Who Communicates With Whom?
Measuring Communication Choices on Social Media Sites

Christoph Stadtfeld
IME Graduate School
Information Services and Electronic Markets
Karlsruhe Institute of Technology (KIT), Germany
Email: christoph.stadtfeld@kit.edu

Abstract—People on social media web sites are connected in many different ways. Communication networks are one important element. When analyzing the communication choices in such social media environments, the existence and dynamics of these different networks should be taken into account. The paper proposes a Markov process framework which includes an exponential random graph model that allows to estimate in how far communication depends on previous communication between actors, on affiliation with media items or other objects, on other actor networks and on actor attributes. A new software was developed to estimate the influence of certain structures in these networks on communication choices.

Keywords—event data; communication networks; social network analysis; social media websites; Markov process; ERGM

I. INTRODUCTION

On social media platforms, people are connected with different people and many different types of media. Some of these media items may be linked to others. The links between people can be formalized by the platform design (like friendship on social network sites) or be defined by informal communication ties. Also, actors are indirectly linked by sharing media items. In such a set of networks, the dynamics of communication, media (or other entity) affiliation and formalized relationships cannot be analyzed separately, because they probably influence one another significantly.

The aim of the paper is to present a framework that allows to describe communication in social media environments. It can be used to define models that explain in how far communication depends on existing communication networks, on media affiliation of actors, on the existence of other types of links, or on actor attributes. Network structures that surround sender and recipient of communication messages are measured and evaluated for observed data streams with a non-linear regression model. A Maximum Likelihood estimation returns parameters for these structures that can then be interpreted. Framework and estimation process are part of a new software.

The proposed framework uses an exponential random graph model (ERGM) to describe the choice of communication recipients. This class of models is commonly used in social sciences to analyze static networks or networks based on panel data as part of actor driven models (see [1], [2], [3], [4]). In this paper the underlying data, however, is not panel data but event data. Events are dyadic, time-stamped, directed interactions from actors to another entity. These recipients can be actors as well or other nodes, like media items. Examples for events are phone calls, e-mails, new friendship links, a new movie upload or an actor tagging an image. This type of data is increasingly available and some new approaches try to describe network dynamics based on this data type (see [5], [6], [7]). Analyzing these very detailed event streams allows better insights into the dynamics of communities, like those in social media environments.

The framework has been implemented in Java and allows to estimate how structural conditions in different networks influence the choice of communication recipients. Dependent on the research question or the business application, dynamic models can be defined. The general process is as follows: After transforming the data into a valid input format (see section II-A), the user defines how events change graphs that represent to be defined networks (section II-B). Typical rules change ties or tie values if an event takes place and determine how networks deterministically evolve over time. Then, the available data and the rules are mapped to a list of static graphs for those points in time when communication events take place. For these events, the network structures that surround sender and recipient can then be measured (section II-C) for all active recipients within the analyzed community. As part of a stochastic Markov process (section III-A), the decisions about event recipients are described with a probability function that depends on these structures (section III-B). They are weighted with a set of parameters for that the Maximum Likelihood optimum can then be found (section III-C). These parameters then indicate whether structures increase the probability for communication. The general activity of actors and the intensity distribution of events can as well be estimated with the software. Some observations about the computational complexity are given. Section IV concludes this paper and shortly discusses future directions of research.
II. MEASURING EVENT NETWORKS

The goal of the later introduced model is to evaluate network structures that influence communication. In this section the underlying data format is explained, then it is shown how it can generically be transformed into a set of graphs and finally how local environments of these graphs can be described with network statistics.

A. Communication Streams

The analyzed data format is a stream of dyadic, time-stamped and directed events. Formally, it is named \( \Omega \) with elements \( \omega_1, \ldots, \omega_m \in \Omega \). Each event \( \omega \) is a quintuplet

\[
\omega = (time, type, sender, recipient, intensity).
\]

Examples for events are calls between actors or the emergence of new links between an actor and an image on an image community web site. Each event has a timestamp, a type (like call or new image link), a sender (in both cases an actor), a recipient (an actor or like in the image case a media entity) and an intensity (like the length of a call). The values of these entries of a certain event are w.l.o.g. addressed as \( \omega.sender \).

As each event has a timestamp, the elements of stream \( \Omega \) can be sorted chronologically. \( \Omega = \{\omega_1, \ldots, \omega_m\}, \forall \omega_k, \omega_l: k < l \Leftrightarrow \omega_k.time \leq \omega_l.time \). The time between two consecutive events \( \omega_{k-1} \) and \( \omega_k \) is noted as \( \delta_k \):

\[
\delta_k = \omega_k.time - \omega_{k-1}.time. \tag{2}
\]

A part of an exemplary event stream is given in table I.

