Four factors influencing effectiveness in email communication networks

Ofer Engel[*]
Department of Management
London School of Economics and Political Science
Houghton Street, London WC2A 2AE, UK
o.engel@lse.ac.uk

Abstract

This paper develops a method to analyse email communication networks in terms of the effectiveness of emails to elicit a reply. Four factors are considered: the email's sender (sender effect), the email's recipients (recipient effect), specific properties of each sender-recipient dyad (dyad effect), and specific properties of the email (stimulus effect). A multilevel model is developed and applied to an email communication dataset. The fitted model suggests that dyad and stimulus effects are more important sources of variability in reply rates than sender or recipient effects. Moreover, the correlation between sender and recipient effects is found positive when taken in isolation, but vanishes when accounting for the dyadic effect. The strength of this method lies in the general and systematic way it provides to evaluate organization members and their ties with respect to their influence on others to reply. From a socio-theoretical perspective it operationalizes a co-evolution process that operates not only at the meso-level (nodes and dyads), but also between the meso-level and the micro-level (transactions).

1 Introduction

Email communication consists of chains of related social transactions. Each email may serve as an invitation or stimulus for the next email in the chain, possibly setting in motion a series of related emails that bounce back and forth between actors. At least for a certain period of time, this process may activate social ties and realize network configurations. This paper probes the mechanisms underlying this process, what could be seen as a 'collective action' of sorts.

The basic building block of such a process is the email that acts as a 'stimulus' for subsequent and related emails. Whether such a 'stimulus' is effective in marshalling a reply depends on various factors. Four factors are considered in this paper; (1) properties of the email's sender (sender effect), (2) properties of its recipients (recipient effect), (3) properties unique to each sender-recipient dyad (dyad effect) and (4) properties of the email that may or may not trigger a reply (stimulus effect).

The sender and recipient effects refer to the specific properties of the senders and recipients of the original stimulus email. A high *sender effect* is a property of the sender, whose emails tend to be highly effective in eliciting replies. This property may reflect a unique role or reputation of the sender, whose emails are rarely left unanswered. By the same token, a high *recipient effect* is an attribute of recipients who are generally more responsive than other actors in the network. The *dyadic effect* allows for variations between different dyads that cannot be attributed to single actors. Finally, the *stimulus effect* allows for specific features of the email itself to stimulate replies to a degree that cannot be reduced to the actors or dyads. A high stimulus effect sets an email apart from other emails sent by the same sender to the same set of recipients: it may reflect the email's unique

[*] Currently visiting at Harvard Faculty of Arts and Sciences (unofficial status).

content, its timing, or a signal indicating whether or not the sender is awaiting a reply. Whatever the reasons, the method proposed here allows the identification of highly effective (or ineffective) actors, ties and emails in a systematic manner.

In what follows, an well studied approach in network statistics is adapted to model reply rate as described above. The data is described and three different models are fitted to the data. The models partition variance in the reply rate into the four factors. The findings suggest that dyad and message effects account for a greater portion of the variance, than actor level effects. The paper concludes with a discussion of the general applicability of this method to map members of a network, their ties, and the emails they send and receive. It also touches on the theoretical significance of the method from a sociological/theoretical point of view.

2 Method

This section adapts the $p2$ model which is well known in the literature of statistics of social networks [1, 2, 3, 4, 5]. The relevance of the model to the problem at hand stems from its multilevel approach: the outcome variable is defined at the level of the dyad, for example, it could be a binary variable denoting whether a tie exists or not [1] or it could be a continuous variable denoting the strength of a tie[5]. The $p2$ model regresses the outcome on random and fixed effects at the level of the nodes. Since multiple dyads can be associated with each node, the dyad level is nested within the level of the node.

