Social Networks 33 (2011) 258–272

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Analyzing event stream dynamics in two-mode networks: An exploratory analysis of private communication in a question and answer community

Christoph Stadtfeld[a,b,*], Andreas Geyer-Schulz[b]

[a] Information Services and Electronic Markets, Karlsruhe Institute of Technology, Karlsruhe, Germany
[b] Department of Sociology/ICS, University of Groningen, The Netherlands

ARTICLE INFO

Keywords:
Two-mode networks
Actor oriented modeling
Event streams
Multinomial logit model
Markov process
Question and answer communities

ABSTRACT

Information about social networks can often be collected as event stream data. However, most methods in social network analysis are defined for static network snapshots or for panel data. We propose an actor oriented Markov process framework to analyze the structural dynamics in event streams. Estimated parameters are similar to what is known from exponential random graph models or stochastic actor oriented models as implemented in SIENA. We apply the methodology on a question and answer web community and show how the relevance of different kinds of one- and two-mode network structures can be tested using a new software.

1. Introduction

Question and answer (Q&A) communities (like *Yahoo! Answers*, *CosmIQ, Answers.com*) have become very popular in the web. People can easily (often even without registration) pose arbitrary questions. Members of these communities try to answer these questions quickly. Often, the only obvious incentives to answer questions are virtual points given to people who answer many questions. The more points someone has, the higher is his/her virtual ranking (e.g. ranging from *Newbie* to *Albert Einstein*). But are there any other effects that make people stay in these groups? Are there, for example, community structures that can be revealed when looking at how actors write private messages to others? Or is most of this private communication just functional and related to questions, like to provide further explanations or to say thank you if someone answered a question?

Actor oriented models are a good way to investigate tie changes in social networks dependent on structures in networks. Snijders (2005) introduced a class of models that is usually applied with panel data of binary network snapshots. The emergence of network structures can also be assessed on cross-sectional network data using the class of exponential random graph models (ERGM, see Wasserman and Pattison, 1996; Snijders et al., 2006; Robins et al., 2007). Here, the view is not actor oriented, but rather a global view on the network data. The standard class of ERGM has been extended so that it can handle multi-mode networks (see Wang et al., 2009). Beside models for cross-sectional data and panel data there is new research about the analysis of event stream data with dyad oriented models (Butts, 2008; Brandes et al., 2009; Stadtfeld et al., 2010). The increasing availability of event stream data allows to estimate structural models on this type of data as well. Event stream data incorporates a high amount of information that can be exploited. Algorithmic improvements in preprocessing and the estimation of local models make the application of such models feasible for long data streams and big networks.

In this paper, we present and apply a Markov process model framework to understand what drives the dynamics of private messages sent between actors in a Q&A community. Actor decisions about private communication tie formation and updates are conceptually described as a two-level decision process (for technical reasons, a third level is later added to the model). First, actors have a personal activity rate that influences the decision when to write a message at all. In case they decide to write a message, second, actors have to choose a receiver of the private message. This second decision about private message receivers is modeled as a multinomial logit model. This model expresses whether endogenous one-mode communication structures and two-mode affiliation structures have an influence on the choice of receivers in the community dataset at hand. A new java software package called *ESNA*[1] (event based social network analysis) was developed to estimate the parameters of this model framework.

Section 2 introduces the case study, a dataset of a big German speaking Q&A community, and identifies four phases of the

* Corresponding author at: IME Graduate School, Karlsruhe Institute of Technology, Karlsruhe, Germany. Tel.: +31 50 363 6950.
E-mail addresses: christoph@stadtfeld.net (C. Stadtfeld), andreas.geyer-schulz@kit.edu (A. Geyer-Schulz).

[1] More information can be found at http://www.em.uni-karlsruhe.de/ref/esna.

0378-8733/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2011.07.004

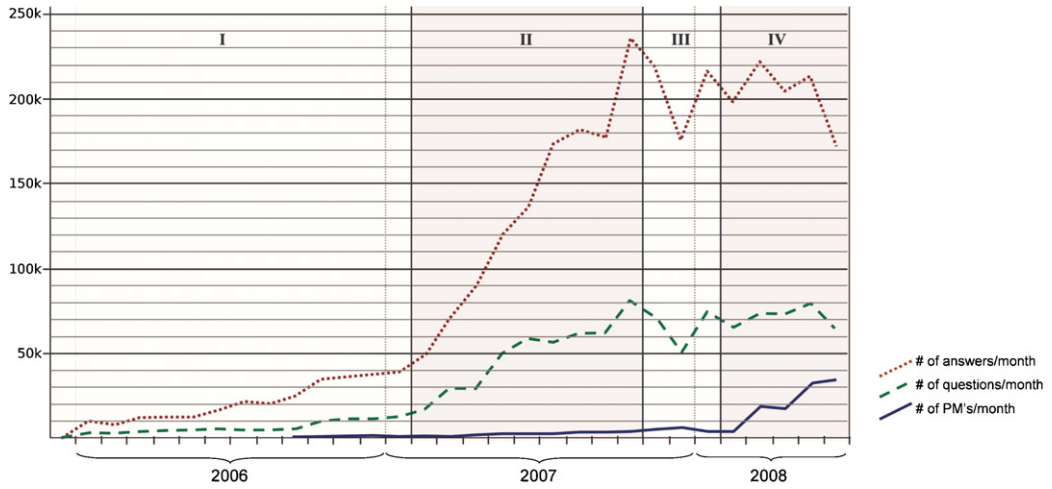


Fig. 1. Number of three different event types (questions, answers and private messages (=PM's) per month over the whole observed period (1K = 1000 events). Four phases were identified: (I) *Initialization*, (II) *Growth*, (III) *Stabilization* and (IV) *Community Growth*).

community development. The event stream of the case study is explained. Section 3 introduces the Markov process model that is used to analyze the data stream of the case study. It is introduced as a a composition of actor-driven choice models. Network update rules and network structures that potentially influence actor decisions are shown. The model is compared to other related models. In Section 4 parameter estimates for the last three sub-phases of the community development are given as well as information about the estimation algorithms. Section 5 discusses the parameter estimates of the multinomial receiver choice sub-model, as those results are especially interesting when trying to understand to whom people write private messages within the analyzed Q&A community. Finally, Section 6 summarizes and gives an outlook on further research.

2. Case study

A dataset of a big German Q&A community is analyzed to demonstrate the potential of an exploratory analysis with the Markov process framework introduced in Section 3. Some characteristics of the dataset at hand will be presented in Section 2.1. The event stream will be explained in Section 2.2.

2.1. The dataset

The dataset describes a time span from December 2005 to June 2008. Regarding their communication behavior, people in the Q&A community behave very different from other online social communities. First, the total number of members is big, but only a small subset of all actors is "active" at the same time, because a lot of people only pose one question and leave quickly. There are 416,879 activated user accounts, but 329,055 (79%) of them are "light accounts" that are just used to pose questions, but cannot be used to write or receive private messages. This implies that 87,824 (21%) actors are considered as the set of actors in the dataset. Second, the communication within the community is assumed to be influenced by questions. A virtual rank in the community is only based on how often and how good a member answers questions.

There are 946,603 questions in the dataset with 2,996,446 answers. Although the dataset starts in December 2005, private messages are only logged since August 2006. Fig. 1 shows how the activity concerning questions, answers and private messages changes over time.

The *x*-axis shows the different months, beginning with December 2005. The dotted line represents the number of answers, the dashed line the number of questions and the solid line the number of private messages sent. The *y*-axis gives the number of events (1K = 1000 events). Generally, the activity in the Q&A community increased. From this first visualization, four different phases were identified as shown in Fig. 1.

In phase I there is only little activity in the dataset with a rather low growth rate (from December 2005 to the end of January 2007). The number of private messages is low: In the first months, the number of messages does not exceed a few hundred messages. This first phase of the Q&A-Community is called *Initialization*.

Phase II is identified between February 2007 and October 2007 and is characterized by a rapidly increasing amount of questions, answers and a slow increase of the number of private messages. Therefore, it is called *Growth*.

In phase III, the numbers of questions, answers and private messages seem to have reached a more or less stable level. Although there is a lot of variance between the months (probably caused by Christmas time), the total number is always about 65,000 for questions and 210,000 for answers. The number of private messages is stable at a level of about 5000 per month. Phase III ranges from November 2007 to February 2008 and is called *Stabilization*.

Phase IV is probably the most interesting one regarding the dynamics of private communication, because the number of private messages rapidly increases, while the number of questions and answers is relatively stable. Phase IV ends – like the whole dataset – at the beginning of June 2008. The values for this last month are extrapolated as it was not completely logged. It will be tested, whether community effects are the reason for this sudden and significant increase of messages from an average of about 5000 per months to more than 30,000 messages in the last completely observed months. This phase is therefore called *Community Growth*. Whether this name is suitable (because the increasing number of messages is based on "community structures") will be tested in this paper.

2.2. Event stream

Events are any kind of directed, dyadic interaction between two nodes in a network for which at least a time-stamp is defined. Events may include more information, like an event type or an event intensity. The dataset includes events of different types, that can

Table 1
Part of the analyzed event stream (with fictitious names).

Time	Sender	Receiver	Type
2007-07-07 14:10:47	Anke	283613	question_opened
2007-07-07 14:10:51	mov_81	283604	answer
2007-07-07 14:11:00	doc-LE	283604	answer
2007-07-07 14:11:16	Snooker01	283600	answer
2007-07-07 14:11:19	mrs_incredible	270053	question_closed
2007-07-07 14:11:31	Nekoy	doc-LE	message
2007-07-07 14:11:42	Anke	283614	question_opened

be used to describe the changes in three different networks. The change in these networks is well defined by the event stream and a set of change rules. As the state of these networks is known for each point in time, this approach processes a lot more information than, for example, models using aggregated panel data.

To analyze the private message dynamics in the database, four different event types were identified and transformed into an event stream with more than 5 million entries. Each of these entries (a row in the resulting database table) describes one event and consists of a time-stamp, a sender, a receiver and an event type. An exemplary snapshot of the event stream is given in Table 1.