<table>
<thead>
<tr>
<th>( \omega.time )</th>
<th>( \omega.type )</th>
<th>( \omega.sender )</th>
<th>( \omega.recip. )</th>
<th>( \omega.intens. )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 13:01:32</td>
<td>new image</td>
<td>Actor1</td>
<td>Image1</td>
<td>–</td>
</tr>
<tr>
<td>2 15:10:51</td>
<td>message</td>
<td>Actor1</td>
<td>Actor2</td>
<td>important</td>
</tr>
<tr>
<td>3 15:17:00</td>
<td>likes image</td>
<td>Actor2</td>
<td>Image1</td>
<td>–</td>
</tr>
<tr>
<td>4 15:18:16</td>
<td>call</td>
<td>Actor2</td>
<td>Actor3</td>
<td>144 sec.</td>
</tr>
<tr>
<td>5 15:19:19</td>
<td>likes image</td>
<td>Actor3</td>
<td>Image1</td>
<td>–</td>
</tr>
<tr>
<td>6 15:21:31</td>
<td>message</td>
<td>Actor3</td>
<td>Actor1</td>
<td>normal</td>
</tr>
<tr>
<td>7 16:12:42</td>
<td>new friend</td>
<td>Actor1</td>
<td>Actor3</td>
<td>–</td>
</tr>
</tbody>
</table>

Table I

PART OF AN EXEMPLARY DATA STREAM WITH COMMUNICATION AND MEDIA AFFILIATION EVENTS AND ONE RELATIONSHIP EVENTS

There are three general groups of events: communication events, media affiliation events and relationship events. Furthermore, these groups can consist of different types. In case of communication, typical types are private mail, chat messages or calls. Typical media affiliation events are creating new media entities, establishing new connections or dissolving a connection. Relationship events are, for example, the change of friendship ties in a social community. In table I rows 2, 4 and 6 contain communication events, rows 1, 3 and 5 contain media affiliation events and row 7 a relationship event.

Communication events and relationship events occur between actors \( \in A \), the set of all actors. Communication events usually have an intensity, like the length of a call or the importance flag of a message (see column 5 and rows 2, 4 and 6 in table I). Media affiliation events have a sender \( \in A \) and a recipient \( \in M \), the set of Media entities.

B. Transforming communication events into networks

Communication is influenced by the state of different networks, like a communication network representing recent communication and networks that represent media affiliation of actors. Transformation rules are necessary to define how event streams create and change networks over time. Communication events, for example, increase tie values in communication graphs, while media affiliation events connect or disconnect actors with media entities.

The different networks are represented by a set of graphs \( X = \{X_1, \ldots, X_m\} \) with realisations \( x = (x_1, \ldots, x_m) \). \( x^k \) is a realisation of random variable \( X \) after event \( \omega_k \). An entry \( (i,j) \) of matrix \( x_t \) describes the tie from \( i \) to \( j \) and is noted as \( x_{tij} \). Each graph consists of a set of nodes \( N \), \( N = \{N\} \) in case of communication and relationship graphs and \( N = \{N \cup M\} \) (the sets of actors and media entities as introduced in section II-A) in case of media affiliation graphs. Media affiliation graphs are usually binary, bipartite and undirected graphs while communication graphs are weighted, directed and non-reflexive graphs. Relationship graphs are usually binary graphs with – like in case of communication graphs – a node set \( N = \{N\} \). They can be both directed or undirected.

The state \( x \) is changed by a stream of events using an ordered set of rules \( R \). The graphs are defined for each point in time, but only the points in time directly after an event are evaluated in a model. The state \( x^{k+1} \) directly after the \( (k + 1) \)-th event can be calculated recursively using the state \( x^k \) after the previous event, the last event \( \omega_{k+1} \) and its previous event \( \omega_k \):

\[
x^{k+1} = R(x^k, \omega_{k+1}, \omega_k). \tag{3}
\]

Most rules are triggered by events. Examples are updating tie values in the communication network or the establishing or dissolving of connections between actors and media items. Other rules express the decay of tie values over time. Function \( R \) from equation 3 includes an ordered set of such rules. Decay rules are usually applied before the event triggered update rules of an event \( \omega_{k+1} \).

