Predicting Network Events to Assess Goodness of Fit of Relational Event Models

Laurence Brandenberger[1,2]

[1] Chair of Systems Design, ETH Zurich, Weinbergstr. 56/ 58, CH-8092 Zurich, Switzerland. Email: lbrandenberger@ethz.ch
[2] Institute of Political Science, University of Bern, CH-3012 Bern, Switzerland

Abstract

Relational event models are becoming increasingly popular in modeling temporal dynamics of social networks. Due to their nature of combining survival analysis with network model terms, standard methods of assessing model fit are not suitable to determine if the models are specified sufficiently to prevent biased estimates. This paper tackles this problem by presenting a simple procedure for model-based simulations of relational events. Predictions are made based on survival probabilities and can be used to simulate new event sequences. Comparing these simulated event sequences to the original event sequence allows for in depth model comparisons (including parameter as well as model specifications) and testing of whether the model can replicate network characteristics sufficiently to allow for unbiased estimates.

Keywords: dynamic network, goodness of fit, prediction, relational event model

1 Introduction

Dynamic networks—networks that evolve over time—can be analyzed using relational event models (REMs). First presented by Butts (2008), these models examine how sequences of relational events progress through time. Each of these events represents an edge (or tie) forming in a network at a distinct point in time. This flexible and dynamic form of network inference can be used to examine how actors behave in changing network settings. Examples of event networks include email communications (DuBois, Butts, and Smyth 2013), parliamentarians bargaining over new regulations (Desmarais *et al.* 2015; Brandenberger 2018b), patient transfers between hospitals (Kitts *et al.* 2016), or individuals interacting online (Welbers and de Nooy 2014; Quintane *et al.* 2014). The additional information regarding the timing of events allows for a more precise estimation of popular network effects such as popularity, triadic closure or homophily effects. Inference on how networks evolve over time can be gained from combining network effects with statistical models from survival analysis, such as stratified Cox models which can be estimated through conditional logistic regressions (Andersen and Gill 1982).

However, estimated parameters of these models may suffer from a form of omitted variable bias if the endogenous dependencies are not specified correctly and/or sufficiently, resulting in a misspecification of the joint likelihood of the model (Butts 2008, 168). In other words, if the endogenous properties of the event sequences are poorly captured by the model terms, the estimated effects on event occurrence are unreliable. Standard approaches to detecting omitted variable bias in survival models, such as information criteria or precision-recall (PR) curves, fail to detect shortcomings in the network specifications and cannot give guidance as to which endogenous terms should be included to improve fit.

Author's note: The author would like to thank Philip Leifeld for helpful comments and conversations as well as participants of the conference panel 'Modeling Network Dynamics II: Time-stamped Network Data' at the Third European Conference on Social Networks in Mainz (September, 2017) for helpful questions and comments. The author would also like to thank two anonymous reviewers and the editor for their valuable comments. LB carried out parts of this research while at the Swiss Federal Institute of Aquatic Science and Technology (Eawag).

Political Analysis (2019)
vol. 27:556–571
DOI: 10.1017/pan.2019.10

Published
29 April 2019

Corresponding author
Laurence Brandenberger

Edited by
Jeff Gill

This paper addresses these shortcomings and presents a simple approach to predicting relational events as well as goodness of fit measures to evaluate the choice of sufficient statistics for REMs. Predictions are based on survivor probabilities and are calculated from model parameters and endogenous network statistics. New events are chosen from a set of potential events—or risk set—based on their survivor probabilities. Events with lower survivor probability have a lower chance of surviving the present event and are more likely to occur. After one or more events are chosen to occur at a distinct point in the event sequence, subsequent events are predicted in the same manner—however, their endogenous network statistics change due to the newly chosen previous events. These newly simulated sequences can then be used to compare different models and their predictive power as well as check if network dynamics are captured satisfactorily. Furthermore, the approach can be used to examine shortcomings in the specification of the endogenous network statistics. By comparing network characteristics of the simulated sequences with the original event sequence the approach can show which network characteristics are captured well by the simulation and which are lacking. The REM can then be complemented with additional endogenous network terms to achieve a better fit and model network dependencies in the event sequence more adequately.

The paper starts off with a short introduction to REMs. After the discussion about shortcomings of conventional goodness of fit tools for these statistical models, the new prediction procedure for REMs is presented. A simulation is used to verify the validity of the new prediction procedure. Afterward, the prediction procedure is applied to a REM on a political debate (Leifeld 2016; Leifeld and Brandenberger 2019) and three distinct goodness of fit tests are presented.

2 Predicting Relational Events Using Survivor Probabilities

2.1 Relational event models: an overview

Relational event models build on survival analysis and use network dependencies in event sequences to estimate which factor expedite event occurrence. A sequence of events represent micro-steps in a dynamic network and consist of a sender node, a target node and exact or ordinal timing. REMs are used to estimate the effect of past network events on future events whilst controlling for exogenous factors. Since all network changes in the model are reflected in the event sequence, these events can be considered conditionally independent of one another and can therefore be analyzed using conventional regression models (Butts 2008; Lerner *et al.* 2013).

The general idea behind REMs is that event occurrence is modeled using a piecewise constant hazard model (Butts 2008). The likelihood that an event or a number of events $n_{ij}(t)$ take place on a dyad (i, j) within the time interval t is given by the hazard rate $\lambda_{ij}(t)$. The hazard rate is then multiplied by the survival function $\exp(-\lambda_{ij}(t))$, which captures all events that could have occurred at time t yet did not (see Butts 2008, 161–3 and Lerner *et al.* 2013, 18–9):

$$p(n_{ij}(t)) = \frac{\lambda_{ij}(t)^{n_{ij}(t)} \exp\left(-\lambda_{ij}(t)\right)}{n_{ij}(t)!}. \tag{1}$$

The probability density of the event sequence multiplies the likelihood over all dyads and all time intervals t_1 to t_N.

$$f_\lambda\left(E, \theta^\lambda\right) = \prod_{t=t_1}^{t_N} \left(\prod_{ij \in D_{act}(t)} \frac{\lambda_{ij}(t)^{n_{ij}(t)}}{n_{ij}(t)!} \right) \exp\left(- \sum_{ij \in D} \lambda_{ij}(t) \right), \tag{2}$$

where $D_{act}(t)$ represents all dyads in which at least one event occurred over the entire event sequence and D represents all possible events that could have potentially occurred (Lerner *et al.* 2013, 18–9). For a more detailed derivation and specification of the probability density function, see Lerner *et al.* (2013, 14–9) or Butts (2008, 161–3).

