The Micro-Foundations of Email Communication Networks

Ofer Engel

London School of Economics and Political Science

Department of Management
Information Systems and Innovation Group

Thesis submitted for the degree of

PhilosophiæDoctor (PhD)

2013 June

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it.) The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without the prior written consent of the author. I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

Abstract

The popular and scientific literature has been discussing the advent of 'big data' with a measure of excitement and apprehension. For the first time in history, it seems, every breath we take, every move we make, someone's watching us. But beyond their unprecedented volumes and the anxieties they raise, new communication data have a less obvious aspect, in so far as they are (arguably) of a fundamentally different *kind*, compared to traditional network datasets.

Traditionally, social network data describe relationships between individuals; quasi-static social ties such as friendship, trust, kinship and employment relations. But when they are used to model digitally mediated communicative transactions, the connections are of a different nature. Instead of representing stable social ties, transactions (such as emails, text messages and phone calls) constitute sequences of short-lived events, with each transaction being a possible response to a preceding one and a potential stimulus to the next.

The point of departure of this dissertation is the distinction between the topology of the tie structure and the temporal structure of sequences of communicative transactions. Theoretically, the dissertation explores mechanisms of co-evolution between these two structures at three levels of aggregation: (i) the *macro-level* consisting of the network itself or substructures within it, the level of an organization or a community as a whole; (ii) the *meso-level* consisting of nodes and social ties; and (iii) the *micro-level* consisting of sequences of interrelated communicative transactions. On the one hand, networks, individuals and ties are seen as the backdrop against which sequences of transactions unfold. On the other hand, transactions are considered to have (cumulative) consequences on the evolving structure of social ties and the network at large.

Methodologically, the thesis uses a publicly available dataset consisting of email transactions within Enron, an American energy and services company, during the few months of its bankruptcy. Two methods are applied to identify and explore the mechanisms. First, the dataset is disaggregated into various types of email transactions, revealing how different transactions contribute to various structural properties of the network. Second, a multilevel analysis approach is used to reveal how structural and transactional mechanisms combine to elicit new communicative transactions on the part of email recipients.

The mechanisms identified in the empirical chapters challenge received wisdom about the nature of social networks and their link to the notion of social (trans)action while at the same time addressing practical problems faced by network modellers who need to construct networks out of digitally mediated transaction datasets. In addition, the findings raise general questions about new types of data and the consequences they may have, not only for the field of social networks, but also for popular ways of thinking about 'the social' and ways of intervening in its course.

Acknowledgements

It would be very difficult to name all the people who have contributed, either directly or indirectly, to the journey that culminated in the production of this thesis. First and foremost I would like to thank my supervisors Carsten Sørensen and Peter Abell for their unabated support and intellectual generosity. I would also like to thank faculty members of the Department of Management at the London School of Economics, fellow students and others for stimulating conversations that sometimes erupted into animated but always friendly and fruitful exchange: Aaron Martin, Adebusoye Anifalaje, Aleksi Aaltonen, Attila Marton, Cohen Simpson, Edgar Whitley, Jannis Kallinikos, Leslie Willcocks, Lida Khalili, Magda Hercheui, Maha Shaikh, Maryam Ali Ficociello, Nuno Barros De Oliveira, Peter Erdélyi, Silvia Masiero, Shirin Madon, Susan Scott, Vasiliki Baka, Wifak Houij-Gueddana and Will Venters. I would also like to thank the network group of Nuffield college at the University of Oxford and to those associated with the group, especially Tom Snijders, Philippa Pattison, Garry L. Robins and Bernie Hogan for the assurance that what I had to say was both cogent from a mathematical point of view and relevant for the field of social networks.

My gratitude extends to my family and my friends who accompanied me on this journey, even in times of uncertainty and doubt: for your patience, your love and unyielding support. To my parents Avital and Avner Engel, to my brothers Amir, Jonathan and Michael, as well as to Rachel Engel, Elke Engel, Joelle Abi-Rached, Abdul Hay, Nomy Borenstein and Gabriel Wolkenfeld, you are a part of who I am. To August and Yael Wiedmann, to the Shehadis, to Greta Berlin and 'The Apples', to Mohammad Arafat, David Melot and Linda Feldman, who were there for me, unflinchingly, when I needed you most. I now read this list of names and imagine you marching before my eyes, the extraordinary women and men who touched and shaped my life, feeling that somewhere along the line, I must have been blessed.

Contents

List of Figures

List of Tables

Glossary

Adjacency matrix A representation of the network in matrix form.

Anatomical vs Functional networks Two terms adopted from the field of neural networks, the anatomical network consists of a set of entities and all the physical connections between them (such as interconnections of nerve fibers in the brain.) Functional networks are sub-networks of the anatomical one, consisting of the connections between entities that are activated for the accomplishment of specific tasks.

Balance A set of theories claiming that individuals strive for a state of coherence between the cognitive and affective states they ascribe to their various relationships. Insofar as a relationship reflects and constitutes a part of the identity of an individual, the theory states that individuals manage their relationships in a manner that increases their sense of consonance and decreases the dissonance between these parts.

Degree Used in one of two meanings: the degrees of separation between two nodes refers to the number of nodes spanned by the shortest path connecting the two. The degree of a node in a network refers to the number of relational ties associated with that node. Consequently, the degree distribution is the distribution of the number of relational ties associated with each node in the network. In directed networks, a distinction is made between the *in-degree* of a focal node and its *out-degree*, referring to the number of incoming or outgoing directed-ties incident to the focal node.

Density The proportion of connected *dyads* to the total number of *dyads* in a the network. A *complete graph* has a density of 1, and a graph with no ties has a density of 0.

DMTD Digitally Mediated Transactions Datasets, often referred to as ***Big Data***. These datasets are many orders of magnitude larger than *TND*, exhibiting high temporal and spatial resolution of the data (Borge-Holthoefer *et al.*, 2013). Compare with *TND*.

Dyad Any unordered pair of nodes in the network. Each network of $n \geq 2$ nodes has exactly ${}^{n(n-1)}\!/_{2}$ dyads.

ERGM Exponential Random Graph Models - a class of statistical models that account for the presence (and absence) of relational ties in terms of local tie based structures, such as reciprocated ties, *degree* heterogeneity and local transitivity.

Functionalism A theoretical doctrine according to which social phenomena (at the macro-level) exist by virtue of their features, whose function it is to sustain the group, protecting its integrity from internal or external threats. Compare with *Path-Dependence*.

Graph A mathematical representation of a network, $G(N, E)$ with a set of nodes $N = \{1, 2, ..., n\}$ and a set of edges $E = \{\{i_1, j_1\}, \{i_2, j_2\}, ...\}$, each of this set's members describing an association between two nodes in the set N. A complete graph is one in which every *dyad* is connected.

Homophily A feature of social networks, whereby connected individuals are likely to share similar traits, also known as *assortative mixing*. Two mechanisms explain this feature; associated individuals can develop similar traits (*Influence*), or similar traits of individuals can bring them to associate with each other (*Selection*.)

Macro-Micro link Two different links between the macro and the micro are discussed in this work. A ***definitional*** link refers to macro properties defined in terms of the micro (also known as constitutive, supervenience, analytical or aggregational link.) A ***contingent*** link refers to macro properties that are influenced but not logically defined by micro-properties (also known as empirical, synthetic or causal link.)

Network events Consist of tie formation, tie dissolution and the changing properties of individual nodes.

Path A sequence of distinct nodes and ties, in which each node is incident with the ties following and preceding it in the sequence.

Path-Dependence A theoretical doctrine according to which social phenomena (at the macro-level) exist by virtue of a particular sequence of historical events, contingent occurrences that were not necessarily determined on the basis of prior historical conditions alone (Mahoney, 2000). Compare with *Functionalism*.

Popularity Well connected individuals 'attract' more new ties than less connected ones, an effect in the network that is responsible to a considerable heterogeneity in the *degree* distribution.

Power distribution A distribuiton that follows a power law $\Pr(X > x) \sim x^{-(\alpha+1)}$ where $\alpha > 0$. The distribution is positively skewed and popularly known for its *fat tail*, a tail much fatter than that of the normal distribution. Closely associated with the *Lotka's law*, *scale free*, *Pareto* and the *Zipf* distributions.

QAP Quadratic Assignment Procedures, a bootstrapping approach to allow statistical inferences by comparing different networks among the same set of nodes.

Reciprocity A feature in a network by which connected nodes reciprocate favours or exchange information in both directions. The reciprocity of a network could be measured either by calculating the proportion of reciprocating dyads or by using *QAP* between the network's *adjacency matrix* and its inverse. Closely related to *symmetry* and *mutuality*.

Supervenience The set of properties A supervenes on a set of properties in B if there cannot be an A difference without a B difference. When the macro is uniquely defined by the distribution of a set of micro states, one speaks of the macro supervening over the micro.

Tie-Interdependency The notion that a relational tie has an effect on other ties in its vicinity. Examples include *popularity effects*, *homophily* and *triadic effects*.

Ties vs. Transaction-Patterns A distinction between two types of associations between individuals; a *social tie* refers to the social relationships existing between individuals (e.g. mutual trust, level of intimacy, expectations, conventions) whereas a *transaction-pattern* denotes the existence of transactions between them. Whereas the former is a meso-level construct that is *contingent* on the micro-level transactions, the latter is a meso-level construct *defined* by, and completely reducible to, the sequence of transactions. See also the *Macro-Micro link* and *Anatomical vs Functional networks*.

TND Traditional Network Data, network datasets commonly elicited by survey and questionnaire methods (Marsden, 2011). Compare with *DMTD*.

Transitivity A feature of the network expressed by the adage 'friends of my friends are my friends.' In a network with high transitivity, any two nodes sharing common contacts tend to be associated directly. High transitivity is sometimes associated with social capital, collaboration and a sense of equality, whereas low transitivity can be associated with hierarchy and inequality. Closely related to *closure*, *triangulation*, *clustering* or *balance*. The level of transitivity can be measured by the use of clustering coefficients or through *ERGM*.

Triad Any unordered set of three nodes in the network. Each network of $n \geq 3$ nodes has exactly ${}^{n(n-1)(n-2)}/_{6}$ triads. Triads may be connected or disconnected. A connected triad (with at least two ties) is known as a *triplet*.

Triadic effects A set of network effects which involve three nodes. Examples include *transitivity* or *balance*.

Two-mode network A network that involves two types of nodes, members of one type directly affiliated with members of the other type but not with one another. Examples include actors affiliated with films, individuals with social events, or directors with boards of companies. Also known as affiliation or bipartite networks.

1

The Heartbeats of Social Networks

I am resolved to keep afar
Wherever Love's attractions are;
The man of sense, as I detect,
Is ever shrewd and circumspect.

I have observed that love begins
When some poor fellow for his sins,
Thinks, it is thrilling, ever so,
To gaze on cheeks where roses glow.

But while he sports so joyfully
With not a care to mar his glee,
The links are forging, one by one,
And he's enchained, before he's done.

So there he is, deluded fool;
Stepping benignly in the pool
He slips, and ere he can look round
He's swept along the flood, and drowned.

Ibn Hazm Ali Ibn Ahmad (994 - 1064) (1981)

1.1 Introduction

Written over a thousand years ago, this short poem by the great Muslim polymath, Ibn Hazm Ali Ibn Ahmad of Córdoba, makes a couple of interesting observations, setting the stage for the problem addressed in this dissertation. First, the poem makes a distinction between short term interactions and the longer term commitments that bind human beings. The interactions consist of quick gazes, sweet words and innocent exchanges between the lovers, followed willy nilly by the crystallization and consolidation of a persistent bond, the 'chains' of a relationship consisting of expectations, responsibilities and obligations.

The speaker in the poem suggests a (causal[1]) connection between these two levels of social associations: short-lived transactions morph into full-fledged ties. But despite their interdependent nature, transactions and ties have very different properties[2] in terms of their purpose, their consequences and their longevity. The transactions have a utility[3] - they contribute to a feeling of joy in the heart of the lover, without bringing about any 'care to mar his glee.' But the lightweight and playful nature of these transactions are deceiving, because behind their backs, like an evil conspiracy, heavyweight ties are forged and the lovers find themselves ensnared in a system that is already well established, 'swept along the flood, and drowned.' No longer are they the agents of their own destiny, no longer do they enjoy the liberty of choice. Now there are other, external powers that are at play, their effect not unlike natural forces, social institutions now guide their actions; the expectations of society, the commitment to a family, the obligation to a career, and a responsibility to lead a structured and secure life. Unawares, the utility functions from the field of economics transform into brute 'social facts.' (Elster, 1989c)

One last observation from the poem is relevant to the theoretical discussions (Chapters 2 and 6), the warning that the fulfilment of the lovers' short term desires may be detrimental to the prospect of their long term happiness. The flip side of the same advice was developed systematically over a thousand years later by the distinguished economist and writer John Kay. In his book *obliquity*, Kay (2012) advises his readers to achieve happiness only indirectly, as a side-effect of

[1]For a debate on the type of connection between these two levels see section 2.2.

[2]Gibson (2005) talks about transactions and ties constituting 'orthogonal [!]) dimensions of social organization, with [transactions] unfolding sequentially, and social networks [of ties] extending (in a certain sense) spatially.'

[3]It is precisely the utility that is associated with social actions, that motivates the use of the word 'transaction' in order to describe the social action (Elster, 1989c). More on the nomenclature below and in Chapter 2.

actions that are motivated by reasons other than the pursuit of happiness. Both Ibn Hazm and Kay challenge a naive relationship between intentions, actions and consequences. Social actions, even if motivated by intention and purpose, bring about unintended consequences, and consequences are attributed ex-post to intentions that never existed (Merton, 1936). In what follows, I explore the possibility that network structures exist not only thanks to the advantages they bestow on individuals in the network (Coleman, 1988), but possibly also as a result of a path-dependent evolution, the unintended consequences of transactions that bring short-term benefit to the actors initiating them (Elster, 1989a). Network inefficiencies (Raub & Weesie, 1990) may thus be explained by path-dependency rather than functionalist accounts, ancestral traits rather than adaptive ones.

Short term transactions and long term ties are two levels that exist in many different systems, not only in the system of the relationship between lovers (Luhmann, 1998). Consider biological systems such as the cardiovascular system for example, a set of blood vessels, organs and tissue whose primary function is to provide adequate amounts of oxygen and nutrients to the metabolizing tissues, at the same time ensuring the removal of carbon dioxide and other metabolic waste products (Hicks, 2002). Here too we can distinguish between a structural system consisting of a semi-static network of arteries, veins, etc. on the one hand, and the actual flow of matter through the system on the other. The cardiovascular network itself is more or less durable, constraining, facilitating and synchronizing the flow of blood according to the needs of the living body. The flow is sustained and regulated by discrete like events such as each and every heartbeat, or the widening and narrowing of blood vessels as a reaction to environmental changes.

This metaphor provides additional insights; first, the network structure has a functional dimension. By this I mean to say that the reason for its existence is its properties, by virtue of which the living body is sustained. That said, some cardiovascular structures are inefficient and their properties cannot be explained by recourse to their function alone, but rather by a path-dependent, evolutionary process (Hicks, 2002). Furthermore, this metaphor highlights the notion that though they operate interdependently, the network and the flow are two separately existing phenomena. To demonstrate this, consider the existence of blood vessels of a dead body, after the heart ceases to pump and blood no longer flows. When network structure exists without flow, one can no longer argue that the former is analytically reducible[1] to the latter. It is a matter of debate as to whether or not this metaphor applies to the social realm as well; whether social structure is analytically reducible to transactions between human beings or whether these could be conceptually separated, the structure enjoying *sui generis* existence. The question of logical reducibility of the structure to the flow is an ontological one, a key point in the debate between Durkheimian realism and Tardian nominalism which is discussed in greater detail in Chapters 2 and 6.

Before demonstrating why the distinction between transactions and ties is relevant for the study of social networks today, I would like to introduce a second metaphor, this time taken from neural networks in the brain. Neuro-scientists make a distinction between two types of brain connectivity (Rieke *et al.*, 1997; Shalizi *et al.*, 2006). The anatomical network consists of persistent, semi-static connections between neurons, bound together through physical nerve fibers in the brain. Another type of connectivity can be defined through coordinated behavior, the network of neurons that tend to synchronize and correlate activity. Even if a group of neurons are connected to one another physically, their activity can be correlated for the accomplishment of one type of task or uncorrelated when accomplishing a different type of task. Thus, the same anatomical structure of neural networks can be mapped to different 'functional networks,' depending on the cognitive task at hand. The point again is to emphasize the conceptual distinction between the network as an infrastructure in which flow can potentially take place, and the actual flows, activations or transactions that occur within the network.

There is really nothing new in the realization that social networks exist as a semi-static structural entity, within which sub-units can be active whereas others are not, exchanging information at different times and for various purposes (Mitchell, 1969). However, it is fairly recent that systematic work has been carried out with the intention to analyze the temporal structure of network flow and its

[1]Saying that A is analytically reducibile to B means that there A *supervenes* on B and that there is nothing in A properties that is *over and above* B properties. See section 2.2 for a discussion on this matter.

consequences (Holme & Saramäki, 2011). A growing body of literature is now examining these issues, revealing new and exciting puzzles. In what follows, I will demonstrate some these issues, highlighting their interdisciplinary nature.

1.2 Transactions and Their Consequences

Traditionally, social networks are applied mainly to represent a topological and durable structure of interconnected individual elements (Barnes, 1954; Mitchell, 1969). From employees embedded in organizational hierarchies to ties of friendship, kinship and trust, data describing complex networks is represented in the literature through so called *graphs*. A graph is a mathematical object consisting of a set of nodes and a set of ties, each node representing the basic units of the system, and each tie representing a pair of connected units. To the naive observer, a graph is merely a model of the data, supposedly[1] carrying only modest theoretical luggage (Morgan & Morrison, 1999). Adding theoretical propositions to the model paves the way for the classification, prediction and optimization of the system's behavior (Wasserman & Faust, 1994; Barrat *et al.*, 2008; Pastor-Satorras & Vespignani, 2010; Serrano *et al.*, 2009; Vespignani, 2009).

In contrast to the durability of the social network, the association between individuals communicating via email, text messages or phone calls, is performed for a short length of time; the meaning and significance of these transactions often unknown to the observer (Holme & Saramäki, 2011). Was an arbitrary email sent by mistake, or does it signify a meaningful connection between sender and recipient, traces in the data of a relationship relevant to the study at hand? Can one attribute any meaning to an email without knowing its context or even content? Moreover, how should one account for the various ways in which a tie is activated, and what consequences might temporal patterns of tie activation have?

The first example for the importance of temporal patterns in networks concerns the issue of transitivity (Holme & Saramäki, 2011; Rocha *et al.*, 2010; Moody, 2002). Standard graph theory tells us that given a triplet, A, B and C, where one

[1] The role of the model and the theoretical assumptions it conceals is both a matter of debate in the literature (Morgan & Morrison, 1999; Hacking, 1990; Mitchell, 2009) and one of the critical points raised in the dissertation - see chapters 3.

path[1]exists between A and B and another between B and C, there must also be a path between A and C. Less formally, consider three actors, one of whom acts as a broker (B) between the other two (actors A and C.) The latter two actors can both exchange information directly with their broker. But even if those two actors cannot exchange information directly with each other, one might conclude that they can exchange information indirectly via their broker. Thus, networks have a property of transitivity. However, consider what happens if (A, B) can exchange new information only ***after*** actors (B, C) do so. If the temporal sequence of communication is consistent, then no new information can ever propagate from A to C, violating transitivity under this regime of information transaction patterns. Without taking the temporal patterns into account, the study of diffusion of new information may yield faulty conclusions (Kempe *et al.*, 2000; Moody, 2002).

It is one thing to take a static network and to study its attributes; the types of nodes and edges, the role they play in the network and the properties of their subunits. But the research of networks often deals with questions about dynamic processes that unfold in the context of interconnected individuals. The classic example is the study of diffusion of viruses (or technology, information etc.) within a network of connected actors (Coleman *et al.*, 1957), where the ties between individuals are conceptually separated from the discrete events by which one individual is infected by another. More generally, there is a distinction between the underlying static network of social ties and the set of events that the network facilitates or constrains. When is the temporal nature of transactions most important to take into consideration? When can it be ignored? Holme & Saramäki (2011) argue that the temporal nature of interactions should be taken into account when temporal structures are 'not too random or too regular.' The authors warn that in those cases, ignoring the temporal patterns of transactions could result in the loss of explanatory power, possibly leading to faulty results.

To illustrate their point, Holme & Saramäki (2011) note the role of 'bursty' patterns of human interactions in the propagation of sexually transmitted diseases. In a milestone publication in *Nature*, Barabási (2005) explains why certain types of human activity follow a random pattern, whereas others do not. Random events are typical of traffic flow and incoming calls in a call center, for example, each

[1]For abbreviations and nomenclature see the glossary on page xii.

event occurring independently of past ones. In a call center, for example, the probability for a call within the next 5 minutes does not depend on whether or not the last call occurred just a minute ago or half an hour ago. In contrast to random events, 'bursty' ones do exhibit interdependency over time. Consider the act of sending out emails. A relatively long period of inactivity could be followed by a short sessions in which users send out several emails in a quick succession. Sending emails is just one example of many types of task oriented and cognitive dependent activities that exhibit 'bursty' regularities. Other examples include web browsing, library visitation, ratings of movies, mobile communications and trade transactions (Barabási, 2010; Song *et al.*, 2010).

Understanding the temporal distribution of social activities is crucial for the study of how things spread in a collective, and specifically to the study of sexually transmitted diseases (Rocha *et al.*, 2010). In each sexual encounter between an infected and a non-infected individual, only a small dose of the pathogen is transmitted to the uninfected individual. In and of itself, this dose is usually not enough to infect the receiver and its concentration in the blood decays exponentially over time. Hence, a small number of sexual encounters, randomly distributed over a period of time, would allow the level of the pathogen in the blood to decrease enough between encounters, never reaching the critical threshold necessary for infection. In contrast, consider an uninfected individual practicing abstention over most of the same period of time, followed by a succession of sexual encounters. In this case, the concentration of the pathogen in the blood may not have enough time to decay, building up and much more likely to reach the critical threshold that results in infection. Thus, the propagation of the disease depends not only on the structure of the network or the number of sexual encounters within a given period, but also on the temporal distribution of these events (Rocha *et al.*, 2010; Holme & Saramäki, 2011).

Another issue has to do with transactions that involve more than two actors at any one time. Networks represent connections between pairs of actors, since each tie connects exactly two actors. But social study becomes most interesting beyond two actors and just one single tie (Simmel, 1964). More specifically, the study of interdependencies between one pair of actors and another pair is arguably at the very core of what it means to study networks. To address issues of tie-interdependencies, several methods have been developed in the literature, some of which present useful but ad-hoc techniques (Eckmann *et al.*, 2004; Kossinets & Watts, 2006; Palla *et al.*, 2007) whereas others (Snijders *et al.*, 2006; Brandes *et al.*, 2009; Butts, 2008) are more general but mathematically onerous and not always suitable for analysis, especially for large and complex datasets of social exchanges. This is unfortunate because in many cases interdependencies are explicit in the empirical data itself, they can be read off the data and need not be derived. For example when more than two individuals interact, the transactions between one pair could bear upon another. However, for reasons that are discussed in chapter 3, it is precisely this information about interdependency between distinct pairs that is discarded in the process of aggregating the data and encoding it in a way that is amenable to network analysis.

Though the exchange of information can be done on a one-to-one basis, many such transactions involve a message that is broadcast to anyone that might listen. Studies of communication networks include spreading messages in blogs (Kumar *et al.*, 2005; Adar & Adamic, 2005) and microblogs (Java *et al.*, 2007; Kwak *et al.*, 2010). Between one-to-one and broadcast messages, there is an intermediate form of transactions that occurs within bounded groups. Examples include the exchange of information in meetings (Gibson, 2005) or e-mails sent to multiple recipients (Zhou *et al.*, 2005; Engel, 2011; Liben-Nowell & Kleinberg, 2008). The process of handling the data in question, aggregating it and disaggregating it to obtain the final network model, has been shown to have substantial consequences on the network model and on the results of the research (Engel, 2011; Quintane & Kleinbaum, 2011; Krings *et al.*, 2012). As of today, there are no standard practices or accepted considerations, how to transform data into a network model, an open problem that presents urgent methodological and theoretical challenges.

The examples above demonstrate some of the limitations and frontiers in the state of the art of social networks. They are especially onerous when studying communication networks (Monge & Contractor, 2003), when models are derived from digitally mediated communication exchanges. This is less of a concern when the object of study is a static or slowly evolving network and more of a problem in situations where the network is derived from data consisting of event-type encounters, exchanges and transactions. Traditionally, researchers overcome this problem simply by aggregating the contact sequences. The implied assumption here is that a relationship is equivalent to a set of encounters. In some cases this is certainly an acceptable approximation. But, depending on the objective of the research, this method may have serious limitations, as shown in chapters 3 and 4.

Studying networks of ties and transactions is an interdisciplinary endeavor. For example, the study of sexual transmitted diseases involves not only the biological mechanisms by which infection occurs, but also the social, psychological and even political (Brandt, 1985) (!) processes that govern human sexual activity (Krieger, 1994). The interdisciplinary nature of the problem isn't evident only within a study, but also across studies (Holme & Saramäki, 2011) and like many such problems, challenges and opportunities arise when researchers of different fields need to address similar problems. General issues include the development of an appropriate vocabulary and an adequate way to visualize the temporal nature of transactions associated with networks, graphically demonstrating the relevant statistics that are most suitable to characterize the system and to compare between them. Thus, where I use the term transactions, other authors use different concepts interchangeably that refer basically to the same thing, concepts such as 'events' (Brandes *et al.*, 2009), or 'relational events' (Butts, 2008), 'edge-activation' or 'temporal networks,' (Holme & Saramäki, 2011) 'functional networks,' (Shalizi *et al.*, 2006) 'interactions' and 'social action' (Gibson, 2005) etc. A vocabulary that is not standardized can make it taxing to read related literature in and across fields. In addition, in order to benefit from insights in one field and adapt it to another, the reader needs to attain some degree of proficiency in an otherwise unrelated field. And yet social scientists, physicists, epidemiologists, neuro-scientists and others need to learn from one another in order to enrich the understanding of their own field.

1.3 Dissertation outline

Against the backdrop of the problem presented above, this work is guided by a research question that addresses the ***links between social transactions (at the***

micro level) and networks of social ties (at the macro- and meso-level) .

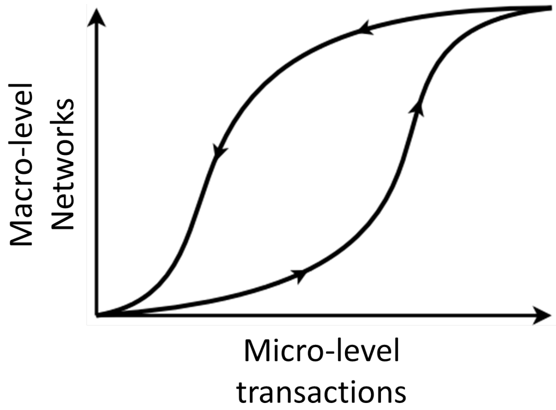


Figure 1.1: Social Hysteresis - micro-level interactions and macro-level networks and ties

Now, before narrowing down the question and qualifying it a bit, I would like to illuminate it by presenting an example of a possible answer, the idea of a social hysteresis suggested by Elster (1976) and depicted in figure 1.1. It is best to explain the idea by recalling again the two lovers from Ibn Hazm's poem. They meet each other for the first time at the graph's origin and start exchanging their bashful gazes, slowly climbing along the lower path of the hysteresis curve. At first, the path is delightfully convex, each of their interactions carries very little baggage in terms of commitment and mutual obligations, the bond between them builds much more slowly than the frequency and emotional intensity of their interactions. But an unfortunate turn of events changes all that - maybe they have been spotted by someone else, friends or relatives, and the path curves upwards, becoming uncannily steep; each small interaction carries with it more expectations as the prospect of marriage and an institutionalized relationship looms near.

The wedding represents the climax of the pairs' commitment and level of interaction, after which the fall back to the origin is all but inevitable, this time taking the upper path of the hysteresis curve: micro-level interactions decrease in emotional intensity and slow down in frequency, at first without apparent damage in commitment, trust and mutual respect, later with palpable effect on the relationship.

Unfortunately, this dissertation will not provide such a neat answer to the question, primarily because it is very hard to measure the social tie in a direct way. It is possible to show empirically that ties do have explanatory power (see in Chapter 4 and 5), but it is difficult to learn a great deal about their properties in detail. In addition, it focuses not just on a pair of lovers but on a network of members in an organization, engaging in email transaction, emails being just one of of several modes of communication. By focusing on networks in organizations and the technology that mediates transactions between individuals, the dissertation strives to contribute to the literature of organization studies and information systems. On the one hand, organization studies promotes the understanding of structures and processes within (and between) organizations, addressing, among others, problems of coordination and control, hierarchy, power, boundaries, performance and evolution of organizations. In so doing, it departs from traditional neo-classical economic approaches that treat firms as profit-maximizing black-boxes, optimally adapting to fit their (exogenously) changing environment (Coase, 1937).

Somewhat related to organization studies, the study of information systems strives to understand the role and consequences of information and communication technologies (ICT) in communities, organizations or other social contexts. It addresses problems of usability, adoption and unintended consequences of ICT, creating a corpus of knowledge that focuses on 'the development of IT-based services, the management of IT resources, and the use, impact, and economics of IT with managerial, organizational, and societal implication' (MISQ, 2013). In so doing, it departs from traditional approaches that treat technology as a black-box deployed to increase the efficiency of carrying out tasks.

