On the Evolution of Statistical Evidence in Dynamic Social Networks

Joris Mulder & Roger Leenders

Tilburg University, the Netherlands

Seminar at Social Networks Working Group, UC Irvine, 10/24/17

Outline

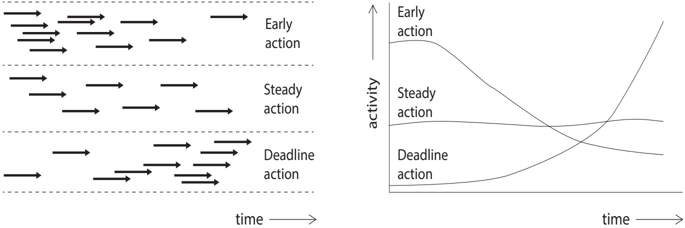
1. Dynamic social networks of colleagues in a large organization
2. Relational event model
   - Model
   - Capturing the dynamic nature using a moving window
   - Analysis I of email data: Estimation
3. Quantifying Statistical Evidence Using the Bayes Factor
   - Methodology
   - Analysis II of email data: Evolution of Statistical Evidence
4. Conclusions and furter research

Outline

Dynamic social networks

The concept of time

- The concept of time (e.g., speed, pacing, rhythm) is often not explicitly included in social theories (merely "**X causes Y**", Monge, 1991; Leenders et al., 2016).
- If there is an effect, how quickly does it have an effect? How long does it have an effect? And does the effect change linearly or nonlinearly?
- The activity of the network is generally not constant over time.



Dynamic social networks

Social network of colleagues

- How do colleagues share and seek information from each other in large organizations?
- Is this behavior of information-sharing and information-seeking dynamic in nature? Yes!
- How is the process affected by hierarchy, team membership, location?
- How is the process affected by interventions?
- How can we quantify statistical evidence between hypotheses about this continuous process?

Data

- Approx. 14,000 email messages were sent in 2010 between approx. 2,500 colleagues in a large organization.
- The emails contained information about new developments and technologies.

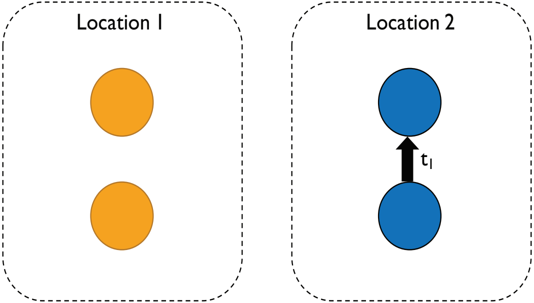
Drivers of the interaction process

Endogenous and exogenous effects

- The interaction process is driven by endogenous effects and exogenous effects.
- Examples of endogenous effects include inertia, reciprocity, transitivity.
- Examples of exogenous effects include hierarchical position, location, team membership.
- These effects are time-dependent.