C. Influences on Communication

Communication in networks is usually not equally distributed over all possible communication ties but depends on existing network structures of the set of graphs \( X \) and on actor attributes. It can be shown that these structures help to explain parts of the observed communication. It is possible to estimate the influence
of network structures on each event type. In this paper, however, only the dynamics of communication events are examined. Some structures might increase the probability of communication while others might decrease it. Usually, the values of all equivalent structures in which sender and recipient are embedded is summed up to get a structural statistic. Structures in weighted networks can be measured by choosing the weight of the weakest tie.

There are many possible influencing factors on communication that can be expressed as structural network statistics. For example, people often tend to interact with whom they have interacted before. In social media environments, existing “formalized” relationships or the affiliation with the same media entity can help to understand communication behavior. People might start to talk to each other because they like the same pictures on a picture sharing web site. Communication can be explained with previous communication - these endogeneous variables are explained in section II-C1. The dependence of communication on media affiliation (in general: entity affiliation) is an exogeneous effect and is discussed in section II-C2. The dependence of communication on other networks than the communication network, like friendship networks, is discussed in section II-C3, the dependence on actor attributes in section II-C4. Other effects that combine effects of the mentioned types are also possible and are introduced in section II-C5.

1) Endogeneous Communication Statistics: Parts of communication can be explained endogenously by previous communication. Figure 1 shows five different structures that may surround sender and recipient of a communication event. Of course, many others are possible. An overview of possible (undirected) three-node structures is given in [8, p.566]. The idea to test only “simple” structures follows from the introduction of Markov graphs [9] and advancements of this approach.

Arrows express that some communication has already been observed on the tie, the value of these weighted structures is not given. The letter symbol indicates the tie on which the current event takes place and its direction. Sender (the upper left actor) and recipient (the upper right actor) are always part of the measured structure.

Figure 1(a) straightforwardly counts the value of the new tie. A high importance of this structure (weighted with a high positive value in the later introduced regression model) would indicate that actors tend to reuse existing communication paths. If it was negative, it could be concluded that people like to have as many contacts as possible. The opposite effect would measure the outdegree of the sender. Figure 1(b) measures the tendency for reciprocity in communication. Figure 1(c) includes a third actor and expresses the tendency of communicating with actors who have a high indegree. Similarly, the outdegree of the recipients in the dataset could be measured. Depending on the actually observed communication network these concepts can be named popularity and activity or Communication Hub and Communication Authority. Note, that the statistics count all observed structures over all third actors, so it really returns an equivalent to an absolute degree measure. The final two structures in figures 1(d) and 1(e) measure whether there is a tendency for communication with actors that are inside a triangle structure: either a transitive triangle or a complete circle. Note, that equivalent structures with changed tie directions are possible and can also be counted to get the statistics value. As mentioned before, there are many possible structures. When defining a model, sub graph structures (like transitivity includes popular recipient) should always be included.

All these structures are typically implemented as binary structures in longitudinal actor-oriented models like SIENA (see [10]). There is also new research on including (binary) multi network statistics in ERGMs (see [11]) similar to the simple cases in figure 3.

2) Entity Affiliation Statistics: In social media environments, communication may depend on media items that actors are connected with. For example, people may talk about videos, images or texts, although they have no common communication history. Therefore, it is proposed to test the effect of these structures as well if the data is available. Figure 2 shows two different statistics that could be tested for different media or content types.

Interpretations of these effects strongly depend on the context. The second structure in figure 2(b) expresses whether there is an increased probability for communication if people are connected to the same media entities. It can be interesting to compare the effects of different types of media.

3) Other Actor Network Statistics: Communication is also influenced by other networks beside the communication networks. Two examples are given in figure 3. The time
since the last event on a tie may be encoded in a graph. Then it is possible – without violating the Markov property of the later introduced model – to test whether time since the last event influences future communication. Communication often depends positively on “formalized” networks like friendship relations on social media sites.

4) Actor Attribute Statistics: The exemplary structures in figures 4(a) and 4(b) can be part of a model, if the effect of attributes on communication is of interest. It can, for example, be tested whether similar sex or age has a positive or negative effect on the choice of recipients.

5) Combined Statistics: All the introduced effects (and other related effects) can be combined arbitrarily. Combination of statistics can well be used if communication is only analysed for actors in certain structures (like media affiliation of recipients) or with certain attributes (like when genders shall be compared). Friendship, for example, can also be used as a variable for combined structures, in case only communication between friends is supposed to be analyzed. But it should be kept in mind that parameters are not linearly independent due to subgraph relations. Especially in these cases, it is recommended or in some cases it is even mandatory to test sub structures of statistics as well. When including very complex communication statistics this might therefore lead to problems with the model specification.