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

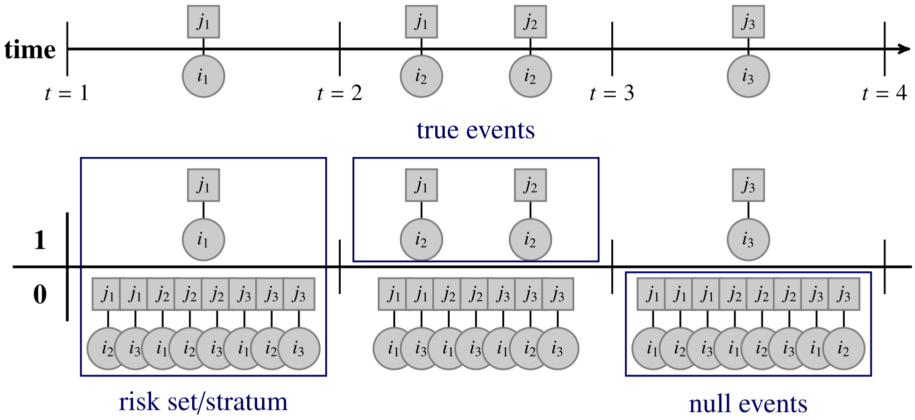

Figure 1. Stratified Cox regression set up of a discrete-time event sequence. For each event in the event sequence a risk set of all events that could possibly occur at said time t is defined which forms a stratum. The risk set consists of true events, i.e., events that took place at time t, and null events, i.e., events that did not occur at time t. Examples of risk sets, true events and null events are encircled in blue in the figure.

A stratified Cox regression can be used to model the effects endogenous or exogenous variables have on the hazard rate. Figure 1 illustrates the basic idea behind a stratified Cox regression with constant event times. For each distinct point in time the risk set D is formed, consisting of events that occurred at said time point (true events) and event that did not occur (null events).

The main distinction of REMs to conventional survival models is the endogenous network statistics that are used as independent variables to explain (together with other exogenous factors) event occurrence. These endogenous statistics are calculated as time-varying covariates and capture network patterns that are assumed to expedite event occurrence, i.e., they capture how nodes react to changes in their surrounding network.

Endogenous network statistics are calculated for each true and null event over all past events $E = (e_1, e_2, \dots, e_n)$:

$$G_t = G_t(E) = (A, B, w_t). \tag{3}$$

A weight function w_t can be applied to each past event in order to account for memory loss or the general passage of time (Brandes, Lerner, and Snijders 2009; Lerner *et al.* 2013). For instance, the weight function can use an exponential decay so that more recent events are given more weight than more distant events:

$$w_t(i,j) = \sum_{\substack{e:a_e=i,b_e=j,\\ t_e<=t}} |w_e| e^{-(t-t_e)\left(\frac{\ln(2)}{T_{1/2}}\right)} \frac{\ln(2)}{T_{1/2}}. \tag{4}$$

The weight function w_t sums over all past events e, i.e., events that occurred before the current time t and that consist of a sender i (with a_e denoting this specific sender involved in an event e; $i = a_e \in A$), a target j ($j = b_e \in B$) and occurred at time t_e; w_e can be included if events are not considered of equal importance and weighted accordingly. Each past event is exponentially weighted by how long ago the event took place ($t - t_e$). The half-life parameter $T_{1/2}$ can be used to adjust the down-weighting of the event, with a larger half-life indicating slower rate of decay and therefore giving less weight to the passage of time (Lerner *et al.* 2013).

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

The weight function w_t is used in the calculation of endogenous network statistics I, such as sender activity for instance:

$$\mathtt{I}_{\mathtt{senderActivity}}(G_t, a, b) = \sum_{j \in B} w_t(a, j). \tag{5}$$

The statistic captures to what extent actor a ties to b at time t because a has been active in the (recent) past on different target nodes j.

Different network configurations, such as triadic or four-cycle closure, homophily effects or endogenous similarities, can be translated into the REM framework and adapted to heed the timing of events.

However, model specification can pose a challenge for two distinct reasons. First, operationalization of theory-based mechanisms into endogenous network statistics is not always straightforward (see for instance Leifeld and Brandenberger 2019) and is further complicated by the addition of the temporal dependencies, which add another layer of complexity. Second, if these operationalization problems can be overcome, it is difficult to know from the estimated models themselves, whether network dependencies have been captured well by the endogenous network statistics included in the model. Butts (2008) stresses the importance of capturing *all* network dependencies among the events in the event sequence in the form of endogenous network statistics to receive unbiased estimates. If some endogenous network statistics are neglected, other statistics may be over- or underestimated and lead to faulty interpretations of results.

One way to overcome these challenges of model specification is through appropriate goodness of fit tests. The next section addresses shortcomings of conventional goodness of fit statistics and proposes an alternative way of assessing fit via simulated event sequences. These goodness of fit tests from simulated sequences can be used to assess whether the specified model terms capture the network characteristics adequately. And more importantly, these goodness of fit tests can be used to figure out which network dependencies are currently not modeled sufficiently so that they can be included in future models and prevent biased estimates.

2.2 REMs and goodness of fit statistics

REMs only produce unbiased estimates if they fully capture the endogenous network formation processes (Butts 2008, 168). In other words, if a dynamic network is theorized to include some form of network closure, an appropriate model term capturing this closure needs to be controlled for to ensure other model terms are not over- or underestimated. To fulfill this requirement for complete capture of endogenous processes in the event sequence, goodness of fit statistics are helpful in guiding model term selection and model specification.

One way of examining model fit is to calculate receiver operating characteristic (ROC) curves or PR curves (Davis and Goadrich 2006). While these tools are helpful for evaluating cross-sectional network analyses or dynamic network analyses using snapshots of networks over time (Leifeld, Cranmer, and Desmarais 2018), they do not fare well with REMs. The problem lies in the calculation of the precision, or positive predictive value. Precision is defined as the fraction of selected events from the risk set that actually occurred in the specified stratum (i.e., true positives divided by the sum of true positives and false positives). Since each stratum only represents one micro-step in the dynamic network, the precision value may be too restrictive. Take for instance a dynamic model where sender activity is one of the dominant traits. At time $t = 1$, sender i_1 engages in target j_1. At time $t = 2$, sender i_1 then engages in target j_2. At time $t = 3$, sender i_1 then engage in target j_3. The precision value calculates the power of the model by checking how many events can be correctly identified per stratum, i.e., per unit of time. Since both events are relatively similar and

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

Table 1. Terminology for simulated and original events.

	Risk set	
	True event	Null event
Original event sequence	Original true event	Original null event
Simulated event sequence	Simulated true event	Simulated null event

close together in time, the precision value can be low because its calculation does not allow for small temporal errors, i.e., selecting (i_1, j_3) at time $t = 2$ instead of (i_1, j_2).

Furthermore, PR and ROC curves do not give any indication on whether network dependencies have been captured in a satisfactory way. This holds true for other forms of fit parameters, such as the Bayesian information criterion (BIC). Both goodness of fit statistics can give no information on whether endogenous patterns in the data are sufficiently controlled for. Additional goodness of fit statistics for REMs are necessary to check network dependencies, the timing of events as well as predictive power of the model.