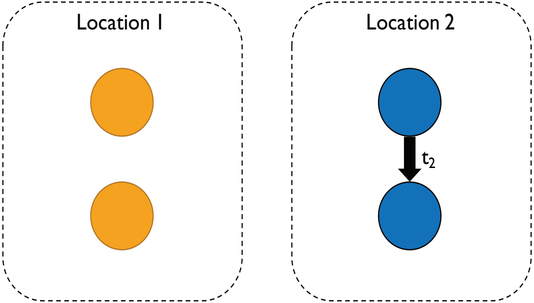
Email messages

Message at time t_1



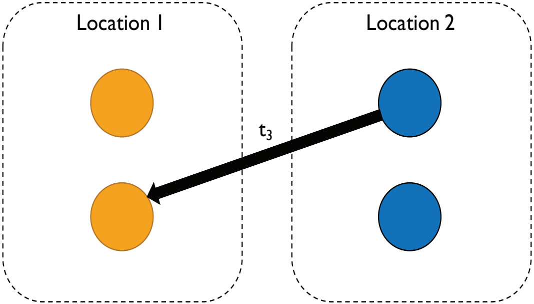
Email messages

Message at time t_2



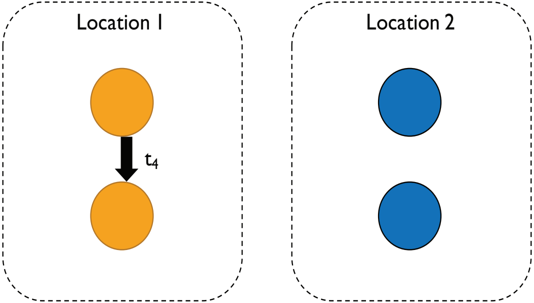
Email messages

Message at time t_3



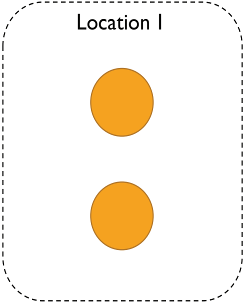
Email messages

Message at time t_4



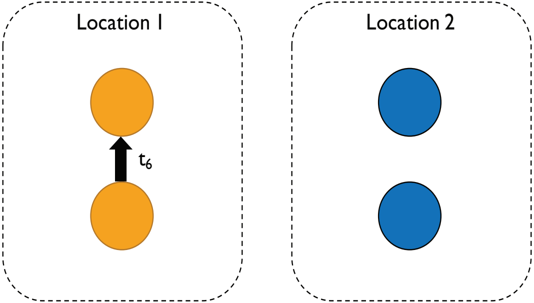
Email messages

Message at time t_5



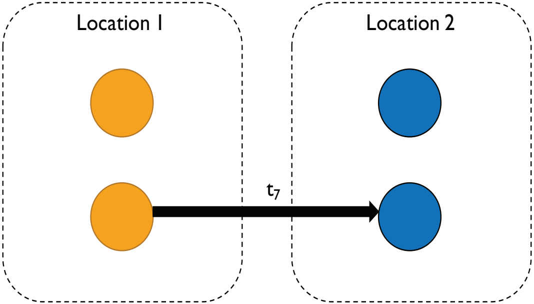
Email messages

Message at time t_6



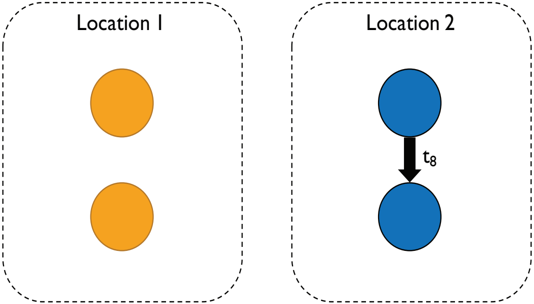
Email messages

Message at time t_7



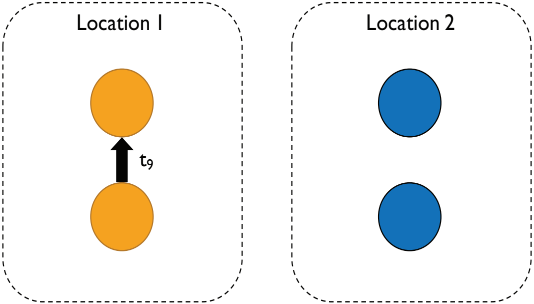
Email messages

Message at time t_8



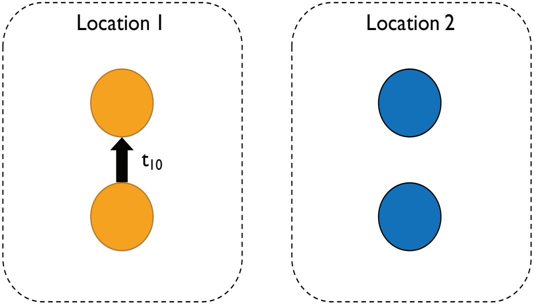
Email messages

Message at time t_9



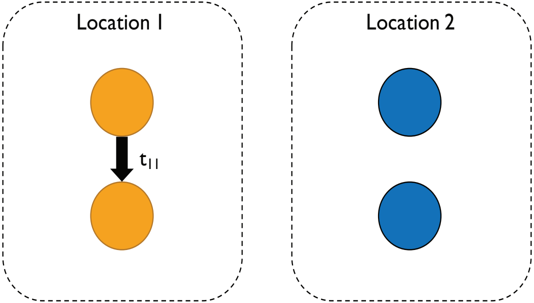
Email messages

Message at time t_{10}



Email messages

Message at time t_{11}



Outline

Relational event model

Relational events

- Event = {sender, receiver, time, weight, type, modality, ...}
- Events are recorded in an event list or 'relational event history'.