To include social transactions in the model, consider the following adaptation of the basic $p2$ model described above. Instead of having an outcome variable at the dyad level, the outcome variable is at the level of the single sender-recipient transaction [6]. Since multiple transactions may be exchanged within each dyad, transactions can be seen as nested within the dyad and within its actors. Specifically, the model developed below has a binary outcome variable reflecting whether or not a 'stimulus' email has prompted a reply from its recipients.

In the first and simplest form of the model, this outcome is regressed on the sender and recipient random effects. The second model adds an estimation of a dyad-level random effect to test how the reply rates vary not only between actors, but also between different dyads of which actors are part. The third and last model adds an estimation of the effect of each particular stimulus along with fixed indicators unique to the stimulus and the dyad.

The email communication data set used in the current study consists of a snapshot taken from a well documented version of the Enron email corpus [7]. A group of 71 highly active and well connected email users were selected from the dataset. Within this group 396 (unordered) dyads have been identified, such that its members have exchanged at least one email. The data set includes 2973 email messages sent within a span of three months. The number of recipients in every email ranges from one to eighteen, a large minority of the emails addressed to a single recipient. The number of replies identified in the dataset is 540, which means that the overall proportion of replies is 12.9%.

2.1 Modelling actor, dyad and stimulus effects

The outcome variable y_{ij} denotes the existence of a reply from the i^{th} recipient of the j^{th} email back to its sender (1 denotes a reply, 0 no reply). Fortunately, there is a rather plausible way to establish the existence of a reply. For every email in the dataset and each of its recipients we search for a subsequent email that is sent from that recipient back to the sender and bears a subject-line identical to that of the stimulus (ignoring prefixes like *'re:'* or *'fwd:'* or combinations thereof)[1]. The outcome variable y_{ij} is assumed to have a Bernoulli distribution:

$$y_{ij} \sim \text{Binomial}\ (1, \pi_{ij})$$

We assume a logit link function from the probability π_{ij}, related to the predictors X_{ij}^T specific to the outcome through a vector of fixed parameters β and four random effects consisting of the effect of

[1] It is of course possible that email recipients change the subject when they reply (a false negative), or hit reply on an email instead of searching for the address of the actor they would like to communicate with (a false positive)

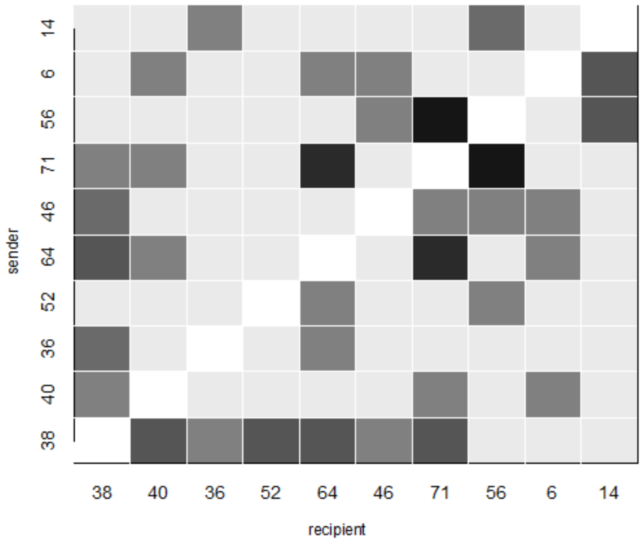

Figure 1: Aggregate rate of reply among a subset of the actors in the data set. The vertical axis denotes senders of stimuli emails, the horizontal axis denotes the recipients. The darker colors denote a higher rate of reply from recipients to senders. Note that the matrix has a weak but noticeable tendency to symmetry.

the stimulus email itself u_j^{stim}, its sender $u_{s[j]}^{sender}$, its recipient $u_{r[i,j]}^{recip}$ and the sender-recipient dyad $u_{d[r,s]}^{dyad}$ where j, s, r and d correspond to unique identifiers of the email, the sender, the recipient and the undirected dyad. The latter is subject to the constraint $d[r,s] = d[s,r]$.

$$\text{logit}(\pi_{ij}) = \text{logit}(\frac{\pi_{ij}}{1-\pi_{ij}}) = X_{ij}^T\beta + u_{s[j]}^{sender} + u_{r[i,j]}^{recip} + u_{d[r,s]}^{dyad} + u_j^{stim}$$

Where the two residuals for any particular actor k are assumed to be correlated (the inverse covariance matrix Σ^{-1} is assigned a prior Wishart distribution with two degrees of freedom). Random effects associated with different actors are not assumed to be correlated, and the other effects are likewise assumed to be uncorrelated.