Within this disciplinary framework, the dissertation examines the mutual influences between networks of ties and email transactions exchanged between its members, albeit somewhat obliquely (Kay, 2012). Because of its circular nature (transactions affecting network and being affected by them) a direct approach could follow the footsteps of Manski (1993, 1995), in his investigation of so called 'endogenous' or 'correlation effects' . But this dissertation adopts a less direct, less general approach and a more empirically driven one; instead of testing the mutual influences between ties and transactions directly, the dissertation proceeds by focusing on a practical issue that faces many researchers: given a dataset of email transactions, what are the alternative options to construct a network model? By investigating this question, by exploring different avenues and comparing between them, the intention is twofold; first, to tackle a pressing practical issue and second, to allow theoretical issues to arise in the process and to focus on each as it emerges. This is an exploratory approach, to some extent data-driven, making use of an extract from a well studied email dataset known as the Enron corpus (Shetty & Adibi, 2004).

The theoretical and the methodological approach chosen has different names and various interpretations (Udehn, 2002), but in what follows I use the term micro-foundations. Micro-foundations is becoming an increasingly important theme in studies of organizations in the past decade or so. Its basic premise is that an adequate explanation of collective phenomena needs to involve explanatory mechanisms of the participants in those phenomena (Miller, 1978; Udehn, 2001, 2002). Scholars are paying substantial attention to explanatory mechanisms that involve individual incentives, preferences, opportunities and action in understanding issues such as financial performance (Abell *et al.*, 2008), resource value (Lippman & Rumelt, 2003b; Foss & Foss, 2005), value appropriation (Lippman & Rumelt, 2003a; Coff, 1999; Barney, 2001), inertia (Kaplan & Henderson, 2005), strategy implementation (Barney, 2001), firm-level heterogeneity (Gavetti, 2005), factor market dynamics (Makadok & Barney, 2001) and other properties of organizations[1]. In their search for the micro-foundations of collective phenomena, researchers of organizations follow the steps of a similar research projects in macro-economics (Leijonhufvud, 1967; Colander, 1993; Carlin & Soskice, 2006) and rational choice sociology (Elster, 1989c, 1998; Coleman, 1990; Abell, 2003a).

Having touched upon the research question and the theoretical and methodological issues at stake, I turn now to outline the next chapters.

[1]For a review of micro-foundations based studies of organizations, see (Felin & Foss, 2006; Foss, 2010)

1.3.1 Chapter 2 - Theoretical foundations

The next chapter provides a literature review and a theoretical background for the rest of this work. The core question guiding the chapter is how to apply the theoretical framework of micro-foundations to the study of networks. The motivation is that despite the exploding availability of cheap and large-scale electronic datasets describing human communication transactions (Watts, 2004a, 2007; Lazer *et al.*, 2009b), network researchers continue to rely primarily on traditional questionnaire-based data collection methods to test and advance substantive network theory (Quintane & Kleinbaum, 2011; Marsden, 2005, 1990). Moreover, comparative studies report systematic differences between social network data extracted from observed interactions and those extracted from survey-data: people report one set of social ties, but their interactions suggest a different set of associations (Krackhardt, 1987; Killworth & Bernard, 1980; Bernard *et al.*, 1980, 1984, 1990).

There are several issues here at stake, one is a methodological problem (imperfect recall and report bias for example.) The second is the conceptual difference between ties and transactions. This chapter focuses on this conceptual difference that makes reported relationships more amenable to the traditional study of social networks. This does not mean that digital transaction data do not have the potential to inform and contribute to theoretical developments in the field (Lazer *et al.*, 2009b; Wimmer & Lewis, 2010; Szell & Thurner, 2010), but it suggests that there may be a need for conceptual developments if the datasets are to be used to inform the study of networks.

The motivation for engaging in micro-foundations is hence the possibility that it could provide a bridge between the level of interaction and the level of network structures. The chapter opens by clarifying the two ways in which these two levels of human associations are conceptually linked. Furthermore, it suggests a model that resonates with the principles of micro-foundations, at the same time bridging the divide between social transactions at the micro level, tie level attributes and events at the meso-level and network topology at the macro-level.

This conceptual issue is then anchored in a long standing controversy in the social sciences, epitomized in the momentous debate concerning the nature of sociology and its relation to other sciences, a debate that took place between Gabriel Tarde and Émile Durkheim at the École des Hautes Études Sociales in 1903 (Vargas *et al.*, 2008; Karsenti, 2010). Attempts to synthesize between Tarde's nominalism and Durkheim's realism can be found in the works of Marcel Mauss, Georg Simmel and Max Weber, all of whom contribute to the development of definitions used in the dissertation: social (trans)actions and social ties, networks, structure and tie-interdependency.

The second part of the chapter categorizes the body of social networks research into three groups, according to the way it deals with dynamical processes. The first group consists of studies of the static network. This type of work does not focus on dynamics at all, but even here there are interesting links between the different levels of analysis. The example provided is the strength of weak ties hypothesis, according to which properties of the tie are related to the topology of the network surrounding the tie. This is a fascinating and unintuitive example of a link between different levels of analysis.

The second body of work focuses on so called *network events*. These are mechanisms that operate at the level of the tie and the individual, consisting of tie formation, tie dissolution and changing properties of individuals. Three main effects explain these network events; the popularity effect (Snijders *et al.*, 2006), homophily (McPherson *et al.*, 2001; Robins *et al.*, 2001a,b), and triadic effects (Snijders *et al.*, 2006; Feld, 1981). All of these mechanisms are examples of tie interdependencies, but also examples of a micro-foundational mechanisms in which the act of tie creation is made at the local level, giving rise to macro patterns at the level of the network as a whole. However, from the point of micro-foundations, the lowest level in which anything happens is still the level of the tie. The meso-level of ties and individuals interacts with the macro-level of network topology. But creating or dissolving ties is altogether a different type of event to that of sending out emails or engaging in a transaction. The difference is in the level of analysis. Conceptually, a social tie is located at a higher level of analysis than a social transaction, since every tie is associated with multiple transactions. This body of work does not emphasize the micro-level of social transactions.

The third body of work explicitly recognizes and acknowledges the existence of social transactions as a social phenomena, and sets off to develops strategies how to deal with it. One strategy defines the network in terms of observed patterns of transactions. The proponents of this strategy claims quite explicitly that for the purpose of their research, there is no difference in kind between social transactions and social ties. Another strategy is to acknowledge the different levels of analysis, looking for mechanisms that bridge between these levels. It is this third strategy that the current work adapts and applies in the empirical chapters. The chapter ends with a formulation of the research question that seeks to identify mechanisms of mutual influence between the micro-level of social transactions, the meso-level of ties and nodes, and the macro-level of network topology.

1.3.2 Chapter 3 - Forging networks from their micro-foundations

The previous chapter ends with a theoretical question that can be turned into a methodological one. Given a dataset of social transactions (phone calls, emails etc), how should one construct a social network? To answer the question, a comparison is made between social network models that are based on traditional, survey based data, and social networks based on digitally mediated transaction data. Three issues are discussed. First, the technical issue and the sheer volume of the data associated with digitally mediated transactions. This volume has implications on the network model that are both methodological and theoretical. For example, networks that are based on digitally mediated transactions tends to be much more dense than traditional network models. This density has implications for the interpretation of network parameters that are known to interact with density.

A second issue is that of relevancy. Whereas in survey based methods, it is the respondent that judges the relevance of her reported connections, in digitally mediated transaction data, the scientist must judge this issue. Finally, the nature of interdependency is quite different and the links between one tie and another are altogether of a different kind than the links between one transaction and another. All these different aspects must be taken into account when constructing the network model.

The methodological chapter spells out how existing studies go about constructing network models from raw data, focusing on the treatment of email communication datasets. Those are typically transformed into networks in which nodes designate email users and ties connect nodes if an email has been dispatched from one of them to the other. Each email is treated as a star shaped 'ego-network' (Freeman, 1982) with the sender in the center of the star, connected by an edge to each of its recipients. The ego-networks are then interlaced and superimposed on top of one another, creating the entire network. Unfortunately, this method of extracting sender-recipient pairs from multiple recipient emails discard important information regarding the nature of affiliations and the process by which they come about . Specifically, recipient lists of a user's outgoing emails delineate meaningful organizational units so that being co-recipients on a single email is a stronger indication of affiliation than being recipients of separate emails from the same user. Whereas the first may designate a part of a collaborative effort (such as a group discussion,) the second suggests a series of independent transactions.

The core of the chapter discusses several methods, by which networks can account for phenomena that are located at the micro level of transactions. The first method uses tie-strength to reflect the temporal patterns of transaction. The idea is that every email between two users reflects and re-affirms the tie between them, so that the more emails are sent between two users, the more meaningful is their tie. The key advantages for using the strength of ties is that this method is relatively straightforward, both conceptually and mathematically. It uses more of the data than traditional methods, enriching the network model and representing more of the original network data in a way that is compatible with existing methods of network analysis. Accounting for non-binary edge values is a natural extension of the existing toolkit of network analysis; centrality and density for example, group search algorithms and clustering indicators have been extended in the literature to account for tie-strength (Opsahl & Panzarasa, 2009). In Chapter 4, an attempt is made to validate this method, showing the benefits of using idiosyncratic properties of the email artefact to calculate tie-strength. The downside of this method is that much of the important information is still discarded, especially with regards to the issue of tie interdependencies.

Other methods include the disaggregation of the network data into sub-networks, in order to control for confounding variables that are hidden when the entire dataset is aggregated. An example of this method is the notion of taking snapshots of the dataset. Other methods include event networks and the bipartite network.

Whereas the the methods of disaggregation and tie-strength are applied empirically in chapter 4, chapter 5 develops a method that resonates directly with the theoretical framework of micro-foundations. This method is based on multi-level analysis (Snijders & Kenny, 1999; Snijders & Baerveldt, 2003; Courgeau, 2003) and applies them to the two email transactions dataset. The method is especially interesting from a theoretical point of view, because it views a transaction as a potential product of properties at different levels of aggregation. A given transaction can be triggered by preceding transactions (at the micro-level), it is affected by the individuals involved (sender and receiver(s)), and by the history of their relationships (at the meso-level.) Finally, it is affected by the macro-level properties of the system. Teasing out the sources of variance among the different levels of aggregation is important theoretically, but it has different applications as well. For example, the method can be used to compare within each of these factors, e.g.,for comparing email users to one another in terms of the way they use emails. The method can also be used to compare between the factors. This could be important for those interested in marketing and customer relations, for example. Say a customer reacts to a certain message sent by the marketing department. It would be of interest to learn the reason for this 'success,' why did the message elicit a response? How much importance should be attributed to the sender's properties ('sender effect?') Is her role or reputation such, that her mails are rarely left unanswered? How much should be attributed to the overall responsiveness of the customer ('recipient effect?') What is the relative importance of the specific tie between the sender and her customer ('tie effect?') What is the relative importance of the message to elicit a reply ('message effect?')

Multilevel analysis is a method that offers ways to think about these questions. Its downside is that it is not adequate for large datasets and that triadic effects are not really taken into account. In fact, the data is not treated as a full fledged network but as a hierarchy, that consists of one-to-one transactions, ties and individuals. Whether it is possible to extend the model in future work to account for triadic effects is an open question. The methodological section ends with a discussion of how extant literature on email communication technology informs the different methods discussed and how the development of these methods can contribute to the extant literature on information systems and organizations research.

1.3.3 Chapter 4 - From Micro to Macro: How emails contribute to network structure

The first empirical chapter sets out by disaggregating the network into sub-networks in order to control for a confounding variable which depend on properties of the email. The empirical data yields interesting associations between the email properties and structural properties at the level of the network. Specifically, the emails are sorted into a spectrum according to the number of recipients, few recipients are associated with private emails, and numerous recipients with broadcast emails. The motivation is to explore the significance of the medium of transaction, email messages in this case, to reveal different structural patterns in the data. The intuition here is that an email sent to a single recipient has a different meaning, function and consequences than sending an email to multiple recipients.

To test this proposition, several different networks are constructed using the same group of 254 users over the same period of time, but each network was based on emails with a different number of recipients. Comparing networks derived from single-recipient emails (private transactions) with networks derived from multiple recipient emails (broadcast transactions) reveal systematic differences. First, reciprocity is greater in private transactions than in broadcast transactions. Second, broadcast transactions reveal a structure that is more clustered and tightly knit, whereas private transactions reveal a structure that is less clustered and more hierarchically structured. Finally, the degree distribution shows substantial differences between private and broadcast emails, and between in-degree distribution and out-degree distribution.

These findings open theoretical questions as to the social mechanisms that explain them, an issue that is addressed in the analysis and discussion chapter (chapter 6.) A second insight is that the way people use the same technology matters, that different uses of technology reveals various networks, suggesting that within a community there exists not one, but multiple 'layers of structure' existing side by side. A third insight is that the traditional method of constructing social networks from email dataset is problematic, in the sense that it discards important information regarding interdependencies among ties.

These findings prompt further investigation of various types of degree distributions within email mediated transaction datasets, all of them exhibiting a starkly skewed distribution.

One of the aims of the chapter is to use tie strength in order to construct network models that incorporate more of the information captured in the empirical dataset of email transactions. The findings indicate that the resulting network model is sensitive to decision made in the process of constructing networks. Moreover, critically evaluating the process of network construction reveals interesting patterns that demand explanation: the curious distribution of email production, consumption and dissemination being one example. 'Layers of structures' being another. Chapter 6 attempts to suggest explanations to these findings in the form of mechanisms that operate at the micro-level. But before using theory to make more sense of the data, let us turn to a third method for analyzing the data empirically.

1.3.4 Chapter 5 - From Macro to Micro: Four factors influencing email reciprocity

The second empirical chapter applies the method of multilevel analysis introduced in the methodological chapter the dataset. The intuition that guides this method is that an email transaction is not necessarily a one-off event, but is embedded in chains of related transactions, not unlike the 'chains of interactions' idea developed by Erwin Goffmann and his students (Collins, 2004). According to this view, each email transaction is potentially a stimulus for the next email in the chain, as well as a possible response to a previous transaction. Thus, each email is one transaction that must be understood in the context of the chain of transactions, emails bouncing back and forth between the users in the network, people collaborating or discussing a problem and together reaching a form of consensus, thereby re-affirming or reconfiguring the nature of their ties.

Transaction sequences and social networks are conceptually distinct and not easily reconciled, two different programs of research, a division of labour within the social sciences themselves between 'interactionists' and researchers of 'networks,' two different camps sharing a sense of mutual suspicion (Gibson, 2005). This chapter attempts to follow the advice of Gibson (2005, 2008), searching for the intersections between these two sub-disciplines. Thus, an email is not only an indication of a network tie, but also as a discrete bead in a stimulus-response chain of interaction.

For the purpose of this study, an email is considered 'effective' to the extent that it succeeds to elicit a reply from a recipient. For every email sent from A to B, a reply email is searched for, one that was sent from B to A at a later time and shares the exact same subject (ignoring subject prefixes such as 'fwd' or 're' that are added automatically by the email client software.) For each email from A to B, the binary outcome variable measures whether the recipient B has replied. Four effects are estimated. The sender effect measures to what extent emails from a specific sender, A, are likely to receive responses. Some senders are found to be much more (or less) effective in eliciting replies than other senders, independent of who the recipients are. The recipient effect measures the responsiveness of each recipient to emails. Some of recipients are found to be much more (or less) effective in replying to emails than others. The tie effect measures to what extent a specific tie between a pair of users is responsible for a high level of responsiveness. It is possible that two users are much more (or less) responsive to each other than they are to the ties they have with other people. The fourth effect is the effect of the message itself – perhaps there is something in the message itself that is responsible for a high level of responsiveness from its recipients.

The findings show that all these four effects are important, to a varying degree. Besides the importance of teasing apart effects at different levels of aggregation, the model validates the findings in the previous chapter, showing that the number of recipients is significant and inversely proportional to the probability of eliciting a reply. Second, there are several practical explorations that ensue from these findings. For example, one dataset shows that sender effects and recipient effects of the same individuals are positively correlated, while in the other dataset they are negatively correlated. To understand exactly why we get different findings in different networks, we would require access to more data about the organizations, the individuals working in it and their organizational roles. But for our purposes it is enough to note that some findings are consistent between networks and other findings vary. In this way effects at all three levels of aggregation were identified, each playing their role in determining human action: effects at the micro-level of past email transaction, effects at the meso-level of the ties and nodes, and finally effects at the level of the network itself.

.

1.3.5 Chapter 6 - Discussion and Conclusion

After presenting the findings, this chapter turns back to the theoretical framework discussed in chapter 2, attempting to apply it to the empirical results in light of its micro-foundations. The chapter opens by addressing a general critique to the whole work, and explaining the different types of links existing between the micro-level of social transactions and the network level. It then reviews the findings in the empirical chapters and associates each of the mechanisms identified with the the appropriate component in the theoretical framework.

The following section reviews the contributions critically, in three stages. First, through a systematic comparison between network models that are constructed from survey data and those based on transaction data. Examples from the empirical findings serve to enrich and illustrate the the argument. Theories that work in one way at the level of traditional network models, seem to work in a different way on the level of transactions, thanks to the way people interact with the technological medium. Some of the concepts have slightly different meaning on those two levels, and though the notion of interdependency plays an important role at the network level, on the transaction level it is what motivates the actor to engage in the transaction. The chapter closes with a review of the methodological contributions and a critical assessment of the principles of micro-foundations and their potential contribution to social networks.

1.4 Summary and Reflections

The invention of the motion pictures has revolutionized the way we perceive the idea of movement (Canales, 2010). By breaking down motion into a discrete sequence of shots in a way that the human eye was not capable of doing, mysteries of processes and change have been solved, replaced by new kinds of puzzles. An example thereof is the historic case of *Sallie Gardner at a Gallop*, when the photographer Muybridge settled a long standing puzzle regarding a horse's racing-speed gait with a series of progressively clearer, single photographs of Leland Stanford's trotter, Occident.

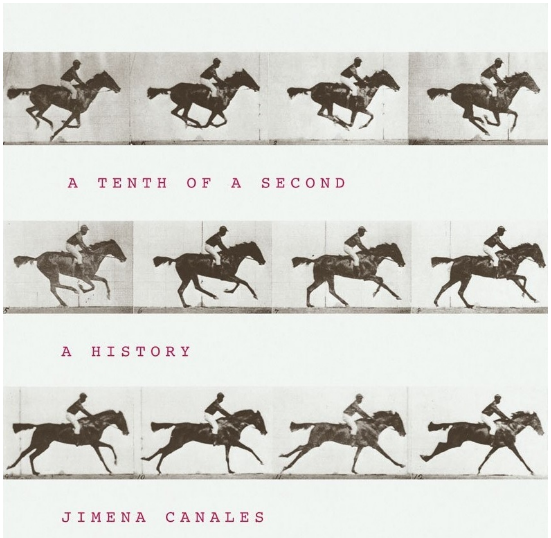


Figure 1.2: Advances in technology influences the perception of time and motion - as described in *A Tenth of a Second* by Canales (2010)

Relevant to this discussion is the controversy triggered by the development of cinematographic techniques along with the availability of increasingly sensitive film, a debate about the reducibility of our macro experience of time into its smaller micro units. Philosophers such as Henri Bergson insisted that 'real movement, real change, and real events escaped between the static intervals of time used in the sciences,' (Canales, 2010, 17) whereas Gaston Bachelard, a scientist, philosopher and poet worried that the 'stroboscopic era' challenges thinkers to question 'the ease . . . of correspondence between 'real' phenomenon and the instrumental phenomenon of stroboscopy'(Bachelard, 2000).

Today we face again the dissection of continuities into discrete micro events. It is probably the first time in the history of our species that our most ephemeral, spontaneous and impulsive social actions are being inscribed forever on tape, recorded digitally, organized and available for research. Of course, the kind of research that is being carried out on our own society has a different kind of effect than the research made on a horse's gait, since the latter is indifferent to our knowledge about it, whereas the former is at the same time the object of our observation but also a subject, and as such, rather sensitive to the results of scientific enquiry. As a range of intellectuals starting from Marcel Mauss, Emile Durkheim (Bloor, 1982) and all the way to Ian Hacking (1995) repeatedly remind us, scientific assertions feed back and influence people's system of beliefs, values and priorities. At the very least one could expect people to be influenced by the growing awareness (and perhaps concern,) that each of their communicative transactions are being increasingly monitored, as the Hawthorne effect famously demonstrates.

These changes consist of two interconnected processes. First, we can now zoom into the sequence of events that connect lightweight transactions with heavyweight social structures, a process that was articulated ever so poignantly by the great Ibn Hazm one thousand years ago, but whose 'nuts and bolts' remain a mystery to us until this very day. The second process is the one by which the availability of communication data and the effort to articulate and gradually solve puzzles surrounding it push both popular and academic thinking to use the network metaphor to understand processes in the world, to articulate political/social sequences of events, and to influence their outcomes. This dissertation begins to touch upon these two interconnected processes.

2

Theoretical foundations

> You cannot step into the same river twice.
>
> Heraclitus quoted by Plato in Cratylus, (2003, Fragment 41)

> Does Mr. Durkheim think that social reality is anything other than individuals and individual acts or facts? If you believe that ... I understand your method which is pure ontology. Between us is the debate between nominalism and scholastic realism. I am a nominalist. There can only be individual actions and interactions. The rest is only a metaphysical entity, mysticism ... Il n'y a de réalité que dans l'action'
>
> Gabriel Tarde [1908] (2010, p 140)

> Those who arrive at Thekla can see little of the city beyond the plank fences, the sackcloth screens, the scaffoldings, the metal armatures, the wooden catwalks hanging from ropes or supported by sawhorses, the ladders, the trestles. If you ask, 'Why is Thekla's construction taking such a long time?" the inhabitants continue hoisting sacks, lowering leaded strings, moving long brushes up and down, as they answer, 'So that its destruction cannot begin.' And if asked whether they fear that, once the scaffoldings are removed, the city may begin to crumble and fall to pieces, they add hastily, in a whisper, 'Not only the city.'
>
> Italo Calvino (1978, p 127)

2.1 Introduction

The purpose of this chapter is to present the theoretical framework of micro-foundations, and to explore how it may contribute to the field of social networks. Micro-foundations has been increasingly attracting scholarly attention over the past decade or so (Demeulenaere, 2011) in particular in the field of organization studies (Felin & Foss, 2006; Abell *et al.*, 2008). Since it has been applied to so many other areas of organization studies, we might want to consider the application of micro-foundations to the field of social networks as an interesting intellectual experiment, testing the usefulness of the theory in shedding light on some of the findings established in the field. Another reason to consider micro-foundations stems from the common language it shares with economics. If the theory facilitates the understanding and interpretation of social network findings, one may be inclined to start thinking of the prospect for using it to pave the way for a unified social science, a bridge between economics and sociology (Abell, 2003a).

Micro-foundations also provides a useful way of organizing and guiding empirical research, not so much thanks to its ability to make predictions about outcomes but as a way to organize social explanations and entities, suggesting what a social explanation needs to entail and how different explanations stand in relation to one another to form what one may call a grand-theory of social change (Boudon, 1986).

A further reason to turn to micro-foundations in the context of networks, is that this framework arguably provides a way of addressing a long standing debate in the field, the debate about the nature and essence of social ties. Various answers have been suggested to this question (Borgatti *et al.*, 2009; Podolny, 2001), but common to them all is the distinction between the conceptualization of ties as durable, semi-static *structures* connecting between entities and the notion of momentary action, movement or *flow*. Whether structure or flow, the ontological status of ties seems to belong to one of those insoluble debates one could trace back to a famous aphorism coined in ancient Greece (see opening quote for this chapter.) I am referring here to the puzzling relationship between that which is durable and that which is changing, between the essence of a river as a stable idea in the mind and its defining character of flowing matter. Granted that ontological questions like this will surely continue to boggle the mind, but in what follows I would like to assess the claim that micro-foundations could perhaps provide a tentative way to grapple with this question or at least represent it in a coherent manner.

The chapter is divided into two parts: in the first part, the principles of micro-foundations are briefly demonstrated through the Coleman 'boat' diagram. This diagram illustrates the distinction between different levels of social aggregation and the possible explanatory links connecting between them. To anchor the problem in a wider intellectual project, we embark on a brief excursion back to a debate between Gabriel Tarde and Émile Durkheim, followed by some helpful notions about social action from Marcel Mauss, Georg Simmel and Max Weber, finally reaching the very founding fathers of the field of social networks.

The second part of the chapter is an attempt to organize a few of the more famous social network findings into categories organized according to the Coleman diagram. This exercise helps us assess the utility of the framework of micro-foundations and it is here that we will find some open questions regarding the links between social action and networks, questions that will guide the empirical investigation.

2.2 Coleman's boat and micro-foundations

Coleman's (1990) diagram (figure 2.1) describes four types of social explanations, all of them derived from a basic distinction between two levels of analysis or levels of aggregation, the macro-level of the collective and the micro-level of individuals.[1] The first mechanism, also known as the 'situational mechanism' (Hedström & Swedberg, 1998, p 22) is represented in arrow number 1, accounting for the manner in which social conditions affect the way people might interact with one another.[2] Clearly, the type of interactions people engage in depend on the context in which they find themselves. Romeo and Juliet would most probably

[1]Some reject this distinction, claiming that there is no difference in kind between micro-level entities and macro-level ones. Granted it is true that every collective is linked with multiple individuals, but by the same token every individual is also linked with multiple collectives (Latour *et al.*, 2012). We would therefore expect some symmetry between the level of the macro and the level of the individual, a symmetry that is not apparent from Coleman's diagram. This is an important critique we shall return to in Chapter 6, discussed in light of the empirical findings.

[2]It is interesting to note that Erwin Goffman did not seem to think that relational structures would have much affect on behaviour, and to the extent it did, he suggests, it would be likely to result in a very simplified representation of that structure. He justifies his claims by arguing that face-to-face interactions are too demanding of the attention of participants to allow for much preoccupation with external factors (Gibson, 2005).

not have had the chance to interact, if it were not for Capulet's masquerade ball, a context that allowed them to shed off certain parts of their identity (affiliations with rival families,) while at the same time liberating other parts (heightened hormone levels.) In the social sciences, one could investigate how the size of a community or the level of its diversity affects the way people interact. More specifically in the field of social networks, the network of relationships is often taken as an exogenously given structure providing the backdrop for interaction, shaping who might interact with whom, and in what form.

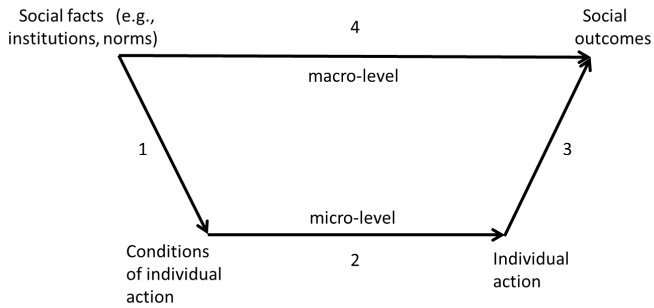


Figure 2.1: Coleman's diagram: basic version - Four types social mechanisms

The second arrow is sometimes called 'action formation mechanisms' and it represents the actual interactions at the micro-level. It consists of individuals engaging in social events, communicating, bargaining, exchanging ideas, struggling to reach an agreement, etc. These interactions might have certain results, yielding changes at the micro-level; people might decide to purchase a new product, change their political affiliation or quit smoking. More generally, this mechanism is associated with the way an individual's action might be shaped by prior interaction with others.

Finally, the set of actions at the micro-level might then have (cumulative) consequences at the macro-level, either reinforcing or changing the properties of the system as a whole. For example, people's action at the micro-level could influence the group's identity. This link between the results of micro-level interactions and systemic outcomes at the macro-level, is depicted in the diagram as arrow number 3. Links of this type could be one of two types (Abell *et al.*, 2010):

1. *Definitional:* a macro-level attribute is defined by micro-level attributes and logically determined by them. For example, the size of the population is simply the number of individuals in it, the outcome of an election poll is determined by the ratio between the sums of the votes cast for each party in the ballots, the Global Domestic Product is simply the sum of the outputs of every class of enterprise etc. As we shall see in section 2.3.1 Granovetter (1973) proposes a definitional link between the strength of a tie and the transactions associated with it. This type of micro-macro link is sometimes referred to as *supervenience* (Hedström & Bearman, 2009), *constitutive*, *analytical* (Lazarsfeld & Menzel, 1961) or *aggregational*. The difference between the micro-level properties and the macro-level is one of order, not one of kind.

2. *Contingent:* Macro-level attributes are linked to micro-level ones, but the relationship is not necessarily or logically determined. To describe this situation, one might say that the macro is 'something over-and-above' the sum of micro-states, where the difference between the 'sum' and the resulting 'macro' state is illustrated by a time lag between the moment in which micro-states have reached a certain distribution and the moment that the macro has 'caught on.' The slight slope of arrow number three, rising not only upwards but advancing slightly to the right as well, expresses this time lag, suggesting the macro needs to 'catch up' with the micro and is therefore not logically defined by it, but has a trajectory of its own, empirically contingent on the micro but not logically defined by it. One could logically (or even empirically) conceive the same micro-state associated with different macro-states, so that the relation between the macro and the micro is not that of supervenience (see figure 2.2 for an illustration.) An example could be the price of a product, a macro-level feature that is the outcome of a processes of negotiation between sellers and buyers. In some types of transactions, such as the stock-prices, the price is calculated directly by an algorithm that depends on the aggregate number of buyers and sellers. In this case the micro-macro link becomes definitional and the Coleman diagram becomes practically squared (ignoring the time-lag it takes to calculate

the stock price.) But when economists speak of price-stickiness, they may refer to the process where suppliers have market power and are hence 'price makers', while consumers are 'price takers.' Demand might be low but for some exogenous reason a macro-result is obtained which is not uniquely defined by the state of the various actors. This type of link is sometimes referred to as *empirical*, *synthetic* or *causal* or *global* (Lazarsfeld & Menzel, 1961). The difference between the micro-level and the macro-level property is one of kind, not of order.