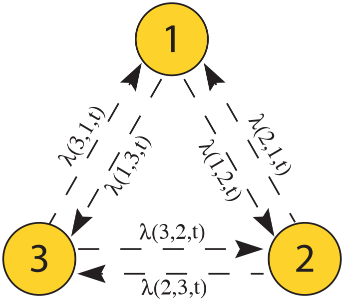
Relational event model of Butts (2008)

- Key feature: Model the timing of social interactions.
- Related to Cox's proportional hazards model.
- The time until the next relational event from sender s to receiver r, where the pair (s, r) lies in the risk set $\mathcal{R}_t$, is modeled using an exponential distribution with rate parameter $\lambda(s, r, t)$.
- The rate at time t is modeled as a log linear function of time-dependent endogenous and exogenous covariates, e.g.,

$$\lambda(s, r, t) = \exp\{x_{s,t}\beta_{inertia,t} + x_{reciprocity,t}\beta_{reciprocity,t} + x_{hierarchy,t}\beta_{hierarchy}\}$$

Relational event model

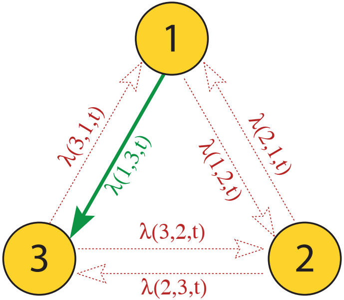
Likelihood



- Time till the next event: $t_m - t_{m-1} \sim \text{Exp}\left(\sum_{(s',r')\in\mathcal{R}_t} \lambda(s',r',t)\right)$.
- Which relational event: $P(s,r) = \frac{\lambda(s,r,t)}{\sum_{(s',r')\in\mathcal{R}_t} \lambda(s',r',t)}$.
-

Relational event model

Likelihood



- Time till the next event: $t_m - t_{m-1} \sim \text{Exp}\left(\sum_{(s',r')\in\mathcal{R}_t} \lambda(s',r',t)\right)$.
- Which relational event: $P(s,r) = \frac{\lambda(s,r,t)}{\sum_{(s',r')\in\mathcal{R}_t} \lambda(s',r',t)}$.
- Events that did not occur are right-censored (DuBois et al. 2013).

Capturing the dynamic nature

Moving window

- A moving window was used to get insights about the dynamic nature of the effects.
- The moving window was set to 60 days. This implies that only the email messages that occurred within the last 60 days we taken into account to fit the model.
- A moving window can be seen an operationalization of the memory of the process.

Capturing the dynamic nature

Moving window

- A moving window was used to get insights about the dynamic nature of the effects.
- The moving window was set to 60 days. This implies that only the email messages that occurred within the last 60 days we taken into account to fit the model.
- A moving window can be seen an operationalization of the memory of the process.

Capturing the dynamic nature

Moving window

- A moving window was used to get insights about the dynamic nature of the effects.
- The moving window was set to 60 days. This implies that only the email messages that occurred within the last 60 days we taken into account to fit the model.
- A moving window can be seen an operationalization of the memory of the process.

Capturing the dynamic nature

Moving window

- A moving window was used to get insights about the dynamic nature of the effects.
- The moving window was set to 60 days. This implies that only the email messages that occurred within the last 60 days we taken into account to fit the model.
- A moving window can be seen an operationalization of the memory of the process.