2.3 Predicting subsequent events

One way of examining model fit is to create artificial event sequences based on a specified REM. The approach mirrors the goodness of fit tests used in Exponential Random Graph Models (and temporal extensions thereof; see for instance Hunter, Goodreau, and Handcock 2008 or Hunter *et al.* 2008) in the sense that new event sequences are simulated and compared to the original event sequence. Comparisons of the timing of events as well as whether or not other network dependencies have been captured well can be tested using such simulated sequences. For the sake of clarity, Table 1 disentangles nomenclature for true and null events for the original and the simulated sequences. *Simulated true events* refer to newly predicted events in the simulation and *original true events* refer to events in the original data set.

```
input  : Network of past events $G_t(E)$
output: New event sequence $E_{sim}$ starting at time $t_i$ until $t_s$ (where $s$ denotes the
        number of simulated strata)
for $i \leftarrow 1$ to $s$ do
    define the risk set $D_{t_i}$ for the stratum at $t_i$;
    calculate endogenous network statistics $I$ based on $G_t(E)$;
    determine the number of events $d_i$ (fixed for all $i$ or dynamic);
    calculate baseline hazard $\hat{h}_0(t_i)$ for $D_{t_i}$ using Equation (6);
    calculate survivor probability $\hat{S}_i(t_i)$ for every event in $D_{t_i}$ using Equation (7);
    sample $k$ events from $1 - \hat{S}_i(t_i)$ to determine simulated true events in the new
      stratum $D_{t_i}$;
    while $k > d_i$ do
        randomly select one event in $E_k$ and toggle it (set to null event);
    end
    append new stratum $D_{t_i}$ to $G_t(E)$;
end
```

Algorithm 1: Procedure for predicting relational events

A simple procedure is proposed here to simulate new relational event sequences (see Algorithm 1) and each step is explained in more details below. The proposed procedure can be used for continuous-time (or exact-time) event sequences as well as for ordinal-time event sequences.

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

Step 1: Prediction setup. Relational events can be predicted either as within-sample or out-of-sample predictions. For within-sample predictions, a REM is fitted over the full range of the event sequence. For the prediction the original event sequence is then cut at a distinct stratum and subsequent events are simulated from there on and are later compared to the subsequent original event sequence. For out-of-sample predictions, a REM is fitted over a portion of the event sequence and new predictions are appended to the fitted portion of the data. The simulated sequences are later compared to the excluded sample.

Step 2: Defining the risk set for the new stratum and calculating endogenous network statistics. In order to predict next events, new strata have to be built. The stratum represents all possible events that can occur at a given point in time. Depending on whether or not the event sequence presents as time-dependent, the strata can be either static (i.e., the same for each prediction round) or dynamic (changing for each prediction round). A dynamic risk set or strata is necessary for event sequences where events can only occur at specific points in time or where events can only occur once over the entire event sequence. For instance, a member of Congress a can only sign their support for a bill b once. This action cannot be repeated and should therefore not be included in the risk set in the following stratum if the event occurred in the present stratum (for an overview over dynamic and static risk set definitions, see Brandenberger 2018b).

After the new stratum is defined, endogenous network statistics (and time-varying covariates) need to be calculated. These endogenous statistics depend on all past events, so whenever new events in a stratum are predicted, these simulated true events (i.e., the newly predicted events in the simulation, see Table 1) will subsequently affect all other events as they become part of the network of past events G_t.

Step 3: Defining the number of events. In order to calculate survivor probabilities, the number of events d_i (also known as the number of deaths in survival analysis terms) has to be defined. The overall mean number of events per stratum can be used to define d_i or a moving average can be defined to allow for more dynamic event occurrence. It is important to note that over- or underestimating the number of events does not affect the calculation of the survivor probability much (see Table 2 in Section 3).

Step 4: Calculating the baseline hazard for the strata.

The baseline hazard in a Cox model is defined as

$$\hat{h}_0(t_i) = \frac{d_i}{\sum_{j,t_j \geq t_i} \exp(x_j \hat{\beta})}, \tag{6}$$

where d_i is the number of events that occur at t_i for sender i, x_j is the covariate vector and β the estimated coefficients of the model. The baseline hazard is constant for all events in a given stratum and is therefore not relevant for the estimation of the coefficients. It can be used, however, to calculate survivor probabilities (Cox and Oakes 1984, 107).

Step 5: Calculating the survivor probability for each event in the stratum. The probability of an event surviving to the next point in time is defined as

$$\hat{S}_i(t) = \hat{S}_0(t)^{\exp(x_i \hat{\beta})}, \tag{7}$$

where $\hat{S}_0(t) = \exp(-(\sum_{j,t_j \leq t} \hat{h}_0(t_j)))$.

This survivor probability reflects the probability that an event does not occur at time t. It is calculated for each event in the stratum and determines whether or not an event is likely to occur or not.

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

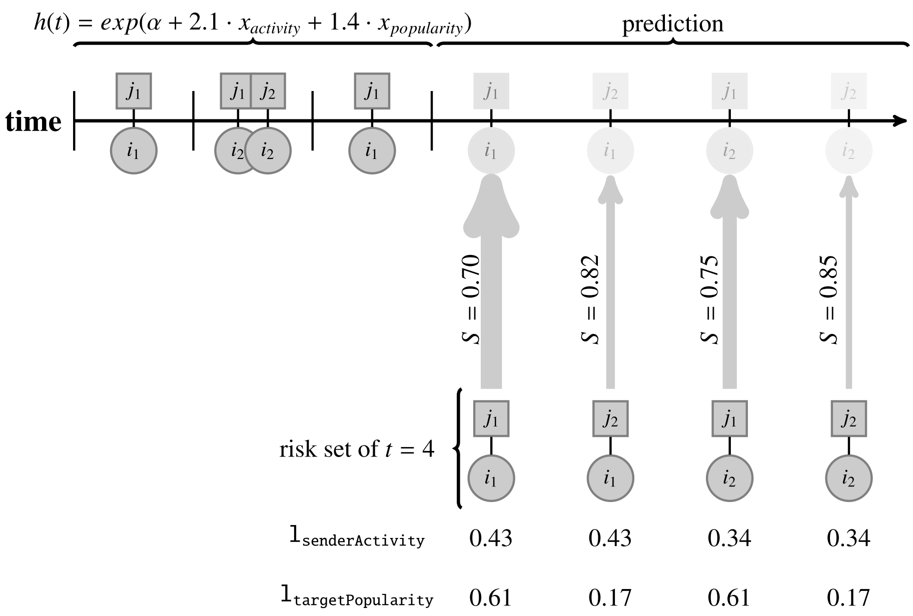
$h(t) = exp(\alpha + 2.1 \cdot x_{activity} + 1.4 \cdot x_{popularity})$
prediction
time
j_1
i_1
j_1 j_2
i_2 i_2
j_1
i_1
j_1 j_2 j_1 j_2
i_1 i_1 i_2 i_2
$S = 0.70$
$S = 0.82$
$S = 0.75$
$S = 0.85$
risk set of $t = 4$
$1_{\texttt{senderActivity}}$ 0.43 0.43 0.34 0.34
$1_{\texttt{targetPopularity}}$ 0.61 0.17 0.61 0.17

Figure 2. Illustration of event prediction at $t = 4$. The risk set for $t = 4$ is specified. For each event in the risk set the endogenous network statistics are calculated. Using the specified model, the baseline hazard (with $d_i = 1$) and survivor probabilities are calculated. The event with the lowest survivor probability is most likely to occur next (arrow width and event opacity).