One way to adjudicate between the definitional and the contingent link is to ask oneself the following questions: could there be a time-lag between changes at the micro-level of analysis and changes at the macro-level? Could the macro-level property be separately apprehended (measured or conceived,) independently of objects at the micro-level? If any such time-lag could be identified and if the macro-level property could be apprehended separately, we have an indication of a contingent, not a definitional link between the micro and the macro.

An example of how this theoretical framework could work in practice can be illustrated in the study of the dissemination of new pharmaceutical products (Coleman *et al.*, 1957). When studying the adoption of the product in a population of hospital doctors, the diffusion process is found to have a distinctive sigmoid (i.e., S-shaped) pattern. In the beginning, adoption of the new product is relatively slow, with the number of doctors adopting the product increasing at a modest rate. But with time, adoption rate accelerates and reaches a maximum when roughly half of the doctor population have been won over. Adoption rate then slows down steadily until almost every doctor adopts the new product. Depicted on a graph in which the x-axis reflects the progression of time and the y-axis reflects the number of doctors adopting the product, the process is expressed as a sigmoid curve.

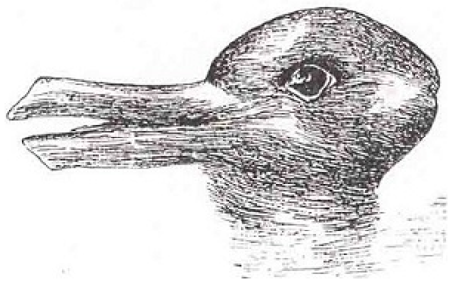


Figure 2.2: Illustrating a contingent macro-micro link - A single configuration at the micro-level (white and black pixels on the page) can be mapped into two macro-level meanings, a duck or a rabbit.

The Coleman diagram suggests a way of analyzing the process of diffusion. The population of doctors are connected to each other through an (exogenously given) network. Contingent on the contacts they have (arrow number 1) they begin to converse, exchanging their experiences and engaging in a debate about the new product, its advantages and disadvantages. The result of this interaction (arrow number 2) is a decision made at the level of the individual doctor, whether to adopt the new product or not. A simple aggregation yields the ratio of 'converted' doctors (definitional link between the macro and the micro), a number that represents a 'social outcome' at any given time. The features of the graph represent the system as a whole. Mathematically, the differential equation expressing this interaction model looks like this:

$$\frac{dx}{dt} = \alpha x\,(N - x)$$

Where the rate of conversion $\frac{dx}{dt}$ depends on x, the number of those converted to the new product and N, the total number of individuals in the system.

Surprisingly, the sigmoid pattern is only characteristic of hospital doctors, and not of those doctors working in their own practices. In the latter population, a different pattern arises, with a very rapid uptake up front, and the rate of adoption decreasing steadily with time and approaching zero when almost all of the doctors have adopted the new product. Graphically, an arc-shaped curve is obtained, not a sigmoid. Thus the characteristic structure of diffusion (which is a macro-level feature of the system as a whole) depends on the conditions of interaction for the population of doctors. This process is expressed mathematically as follows:

$$\frac{dx}{dt} = \beta\,(N - x)$$

According to Coleman's (1957) paper, the difference between the two populations of doctors is a result of the different ways in which their networks are structured. The structure of relationships in the hospital allows for more interaction between those doctors who have tried the drug and those who have not. Hospital doctors are embedded in a network that gives them access to information that is in a sense more reliable, namely the opinion of colleagues who accumulated experience using the drug. The network structure of doctors working in their own practices is different in that their network of professional relationships is much more limited. Thus, it is not easy for them to get reliable information from disinterested parties such as colleagues in a hospital, forcing them to rely less on the process of interaction and exchange and more on publicly available sources of information.

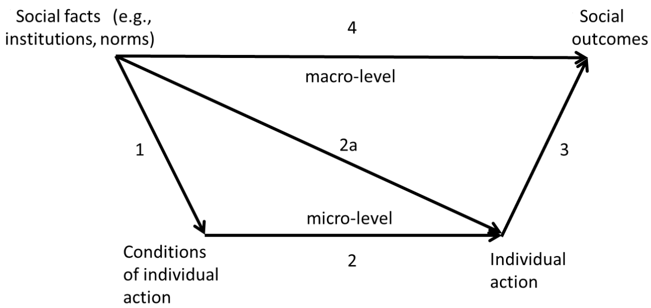


Figure 2.3: Coleman's diagram: with and without interactions - Five types of social mechanisms

The distinction between these two populations can be illustrated in the following elaboration of Coleman's diagram, shown in figure 2.3. The decision whether to adopt a new technology is made at the level of the private doctor, but it could be driven by two mechanisms: one could involve a process of interaction and negotiation between doctors in a hospital setting (arrow number 2.) The second could involve features of the entire collective, such as publicly available information or simply the fact that the proportion of 'converts' in the population has reached a certain threshold. Arrow 2*a* in the diagram reflects the possibility that choices are made independently of interaction and only contingent on certain properties of the system as a whole (Abell, 2003a).

Now, before turning to another elaboration of the basic Coleman diagram, a few words about the arrow number four, a mechanism that describes the change between one macro state and the next without involving decision making and actions of individuals. Several authors argued against a theoretic possibility for such a mechanism to occur, whereas proponents of such a process that operates 'sui generis' at the macro-level include, in the first instance, Émile Durkheim. There is a certain appeal in entertaining the possibility for such a mechanism, for it suggests a symmetry between macro- and micro-level entities, implying that there is in fact no difference in kind between these entities. An interpretation of Durkheim's logic of macro-macro level mechanisms was presented by Blau

(1972) in the context of a theory that links the effect of size (macro-level feature) on structural differentiation (also macro-level feature):

> Another assumption is implied here: the prevailing characteristics of organizations . . . can be explained in terms of the influence of antecedent conditions in organizations (or their environment) without reference to the psychological preferences or decisions of individual managers, because these social conditions greatly restrict the options of managers who pursue an interest in efficient operations. This principle derives from Durkheim (1938: 110): 'The determining cause of a social fact should be sought among the social facts preceding it and not among the state of individual consciousness'

It is important to understand what is at stake here, and to question what meaning we would like to attribute to arrow number 4 type explanations. Following the logic of functionalism, social actors do not play an important role when they make decisions that would ultimately lead to the optimal performance of the organization as a whole. I think it is not controversial that organizational changes occur as a result of decisions by individuals, managers or other parties, and that the ideas in the minds of these actors (at the micro-level) are part of the process that yield these changes. However, insofar as choices are made so as to optimize the macro-features of the organization as a whole, and assuming that only one option is the most effective, that is the decisions that actors are predetermined to make. The logic of functionalism is therefore coherent with mechanisms that operate at the macro-level, when an option is chosen because of its functionality with respect to its beneficial consequences at the level of the system as a whole. The properties of the individuals and the process of interaction at the micro-level is thus of no consequence, and the macro-conditions determine the macro-outcome, *as if* they were independent of micro-social processes. I think that if we were to interpret arrow number four in this light, we would not need to reject the possibility that some social phenomena indeed proceed along this mechanism. This interpretation coincides with Durkheim's own defence of his view of social facts, whereby;

> The totality of beliefs and sentiments common to average citizens of the same society forms a determinate system which has its own life; one may call it the collective or common conscience. No doubt it has not a specific organ as a substratum, it is by definition diffuse in every reach of society. Nevertheless it has specific characteristics which make it a distinct reality. It is, **in effect,** independent of the particular conditions in which individuals are placed; they pass and it remains (emphasis added, Durkheim, 1997 [1893])

Notice the interjection 'in effect.' Durkheim isn't proposing a process that is fundamentally detached from humans, but that *in effect* and *consequence,* proceeds *as if* it were 'independent' of variations among individuals. It is hard to object to the notion that there exist certain constraints to which members of a group are all subjected, independent of possible variations between them. To the extent that such micro-level variations exist, it is conceivable that they have no bearing on the outcome at the macro-level. Understood in this way, I am not sure we need to reject, a-priory, the possibility of such macro-to-macro transitions.

A last variant of the Coleman diagram that is used in this dissertation introduces a meso-level of analysis into the picture (see figure 2.4). The idea here is to distinguish between the meso-level of tie formation and the micro-level of social actions. It is here that we first introduce the distinction into the Coleman diagram, differentiating between ties as durable structures connecting people, and the actual flows associated with these ties. I will attempt to use the Coleman diagram to assess how these two dimensions relate to one another, and what sort of explanations we need in order to bring them together into a unified theory of social networks.

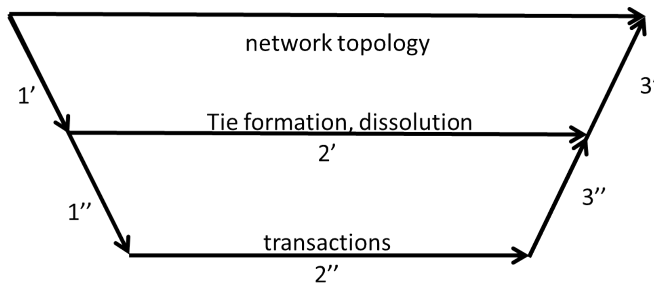


Figure 2.4: Coleman's diagram: introducing the meso level - Six different social mechanisms

2.2.1 Social (trans)actions

The links between social ties and the social actions with which they are associated, necessitate a short excursion into the nature of social action. The literature on this topic is vast (Danto, 1973), and mostly irrelevant for the purposes of this investigation. However, I shall proceed by trying to identify the defining characteristics of a social action. Perhaps one of the first scholars to make the social action a pivot of social theory was Gabriel Tarde (quoted by Vargas *et al.*, 2008):

> What is or rather what are social facts, the elementary social acts, and what is their distinctive character? [...] The elementary social fact is the communication or the modification of a state of consciousness by the action of one human being upon another. [...] Not everything that members of a society do is sociological. [...] To breathe, digest, blink one's eyes, move one's legs automatically, look absently at the scenery, or cry out inadvertently, there is nothing social about such acts. [...] But to talk to someone, pray to an idol, weave a piece of clothing, cut down a tree, stab an enemy, sculpt a piece of stone, those are social acts, for it is only the social man who would act in this way; without the example of the other men he has voluntarily or involuntarily copied since the cradle, he would not act thus. The common characteristic of social acts, indeed, is to be imitative. [...] Here is, then, a character that is clear cut and what is more, objective.

What defines something as an action for Gabriel Tarde, is the principle of imitation, which is closely related to the idea of diffusion discussed above. A different approach, and the one more popular today, is to define an action in terms of reason, intention and purpose. This idea was neatly elaborated by Marcel Mauss (2000, p 21,59), in his famous study of the function and consequences of the action of gift-giving.

> Our festivals [in New Caledonia] are the movement of the hook that serves to bind together the various sections of the straw roofing so as to make one single roof, one single world [...] The gift is therefore at one and the same time what should be done, what should be received, and yet what is dangerous to take. This is because the thing that is given itself forges a bilateral, irrevocable bond.

Thus conceptualized, the act of giving a gift is not simply a disinterested act of generosity. It is a calculated transaction, an investment, a way of dealing with a possible turn of events in a risky and unknown future. There is a reason and a purpose for this transaction. It creates a form of social debt, and is therefore (possibly) bound with another future action, the action in which this debt will be paid back. These two defining properties of social action, *purpose* and *interdependence*, will be further elaborated in a moment, but before doing so let us note a second interesting point made in this quote, namely the link between the social transaction of giving a gift, and the tie that is being forged as a consequence. The transaction and the tie are not the same thing, and 'tie formation' is not considered here a social act, in and of itself, but a consequence of the transaction(s), coupled by accepted norms of reciprocity.

This second point, about the link between the meso-level of tie-formation and the micro-level of social transactions is further highlighted in the work of Georg Simmel (1908). Consider the following qoute:

> The large systems and super-individual organizations that customarily come to mind when we think of society, are nothing but immediate interactions that occur among men constantly, every minute, but that have become crystallized as permanent fields, as autonomous phenomena. As they crystallize, they attain their own existence and their own laws, and may even confront or oppose spontaneous interaction itself.

Again we have the separation between the 'immediate interactions that occur ...constantly' and the crystallization of structure, possibly as a side effect of these interactions. Again we have a distinction between momentary events and that which has become rigid, reminiscent of Heraclitus' distinction between the flow of the river and the stable idea of a river. These here are two separate levels of analysis.

Notice the difference between Mauss and both Simmel and Tarde, regarding the origin of action. Whereas Mauss thinks that gifts have one clear *purpose* which is to 'forge' the bond, neither Simmel nor Tarde speak of a purpose for action. Simmel's actions could have a completely different purpose. But the consequences, whether intended or not, are the same: the forging and crystallization of structure, organization, and laws.

In contrast to Tarde and Mauss, Simmel does not speak of the defining properties of actions. However, on this issue Max Weber has an important insight: 'We shall speak of 'action' insofar as the acting individual attaches a subjective meaning to his behaviour ...an action is social insofar as its subjective meaning takes account of the behaviour of others and is thereby oriented in its course' (Weber, 1978, p 4). At the very least, an action needs to have a 'purpose' in order to be social. James Coleman, following Weber, claims that 'for some purposes in the theory of this book, nothing more than a common sense notion of purposive action is necessary' (Coleman, 1990, p 13). But that is not enough. For an action to have any chance of having the desired effect, it must also be understood by other(s) via common features shared by individuals, by conventions and norms (at the macro-level) that allow for them to interpret their meaning, the intention and purpose motivating them, and the nature of the expectations held by the person who carried it out.

To illustrate what Weber and Coleman have in mind, consider the following distinction between a social action and other forms of non-social behaviour, the distinction between a wink and a blink (or some other involuntary twitch of the eye;) 'two boys fairly swiftly contract the eyelids of their right eyes. In the first boy this is only an involuntary twitch; but the other is winking conspiratorially to an accomplice. At the lowest or thinnest level of description the two contractions of the eyelids may be exactly alike. From a cinematographic film of the two faces there might be no telling which contraction, if either, was a wink, or which, if either, were a mere twitch. Yet there remains the immense but non photographable difference between a twitch and a wink.' (Ryle, 1971, p 480).

The nature of these 'immense' differences was articulated by Michael Oakeshott (1991, p 15), stating that a wink 'is an exhibition of intelligence, a subscription to a *practice* and [motivated by] *reason*,' whereas a blink, he says, 'is a component of a *process* to be understood in terms of a *law* or a *cause*'.[1]

The defining properties of social action are therefore twofold (see figure 2.5.) First, what I would call ***interdependency***, here understood on two levels; in terms of the association between people ('To act is always to act *with* others' Ricoeur, 1984, p 54,) and in terms of the associations between one transaction and another, either within stimulus-response type of exchanges, or in the way every transactions is a token belonging to a type of transactions, in Oakeshott's formulation a 'subscription to a practice,' and in Tarde's formulation an imitation of other token actions (a nominalist like Tarde rejects the notion of abstract types.) So for

[1] I can imagine some counter arguments about the validity of these distinctions. Just like a wink, a blink might also be said to be functional. It is used to spread moisture and remove irritants from the surface of the cornea, and is therefore also oriented to the future similar to the way that a wink is oriented to future transactions. The distinction between reason and cause is one wrought with controversy, but for our purposes not entirely relevant.

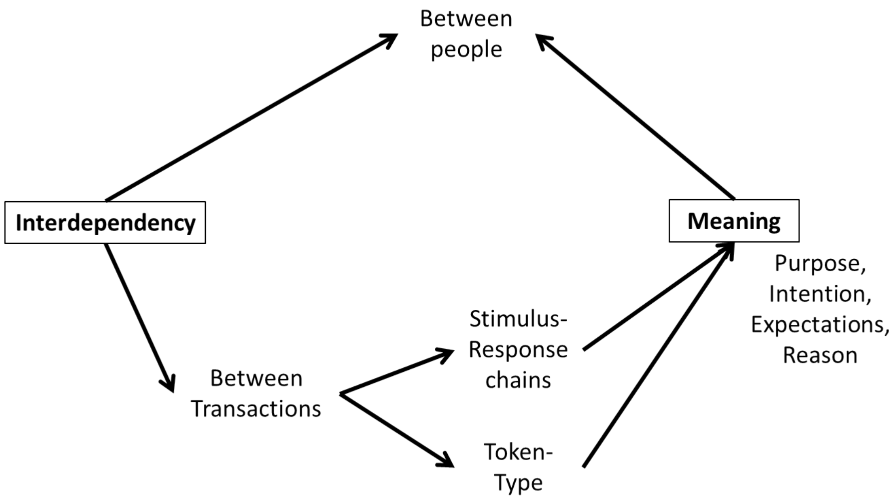


Figure 2.5: Defining features of social transactions - consisting of ***Interdependency*** and Meaning (or ***Purpose***)

example, a token wink is associated with all winks experienced in the past, and those ascribe it with meaning.

Ascribing 'meaning' to a transaction is the glue that connects a stimulus action to a response. Each and every transaction in a stimulus-response chain is ascribed meaning through the chain it constitutes, just as every word in a sentence is given meaning by the sentence which co-constitutes the words of the sentence, for actions cannot be 'understood or explained unless they are related to the actions of others' (Hedström, 2005, p 35). In addition, the meaning of a token transaction is carried from the category of human practices to which it belongs, a category whose meaning is assumed to be shared between people.

The 'meaning' of transaction is associated with its second defining property which I will refer to as ***purpose***, closely related to the notions of reason and intention. Transactions are initiated for a reason, as opposed to the automatic effect of a cause. Something has a reason if it is oriented to an event in the future, whereas something with a cause is oriented to an event in the past.

The fundamental principle in micro-foundation analysis is that social change needs to be analysed in terms of individual transactions. Sociologists adhering to this doctrine include much of the classical German tradition (Weber and Simmel), the classical Italian tradition (Pareto and Mosca) and sections of American sociology (Parsons, Merton and Coleman.) We find related modes of thinking in economics, of which both the classical and the neo-classical variants share the principle economic phenomena can be analysed in terms of an accumulation of elementary individual actions (Boudon, 1986).

It is no surprise that social scientists continue today in the pursuit of this paradigm, especially in regards to the study of social change. There is no reason why the paradigm should be confined to economics, as Max Weber has noted, where it had been widely accepted since Adam Smith's time. It has a universal nature, and according to Boudon (1986) it is 'probably one of the most important discoveries in the modern social sciences (though not always recognized as such).'

2.3 Social Networks in light of the Coleman diagram

> We concentrate on the structural properties of elementary social relations as a first step towards understanding how structure in social relations arises and evolves [...] For our purposes, relations are taken as givens. (Holland & Leinhardt, 1977)

> '...the underlying process for network change is assumed to be located in the network structure' and in the evolving 'characteristics of network members' (Doreian, 2002).

Perhaps Durkheim would embrace the ideas expressed in the quotes above, but I wouldn't be surprised if Tarde, Mauss, Simmel and Weber would be taken aback. According to the network scholars quoted above, the whole project of social action, the entire effort associated with establishing, consolidating and maintaining alliances between partners is rendered all but irrelevant for the evolution of the network.

There are two possible counter-arguments to the claim that network researchers ignore social transactions. The Durkheimians would probably say that such micro-level phenomena are nothing but 'airy chaff, posing little resistance to network effects which, given enough time, will carry the day' (Gibson, 2005). I argued above that such a Durkheimian approach needn't be rejected at the outset, especially when the entities in question follow a long term, consistent functionalist agenda, hell-bent on forming ties or dissolving them, opportunistically taking advantage of situations in order to pursue their goals. But this kind of explanation is limited if we wish to account for ties that form as a consequence of unintended, long term interactions, the kind Simmel described in the quote above and Ibn-Hazm warned of in the poem that opens chapter 1.

Another reply is to argue that it is not true that social network researchers are ignoring social transactions. This position could take one of two forms; first, one could argue that the very formation of ties and their dissolution (Doreian, 2002) are events rather like social action, in the sense that these events are more or less situated in time and place, that they are associated with meaning, intention and purpose and that they involve more than one person. All of these are defining features of a micro-level transaction phenomenon, as discussed above. But upon further inspection it would be peculiar to argue that the formation of new social ties and the act of winking at someone, for example, is the same 'kind' of thing. In contrast to a social action, forming a tie requires the alignment of interests and purposes of two people, not just one. Besides, forming a tie should be classified as a meso-level event, contingent on multiple micro-level transactions including acts of communication, interaction, exchanging of gifts etc. Finally, we still have the problem of accounting for ties that are formed as a side effect of other intentions, the consequence of a path-dependent process.

But even if we do not classify tie formation/dissolution as a social transaction, one could still argue social ties are nothing but a recurring pattern of social transactions (see section 2.3.3). To clarify, let us consider a distinction between two related terms at the meso-level. First, consider the term 'patterns of transactions.' If we have no direct and independent way of measuring a tie, we might just look at the transactions that occur between nodes and define their aggregation as a 'pattern of transactions,' a meso-level entity representing a relation between two nodes and ***defined*** by a bunch of transactions and nothing more. On the other hand, we could now speak of ties as being an entity that can be measured independently of the transactions, and that is therefore empirically ***contingent*** on those transactions. In the following sections I will try to make the case that these two concepts refer to different empirical objects. However, I will also demonstrate that some scholars disagree with the argument I am making, claiming that a tie is nothing more than a pattern of transactions. Such scholars could be considered reductionists, for they believe that any tie can be reduced to a pattern of transactions or that ties and transaction-patterns are the same thing described at a different level.

To explore these ideas further, the rest of this chapter organizes a sample of network studies with respect to their relation to the Coleman diagram as depicted in figure 2.4. Three types of work are considered: 1) some studies focus on static networks without giving much attention to events or change. However, parts of this work do establish important connections between different levels of analysis, 2) others study macro-meso links, consisting of tie formation and dissolution, and how these are related to the network topology, and 3) The third body of work study macro-meso-micro processes operate within all three different levels of networks, ties and sequences of related social transactions, constrained and facilitated by an evolving network structure.

2.3.1 Static Networks

The first body of work studies static properties of networks, often with an eye to features that distinguish social networks from other types of networks (Newman & Park, 2003). Probably the most quoted (Lazer *et al.*, 2009a) example of an unintuitive finding from this body of work is the 'strength of weak ties' hypothesis (Granovetter, 1973), a hypothesis first tested on a large dataset almost 35 years after it was published (Onnela *et al.*, 2007b). The 'strength' of a tie is defined by 'the combination of amount of time, emotional intensity, intimacy and reciprocal services characterizing a tie' (Granovetter, 1973).

This definition, in and of itself, is an example of a link between the meso-level property of the tie and the micro-level of transactions. Notice how it defines a meso-level property of the tie by reference to a combination of elements, both at the meso-level and at the micro-level. At the meso-level we have properties like the 'emotional intensity' and the level of 'intimacy' that actors attribute to their social ties as a *whole*, properties that are probably contingent on lower level transactions, but arguably cannot be reduced to them in a straightforward manner. In contrast, at the micro level, the 'amount of time' actors invest in the tie could be interpreted as a simple aggregation(?) of the time invested in each of the transactions. This interpretation is adopted by certain researches, for example, Onnela *et al.* (2007b) operationalize the strength of the tie as the aggregate amount of time people spend talking to one another on mobile phones. (see below in section 2.3.3.) Adopting the definitions discussed above to distinguish between a tie and a pattern of transactions, one would conclude that this study is not about ties but about patterns of transactions, because the only data used in the study is the data on transactions. Consequently, the relation between the meso-object and the micro-object of transaction is a definitional one. Thus, Granovetter's definition of the strength of the tie is a combination of meso-level and micro-level features, properties that are contingent on transactions and those that are analytically defined by them.

Moreover, the characterization of the strength of the tie is more than a definition - it is an theoretical assertion about the relationship between various attributes of the tie, for example, time invested, intimacy, emotional intensity and the reciprocity of services. There is an assumption here, that the time invested in a tie is (proportionally) related to the intimacy or the reciprocity associated with the tie. Empirical studies haven't always been successful in verifying this theoretical relationship (Kovanen *et al.*, 2010).

According to the hypothesis, strongly tied pairs typically share more friends than weakly tied ones. Thus, people's weak ties lead them away from their ordinary social circles into more distant parts of the network, making those ties crucial for the global connectivity of the network, but also for gaining access to new information. To take this logic to an extreme, people who haven't spoken with one another for a long time have much news to exchange, because their connection is a form of a bottleneck between distant communities, the only one by which this news can travel. In contrast, when calling a close friend, one shouldn't expect to hear news they could not have obtained from other sources.

Although this body of work does not dwell explicitly on processes of change, it does establish interesting links between different levels of analysis. There is an important link between the meso-level of tie strength and the macro-level of the topology of the region in which the tie is embedded. Strong ties are embedded in tightly knit regions and weak ties act as bridges. Tell me the level of clustering of

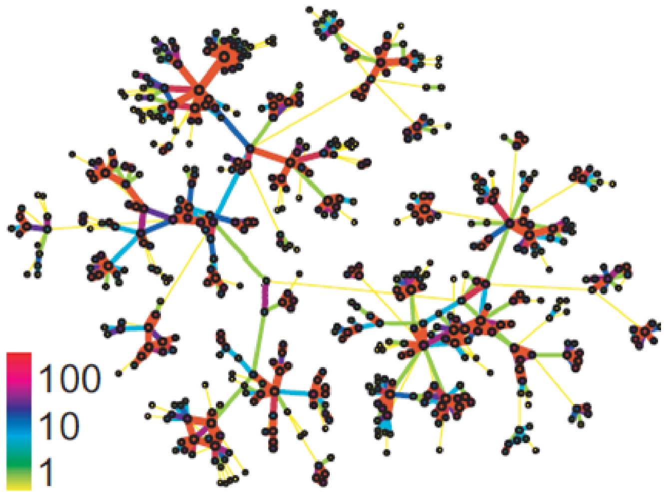


Figure 2.6: Testing the Strength of Weak Ties Hypothesis - Nodes within clusters are connected by 'strong' ties whereas clusters are connected by 'weak ties.' The strength of the tie represented in the figure by color and measured by proxy as the aggregate duration of phone calls between every two nodes (Onnela *et al.*, 2007b)

the region in which a tie is embedded and I shall tell you the strength of the tie (or the other way around.)

Another link exists between the meso- and micro-levels of analysis: first, weak ties (meso-level tie attributes) are the sites in which new information flows (micro-level transactions.) Of course, this micro-meso link has consequences at the macro-level: since bridges are activated relatively infrequently, the social network is poorly designed for the quick dissemination of new information at a global scale. Granted that global social networks do in fact exhibit properties of a small world in the sense that, ultimately, everyone is connected to most everyone else through a surprisingly small number of degrees of separation. However, since the most important ties for connectivity are activated so rarely, in practice diffusion on a global scale is hindered - not only by the structure of the network, but also by the 'burstiness' of activity (Karsai *et al.*, 2011).

A functionalist approach would immediately raise the following question: if they are not 'designed' for the dissemination of new information, why are they structured this way? Investigations of the static properties of social networks suggest that humans tend to departmentalize into clusters, each cluster 'circling its wagons,' to create structures that are instrumental for establishing and maintaining social capital (Karsai *et al.*, 2011; Coleman, 1988; Onnela *et al.*, 2007b).

I would like to introduce another network related study to show how features at the macro-level of analysis could directly impinge on individuals, leading them to make social choices without the process of interaction or conscious deliberation, as in arrow 2*a* in figure 2.3. It is as if macro-properties operate 'behind the backs' (Hedström & Bearman, 2009) of the actors, without them being fully aware of the reasons (or rather causes?) that have led them to choose as they have. I am referring to a study (Bearman *et al.*, 2004) of sexual networks of adolescents in a high school of roughly a thousand students in the Midwestern United States (see figure 2.7). The network has a few properties that are rarely found in social networks. First, it is not a small-world network, having neither the random ties that create short average paths, nor the high level of clustering that is so common to other types of social networks. Triangles, which are all but ubiquitous in friendship networks, are found here only once (close to the upper right corner of the diagram.) The reason being is that a triangle would necessitate at least one sexual relationship between two persons of the same sex, and the study was made on a population which was (reportedly) mostly heterosexual.