Relational event model for email data

Email data

- In the case of 2500 colleagues, the risk set consists of $2500 \times 2499 = 6,247,500$ possible relational events.
- Every relational event results in $6,247,499$ right censored events and one observed event.
- $14,000$ email messages were recorded. This makes a dataset of $6,247,500 \times 14,000 = 87,465,000,000$ events of which $14,000$ are observed and the rest is right censored.
- Hence the complete dataset that is analyzed is huge which complicates the analysis.

Relational event model for email data

Subset of email data

- I took a subset of all the events based on the 55 most active persons in the network. This resulted an relational event history of 1,505 events.
- The risk set consisted of $55 \times 54 = 2970$ possible events.
- Thus the complete data set that is analyzed consisted of $2970 \times 1,505 = 4,469,850$ events out of which 1,505 were observed and the rest were right-censored.
- The rem-package of Butts was quite slow. Analysis was done using the coxph function in the survival-package.

Relational event model for email data

Endogenous covariates

- Recency_send
  - Quantifies if recent sending activity results in future sending activity.
  - Statistic: $\frac{1}{1+\text{TimeUntilLastSendingEmail}}$
- Recency_receive
  - Quantifies if recent receiving activity results in future receiving activity.
  - Statistic: $\frac{1}{1+\text{TimeUntilLastReceivingEmail}}$
- Sender_out-degree
  - Quantifies if node was a highly active sender in the last period, this results in high future sending activity.
  - Statistic: $\log \frac{\#\text{SendingEvents}+1}{i+N}$,
    for node i and N potential events (DuBois et al., 2013).
- Sender_in-degree
  - Quantifies if node was a highly active receiver in the last period, this results in high future receiving activity.
  - Statistic: $\log \frac{\#\text{ReceivingEvents}+1}{i+N}$.

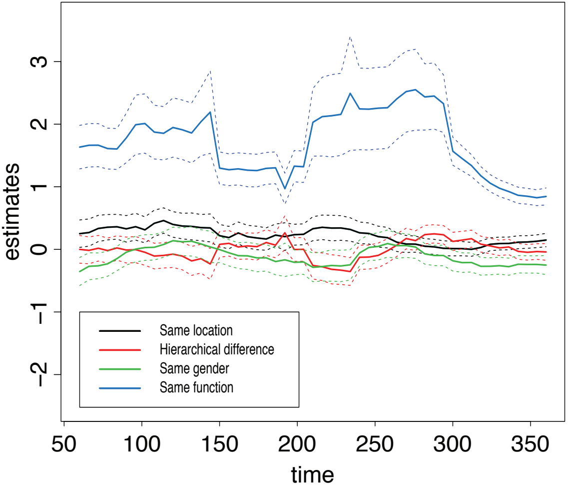
Relational event model for email data

Exogenous covariates

- Same gender (0,1)
- Same location (0,1)
- Same job description (0,1)
- Difference between hierarchical levels $(-3, -2, \ldots, 3)$.

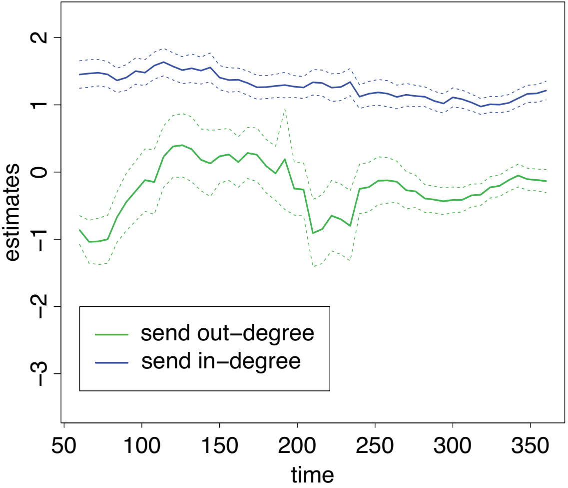
Evolution of dynamic interaction process over time

Some empirical results



Evolution of dynamic interaction process over time

Some empirical results



Outline

1. Dynamic social networks of colleagues in a large organization
2. Relational event model
   - Model
   - Capturing the dynamic nature using a moving window
   - Analysis I of email data: Estimation
3. Quantifying Statistical Evidence Using the Bayes Factor
   - Methodology
   - Analysis II of email data: Evolution of Statistical Evidence
4. Conclusions and furter research

Quantifying Statistical Evidence Using the Bayes Factor

Definition

The Bayes factor between hypothesis H_1 and H_2 is defined as the ratio of the marginal likelihoods

$$B_{12} = \frac{p_1(Data)}{p_2(Data)}.$$

The Bayes factor B_{12} quantifies the relative evidence in the data for H_1 against H_2.

