Langholz B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics*, 23(5), 1749–1778. https://doi.org/10.1214/aos/1176324322

Brandes U., Lerner J., & Snijders T. A. (2009). Networks evolving step by step: Statistical analysis of dyadic event data. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM)* (pp. 200–205). IEEE Computer Society.

Breslow N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1), 89–99. https://doi.org/10.2307/2529620

Bronsteen J., & Fiss O. (2002). The class action rule. *Notre Dame Law Review*, 78, 1419.

Butts C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1), 155–200. https://doi.org/10.1111/j.1467-9531.2008.00203.x

Colizza V., Barrat A., Barthelemy M., Valleron A.-J., & Vespignani A. (2007). Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Medicine*, 4(1), e13. https://doi.org/10.1371/journal.pmed.0040013

Diesner J., Frantz T. L., & Carley K. M. (2005). Communication networks from the Enron email corpus 'It's always about the people. Enron is no different'. *Computational & Mathematical Organization Theory*, 11(3), 201–228. https://doi.org/10.1007/s10588-005-5377-0

Efron B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359), 557–565. https://doi.org/10.1080/01621459.1977.10480613

Fowler J. H., Johnson T. R., Spriggs J. F., Jeon S., & Wahlbeck P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the US Supreme Court. *Political Analysis*, 15(3), 324–346. https://doi.org/10.1093/pan/mpm011

Freeman L. C. (2003). Finding social groups: A meta-analysis of the Southern Women data. In R. Breiger, C. Carley, & P. Pattison (Eds.), *Dynamic social network modeling and analysis: Workshop summary and papers* (pp. 39–97). National Research Council, The National Academies Press.

Gibson D. R. (2005). Taking turns and talking ties: Networks and conversational interaction. *American Journal of Sociology*, 110(6), 1561–1597. https://doi.org/10.1086/428689

Guimera R., Uzzi B., Spiro J., & Amaral L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722), 697–702. https://doi.org/10.1126/science.1106340

Hâncean M.-G., Lerner J., Perc M., Ghiţă M. C., Bunaciu D.-A., Stoica A. A., & Mihăilă B.-E. (2021). The role of age in the spreading of COVID-19 across a social network in Bucharest. *Journal of Complex Networks*, 9(4), cnab026. https://doi.org/10.1093/comnet/cnab026

Hertz-Picciotto I., & Rockhill B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, 53(3), 1151–1156. https://doi.org/10.2307/2533573

Hollway J., & Koskinen J. (2016). Multilevel bilateralism and multilateralism: States' bilateral and multilateral fisheries treaties and their secretariats. In *Multilevel network analysis for the social sciences* (pp. 315–332). Springer.

Kalbfleisch J. D., & Prentice R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2), 267–278. https://doi.org/10.1093/biomet/60.2.267

Keila P. S., & Skillicorn D. B. (2005). Structure in the Enron email dataset. *Computational & Mathematical Organization Theory*, 11(3), 183–199. https://doi.org/10.1007/s10588-005-5379-y

Keogh R. H., & Cox D. R. (2014). *Nested case-control studies* (Chap. 7, pp. 160–190). Institute of Mathematical Statistics Monographs. Cambridge University Press.

Kim B., Schein A., Desmarais B. A., & Wallach H. (2018). 'The hyperedge event Model,' arXiv, arXiv:1807.08225, preprint: not peer reviewed.

Kuhn J., Younge K., & Marco A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1), 109–132. https://doi.org/10.1111/rand.v51.1

Lazer D., Pentland A., Adamic L., Aral S., Barabasi A.-L., Brewer D., Christakis N., Contractor N., Fowler J., & Gutmann M., *et al.* (2009). Computational social science. *Science*, 323(5915), 721–723. https://doi.org/10.1126/science.1167742

Leenders R. T. A., Contractor N. S., & DeChurch L. A. (2016). Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review*, 6(1), 92–115. https://doi.org/10.1177/2041386615578312

Lerner J., Bussmann M., Snijders T. A., & Brandes U. (2013). Modeling frequency and type of interaction in event networks. *Corvinus Journal of Sociology and Social Policy*, 4(1), 3–32. https://doi.org/10.14267/issn.2062-087X

Lerner J., & Lomi A. (2017). The third man: Hierarchy formation in Wikipedia. *Applied Network Science*, 2(1), 1–30. https://doi.org/10.1007/s41109-017-0043-2

Lerner J., & Lomi A. (2020). Reliability of relational event model estimates under sampling: How to fit a relational event model to 360 million dyadic events. *Network Science*, 8(1), 97–135. https://doi.org/10.1017/nws.2019.57

Lerner J., & Lomi A. (2022). A dynamic model for the mutual constitution of individuals and events. *Journal of Complex Networks*, 10(2), cnac004. https://doi.org/10.1093/comnet/cnac004

Lerner J., Lomi A., Mowbray J., Rollings N., & Tranmer M. (2021). Dynamic network analysis of contact diaries. *Social Networks*, 66, 224–236. https://doi.org/10.1016/j.socnet.2021.04.001

Lerner J., Tranmer M., Mowbray J., & Hancean M.-G. (2019). 'REM beyond dyads: Relational hyperevent models for multi-actor interaction networks', arXiv, arXiv:1912.07403, preprint: not peer reviewed.

Mulder J., & Hoff P. D. (2021). 'A latent variable model for relational events with multiple receivers', arXiv, arXiv:2101.05135, preprint: not peer reviewed.

Newman M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl_1), 5200–5205. https://doi.org/10.1073/pnas.0307545100

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. https://doi.org/10.1111/rssb.2013.75.issue-5

Radicchi F., Fortunato S., & Vespignani A. (2012). Citation networks. In Scharnhorst, A., Börner, K., & Besselaar, P. (Eds.). *Models of Science Dynamics* (pp. 233-257). Springer.

Schecter A., & Quintane E. (2021). The power, accuracy, and precision of the relational event model. *Organizational Research Methods*, 24(4), 802–829. https://doi.org/10.1177/1094428120963830

Simmons B. A., & Hopkins D. J. (2005). The constraining power of international treaties: Theory and methods. *American Political Science Review*, 99(4), 623–631. https://doi.org/10.1017/S0003055405051920

Snijders T. A., & Lomi A. (2019). Beyond homophily: Incorporating actor variables in statistical network models. *Network Science*, 7(1), 1–19. https://doi.org/10.1017/nws.2018.30

Therneau T. M. (2015). *A package for survival analysis in S*, version 2.38.

Therneau T. M., & Grambsch P. M. (2013). *Modeling survival data: Extending the Cox model*. Springer Science & Business Media.

Verspagen B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, 10(01), 93–115. https://doi.org/10.1142/S0219525907000945

Vu D., Lomi A., Mascia D., & Pallotti F. (2017). Relational event models for longitudinal network data with an application to interhospital patient transfers. *Statistics in Medicine*, 36(14), 2265–2287. https://doi.org/10.1002/sim.v36.14

Vu D., Pattison P., & Robins G. (2015). Relational event models for social learning in MOOCs. *Social Networks*, 43, 121–135. https://doi.org/10.1016/j.socnet.2015.05.001

Zhou Y., Goldberg M., Magdon-Ismail M., & Wallace A. (2007). Strategies for cleaning organizational emails with an application to Enron email dataset. In *5th Conference of North American Association for Computational Social and Organizational Science*. North American Association for Computational Social and Organizational Science.