Besides triangles, the next order of closed cycles (and the minimal one possible in a heterosexual population) is cycles of order four, representing a situations of pairs switching partners, say from ($\langle \text{Male}_1, \text{Female}_1 \rangle$, $\langle \text{Male}_2, \text{Female}_2 \rangle$) to ($\langle \text{Male}_1, \text{Female}_2 \rangle$, $\langle \text{Male}_2, \text{Female}_1 \rangle$). Such a switch would constitute a cycle of length four. From the point of view of one of the males, he has formed a partnership with his ex-girlfriend's current boyfriend's ex-girlfriend. However, we find no such cycles of order four in the network, and there appears to be a norm that makes people avoid such pair swapping excercise. This avoidance most probably springs from an analog to an incest taboo (Moody, 2009), since a relationship that would close a four-cycle appears too intimate. Dynamically this is an interesting type mechanism since the prohibition law develops over time, a partner that would be acceptable at time t_1 is prohibited in time t_2. The prohibition is localized in place and time, but it has an effect on the macro-properties of the network, increasing the length of cycles, making it different from 'small world' networks and

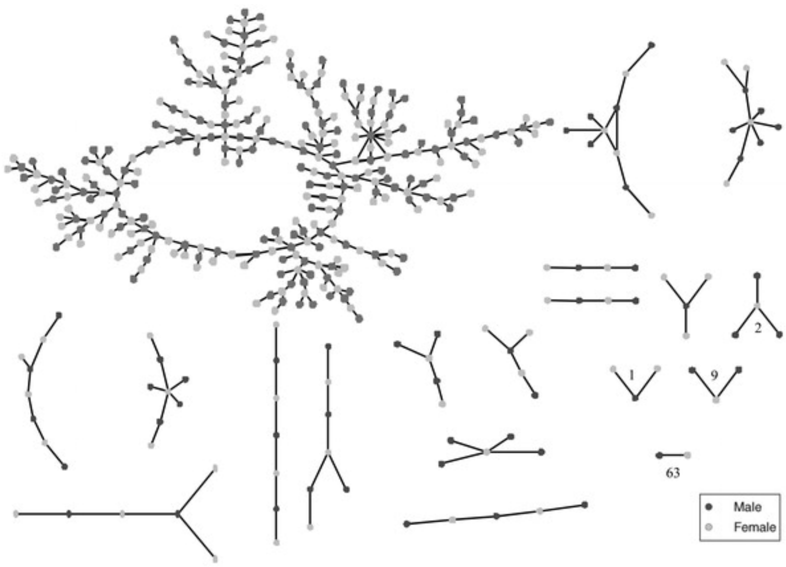


Figure 2.7: Romantic network of adolescents in an American high-school - the structure of sexual relationships observed over a period of eighteen months, every node representing an individual, ties represent sexual/romantic relationships. Taken from Bearman, Moody & Stovel (2004)

hence prolonging the time for sexually transmitted diseases to diffuse throughout the community.

Thus, the first body of work treats the network as a relatively static structure, one that does not change much over time. Any micro-transaction that takes place in the network is assumed to unfold in an exogenously given structure, pre-determined and relatively unchanging.

2.3.2 Macro-Meso links

But social networks do evolve, people who were once strangers or distant acquaintances become friends, and close friends drift away. Thus, the second body of work studies the evolution of networks themselves, the mechanisms that govern the likelihood of changes at the level of ties, meso-level events that may have consequences at the macro-level of the network. These so called 'network events' are of two types; first, the 'event' by which properties of individuals change when they acquire new behaviour patterns, changing their smoking, eating or drinking habits, altering their political affiliations etc. These changes are understood to be influenced by people's position in the network, and are therefore known as *influence* models (Friedkin & Johnsen, 1999; Robins *et al.*, 2001b). In contrast to influence, *selection* models consist of mechanisms explaining how people select their friends, a dynamic consisting of changes in the status or the properties of the relationship between two individuals. People can form new ties, disband an old ones (Robins *et al.*, 2001a) or else change the affect attached to existing ties, say from positive feelings about someone to negative ones (Doreian, 2002). Notice that the literature does not refer to these as 'actions,' but as network events.

Clearly, people's personal networks change most extensively when they change their life's circumstances, when they move between jobs or relocate for example. Such changes provide an opportunity to study in detail the process of tie formation and dissolution. They also provide an opportunity to compare between one's personal network in different contexts. Since the same person establishes relationships in different settings (different conditions for individual action, arrow number 1 in the Coleman diagram,) we might expect certain things to change and others to remain the same. One thing that remains the same, at least at first, are the properties of the person who moved. Specifically, consider the 'social brain' hypothesis (Dunbar, 1998; Dunbar & Shultz, 2007), the theory that states that people's brains are wired with stable cognitive properties that influence the structure of their personal network. These properties are unique to individuals, and may differ from one to the other in terms of an individual's preferred size of their personal network for example, or in terms of the way individuals prefer to allocate time to their different friends.

The social brain hypothesis suggests that geographical relocation would have an effect on personal networks in the sense that the identity of one's contacts might change, but the network could still preserve its original structure. This hypothesis has been actually confirmed in a rather interesting empirical study (Saramaki *et al.*, 2012) of 30 students who just completed school and moved away from home to attend a university elsewhere.

A considerable variance between individuals was identified, in terms of the distribution of time each allocated to their alters; some prefer to have just one or two best friends and allocate most of their time to them, and hardly any time to anybody else. Others allocate time more evenly between numerous acquaintances. After the relocation, the composition of the personal networks changed substantially, with many new people entering and old ones leaving it. However, each individual maintained the way they distribute their time among their friends. Thus, each individual seems to carry a 'social signature' in the way they distribute time among their alters, signatures that vary remarkably between individuals. And though life circumstances may change the identities of those they interact with, these signatures stay surprisingly persistent.

The structure of the network at the macro-level is therefore contingent on the possibility of matching between different types of structure of cognitive fingerprints. Relocation of a group of people into a new context is like a shock to the system at the macro-level, the upper left corner of the Coleman diagram. There follows a negotiation between people, and the process of friendships formation is constrained by the cognitive structure that each individual brings into the new context, resulting in a final network in which the constraints of each of the individuals is fulfilled, more or less.

Relocation is an exogenous event that prompts all three types of network events: tie formation, dissolution and changes in one's character. But there are other types of mechanisms that are responsible for such events, three of the most well studied ones consist of the popularity effect, homophily and triadic effects.[1]

In all three types of mechanisms, an existing network structure is exogenously given at the upper right hand corner of the Coleman diagram, consisting of individuals, their properties and the ties that connect them. The different mechanisms mobilize the occurrence of events across the network, at times working in the same direction to strengthen certain topological properties, at times working in opposite directions (Wimmer & Lewis, 2010). These local changes then have cumulative effects at the macro-level of the network. Thus the upper part of figure 2.4 unfolds, linking meso-level network events with the evolving macro-structures.

[1]For abbreviations and nomenclature see the glossary on page xii.

2.3.3 Macro-Meso-Micro links

The preceding section covered mechanisms labeled 1′, 2′ and 3′ in Coleman's diagram 2.4. What remains is to complement the picture with the bottom part of the diagram. The question here is how to capture both levels at once - the (possibly changing) topology of the social network and the sequences of related social transactions.

Consider what we can learn by focusing on the exact moments in which people initiate telephone calls. These are known to reveal sudden bursts (Barabási, 2005) of activity: patterns that are neither completely regular nor random, but are characterized by long periods of silence followed by quick succession of action, producing inhomogeneous distributions over time. But this type of research ignores the defining properties of social transactions and particularly the links between them, the causal chain of stimulus and response. For example, receiving a message from someone can be interpreted as a stimulus that may be followed by some kind of response, either by way of replying to that message, or by forwarding it on to a third person, or by any other type of social transaction one might want to think of. This interdependency between transactions is lost if we simply aggregate all of them and study their distribution in time.

Thus, we are looking for research that explicitly acknowledges the interdependency between transactions and studies the relationship between them. Before demonstrating appropriate examples, let us turn to strategies used by network scholars when addressing the issue of social transactions.

The Durkheimian Strategy: Ignoring Transactions

The first strategy was already introduced in the quotes opening section 2.3 (Holland & Leinhardt, 1977; Doreian, 2002). The subject matter of social networks is presented as the social, durable structure, an entity that follows its own laws independently of micro-social interactions. This was the favoured strategy, quite explicitly stated by the founding fathers of the field of 'social networks.'

It is widely accepted (Mitchell, 1969; Caulkins, 1981) was one of the first scholars to have explicitly used the term 'networks' in order to denote a social field was John Barnes. An anthropologist who recently passed away, Barnes was also the keynote speaker at the 1982 Sunbelt International Social Network Analysis Conference. But he is not only credited with a systematic and rigorous use of the term ‘social networks,’ he is also one of the first (Mitchell, 1969) to make a clear distinction between ties and transactions. He writes, and Mitchell (1969) quotes him on this, that the ‘social network’ is a concept applied to ‘what is left behind when we leave out groupings and *chains of interaction*’ (my emphasis.) The reason he gives for leaving interaction out of networks boils down to their ephemeral nature: ’These units’, he says ’do not necessarily persist through time, nor does their membership remain fixed’ (Barnes, 1954). Therefore they do not qualify to be relevant to the network. One can almost hear the voice of Durkheim echoing through these words, claiming that the macro evolves *in effect* independently of variations at the level of individuals.

Figure 2.8: Portrait of John Barnes The front cover of his autobiography Barnes (2008)

Mitchell agrees with Barnes that the concept of social networks should be kept analytically separate from the concept of social interactions. He develops this idea further and discusses a controversy in the literature regarding the question of how we should understand the ‘content’ of a social tie, while dismissing arguments advanced by other scholars who insist on the relevance of the ‘flow of information’ for the study of social networks. Instead, he suggests that the study of networks should be limited to ‘the normative context in which interactions takes place,’ this ‘normative context’ consisting of interpersonal expectations that regulate transactions and their permissible interpretation.

Definitional Strategy: Patterns of transactions

Consider how the following definition of communication networks compares to the ideas expressed above:

> ‘Communication networks are the patterns of contact that are created by the flow of messages among communicators through time and space [a message refers] to data, information, knowledge, images, symbols and any other symbolic forms that can move from one point in a network to another or can be co-created by network members’ (Monge & Contractor, 2003)

Ostensibly, it is precisely what Barnes and Mitchell hoped to leave outside of social fields of networks, that Monge & Contractor (2003) bring right back in. Moreover, they seem to adopt the opposite extreme to Durkheim, the idea that there is nothing ***but*** transactions, and that social networks are ***defined*** by the aggregate of all transactions observed within a specified time-window.[1]

One could now apply the distinction between ties and transaction patterns (as introduced in section 2.3,) concluding that Monge & Contractor (2003) are not speaking of social ties but of transaction-patterns, i.e., a definitional conceptualization of the link between the meso- and the micro level of analysis. The meso-entity thus defined supervenes on the micro, with no change in any of the meso-features logically conceivable without a change at the micro-level. This approach is shared by different scholars throughout the decades. Consider for example the claim made by the great George Caspar Homans in relation to groups (Homans, 1951, p 84): 'a group is defined by the interactions of its members.' Only interactions and nothing 'over and above' interactions. And if we take the tie to be nothing other than a group of two, we see that Homan's definition of the group collapses into something similar to what Monge & Contractor (2003) have in mind. Max Weber also seems to belong to the group, when he says that 'the social relationship ... thus consists entirely and exclusively in the existence of a probability that there will be a meaningful course of social action' (Weber, 1978, p 26).[2]

This approach raises two concerns, one theoretical and the other methodological. If the social object we focus on is just a pattern of transactions, what meaning should we ascribe to this pattern? What does it represent? What kind of consequence does it have? Is it really the pattern, in and of itself, that is the object of interest? Or is it something else that covaries with this pattern?

One answer could be that transactions-patterns should be treated as *indication* of underlying social ties. The object of investigation is therefore not the transac-

[1]But perhaps this is a misinterpretation: the authors say that 'patterns of contact' are 'created.' What is unclear is how are these patterns created and what are the consequences of this creation? Are they simply created in the minds of the scientist observing and looking for them in the data? Or are these patterns created also in the minds of the actors in the network and affecting their actions? Unfortunately, Monge & Contractor (2003) do not explore these questions.

[2]Whether 'probability' of transactions and their aggregation amount to the same thing is an open question, but I suppose that they are very much closely related.

tions per-se, nor a recurring pattern in which they appear (although that could also be of interest, such as for example in the highly influential work by Barabási (2005); Barabasi (2009)), but the latent, more meaningful social ties, or some other latent object that covaries and becomes manifest only indirectly through observable transactions. Indeed, it seems reasonable to assume that when people interact frequently, they may know each other and may be related to each other in a meaningful social bond. This argument chimes with the principle that social ties are not merely defined by transactions, and that despite measuring transactions the real interest lies in more meaningful social ties (see section 2.2 and the related discussion in section 6.2.)

But recall that transactions were defined in terms of their interrelatedness, each transaction a link in a chain of related transactions. The process of transforming the transaction data-set into a set of nodes and transaction-patterns whitewashes precisely what made the transactions 'social' in the first place, emptying them from their purpose and 'meaning' in the Weberian sense of the word. If we simply aggregate the transactions into streams of messages without attention to sequences of social action, it is not clear that we arrive at 'rock-bottom explanations,'[1] which is what the doctrine of micro-foundations requires.

A related problem on a methodological level will be elaborated in the next chapter, but in essence it is this: there are numerous ways to transform data streams of transactions into transaction network models. This is because it is not clear what types of transaction-patterns are relevant. As we shall see, collapsing the rich properties of transactions into network models requires the modeler to make numerous ad-hoc decisions, and there is no standard against which one could validate whether one process of model construction is superior to another.

Third Strategy: Taking Transactions Seriously

How can we study the topology of social networks while taking not only individuals and ties into account, but also the transactions between them? I shall begin by making a controversial argument for the difference *in kind* between social ties

[1]The idea that micro-foundations should reach 'rock-bottom' explanations that involve individuals and their social action was advanced by Watkins (1957)

and social transactions, and then briefly review some studies that look not (only) at the aggregate of transactions but about their interrelatedness.

It was Georg Simmel in the quote above (see page 35) who contrasts most clearly between social transactions and ties, when he speaks of 'super individual systems' (such as ties), developing their own laws, and even 'confronting and opposing interaction itself.' Conceptualizing the tie as a channel through which transactions flow, makes it surprising that such a channel serves not to facilitate the flow of transactions but, quite the contrary, to block them. Proving that ties block transactions could thus be seen as a rejection of the claim that ties and transaction-patterns refer to the same thing. But Simmel's assertion can be empirically illustrated using an impressive study of low-income African Americans from a mid-western city in the United States (Smith, 2005), an investigation of the causes for low-employment rates in these communities. Previous studies have challenged a widely held assumption, that continued unemployment is partly due to the community's isolation from sources of information and influence among those employed. Members of the community were found to be well connected with employed individuals, friends and relations who could provide timely information regarding job opportunities in their own workplaces. They *could* provide that information, but chose not to do so.

Those employed did not act to improve the employment situation within their network; they were often reluctant to wield their influence or even to provide information about job opportunities. The reason for this was that they were concerned that job seekers in their networks might act 'too irresponsibly on the job, thereby jeopardizing contacts' own reputation in the eyes of their employers'. Smith (2005) concludes that one primary reason for sustained unemployment in the community is not the lack of potentially useful relationships, but the failure to mobilize these relationships and to realize the social capital for unemployed individuals.

And so we arrive full circle back to Simmel's claim about ties that stop information flow, since precisely such flow was throttled between the employed and unemployed individuals, merely because of the concern of the employed individuals to maintain the integrity of their ties with their employer. This case provides motivation for a distinction *in kind* between social ties and transaction-patterns. '[R]elationships have ontological status even when they are not being directly acted upon. Two people, for instance, can be considered 'friends' even when they are not interacting' (Gibson, 2005). The contrary may also hold true, when social transactions take place without being associated with a meaningful social bond. The separateness of the two concepts is neatly expressed in the title of Adams (2010) letter to the Proceedings of the National Academy of Sciences (PNAS): 'Distant friends, close strangers.' Here the words 'close' and 'distant' refer to physical proximity, the four-word-title suggesting that face-to-face interactions, just like other forms of human contact, is analytically separate both from the notion of social relationships, and from the normative commitment and social bond this notion entails.

Finally, recall the three popular mechanisms operating at the meso-level of the tie, namely popularity effects, homophily and triadic closure (Snijders *et al.*, 2006). Despite the frequent referral to them in the social networks literature, many are still worried with the 'inadequacy' (Snijders *et al.*, 2006; Newman, 2003) of these mechanisms, partly because they are still not well understood in terms of the micro-social processes that give rise to, and sustain them. It is infrequent that people 'decide' to establish new relationships or dissolve old ones, since changes in the status of ties are themselves a meso-level event at the level of the tie, the result of a culmination of micro-events involving various opportunities, choices, circumstances and mutual social transactions (Lazarsfeld & Merton, 1954; Hartup & Stevens, 1997). Even the literature on friendship (Hartup & Stevens, 1997) makes an important distinction between deep-structure (based on reciprocated emotions or perceptions) and surface structure (social transactions), and exploring the interaction between these two levels implies that they are not only interdependent, but that each enjoys some level of (ontological) autonomy.

How would one go about exploring social networks and the underlying sequences of social transactions? One way to go about it is to investigate those sequences of transactions in and of themselves, such that A calls B, prompting a call from B to C, prompting the latter to return the call etc. One such study by Kovanen *et al.* (2013) followed sequences of related transactions, investigating how these depended on attributes of actors engaging in them. For example, transaction chains consisting of men were found to be invariably shorter and less complex than those consisting of women.

Another group of studies (Butts, 2008; Brandes *et al.*, 2009; de Nooy, 2011) seeks to predict how sequences are likely to unfold given a data-set of transactions. These studies model the expected waiting times between successive transactions, depending on the identities of actors who may initiate future transactions. Though the models predict temporal dependencies between transactions, they, too, do not consider the links of meaning between them. That is to say, when transactions follow one other in succession, we do not know whether one has prompted the next, or whether there is any interdependence between these transactions from a socially meaningful perspective. Analytically, the objective of these studies is to develop new statistical methods for the analysis of transaction data-sets.

A common feature in this type of work is the focus on transactions as events, just like network events. To the extent that there is reference to a distinction between ties and transactions, this is taken to be one of order, not one of kind (this is made explicit in a paper by Butts, 2008, p 191-192). Consequently, there is no attention to the study of co-evolution processes that operate between the meso-level of ties and the micro-level of transactions.

A second line of inquiry takes the network of relationships as exogenously given and follows the way transactions unfold within this network. Diffusion studies would arguably fall into this category, depending on how we might understand the term diffusion. Many such studies, specifically those that explore the spreading of disease, follow the way a certain property of a focal actor's friends affect changes in properties of the focal actor. But I am not sure we would want to say that changes of an actor's properties should be considered a 'social action.' Gabriel Tarde would disagree to some extent, recall from the quote above (page 34) that a transaction is '...the modification of a state of consciousness by the action of one human being upon another.' And though infection by a virus might modify one's state of consciousness, I do not think that this is what Tarde has in mind when he speaks of a social transaction. What is lacking from the infection is the second defining property of social action, and that is the notion of purpose and reason.

Another kind of research that might fall into this line of inquiry belongs to a branch of social psychology known as the study of expectation states. These studies are based on the premise that social actions are shaped, facilitated and constrained by cognitive states with distinctive 'structures' (Balkwell, 1991). For example, the rate and nature of utterances in conversations have been shown to be shaped by properties of social relationships such as authority and deference (Shelly & Troyer, 2001; Johnson, 1994), or the degree of familiarity and intimacy (Boxer, 1993). In a particularly interesting study, David Gibson (2005) devised a classification for turn-taking in a conversation (which he calls participation-shifts.) He traced the discussions of ten groups of managers who frequently work together and concludes that the pattern of their turn-taking is contingent on their relative positions in a network of friends, co-workers, and reporting relationships.

Although findings seem to vary from paper to paper, the study of interactions provides strong evidence that the process of social exchange hinges on 'the structure of the interaction' (Hedström, 2005). In other words, exogenously given social ties govern practices of communication and interaction, the mechanism represented by arrow number 2 in the Coleman diagram (figure 2.1.)

The other direction of influence from transactions to ties (arow number 3) has been studied to some extent in the context of the formation of dominance hierarchies (Fararo *et al.*, 1994; Skvoretz & Fararo, 1996). These studies formulate and test the mechanisms by which dominance relationships are shaped by a succession of dyadic encounters between animals. However, some scholars are concerned that these studies follow a 'stylized account of interaction', and do not incorporate 'insights into conversational rules' that regulate the transactions of human societies (Gibson, 2005).

One rather remarkable exception is a paper (Liben-Nowell & Kleinberg, 2008) that investigates the formation of diffusion networks from a detailed study of micro-social transactions. The authors trace a process by which massively circulated Internet chain letters spread on a person-by-person basis. To their suprise they discover that the network representing the flow of chain letters is very different from the 'small-world' network one would expect. Instead, the network progresses in a narrow but very deep tree-like pattern, continuing for hundreds(!)

of steps, as depicted in figure 2.9. In one of these networks, the median distance to the root over all nodes was nearly 300, and more than 90% of the nodes had exactly one child.

Assuming that the signatories of the chain-letter were connected in a typical 'small-world' network, how did this diffusion structure come about? Why doesn't it resemble the structure of the social network from which it springs? What could have possibly been the series of actions taken by the email users, so as to produce this unusual structure? Even after modelling a network in which only a fraction of the recipients forwarded the chain letter to their friends, this tree type structure could not have been obtained.

The researchers then added a few extensions to the basic model of diffusion, emulating the way people use the technology of emails. First, they modelled the asyncronous nature of emails, having each recipient wait a length of time before acting on the message. Second, they introduced three types of responses on the part of recipients: the recipients could either discard the incoming mail, they could forward it to their contacts or they could hit 'reply-all' and group-reply to the set of corecipients on the original email message they received.

These two extensions had a 'serializing' effect in networks with tightly knit regions, because multiple recipients of an incoming email would act on the message sequentially. The first might add her name to the list and only then forward it to the second. The second receives the list twice would ignore the first message with the shorter list of names, adding her name to the longer list and forwarding it on. Consequently, instead of having many lists with various sequences, we obtain lists that become incrementally longer. This model of how people use the email message to interact, along with the asynchronous nature of social transactions, produces 'runs' of nodes in which each node has exactly one

Figure 2.9: Chain letters produce unusual network structures - not small world networks but trees (Liben-Nowell & Kleinberg, 2008)

child, precisely the structure that empirically observed in the data. There is an explicit distinction here between the structure of the social ties (who is connected to whom) and the structure of transaction-patterns (which transactions are ignored and which are acted upon). The first structure has the features of a small world, while the second has the has a structure resembling a ring lattice (Watts & Strogatz, 1998). This study takes all the ingredients of the Coleman diagram in figure 2.1. It takes the small-world network as exogenously given, adds the micro-social rules of interaction that depend on the specific technology of emails, and yields a diffusion structure of a very different kind than the one that originated it.

2.4 Summary and Reflections

This chapter introduced some of the key theoretical concepts in the literature of micro-foundations, focusing on the Coleman diagram and the way it organizes different types of social explanations. Two types of macro-micro links were discussed, the *contingent* and the *definitional* link. The theory was then applied in the context of social networks, where a distinction was made between two types of meso-level entities: *social ties* that are contingent on transactions, and *transaction patterns* that are defined by them. With this analytical framework in place, three types of network studies were identified:

1. ***static networks*** The first type of work focuses on static networks of social ties, searching for interesting patterns in the data. Already here we identify micro-macro links, for example the link between the strength of a tie (a property at the meso-level) and the topology of the region in which it is embedded (a property at the macro-level.)

2. ***macro-meso*** The second group of studies investigate network dynamics, but limit themselves to three or four 'network events,' consisting of tie formation, tie dissolution and the changing properties of ties and individuals. This type of work includes virtually all the empirical work based on longitudinal panel waves of traditional network datasets.

3. ***macro-meso-micro*** The third body of work recognizes and acknowledges the existence of underlying social transactions. Within this body of work we identify three strategies to deal with transactions. The Durkheimian strategy acknowledges transactions but maintains they are irrelevant for networks. This strategy was supported and argued for by the founding fathers of social network analysis. The second strategy focuses on transaction patterns (the link between the micro and meso is *definitional*.) The third strategy is to take transactions seriously and study interdependencies among them. Social ties were understood as *contingent* on transactions, and their structure was qualitatively different from the transaction-patterns.

The network literature that links the micro and the macro is vast, but there are only diffuse attempts to articulate precisely how the structure of social ties give rise to patterns of interrelated communication transactions, and how those in turn shape network structures (co-evolution mechanisms.) A full-fledged and systematic synthesis of the 'interaction order' (Goffman, 1983) and what we may call the 'network order' is difficult to come by. One reason is perhaps the difficulty in gaining access to independent data on transactions and social-tie data (for an exception see Quintane & Kleinbaum, 2011). Two such sources of data would allow for a direct test of the hypotheses involving both levels (although this is not necessary for studying such micro-macro link, as will be argued in Chapter 6.)

Another reason suggested by Gibson (2008) is based on institutional inertia, driving a wedge between the perspectives of 'interactionists' and network theorists. Network analysis is more amenable to quantitative methods (Gibson, 2005) because much of it is concerned with the effect of social network structures on properties of individuals (Burt, 2001; Podolny & Baron, 1997), issues that lend themselves to statistical methods of graphs. In contrast, the study of sequences of social transactions and interaction has traditionally come under the purview of more qualitative research (Gibson, 2005). Historically, the reason for this has probably been the difficulty to access systematic data-sets of social transactions, precisely because of their ephemeral nature. However, this is now rapidly changing and statistical methods are gradually being applied for the analysis of large transaction data-sets (Lazer *et al.*, 2009b; Watts, 2007). However, laying the foundations for a project that bridges this divide is intellectually attractive for reasons that were highlighted in the beginning of this chapter, and as we shall see in the following chapters, they have practical, methodological and theoretical merit.

To fill this gap in the literature, the 'big question' driving this dissertation asks us to identify mechanisms of co-evolution between communication transactions and network structures in the context of email communication. Specifically, we are asked for an empirical account of the links between three levels of analysis: the macro-level of network topology (notions of density, transitivity or even centrality,) the meso-level of the individual and the tie (their properties, for example) and the micro-level of the individual engaging in social transactions. To get a handle on this big question, the next (methodological) chapter asks a straightforward and practical question: how should one go about constructing a network model, given a set of digitally mediated transactions?

3

Forging networks from their micro-foundations

In Ersilia, to establish the relationships that sustain the city's life, the inhabitants stretch strings from the corners of the houses, white or black or gray or black-and-white according to whether they mark a relationship of blood, of trade, or authority, agency. When the strings become so numerous that you can no longer pass among them, the inhabitants leave: the houses are dismantled; only the strings and their supports remain.

Italo Calvino (1978)

Networks are phenomenological realities as well as measurement constructs.

Harrison White (1992, p 127)

3.1 Introduction

Given a dataset of transactions between individuals within an organization, how should one construct a social network? This question is not merely an academic excercise, but one facing practitioners whose job it is to analyze networks. To illustrate, consider a problem I came across in a professional context. 'Real Impact Analytics' is a start-up company based in Brussels and specializing, among others, in the mining of data produced by mobile network operators in West-Africa. Part of the analysts' job is to transform mobile communication data into network models for purposes of visualization and analysis. To their surprise, the analysts found that their models had the unusual property of very low reciprocity, much lower compared to similar mobile communication networks in developed countries. Large number of mobile users engage in calls wherein one user invariably initiates the calls, the other never returning it. It turns out that many of those calls are carried out between city workers and their poorer friends and relatives in rural areas who could not afford to initiate the call. The medium of communication implements a price-model which, in the presence of inequality between caller and callee, has an effect of producing highly asymmetric transaction network models.

This situation presents a challenge for the network modelers, not unlike the problem presented in figure 2.9, where the network of transaction patterns does not reflect the expected social network associated with the transactions. Technically, modelers could take every call made and present it as a tie in the network, but this would yield a very tightly knit network with transaction-patterns nearly homogeneously distributed. This problem is typical to mobile communication networks, and in most cases an adequate solution consists in removing non-reciprocated transactions. However, doing the same thing in the West-African context would yield a highly disconnected network with very low transitivity, and again a social network of transaction patterns with features very unlike the small-world networks one would expect to find in a social context. The West African datasets presents a special context that raises practical challenges for model construction and analysis.

Figure 3.1: Real Impact Analytics - was founded in 2009 by Sébastien Deletaille and Loïc Jacobs van Merlen. The company specializes in data mining, strategy and business consulting for telecoms in West-Africa and other developing countries.

This chapter explores the methodological issues that are associated with the gap between networks of social ties and networks of transaction patterns. The chapter consists of a general methodological discussion and a more focused discussion relevant to the empirical chapters. The general discussion reviews the literature with an eye to the way researchers have addressed the gap between social ties and transaction patterns. As we have seen in the previous chapters, theoretical work on this issue can be traced to the very beginnings of the field of social networks, and in the 70's and 80's, researchers have accumulated numerous empirical insights into this issue. However, as we shall see, this has become an urgent practical and methodological issue only since the surge in the availability of large data sets of social transactions. Building on these general methodological issues, the discussion then focuses on the data and methods used in the empirical chapters, namely the famous email dataset from ENRON just before the company went bankrupt.

In one important sense, the question of how to construct a network model is more pressing for digitally mediated transaction datasets (DMTD) than for questionnaire based, traditional network datasets (TND.) Granted, the latter have their own set of issues: how to collect the data, how to design the questionnaire so as to minimize recall and bias issues, etc. The challenges are huge, but they were mostly located in the process of data collection (Marsden, 2011; Hogan *et al.*, 2007; Carrasco *et al.*, 2008). However, once the data is available, it is already formatted in network form, and the process of modelling is straightforward for the simple reason that there are not many available alternatives nor is there discretion needed on the part of the modeller. The process of eliciting the data is designed in such a way, that the responses feed directly into the network model, and it is the interviewees who need to make the effort, to transform their experiences and judgements into answers that are already tailor-made for the network model.

The reverse happens in the construction of network models from DMTD. Here, the collection of the dataset is practically a non-issue. Huge datasets are produced in great quantity and detail, gushing out of machines as a by product of the auditing processes. Unfortunately, the datasets are not formatted in network form, but in a form that is optimized for the purpose of billing, diagnostics, monitoring, maintenance and control of information and communication technology (ICT) itself. But these objectives are not the only factors shaping the data. There are additional technical constraints as well as legal and ethical ones (such as demands on users' privacy,) data corruption issues, data redundancies and other unintentional side-effects that shape the data (see section 3.4.2.) The result is a data-structure that may or may not coincide with the choices that would be optimal for the research of social networks. Researchers must make do with the available data, accepting its given format, forming research questions that cater to the opportunities available in new types of data. Some (Pisani, 2010) have argued that this kind of data guides research to be more data-driven than theory driven.