In this dataset, senders are always actors, while receivers may either be actors or questions. The first event type is *question_opened* which indicates that an actor poses a question (which is identified by a unique number). Event type *answer* shows that an actor responds to a question, while *question_closed* indicates that a question is closed either by the question opener, by an administrator, or because the maximum question lifetime of 7 days has been reached. Though the different event streams are not independent, this paper focuses on the dynamics of the fourth event type *message*, which shows that one actor writes a private message to another actor.

3. Modeling the case study

The decision making of actors regarding private message sending can be modeled as a Markov process with three different levels of decisions. First, this process has to decide, which actor writes a private message. Second, the conceptual decision about the receiver is split into two sub-decision: It is decided whether the possible receiver should be *active* or *non-active*. This "decision" is included for computational reasons since it considerably decreases the size of the multinomial choice model and reflects the fact that many actors leave the community after a short while. If the receiver is of non-active type the chosen sender picks the sender with equal probability. If the receiver is of active type, then the chosen sender decides whom of all *active* actors he or she sends the private message. This decision depends on the network structures that surround sender and receiver. Whether certain structures are relevant for actors' decisions can be tested with a multinomial logit model based on the observed behavior in the dataset. The network structures are a result of all events having been observed in the past and a set of change rules. These rules are briefly explained in Section 3.1. Section 3.2 shows how the regression statistics look like that are tested for influence on the decision process of actors. Section 3.3 introduces a heuristic that distinguishes active and non-active actors. In Section 3.4 the global Markov process is modeled. It consists of many individual decision processes that are explained in Section 3.5. The econometric evaluation of the model is based on certain assumptions that are listed in Section 3.6. Section 3.7 compares the model to stochastic actor-oriented models for longitudinal data sets, to exponential random graph models and to other event-based models.

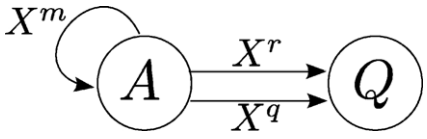


Fig. 2. Three different graphs are defined on the two node sets A (actors) and Q (questions). The weighted one-mode graph X^m represents recent private communication. X^q and X^r are binary two-mode networks and show which actors have asked questions (X^q) and which actors have responded to them (X^r) on the platform.

3.1. Transforming events into graphs

A state of the whole process is named x. x is a realization of a random variable X. x is defined by the state of three graphs at a certain point in time. These graphs are defined on two sets of nodes (two modes), which are the set of actors A (with elements $a_1, a_2, \ldots$) and the set of questions Q (with elements $q_1, q_2, \ldots$). The three graphs describing private messages, question asking and responses to questions are named X^m, X^q and X^r. A realization of x equals (x^m, x^q, x^r). X^m reflects the recent intensity of message writing between actors and has directed, weighted ties in $\mathbb{R}_0^+$. X^q shows which actors have posed a question that has not been closed, yet. X^r is a similar graph that connects actors with active questions they have responded to. The last two of these graphs have directed, binary ties and are bipartite two-mode graphs (affiliation networks). Fig. 2 shows how the node sets are connected by the three different graphs.

The event stream is an ordered sequence Ω with elements $\omega_1, \omega_2, \ldots, \omega_v, \ldots, \omega_{|\Omega|}$. If the position within the sequence is not of interest, the elements will just be named ω. The variables $\omega.time$, $\omega.sender$, $\omega.receiver$ and $\omega.type$ indicate the four attributes of events as introduced in Section 2.2 and shown in Table 1. Depending on these characteristics, events change certain ties of the graphs that define the Markov state X.

An overview of the used probabilistic (event triggered) change rules is given in Table 2. If an event of type *question_open* is observed, a new tie from an actor node to a question node is added to the two-mode network X^q. If this question is responded to (event type *answer*), a binary tie is added between the answering actor and the question node in graph X^r. If the question is closed (event type *question_close*), all attached ties are removed from graph X^q and X^r. X^m is a weighted graph (although the later used statistics dichotomize the observed ties). So, if an event of type *message* is observed in the data stream, the corresponding directed network tie from message sender to recipient is increased by 1 ($x_{ij}^{m\prime} = x_{ij}^m + 1$).

But even if no event takes place, the values of ties change due to deterministic, time dependent processes. In this case only one deterministic change rule is applied. The tie values of the private message graph X^m decrease over time with an exponential decay function (see Greiner et al., 1993). Introducing such a natural decay function seems reasonable, as otherwise the communication intensity between actors could only increase. Even, if there was no communication between actors for very long periods, this value would remain stable and we would still consider the private communication level as being very high.

Table 2
Overview of probabilistic (event triggered) graph change rules.

		Event types			
		question_open	*question_close*	*answer*	*message*
Graphs	X^m	–	–	–	tie value + 1
	X^q	add tie	remove tie	–	–
	X^r	–	remove ties	add tie	–

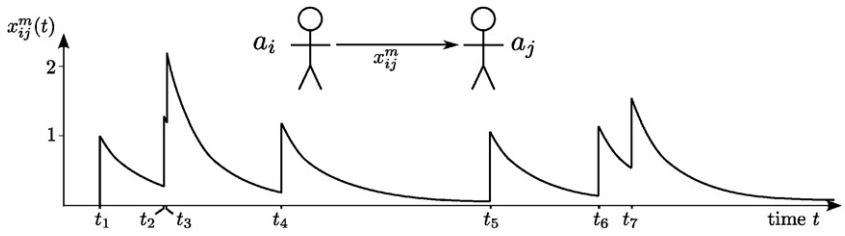


Fig. 3. Change of a directed private message tie from actor a_i to a_j. At each point in time $t_1, \ldots, t_7$ there is an event ω with $\omega.sender = a_i$ and $\omega.receiver = a_j$. Each event increase the tie value by 1 (probabilistic rule). Between events there is an exponential decay of ties with a fixed half-life (deterministic rule).

In an exponential decay function only a parameter *half-life* $t_{1/2}$ needs to be specified. It gives the time after which each tie value decreases by 50%. Ties that have a value $\leq \epsilon$ are reset to zero. ϵ is a small value, in this paper it was set to 0.01. The half-life was defined as 1 week. This is done for computational reasons to reduce the set of *active* actors that are considered to be potential receivers of messages (see Section 3.3). Note, that a decay plus a threshold value is similar to dichotomizing networks using a threshold in longitudinal models.

Fig. 3 shows how a directed communication tie (representing the recent private message writing intensity) between two actors changes over time driven by events of type *message* with a_i being the sender and a_j being the receiver. Whenever such an event takes place (at times t_1, t_2, …, t_7) the tie value is increased by one. Between the events, the tie value decreases due to the exponential decay.

3.2. Decision statistics

Ties in social networks do not emerge randomly. Existing network structures are often a good predictor for how people connect with others. These structures (see Figs. 4 and 5) can be measured and the resulting decision statistics be understood as independent variables of the receiver choice process modeled by a multinomial logit model (see Section 3.5). Parameter estimates describing the relevance of these structures can be interpreted similarly to estimates in exponential random graph models (ERGM, see Robins et al. (2007)) and SIENA models (Snijders, 2005).

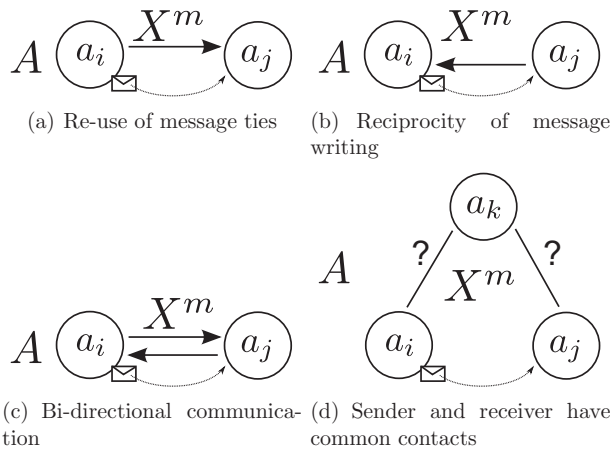
For each decision, this model evaluates the structures in the local environment of the senders and receivers of private messages. Decision statistics are functions $s_d(x, i, j)$ which map the state of the Markov process $x = (x^m, x^q, x^r)$ and the indices i and j of the sender a_i and the receiver a_j into $\mathbb{R}$. Structures can be measured within the communication network itself (endogenous structures) or in other one- or two-mode networks. Structures can in general incorporate actor attributes, multi-network structures or any combination of these (Stadtfeld, 2010).

In this paper, we are interested in whether endogenous (private communication is driven by previous private communication) structures in the message graph and two-mode structures measuring question affiliation influence the choice of event receivers. As mentioned before, only structures in the local environment of sender a_i and receiver a_j in A are evaluated.

3.2.1. Endogenous one-mode statistics

Fig. 4 shows four endogenous one-mode structures on the private message graph X^m.