III. Model Framework

Section II explained how to measure network structures that influence the probability for events between certain actor tuples. A Markov process model is introduced in section III-A that describes the occurrence of communication events as a four level process. These processes are described – using other underlying data structures – as a two level process by Snijders [12]. A less generic but similar three level process can be found in [13]. In this case, the model assumes that each actor has a certain individual propensity to communicate. Then, he chooses whether he wants to communicate with somebody who has recently communicated or has been active in any other way (new media affiliation, new friendships): He decides whether to communicate with an active actor. This second level decision can be introduced to reduce computational complexity in large networks with a high rate of non-actives. If the recipient is active, the choice of the actual recipient depends on network structures as introduced in section II-C (third level decision). Otherwise, the sender just randomly picks a recipient. The last decision is about the event intensity. The model assumes that all four decisions in the process are mutually independent. The third part of the process, the choice of communication event recipients is examined in more detail in section III-B. Some facts about the estimation process are presented in section III-C.

A. Markov Process of Communication Events

The occurrence of communication events in an event stream is described with a Markov process (also called continuous-time Markov chain) (see [14]). Vector \( x \) (see section II-B) includes matrices that represent the state of different graphs and is also the state of the process. Due to time dependent subprocesses (decay of communication ties, time count ties) there is a constant change of the state \( x \). However, in networks with a high activity in event streams, for short time spans they can be assumed to be stable. The process changes that are triggered by communication events, are modelled with a Poisson parameter \( \lambda(\omega; x, \beta) \). As the time stamp of events does not matter any more (it follows from the exponentially distributed time spans of the Poisson process), the quintuplet \( \omega \) is in case of observed communication events replaced by a quadruplet

\[
\omega = (\omega_{\text{sender}}, \omega_{\text{recipient}}, \omega_{\text{type}}, \omega_{\text{intensity}})
= (i, j, \tau, \eta).
\]

The tendency for occurrence of \( \omega \) on \( x \) is expressed by a Poisson parameter \( \lambda(\omega; x, \beta) \) (\( \in \mathbb{R}^+ \)):
\[ \lambda(\hat{\omega}; x, \beta) = \begin{cases} 
\rho(i, \tau)p^+(j; i, x, \beta)p^+(\eta; \tau), & j \in A^+ \\
(1 - p^+)(1 - \frac{1}{|A|})p^+(\eta; \tau), & j \in A^- . 
\end{cases} \tag{5} \]

Every actor \( i \) has a general activity for events of communication type \( \tau \). This is expressed by the actor related Poisson rate \( \rho(i, \tau) \). If an actor starts a communication event of a certain type, he has three more decisions to make which are expressed by the probabilities \( p^+ \) (is recipient active?), \( p^+ \) (the recipient is active: who is it?) and \( p^+ \) (for the given communication type choose an intensity).

\( p^+ \) is a Bernoulli probability. If \( p^+ >> (1 - p^+) \), most of the recipient choices \( p^+(j; i, x, \beta) \) events can be explained with the model. Otherwise, the set of active actors should be defined differently.

Given that \( i \) is the event sender, the process state is \( x \) and a certain parameter vector \( \beta \), \( p^+ \) returns the probability for choosing actor \( j \) over all other possible and active recipients.

Finally, \( p^+(\eta; \tau) \) determines the intensity \( \eta \) of the event type \( \tau \). The intensity may depend both on actor attributes of sender and recipient and network structures. However, it has to be made sure that these influencing factors are independent from the attributes that have an influence on the first three decisions as the three probabilities are mandatorily mutually independent.

Summarized, the four level process for communication events is as follows:

1) Some actor \( i \) gets active with event type \( \tau \) dependend on personal activity rate \( \rho(i, \tau) \)
2) Is the recipient active? (\( j \in A^+ ? \))
   No: Equal distribution for each non-active recipient  
   Yes: Choose recipient \( j \) dependend on \( x \), \( i \) and \( \beta \) (\( p^+ \))
3) Choose the communication intensity (\( p^+ \))

Every user has a certain activity. Some communicate a lot, others hardly ever send events to others. However, when starting a communication event with an active recipient \( j \in A^+ \), the choice of recipient only depends on the network structures that embed sender and recipient and does not depend on the user’s general activity. Also, the intensity of sent messages is independent from the user’s general activity and the chosen recipient. However, this does not mean it cannot depend on tie attributes, for example. It only means that these attributes have to be independent from those used to explain the choice of recipients or the user’s activity. As in social media environments like social network sites often only a small subset of users is active at the same time, the decision making of the stochastic process is extended by one further case discrimination. When getting active, each user decides whether he wants to send the event to a currently active user or not. Possible criteria are whether an actor has recently communicated or has new media affiliations.