This is probably one of the less discussed differences between the research of TND and those based on DMTD, namely the locus of the intellectual effort required, when constructing the social network model: eliciting TND requires most of the effort before and during data collection, both in terms of designing methods for eliciting the data and in terms of the cognitive effort on the part of the interviewees.[1] Analyzing DMTD puts all of the intellectual effort on the network modeler, after the data has been collected. The following section continues to discuss the issues that distinguish between networks based on TND and those based on DMTD.

3.2 From traditional data-sets to new ones

In spite of the development of new methods and increasing availability of DMTD, Marsden (1990), in a recent review of data collection methods repeated his two-decade old claim that network scholars still continue to rely widely on TND to advance substantive network theory (Quintane & Kleinbaum, 2011). The most common way of eliciting TND is through interviews or surveys. This method chimes with epistemic realism: it assumes that there is one single 'social-network' out there in the 'real world,' and that 'access' to it is ideally obtained by asking people about their social relations, sometimes asking them to give an account of the subjective meaning they acribe to network positions and features (Krackhardt, 1987). This epistemic commitment is expressed in the quote by Harrison White (1992) that opens this chapter, pointing to two different 'networks,' one in the real world and one in the minds of scientists that study it. Accordingly, individuals in the real world are busy building their networks (a latent construct,) and while they are at it they leave traces (manifest construct.) These traces are then painstaickingly collected by the network modeller, whose job it is to reverse

[1] But see Bearman & Parigi (2004) for challenges encountered while interpreting survey elicited network data.

engineer the evidence and to reassemble the social network that might have given rise these traces (as in figure 2.9.)

Traditional network datasets are elicited through surveys and questionnaires (Marsden, 2011), filled in by individuals who report their contacts on the basis of the questions posed to them. Each questionnaire is then transformed into a star shaped personal-network. The stars are then aggregated to create the full network. The survey methods raise typical issues common to methodology in the social sciences, such as recall issues, bias, reliability etc. But there are challenges that are particular to methods of eliciting data for the purpose of constructing social network models; the method is costly, survey data is limited in terms of the number participants and the kinds of relationships (trust, friendship, kinship...).

But there are also other challenges, specific to the process of collecting relational data. Interesting among them are the patterns that appear to suggest that respondents think of their alters in terms of affiliation groups and not in terms of one-to-one relationships. When reporting names of acquaintances, respondents tend to group the names they report, each group consisting of interconnected contacts. The pauses they make between the utterance of one name and the next are systematically shorter when the names belong to people within a group than when they are not (Bond *et al.*, 1985). Finally, when names of friends are required, respondents might mention people they do not consider friends, but who are perceived to belong to the group of friends whose members they were reporting (Bellotti, 2008). These patterns have the potential to overestimate the homophily[1]in the networks compared to network models based on the a disinterested observation of human interaction alone (Quintane & Kleinbaum, 2011).

One way to overcome these problems is to compare between network models derived from independent data sources; those from self-reported TND and those from DMTD. Finding that the two network models are similar, or at least that the differences between them are not systematic, would be coherent with epistemic realism, and arguably confirm the view that micro-macro link is definitional[2] between social transactions and ties. Unfortunately, we shall presently see that empirical studies fail to support this view.

[1]For abbreviations and nomenclature see the glossary on page xii.

[2]See section 2.2.

Substantial differences are found between network models that are based on observing social transactions, and those based on people's reports of their social ties. In a series of studies (Bernard *et al.*, 1980, 1981, 1984, 1990; Killworth & Bernard, 1980) conducted in the late 70's by Bernard, Killworth and Sailer (BKS), five different groups were studied, and a comparison was conducted between network survey data and a record of observed interaction. The objective was to discover to what extent people's reports cohere with the real behaviour as observed by the researchers. BKS conclude that 'People do not know, with any acceptable accuracy, with whom they communicate; in other words, recall of communication links in a network is not a proxy for communication behaviour' (Bernard *et al.*, 1981).

Though BKS found substantial differences between network models based on observed and reported data, it was Quintane & Kleinbaum (2011) who spelled out in what way the structural properties of the models differ, and what are the social mechanisms involved. They compared between survey data (based on free-recall) and emails exchanged among a group of 23 individuals in a medium-sized childcare agency operating in the greater New York area. Like in the BKS studies, they too found substantial differences between the network models. Specifically, clustering had an endogenous component in the email network model, whereas it disappears completely from the survey network model once homophily is controlled for. What this means in theoretical terms, whether one of the networks is closer to the 'real' network out there or whether individuals take part in multiple networks, is a question that remained unresolved, although the paper tends to adopt the latter interpretation.

Further substantive and theoretical research (Freeman *et al.*, 1987; Bazerman & Moore, 2008) suggests that people seem to recall social ties associated with long-term, stable and recurrent interaction patterns. Data about human behaviour seem to be more precise because each and every interaction is recorded in a disinterested fashion, especially when they are the product of digitally mediated transactions. However, DMTD has its own set of complications, broadly discussed below in terms of technical, relevance and interdependeny issues.

3.2.1 Technical Issues

The first and perhaps most obvious issue is the sheer size of DMTD. The Enron dataset discussed below (section 3.4.2) includes over 150,000 emails. The mobile phone call dataset studied by Onnela *et al.* (2007b) consisted of 4.6 million individuals connected by 7 million ties. Kleinbaum *et al.* (2008) use a sample of 30,328 employees, sending over a million emails. These fantastic numbers require advanced data-mining skills, programming and technical knowledge of a very different kind than those needed for traditional datasets. In addition, the types of statistical tools available for this kind of data require a substantial learning curve.

One way to overcome these problems is to aggregate the data, and prune it in various ways, attempting to reduce its volume while maintaining as much as possible its network level properties (Serrano *et al.*, 2009). Yet, this process is not straightforward. More often than not, any attempt to reduce the data leads to networks with very different properties (Butts, 2009; Grannis, 2010; De Choudhury *et al.*, 2010).

There are interesting theoretical consequences from the vast amount of data coupled with the fact that this data is not filtered by the actors' own judgment regarding the meaningfulness of their social ties. This is a systematic difference in the density between the two models, the model based on DMTD invariably more dense than the one based on TND. Density is usually taken to be a sign of social cohesion, where norms are well inculcated into the conciousness of its members, and members are well integrated, identifying with the group etc (Coleman, 1988; Friedkin, 2004). But in a DMTD based network, density may be driven by the type of task the group has to preform. It might be simply a result of people fulfilling their organizational roles, and not necessarily the product of internal cohesion.

Finally, both the size and the density of a network model are known to interact with many other types of network measures (Anderson *et al.*, 1999), so that the interpretation of a measure's value in a small and sparse network can be completely different from its interpretation in a large and dense one. A large differences in a measure can be the result of the type of data used to construct the network, not necessarily a sign of substantial differences between the populations (Quintane & Kleinbaum, 2011).

3.2.2 Relevance Issues

One of the advantages of TND is that respondents pre-process the data and filter out irrelevant connections. Emails do not carry any kind of indication to the level of their significance for those involved. We know that respondents tend to recall 'stable' relationships (Freeman *et al.*, 1987), but what is the correct way to operationalize this term in the context of transaction data? How do we extract the stability of a relationship out of a dataset of email messages, for example? Does stability mean regularity of exchange or frequency of transactions? Does it mean that users bother to reply to incoming emails or the amount of effort they put in writing them (e.g., the length of the message)?

Moreover, we may possibly observe a stable exchange of communication that is not perceived as socially significant. Some administration roles or help desks regularly communicate with employees throughout the organization. If we were to take communication frequency as a sign of tie-strength, we might impute relevance to ties that exist only as organizational scaffolding. In the context of a questionnaire it might not even occur to the respondent to mention these contacts. This is because respondents automatically judge and evaluate the social significance of their relations. If we explore the transactions without asking what meaning they have for the actors, we are left with little clue about how and what to remove from the dataset.

Add to this the very low signal to noise ratio that is a consequence of the strength of weak ties (Granovetter, 1973) hypothesis. Weak ties are crucial for the connectivity of the network, and for the distribution of valuable information. However, the transactions associated with weak ties are rare. Consequently, weak ties are both low in frequency and highly relevant. Filtering out low frequency ties can therefore do away with non-redundant, relevant ties. In the context of email communication networks for example, Onnela *et al.* (2007b) shows empirically that an increase in the threshold of the number of transactions necessary for an exchange to count as a tie, does away first and foremost with the weak ties that function as bridges, crucial for network connectivity.

Though I have pointed out the problem of relevance specific to DMTD, one should bear in mind that survey tools also have an analogous problem. Respondents presumably apply some criteria when they judge who to report as a contact; yet the researcher cannot always control these criteria, or even know what they are and hence, it is sometimes difficult to interpret the relevance of the reported ties (Bearman & Parigi, 2004; De Choudhury *et al.*, 2010).

3.2.3 Interdependency Issue

A network based on questionnaire data is comprised of the aggregation of personal networks, each of the respondents choosing their contacts. Consequently, respondents choose contacts without knowing whether those contacts choose them in return. In this sense the answers of the respondents are independent of one another. Moreover, a reciprocal nomination is a sign of a symmetric relationship, one that indicates trust, commitment and social capital (Scott, 1991).

In contrast, by the very definition of the term, transactions are interrelated events, of which both sender and receiver are aware (see figure 2.5). When an email is sent from A to B, B is not only aware of being 'chosen,' she might want to abide by an etiquette according to which recipients ought to reply to their emails. Consequently there is a dependency between an email sent in one direction and the emails sent back in reply. The notion of 'reciprocity' observed in TND has a completely different meaning than the notion of 'reciprocity' observed in DMTD.

The interdependence problem in transaction-data is even more complicated, considering that transactions can involve more than two people. For example, emails could be sent to more than one person, and as the next chapter shows, an email sent to multiple recipients has different consequences than multiple emails, each sent to a single recipient. In a multiple recipient email situation, each recipient is aware not only of being chosen, but also of others being chosen. The option of hitting the 'reply-all' button on the mail client software makes it possible for an email sent to two recipients to trigger a transaction between the two recipients. One way to deal with this issue is to set a maximum threshold of the number of recipients, filtering out emails with recipient number that is greater than this threshold (see in section 3.4.1.) However, this method of filtering might have unwanted consequences on the network's structure, as emails may contribute differentially to network parameters, depending on the number of recipients (See Chapter 4.)

3.3 Miscellaneous Strategies for the Analysis of DMTD

Up to this point, we only spoke of filtering and aggregation of transactions, the two most common ways to construct social network models out of DMTD. In terms of the Coleman diagram described in section 2.2, these basic methods conform to the definitional type of micro-macro link. However, the literature uses more sophisticated methods as well. This section presents some of those methods.

3.3.1 Strength of ties

The easiest way to incorporate more of the information into the network model is to ascribe each tie with a strength attribute, proportional to the frequency of interaction. This method was used in various forms and purposes (Barrat *et al.*, 2004; Newman, 2001b; Diesner *et al.*, 2005) oftentimes (Adamic & Adar, 2005; Eckmann *et al.*, 2004) dichotomizing the strength of the tie and using a threshold values, taking all ties that are below a certain value to be non-existent (this is equivalent to the filtering technique in section 3.4.1.) In one exceptionally interesting paper, this method was used to verify the strength of weak tie hypothesis (Onnela *et al.*, 2007b), as discussed in 2.3.1.

This strategy works best if transactions were uncorrelated and randomly distributed in time. The weights would then represent the probability for a transaction, the defining property of ties according to Max Weber (see section 2.3.3). An alternative is to think of the strength of the tie varying with time, depending on the density of transactions around any moment in time. This notion is captured mathematically by an innovative method used by Palla, Barabási & Vicsek (2007). The strength of the tie between two actors a, b was calculated as follows:

$$S_{a,b}(t) = \sum_i s_i \exp\left(-\lambda |t - t_i| / s_i\right)$$

Where the summation runs over all transactions involving a and b and s_i denotes the weight of event i occurring at time t_i. (The constant λ is a decay coefficient characterizing the particular social system.) Finally, ties are ignored if their strength falls beneath a certain threshold (see figure 3.2.) This method bridges between transactions and ties, taking the discrete character of the latter and transforming it into the continuous one of the former.

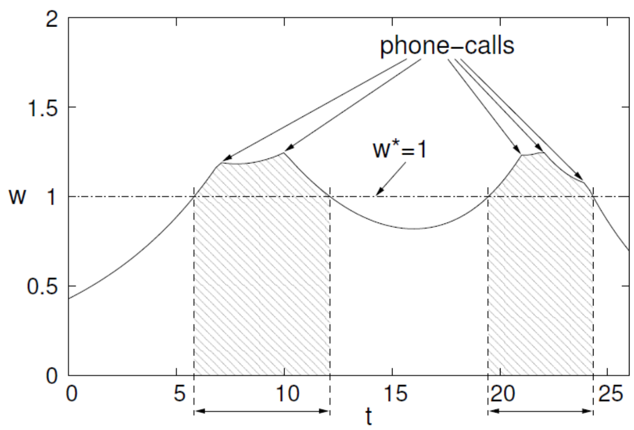


Figure 3.2: Tie weight depending on moment of transaction - for phone-call network (Palla *et al.*, 2007). A threshold of $w^* = 1$ was used, the tie considered present only when its associated strength is above threshold, i.e., within the shaded area.

3.3.2 Snapshot Networks

Another straightforward technique to overcome the tie-transaction gap is the use of snapshots (Moody *et al.*, 2005; Palla *et al.*, 2007; Kostakos, 2009; Miritello *et al.*, 2011). Here time-intervals are defined, and all transactions within an interval are aggregated to form a snapshot network. The result is much like panel waves known from traditional types of longitudinal network datasets (Snijders *et al.*, 2010). Specifically, transactions are grouped into clusters, each cluster associated with an interval. The clusters are exclusive, (such that no transaction is associated with more than one interval,) and exhaustive (such that no transaction exists that is not associated with one cluster.) Mathematically, the networks are represented as a set of graphs $\mathcal{G} = \langle G_0, \ldots, G_t \rangle$, where $G_t = \{V_t, E_t\}$ is the graph of time interval t, and V_t the set of active individuals in that interval, E_t is the ties between two individuals in V_t, such that $E_t \subseteq V_t \times V_t$.

Like with the strength of the ties approach, snapshots are ,most useful when transactions are distributed uniformly over time. But the bursty nature of human transactions (Barabási, 2005) and the interdependency among them makes it difficult to choose adequate time intervals, ensuring that chains of related interactions are all kept together within the same time interval. There are further problems if we take the transaction to be of a non-negligible duration (Pan & Saramäki, 2011).

To resolve this issue, some studies (Morris & Kretzschmar, 1995; Riolo *et al.*, 2001) make use of so-called transmission graphs, sometimes known as concurrency graphs. These depict all dyads in the left-most column, the row associated with each dyad depicts the moment in time (or the interval of time) in which the pair was active. This method is used particularly in epidemiological work, where links represent sexual partners and transactions representing encounters.

3.3.3 Multilevel approaches

From a theoretical standpoint, it makes perfect good sense to study macro-micro phenomenon using the statistical method of hierarchical or multilevel models. Both the theory and the method consider entities that are organized in a hierarchical form, micro-cases embedded in macro-entities: children's achievement in different schools, people's lifespans in different regions etc.

Multilevel methods were also used in the field of network analysis (Zijlstra *et al.*, 2006; Duijn *et al.*, 2004; Snijders & Kenny, 1999; de Nooy, 2011; Lazega *et al.*, 2008), but it has defintely not been mainstream tool: in the entire SAGE Handbook of Social Network Analysis (Scott & Carrington, 2011), the multilevel approach is mentioned less than a dozen times. This is partly because some of the multilevel models used in networks are particularly involved, especially when it comes to data structures characterized by crossed-hierarchies (Snijders & Kenny, 1999; de Nooy, 2011). Furthermore, multilevel models become very complex when accounting for structures larger than dyads. Finally, there are other methods, such as ERGM[1], that can accomplish much of what multilevel approaches

[1]For abbreviations and nomenclature see the glossary on page xii.

are supposed to achieve.

A recent paper (de Nooy, 2011) applied multilevel analysis to model the likelihood for relational events (a critic reviewing a book). But the use of multilevel analysis was not motivated by theoretic argument regarding the micro-macro. It was merely used to overcome a 'technical complication,' the complication being the dependencies between properties of individuals and the likelihood of the event. Thus, some critics are more likely than others to write reviews, and some authors are more likely than others to be reviewed. Thus the likelihood for a specific critic to write a review about a given author is an event, dependent in part on properties of the critic and those of the author. The micro-case is therefore the event of writing a review, and it is 'nested' in the group of all reviews written by a specific critic, and also in the group of all reviews written about a specific author. The critics and authors are individuals, within which multiple micro-cases are nested. Hence the hierarchic structure of the data and the complex patterns of interdependencies.

The second empirical chapter (Chapter 5,) uses this method in order to model the likelihood of receiving a reply to an email. Instead of seeing the interdependency as a technical complication that has to be controlled for, the chapter takes the macro-micro link between the social tie and the social transaction (in this case: a reply to a given email) as the theoretical motivation that justifies the use of multilevel analysis.[1]

3.3.4 Event Networks

This method conforms with the strategy described in section 2.3.3, doing away completely with social ties, and replacing them with networks of of transaction-patterns. The literature here uses the term 'event' to refer to transactions that involve exactly two individuals.[2] Assuming it is of negligible duration, an event from actor i to j at time t_1 would be expressed thus: $e_1 = (i, j, t_1)$. From here, there are several ways to proceed. One way is to develop (Butts, 2008; Brandes

[1]Interestingly, Abell (2003b) also uses multilevel statistical methods to operationalize the Coleman diagram in a study of what he terms 'Narrative Action Theory.'

[2]This means that further elaboration of this method is needed to capture transactions that involve more than two individuals, such as multi-recipient networks.

et al., 2009) models predicting what events are likely to unfold, and what are the parameters that affect this likelihood. The studies investigate questions such as this: are actors more likely to cooperate with those who cooperated with them in the past? Are they more likely to be hostile towards those who were hostile to them in the past? Are they more cooperative towards the friends of their friends.

A second, perhaps more brazen approach is to redefine network concepts (path, centrality, connectivity, density etc.) in terms of these transactions. Take the notion of path between two nodes, i and k for example. A possible path would consist of two events, $e_1 = (i, j, t_1)$ and $e_2 = (j, k, t_2)$ provided that $t_1 < t_2$. It becomes clear very quickly that this adds interesting conditions on the definition of transitivity, and basically every other network concept one could think of.

The event networks have very different properties than the static network in which all events are aggregated to form a (definitional) tie. Two individuals that are connected in the static network might not be connected at all in the event network.[1] Moreover, nodes may be close to one another in the static network, but the events connecting them are so rare that in practice, the time to reach from one node to the other can be very long indeed. On the other hand, two nodes that are very far apart in the static network may be transversed swiftly, considering the rapid rate of the events connecting them. Because of this, diffusion processes may follow paths that are very unlike what one would expect by looking at the static network (see section 2.3.3), and nodes that seem insignificant in the static network may become central for diffusion in the event network.

3.3.5 Bipartite Networks

It is also possible to construct networks with a very high fidelity to the original dataset through the use of so called ***bipartite*** or ***two-mode*** networks. Bipartite networks do not suffer from the limitations of the typical social network in that the translation of the communication data into a bipartite network model is relatively straightforward. There' no need to filter or aggregate the data in order to create it, no need to make ad-hoc assumptions about it or contemplate what is the meaning of the ties or the differences between ties and transactions. In

[1] Although empirical data shows this is rarely the case (Pan & Saramäki, 2011).

this respect, bipartite models are closer to data-models than theoretical models, because they are laden with hardly any assumptions at all. Moreover, they can be expanded and generalized to attach more meaning, and to reflect associations existing not only between individuals and themselves, but also between the sequence of messages.

Bipartite networks involve two distinct types of entities, usually individuals and groups in which they are members. Technically, a bipartite network is a triple $\mathcal{G} = (\mathcal{N}_{\uparrow}, \mathcal{N}_{\downarrow}, \mathcal{E})$ where $\mathcal{N}_{\uparrow}$ and $\mathcal{N}_{\downarrow}$ are two exclusive sets of nodes, and $\mathcal{E}$ is the set of edges that connects between them $\mathcal{E} \subseteq \mathcal{N}_{\uparrow} \times \mathcal{N}_{\downarrow}$.

Bipartite graphs are useful when the association between individual entities is mediated through a second type of entity. Thus, for example, $\mathcal{N}_{\uparrow}$ could designate the set of films and $\mathcal{N}_{\downarrow}$ could designate the set of actors playing in those films. Another example is the network of co-authors, where the set of authors are related to one another through the papers they have co-authored. Likewise, in text analysis, co-occurrence can link sentences with the words they contain. There are other types of networks that are not naturally bipartite, but could be represented as such in order to highlight or visualize certain aspects of the network. Take for example networks of hyper-linked web-pages or protein interaction networks. These networks tend to group into tightly knit communities, or cliques. One could now identify the different communities and assign nodes to the corresponding communities of which they are part. Thus, instead of representing the way nodes relate to one another directly, one could use a bipartite network to represent the way nodes relate to groups. Technically in such a way, every unipartite network could be represented as a bipartite one (Guillaume & Latapy, 2004). The other direction is also possible - every bipartite network could be collapsed into a unipartite network, in which two entities of the same kind are related if they are both linked to the same mediating entity. However, whereas the translation from unipartite to bipartite networks generally adds information to the model, collapsing bipartite into a unipartite model invariably reduces information (Borgatti & Everett, 1997; Koskinen & Edling, 2012). In particular, bipartite networks could be informative about the strength of the tie between two individuals, by representing for example how many events did they both take part in. The strength of the tie between the two individuals could depend, in part, on the number of other taking part in those events. Thus, features of the relationship between two individuals may exist in the bipartite model, but are rendered invisible in the collapsed unipartite model.

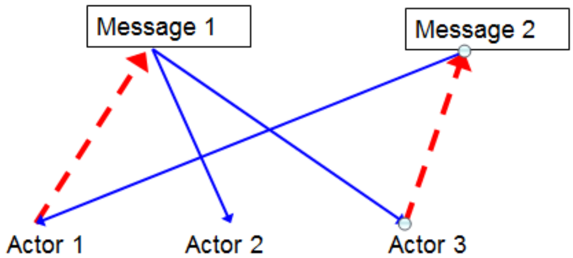


Figure 3.3: Email network model: a bipartite approach - Bipartite network models are better representations of email communication network than the typical unipartite network, but they have unusual properties and are difficult to analyse. In the figure, two messages were sent; message one was sent by actor one to actor two and three, whereas actor three replies by sending message two to actor one.

Bipartite graphs are used in the social network literature to represent both long term, structural affiliations and transient events (Kumar *et al.*, 2008), in general all types of $N : M$ relationships between two types of entities could be presented in bipartite graphs. Perhaps the most common type of study consists of directors on corporate boards, also known as interlocking directorate (Mizruchi, 1996), a bipartite configuration in which affiliation ties connect each board with its directors. Bipartite networks have also been used to represent more transient, ad-hoc links created in a temporal, event like setting. One of the first bipartite network investigated is known as the the Davis Southern Women dataset (Davis *et al.*, 1941), which recorded the participation of a group of a set of women in a set of social events. Another example can be seen in the literature on bibliographics, where bipartite graphs connect authors to papers (Small, 1973) or crime incidents to offenders (Frank & Carrington, 2007). Such micro-level events have a strong link to meso- and macro-level social properties such as the strength of a tie and the topology of the network at large, a notion that has been addressed in Scott Feld's seminal paper on the Focused Organization of Social Ties (1981). However, I am not aware of a paper that sought to bring both levels of aggregation into a single bipartite model, in the context of email messages.

So how would one seek to represent the email dataset in a bipartite network? One way way is presented in figure 3.3. In this model, users and messages constitute two disjointed sets of nodes. Users are not related to one another directly, rather, their link is mediated through the email message. We might be now tempted to start using the methodology developed for bipartite graphs (Opsahl, 2011; Koskinen & Edling, 2012; Wang *et al.*, 2012) to investigate the phenomena observed above: in- and out-degree distributions, reciprocity and transitivity, all within one model, without the need to desegregate the model into sub-models in order to control for the effects of the intermediate technological artefact. We might be even tempted to go wild and add a third kind of entity, the email thread which connects related emails to one another into chains. Research using tripartite networks is rare, but it does exist (Fararo & Doreian, 1984).

Using bipartite graphs to model the network of individuals and the emails they send is rather similar to the temporal network approach discussed above. The advantage of the bipartite graphs is that its design easily allows the representation of multiple-participant transactions such as mulitle-recipient emails, whereas the events defined thus $e_1 = (i, j, t_1)$ need to be further elaborated in order to capture such types of transactions. Unfortunately, the bipartite network suggested in figure 3.3 has unique features and the arsenal of method developed for bipartite networks are ill suited to deal with this model. Specifically, the issue is the distinction between email sender and email recipients, a distinction that does not exist in traditional bipartite graphs as they are used in the research of social networks. In common bipartite graphs, when members of one type of node are affiliated with a member of another type, all members of a particular type acquire the same 'role' vis-a-vis the node they are affiliated with. There is no conceptual distinction among the different nodes of the same type, and no equivalent to the notion of reciprocity or in- and out-degrees.

In the case of emails, each email is associated with exactly one 'sender' and at least one recipient. The story can become even more complicated if we would like to model the different roles recipients can occupy, distinguishing the recipients designated in the *to* field from those designating as *cc* or *bcc* fields. But as we found out above - and as we shall find out in the next chapter, reactions to emails are very sensitive to the number of recipients, for example, and to whether or not a recipient is designated in the *to* field or the *cc* field. These things matter, at the micro-level, but they make the bipartite model rather complex. The hope of finding some kind of progress pursuing this path diminishes the more one thinks of the complications involved.

To sum, there is much appeal in the idea of using bipartite graphs to model links between micro- and macro-level entities at the same time. In fact, one could think of the Coleman 'boat' diagram (Coleman, 1990) discussed in Chapter 2 as such a kind of bipartite graph, connecting micro entities with macro-entities. However, upon closer inspection and quite a few attempts to tackle the problem head on, it turns out to be a very difficult problem, indeed one that deserves a whole dissertation in its own right.

3.4 Email mediated transaction datasets

The previous sections presented in, a rather general manner, the methodological challenges that are involved in the analysis of digitally mediated transaction datasets. The intention was to show the lively debate in the literature concerning the theoretical and methodological link between social ties and patterns of transactions, at the same time demonstrating the numerous methods of dealing with descrepancies between these two levels of analysis. This final section is designed to focus on the type of data that will be used in the following two empirical chapters, namely an email dataset. The section begins with a very brief overview of the research of emails in organizational settings, and then turns to review the way emails have been used to study social networks. There is no attempt here to make a comprehensive review of the social network literature that uses emails, but only to focus on how this literature deals with the distinction between social ties and patterns of transactions. Finally, the ENRON email corpus is introduced, the dataset resource that is subsequently used in the empirical chapters.

Email is probably the oldest and most widely used Internet application for communication and coordination, certainly within many organizations (Dabbish & Kraut, 2006). It's popularity is due partly to positive network effects and partly to some of its key advantages, an easy, free and fast method to communicate in distributed environments (Sproull & Kiesler, 1986). As is often the case with new technologies introduced into organizational settings, emails have triggered a debate between supporters and critics of this type of communication technology. Some see it as a way to increase productivity (Rice & Bair, 1984; Crawford, 1982) whereas others worry about volumes of email encroaching on over-worked employees, leading to 'information overload' (Schultz & Vandenbosch, 1998) and possible decline in productivity (Dabbish *et al.*, 2005; Dabbish & Kraut, 2006). A related worry is that emails are a poor replacement of direct interaction and 'presence availability' (Zwijze-Koning & De Jong, 2005), highlighting the importance of face-to-face interactions for the accomplishment of organizational tasks. A third type of concern is that communication via emails is prone to misinterpretations, increasing the risk of 'uncertainty and equivocality' (Daft & Lengel, 1986). In a particularly interesting study Byron (2008) finds that email increases the risks of communication misunderstanding since recipients tend to misinterpret work emails as emotionally more negatively charged than intended. This is a property specific to email mediated transactions (and perhaps other types of non-synchronous text based messaging systems,) with possible negative impact on identification of employees with their workplace, influencing loyalty, trust, social cohesion and the general reduction of the social capital in the organization.

Studying a form of communication that is deeply entrenched in organizational settings has clear advantages over the study of emerging communication technologies (such as twitter, online social network systems or even instant messaging applications) where experimentation is still rife and norms are still being formed. Furthermore, emails constitute a unique form of communication technology because each email circumscribes a defined group of recipients, not only creating explicit boundaries between those who are 'in the know' and those who are not, but also making recipients aware of these boundaries.

One of the critiques of the exclusive study of email mediated transactions is that limiting the study to this one medium, the study becomes myopic to certain regions of the social networks. This argument is supported by an interesting study about the interaction between the medium of communication and the emotional intensity people ascribe to a relationship (Licoppe & Smoreda, 2005).

However, empirical research suggests that at least in some organizations, there is a correlation between email interactions, face-to-face meetings and telephone calls (Kleinbaum *et al.*, 2008). Moreover, email network is known to be the backbone of task-related exchanges in organizations, and the study of emails should be therefore crucial to the understanding of task related communication (Quintane & Kleinbaum, 2011). Numerous social network studies have used emails as a primary source of data, among them widely cited empirical papers (for example a paper by Kossinets & Watts (2006) with 748 citations according to google scholar at the time of writing.) Since the focus of these papers is so heterogenous, in what follows I seek to review the different ways in which email based network studies construct their models.