Figure 2 illustrates prediction steps 4, 5, and 6 with a simple example sequence. Assume you have an event sequence with three event days and would like to predict an upcoming event at $t = 4$ (for instance to check out-of-sample prediction by holing back the stratum at t_4 and comparing it to the simulated sequence later). First, the risk set has to be defined for $t = 4$. Here, events can be repeated, so the risk set is defined as broad as possible, with *senders* × *targets*. Next endogenous network statistics are calculated for all events in the new strata (with half-life parameter $T_{1/2} = 1$). The baseline hazard is calculated using Equation (6). In the example in Figure 2 the baseline hazard with $d_i = 1$ is $\hat{h}_0(t_4) = 0.061$ (see Supplementary Information (SI) Online for details on the calculations of this example). The survivor probability can then be calculated for each event in the stratum. The model ($t = 1$ to $t = 3$) is specified to include sender activity and target popularity. The relational event (i_1, j_1) has a high `senderActivity` and a high `targetPopularity`. In a model perpetrated positively by sender activity and target popularity, this event is chosen to be the most likely to occur next. Event (i_1, j_1) has to lowest survivor probability $\hat{S}_{(i_1,j_1)}(t_4) = 0.7$ (i.e., a 30% probability of occurring at $t = 4$).

Step 6: Predict new events. New events can be selected by sampling from the stratum and weighting events with their probability to survive to the next time unit. To keep the number of events in check, the sum of all true events should be monitored (see lines 9–11 in Algorithm 1). Since the strata are added to the event sequence, allowing for more than the specified number of events d_i can cause the prediction of the next stratum to spin out of control.

Step 7: Append stratum to the event sequence. At the end of each prediction round, the stratum with the newly defined simulated true events is appended to the event sequence and a new prediction cycle starts again. As such, newly simulated events are used to predict subsequent events and so forth. The appended stratum is incorporated into the network of past events G_t for the subsequent strata. This ensures that the sequence evolves beyond simple one-shot predictions. Instead, dynamic sequences are simulated.

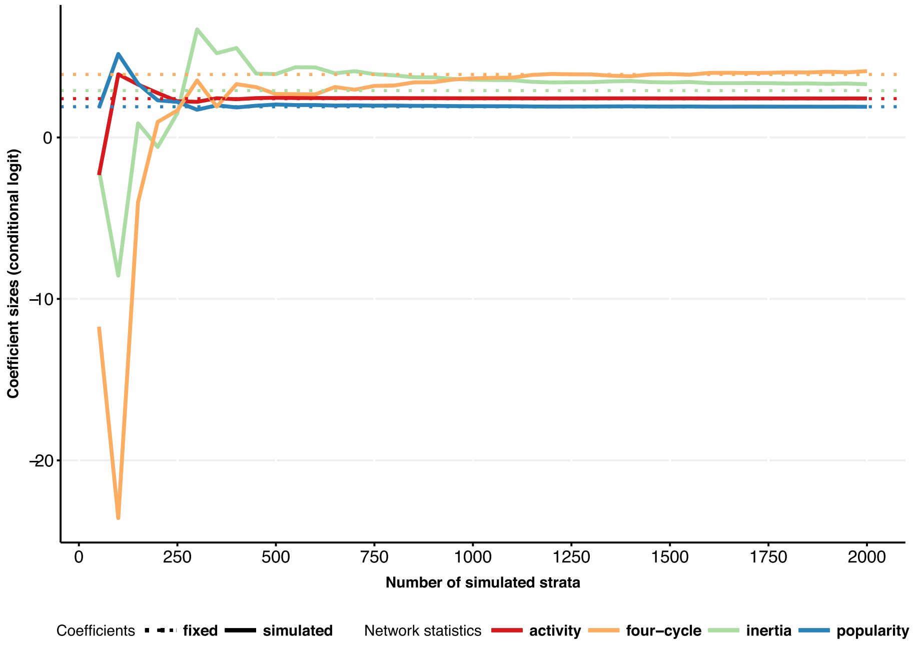

Figure 3. Coefficients from simulated sequences (bold lines) compared to the specified coefficients (dashed lines). The first 50 strata contain 200 random true events. All subsequent events are based on the prediction procedure and coefficients from an artificial Cox model.

2.4 Predicting subsequent events from random events

A simulation test is used to check whether the presented procedure for simulating event sequences is feasible. Given a network of 20 unique sender nodes and 30 unique target nodes, a random event sequence is generated using 50 event days. The resultant 30,000 events $(20 \cdot 30 \cdot 50)$ were randomly sampled into true events (i.e., deaths, d_i) and null events. 200 true events are assigned.

From this baseline, new events were simulated based on an artificially specified REM. The model included inertia, activity, popularity and four-cycle effects. The coefficients for these terms were chosen exogenously. The simulation is run over 1950 strata (resulting in 2000 strata in total). Afterward, conditional logistic regressions were run over different portions of the simulated event sequence to determine whether the coefficients from the simulated sequence reflect the specified coefficients (see SI Online for additional information).

Figure 3 indicates that after around 500 strata, the models start to stabilize around the specified coefficients. The simulation test reveals that the prediction procedure can replicate event sequences based on a specified REM.

3 Illustrative Application: Predicting Statements in a Policy Debate

The following section provides a demonstration of how the prediction procedure can be used to assess goodness of fit of REMs. Predictions are based on two stratified Cox models and compared to each other as well as to random sequences. Different goodness of fit tests are presented to evaluate the simulated sequences.[1]

3.1 Data and models

The example is based on data from a policy debate. Each event represents a statement a particular organization (sender mode) made and got recorded in the press on how best to resolve the financing problem of the German pension system (Leifeld 2016). These proposed policy solutions

[1] All replication materials are available at the *Political Analysis* Dataverse site (Brandenberger 2018a).

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

Table 2. Correlation matrix of survivor probabilities with different definitions of number of events (d_i). Survivor probabilities are calculated for the stratum at $t = 989$. The choice of d_i does not affect survivor probabilities much. Correlations of survivor probabilities with different baseline hazards are all close to 1. Please refer to the SI Online for further details on the calculation of the correlation matrix.

Nevents	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**
10	1								
9	0.9998	1							
8	0.9992	0.9998	1						
7	0.9982	0.9991	0.9998	1					
6	0.9966	0.9980	0.9991	0.9998	1				
5	0.9945	0.9963	0.9979	0.9990	0.9997	1			
4	0.9917	0.9940	0.9960	0.9978	0.9989	0.9997	1		
3	0.9884	0.9911	0.9936	0.9957	0.9975	0.9989	0.9998	1	
2	0.9843	0.9875	0.9905	0.9932	0.9955	0.9974	0.9988	0.9997	1
1	0.9797	0.9833	0.9868	0.9899	0.9928	0.9953	0.9973	0.9988	0.9997

were hand-coded and represent the target mode. The events in the sequence are weighted by a stance dummy, representing agreement for or opposition with a proposed policy solution. The case study as well as the REM are presented in full in Leifeld and Brandenberger (2019).

The data presents an ideal example of the use of REMs to examine micro-dynamics in a social system. Exact- or ordinal-time event sequences offer rich insights into how nodes react to changes in their network. However, they also pose a challenge as theoretical constructs have to be operationalized at the micro-level for fine-grained mechanisms have to be defined to test them. In their paper, Leifeld and Brandenberger (2019) show that REMs can be used to test endogenous coalition formation through policy learning mechanisms. Several policy process theories address the importance of policy learning for coalition formation (see for instance Sabatier and Jenkins-Smith 1993); however, none of them offer concrete insights into how learning takes place (i.e., the mechanism behind it). Leifeld and Brandenberger (2019) operationalize learning as closing four-cycles, where policy solutions are gradually picked up by a focal organization if other nodes in the network, with whom the focal organization has shared policy solutions in the past, postulate them. Similarly, organizations learn from their opponents by repeatedly rejecting solutions they advocate. These two endogenous learning effects are tested using the German pension financing debate as a case study.