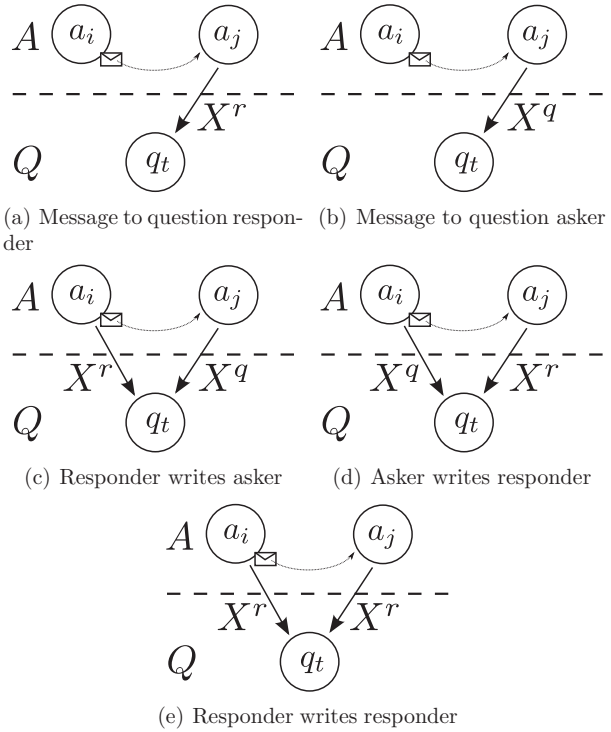
The statistic of the structure in Fig. 4(a) measures whether there is a tendency to re-use message ties, thus to repeatedly communicate with the same actors. Creating new ties is costly and most people have a smaller set of receivers they regularly communicate with, so we assume that this structure will have a positive influence on the probability of choosing the corresponding actor. This effect



(a) Re-use of message ties (b) Reciprocity of message writing

(c) Bi-directional communication (d) Sender and receiver have common contacts

Fig. 4. Four endogenous one-mode network structures that might influence private message receiver decisions. Actor a_i is the sender, a_j the receiver: (a) re-use of message ties, (b) reciprocity of message writing, (c) bi-directional communication, (d) sender and receiver have common contacts.



(a) Message to question responder (b) Message to question asker

(c) Responder writes asker (d) Asker writes responder

(e) Responder writes responder

Fig. 5. Two-mode structures measuring the affiliation of actors to questions. Actor a_i is the sender, a_j the receiver: (a) message to question responder, (b) message to question asker, (c) responder writes asker, (d) asker writes responder, (e) responder writes responder.

is only measured binary – the actual weight of the tie has no effect. Formally, it is defined with the function

$$s_1(x,i,j) = \begin{cases} 1, & \text{if } x_{ij}^m > 0 \\ 0, & \text{else} \end{cases} \quad (1)$$

The statistic of the structure in Fig. 4(b) measures whether actors tend to reciprocate previous, incoming private communication. A positive estimate for this statistic indicates that people prefer communication partners that have written a private message themselves before. It is measured by

$$s_2(x,i,j) = \begin{cases} 1, & \text{if } x_{ji}^m > 0 \\ 0, & \text{else} \end{cases} \quad (2)$$

The third of the endogenous structures is given in Fig. 4(c). It is a combination of the first two statistics and measures whether actors communicate repeatedly. It should, therefore, not be interpreted without regarding the first two statistics. This structure covers all the reciprocated messages from Fig. 4(b) except the very first response to a private message (or similar structures due to a decay of ties). It also covers all re-uses of a tie from Fig. 4(a) except re-uses of ties that are not reciprocated. It is only measured if the first two statistics are measured with a statistic of 1. This structure is, therefore, an interaction of the first two structures. It indicates whether actors tend to communicate bi-directionally with messages going back and forth for a long time, or whether private conversations are rather short. This could be expected in a rather topic oriented online community like a Q&A platform. The statistic is defined as

$$s_3(x,i,j) = \begin{cases} 1, & \text{if } x_{ij}^m > 0 \wedge x_{ji}^m > 0 \\ 0, & \text{else} \end{cases} \quad (3)$$

The fourth structure in Fig. 4(d) may reveal whether sender and receiver of the private message are embedded in community-like structures. The statistic counts the number of actors that sender and recipient are both connected to by previous private messages. The private communication with the counted third actors does not have to be bi-directional. So, it includes all types of transitive triangles, circles and their combinations. It is not taken into account whether a_i and a_j are connected on the message graph x^m. For each third actor, this structure is measured as the binary function f_4. The sum of these measurements (in $\mathbb{N}$) is the statistic of this structure.

$$s_4(x,i,j) = \sum_{a_k \in A \setminus \{a_i, a_j\}} f_4(x,i,j,k) \quad (4)$$

$$f_4(x,i,j,k) = \begin{cases} 1, & \text{if } (x_{ik}^m > 0 \vee x_{ki}^m > 0) \wedge (x_{jk}^m > 0 \vee x_{kj}^m > 0) \\ 0, & \text{else} \end{cases}$$

3.2.2. Two-mode statistics measuring question affiliation

Fig. 5 shows five structures that are measured to test, whether question affiliation has an effect on private communication. All five structures are two-mode structures with sender a_i and receiver a_j in A and questions from the set of questions Q. Structures in the asker graph X^q and in the responder graph X^r are evaluated.

All five structures measure binary ties. The first two structures (Fig. 5(a) and (b)) evaluate whether the receiver is connected to questions. Statistic $s_5(x,i,j)$ measures the tendency to write private messages to question askers, statistic $s_6(x,i,j)$ the tendency to write private messages to question responders:

$$s_5(x,i,j) = \sum_{q_t \in Q} f_5(x,j,t) \quad (5)$$

$$f_5(x,j,t) = \begin{cases} 1, & \text{if } x_{jt}^r = 1 \\ 0, & \text{else} \end{cases}$$

$$s_6(x,i,j) = \sum_{q_t \in Q} f_6(x,j,t) \quad (6)$$

$$f_6(x,j,t) = \begin{cases} 1, & \text{if } x_{jt}^q = 1 \\ 0, & \text{else} \end{cases}$$

The structures in Fig. 5(c)–(e) evaluate whether actors tend to write private messages to receivers who are connected to the same questions as the sender. It is differentiated between private messages from responder to asker (statistic $s_7(x,i,j)$), asker to responder ($s_8(x,i,j)$) and between responders of the same questions ($s_9(x,i,j)$). The statistics are defined as follows:

$$s_7(x,i,j) = \sum_{q_t \in Q} f_7(x,i,j,t) \quad (7)$$

$$f_7(x,i,j,t) = \begin{cases} 1, & \text{if } (x_{it}^r = 1) \wedge (x_{jt}^q = 1) \\ 0, & \text{else} \end{cases}$$

$$s_8(x,i,j) = \sum_{q_t \in Q} f_8(x,i,j,t) \quad (8)$$

$$f_8(x,i,j,t) = \begin{cases} 1, & \text{if } (x_{it}^q = 1) \wedge (x_{jt}^r = 1) \\ 0, & \text{else} \end{cases}$$

$$s_9(x,i,j) = \sum_{q_t \in Q} f_9(x,i,j,t) \quad (9)$$

$$f_9(x,i,j,t) = \begin{cases} 1, & \text{if } (x_{it}^r = 1) \wedge (x_{jt}^r = 1) \\ 0, & \text{else} \end{cases}$$

Note, that more complex structures could be tested as well. Also, it is possible to combine binary structures with weighted structures. More information is given in Stadtfeld (2010).

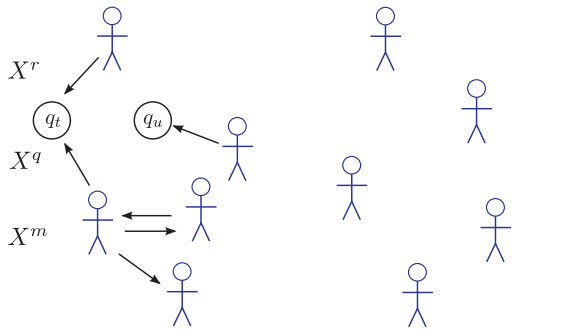
3.3. Active actors

A lot of accounts on the Q&A platform are only used for short time spans, e.g. just to pose one question. Therefore, active actors and non-active actors are distinguished. The set of actors A is split into the subsets A^+, the set of active, and A^-, the set of non-active actors with $A^+ \cup A^- = A$. These sets vary over time. We use a simple heuristic to define the set of active actors based on the three graphs X^m, X^q and X^r. Active actors are those that are connected to a non-closed question (as asker or responder) or have at least one in- or outgoing message tie with a value >0. Formally, for all actors $a_i \in A^+$ the following condition holds for each point in time:

$$\begin{aligned} a_i \in A^+ \Leftrightarrow \quad & (\exists q_k \in Q : x_{ik}^q = 1 \vee x_{ik}^r = 1) \\ \vee \quad & (\exists a_j \in A : x_{ij}^m > 0 \vee x_{ji}^m > 0) \end{aligned} \quad (10)$$

A graphical representation of the idea is shown in Fig. 6.

This heuristic reduces the computational complexity of the estimation significantly as A^+ is much smaller than A and in the model only actors in A^+ are considered as potential receivers of a private message (with high probability). The development of the size of set A^+ over time and an evaluation of the precision of the heuristic are given in Section 4.2.



(a) Active actors (A^+) (b) Inactive actors (A^-)

Fig. 6. It is distinguished between active actors in A^+ and non-active actors in A^-: (a) active actors (A^+) and (b) inactive actors (A^-).

3.4. Markov process

The proposed Markov process models the occurrence of private message events in the event stream (see Table 1) from a global perspective. It is a composition of three individual choice models that are explained in Section 3.5.

Let $\{X(t)|t \geq 0\}$ with state space $\mathcal{X}$ be a Markov process (a continuous-time Markov chain) with right-continuous realizations. The state space $\mathcal{X}$ is defined by all combinations of the possible states of the three graphs X^m (private message graph), X^q (posed questions) and X^r (responses to questions) – see Fig. 2. The Markov process describes updates of ties in X^m due to the occurrence of private message events in the event stream (see Table 1) as state changes. Formally, it holds that the state space is

$$\mathcal{X} = \{(x^m, x^q, x^r) : x^m \in (\mathbb{R}_0^+)^{n\times n}; x^q, x^r \in \{0, 1\}^{n\times m}\} \tag{11}$$

where n is the size of the set of actors A, m the number of questions in set Q and the small x denote concrete realizations of random variables X. Recall that X^m describes a weighted graph on the set of actors, while X^q and X^r are binary and bipartite graphs with ties connecting actors and questions as explained in Section 3.1.

A Markov process (or continuous time Markov chain) is a process "without memory" which means that all relevant information for the next process change is represented by the current state. Therefore, the emergence of new ties in the event stream is assumed to depend only on current network structures in the three graphs and a set of constant parameters. In this case, it does not matter by which sequence of events the graphs actually evolved. However, the state space can be extended to model this fact.