The first and second models estimate only the random effects, but the third model includes three fixed effects: first, a dummy variable indicating whether recipient j was assigned to the *to* field or to the *cc/bcc* fields, where a higher higher of reply-rate is expected from recipients assigned to the *to* field. The second fixed effect is a count of the number of recipients in each email. A lower rate of reply is expected for emails with numerous recipients (such as in the case of 'bulk' emails). Due to the very skewed distribution of this count and its non-linear effect it has been binned, and the order of the bin was used as the predictor in the model. The last fixed effect is the total number of emails exchanged between the actors of each dyad, prior to the email in question. This count is used as a proxy for the strength of the tie between sender and recipient at the moment the email was was sent. 'Stronger ties' are expected to elicit a higher rate of reply from the actors involved.

3 Results and Discussion

The three models were estimated using Markov chain Monte Carlo sampling approach. All models were fitted using JAGS version 2.1.0 [8] running two parallel chains, discarding the first 3000 replicates and basing inference on the next 30000 for each chain. The findings presented in Table 1 provide strong evidence that the four factors are important sources of variability in the effectiveness of emails to elicit replies.

	model 1	model 2	model 3
Constant	-2.06 (0.140)	-2.55 (0.140)	-2.02 (0.270)
To Field			0.52 (0.220)
Number of email recipients (binned)			-0.31 (0.030)
Frequency of email exchange			0.02 (0.003)
σ_{sender}	0.73 (0.11)	0.54 (0.15)	0.47 (0.20)
σ_{recip}	0.71 (0.09)	0.54 (0.14)	0.50 (0.20)
$\rho_{send,recip}$	0.38 (0.15)		
σ_{dyad}		1.09 (0.14)	1.82 (0.31)
σ_{stim}			2.62 (0.52)

Table 1: Crossed multilevel models: Results

The findings in the first model demonstrate a substantial variation between actors, as well as a correlation between sender and recipient effects. This means that, at least in this specific dataset, actors who tend to reply to others also tend to elicit replies from others (this pattern can also be discerned from Figure 1). A comparison of the different actors in terms of their actor level effects is presented in Figure 2. Some actors appear to be particularly effective in triggering replies from others, and some are particularly responsive to others.

The second model suggests that the dyad effect better explains variation in the rate of reply than sender and recipient effects. In this model the estimation of the correlation between sender and recipient effects has become very unreliable and was dropped from the model.

The third model demonstrates that both emails and dyads are important factors governing the variability of the rate of reply, and that these are more important sources of variability than the properties of the actors. Furthermore, the fixed effects operate in the expected direction: (1) recipients in the *to* field are more likely to reply, (2) recipients of multiple recipient emails are less likely to reply and (3) actors bound by stronger ties are more likely to reply to each other. These findings are suggestive, but they open the theoretical question as to what are the precise social mechanisms that govern these effects and their strengths.