A full model and a reduced model are presented here from which within-sample predictions are made. Figure 4 reports the results of both models. The full model contains four-cycle statistics that represent learning mechanisms in policy debates (for additional details, please refer to Leifeld and Brandenberger 2019). The reduced model only contains control variables with inertia, sender activity and target popularity as endogenous network statistics.

Even though the four-cycle learning effects show positive effects in the REM (see Figure 4), it is unclear by how much they improve the predictive fit of the model or if they are even necessary to capture the endogenous process in the debate sequence. For theory-building purposes, the proposed goodness of fit tests can be used to assess the impact of the learning mechanisms in explaining the temporal and network structure of the debate.

The data contains 6,704 statements, made over the course of nine years (1993–2001) and were manually coded from 1,842 newspaper articles using tools from discourse network analysis (Leifeld 2017). The sender mode contains of 245 organizations and the target node contains 69 policy solutions (for additional information on the data, see Leifeld 2016).

In order to test the fit of the two models, within-sample predictions were made and compared to the original sequence. 1,000 distinct event sequences were simulated starting at stratum

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

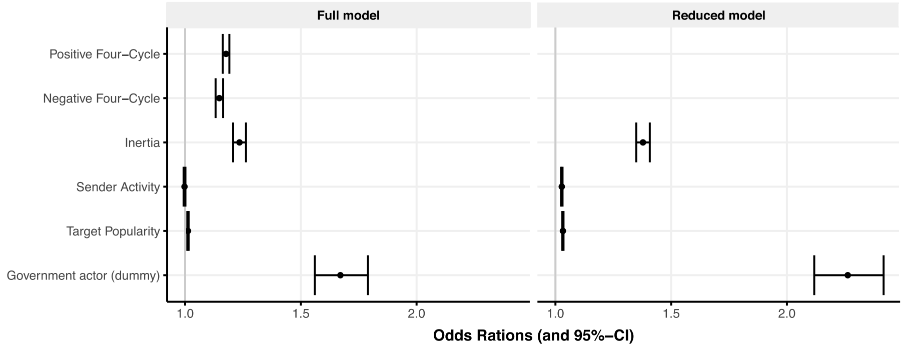

Figure 4. Results of the stratified Cox model on event occurrence. Events correspond to political statements organization made during the German pension debate. Endogenous network statistics are used to capture temporal patterns of interactions in the debate. The full model contains four-cycle statistics to capture learning patterns among organizations. (*Note*: AIC full model = 54219.64, AIC reduced model = 55452.81. Mc-Fadden $pseudo - R^2$ is 0.13 and 0.11 for the full model and the reduced model, respectively. Estimations based on 6044 true events and 877,811 null events.)

$t = 700$ (representing event date March 26, 1999) and ending at stratum $t = 800$ (representing event date September 29, 1999).[2] The risk set for the prediction was defined by all sender–target combinations that were voiced throughout the entire debate $D_{act}(t)$, resulting in a strata containing 1916 events from which next events were selected based on their survival probability. The number of events d_i were set to $d_i = 8$ based on the average number of events occurring between $t = 700$ and $t = 800$ (mean number of events per stratum = 7.8). The choice of d_i barely affects the calculation of the baseline hazard and by extension the survivor probabilities. Table 2 holds the correlations between survivor probabilities of events in a stratum based on different numbers of events d_i. All correlations are close to 1 (see SI Online for additional information). 1,000 event sequences were simulated for both the full model and the reduced model. Additionally 1,000 event sequences were generated based on a random selection of d_i events per stratum.

3.2 Temporal prediction error

In a first goodness of fit test the timing of simulated events is examined. If the model has predictive power, the simulated true events should be closer to the original true events than the simulated null events. This would indicate that the simulation chooses events to occur next that are more closely related to the temporal position of the original true events.

Figure 5 shows density plots of the time between simulated (null and true) events and original true events, i.e., the temporal prediction error. Comparing the distributions and overall means of the temporal prediction error for simulated true events between the two models reveals by how much the model improves if endogenous four-cycle statistics are included.

Figure 6 reports the temporal prediction error for simulated true events only (i.e., excluding null events). The mean temporal prediction error for the full model is 202.25 days. For the reduced model, the error is 224.81. Comparing the two distributions and means to the temporal prediction error of fully random sequences allows an assessment of how the models perform compared to a random baseline. The mean temporal prediction error of the random sequence is 242.78 days and reflects the temporal prediction error of the simulated null events. All in all, the calculations show that four-cycle statistics improve the model drastically.

[2] Please refer to the SI Online for additional information on how the number of simulations affect the prediction results.

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

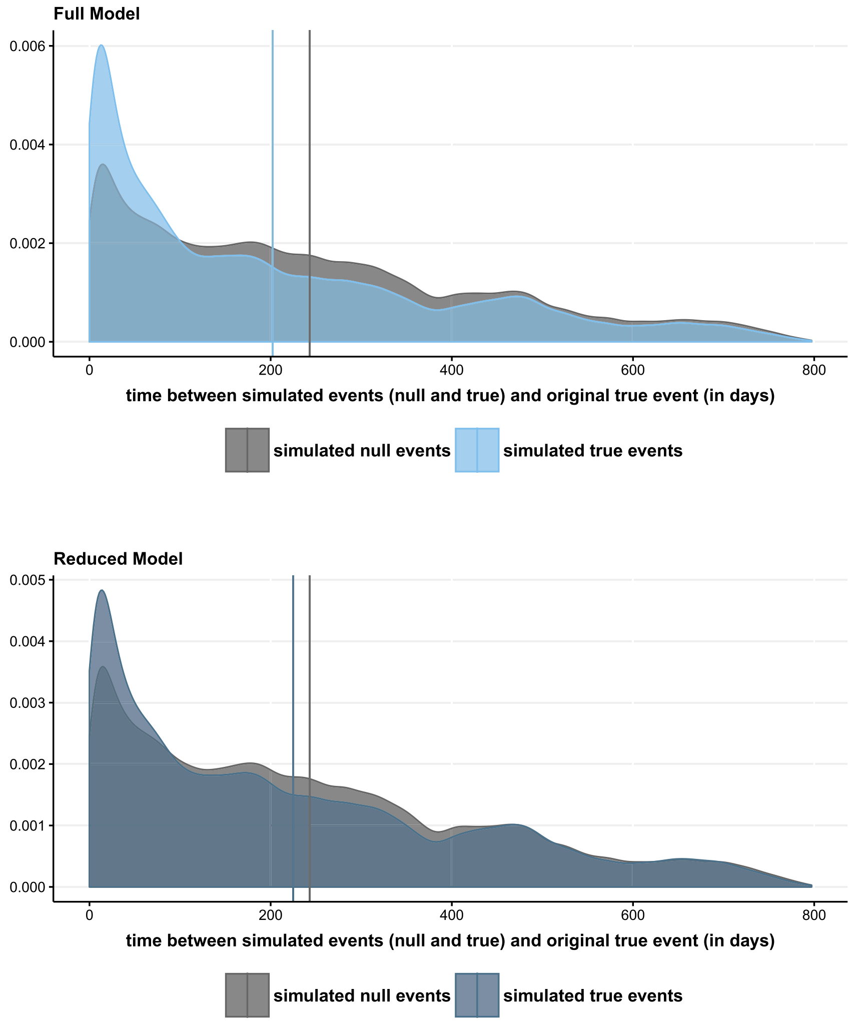

Figure 5. Goodness of fit assessment of the full model through timing of simulated events. The temporal prediction error is calculated by the difference between the original true event and the simulated (true and null) event. For the full model, the mean temporal prediction error for simulated true events is 202.25 days and 242.85 days for simulated null events. For the reduced model, the mean temporal prediction error for simulated true events is 224.81 days and 242.89 days for simulated null events.