For two subsequent private message events $\omega_\nu = (i_\nu, j_\nu, t_\nu)$, $\omega_{\nu+1} = (i_{\nu+1}, j_{\nu+1}, t_{\nu+1})$ the Markov property holds:

$$\begin{aligned} &P(X(t_{\nu+1}) = x_{\nu+1}|X(t_\nu) = x_\nu, X(t_{\nu-1}) \\ &= x_{\nu-1}, \ldots, X(t_0) = x_0) = P(X(t_{\nu+1}) = x_{\nu+1}|X(t_\nu) = x_\nu) \end{aligned} \tag{12}$$

For each possible message event from a sender $a_i \in A$ to a receiver $a_j \in A$ a "tendency" for its occurrence is defined as a Poisson process with a rate λ_{ij} (explained in Eq. (13)). This is similar to the statement that the time between two consecutive messages from a_i to a_j is λ_{ij}-exponentially distributed.

These rates vary with the sending and receiving actors. λ_{ij} can be understood as the propensity of actor a_i to write a private message to a_j. It depends, first, on the general activity of a_i and, second, on the network structures that a_i, a_j and all other potential event receivers are embedded in. Based on these structures, a_i is assumed to make a choice about the receiver. A rather artificial third additional decision on the type of actor is included: It is distinguished between the two cases that the event receiver may be active or non-active at the time of the event. We understand all three decision levels of the Markov process transition rates as driven by individual choices of the sending actor.

3.5. The individual choice model

The transition rates of the Markov process are based on individual choices. They are described by a Poisson parameter $\lambda_{ij}(x; \rho_i, \beta, p^+)$ which models the decision of actor a_i to write a message and to choose actor a_j as the receiver, given the process state x and a set of stable parameters (ρ_i, β, p^+). The process state is only stable for short time intervals, as several not explicitly modeled processes change it: The exponential decay, new questions and answers in the community, and the closing of questions. Therefore, the transition rate is defined as an approximation:

$$\lambda_{ij}(x; \rho_i, \beta, p^+) \approx \begin{cases} \rho_i p_{ij}^?(x; \beta) p^+, & \text{if } a_j \in A^+ \text{ (i)} \\ \rho_i \dfrac{1}{|A^-|}(1 - p^+), & \text{if } a_j \in A^- \text{ (ii)} \end{cases} \tag{13}$$

with ρ_i the parameter which describes the general activity level of actor a_i, p^+ denoting the probability $P(\omega.receiver \in A^+)$, and $p_{ij}^?$ a multinomial logit model describing the choice of receivers and explained in Eq. (14). It depends on x and a weight vector β. The set A^+ changes with the state of the Markov process and can directly be derived from x as explained in Eq. (10). The rationale for the parameters is explained below.

ρ_i is a parameter of a Poisson process. It describes the general activity of an actor a_i regarding the sending of private messages. The parameters ρ_i of this process can be interpreted as the expected number of messages sent by actor a_i in a defined time span. In Eq. (13), the Poisson rate ρ_i is split into sub-Poisson rates of independent sub-processes in two different ways:

(i) For receivers in the set of active actors A^+, ρ_i is split by multiplying with the probability $p_{ij}^?$ from a multinomial logit model which describes a_i's choice of a receiver from this set. This case is weighted with p^+.
(ii) For receivers in the set of non-active actors A^-, ρ_i is split – for reasons of simplicity – into equal sub-rates by multiplying with $1/|A^-|$. This case is weighted with $(1 - p^+)$.

p^+: A case distinction is made depending on whether the receiver of the private message is active ($a_j \in A^+$) or not ($a_j \in A^-$, $A^- = A \setminus A^+$). Probability p^+ is equal to $P(\omega.receiver \in A^+)$. p^+ is assumed to be considerably higher than $1 - p^+$. So, most receivers of private messages are actually active. For all other receivers $\in A^-$ the probability of a selection is just equally distributed. The probability p^+ is a Bernoulli probability and is assumed to be independent from the size of the set of active actors A^+ that considerably changes over time (see results in Fig. 8). For any period in the dataset it is computed by the fraction of the number of messages sent to active actors to the number of all messages.

$p_{ij}^?$: In case of an active receiver, the probability for choosing a specific receiver $a_j \in A^+$ depends on the network structures sender a_i and receiver a_j are embedded in. $p_{ij}^?(x; \beta)$ is a multinomial logit model (see McFadden, 1974; Cramer, 2003, pp. 107–108; Hosmer and Lemeshow, 2000, pp. 260–263) on the set of all active receivers in A^+.

$$p_{ij}^?(x; \beta) = \frac{1}{c^+} \exp(\beta^T s(x, i, j)) \tag{14}$$

with

$$c^{+} = \sum_{a_k \in A^{+}} \exp(\beta^T s(x, i, k)) \tag{15}$$

$s(x, i, j)$ is a vector of network statistics that includes statistic functions like those in Section 3.2 (Eqs. (1)–(9)):

$$s(x, i, j) = \begin{pmatrix} s_1(x, i, j) \\ s_2(x, i, j) \\ \vdots \end{pmatrix} \tag{16}$$

Each statistic $s_d(x, i, j)$ in vector $s(x, i, j)$ is weighted with a corresponding parameter $\beta_d \in \mathbb{R}$. The vector β is unknown but can be calculated using a maximum likelihood estimation. β and $s(x, i, j)$ have the same dimension. The linear function $\beta^T s(x, i, k)$ describing a possible decision is transformed with an exponential function. The resulting value, giving a "weight" for the structures surrounding the sender and the observed receiver is normalized with the characteristics of all those weights that might have occurred, given that a_i decided to write the message to any active actor $a_k \in A^{+}$. This assures that $p_{ij}^{?}(x; \beta)$ is a proper probability distribution. The denominator c^{+} ("+" indicates that only *active* actors are evaluated) is given in Eq. (15). A^{+} directly follows from the process state x and the heuristic in Eq. (10).

3.6. Assumptions for estimating the individual choice models

There are three important assumptions connected to the formulation of this stochastic process as a Markov process:

(1) *No phase transitions in analyzed windows*: From Fig. 1 it follows that there are (at least four) different phases with specific characteristics. The characteristics of a phase should have some influence on the emergence of private message ties and should, therefore, be part of the Markov process state. However, we assume that within a shorter analyzed time window certain parameters are constant and therefore do not need to be encoded as part of the process states. This holds for the individual activity of actors, the probability to write messages to active actors and also for structural effects determining the choice of event receivers. We assume that there is no phase transition within one of the analyzed periods of the Markov process.
(2) *Local homogeneity*: The Markov process is assumed to be homogeneous (to be independent from the concrete point in time t) within an analyzed period as long as it is small enough. This is reasonable at least within phases II to IV (see Fig. 1) as the distribution of possible states would only marginally depend on the initial state with three empty graphs. Growth processes within a phase are not further considered. We assume different behavior patterns between the different phases II to IV and, in addition, homogeneous behavior in each "small" period analyzed.
(3) *Stability in short time spans*: A concrete process state $x = (x^m, x^q, x^r)$ is influenced by a decay of message ties of graph x^m and also by events of other type that change the graphs x^q and x^r. We define the Markov process transition rates only for very short time spans so that we can assume an only marginal relevance of those factors. As the general activity of private message writing is high we consider this a reasonable assumption.

3.7. Related models

The probability $p_{ij}^{?}$ in Eq. (14) and its combination with the activity Poisson parameter ρ_i in Eq. (13) are similar to the stochastic actor oriented model for longitudinal network data (SAOM) introduced by Snijders (2001, 2005) and implemented in SIENA. However, the proposed model in this paper does not explicitly describe a creation and dissolving of binary ties but an update of weighted ties. Dissolving of ties is modeled as an external exponential decay process. A further difference is the local evaluation of network structures from an ego perspective of the event sender while SAOM evaluates all structures in the network. In this paper, we estimate individual Poisson rates, instead of estimating an average activity rate for all actors. The model we propose here can further be extended to estimate multinomial decision models on an actor level as well. As we assume full information about events in the data set the estimation is much more straightforward.

The probability in Eq. (14) is also comparable to the evaluation of network probabilities in exponential random graph models (ERGM, see Robins et al., 2007). Both ERGM and this probability depend on the occurrence of network structures. While ERGM evaluates the global network, the probability $p_{ij}^{?}$ from Eq. (14) has a local view (the local environment of sender and receiver) on network structures. The base line model of ERGM is a random graph, while in this case the parameters are compared to a random decision over all potential actors. In ERGMs, the denominator is a (often not computable) *constant* over all possible outcomes of a graph. Therefore, ERGMs can be interpreted as exponential family models. Due to the homogeneity assumption in the proposed Markov process, the expected outcome of the denominator can be interpreted as a stationary distribution of local environments which is similar to the idea of a normalizing constant.

In Butts (2008) the occurrence of events is described by Poisson rates that can be parameterized in a very flexible way. The proposed *relational event framework* is based on classical event history modeling. As in the framework introduced in this paper, the time intervals between events are exponentially distributed. It is, however, not distinguished between different decision levels. In an exemplary application, Butts uses a time-discrete sub-model which does only take the order of events over time into account. The multinomial likelihoods defined for each event observation in this discrete sub-model look similar to Eq. (14) but are not based on (econometric) choice theory. The relational event framework is not explicitly actor-oriented. The original idea is rather "behavior-oriented" modeling (see Butts, 2008, p. 167). The relational event model does not allow external processes to change the process states like the external decay function in our model.

In this paper we try to separate different decisions. This is very useful in very big datasets as the different sub-models can be estimated separately. This means, however, that in our approach we have to consider quite strong independence assumptions between the different decision levels. It is possible to extend this approach with more sophisticated Poisson rates that also depend on environmental parameters. But these parameters must then be independent from the parameters used in the other sub-models like, for example, the choice of receivers. Other additional sub-decisions, like decisions about event intensities or event types, can easily be included in this framework. Butts (2008) allows modeling several event types in one dynamic model. In Brandes et al. (2009) an extension of the approach in Butts (2008) was introduced that explicitly takes weights of events into account.

