Modelling Communication Dynamics in Social Networks

Amber Tomas
September 28, 2011
Key Points

- Aim is to understand the factors influencing communication patterns
- Model communications as events on edges of a network
- Allow for factors influencing communication to change over time
- Jointly model events on edges and nodes
Questions:

▶ What are the patterns of communication?
▶ How can we explain them?
Questions:

▶ What are the patterns of communication?
▶ How can we explain them?

Statistical modelling allows for:

▶ Analysis of behaviour (testing hypotheses)
▶ Improved understanding of communication systems
▶ Monitoring of communication patterns for changes
Network Event Data

At any time $t$

- There exists a set of actors who may communicate
- There exists a set of possible communications *between pairs of actors*

An “event” is an instantaneous communication between two actors.
Network Event Data
Network Event Data
Network Event Data
Network Event Data
Network Event Data
Every event is recorded as a tuple \((a, b, k, t)\):

- \(a\) — sender
- \(b\) — receiver
- \(k\) — type of event
- \(t\) — time of event
Every event is recorded as a tuple \((a, b, k, t)\):

\[
\begin{align*}
    a & \quad - \quad \text{sender} \\
    b & \quad - \quad \text{receiver} \\
    k & \quad - \quad \text{type of event} \\
    t & \quad - \quad \text{time of event}
\end{align*}
\]

A dataset consists of

1. A sequence of recorded events
2. Information about each of the actors
Example - Enron Email Corpus

- 1.8 million emails sent from employees of Enron between November 1998 and September 2002
- Data includes
  - Time stamp
  - sender
  - recipients
  - type of recipient (to, cc, bcc)
  - subject
  - message content
Butts (2008) proposed a “relational event framework” for modelling network event data.

This models the rate of events between actors conditional on event history and actor characteristics.

Events are assumed to be independent conditional on the past event history.
Models for Event Data

The likelihood of the event history up to the most recent event is given by

\[ p(e_{1:t}) = \prod_{i=1}^{t} \left[ h(t_i | e_i, e_{1:i}) \times \prod_{\epsilon_i} S(t_i - t_{i-1} | \epsilon_i, e_{1:t-1}) \right] \]
The likelihood of the event history up to the most recent event is given by

\[ p(e_{1:t}) = \left[ \prod_{i=1}^{t} \left[ h(t_i|e_i, e_{1:i}) \times \prod_{\epsilon_i} S(t_i - t_{i-1}|\epsilon_i, e_{1:t-1}) \right] \right] \]

It is assumed the hazard is constant given past history, so

\[ h(t) = \lambda \]
\[ S(\Delta t) = e^{-\lambda \Delta t} \]
Models for Event Data

The rate function is modelled as

\[ \lambda(e_t, e_{1:t-1}, X, \theta) = \exp\left\{ \lambda_0 + \theta^T u(e_t, e_{1:t-1}, X) \right\} \]
Models for Event Data

The rate function is modelled as

$$\lambda(e_t, e_{1:t-1}, X, \theta) = \exp \left\{ \lambda_0 + \theta^T \mathbf{u}(e_t, e_{1:t-1}, X) \right\}$$

Considering only the ordering of data leads to the likelihood

$$p(e_{1:t} | \theta, X) = \prod_{i=1}^{t} \frac{\exp \left\{ \theta^T \mathbf{u}(e_t, e_{1:t-1}, X) \right\}}{\sum_{\epsilon_i} \exp \left\{ \theta^T \mathbf{u}(\epsilon_i, e_{1:t-1}, X) \right\}}$$
The rate function is modelled as

$$\lambda(e_t, e_{1:t-1}, X, \theta) = \exp \left\{ \lambda_0 + \theta^T u(e_t, e_{1:t-1}, X) \right\}$$

Considering only the ordering of data leads to the likelihood

$$p(e_{1:t} | \theta, X) = \prod_{i=1}^{t} \frac{\exp \left\{ \theta^T u(e_t, e_{1:t-1}, X) \right\}}{\sum_{\varepsilon_i} \exp \left\{ \theta^T u(\varepsilon_i, e_{1:t-1}, X) \right\}}$$

An important aspect of fitting this model is choosing the statistics $u(e_t, e_{1:t-1}, X)$. 
Choice of Statistics

1. Characteristics of sender $a$ and receiver $b$
2. Tendency for $a$ to send messages
3. Tendency for $b$ to receive messages
4. Tendency for communications between $a$ and $b$
5. Communications involving joint contacts of $a$ and $b$
Choice of Statistics

1. Characteristics of sender $a$ and receiver $b$
2. Tendency for $a$ to send messages (node)
3. Tendency for $b$ to receive messages (node)
4. Tendency for communications between $a$ and $b$ (dyad)
5. Communications involving joint contacts of $a$ and $b$ (triad etc)

Including network statistics allows for incorporation of complex dependencies.
Some Possible Statistics

To simplify computation, an event network is used to compute statistics.