3.3 Precision of event prediction with tolerance

In a second goodness of fit test, the accuracy of simulated events is examined. For each simulated stratum, the simulated true events are compared to the original true events to check if the original true events have been predicted. Matching senders, targets or sender–target combinations can be evaluated. To ensure that the predictions are not judged too restrictively, as is done in the precision measure, a tolerance can be introduced. The tolerance can check whether simulated true events are located within a particular time span.

Figure 7 reports the results of the event matching test. Next to exact matching, correct predictions within a tolerance of 5 days and 10 days are reported as well. The comparison to the random sequence is useful if the risk set is made up of conditional events, i.e., events that occurred at least once over the sequence instead of all possible sender–target combinations. The random sequence provides a baseline against which the models can be checked. For instance, the random sequence predicts 10% of all its sender nodes (exact timing). The reduced model increases

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

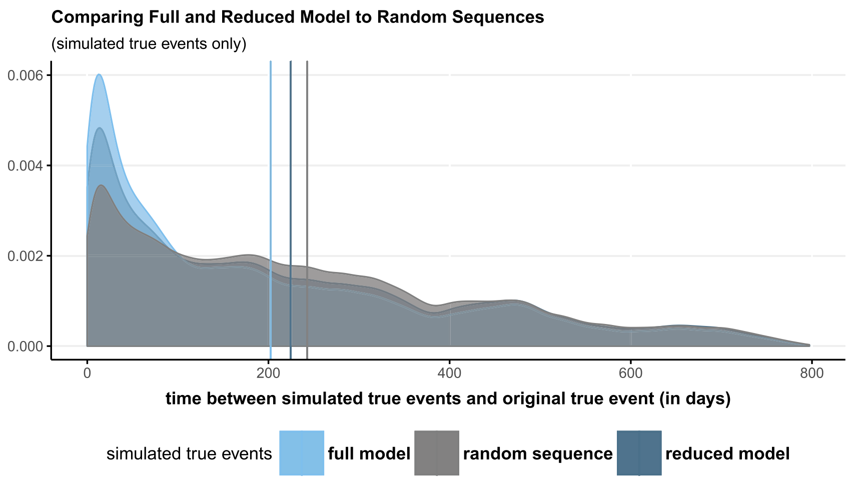

Figure 6. Temporal prediction error of simulated true events for the full and reduced models as well as for random sequences. The temporal prediction error is calculated by the difference between the original true event and the simulated true event. Mean temporal prediction error of the full model is 202.25 days, 224.81 days for the reduced model and 242.78 days for the random sequence.

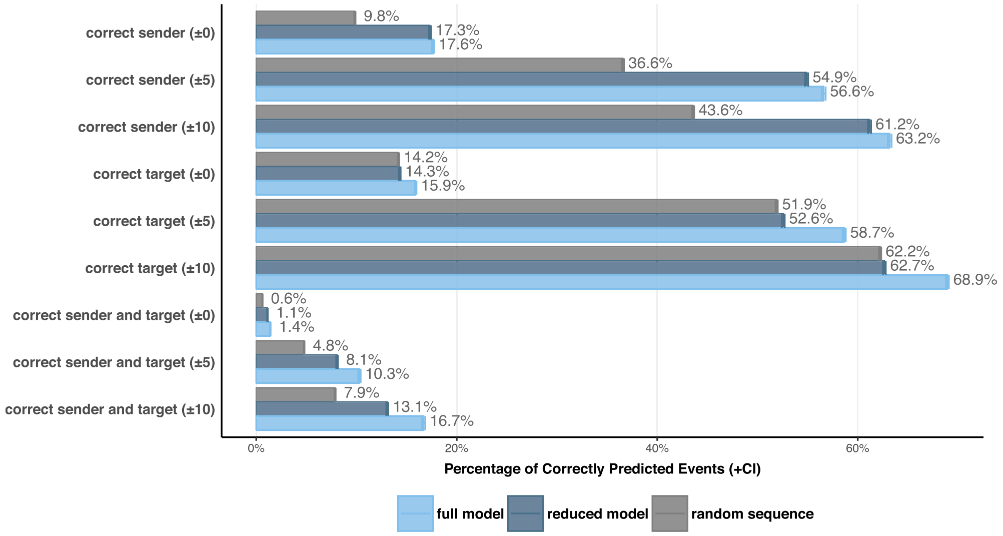

Figure 7. Goodness of fit assessment through event matching. The bar chart shows by how much the reduced and full models improve the prediction of senders, targets and sender–target combinations vis-à-vis the randomly predicted sequence. Confidence intervals are based on 1,000 simulated sequences.

the matching outcomes to 17.3%. Including four-cycles, however, does not improve prediction of sender nodes much compared to the reduced model (17.6%). If a tolerance of 5 event days is used to evaluate matches, the full model improves slightly over the reduced model. Over 50% of all sender nodes are predicted correctly with the reduced and the full model with a tolerance of 5 event days.

Looking at the prediction of target nodes shows that the inclusion of the four-cycle statistics slightly increases the correct prediction of target nodes. Even though the full model reaches almost 70% correct target nodes (with a tolerance of 10 event days), it is important to note that the random sequence is able to get 62% of all targets with the same tolerance. A comparison to random sequences is helpful for event sequences where events repeat themselves.

Most importantly, the specified REM should be able to predict both sender and target nodes together, i.e., the full relational event. Prediction success is low for both the reduced and the

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

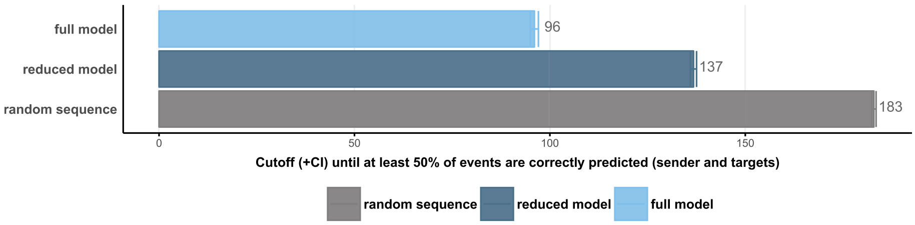

Figure 8. Goodness of fit assessment through event matching. The bar chart shows how large the tolerance has to get until the model reaches 50% correct event (sender–target combinations) predictions. The lower the tolerance, the better the model fit. Confidence intervals are based on 1,000 simulated sequences.

full model (compared to the random sequence) if exact matches are examined. Increasing the tolerance by 5 days increases prediction by a factor of seven (for the full model). Finding 10% of all events within a 10 day span (plus–minus 5 days) is a remarkable feat, since the risk set contains 1,916 unique events and a total of eight events at the most are chosen for each simulated stratum.