4. Estimation and results

The proposed event model is an actor-driven three-level decision process (see Eq. (13)). First, the general actor activity is modeled with the Poisson parameter ρ_i. Second, the probability for choosing an active actor instead of an unconnected actor is given by the probability p^{+}. Third, if an active actor is chosen, the

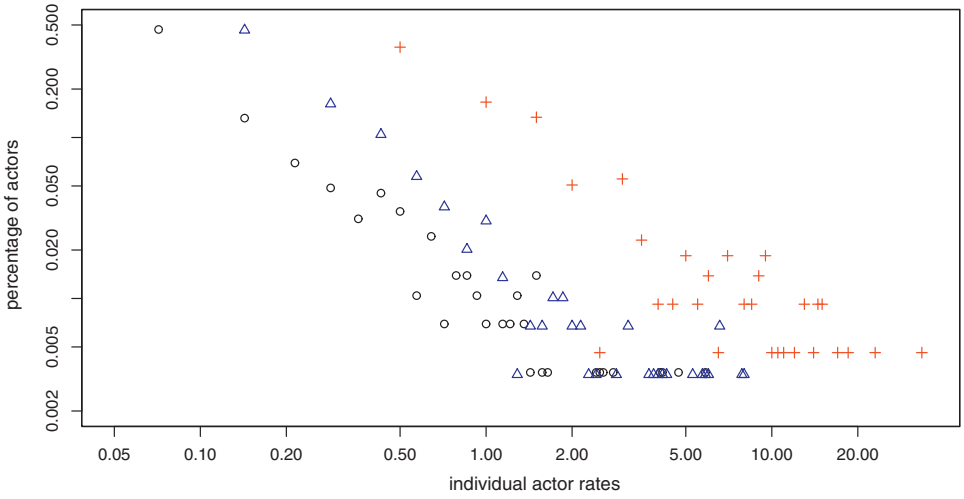


Fig. 7. Relative number of individual Poisson rates in the sub-windows of phase II (○), phase III (△), and phase IV (+). Both scales are logarithmic.

Table 3
Three sub-streams of phases II–IV were selected.

	Phase II	Phase III	Phase IV
Windows starts	August 1, 2007	December 3, 2007	March 10, 2008
Window ends	August 14, 2007	December 9, 2007	March 11, 2008
Length	14 days	7 days	2 days
Messages	1465	1323	1227
Sending actors	288	297	217
Avg. size of A^+	8710.96	6725.97	10,102.90

multinomial choice is given by the probability $p^?_{ij}(x;\beta)$. Results of the parameter estimates are given for each level separately.

As the four phases of the Q&A community seem to have different characteristics (see Fig. 1), a subsequence of each phase II to IV was evaluated separately. The whole event stream has more than five million events including 120,000 private message events. Therefore, it is already sufficient to analyze smaller samples of the stream to get statistically significant results. Also, the process assumes a homogeneity in behavior within shorter time spans as mentioned in Section 3. This makes it reasonable to look at smaller, stable subsets of the stream within three of the four phases. Subsequences of 2 days to 2 weeks were chosen for estimation that included enough message events to get stable results. Analyzing bigger sub-streams is possible though, as memory complexity and computational complexity only increase linearly with the number of estimated events for each simulation iteration step (due to a preprocessing of network statistics). Some characteristics of the three windows chosen are provided in Table 3.

4.1. Estimated actor activities ρ_i

The Poisson rates ρ_i are given separately for each of the defined phases II to IV. Only those actors that wrote a message within such a time span are considered. The estimates were calculated with a maximum likelihood estimation. The time unit is days. The average individual Poisson rates $\overline{\rho_i}$ are presented in Table 4. On average, the actors that send messages in phase II at all wrote 0.372 messages per day. In phase III this rate almost doubled to 0.657 messages per day. In phase IV actors wrote (on average) 2.903 messages per day if they wrote private messages at all. In this phase, we observed a big growth of private messages in the community. Partly, this is reflected by a higher average activity of actors.

In Section 3 the Poisson rates ρ_i were defined as individual parameters of actors a_i. Plots of the individual actor activities in the selected windows are shown in Fig. 7. The x-axis shows individual actor rates, the y-axis gives the fraction of actors. Both axes are logarithmic. The different phases are indicated by three different symbols. It can be seen that the frequency distribution of parameters seems similar (many actors have a low rate and only few actors have very high rates, the logarithm of the curve is almost linear), but the absolute values are significantly higher in the later phases. In phase IV almost 7% of all actors had an activity rate in the range from 10 to 30 expected messages per day. In phase II the highest observed rate was 5.86 only. In phase III the maximum was 7.86 only.

4.2. Probability for choosing active actors p^+

The average probability p^+ of choosing an active actor in A^+ over a non-active actor in A^- was 96.27% in the observed event stream. Of 112,811 messages in the completely logged months from September 2006 to May 2008 only 4208 were observed to be sent to inactive actors, which shows that our heuristic (see Eq. (10)) worked quite well.

In phases II, III and IV the observed probability p^+ had values between 96.92% and 97.67%. The estimates of this Bernoulli probability are shown in Table 5.

The average number of active actors $|A^+|$ per month is shown in Fig. 8. The x-axis indicates the beginning of the months. As in Fig. 1, the different phases are highlighted with different shades of gray. In phase I less than 2000 active actors were in the network. The number increased to an average of almost 10,000 active actors at the end of phase II. We showed in Fig. 1 that also the number

Table 4
Average Poisson rates of all actors that wrote at least one message in one of the phases II–IV (see Fig. 1) with standard errors of estimates.

Phase	$\overline{\rho_i}$	(S.E.)
II	0.372	(0.036)
III	0.657	(0.047)
IV	2.903	(0.116)

Table 5
Observed probabilities of p^+ in the sub-streams of phases II–IV.

Phase	All messages	Messages to any $a_j \in A^+$	p^+
II	1500	1465	97.67%
III	1365	1323	96.92%
IV	1260	1227	97.38%

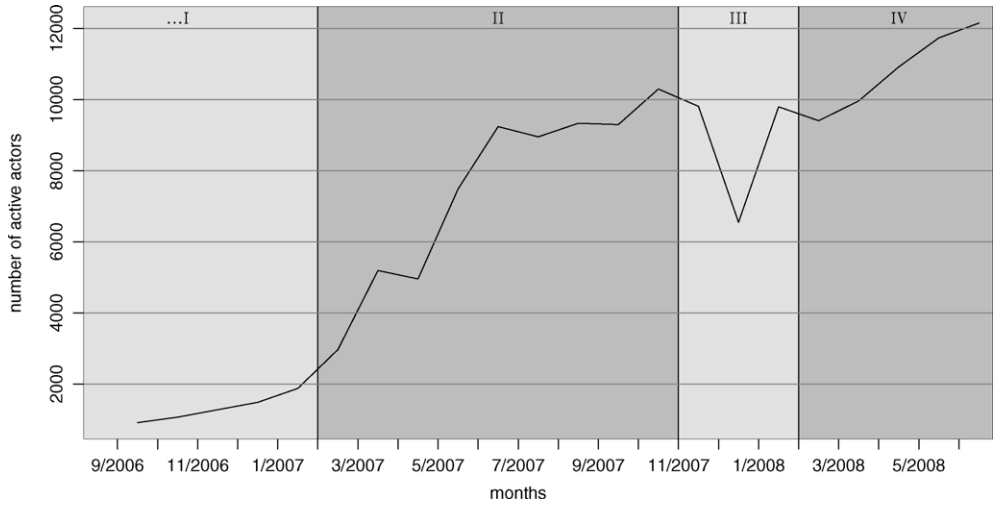


Fig. 8. Number of active actors in the analysis over time. The plotted values are average values of a month. Phases I–IV are indicated by the background color. A tick on the *x*-axis indicates the first day of a month. The total number of actors is $|A| = 87,824$.

of questions and answers increased significantly in this time span. In phase III, the number of active actors varied between 6000 and 10, 000. In December we observed a much lower activity on the platform, probably due to Christmas break. In phase IV, the number of active actors slightly increased with the increasing number of private messages and goes up to about 12, 000 in the last 2 months. The set of actors A is constant and has a size of 87,824.

Fig. 9 shows how the parameter p^+ changes over time. As in Fig. 8 the *x*-axis indicates the months between September 2006 to June 2008. The *y*-axis represents the percentage of messages sent to active actors (the monthly average of p^+). Except from low values at the end of phase I and the beginning of phase II the value is rather stable and always higher than 92.5%. In phase IV which includes most of the messages, the percentage of messages written to active actors is at a even higher level of about 97.5%.

4.3. Choice of event receivers $p^?$

The best fitting parameters $\hat{\beta}$ of probability $p^?_{ij}(x;\beta)$ (see Eq. (14)) are calculated by applying a maximum likelihood (ML) estimation.

$$\max_{\beta} \log L = \sum_{\nu=1}^{|\Omega|} \log\, p^?_{i_\nu j_\nu}(x_\nu;\beta) \tag{17}$$

Ω is the event stream with ordered events $\omega_1, \ldots, \omega_\nu, \ldots \omega_{|\Omega|}$. i_ν and j_ν are the indices of actors $a_{i_\nu} = \omega_\nu.sender$ and $a_{j_\nu} = \omega_\nu.receiver$. The event triggered changes have not yet been applied on x_ν.

For each event $\omega_\nu \in \Omega$ the decision probabilities $p^?_{i_\nu j_\nu}(x_\nu, \beta, A^+_\nu)$ were assumed to depend only on the network structures at that time (being conditionally independent given x). We assumed that those structures have a stable stationary distribution at least for shorter time windows within the event stream and are not significantly influenced by previous events taking place in other environments.

Standard errors of parameter estimates were estimated using a bootstrapping approach with a sample size of 50 (Efron and Tibshirani, 1986).