So, only those communication events are observed in detail that are sent to active users. This reduces the computational complexity.

Probability \( p^+ \) is of particular interest as it evaluates the network structures sender and recipient are embedded in. The next section takes a closer look at this probability function.

**B. Chosing event recipients**

Probability \( p^+ \) is a so called exponential random graph (ERGM) probability function (see [1]) and decides about choice of communication event recipients. It is the core of the process: The fact that communication depends on surrounding network structures is only described with this element of the Markov process. It is defined as follows:

\[
p^+(j; i, x, \beta) = \frac{1}{c^+} \exp \left( \beta^T s(R(x, \hat{\omega}), i, j) \right) .
\]

\[
c^+ = \sum_{k \in A^+} \exp \left( \beta^T s(R(x, \hat{\omega}\{j \rightarrow k\}), i, k) \right) \tag{6}
\]

It models the choice of recipient \( j \) over all other active actors, given that \( i \) is the sender and \( x \) the underlying graphs. Vecor \( \beta \) evaluates a statistics vector \( s \) that includes counts of certain structures (introduced in section II-C). As \( s \) evaluates only local structures of sender and recipient, it depends on some \( x \), \( i \) and \( j \) (\( s(x, i, j) \)). Both vectors \( \beta \) and \( s \) have the same dimension. \( x \) is not directly evaluated, but the state \( x' \) after a certain event \( \hat{\omega} \) has taken place. Transformation function 3 is used. \( c^+ \) is a normalizing constant which assures that \( p^+ \) is a proper probability distribution. It evaluates all those graphs that could have been observed given that \( i \) decided to pick any possible actor \( k \in A^+ \). \( \hat{\omega}\{j \rightarrow k\} \) is therefore equal to \( \hat{\omega} \) except for the fact that the actual recipient \( j \) was changed to \( k \in A^+ \).

The inner part of the function consists of an evaluation of the observed structural parameters of vector \( s \). Each is multiplied with its corresponding weight in \( \beta \) and the sum is exponentially transformed. For positive entries in \( \beta \) holds that the existence of corresponding structures increases the probability if these structures are underrepresented in the constant \( c^+ \). For negative entries holds the opposite. Constant \( c^+ \) represents the randomly possible outcomes. A zero vector \( \beta \) would indicate an equally distributed probability for all recipients. Vector \( \beta \) is the part of probability \( p^+ \) that is not determined and has to be estimated based on observed event data (section II-A) and a set of transformation rules (section II-B). Section III-C briefly explains the estimation procedure.

**C. Estimating Parameters for Choice of Recipients**

When analyzing an event stream, the best estimators \( \hat{\beta}_k \) for the \( k \)-th statistic in parameter vector \( \beta \) are of interest.
They weight different network structures in the neighborhood of event sender and recipient that have in influence on the choice of recipients. Significantly positive or negative estimators indicate that the observed choices differ from random choices as explained in section III-B. The absolute value is harder to interpret but generally high absolute values express particularly strong effects (either negative or positive).

Probability function $p^*$ can be interpreted as a non-linear regression model (see [15]) in which the observed network statistics $s$ are the independent variables and the observed decision is the dependent variable (see [10]). As all recipient choices are assumed to be independent, the likelihood function depending on $\beta$ is the product of all decisions of equation 6. For a given event stream it can be maximized to estimators $\hat{\beta}$ of the regression (see [13], [16]).

IV. SUMMARY AND CONCLUSIONS

This paper aimed on presenting a framework which can be used flexibly to analyze communication in social media environments. It allows to define concrete models that describe how actors decide about communication recipients depending on the network structures sender and recipient are embedded in for certain data streams. A newly developed software can be used to efficiently estimate maximum likelihood estimates, even in big networks. It also provides tests for statistical significance.

In the future, the framework will be extended to describe complete event streams and not just the communication parts of it. It could be tested in how far there are co-evolutionary effects in the dynamics of different event types. Also, event stream analysis allows a detailed look on change in network dynamics by applying the software on sub streams of the available data. Still, this “change of dynamic effects” framework requires a statistical description that allows testing the significance of the detected changes. So far, only visualizations are possible. The software allows to estimate the influence of network structures on communication event dynamics. However, the absolute values of the estimators are hard to interpret as they depend on the general activity in the event stream and structural network parameters, like density. This also leaves room for further research.

There is a broad range of possible applications of the proposed framework, as event data is available in a lot of fields, e.g. in many social media environments. For example, marketing related and sociological research questions can be posed – the estimated model parameters may then help to better understand dynamics of communication in social media networks.

REFERENCES