Alternatively, event matching can be used to examine how wide the tolerance has to be set to reach 50% correct prediction of sender–target events (see Figure 8). This is useful to quickly compare different model specifications, as well as models using alternative risk set definitions or half-life parameters in the time-weighting of past events.

3.4 Checking endogenous network structures

Whilst temporal prediction errors and event matching are useful for comparisons of different model specifications and establishing which model terms should be included in the model, neither can establish if network dependencies have been captured fully, as needed for the REM to estimate unbiased coefficients.

Figure 9 shows network characteristics captured by the simulated sequences compared to the original sequence. The original event sequence and 1,000 simulated sequences are aggregated into a cross-sectional network snapshot (using all events between event days 700 and 800). As is done with other network models when checking fit (Robins, Pattison, and Woolcock 2005; Hunter *et al.* 2008; Cranmer *et al.* 2017), different network characteristics are calculated based on these snapshots, such as degree distributions (for both modes separately). If the REM is able to capture endogenous processes well, then the simulated sequences should reveal similar network characteristics as the original sequence does. In Figure 9, the degree distributions (for the sender and target mode) of the original event sequence are depicted in a bar chart. Degree distributions are then calculated for the 1,000 simulated sequences and plotted as box plots (to show the variation between the different simulations) on top of the bar chart. Comparing the original and simulated degree distributions shows that the fit is not perfect; yet both models show some congruence with the original degree distributions (with the full model performing slightly better than the reduced model; see left two panels in the top row of Figure 9). Adding the four-cycle statistics helped align the degree distributions of both modes. The k-star distribution of the sender mode is slightly overestimated by the REM, though less so in the full model. The k-star distribution of the target mode as well as dyad-wise shared partner distributions and geodesic distances are captured well in the full model and slightly less so in the reduced model. All in all, the model-based simulations are able to capture network properties fairly well.

4 Discussion

Relational event models are a powerful tool for longitudinal network inference. These models are able to model temporal patterns from theoretical concepts and test them on sequences of relational events. However, to ensure the estimated models do not suffer from biases, model fit has to be assessed thoroughly. Without goodness of fit assessment, the interpretation of the results

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

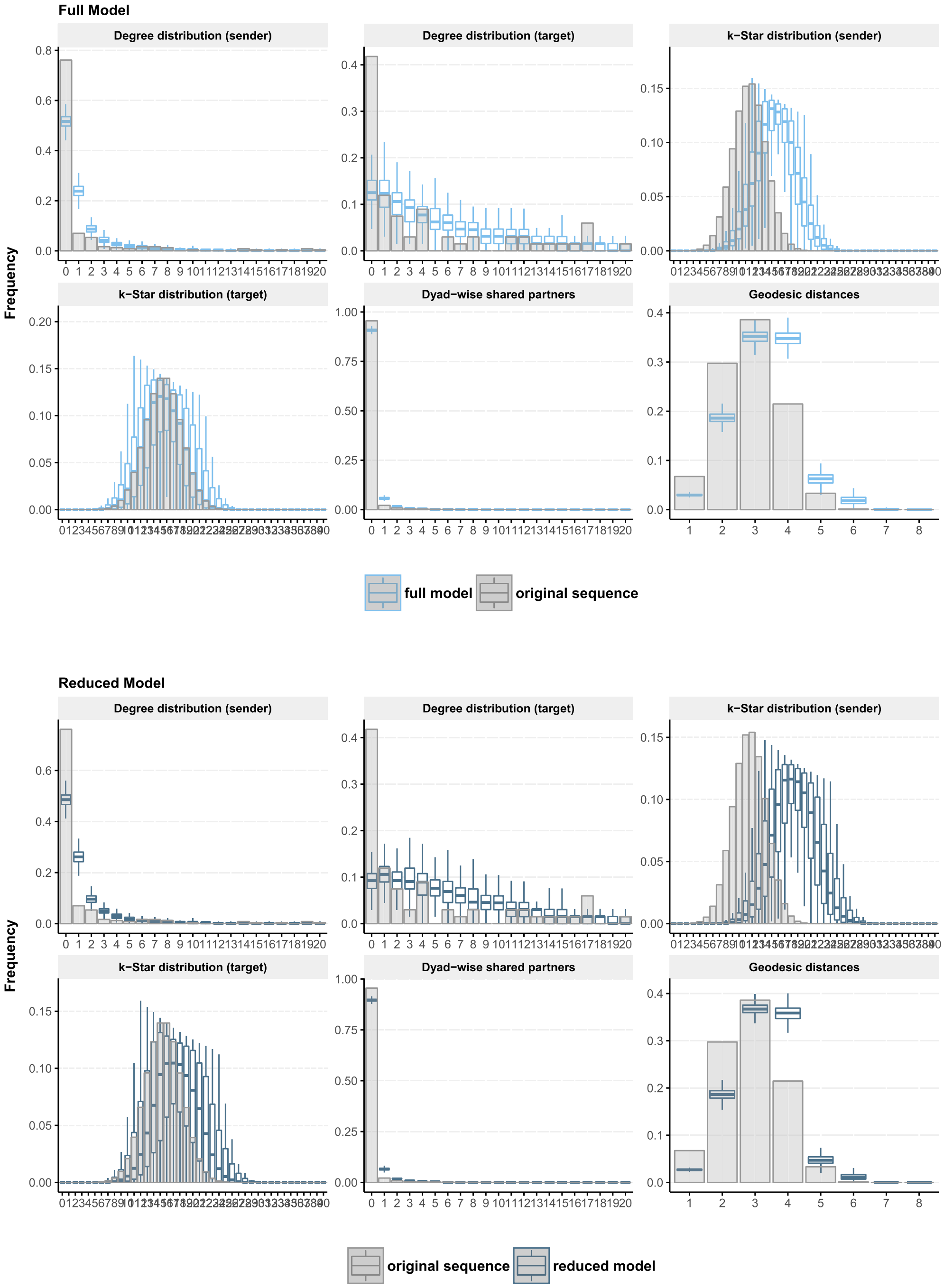

Figure 9. Goodness of fit assessment through comparison of network characteristics. The bar chart represents the distributions from the aggregated network of events between event day 700 and 800 of the original event sequence. The box plot corresponds to the distributions captured in the model-based simulations. The full model captures degree distributions fairly well, slightly overestimates k-star distribution of the sender mode and yields passable results for k-star distributions of the target mode, dyad-wise shared partner and geodesic distance distributions.

warrant little attention as they may not capture the evolving properties of the relational event sequence properly. Furthermore, by evaluating the contribution of individual model terms to the model's predictive power as well as to the improvement of capturing network dependencies, new insights can be gained on the importance of the model term. As REMs are often run on large data sets, model terms almost always exhibit some effect on event occurrence. However, by

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

closely examining the contribution of the model term to the model's fit, the term's importance can be judged more accurately. This is particularly important in cases where the REM serves an exploratory function and tries to operationalize theoretical constructs that have never been empirically tested before. By gauging the importance of these new model terms, these newly operationalized micro-dynamics can also be used for theory-build purposes (e.g., by comparing two alternative operationalizations of a theoretical construct and comparing their individual impact on model fit).