Testing multiple hypotheses

Bayes factors can straightforwardly be used for testing multiple hypotheses. Furthermore, Bayes factors satisfy the transtivity property.

$$\left.\begin{array}{l} B_{12} = 10 \\ B_{23} = 5 \end{array}\right\} \Rightarrow B_{13} = B_{12} \times B_{23} = 50.$$

Quantifying Statistical Evidence Using the Bayes Factor

How to specify the priors?

In order to compute Bayes factors, priors need to be specified for the unknown parameters under each hypothesis.

Priors

Priors reflect our beliefs about the model parameters before observing the data.

Default approach

We consider the situation where prior information is weak. For this reason we want to quantify statistical evidence using default Bayes factors based on default priors where subjective prior specification is avoided.

Quantifying Statistical Evidence Using the Bayes Factor

Default Bayes factors

- In the methodology of Gu et al. (2017), only the ML estimates and the estimated covariance matrix of the estimates are needed.
- The default prior is a scaled version of the posterior (O'Hagan, 1995), such that (i) it contains minimal information and (ii) it is centered at the null value.

Multiple hypothesis test (some Greek...)

Consider the hypotheses: $H_0 : \beta = 0$, $H_1 : \beta < 0$, $H_2 : \beta > 0$, $H_u : \beta \in \mathbb{R}$:

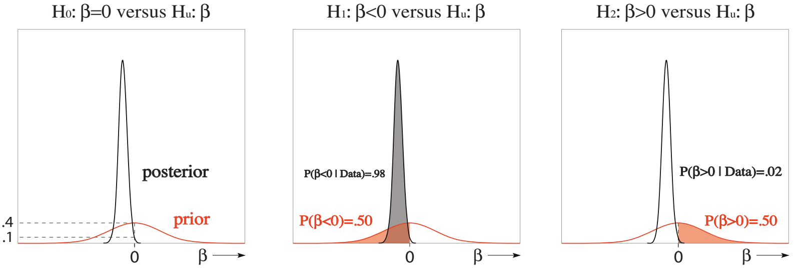
$$B_{0u} = \frac{p(\beta = 0|Data)}{p(\beta = 0)} \quad B_{1u} = \frac{P(\beta < 0|Data)}{P(\beta < 0)} \quad B_{2u} = \frac{P(\beta > 0|Data)}{P(\beta > 0)}$$

Quantifying Statistical Evidence Using the Bayes Factor

Multiple hypothesis test (some Greek...)

Consider the hypotheses: $H_0 : \beta = 0,\ H_1 : \beta < 0,\ H_2 : \beta > 0,\ H_u : \beta \in \mathbb{R}$:

$$B_{0u} = \frac{p(\beta = 0|Data)}{p(\beta = 0)} = \frac{.1}{.4} = .25 \qquad B_{1u} = \frac{P(\beta < 0|Data)}{P(\beta < 0)} = \frac{.98}{.50} = 1.96 \qquad B_{2u} = \frac{P(\beta > 0|Data)}{P(\beta > 0)} = \frac{.02}{.50} = .04.$$