3.4.1 Aggregation and Filtering

When examining the email social network literature, one is overwhelmed by the diverse ways in which authors seek to construct their network models. But after a careful analysis of the literature, two common methods permeate the process of model construction: aggregation and filtering, where some of the filtering is done at a pre-aggregation stage, and some at a post-aggregation stage.

Aggregation proceeds by designating email senders and recipients as nodes of the network. For every email, the node representing the sender is connected by directed ties to each of the nodes that represent the recipients. In other words, each email is represented as a personal-network in the form of a star, with the email's sender at its center, the sender connected to each of the recipients. Note that already in this stage, some of the original information in the dataset is lost, information that is relevant for the formation of the transaction network. To understand the nature of what is lost, consider the difference between sending an email to multiple recipients (***brodcast*** emails) and sending multiple private emails, each to one recipient (***private*** emails.) As demonstrated empirically in the next chapter, there are strong reasons to suspect that recipients react differently to broadcast and private emails. Unfortunately, these two cases become all but indistinguishable in the network model, thanks to the process of aggregation.

The second component in the process is filtering. The motivation here is that some of the data is redundant, or simply noise, and the task is to differentiate between the signal and the noise. Pre-aggregation filtering consists in discarding emails upfront because they are deemed irrelevant. Emails may be judged to be irrelevant if they were sent by mistake, if they are impersonal or if they are sent in bulk. In these cases, they are not seen to represent 'real interpersonal' exchanges (Kossinets & Watts, 2006; Tyler, Wilkinson & Huberman, 2005). Post-aggregation filtering consists in discarding ties, often when they are not symmetric, or if their throughput falls below a chosen threshold. The filtering stage raises further concerns of lost information when taking into account that an email's recipient list is not an arbitrary collection of individuals (Zhou *et al.*, 2005), and that even bulk emails may delineate meaningful organizational units. Table 3.1 illustrates the diversity of methods used to filter email data, the different justifications used by the authors, and the often ad-hoc values of thresholds used in these studies.

Table 3.1: Email Mediated Transaction Data: Strategies of Filtering

Source	**Filtering method and justification**
Eveland & Bikson (1986)	No detail of filtering of emails or links .
Ebel *et al.* (2002)	No filtering of emails or links.
Guimerá *et al.* (2003)	'Bulk e-mails provide little or no information about how individuals or teams collaborate' and hence they were discarded. Bulk emails were defined as those sent to more than 50 recipients.
Gloor *et al.* (2003)	No detail of filtering of emails or links.
Shetty & Adibi (2004)	Non-reciprocated ties are discarded, as well as ties which exchange less than a threshold of 5 emails over the entire period (4 years).
Eckmann *et al.* (2004)	'Mass mailings' are discarded, mass mailings defined as mails with more than 18 recipients. Non-reciprocated ties also discarded.

Continued on next page...

Table 3.1 – *Continued from previous page*

Source	**Filtering method and justification**
Diesner *et al.* (2005)	The data was transformed into a weighted, directed network. Some messages were deleted in response to requests from affected employees. Only a sample of the users was chosen, for which personal details were available.
Adamic & Adar (2005)	An undirected network was constructed based on links between two individuals who have exchanged at least 6 emails in both ways over the period (3 months). Emails with more than 10 recipients were removed completely (these emails are regarded by the authors to be 'mass emails'). The authors justify these thresholds by saying that they 'sought to minimize the likelihood of including one sided communication' or brief email exchanges where individuals 'do not get to know one another.'
Tyler *et al.* (2005)	Messages excluded if sent to more than 10 recipients because these 'were often lab-wide announcements', rather than 'personal communication.' Ties were excluded the number of emails exchanged falls below 30, or if each node sent less than 5 e-mails to the other. The aim was 'to reduce the number of one way relationships.'

Continued on next page…

Table 3.1 – *Continued from previous page*

Source	**Filtering method and justification**
Chapanond *et al.* (2005)	The paper employs two 'noise filtering' techniques and demonstrates that the analysis of the data is sensitive to the filtering technique. The noise filtering techniques used are based on: 1. Thresholds. This method discards links in which less than 30 emails have been exchanged or links in which less than 6 emails have been exchanged in each direction. This is similar to the method used by Tyler *et al.* (2005) using different threshold values. The following justification is given to this practice: 'by removing edges with small number of emails we enhance the real connection between people; the edges with small number of emails are considered as noise here. We are also interested in the interaction between people. The threshold we use to construct the undirected graph emphasizes an interaction by considering two-way communication.' 2. Eigenvalue decomposition. This method shows that the adjacency matrix has a low rank approximation. Explain more: what is an eigenvalue decomposition, what does it mean that there is a low rank approximation etc
Kossinets & Watts (2006)	Emails with more than 4 recipients are discarded "to ensure that our data do indeed reflect interpersonal communication as opposed to ad hoc mailing lists and other mass mailings"
Braha & Bar-Yam (2006)	"To consider only e-mails that reflect the flow of valuable information, spam and bulk mailings were excluded using a prefilter . . . We report results obtained by treating the communications as an undirected network, where e-mail addresses are regarded as nodes and two nodes are linked if there is an e-mail communication between them."
Onnela *et al.* (2007a)	". . . the mobile phone data is skewed towards trusted interactions, i.e., people tend to share their mobile numbers only with individuals they trust. Therefore, the [Mobile Call Graph] can be used as a proxy for the underlying social network."

Continued on next page. . .

Table 3.1 – *Continued from previous page*

Source	Filtering method and justification
Kleinbaum *et al.* (2008)	"We focus our analyses on e-mails that are sent to four or fewer recipients. In the core models, we exclude sender-to-BCC pairs ... Imposing these screens shrinks the data set by almost an order of magnitude to 13 million e-mails."

The table demonstrates a lively discussion and a diverse set of considerations regarding the standards required for network construction from email datasets. The decisions made by data modelers are important, because, as De Choudhury *et al.* (2010) demonstrate, choosing different strategies yield substantial differences between the resulting network models. It is therefore not surprising that these authors are alarmed that the question is rarely raised, regarding the different options modelers have when they construct their network models. They experiment by varying the thresholds on the minimal rate of transactions a dyad should exchange in order for it to be defined as a tie. Then they search for the threshold that maximizes homophily in the network, finding that the optimal range for the threshold is the same across different email datasets.

3.4.2 ENRON email dataset

From a more general discussions on emails, we now turn to the Enron email dataset, on which the empirical chapters are based. The ENRON dataset is probably the largest email corpus that is publicly available, consisting of corporate mails collected by the Federal Energy Regulatory Commission (FERC) during the judicial proceedings against the ENRON corporation. The very unusual circumstances in which these emails were exchanged makes it an attractive resource for organization studies, yet from a social network perspective Diesner *et al.* (2005); Diesner & Carley (2005) did not find any evidence for idiosynchratic features that distinguish this corpus from other email communication networks (and hence the subtitle of their paper: "It's Always About the People. Enron is no Different.") And though there is no clear justification for generalizing findings from this dataset to other organizational settings, I could not find any existing studies of the dataset that suggest that, from a social network perspective, there is anything highly unusual or unique about the ENRON dataset.

In the year 2002, as ENRON was fighting its last legal battles, the FERC decided to make a number of the ENRON related emails exchanged in the period between 1998 and 2002 available to the public. The original version of the dataset consisted of 619,449 emails found in different folders (among them the inbox and outbox) of the mailboxes of a group of 158 Enron employees. Some of these emails were exchanged between the members of this group, but most of the emails were exchanged between group members and other ENRON employees or individuals outside of ENRON. Consequently, the number of individuals in the dataset exceeds 158, reaching dozens of thousands of email-users, some of which are ENRON employees and some are not. At first, the data was made available in an mbox style format, with each message in its own text file (Rowe *et al.*, 2007), the data exhibiting a number of integrity problems and data corruption issues.

Consequently, a number of research groups worked to correct the integrity issues, making multiple different versions of the dataset available. Like Diesner & Carley (2005); Rowe *et al.* (2007), this thesis uses the version made available by Shetty & Adibi (2004). This group deleted corrupt data and fixed some of the integrity issues having to do with empty or illegal email names, as well as empty, blank or bounced messages. Duplicates were also removed. Invalid email addresses were converted to the form user@enron.com whenever possible (i.e., recipient is specified in some parse-able format like "Mary K. Smith") and to no_address@enron.com was assigned when no recipient was specified. Several researchers (Carenini *et al.*, 2005) have indicated that a number of emails were lost, either in the process of collecting the dataset or while preparing it for the public.

Numerous studies have been published using this dataset. Diesner *et al.* (2005) found that during the crisis, the personal network of the employees increased, and became more varied with respect to the formal roles of contacts. People who were previously disconnected began to engage in intense communication, transcending organizational barriers that were in place before the crisis. Several other studied changes in structure or activity during the period of the crisis (Collingsworth & Menezes, 2009; Tang *et al.*, 2010; Uddin *et al.*, 2011; Strite, 2013). And though some changes have been identified, it is hard to say that we know how to identify a crisis by looking at the dynamics of a communication network alone.

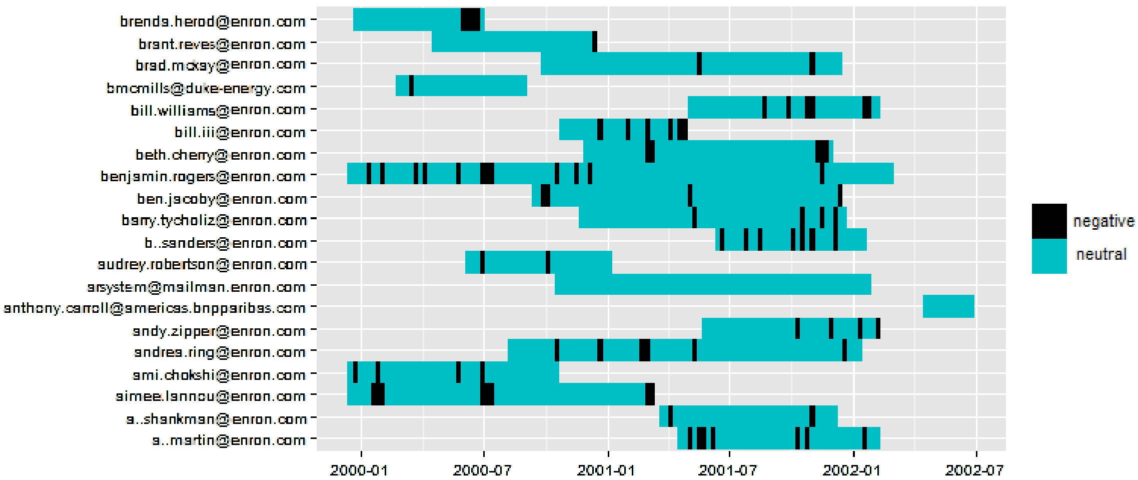


Figure 3.4: Sentiment analysis in Enron Emails - each horizontal bar represents an average sentiment associated with an email sender. The black marks emails where the sender's sentiment are negative (Strite, 2013)

It is probably safe to claim that most of the research done on the Enron dataset had very little to do with the unique historic context of the organization. Some researchers used the dataset to develop new algorithms (Rowe *et al.*, 2007) or software (Frantz & Carley, 2008). Others studied it from a Natural Language Processing perspective (Diesner *et al.*, 2005; Klimt & Yang, 2004). The empirical chapters in this thesis continue in this tradition of using the dataset to explore substantive and theoretical issues that are not related specifically to the ENRON affair. Note that in order to answer the research question, the issue of representativeness of the dataset is not entirely relevant. The current study is not an enquiry of social-phenomena per-se, and so it is not quite important to assess to what extent the described mechanisms are universal. The objective is merely to describe mechanisms that operate between the micro, meso and macro levels. It is entirely probable that the unique context of the ENRON environment would have some effect on the kind of micro-level transactions of email exchange, creating transaction patterns that are unique in this context. But the problem of generalization is independent of the question of the micro-macro link, and since this work focuses on the latter, the elaboration of the former is outside of the scope of this work.

3.5 Summary and Reflections

In one of his most popular essays, Isaiah Berlin (2013) recalls the ancient Greek poet Archilochus: 'the fox knows many things, but the hedgehog knows one big thing.' Berlin uses this to distinguish between two types of intellectual endeavours. Hedgehog intellectuals specialize and zoom-in on one problem, following one big idea during their entire career (Berlin's examples include Plato, Dante, Pascal, Hegel, Dostoevsky, Nietzsche.) Fox-like thinkers draw on various experiences, mix and match, diversify and adopt without settling down on a particular, single idea (Berlin's examples include Aristotle, Shakespeare, Montaigne, Goethe, Pushkin.)

It seems that today, network modellers are more like foxes than hedgehogs. The wide spectrum of methods being employed to construct networks is mind boggling. It is very possible that the foxy nature of modellers is here to stay, since each context seems to require its own unique strategy. As the introduction to this chapter has shown, a strategy that works for transactions in a developed context does not necessarily work for a developing context.

That said, there seem to be common premises in much of the literature, such as the hidden assumption that the network is constructed not once but twice: first by the actors themselves, in organizations, communities and social movements, their members interacting, communicating, exchanging gifts for obligations, consolidating their own network and testing the reliability and strength of their ties. The network is constructed for the second time when it is studied, when its researchers re-assemble the evidence to produce the network model, distributing questionnaires, conducting interviews, organizing the data, collecting it from different sources, adjudicating between signal and noise and cross validating their assumptions to establish that the network ties they have come up are useful for research.

One concern raised throughout this chapter is the worry that in the haste to aggregate and filter, some of the traces in the data are wiped out, testimonies as to how the actors built their own networks. A related concern is that minute decisions on the part of the modeller yield very different types of network modells Grannis (2010); De Choudhury *et al.* (2010) Consequently, the next chapter embarks on the path of disaggregation, separating email transactions into different kinds and noting how different types of transactions lead to different types of network structure.

Like other models in science, network models function as mediators between data and theory. The literature on the philosophy of science has much to say about the role of models for the advancement of the scientific project (Morgan & Morrison, 1999). Models have several defining attributes. First, the process of their *construction* gives them a sense of 'autonomy.' It is tempting to think that models are nothing but the re-organization of data or a reformulation of the theory, but as this chapter has shown, many theoretical assumptions go into the construction of the network model, specifically in the process of aggregation and filtering described above. It is precisely this construction process that makes the model autonomous in the sense that it is more general and abstract than the data from which it came, yet not the full fledged theory in its own right. The model becomes a self-sufficient construct by virtue of including elements of both, and only by being separated from them, can it act as a mediator between the theory and 'the world.'

A second defining attribute of the model is its *function*. Like other tools, network models have a purpose. They serve as a testing ground for various theoretical propositions about a certain empirical reality. This purpose is achieved by way of *representation* of some aspect in the world or some dimension of a general theory. And finally, it is this type of representation that allows the model to function as the facilitator of *learning*.

4

From Micro to Macro: How emails contribute to network structure

> In theory, theories exist. In practice, they do not.
>
> Bruno Latour (1988, p 178)

> HAMLET (Drawing his sword:) 'How now! a rat? Dead, for a ducat, dead!' (Stabs through the arras.)
>
> William Shakespeare, Hamlet: Act 3, Scene 4

4.1 Introduction

This is the first of two empirical chapters, an exploratory, data driven chapter, whose objective is to test mechanisms that operate at the micro-level of transactions, with effects on the macro-level network structure. As described in the previous chapter (see section 3.3.2), the process of aggregating (email) transaction to construct network models, whitewashes some of the important attributes of the raw transactions. One way to make these attributes visible again is by disaggregating the dataset into separate groups, each group consisting of transactions of a certain type. After creating separate network models from each group of transactions, it is then possible to compare between the structural features of each of the network models. The rationale behind this method is that disaggregation reveals patterns that are concealed by the aggregate network.

The chapter is divided into three parts. The first part demonstrates that the number of recipients on an email is a feature of the transaction at the micro-level that has important consequences, both at the meso-level and the macro-level of transaction patterns. One hypothesis that explains this effect is that the number of recipients on an email is related to the strength of the (unobserved) tie connecting between sender and recipient (see section 6.3.2). To demonstrate how this hypothesis could inform the construction of better network models, the second section explores various distributions that are contingent on the number of email recipients, and the third section offers a novel method for constructing network models based on those distributions.

There is one feature of email transactions that is whitewashed in the process of aggregation and therefore often neglected from social network studies of emails (but cf. Zhou *et al.*, 2005; Liben-Nowell & Kleinberg, 2008), this feature being the number of recipients on a given email. Consequently, this chapter opens by disaggregated the email dataset according to the number of email recipients.[1] The motivation for this is twofold. First, as shown in section 3.4.1, one of the standard ways used by network modelers to filter out the data is to remove emails whose recipient number is above a certain threshold (broadcast emails.) The typical justification cited is that broadcast emails are likely to represent bulk or spam messages rather than meaningful inter-personal ties. Therefore, they are not relevant for a network model that should represent meaningful social ties. However, it is unclear from the literature what consequences this type of filtering may have for the network structure (Zhou *et al.*, 2005). Second, the intuition is that unlike private emails, broadcast emails are used for group discussions, coordination and collaboration. If this is so, we might expect to see different structures associated with private and broadcast emails.

Focusing on this particular issue makes the discussion less general and more sensitive to the distinctive properties of the email as a unique form of communication medium. I can think of no other communication medium that consists of a single author and multiple recipients, the recipients themselves organized into three different categories; those listed in the ***'to'*** category are the immediate addressees of the message. Those who need to be aware of the existence of the

[1]In what follows I use the term *broadcast emails* to denote emails with relatively numerous recipients and *private emails* to denote those with few recipients.

message, but need not necessarily act upon it, are typically listed in the ***'cc'*** category. Finally, the ***'bcc'*** category lists recipients who remain invisible to the rest of the actors in the scene, like Polonius eavesdropping behind the curtain to an unfolding drama. Thus, the email is designed to allow for the organization of actors into an impressive array of roles, the senders signaling their expectations for the kind of reactions they hope for from their recipients. Moreover, the email itself is embedded in the context of a sequence of other transactions known as a 'thread,' the context that gives the message its meaning, reminiscent of what Erwin Goffman called *the structure of a situation*: a set of constraints and affordances that guide the development of the sequence of communication transactions. The specific nature of the structure of a situation is a key factor that can determine the trajectory of social processes and their outcome on the macro-level[1].

A second reason to pay close attention to the technology is that the interpretation of the data hinges, to a large extent, on the way people use the technology. In a study of mobile communication networks, for example, Kovanen *et al.* (2010) had to split the users of mobile services in his dataset into two different groups, depending on the way they paid for the communication services: prepaid users who pay for usage before making calls and postpaid users who pay afterwards. Different ways of paying for mobile usage is correlated with substantial differences in network statistics between the two groups of users in terms of their degree distribution and reciprocity. To control for this effect, the groups were analysed separately.

Just as in the case of mobile communication data, the study of email data requires careful consideration of the way technology is used. Consider the role of hubs (nodes from which emails are sent to an unusually large number of recipients) and authorities (nodes that receive emails from an unusually large number of senders,) for example. Like typical social networks with directed ties (*arcs*), email communication networks have both hubs and authorities. But Eckmann *et al.* (2004) rightly point out that in contrast to the central roles of hubs and authorities in social networks, in the email graph one should handle them with

[1]Raymond Boudon (1986, p 32-35) poignantly demonstrates this in the context of diffusion of innovations, where the rate of diffusion is a macro-level feature of the system, its change over time hinging on what Boudon calls 'the structure of the situation.'

some suspicion. Authorities might be service desks whereas hubs could be machines, mass mailers or administrators that distribute organizational announcements, going out to many users. Thus, the importance of hubs and authorities hinges to some extent on whether the object of inquiry should include only thematic issues or administrative ones as well. That said, email hubs and authorities are, of course, important when studying the diffusion of viruses for example.

There is a third reason for focusing on the idiosyncrasies of a specific technological medium, and that is to respond to a pressing concern within the community of information systems and organization researchers, a concern articulated in a seminal paper by Orlikowski & Iacono (2001). In this chapter, the authors argue that the study of information systems has long been preoccupied with a research program that treats the technology as a black box; either by handling it in very general terms, or by accepting it as a given, deterministic agent of change, ignoring the different modes in which it can be used, regardless of any of its unintended consequences and the context in which it operates. This program, they lament, has contributed precious little to the advancement of a theoretical nature. To correct this sad state of affairs, the authors ask researchers to look into detailed features of technological artefacts and to be sensitive to unforeseen patterns of human-machine interaction. Focusing on the actual practice of designing, developing, using and interacting with the technological artefact itself, they assert, could be used as a 'methodological device,' a point of departure through which researchers could gain greater theoretical insights into the operation of an organization as a whole.

The email communication dataset used in the current study consists of a snapshot taken from the famous Enron corpus (Shetty & Adibi, 2004). The chosen period spans the months of September to December 2001, as this was the most dramatic period for Enron as an organization and the most active in terms of the frequency of email exchange (Diesner *et al.*, 2005). Despite the concern that this period might represent an exceptional moment from a communication point of view, previous research gives us little reason to believe that this is likely to bias our results in any systematic manner. Moreover, in choosing this period the data could be validated against previous work that focused on the same period of this dataset (Davis *et al.*, 2007).

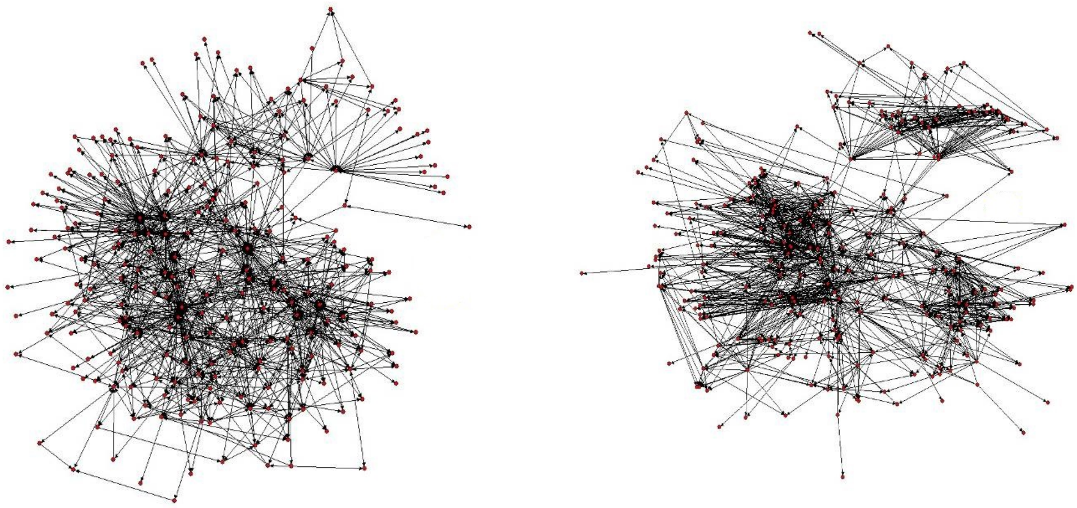
Standard practices of data cleansing were employed to clean the raw data. For example, when it was established that an individual was using separate email accounts, the accounts in question were united and represented as a single node in the network (as described for example in section 2.1 of Chapanond, Krishnamoorthy & Yener (2005)). All emails were removed, that involved email users (either senders or receivers) that were not ENRON employees. In order to simplify the analysis, no distinction was made in this study between different recipient fields (*to*, *cc*, or *bcc*). If any recipients were listed more than once in the list of recipients, their name was retained only once. If the sender of an email was also on the list of recipients, the user was removed from that list. Duplicates were identified and removed. The result of this arduous cleaning process was a dataset of $35,964$ emails. The constructed network consists of $9,818$ ENRON employees interconnected by $68,409$ directed arcs.

4.2 Disaggregating the email dataset

The distinction between anatomical and functional networks is adopted from the work on neural networks (Shalizi *et al.*, 2006), where subgroups of connected neurons synchronize their activity in a way that depends on the cognitive task at hand, every task excites a different sub-network of cells. Thus, the same anatomical network is associated with multiple functional ones. In a similar vein, studies of email communication networks (Eckmann *et al.*, 2004) demonstrate how regions of the network are activated in synchronized activity, revealing sub-structures that are not apparent from the network of aggregated emails. The same individuals might participate in different positional roles in the two sub-networks. For example, a central node in one sub-network could be marginal in another.

Various studies use this method of disaggregating the email dataset into separate groups, forming sub-networks from each group separately. Each study does this for different purposes. Kovanen *et al.* (2010) analyses postpaid and prepaid mobile calls separately, finding that each subnetwork exhibits different structural patterns. Eckmann *et al.* (2004) and Braha & Bar-Yam (2006) disaggregate the dataset into groups of emails sent within relatively short intervals from one another. Both papers report how the sub-structures and positions of nodes in the network vary between one sub-network and the other, the aggregate network deviating in its structure substantially from the structure of each sub-network.

Continuing in this tradition, the following empirical investigation disaggregates the email dataset into groups, each group consisting of emails with a certain range in terms of the number of recipients. If broadcast and private emails have different functions, the disaggregated network models represent different functional networks. Two functional networks are depicted in figure 4.1.



(a) Network constructed from single recipient emails. One recipients per email. Connecting the nodes are 1471 directed ties (*arcs*), reciprocity is 43.9%, transitivity is 16.9%. *Notice the open, fan like structures in the network graph.*

(b) Network constructed from multiple-recipient emails. $20 - -50$ recipients per email. Connecting the nodes are 1327 directed ties (*arcs*), reciprocity is 8.9%, transitivity is 31.3%. *Notice the closed triangle like structures in the network graph.*

Figure 4.1: Two networks constructed from two types of emails. All emails sent and received between members of a group of 254 users within the same period of three months, each network constructed from emails with a different range of number of recipients

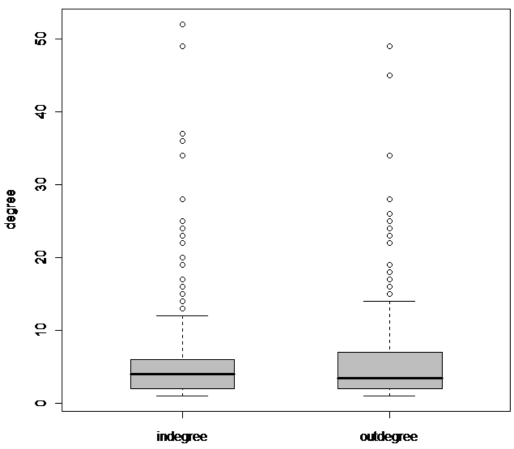
Both models in this figure are based on emails exchanged among 254 individuals sampled from the Enron dataset (more about the method of sampling the data below.) The network on the left is based on single recipient emails only. All 254 users have either received or sent single-recipient emails from others in the group. The network on the right is based on multiple recipient emails, the number of recipients ranging between 20 and 50, and again all users have either received or sent emails of this kind from others in the network. All the emails were sent within the same time frame, the months of September to December of 2001.

The two network models consist of the exact same individuals, sending and receiving emails within the same period of time. Both networks have a similar density, one with 1471 and the other 1327 ties, yet the differences between the two networks are striking; compared to the broadcast email network, the level of reciprocity is much greater on the private email network, the level of transitivity much lower.

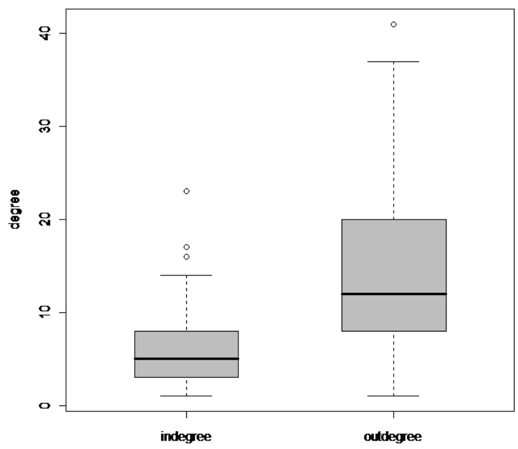
The incoming and outgoing degree distributions of the two networks are shown in figure 4.2. Again we see a marked difference between the two networks. All four degree distributions are positively skewed, which means that although most of the degrees are within a relatively narrow range, there are some individuals with an anomously large number of contacts, up to five times more than typical numbers within this range. But despite this similarity, it is easy to identify differences between the networks in terms of the degree distribution.[1] Two interesting features stand out when comparing the degree distributions. The first is the difference between the degree distributions of the private and broadcast emails, the former being much narrower than the latter. Also, note the difference between the indegree and the outdegree distributions of the broadcast emails, again the former being much narrower than the latter. What explains these differences?

Let's start with the difference in degree distribution between private and broadcast emails. This is especially curious taking into account that the dataset includes an order of magnitude more private messages than broadcast messages (see table 4.1). Granted, the number of recipients on each of the broadcast messages is an order of magnitude greater than the number of recipients on the private messages. Hence, these effects should cancel-out, more or less, since the number of sender-recipient pairs is of the same order of magnitude. The differences in network structures suggest that there are different social mechanisms at work for private

[1] Note that some individuals in the broadcast network on the right have low levels of outdegree, some less than the minimal number of recipients on emails. It may well be that some of the recipients are not in the network at all.



(a) Single recipient email network: Degree distribution



(b) Multiple recipient email network: Degree distribution

Figure 4.2: Degree distribution of the networks appearing in 4.1

emails and broadcast ones. The decision to send email messages is governed by a set of interests and norms: private messages are sent to a smaller subset of contacts than public messages, and there are many people with whom contact is materialized *only* in the context of public messages. Like parties in which one wants see and to be seen, broadcast emails seem to realize ties that are not materialized in private settings, whereas private messages realize relationships that are more intensive in terms of the frequency of transactions, designating perhaps stronger and more meaningful ties. This imputed link between the number of recipients on an email (at the micro-level) and the **strength** of a tie (at the meso-level) is an issue I shall return to below.