The weight of an edge is an exponentially weighted aggregation of past events (Brandes et al).
Some Possible Statistics
Model Fitting

Model fitting can proceed as follows:

1. For every event
   1.1 calculate relevant statistics
   1.2 calculate contribution to likelihood
   1.3 Update weighted event network

2. Estimate parameters using maximum likelihood
Model Fitting

Model fitting can proceed as follows:

1. For every event
   1.1 calculate relevant statistics
   1.2 calculate contribution to likelihood
   1.3 Update weighted event network

2. Estimate parameters using maximum likelihood

Alternatively, it is natural to update the parameter estimates sequentially.
Model Fitting

The main difficulty is in specifying appropriate statistics, and correctly interpreting the fitted model.
Example - Enron Data
Example - Enron Data

ENRON

SUCCESSFUL
COMPANY
Example - Enron Data

ENRON
SUCCESSFUL COMPANY
REALITY
Example - Enron Data

ENRON
SUCCESSFUL
COMPANY
REALITY
Accounting
Shenanigans
Example - Enron Data

Considered a subset of email messages
- Only consider addresses ending in @enron.com
- Only consider messages sent to \( \leq 5 \) receivers
- Only consider actors who send and receive at least 2 messages
- Leaves 134 actors
Aggregated Data
Aggregated Data
Aggregated Data
Aggregated Data
Rate of Sending

(left) Distribution of sending over time

(right) Distribution of sending by day
Candidate Events
Choice of Statistics

![Scatter plot showing the number sent vs. number received for various individuals.](image)

- Need to control for actor tendency to send/receive messages.
Choice of Statistics

Need to control for actor tendency to send/receive messages
Choice of Statistics

- “Inertia” $w_{ab}$
- “Reciprocity” $w_{ba}$
- Sender “out-degree” $\sum_h w_{ah}$
- Triad effects:

![Triad diagrams](transitivity_sharedtriad.png)
Choice of Statistics

- role (manager, lawyer, executive, other)
- role homophily
- new member
Results

![Graph showing time series of various metrics]

- **E[β]**
- **Inertia**
- **Reciprocity**
- **Sender out degree**
- **Receiver out degree**
- **Sender in degree**
- **Receiver in degree**

The graph displays the time series of these metrics over time, with time ranging from 0 to 6000 and the metric values ranging from 0 to 50.
Results

- Inertia
- Reciprocity
- Sender's Out Degree
- Receiver's Out Degree
- Sender's In Degree
- Receiver's In Degree

![Graph showing the evolution of various network metrics over time.](attachment:image.png)
Results

![Graph showing various network metrics over time](image-url)

- $E[\beta_i]$
- inertia
- reciprocity
- sender.out.degree
- receiver.out.degree
- sender.in.degree
- receiver.in.degree
- transitivity
- shared.triad
- joint.triad

Axes:
- Time (x-axis)
- $E[\beta_i]$ (y-axis)
Data is available for a large hospital network, including

- Patient ward transfers (time, from ward, to ward, patient condition)
- Tests for hospital acquired infections (bug, time of test, time of result, result, patient info)
Example - Hospital Acquired Infections

Data is available for a large hospital network, including

- Patient ward transfers (time, from ward, to ward, patient condition)
- Tests for hospital acquired infections (bug, time of test, time of result, result, patient info)
- 14 000 transfers (11 000 discharged or died)
- 22 000 bug tests
- 2 700 bug positives
Data is available for a large hospital network, including

- Patient ward transfers (time, from ward, to ward, patient condition)
- Tests for hospital acquired infections (bug, time of test, time of result, result, patient info)

- 14,000 transfers (11,000 discharged or died)
- 22,000 bug tests
- 2,700 bug positives

We might be interested in the joint dynamics of infections and ward transfers.
Hospital Data
Can model the joint dynamics as

\[ p(e_{1:E}, v_{1:V}|\theta, \gamma) = p(e_{1:E}|\theta)p(v_{1:V}|\gamma), \]

where

\[ p(v_{1:V}|\gamma) = \prod_i \frac{\exp \{\gamma^T u'(e_{t<i}, v_{t<i})\}}{\sum_{\nu} \exp \{\gamma^T u'(e_{t<i}, \nu_{t<i})\}} \]
Hospital Data

Statistics might include

- number recently admitted to ward
- number on ward
- infection history of ward
- infection history of source wards
- transfer patterns and rates between wards
Conclusions

Network event data is becoming more common. Analysis can help improve understanding of the factors driving the communication dynamic.
Conclusions

Network event data is becoming more common. Analysis can help improve understanding of the factors driving the communication dynamic.

Possible applications include

- Business and emergency communications
- Political dynamics
- Personal communications

New studies should be designed with the possibilities for network analysis in mind.
Network Models

- The appropriate model for any data set will depend on what we are interested in.
- Suppose interested in the probability of an edge between two actors, given personal characteristics, network data.
- The probability of a tie between actors $A$ and $B$ might depend on
  - which other ties are present (degree, reciprocity, transitivity, etc)
  - actor characteristics (age, sex, role, etc)
Network Models - ERGM

\[ P(X = x) \propto \exp\left\{ \sum_k \beta_k s_k(x) \right\}, \]

where \( \beta \) is a vector of parameters and \( s(x) \) is a vector of statistics.

- implies edge probabilities
- allows testing of hypotheses regarding significance of actor or network statistics
- normalizing constant causes computational difficulties
- requires a single observation of the network
ERGMs and Event Data

We could aggregate all events to produce a “communication network”, i.e. who had communications with whom at any point in time. This allows investigation of questions such as

- Is transitivity a significant influence on the structure of the network?
- How many contacts do people have?
- Are communications reciprocated?
- Are communications more likely among actors of similar gender/age/ethnicity etc?

But this tells us nothing about the *temporal* aspects.
Stochastic actor-oriented network models (SAOMs) have been developed to model longitudinal network data. We could aggregate event data in sequential time intervals to produce such data.

- Allows testing of temporal hypotheses, eg did A become a manager because he was communicating with manager B, or did A only begin to communicate with B after obtaining a promotion independently?
- Results may be sensitive to the periods of aggregation
- Tells us little about communication rates or patterns over short time intervals.