This paper proposes adapting tools from survival analysis to generate model-based simulations. A simple simulation procedure is presented that uses survivor probabilities to simulate new events. Most importantly, these simulated new events are then used to predict additional events, forming a simulation of a dynamic event sequence based on a previously set REM.

By repeating these simulations multiple times a bushel of simulated event sequences can be used to check model fit by comparing them to the original sequence. Three distinct goodness of fit tests were presented in a real world data example. The first test captures temporal prediction error and checks whether simulated true events are chosen in close temporal proximity to their original true event. The second test evaluates whether the simulated true events match the original true events, i.e., checks the precision of the model. By allowing a temporal tolerance in the calculation of the precision, the precision measure can be relaxed to accommodate the dynamic and endogenous network properties of the relational event sequences. A third goodness of fit test evaluates the network properties that the simulated event sequences entail and compares them to the properties found in the original sequences. This goodness of fit test is adapted from cross-sectional network models, where the comparison of network characteristics from the original network to simulated networks is a standard procedure in evaluating the fit of a cross-sectional network model (Hunter *et al.* 2008; Cranmer *et al.* 2017).

These model-based simulations can be used to assess parameter specifications by evaluating whether endogenous network statistics are specified sufficiently to reproduce network dependencies and structures in the original sequences. If network dependencies are not reproduced sufficiently well, these goodness of fit tests offer concrete help as to which aspect of the network dependencies are underdeveloped in the model. For instance, if the original network exhibits strong tendencies toward triadic closure and the simulated sequences cannot reproduce them, a model term measuring triadic closure should be added to the REM to improve model fit. These model-based simulations can be further used to compare different model specifications, such as alternative ways of risk set definitions or checking which half-life parameter in the calculation of endogenous network statistics best fits the data.

There are some limitations to the presented procedure that future research should address. The most important limitation is the computational burden that the simulation procedure evokes. There are four variables that influence the computational burden: first, the size of the respective risk set; second, the length of time of the simulated sequence; third, the starting position of the simulation in the original event sequence; and fourth, the complexity of the REM, i.e., the number of endogenous network statistics and their complexity. For every event in the risk set, endogenous network statistics have to be calculated. These calculations can be computationally intense if the network of past events G_t is large. While some computations of endogenous network statistics are fairly simple (e.g., inertia), others require more time (e.g., four-cycles). Future research should examine the effects of sampling on the calculation of endogenous network statistics and the REMs itself. Sampling from the risk set, for instance, could alleviate some of the computational burden in the simulation process as well.

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press

Funding

This work was supported by the Swiss National Science Foundation (grant number 149410).

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2019.10.

References

Andersen, P. K., and R. D. Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." *The Annals of Statistics* 10:1100–1120.

Brandenberger, L. 2018a. "Replication Files for: Predicting Network Events to Assess Goodness of Fit of Relational Event Models." https://doi.org/10.7910/DVN/GM5SYQ, Harvard Dataverse, V1.

Brandenberger, L. 2018b. "Trading Favors - Examining the Temporal Dynamics of Reciprocity in Congressional Collaborations Using Relational Event Models." *Social Networks* 54:238–253.

Brandes, U., J. Lerner, and T. A. Snijders. 2009. "Networks Evolving Step by Step: Statistical Analysis of Dyadic Event Data." *2009 International Conference on Advances in Social Network Analysis and Mining*, 200–205. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Butts, C. T. 2008. "A Relational Event Framework for Social Action." *Sociological Methodology* 38(1):155–200.

Cox, D. R., and D. Oakes. 1984. *Analysis of Survival Data*. London: Chapman and Hall.

Cranmer, S. J., P. Leifeld, S. D. McClurg, and M. Rolfe. 2017. "Navigating the Range of Statistical Tools for Inferential Network Analysis." *American Journal of Political Science* 61(1):237–251.

Davis, J., and M. Goadrich. 2006. "The Relationship Between Precision-Recall and Roc Curves." In *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. ACM.

Desmarais, B. A., V. G. Moscardelli, B. F. Schaffner, and M. S. Kowal. 2015. "Measuring Legislative Collaboration: The Senate Press Events Network." *Social Networks* 40:43–54.

DuBois, C., C. Butts, and P. Smyth. 2013. "Stochastic Blockmodeling of Relational Event Dynamics." In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 238–246. Available at http://proceedings.mlr.press/v31/dubois13a.pdf.

Hunter, D. R., S. M. Goodreau, and M. S. Handcock. 2008. "Goodness of Fit of Social Network models." *Journal of the American Statistical Association* 103(481):248–258.

Hunter, D. R., M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. 2008. "Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24(3):nihpa54860.

Kitts, J. A., A. Lomi, D. Mascia, F. Pallotti, and E. Quintane et al. 2016. "Investigating the Temporal Dynamics of Inter-Organizational Exchange: Patient Transfers Among Italian Hospitals." *American Journal of Sociology* 123(3):850–910.

Leifeld, P. 2016. *Policy Debates as Dynamic Networks: German Pension Politics and Privatization Discourse*. Frankfurt am Main: Campus.

Leifeld, P. 2017. "Discourse Network Analysis: Policy Debates as Dynamic Networks." In *The Oxford Handbook of Political Networks, Chapter 12*, edited by J. N. Victor, M. N. Lubell, and A. H. Montgomery, 301–326. Oxford: Oxford University Press.

Leifeld, P., and L. Brandenberger. 2019. "Endogenous Coalition Formation in Policy Debates." Preprint, arXiv:1904.05327.

Leifeld, P., S. J. Cranmer, and B. A. Desmarais. 2018. "Temporal Exponential Random Graph Models With Btergm: Estimation and Bootstrap Confidence Intervals." *Journal of Statistical Software* 83(6):1–36.

Lerner, J., M. Bussmann, T. A. Snijders, and U. Brandes. 2013. "Modeling Frequency and Type of Interaction in Event Networks." *Corvinus Journal of Sociology and Social Policy* 4(1):3–32.

Lerner, J., N. Indlekofer, B. Nick, and U. Brandes. 2013. "Conditional Independence in Dynamic Networks." *Journal of Mathematical Psychology* 57(6):275–283.

Quintane, E., G. Conaldi, M. Tonellato, and A. Lomi. 2014. "Modeling Relational Events. A Case Study on an Open Source Software Project." *Organizational Research Methods* 17(1):23–50.

Robins, G., P. Pattison, and J. Woolcock. 2005. "Small and Other Worlds: Global Network Structures From Local Processes." *American Journal of Sociology* 110(4):894–936.

Sabatier, P. A., and H. Jenkins-Smith. 1993. *Policy Change and Learning: An Advocacy Coalition Approach*. Boulder: Westview Press.

Welbers, K., and W. de Nooy. 2014. "Stylistic Accommodation on an Internet Forum as bonding: Do Posters Adapt to the Style of Their Peers?." *American Behavioral Scientist* 58(10):1361–1375.

https://doi.org/10.1017/pan.2019.10 Published online by Cambridge University Press