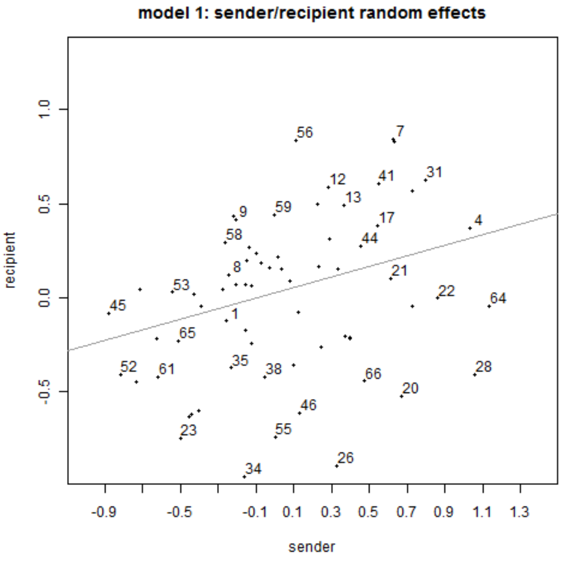

Figure 2: Comparing sender and recipient random effects.

4 Conclusion

This paper offers a general method to extract valuable information about latent properties of actors, ties and messages from email data sets. It allows the comparison within a level of abstraction (e.g., comparing actors as in Figure 2), and between levels of abstraction (e.g., comparing the effect of ties, nodes and stimuli). Substantially, that ties are a more important source of variability than nodes makes email effectiveness rather like a form of social capital, that 'inheres in the structure of relations between actors' [9] rather than being a property of the individuals themselves.

Conceptually, this method is innovative on two accounts: first, it considers not only how one actor is connected to another in a network, but also how one transaction triggers another in a sequence [10, 11]. Second, it suggests a way to develop models of co-evolution processes that operate not only at the meso-level (nodes and dyads) [12], but also between the meso-level and the micro-level (transactions).

Arguably, the distinction between social transactions and social ties is also at the very heart of the departure of 'computational social science' [13] from other studies of social networks. Whereas the former focuses on networks of interaction and communication [14], the latter focuses on social relationships (friendship, kinship etc) [9]. By combining dyads, actors and transactions into the same framework, this paper argues for the potential in bridging these two agendas.

References

[1] B J H Zijlstra, M A J Duijn, and T A B Snijders. Model selection in random effects models for directed graphs using approximated bayes factors. *Statistica Neerlandica*, 59(1):107–118, 2005.

[2] M A J van Duijn, J T Busschbach, and T A B Snijders. Multilevel analysis of personal networks as dependent variables. *Social Networks*, 21(2):187–210, 1999.

[3] M A J Duijn, T A B Snijders, and B J H Zijlstra. p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2):234254, 2004.

[4] B J H Zijlstra, M A J van Duijn, and T A B Snijders. The multilevel p 2 model. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1):42–47, 2006.

[5] T A B Snijders and D Kenny. The social relations model for family data: A multilevel approach. *Personal Relationships*, 6(4):471–486, 1999.

[6] Wouter de Nooy. Networks of action and events over time. a multilevel discrete-time event history model for longitudinal network data. *Social Networks*, 2010.

[7] J Shetty and J Adibi. The enron email dataset database schema and brief statistical report, 2004.

[8] M Plummer. Jags: Just another gibbs sampler, 2004.

[9] James S. Coleman. Social capital in the creation of human capital. *The American Journal of Sociology*, 94:S95–S120, 1988.

[10] D R Gibson. Taking turns and talking ties: Networks and conversational interaction. *American Journal of Sociology*, 110(6):1561–1597, May 2005.

[11] C T Butts. A relational event framework model for social action. *Sociological Methodology*, 38(1):155–200, 2008.

[12] T A B Snijders, C E G Steglich, and M Schweinberger. Modeling the co-evolution of networks and behavior. *Longitudinal models in the behavioral and related sciences*, page 4171, 2007.

[13] D Lazer, A Pentland, L Adamic, S Aral, A-L Barabasi, D Brewer, N Christakis, N Contractor, J Fowler, M Gutmann, T Jebara, G King, M Macy, D Roy, and M Van Alstyne. Computational social science. *Science*, 323(5915):721–723, February 2009.

[14] P R Monge and N S Contractor. *Theories of Communication Networks*. Oxford University Press, New-York, 2003.