Quantifying Statistical Evidence Using the Bayes Factor

Posterior probabilities

The Bayes factors can be used to update prior odds of the hypotheses to obtain posterior odds:

$$\frac{P(H_0|Data)}{P(H_1|Data)} = B_{01} \times \frac{P(H_0)}{P(H_1)}.$$

Three hypotheses

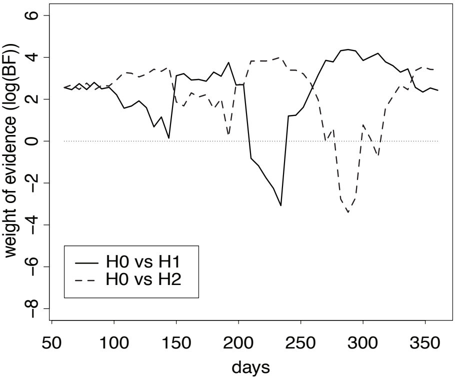
If we assume the three hypotheses, H_0, H_1, and H_2, to be equally likely a priori, then the posterior probabilities can be computed as

$$\begin{aligned} P(H_0|Data) &= \frac{.25}{.25 + 1.96 + .04} = .11 \\ P(H_1|Data) &= \frac{1.96}{.25 + 1.96 + .04} = .87 \\ P(H_1|Data) &= \frac{.04}{.25 + 1.96 + .04} = .02. \end{aligned}$$

The Evolution of Statistical Evidence

Hypothesis test I

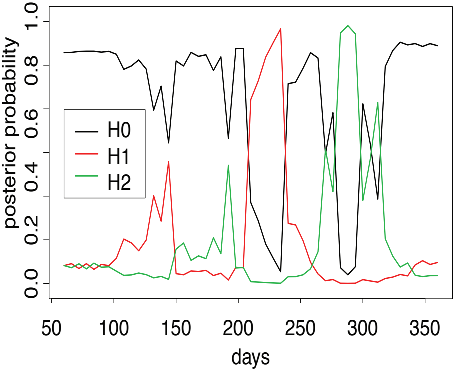
- H_0 : no hierarchical effect
- H_1 : bottom-up behavior
- H_2 : top-down behavior



The Evolution of Statistical Evidence

Hypothesis test I

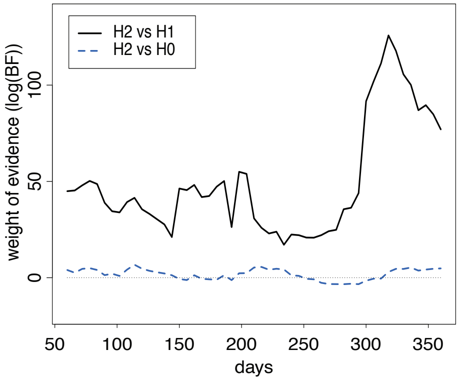
- H_0 : no hierarchical effect
- H_1 : bottom-up behavior
- H_2 : top-down behavior



The Evolution of Statistical Evidence

Hypothesis test II

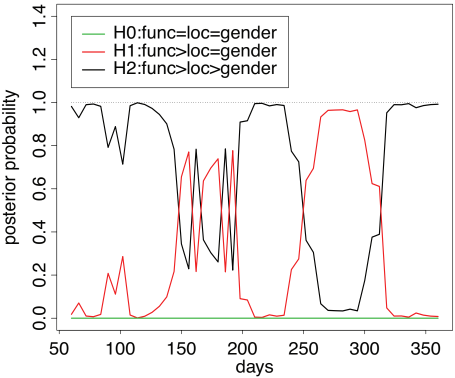
- H_0 : function = location = gender
- H_1 : function > location = gender
- H_2 : function > location > gender



The Evolution of Statistical Evidence

Hypothesis test II

- H_0 : function $=$ location $=$ gender
- H_1 : function $>$ location $=$ gender
- H_2 : function $>$ location $>$ gender



Outline

Conclusions

- Relational event model is useful for modeling email data in social networks of colleagues.
- Moving windows is useful to get insight about the dynamic behavior over time.
- The moving windows can be seen as a form of process memory.
- Default Bayes factors can be used to see how statistical evidence evolves over time.
- Still many open problems...

Future work

- Incorporate more realistic forms of process memory such as half-life periods (e.g., Brandes et al., 2009), and estimate/test half-life periods.
- Improve efficiency of statistical computing.
- Posterior predictive checking to evaluate model fit.
- ...