The log likelihood function in Eq. (17) is concave and can therefore be estimated using a Newton-Raphson algorithm (see Deuflhard, 2004). The software for network statistics preprocessing and estimation was developed in Java, partly using software packages from Apache Commons.[2]

The estimation results for the sub-windows in phases II to IV are shown in Tables 6, 7 and 8. The figure references and the names of the statistics are given in the first column. Nine different models were tested. The first model only includes the one parameter that improved the log likelihood most compared to a random decision model. This base line model includes no parameters and returns a probability of $1/A^+_\nu$ (uniform probability distribution over all potential receivers) for each event ω_ν in the window. The log likelihood of these random decision models is given in the captions of the tables. The additional eight parameters were included stepwise by the additional improvement of the log likelihood (forward selection, see Miller, 2002). The log likelihood of a model is shown in the first row under the model together with Akaike's information criterion (AIC, see Akaike, 1974). In the second row, for each model the deltas of the log likelihood ($\Delta \log L$) and the AIC (Δ AIC) are given. The values are compared to the model with the highest log likelihood and the model with the lowest AIC. In all three tables most parameters are significant with a level of $p < 0.001$ in most cases. Less significant parameters are italicized, non-significant parameters are marked with an "x". More details are given in the captions of the three Tables 6, 7 and 8. The best model regarding the AIC is highlighted by a gray background. This means that all subsequent models with more parameters (and a higher log likelihood) did not further reduce the AIC. The best model has the minimum AIC which is defined by $(-2\log L/n) + (2k/n)$ with k is the number of parameters and n the number of private messages in the observed window (see Table 3). The AIC values are given in the same row as the model log likelihood.

For all models the fit seems to be quite good. Compared with the log likelihood of the random decision reference models (II: −13, 290.974; III: −11, 660.567; IV: −11, 313.649) all models in Tables 6, 7 and 8 have a considerably higher log likelihood (in the range from −3993.563 to −6111.000). This indicates that the models made a contribution towards explaining the receiver choices of the private communication behavior. The estimates are discussed in Section 5.

[2] http://commons.apache.org/.

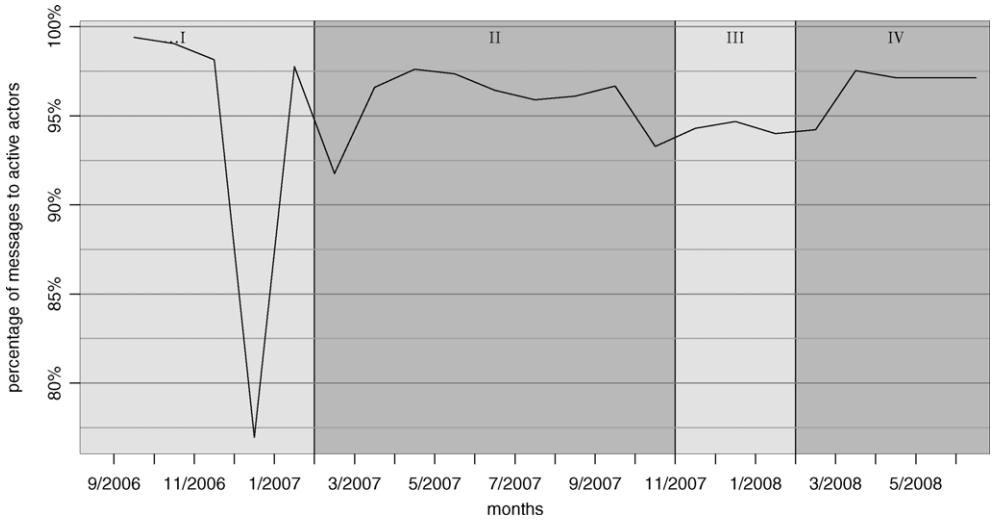


Fig. 9. Percentage of messages sent to active actors (p^+) per month. Phases I–IV are indicated by the background color. A tick on the x-axis indicates the first day of a month.

5. Discussion of receiver choice parameters

In all sub-windows of phases II to IV we discovered both significant endogenous effects and significant question affiliation effects. Most statistics turned out to be significant on a very high level. In total, only three statistics were not included in a best fitting model. The most important effect in all phases turned out to be the tendency of actors to re-use ties. Reciprocity and bi-directional communication were always relevant but seemed to be more common in the last analyzed phase. Common contacts also increased the probability for communication. The two-mode structures helped a lot to explain the model variance as well. In the first two analyzed phases, question affiliation structures were even the second and third most important independent variables contributing to model fit.

In the following, we discuss the findings for endogenous one-mode statistics and two-mode statistics measuring question affiliation separately.

Table 6
Nine models with detailed parameters for a sub-window of phase II. The last two models were excluded as the additional parameters are insignificant and the additional log likelihood improvement does not reduce the AIC. The random decision log likelihood is −13, 290.974. Most parameters are significant with $p < 0.001$. Estimates with a significance level of $p < 0.01$ only are *italicized*. Parameters with a lower significance level are indicated by a "x".

Figure	Name	Model II-1		Model II-2		Model II-3	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	9.071	0.113	8.711	0.123	8.598	0.142
5(a)	Message to responder			0.008	7.E-4	*0.008*	*6.E-4*
5(d)	Asker writes responder					0.275	0.080
log L/AIC		−6111.000/8.344		−5862.794/8.007		−5822.067/7.952	
Δ log L/Δ AIC		−430.774/+0.579		−182.568/+0.242		−141.834/+0.187	

Figure	Name	Model II-4		Model II-5		Model II-6	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	8.508	0.112	8.425	0.159	8.605	0.141
5(a)	Message to responder	0.008	6.E-4	0.008	6.E-4	0.007	9.E-4
5(d)	Asker writes responder	0.253	0.07	*0.249*	*0.094*	*0.232*	*0.072*
4(d)	Common contacts	0.114	0.020	0.102	0.020	0.113	0.020
4(b)	Reciprocity			0.663	0.190	4.837	0.429
4(c)	Bi-directional comm.					−4.423	0.482
log L/AIC		−5787.983/7.907		−5773.338/7.834		−5687.623/7.773	
Δ log L/Δ AIC		−107.754/+0.142		−93.112/+0.069		−7.397/+0.008	

Figure	Name	Model II-7		Model II-8		Model II-9	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	8.574	0.151	8.567	0.135	8.565	0.154
5(a)	Message to responder	0.007	6.E-4	0.007	7.E-4	0.007	7.E-4
5(d)	Asker writes responder	*0.224*	*0.077*	0.232	0.070	0.230	0.069
4(d)	Common contacts	0.108	0.02	0.108	0.025	0.108	0.024
4(b)	Reciprocity	4.860	0.471	4.863	0.521	4.861	0.321
4(c)	Bi-directional comm.	−4.448	0.564	−4.435	0.554	−4.433	0.405
5(b)	Message to asker	*0.011*	*0.004*	*0.014*	*0.005*	*0.014*	*0.005*
5(c)	Responder writes asker			−0.025	0.022x	−0.026	0.028x
5(e)	Responder writes responder					0.004	0.012x
log L/AIC		−5681.133/7.765		−5680.348/7.766		−5680.226/7.767	
Δ log L/Δ AIC		−0.907/0.000		−0.122/+0.001		0.000/+0.002	

Table 7
Nine models with detailed parameters for a sub-window of phase III. The last model was excluded as the additional parameter is insignificant and the additional log likelihood improvement does not reduce the AIC. The random decision log likelihood is −11, 660.567. Most parameters are significant with $p < 0.001$. Estimates with a significance level of $p < 0.05$ only are *italicized*. Parameters with a lower significance level are indicated by a "x".

Figure	Name	Model III-1		Model III-2		Model III-3	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	8.648	0.170	8.203	0.149	8.176	0.156
5(e)	Responder writes responder			0.360	0.036	0.281	0.036
5(d)	Asker writes responder					0.340	0.049
log L/AIC		−5697.398/8.614		−5308.846/8.028		−5220.517/7.896	
$\Delta \log L/\Delta$ AIC		−721.462/+1.079		−332.910/+0.493		244.581/+0.361	

Figure	Name	Model III-4		Model III-5		Model III-6	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	8.067	0.172	8.220	0.190	8.246	0.201
5(e)	Responder writes responder	0.26	0.030	0.256	0.036	0.189	0.033
5(d)	Asker writes responder	0.330	0.051	0.340	0.050	0.294	0.039
4(b)	Reciprocity	1.477	0.236	5.944	0.414	6.447	0.460
4(c)	Bi-directional comm.			−4.729	0.410	−5.464	0.463
4(d)	Common contacts					0.065	0.017
log L/AIC		−5166.512/7.816		−5076.189/7.681		−5018.122/7.595	
$\Delta \log L/\Delta$ AIC		−190.576/+0.281		−100.253/+0.146		−42.186/+0.060	

Figure	Name	Model III-7		Model III-8		Model III-9	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	8.221	0.145	8.081	0.165	8.080	0.186
5(e)	Responder writes responder	0.166	0.036	0.112	0.038	*0.111*	*0.041*
5(d)	Asker writes responder	0.280	0.040	0.293	0.043	0.294	0.042
4(b)	Reciprocity	6.411	0.335	6.234	0.335	6.227	0.411
4(c)	Bi-directional comm.	−5.348	0.432	−5.174	0.415	−5.156	0.447
4(d)	Common contacts	*0.058*	*0.020*	*0.060*	*0.026*	*0.059*	*0.024*
5(c)	Responder writes asker	*0.234*	*0.074*	*0.242*	*0.069*	*0.226*	*0.086*
5(a)	Message to responder			0.004	0.001	0.004	7.E-4
5(b)	Message to asker					0.005	0.007x
log L/AIC		−4996.916/7.564		−4976.391/7.535		−4975.936/7.536	
$\Delta \log L/\Delta$ AIC		−20.980/+0.029		−0.455/0.000		0.000/+0.001	

5.1. Endogenous one-mode statistics

In all three windows we observed highly significant estimates of the three *dyadic*, endogenous one-mode statistics: *Re-use of ties* and *Reciprocity* were always positive, while *Bi-directional communication* was negative. It is interesting to see that *Re-use of ties* always explained most of all parameters and as soon as *Reciprocity* was included in a model, the statistic *Bi-directional communication* increased the log likelihood significantly as the next included parameter (and increased the significance of the first two). As mentioned in Section 3.2, the third structure is interpreted as an interaction effect.