This suggests fewer senders of broadcast emails than receivers, so that only a subset of the group sends out the bulk of broadcast emails - in that case, that subset would have rather large out-degree, but the in-degree of everyone will be relatively limited to those senders. The different degree distribution could thus reflect two roles existing in the organization - a subset of those who tend to send out messages to numerous other.

A final test for the similarity between these two networks consists of a quadratic assignment procedure (QAP) (Krackardt, 1987) using the Ucinet6 software for windows (Borgatti *et al.*, 2002) with 2000 permutations yielding an estimated correlation of 0.23 $(p < 0.001)$, a magnitude that is rather modest, compared for with other QAP correlations found in the social network literature. Consider for example correlations between networks of self-reported relationships and networks reflecting observed interactions between people. When Quintane & Kleinbaum (2011) compare email communication networks to network data elicited through survey procedures, they find a QAP correlation of 0.35 $(p < 0.01)$. Other QAP correlations between observed interactions and self-reported survey data yield values in the range 0.29 and 0.46 (Quintane & Kleinbaum, 2011). In this context, the correlation we find is rather low, albeit significant. This is not surprising given the descriptive summary statistics presented above, but QAP is an statistical inference method, establishing that these correlations are unlikely to have been the consequences of stochastic effects in the data.

It is possible to continue the exploration and build Exponential Random Graph Models (*ERGM*[1]) as described in Quintane & Kleinbaum (2011), but an attempt to evaluate such a model using XPNet software (Wang *et al.*, 2006) has failed to reach convergence, perhaps because of the relatively large network. The advantage of using ERGM would have been to point out what type of local structures (such as triangles, symmetric configurations etc.) are unique to each of the two networks, thus confirming through inferential statistics in what ways do the observed networks deviate significantly from random networks. Moreover, ERGM could control for the differences in density when assessing the difference in reciprocity and transitivity. But it is very unlikely that the network density is responsible for the different structures observed above, and this for two reasons: first, compared to the difference in reciprocity and transitivity, the difference in density between the two networks is small (1471 vs. 1327 directed-ties yields about 10% difference in network density), and second, if density alone was responsible for the difference we would expect the more dense network to have both a higher levels of reciprocity and a higher level of transitivity, which is not the case.

[1]For abbreviations and nomenclature see the glossary on page xii.

4.2.1 A tale of seven networks

To further generalize the findings about a link between email recipient number (at the micro-level) and network topologies (at the meso- and macro-level,) more networks of the same group of users were constructed and compared. Before presenting the results, I'd like to describe in greater detail how the group of 254 members of the group were chosen. The aim was to find a group of individuals, all of which are connected to one another by emails with a large range in the number of recipients.

The users were selected following an unusual and ad-hoc process. The problem was how to find a group of connected individuals, who contacted each other using emails that vary widely in the number of their recipients. For example, an individual who sends only private emails is not interesting for the purpose of this study, because it does not allow the comparison between networks constructed from private and broadcast emails. I experimented with a wide range of thresholds in order to arrive at groups of emails that would be comparable to one another. For example, identifying employees sending or receiving emails with 30 recipients or more, brings down number of individuals in the dataset from the original $9,818$ down to 503. From this group, a subset of users can be chosen such that all group members participate in single-recipient email transactions. Thus, it is possible to arrive at a subgroup whose members: a. sent or received a message via a single-recipient email, and b. sent or received a message via an email with 30 recipients or more. From this subgroup, an even smaller subgroup can be chosen so that all its members sent or received emails of $2-3$ recipients, etc.[1]

This procedure was repeated for different ranges of recipients, resulting in the above mentioned group of 254 individuals. Aggregating all emails sent and received within the specified time period, an aggregate network of density 0.08 was formed, its level of reciprocity reaching 0.41 and a global clustering coefficient of 0.44.

At this stage, all emails sent and received within the group of 254 were grouped according to the number of their recipients. The first group included all the

[1]Comparing these numbers with the literature is very difficult, since the very few network studies that account for recipient lists (eg. Zhou *et al.*, 2005; Liben-Nowell & Kleinberg, 2008) do not give numerical details of the lists they analyzed.

single-recipient emails. This group had the largest number of emails. The second group included all emails with two or three recipients and so forth. A network was then constructed from each of the seven groups described in table 4.1.

# of Recipients	# of Emails	# of unique sender-recipient combinations
1	$6,140$	$1,471$
$2-3$	$2,017$	$1,292$
$4-7$	956	$1,325$
$7-13$	87	$1,266$
$13-21$	353	$1,215$
$20-50$	436	$1,327$
$40-285$	318	$1,278$

Table 4.1: Seven sub-networks

The range of the number of email recipients in each group was chosen in such a way so that the resulting networks would be comparable in terms of their density. For example, the first network is based on emails with single recipients sent between members of the group. $6,140$ such emails were exchanged, but only $1,471$ had a unique combination of sender and recipient. Thus a directed, non weighted network was constructed with exactly $1,471$ directed ties connecting all the 254 members of the group. The second network was based on emails with $2-3$ recipients only. $2,017$ such emails were exchanged constituting $1,292$ distinct ties connecting the same 254 users. Seven networks were thus constructed based on the same 254 nodes. Each network contains a comparable number of ties (around $1,200$ to $1,400$ ties.)

The first and sixth of these networks were compared in the first part of this section. Let us now compare between the first and the second network, expecting that these two will be much more similar. To make the comparison, another QAP was carried out between the network based on single recipient emails and the one based on 2 to 3 recipients. As before, the QAP procedure was conducted using the Ucinet6 for windows software (Borgatti *et al.*, 2002) with 2000 permutations, yielding an estimated correlation of 0.49 $(p < 0.001)$. This is a decent correlation, as far as QAP of social networks go, and we find that the similarity between the first and the second network is much greater than the similarity between the first and the fifth network, where the estimated correlation using QAP was only $0.23\ (p < 0.001)$.

Three measures were calculated for each of the seven networks; transitivity was measured via the global clustering coefficient, which is the proportion of closed *triplets* to the total number of triplets in the network (Opsahl & Panzarasa, 2009).

$$C = \frac{\text{number of closed triplets}}{\text{total number of triplets}} = \frac{3 \times \text{number of triangles}}{\text{total number of triplets}}$$

Where a triplet is a group of three connected nodes. For example, if A is connected to B and C, the three nodes are considered a triplet. A distinction is made between an open triplet (in which B and C are not connected.) and a closed one (B and C are connected.) For obvious reasons, every triangle is considered to be three closed triplets. Thus, star forms with a centre and n edges have $n\,(n-1)\,/2$ triplets, all of which are open, the clustering coefficient of the graph is hence zero. A complete graph with n nodes has $n\,(n-1)\,(n-2)\,/2$ triplets, all of which are closed and hence a clustering coefficient of one.

Reciprocity was measured in two ways: the first was a straightforward method, the proportion of symmetric ties to all ties. As an additional measure of reciprocity, each network was compared with its transposed matrix, and the Pearson correlation coefficient was calculated using QAP as described above. The results are presented in figure 4.3 and figure 4.4.

The first result is therefore a confirmation and expansion of the finding established in the first part of this section, namely that emails with an increasing number of recipients contribute a decreasing proportion of reciprocated ties of the total network. Both measures of reciprocity point to the same trend, where the proportion of reciprocated ties decreases with the increase in the emails' recipient number.[1] There are several possible explanations for this trend, as will

[1] The disaggregation could potentially lead to an underestimation of reciprocity. For example, consider a situation where two users exchange emails, those going in one direction all broadcast those in the other all private. In such a situation, the tie between both users would be reciprocal

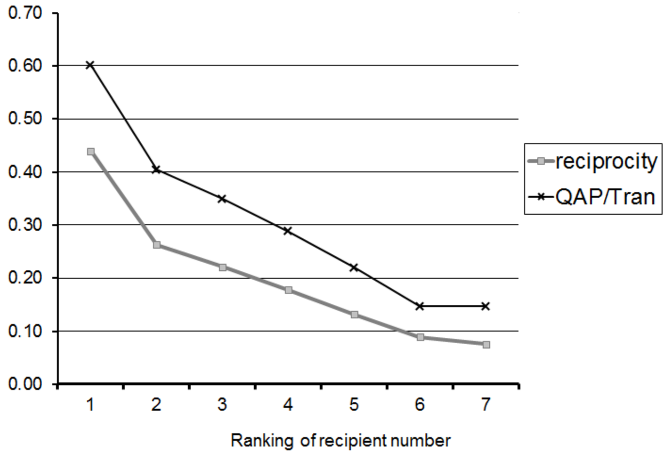


Figure 4.3: Comparing between two reciprocity measures - Reciprocity and a QAP procedure of the network matrix with its transposed as a function of the number of recipients in the emails connecting 254 individuals

be discussed in greater detail in the section 6.3.2. But for now, one might want to consider an explanation that is grounded in email etiquette - it might be considered more acceptable to ignore broadcast messages and not to reply to them, whereas it might be less acceptable to do so with a private email. In fact, not replying to private emails could be regarded as rude, whereas broadcast emails may be seen as a nuisance to be ignored, which is why they are sometime referred to as spam. Thus, norms that guide behavior at the micro-level could have implications at the aggregate level connecting single transactions to their collective outcome at the aggregate level of the communication network.

The second result is that emails with an increasing number of recipients contribute a proportion of closed triplets which increases at first and subsequently decreases. This could be explained if each email delineates a bounded group of recipients, and if relationships among recipients of the same email are more likely to occur than relationships between recipients of different emails. If this is the case, larger recipient lists have the potential to create greater proportions of

in the aggregate network but would not be reciprocated in the each the disaggregated networks. However, it is unlikely that this effect is substantial as the level of reciprocation of the single-recipient email network is very close to the reciprocation of the aggregate network. It follows that the number of ties whose reciprocation has been severed through disaggregation remains relatively low.

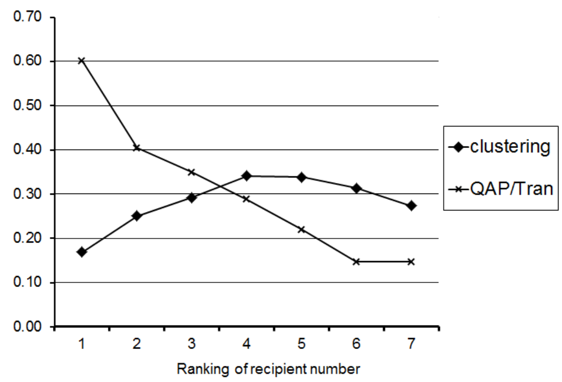


Figure 4.4: Reciprocity and clustering measures - Reciprocity and the clustering coefficient as a function of the number of recipients in the emails connecting 254 individuals

closed triplets. At first, this potential is realized, explaining why bigger groups contribute increasing levels of triplet closure. Above a certain threshold, emails sent to a great number of people make them less likely to know each other, explaining the waning proportion of triplet closure.

In what follows, I'd like to confirm the hypothesis that if a focal node sends an email to two recipients, they are more likely to connect to each other directly, compared to having received separate emails. In other words, being co-recipients is a stronger indication of tie formation two recipients of distinct emails from the same node. To test this proposition, all ties were divided into three types of what may be called *co-citation* categories. Given a combination of two nodes A and B, this *dyad* is said to be co-cited if both nodes have had email contact to a third node. Consequently, there are three types of co-citation: *strong co-citation* consists of dyads, whose nodes are co-cited in a single email. This means that there exists at least one email sent from some third node C, in which both A and B are co-recipients. In contrast, *weak co-citation* consists of dyads with co-cited nodes, but not even one email could be found in which A and B are both recipients. In other words, despite existing the existence of at least one node C sending emails to or receiving then from A and B, no email was sent from C that addresses both A and B concurrently. The third group consists of dyads which were not co-cited at all. This means that there exists no node C sending emails to or receiving them from

A and *B*. All connected dyads in the seven networks were classified according to these three categories, the results of which are represented in figure 4.5.

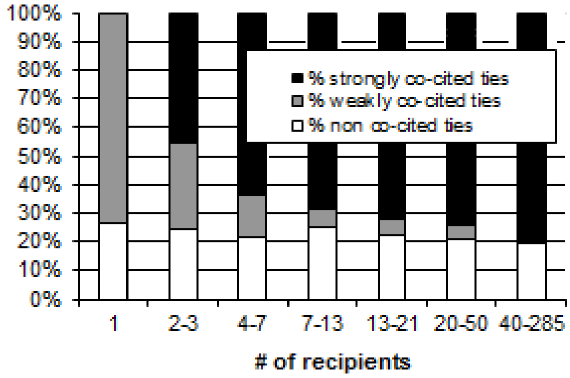


Figure 4.5: Classification of ties into categories of co-citations - All ties of the seven networks of table 4.1 were classified into three types of co-citation

In all seven networks, about 20% of the connected dyads are not co-cited at all. The nodes associated with these ties share no common acquaintances. They could be bridges connecting disparate groups, or else one or both of the nodes may have no other connections. Moreover, by the very definition of co-citation, networks based on single-recipient emails can have dyads co-cited in a weak form only. This is simply because having co-citation of the stronger form would require there to be emails with more than one recipient.

The interesting feature of figure 4.5 is that the more recipients in the emails, the greater the proportion of strong co-citations relative to weak ones. This suggests that as we move from more private to more broadcast emails, ties are more likely to be strongly co-cited. Now, this could of course represent the actual distribution of co-citations of the dyads themselves. Maybe we see more strong co-citations because a greater percentage of the dyads are strongly co-cited. To investigate this point a bit further, all co-cited dyads in all seven networks were classified into two forms of co-citations - strong and weak co-citations, sown in figure 4.6

This figure shows clearly that despite increase in the number of email recipients, there remains a substantial proportion of weakly co-cited dyads, at least 20% but in all but one network 40% or more. Thus, it seems unlikely that the connected and co-cited dyads are merely a random sample of all co-cited dyads,

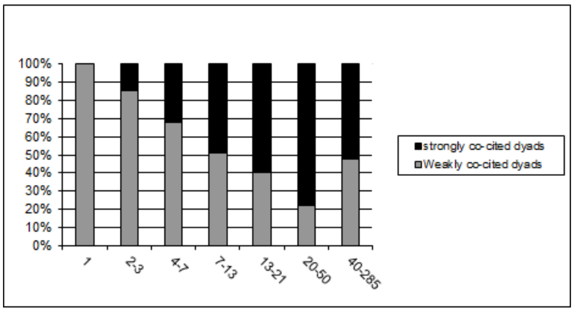


Figure 4.6: Classification of all co-cited dyads into two categories of co-citations - All co-cited dyads in the seven networks of table 4.1 were classified into strong and weak co-cited dyads

since, if they were, we would expect to see many more weakly co-cited ties. The conclusion of this finding is that an email sent to two or more people is a strong indication that these two people themselves are connected. The causal link here is of course unknown - it is possible that a mail was sent to these people because the sender knows that they are connected, and it is possible of course that the mail itself is the medium by that prompted the connection between the two.

Be it as it may, the finding can be seen as an expansion on the principle of transitivity known from social networks; yes, the friends of my friends are likely to be my friends. But this likelihood increases greatly, if my friends interact with their friend *and* with me at the same time, via an email sent concurrently to both of us. Here the key is to highlight the importance of the structure of interactions, over and above the existence of networks of human relationships (Feld, 1981). The tendency towards transitivity at the level network topology is important of course, but the theoretical principle of micro-foundations encourages us to seek the explanations for this tendency at the order of interactions.

The third result presented in figure 4.7 has to do with the distribution of in- and out-degrees of the seven networks. Though the distributions of in-degrees is relatively consistent across the private-broadcast spectrum of networks, the disparity between the in-degree and out-degree widens the more we move from private email networks to broadcast ones: whereas in private emails both in- and out-degree distributions maintain a relatively narrow (though positively skewed) distribution, in broadcast emails we observe a group of people who send out

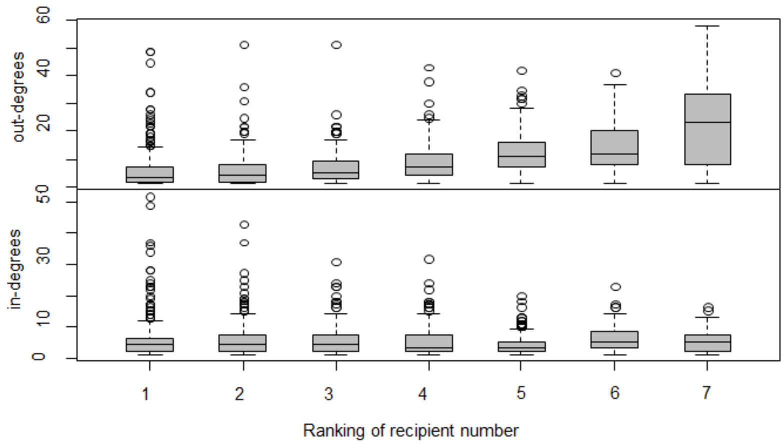


Figure 4.7: Degree distribution of seven disaggregated networks - Comparing the in- and out-degree distributions, notice how the distribution becomes wider moving from private to broadcast emails. Also note that the differences are more pronounced in the out-degree distribution, indicating, perhaps, a separation of roles within the organization

emails to a large number of recipients, though many receive emails from a relatively small number of others, many of whom do not bother to reply in the form of broadcast emails. They receive the emails but do not participate in a discussion. The third finding chimes with the asymmetric nature of broadcast emails on the one hand, on the other making a case for a partition of roles, a distinction between those who participate in the dissemination of multi-recipient emails and those who do not.

4.3 Fat tail distributions in email communication networks

One of the possible explanations for the mechanisms described above has to do with the strength of the (unobserved) social tie. Recall from section 2.3.1 that the strength of the tie is defined as a combination of micro-level properties of transactions (such as the amount of time individuals dedicate to their transactions and the level of reciprocity) as well as meso-level properties (such as the level of intimacy individuals associate with the tie as a whole). Since the dataset only contains email transactions, we do not have access to independent properties of the tie as a whole. However, the mechanisms above suggest that the strength of the (unobserved) tie is also connected to the number of recipients with which an email transaction is associated. For example, if the only type of communication between two individuals is on the basis of broadcast emails (and never private emails), we might infer that the associated tie connecting them is relatively weak. This hypothesis opens the way to the development of methods to construct network models, these methods take into account the number of email recipients (see section 4.4). But in order to develop these methods, this section explores some of the distributions associated with email recipients.

When Newman (2001a) studied the production of scientific articles he reproduced and developed a pattern first published by Lotka (1926), known today as Lotka's law. The law states that the number of publications per author is distributed as a power law, as expressed in equation 4.1. This distribution determines the probability that a given author would publish any number of times. An important characteristic of the distribution is that it is a highly skewed one, especially when α is small, a 'fat tail' distribution in which 'typical' values cover a very large range indeed, usually several orders of magnitude. This characteristic is very unusual for many distributions found in nature and in the social sciences, those known as 'normal' distributions, most of their values centring around a relatively limited range of 'typical' numbers. Variables distributed normally include human weight, for example, the rate of suicide in a community, disease etc. Human height, for example, typically ranges between 1.5 meters and 2 meters. A height of dozens of meters is simply unthinkable. In contrast, the number of publications penned by a single author can range from one to a many of dozens. Power distributed variables do not exhibit a typical number or range, and their variance often very large and their mean of little use or relevance.

$$p(x) \propto x^{-(\alpha+1)} \tag{4.1}$$

This is of course an approximation, since x represents a random variable, representing the number of papers published by an author, the number of recipients in an email, the number of co-authors of a publication etc. All these variables are both discrete and bounded. Treating a discrete variable as if it were continuous is a common practice and, in this case, does not present a mathematical difficulty. However, the boundedness of the variable is of consequence to the normalizing constant. Thus, without loss of generality, the normalizing constant C in the following equation can be calculated.

$$p(x) = Cx^{-(\alpha+1)} \quad \text{for } x \in (a,b) \quad \alpha > 0$$

Where the normalizing factor C can be calculated by integrating the distribution over the range $x \in (a,b)$ and equating the result to one. The normalized distribution function is now presented in 4.2.

$$p(x) = \frac{\alpha}{a^{-\alpha} - b^{-\alpha}} x^{-(\alpha+1)} \tag{4.2}$$

One way to explore the power distribution is to simulate it using the inverse of a cumulative distribution function (CDF). This is perhaps the easiest methods for sampling from a given distribution, a method described for example by Jackman (2009, p 153). Consider a sample needed for a general distribution $p(x)$, where $x \in (a,b)$. The CDF is defined such that $CDF(u) = \Pr(X \leq u) = \int_a^u p(x)\,dx$, CDF being a function that maps from (a,b) unto the unit probability interval $CDF : (a,b) \rightarrow (0,1)$. Using the following algorithm, it is relatively straightforward to sample from $p(x)$ provided that the inverse CDF exists such that $CDF^{-1} : (0,1) \rightarrow (a,b)$, and provided that it is a computable function.

Inverse CDF sampling algorithm:
begin for $t := 1$ to T do
    sample $u^{(t)} \sim UNIFORM(0,1)$
    $x^{(t)} \leftarrow CDF^{-1}(u^{(t)})$ od
endfor

We can now develop the inverse cumulative function CDF^{-1} associated with 4.2. First, the CDF:

$$CDF(x) = \frac{\alpha}{a^{-\alpha} - b^{-\alpha}} \int_a^x \hat{x}^{-(\alpha+1)}\,d\hat{x} = \frac{a^{-\alpha} - x^{-\alpha}}{a^{-\alpha} - b^{-\alpha}}$$

Reorganizing and replacing $CDF(x)$ by u, a uniformly distributed random variable ranging between 0 and 1, yields:

$$x = \left[a^{-\alpha} - u(a^{-\alpha} - b^{-\alpha})\right]^{-1/\alpha}$$

Now all we need to do is to sample from the uniform distribution, yielding a power distribution on x. Notice that x is a monotonous increasing function of u, approaching a (b) as u approaches 0 (1). Plotting the power distribution $x^{-1.5}$ defined on the range $(1, 100)$, figure 4.8 shows the exact distribution function (DF) as in equation 4.1 and a histogram of a simulated sample of this distribution using the inverse CDF method described above. One problem which remains is how to fit of this distribution given the data alone. To address this problem I shall now introduce the reverse CDF function.

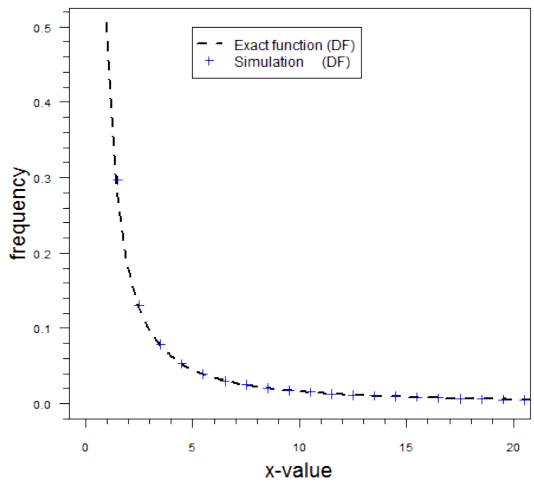


Figure 4.8: Sampling from a power distribution - A power distribution with exponent 1.5, ($\alpha = .5$) and valid values of x ranging between .61 and 100

The problem of course is that we have a distribution that seems to follow a power law, and the objective is to fit and estimate α. Several methods of fitting a power law are described in Newman (2005), methods based on taking the log of both sides of equation 4.1, making it equivalent to $\log p(x) \propto -(\alpha+1)\log x$, a linear relationship with slope $-(\alpha+1)$ on the log-log scale. Consequently, fitting a power law to a distribution amounts to regressing the logarithm of the observations against the logarithm of frequencies, the estimated slope of the curve being $\alpha + 1$. A diagram of the power distribution with exponent 1.5, ($\alpha = .5$) approaches a straight line in 4.9. However, as the figure shows, there is a difficulty with this approach, namely that for large values of x, the probability for an observations is low. This results in a noisy curve on the the tail of the distribution, a region where the power-law function dwindles and the number of observations is small.

One method of dealing with the large fluctuations at the tail is to bin the observations, either by having all bins equal in size, or by increasing the size of the bins where observations become scarce. Increasing the range of the bins decreases their 'resolution' but allows each bin to capture observations with greater probability. An alternative method to binning is to use the reverse cumulative distribution function ($CDF_{reverse}$), where the `y-axis` denotes the probability of observing a random variable equal to or greater than the the observation in the `x-axis`. The reverse CDF would then be expected to be.

$$\begin{aligned} CDF_{reverse}(x) &= \int_x^b p(\hat{x})\, d\hat{x} \\ &= \frac{a^\alpha}{b^\alpha - a^\alpha}\left[(b/x)^\alpha - 1\right] \end{aligned}$$

When x is at its minimum permitted value, $CDF_{reverse}(x = a) = 1$, since all observations are above that minimum value. By the same token $CDF_{reverse}(x = b) = 0$, since none of the observations are greater than the upper boundary of the distribution. We saw above that the log of the DF was perfectly linear. We shall now see, that $CDF_{reverse}$ is approximately linear.

$$\begin{aligned}\log(CDF_{reverse}(x)) &= \log \frac{a^{\alpha}}{b^{\alpha}-a^{\alpha}} + \log\left[(b/x)^{\alpha}-1\right] \\ &\approx \log \frac{a^{\alpha}}{b^{\alpha}-a^{\alpha}} + \log(b/x)^{\alpha} \qquad \text{for } (b/x)^{\alpha} \gg 1 \\ &= \log \frac{a^{\alpha}b^{\alpha}}{b^{\alpha}-a^{\alpha}} - \alpha \log x\end{aligned}$$

Note that the approximation works best when $b \gg x$ and when α is large compared to zero. When these conditions hold, the log-log graph of $CDF_{reverse}$ is approximately a straight line, just like the log-log graph of the DF. But unlike the DF, each point on the graph represents not only one, but several observations, minimizing the noise on the tail of the distribution.

Both the distribution function (DF) and the reverse CDF are plotted in 4.9. The figure clearly demonstrates that the sample of the distribution function is much noisier than the cumulated distribution. Moreover, as long as $x, a \ll b$, the $CDF_{reverse}(x)$ follows a linear function. If instead of the density distribution, we were to plot the cumulative density distribution $P(x)$, we would obtain a more stable log-log plot with a slope of $-(\alpha - 1)$, as can be seen from the following derivation:

In other words, just like the original density distribution, the cumulative distribution function $P(x)$ follows a power law, but with a different exponent which is 1 less than the original exponent. Thus, if we plot $P(x)$ on logarithmic scales we should again get a straight line, but with a shallower slope. However, thanks to the cumulative nature of the data, we would not expect the graph to bounce to zero every time no count has been observed.

Additionally, we can now study not only the distribution of email production, we can also study the distribution of senders and recipients. Thus, let us define email production as the total distribution of the number of emails sent by any single user. Email consumption can now be defined as the distribution of the number of emails received by any single user. Email dissemination is now the number of recipients per email.

This may seem surprising at first, but figure 4.11 makes it clear that the number of messages sent is much smaller than the number of mails received. This is

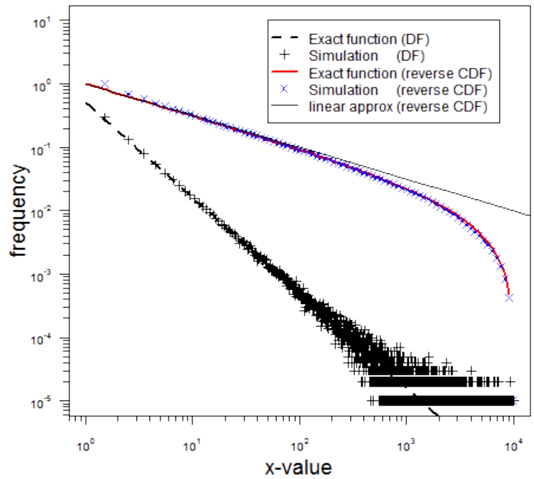


Figure 4.9: Sampling from a power distribution - A power distribution with exponent 1.5, ($\alpha = .5$) and valid values of x ranging between .61 and 100

because all emails are sent from a single account, but some can land in more than one inbox, specifically when sent to more than one recipient. This becomes perhaps intuitive when one notices that people's in-boxes tend to be contain more mails than people's out-boxes.

Both email production and consumption are found to be highly right-skewed distributions spanning four orders of magnitude. The finite time window makes it impossible for these distributions to perfectly fit a power law. However, it is possible to fit the data with a power law with an exponential cut-off as described by Newman (2001a). At least the first two orders of magnitude of the distributions nicely fit power laws with exponents -2.01 for email production and -1.66 for email consumption. This result may be seen as a generalization of Lotka's law for email use. Moreover, dissemination was measured as the distribution of the number of recipients per email (see figure 4.10). Fitting a power law to the first two orders of magnitude of this distribution yields an estimate of the exponent to be 1.86.