The statistics *Re-use of ties* and *Reciprocity* are the two main effects and *Bi-directional communication* models the interaction between the two main effects. Therefore, the interpretation of the third effect has to take the estimates of the first two effects into account. This can best be understood by defining an "equivalent" model with two new statistics replacing *Re-use of ties* and *Reciprocity*:

In this equivalent model, *Re-use of ties* ($s_1(x, i, j)$) can be substituted by an effect that measures the re-use of a tie only if there is no in-coming tie from the sender. This effect is named s'_1. The standard *Reciprocity* effect ($s_2(x, i, j)$) can be replaced with an effect that measures reciprocity only if there is no re-usable communication tie from sender to receiver. This effect is named s'_2. Eqs. (18)–(23) show how these effects are formally defined.

$$s_1 := s_1(x, i, j) \tag{18}$$

$$s_2 := s_2(x, i, j) \tag{19}$$

$$s_3 := s_3(x, i, j) \tag{20}$$

$$s'_1 := s_1 - s_3 \tag{21}$$

$$s'_2 := s_2 - s_3 \tag{22}$$

$$s'_3 := s_3 \tag{23}$$

Table 9 shows how the different dyadic statistics measure the four possible states of the communication dyad between sender a_i and receiver a_j in the private message communication network x^m.

The four rows show first a complete dyad with ties in both directions, then, second and third, two dyads with just one directed tie either from sender to receiver or from receiver to sender, and, fourth, an empty dyad with no positive communication tie. It can be seen that in a model with the three statistics s'_1, s'_2, s'_3 (the last three columns of Table 9) the three non-empty structures in the first three rows are measured disjointly. This has an effect on the interpretation of parameter estimates as we will demonstrate in the following.

In Table 10 we compare the estimates (without s.e.) of model IV-3 from Table 8 (with statistics s_1, s_2 and s_3) with an equivalent model IV-3′ (with statistics s'_1, s'_2 and s'_3) which is based on the "equivalent" definition given above. The results are shown in two sub tables.

In the first variant the interaction effect acts as a correction of the two main effects. In the second variant the three statistics measure disjoint structures (see Table 9). Therefore, the estimates measure the influence of each dyadic structure on receiver choices separately. The log likelihood of both models is the same.

What we learn from this comparison is that in a model with structures s_1, s_2 and s_3 included (model IV-3, for example) the first two structures can – as in model IV-3′ – be interpreted as the influence of non-bi-directional structures on the choice of receivers.

Table 8
Nine models with detailed parameters for a sub-window of phase IV. The random decision log likelihood is −11, 313.649. Most parameters are significant with $p < 0.001$. Estimates with a significance level of $p < 0.01$ only are *italicized*.

Figure	Name	Model IV-1		Model IV-2		Model IV-3	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	8.820	0.112	6.306	0.522	7.967	0.204
4(b)	Reciprocity			3.692	0.529	8.380	0.224
4(c)	Bi-directional comm.					−5.984	0.276
log L/AIC		−5242.867/8.547		−4592.861/7.490		−4186.123/6.828	
Δ log L/Δ AIC		−1249.304/+2.032		−599.298/+0.966		−192.560/+0.304	

Figure	Name	Model IV-4		Model IV-5		Model IV-6	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	7.875	0.223	7.665	0.218	7.645	0.181
4(b)	Reciprocity	8.213	0.275	8.051	0.218	7.948	0.257
4(c)	Bi-directional comm.	−5.877	0.303	−5.72	0.264	−5.672	0.265
5(c)	Responder writes asker	0.473	0.059	0.474	0.060	0.332	0.068
5(a)	Message to responder			0.004	6.E-4	0.004	7.E-4
5(b)	Message to asker					0.033	0.007
log L/AIC		−4105.352/6.698		−4067.848/6.639		−4046.842/6.606	
Δ log L/Δ AIC		−111.789/+0.174		−74.285/+0.115		−53.279/+0.082	

Figure	Name	Model IV-7		Model IV-8		Model IV-9	
		$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
4(a)	Re-use of ties	7.580	0.234	7.525	0.197	7.516	0.243
4(b)	Reciprocity	7.997	0.249	7.967	0.259	8.015	0.248
4(c)	Bi-directional comm.	−5.744	0.292	−5.727	0.289	−5.763	0.293
5(c)	Responder writes asker	0.319	0.078	0.300	0.082	0.332	0.084
5(a)	Message to responder	0.004	8.E-5	0.003	9.E-4	0.005	9.E-4
5(b)	Message to asker	0.034	0.008	0.033	0.005	0.029	0.006
4(d)	Common contacts	0.082	0.018	0.082	0.016	0.085	0.019
5(d)	Asker writes responder			*0.224*	*0.071*	0.258	0.060
5(e)	Responder writes responder					−0.114	0.026
log L/AIC		−4028.618/6.578		−4012.096/6.553		−3993.563/6.524	
Δ log L/Δ AIC		−35.055/+0.054		−18.533/+0.029		0.000/0.000	

This follows from the fact that the values of the first two parameters in model IV-3 and IV-3′ are the same (see Table 10). The third parameter of model IV-3 can best be understood by interpreting it as a correction of the sum of the first two parameters which is 7.967 + 8.380 − 5.984 = 10.363. This is exactly the estimate of the complete sender-receiver-dyad on the choice of event receivers in the disjoint alternative model. Therefore, bi-directional pure communication is actually best be understood as enforcing communication choices although the correcting parameter is negative.

Although a model with the equivalent statistics s'_1, s'_2 and s'_3 would have been more straightforward to interpret we chose to use the non-disjoint statistics. The reason is that with no bi-directional communication statistic s_3 or s'_3 included in a model statistics s_1 and s_2 explain more of the overall model (they generate a higher log likelihood). In the following, the interpretation of the absolute dyadic parameter estimates will be discussed.

The absolute values of the three dyadic parameters can be interpreted by comparing the probability for the choice of receivers *with* a certain structure to the choice of receivers without any dyadic structure (the fourth row in Table 9). The probability of sender a_i for choosing a certain receiver a_j over any other receiver was

Table 9
Overview of the values for the dyadic statistics including the three alternative statistics s'_1, s'_2 and s'_2 (= s_3) for each possible state of the dyad between event sender a_i and receiver a_j in the private message communication network x^m. In the figures on the left a crossed arc indicates that this tie is explicitly missing in the measured dyad. All other directed ties have a value >0.

dyad state	s_1	s_2	s_3	s'_1	s'_2	s'_3
	1	1	1	0	0	1
	1	0	0	1	0	0
	0	1	0	0	1	0
	0	0	0	0	0	0

Table 10
The two sub tables show estimates of model IV-3 and an equivalent model IV-3′ in which the statistics s_1, s_2 and s_3 were replace with statistis s'_1, s'_2 and s'_3 (equals s_3) as introduced in Eqs. (18) to (23). The sum over all parameters from model IV-3 is 10.363 which equals the parameter $\hat{\beta}'_3$ of statistic s'_3 in model IV-3′.

Model IV-3	
$\hat{\beta}_1$	7.967
$\hat{\beta}_2$	8.380
$\hat{\beta}_3$	−5.984

Model IV-3′	
$\hat{\beta}'_1$	7.967
$\hat{\beta}'_2$	8.380
$\hat{\beta}'_3$	10.363

defined in Eq. (14) as $p_{ij}^{?}(x;\beta)$. Compared to the base line decision with no dyadic structures it is θ times higher assuming that all other structures influencing the decision are equivalent:

$$\begin{aligned} p_{ij}^{?}(x;\beta) &= e^{\beta_1 s_1+\beta_2 s_2+\beta_3 s_3} \times \frac{e^{\beta_4 s_4+\dots}}{c^{+}} \\ &= \theta \overbrace{e^{\beta_1\times 0+\beta_2\times 0+\beta_3\times 0}}^{=1} \times \frac{e^{\beta_4 s_4+\dots}}{c^{+}} \\ \Leftrightarrow \theta &= e^{\beta_1 s_1+\beta_2 s_2+\beta_3 s_3} \end{aligned} \tag{24}$$

As stated before, people in the dataset tended to re-use ties, they tended to reciprocate and they even more tended to communicate within stable bi-directional communication patterns. In model IV-3, for example, the existence of a re-usable tie that was not part of a bi-directional communication structure (row two in Table 9) increased the probability times $\theta = e^{7.697} = 2201.73$ compared to a receiver without any dyadic structure.

If there was an incoming message tie from an actor the sender had no outgoing tie to (as in row three of Table 9), the probability for choosing this actor was $\theta = e^{8.380} = 4359.01$ times higher than in the base line model.

If there were both an incoming and an outgoing tie (row one in Table 9), the probability for choosing such a receiver was $\theta = e^{7.697+8.380-5.984} = e^{10.363} = 21,099.40$ times higher which is more than in one of the dyads with just one private message tie.

For the used dyadic statistics holds: Only if the negative value of parameter *Bi-directional communication* would have had a higher absolute value than one of the two other dyadic structures, we could have inferred that certain "unclosed" dyads were preferred over complete communication dyads. This was not the case in any of the models. This implies that bi-directional communication with repeated message writing in both directions was preferred to short conversations with just one private message written in each direction. Still, repeated message writing to the same receivers without reciprocation and reciprocating incoming message events *once* were very important predictors of receiver choice behavior. We argue that all these observations indicate that private communication in the dataset was not only functional, like to give additional information about questions or to say "thank you" if a question was answered but has a (dyadic) social component.

The absolute values of the dyadic parameters differ between the window. While *Re-use of ties* is similar in the best models of the three phases but decreases slightly (from 8.574 (0.141) to 7.516 (0.243)), *Reciprocity* significantly increases over time. The estimate is 4.860(0.471) in phase II, grows to 6.234(0.335) in phase III, and hits 8.015(0.248) in phase IV. Together with the only slightly decreasing statistic *Bi-directional communication* (from −4.448(0.564) in phase II to −5.763(0.293) in phase IV) we conclude that actors tend to communicate bi-directionally more and more: Compared to choosing a potential receiver that is not connected to the sender in the message graph, the probability for choosing a receiver connected bi-directionally was 7990.43 times higher in phase II, 9330.09 times higher in phase III and 17, 465.80 times higher in phase IV. We found those values by adding all three dyadic endogenous statistics in the best fitting models of each phase.

The observed increase of bi-directional communication is underlined by the fact that the relative importance of these parameters in the models increased. *Reciprocity* and *Bi-directional communication* were only the fifth and sixth statistic included in phase II. In phase III they were included as fourth and fifth statistic. In phase IV, finally, the three dyadic statistics were the three most important variables in the model. This means that dyadic structures were observed (and could be well explained) in an increasing number of cases. However, this change of relative importance could also be related to an interaction effect that we did not measure.

This supports the hypothesis that actors increasingly write messages within closed dyads. Partly, this effect can be explained with the higher rate of written messages per person that prevents a decay of ties. However, this effect alone is probably less significant, as the minimum time before a tie is removed from the dataset is more than 6 weeks and in most cases there is an earlier tie update if actors communicate bi-directionally. The absolute value of ties was not considered as long as the tie was not removed from graph x^m because it had decayed under a certain threshold. We argue that these observations indicate an "emergence" of social (in contrast to functional) behavior over the life-time of the Q&A community.

The three dyadic structures are the only statistics in this analysis that are rather independent from the state of the process. They are not influenced by varying factors like network density or number of active actors. They are also not strongly affected by the general activity in the dataset. So, we can always compare these parameters to a decision without the corresponding structure and thereby give an interpretation of the absolute values and its changes over time as long as possible interactions with other effects are kept in mind.

This is different with the endogenous statistic *Common contacts*. It was significantly positive (with varying significance levels) in all models. This means that actors tended to communicate with others who they had (many) common contacts with. The absolute parameter is harder to interpret. The probability for choosing a receiver increases in model II-7, for example, with each additional common contact by 11.4% ($e^{0.1805} = 1.114$).

However, in this model we implicitly assume that changes of the probability depend linearly on the number of common contacts. This is probably not the case. So, if we observe different numbers of common contacts in the local environments in different windows of the event stream the absolute value of the estimate is influenced by that fact and is therefore less straightforward to interpret.

The rank of the *Common contacts* statistic decreased from rank 4 in phase II to rank 6 in phase III to rank 7 in phase IV. The reason could be an interaction with the dyadic statistics with increasing rank. The inclusion of this parameter, however, never had a huge effect on the estimates of the dyadic statistics (or any other statistic), so we assume that the interaction is not too strong. In general, however, the existence of common contacts was an important predictor for communication choices. We argue that this indicates a social component in private communication behavior on the platform.

5.2. Two-mode statistics measuring question affiliation

Whenever one of the five two-mode statistics was significant at all it had a positive weight in almost all models. Only three times (in models II-7, II-8 and III-9) structures were insignificant. We observed a tendency for communication with askers or question responders and also a high tendency for communication between actors connected to the same questions in any way. This supports the hypothesis that private communication on the platform was also driven by question affiliation – affiliated receivers were preferred over others, especially if the sender was connected to the same question. The more question affiliations were counted, the higher was the probability for choosing the corresponding receiver.

The only exception is the statistic *Responder writes responder* in the last model IV-9. Here, a negative effect was observed. Once again, the reason could be an interaction with other effects. We observed that in model IV-9 the significance of the estimates of *Message to responder* and *Asker writes responder* increased (compared to model IV-8) when the negative effect *Responder writes responder*

was included. This is similar to what happened in case of dyadic, endogenous statistics.

The rank of the two-mode statistics changed a lot between the three analyzed windows. This is not surprising, as these structures are not independent. We cannot say whether the general tendency for certain types of question related private messages increased or decreased over time. All we learned was that in general the endogenous structures explained more in the last phase compared to the included two-mode structures. Still, beside *social* aspects in private communication choices question affiliation as a rather *functional* parameter was important for explaining private communication choices in the analyzed web community.

6. Conclusions and further research

In this paper the structural dynamics of private message communication in a German speaking Q&A community were analyzed. We introduced the dataset and the event stream and defined four different phases in the community development. A generic actor oriented Markov process model was introduced that can be applied to describe event formation in social environments. To demonstrate the application, we analyzed the sending of private messages within the Q&A community. The model was constructed as a three-level decision process. First, actors were assumed to decide about the time of private message sending based on individual Poisson rates. Second, in case of sending a private message they were assumed to choose whether to send an event to a currently active actors. This decision level was included to take into account that only a smaller subset of all actors was active in the community at the same time and would therefore be considered as a communication partner. The third decision is about the choice of private message receivers. This last decision was modeled as a multinomial logit model. Different endogenous communication structures and two-mode question affiliation structures were included as independent variables. We estimated different models (using a newly developed software package) to learn how these structures influence the decision about receivers of private messages. We also tried to find out whether we could identify differences in the different phases of the community.

It turned out that private communication dynamics in the analyzed community depended on dyadic and triadic endogenous structures in the private message graph, but also on two-mode question affiliation structures of senders and receivers. We found, for example, a high tendency for repeated private communication with the same actors, for bi-directional private communication, for triadic private communication structures and for choosing receivers that are answering or asking questions.

It could be shown that the estimates are slightly different in the different phases of the community and explain a different amount of the overall variance. Dyadic, endogenous effects seemed to get more relevant in the later phases. We learned that private communication in the analyzed community was driven both by social structures and functional aspects.

We could show how the proposed model framework can be applied on big event streams with different types of events and modes. Possible extensions of this framework were mentioned. Network structures, for example, can also incorporate the values of ties, actor attributes or multi-network structures. The multi-level decision process can be extended by decisions about different event types or event intensities. The simple actor activity rate used in this paper can also be parameterized to model, for example, the influence of structures, attributes or time on actor activity. Due to the richness of event stream data, the multinomial decisions could be estimated on an individual level if the research question was targeting individual behavior patterns instead of general group behavior. The current model only describes a small part of the overall dynamics in the Q&A community. It could, for example, be extended to measure co-evolutionary dynamics of private messages, questions and answers.

In future work we plan to apply a more structured model fitting algorithm. There is a huge number of possible independent variables in structural network models with many possible interactions. It would be interesting to find an algorithm that uses structural dependencies in the graphs to explore the space of possible network structures in a more sophisticated way.

Furthermore, we want to test the methodology on more, interesting datasets to learn more about robustness, interpretation of results and good model fitting strategies. Some of the mentioned extensions of the model framework may make sense when modeling different event stream datasets.

One advantage of event stream analyses is that the analyzed periods do not have to be predefined by an experimental setting but can be chosen ex post. If it was possible to define standardized structures that do not strongly depend on how the networks look like at a certain point in time, sliding window analyses could be applied to reveal the periods where structural breaks or slow changes in the underlying structural dynamics occur.[3]

Acknowledgements

The first author acknowledges support from the Deutsche Forschungsgemeinschaft (DFG), Graduate School IME at Karlsruhe Institute of Technology.

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 215453 – WeKnowIt.

The authors thank Michael Ovelgönne and Otto Allmendinger for supporting the data preprocessing.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723, 12.

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

Butts, C.T., 2008. A relational event framework for social action. Sociological Methodology 38 (1), 155–200.

Cramer, J., 2003. Logit Models from Econometrics and Other Fields. Cambridge University Press, Cambridge.

Deuflhard, P., 2004. Newton Methods for Nonlinear Problems. No. 35 in Springer Series in Computational Mathematics. Springer-Verlag, Berlin/Heidelberg.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1 (1), 54–77.

Greiner, W., Neise, L., Stöcker, H., 1993. Thermodynamik und Statistische Mechanik, 2nd ed. Vol. 9 of Theoretische Physik. Frankfurt am Main, Verlag Harri Deutsch, Thun.

Hosmer, David, W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press Inc., New York, pp. 105–142, Ch. 4.

Miller, A., 2002. Subset Selection in Regression, 2nd ed. Vol. 95 of Monographs on Statistics and Applied Probability. Chapman & Hall, Boca Raton.

Robins, G., Pattison, P., Kalish, Y., Lusher, D., 2007. An introduction to exponential random graph (p^*) models for social networks. Social Networks 29 (2), 173–191.

Snijders, T.A.B., 2001. The statistical evaluation of social network dynamics. In: Sobel, M., M.P., B. (Eds.), Sociological Methodology, vol. 31. Boston and London, Basil Blackwell, pp. 361–395.

Snijders, T.A.B., 2005. Models for longitudinal network data. In: Carrington, P., Wasserman, J.S.S. (Eds.), Models and methods in social network analysis. Vol. 27 of Structural Analysis in the Social Sciences. Cambridge University Press, New York, pp. 215–247 (Chapter 11).

[3] Some helpful remarks by two anonymous reviewers contributed to this idea.

Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S., 2006. New specifications for exponential random graph models. Sociological Methodology 36 (1), 99–153.
Stadtfeld, C., 2010. Who communicates with whom? Measuring communication choices on social media sites. In: In: Proceedings of the 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 564–569.
Stadtfeld, C., Geyer-Schulz, A., Waldmann, K.-H., 2010. Estimating event-based exponential random graph models. In: Dreier, T., Krämer, J., Studer, R., Weinhardt, C. (Eds.), Information Management and Market Engineering. Vol. 2 of Studies on eOrganisation and Market Engineering. KIT Scientific Publishing, pp. 79–94.
Wang, P., Sharpe, K., Robins, G.L., Pattison, P.E., 2009. Exponential random graph (p^*) models for affiliation networks. Social Networks 31 (1), 12–25.
Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and p^*. Psychometrika 61 (3), 401–425.