This result is interesting for several reasons. In terms of human production of intellectual or symbolic resources, Lotka's law has been tested numerous times in the past for the production of co-authored papers (Newman, 2001b) and the pro-

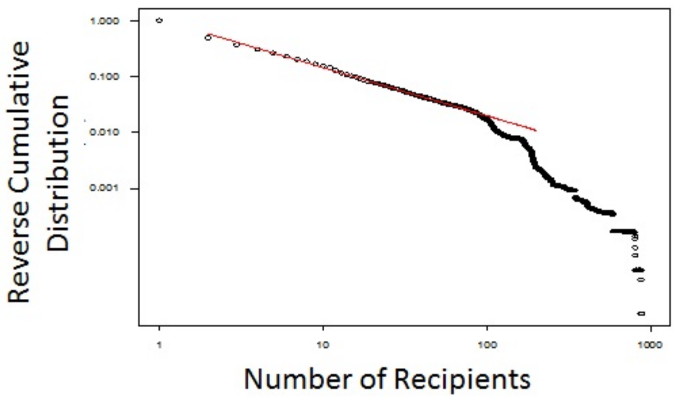


Figure 4.10: Dissemination: Distribution of email recipients - The cumulative distribution of email recipients per email message (presented in a log-log scale)

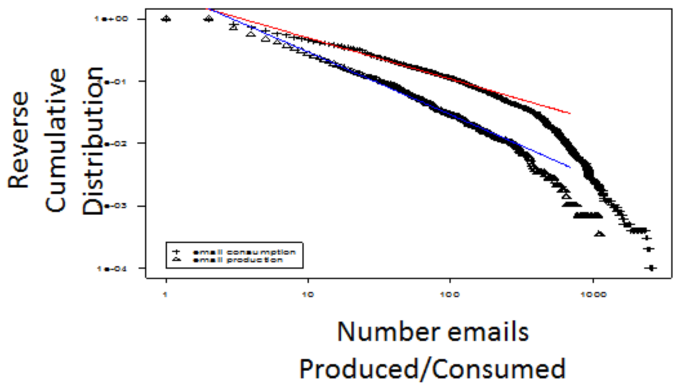


Figure 4.11: Production and Consumption: Distribution of emails produced and received - The cumulative distribution of emails produced per user and emails received per user (presented in a log-log scale)

duction of open source software (Newby, Greenberg & Jones, 2003). However, to the best of my knowledge, this is the first reported attempt to extend Lotka's law to patterns of production and consumption of email communication. Moreover, email production and consumption seem to be compatible with Lotka's law regarding the production of scientific journal articles. This means that many email users produce and consume a relatively modest number of emails, but also that there is a significant number of users who act as hubs of email production and consumption. The comparison between the production of emails and the production of scientific articles suggests that methods and theories developed for the analyses of networks of scientific collaboration (Newman, 2001b,a) could be tested on networks of email communication. Note however, that there is a fundamental difference between production (of emails or papers) and the consumption of emails. In the first case the choice of authoring a paper is done by the author; email consumption is a choice of a different subject than the one observed, namely the senders of the emails. This may have theoretical implications when generalizing Lotka's law to email consumption. Finally, the dissemination distribution shown in figure 1 suggests that there exists a preferential attachment of recipients to emails. Groups of email recipients seem to confirm power law distributions found in the sizes of cities, organizations and other social groups or entities (Adamic *et al.*, 2000; Adamic & Huberman, 2002; Newman, 2005). The extent to which recipient lists also delineate meaningful organizational or functional units is further explored in the next section.

4.4 Calculating the Strength of the Tie using the number of recipients

In the previous section the email dataset was described along dimensions relating to email usage: production, consumption and dissemination. This analysis, it was argued, could motivate further exploration of the relations between email messages and the groups they circumscribe. Two fundamental measures are being used: reciprocity and clustering. Reciprocity is measured as the proportion of the number of symmetric ties to the total number of ties. For the tie to be symmetric, at least two emails must be sent between the nodes, one email in each direction.

The second measure is the global clustering coefficient (or the level of transitivity, see Wasserman & Faust (1994, p. 243)). It gives an indication to what extent the network exhibits aggregations of highly dense clusters. It is defined as the ratio of the number of closed triplets to the number of all existing triplets. A triplet is any group of three nodes in which at least two of the three node-pairs are connected. It is considered closed if it is fully connected, otherwise it is open.

Both reciprocity and clustering are measures which attain higher values in social networks than what would be expected by randomly distributing ties in a network (Holland & Leinhardt, 1970). These two measures are also related to each other through the strength of weak ties hypothesis (Granovetter, 1973) . According to this hypothesis strong social ties are ties in which actors invest considerable amount of resources. Furthermore, they tend to be reciprocated and to be embedded in cohesive structures. This section explores how these measures also depend on the conditions of interaction, specifically on the number of recipients in the emails.

Does the number of recipients in an email give an indication to the kind of relationship existing between the sender and each recipient? One conjecture could be that two users are more strongly tied if they exchange 'private' emails, in contrast to more weakly tied nodes which exchange only multi-recipient messages. A possible explanation could be that sending an email to fewer recipients creates a stronger obligation on each of the recipients to reply. Thus networks formed from private emails may have a higher level of reciprocity, perhaps consisting of stronger ties. Emails with fewer recipients may lead to higher levels of reciprocity, but do they also signify stronger ties? A claim along these lines has been made in the context of networks of scientific collaboration (Newman, 2001a; Börner, Dall'Asta, Ke & Vespignani, 2005). For example, Newman (2001a) claims that 'it is probably the case [. . .] that two scientists whose names appear on a paper together with many other coauthors know one another less well on average than two who were the sole authors of a paper'. To account for this effect, Newman let each coauthored paper contribute a certain weight to the valued tie connecting each author to each of the other coauthors. This weight is inversely proportional to the number of those coauthors, so that if a scientist collaborates with $n-1$ other coauthors, on average, that scientist is acquainted with each of them $\frac{1}{n-1}$ times as much as if he or she were collaborating with just one coauthor. This idea could be easily adopted to email communication datasets. Since all the ties extracted from email k relate the sender i with recipients $j = 1, 2 \ldots n_k$, we would expect the weight of the directed tie between sender i and receiver j to be given by equation 4.3.

$$w_{ij} = \sum_k \frac{\delta_{ij}^k}{n_k} \tag{4.3}$$

Where δ_{ij}^k adds a contribution to the sum only if user i sent email k to recipient j. Consequently, it is defined as follows,

$$\delta_{ij}^k = \begin{cases} 1 & \text{if user } i \text{ sent email } k \text{ to recipient } j \\ 0 & \text{otherwise} \end{cases}$$

One important distinction between email networks and networks of scientific collaborations is that the former are directed whereas the latter are not. This makes the study of email communication useful because directionality of email ties makes it possible to test reciprocity, and since 'reciprocal services' are related to the 'strength of ties' (Friedkin, 1980; Granovetter, 1973), it is possible to put equation 4.3 to the test by comparing groups of ties with similar tie-weights, and examining whether an increasing average of tie-weights is correlated with an increasing proportion of reciprocity. To achieve this, the original $68,409$ directed arcs were ordered in increasing weight and divided into 50 bins, each bin consisting of nearly the same number of arcs (about 1368 arcs in each bin) of equally ranked strength. The proportion of reciprocated ties was calculated within each group and was used as a response variable in a simple logistic regression model, where the explanatory variable was the weight ranking of the ties. The fitted model was significant at the 0.001 level, with an increase of one rank in tie strength explaining an increase of 13% in the log odds for tie reciprocation (see figure 4.12). Similar relationships were found when emails were limited only to those with a number of recipients below 20 or even 15 recipients per email.

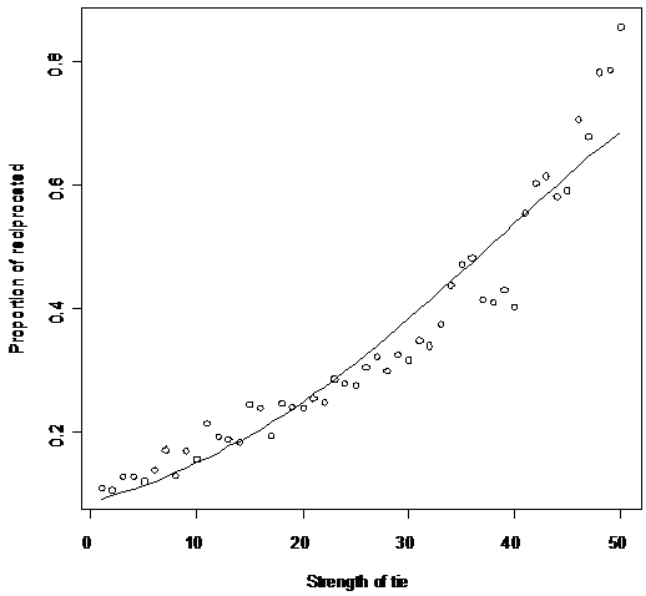


Figure 4.12: Reciprocity by strength - Reciprocity is explained by the ranking of tie strength in comparable sized sub-networks

Note that if an arc is reciprocated with another arc of very different strength, they would fall into different bins and would both count as asymmetric ties. Thus we are not testing reciprocity per-se but mutuality (Monge & Contractor, 2003, p 40), i.e., the extent to which a directed tie is reciprocated by a tie with the same rank of strength. Mutuality is of course a stronger version of reciprocity. Note also that the weights calculated in equation 4.3 increase with an increasing number of emails sent between two actors and decreases with an increasing average number of recipients per email. To rule out the possibility that reciprocity was explained mainly by the relative frequency of emails sent between two actors, it was important to test the contribution of recipient number to the effect on reciprocity. To this end, the email communication dataset was reshuffled and then compared to its original version. The reshuffling proceeded in the following manner: all outgoing emails sent by each user were identified and grouped together. For each group, recipients of distinct emails were swapped at random. As a result, users who received mostly private messages in the original network could now be found in emails with sizable recipient lists (and vice versa). However, the reshuffling process did not change most of the dataset's global properties: the number of ties, the number of users and emails all stayed the same, as well as the distribution of production, consumption and dissemination of emails. From the point of view of traditional approach to network construction from email communication data, nothing has changed in the network model. Nevertheless, the reshuffling process had a substantial effect on the weights of ties defined in equation 4.3. When comparing the correlation between tie weight and reciprocity, both networks exhibit significant correlations, because the relative frequency of emails sent between two nodes is significantly correlated with reciprocity. However, the variation of network reciprocity explained by tie-weights decreases substantially in the reshuffled network, indicating that the weights calculated from the original network better explain patterns of reciprocity.

To show the utility of the equation 4.3, a test was carried out to confirm the strength of weak ties hypothesis (Granovetter, 1973). According to this hypothesis, strong ties are embedded in a tightly knit environment. If this mechanism is at work in the dataset, we would expect that two individuals connected by stronger ties would tend to have more mutual contacts as compared to indivduals connected by weaker ties. The ratio of mutual contacts M_{ij} in the neighborhood of two connected individuals i and j can be quantified according to an equation suggested by Onnela *et al.* (2007b), an equation used in a similar context, namelly the confirmation of the strength of weak ties hypothesis in a study of mobile phone transactions:

$$M_{ij} = \frac{n_{ij}}{(k_i - 1) + \left(k_j - 1\right) - n_{ij}} \tag{4.4}$$

Where n_{ij} is the number of common neighbours of i and j, and $k_i \left(k_j\right)$ denote the degree of nodes $i\,(j)$. If the two nodes share no common neighbors, then $n_{ij} = 0$ and therefore $M_{ij} = 0$. Otherwise, since their degrees are k_i and k_j respectively and since they are connected to each other, they have $k_i - 1$ and $k_j - 1$ spare degrees for other contacts. If all their contacts are mutual we have $n_{ij} = k_i - 1 = k_j - 1$ and $M_{ij} = 1$. An attempt to verify the strength of weak tie hypothesis by correlating the strength of the tie[1] against the overlap has failed to show significance. This means that the hypothesis could not be rejected, that tie strength and high levels of mutual contacts are correlated. The strength of weak ties hypothesis in this particular dataset could not be confirmed.

4.5 Summary and Reflections

On one level, the findings in presented in this chapter are self-evident for anyone who uses emails within organizational or communal settings. You feel more compelled to reply to a private, personal email than to an email sent to many recipients, simply because a personal email is often an invitation to reply. It is also a feature of email client software that enable a *reply all* action, by which a recipient of a broadcast email can simply send a message back to the sender and to all other recipients of the original message. It is enough that a few of the recipients of a broadcast email will hit *reply all*, that the transitivity of the group will increase. The effort required to contribute transitivity to the network of transaction-patterns is hence by far lower in the case of broadcast emails compared to the case of private emails.

However, attempting to present these obvious behaviour patterns in network form, we find results that are not completely intuitive, indeed results that challenge basic social network theoretical claims, such as the strength of weak ties. Here is are the summary of the results:

1. *Degree Distribution.* As is commonly the case in social networks, all degree distributions are positively skewed. However, the degree distributions of broadcast emails are much more skewed than that of private ones, and the out-degree distributions of broadcast emails are much more skewed than the in-degree distributions, an effect that increases as the number of recipients grow.

2. *Reciprocity.* Private emails contribute to reciprocity in the aggregate network, much more than broadcast emails.

[1] The strength of the tie was calculated as the average of the strength of both arcs $\frac{w_{ij}+w_{ji}}{2}$ where w_{ij} was calculated according to 4.3.

3. *Transitivity.* Broadcast emails contribute to transitivity in the aggregate network, much more than private ones.

Since the traditional methods of constructing social networks from email datasets would not reveal these phenomena, two new methods of constructing social networks were developed. The first, based on bipartite graphs seemed to be promising at first, since it incorporates virtually all the information of the original data. However, the complexity of the graph and its unique features do not make it amenable to be used with traditional network methods. The second method was based on incorporating some of the information from the original dataset, through the usage of the notion of the strength of ties. This method was verified against two theoretical propositions: the strength of the tie is found to be strongly correlated with the likelyhood of it being reciprocated. The strength of weak ties hypothesis was shown not to hold for this dataset, but this was expected given the prior findings.

All these results would not be obtained if the exchange of emails was carried out at a random fashion, and for every deviation from randomness should be accounted for by some form of social mechanism that governs email transactions. The next Chapter will attempt to apply a theoretical analysis introduced in Chapter 2 and Chapter 3, on the results presented in this chapter.

5

From Macro to Micro: Four factors influencing email reciprocity

Rigid, the skeleton of habit alone upholds the human frame
Virginia Woolf (2012)

[Structures are] 'machines for the the suppression of time.'
My paraphrase on Claude Lévi-Strauss (Giddens, 1979)

The rumours then spread, like the spillway plunging down the street to the jetty, fanned out, here and there moving more swiftly and forming new branches, elsewhere coming to a standstill and drying up ... And so the rumours were transformed, further embroidered upon or attenuated, sometimes even refuted. Yet they persisted as the cocoon for a single statement, concealing its larva within, and no one knew what might yet come creeping out. The statement was: Naso [Ovid] is dead.
(Ransmayr, 1990)

5.1 Introduction

This second empirical chapter explores a simple question: what is the likelihood of receiving a reply to a given email? Investigating this question continues the line of empirical inquiry begun in the previous chapter, but with noticeable differences:

1. ***Explicit links between transactions.*** The theoretical discussion about social transactions (see section 2.2.1) highlighted one of their defining features, namely the notion that they do not merely reflect or reinforce links between

people, but that the transactions themselves are linked to one another. In an exchange of emails, for example, senders and recipients are associated with one another, but so are the messages, each message a possible response to a previous one and a potential stimulus to the next in a chain of related messages. This interdependency between separate social transactions is part of what makes them 'social.'[1] In the previous chapter these inter-transactional links are inferred implicitly, as a way to explain the empirical findings. To exaplain why private emails contribute more reciprocity and less transitivity than broadcast emails, an explanatory mechanism was suggested according to which stimulus-response links depend on whether the stimulus was a private or a broadcast message. But this was an inferred explanation, not directly observed in the dataset. In contrast, by asking what elicits a reply to an email, this chapter seeks direct observation and explicit measurement of stimulus-response links.

2. ***Emphasis on macro-to-micro mechanisms.*** The previous chapter seeks to explain structural topology at the level of the network by recourse to the way people responded to incoming emails. The emphasis there is on micro-to-macro type mechanisms. In contrast, by asking what contributes to the likelihood of eliciting a reply, the current chapter takes the structure of the network, the individuals, ties and even the stimuli transactions as exogenously given, and seeks to trace how a recipient makes the decision about the (best) course of action to take, whether to reply or not.

3. ***The nature of ties.*** This chapter continues the previous one on yet another level, holding that communication transactions observed in the empirical data are linked to social ties and their properties. For example, the previous chapter made an argument that private emails were associated with stronger ties, whereas broadcast emails with weaker ones. This (imputed!)

[1]Recall the argument from the theoretical chapter (Chapter 2) about the difference between a social transaction and what we might call a non-social 'behaviour.' The example given in this context was the difference between a 'wink' and a 'blink,' the former regarded as 'social' because it was done with 'reason and purpose,' an invitation for a subsequent and related social action. What makes an action 'social' we concluded, is its potential to be causally linked to subsequent and/or prior social actions.

association between tie strength and email type explains both higher levels of reciprocity in private emails and the narrower out-degree distribution of the networks from which they are constructed. This chapter examines the properties of ties that contribute to the rate of reply to a given email. In both of the empirical chapters, properties of the tie is a higher level of analysis serving to explain a phenomenon under observation.

The rest of the Chapter is organized in the following way. The next section describes how chains of stimulus-response type messages are operationalized in this email dataset. This is followed by a description of the multilevel model and the way the data was sampled, as well as the results. The chapter ends with a summary and some reflections.

5.2 Operationalizing stimulus-response chains

The point of departure is the notion that email communication are chains of interrelated social transactions. Just like conversational exchanges (Gibson, 2000, 2005) in which people take turns producing speech utterances in a sequence, each email serves 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 times involving more people in the process, at times leaving some actors out. What we have here is something akin to a process of diffusion, but slightly more general. Diffusion processes typically change the properties of individuals, for example when a disease of some sort spreads, people become carriers of a certain virus. But in this case we are talking more generally about a process that involves collaboration, discussion, advice or any type of social intercourse that transcends the level of a single message and extends over a certain length of time. While exchanging ideas, participants might even move from one topic of discussion to another. There is no requirement for anything tangible to actually spread in the network, only for a sequence of stimulus-response between transactions, like the process of falling dominos, ongoing, until it stops after a certain period of time (Barabási, 2005). This chapter probes the mechanisms that underlie this process and models the conditions that facilitate its unfolding.

Empirically, the basic building block for such a process - the 'micro-case' if you will, is the link between an email 'stimulus' and a subsequent email 'response.' Note here the difference between these two terms; *response* and *reply* Goffman (1976), the latter being a more general case of the former. Upon receiving an email stimulus, a recipients' response could be an email reply to the its sender. But it could be somethig completely different. In its most general form a response could be one of the following; (1) hitting the 'reply' control button dispatches a reply exclusively to the sender of the stimulus mail, (2) hitting the 'reply all' control button, dispatches a message to the sender of the original message and to all co-recipients, (3) hitting the 'forward' control button dispatches an email to any third party, and (4) some completely different kind of response that has nothing to do with emails, perhaps. Let us discuss the first three types of responses in greater detail.

1. The first type of response contributes to reciprocity in the network without changing the level of transitivity (when measuring transitivity, the direction is ignored in which the message is sent.)

2. A reply-all response contributes to the reciprocity in the network, too, but it also contributes to network transitivity. Consider a group of n members. One of the members sends an email to all the other $n-1$ members of the group, creating a star like network with zero transitivity and zero reciprocity. Now, one of the recipients of this first email hits 'reply-all,' effectively creating $n-1$ copies of her reply, one landing in the inbox of the sender of the original email, and one landing in each of the other $n-2$ members of the group. This transaction contributes a single reciprocated tie (the one connecting her with the sender of the first email) and an additional $n-2$ closed triangles to the network (all triangles sharing the same base, it being the tie connecting the sender of the first email with the sender of the second.) When a third group member hits 'reply-all,' she contributes two reciprocated ties (the ties connecting her to the first and to the second senders.) In addition, $2(n-2)$ new triangles appear. If all recipients

hit 'reply-all,' the graph becomes complete - each and every dyad has exchanged messages in both directions, both the transitivity and reciprocity of the networks have reached the maximum value of 1.

3. Besides the 'reply' and the 'reply-all' responses, the 'forward' response may or may not contribute to the network's reciprocity and to its transitivity, depending on other, possibly unrelated transactions. Moreover, there are of course hybrid transactions; one might hit 'reply-all' and then remove some of the recipients or even add some that were not in the original list.

Of these different responses to an email stimulus, this chapter focuses on the first two. A technical problem is how to establish that one email is a reply to the other. This problem is addressed by searching for common subject fields: if two messages share the same subject field, the first sent from A to B and the second sent at a later time back from B to A, these two emails are defined as related.[1] Thus, three conditions must be met (1) stimulus and response emails must have the same subject line (ignoring prefixes such as 'Re') (2) the identity of the sender and the recipients in the stimulus and response should appear in reverse, and (possibly) (3) stimulus and response emails should not be separated by a long time interval (Gibson, 2005).

Now, whether or not a 'stimulus' is effective in eliciting such a response depends on various factors. Four factors are considered; (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). Some of these factors are known from studies of the dynamics of speech exchanges and small group

[1] Two types of concerns might be raised here. A *false-negative* occurs when someone replies to an email, completely changing its subject title. This would most probably be considered a reply, but it would not be identified as one since the messages do not share a common subject field. Perhaps a more serious issue is the *false-positive,* when people exchange completely unrelated emails that share the same subject field; either because they initiated a new email by hitting the reply control button associated with an old, completely unrelated email, without bothering to change the subject field. Another possibility is that people use some generic subjects (such as 'stuff', or 'hi'.) To control for this issue, the subject lines were read and in case of doubt the removal of unrelated emails was carried out. However, we should consider the possibility that both concerns might have an effect on the results.

research (cf: Gibson, 2000, 2005; Goffman, 1976). In what follows, we focus on email communication and seek to isolate and compare the marginal effects of these different factors.

The sender and recipient effects refer to the specific properties of the senders and recipients of the original stimulus email. Certain individuals may have a high *sender effect* if most of their emails tend to be highly effective in eliciting replies. A high sender effect can simply be explained by the organizational position of a sender. For example, a sender high in the organizational hierarchy might be too important to ignore, so recipients tend to reply to their emails more frequently than to emails sent by others. By the same token, a high *recipient effect* is an attribute of recipients who are generally more responsive than other actors in the network. This could happen, for example, if an actor occupies a role that demands her to be highly responsive to incoming emails.

After controlling for the effects of the individuals, most interesting is the *dyadic effect*, allowing for variations between different dyads that cannot be attributed to single actors. Two actors might have average sender and recipient effects, but when they send each other emails, they may be much more likely to hit reciprocate than otherwise. The dyadic effect allows for certain pairs of actors to be more (or less) responsive to each other's emails relative to their mean responsiveness. This could be due to unique features of their relationship such as its history. It could also be due to certain issues that bind them to each other (or separate them from one another,) something special in the 'social tie' connecting the two individuals, that make them more likely to act in a certain way to one another, differently than what one might expect otherwise.

Finally, the *stimulus effect* allows for specific features of the message 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 each of the other recipients: it may reflect the email's unique content, its timing, or a specific signal indicating whether or not the sender is awaiting a reply. Whatever the reasons, the method proposed here allows to tease out and identify highly effective (or ineffective) actors, ties and emails in a systematic manner.

Whereas sender and recipient effects may depend, in part, on the absolute positions of actors in an organization, dyadic effects account for their relative positions. Finally, stimulus effects account for the idiosyncratic properties of specific transactions. The stimulus effect is an attempt to identify transactions with consequences that cannot be explained merely by the average behaviour of the actors involved or their relationships[1].

Estimating the variance that is associated with each of these four factors allows us to judge the relative salience of these factors. To this end, multilevel models are commonly used (see section 3.3.3): a statistical method that partitions variance when the data is structured hierarchically. Such a structure is characterizes the sender, who may send multiple emails, the recipient, who may receive multiple emails, the sender-recipient dyad, which may be a conduit for multiple emails, and the email, which may be associated with multiple replies from each of its recipients. If we find one factor to be more important than the others, that is - if that factor is associated with a larger variance, we may suggest that this factor is critical for the context of the unfolding of email exchange in this organization.

The following sections give a brief overview of the use of multilevel data models, with an emphasis on cross classification and multiple roles. A descripiton of the sampling of the dataset (section 3.4.2) follows together with the results of the fitted model.

5.3 Modeling replies using multilevel analysis

The method of multilevel analysis described in section 3.3.3 refers to a family of regression estimation methods that trace the different sources of variability in the data. It is commonly applied to cases (the micro-level) that are nested within one or more categories (the macro-level.) The objective is to separate the variability observed in the data to the variability explained by each type of category.

Simple uses of multilevel models consist of micro-level cases that are nested within macro-level classes. The classical example is that of children's achievements in school. The variability between children could be large when taken as a group. But separating them into sub-groups allows one to tease out what part of

[1]Allowing for a consideration of Goffman's famous dictum (Goffman, 1967): 'Not men and their moments, rather moments and their men'

the variability between students is due to variations between schools, and what part is due to differences between students within the schools.

Crossed classified models (Goldstein, 2011, Chapter 12) are used when micro-cases are embedded not only into one but into two types of categories. Consider for example students' achievements in their finals. Suppose that the final examinations are standardized country-wide, each year in all schools throughout the country, students present the exact same battery of exams. Now, as before, the variability could be separated according to schools, the variability between students' achievements within the school may be smaller than the total variability of students' achievements.

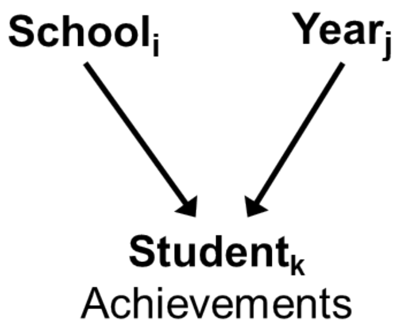


Figure 5.1: Crossed classified model: student's achievement nested in years and schools - A student's achievement in nation-wide, yearly standardized exams constitutes a micro-level case, each such case nested within a specific year in which the exam was taken, and within the school in which the student was studying.

In addition to the different properties of schools, the exams vary from year to year. Hence, variability in student performance could also be attributed to the years in which an exam was taken, the variability between students' achievements within each year smaller than the total variability. Each school contains a large group of micro-cases and each year contains a large group of micro-cases, but the years and the schools are not nested within one another, the relationship between years and schools is not like the relationship between students and schools, the latter being a one-to-many relationship and the former being many-to-many. Thus, crossed classified multilevel models comprise of cases that are nested within cross-classifications of two or more differing hierarchies. Figure 5.1 depicts the cross-level hierarchy.

Just like a student's achievements are partly a product of properties of the school in which she studied and the year in which she took it, multilevel analysis of networks (Snijders & Kenny, 1999; Snijders & Baerveldt, 2003; Lazega *et al.*, 2008; van Duijn *et al.*, 1999) treat tie level attributes as if they were partly the product of the two actors with which the tie is associated, as depicted in figure 5.2. The hypothesis here is that all ties incident to one actor have something in common, their properties influenced by the actor and the variability between them smaller than the overall variability of all the ties.

There are, however, two crucial differences between these two examples, students nested in years and schools, and ties nested within pairs of actors. One difference is that within each school × year combination there are multiple students (micro-cases,) whereas in networks every pair of connected actors define only one tie. That, however, does not present any technical difficulties. The more serious problem is that schools and years do not refer to the same entities, whereas a sender of a tie and its receiver could refers to the exact same entity. This requires the use of multiple roles models (Snijders & Bosker, 2011, p 161-162), to allow for a possible correlation between two attributes of the same actor, once acting in the role of a receiver and once acting in the role of a sender.

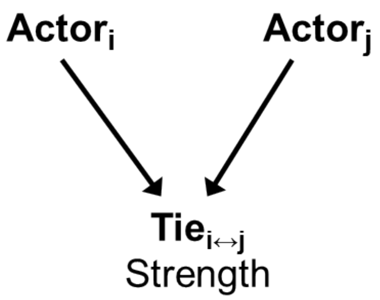


Figure 5.2: Crossed classified model: tie's nested within individuals as in (Snijders & Kenny, 1999) - The tie's strength constitutes a micro-level case, each case nested within the group of one actor's ties and the group of the second actor's ties.

The multilevel model is constructed such, that each actor can have multiple ties, the tie variable is treated as a micro-level case, nested within the set of ties belonging to each actor. Actors are therefore treated as the macro-level entities. We now want to adapt this model into our context, including not only ties between actors but also multiple transactions associated with every tie.

The outcome is a binary variable, namely whether or not a stimulus email has prompted a reply from each of its recipients. Since the time-window of observation is constrained, there is a danger that a stimulus sent close to the end of the time-window will elicit a reply that will not be observed. To fix this problem, two time-windows have been defined, both beginning at $00:00$ hours on September 1, 2001 but the first ending two weeks before the second, the second ending on $23:59$ hours on December 31st, 2001. Stimuli were limited to all emails sent within the first time-window, and replies were identified in the second time-window. Thus, even if a stimulus was sent at the very last moment of the first time-window, its reply could be identified, provided that it was sent within two weeks after the stimulus was sent.

In the first and simplest model, this outcome is modelled against the sender and recipient random effects. Thus for each actor an estimate is made of a 'global' sender effect and a 'global' recipient effect. Similar to the study described above (Snijders & Kenny, 1999), each transaction is associated with two macro-level groups: the group of email transactions sent by the sender and the group of email transactions addressed to the recipient. Each of these groups could have some common features, such that the variability within the group is smaller than the total variability in reply rate. In addition, this model estimates the correlation effect that is necessary for multiple role models.

The second model adds an estimation of a dyad effect to test how the decision to reply varies not only between actors, but also between an actor's different relationships. In this model we need to drop the correlation between the sender and recipient effects, otherwise the model is overspecified. The third and last model adds an estimation of the effect of each particular incoming email stimulus along with fixed indicators unique to that stimulus.

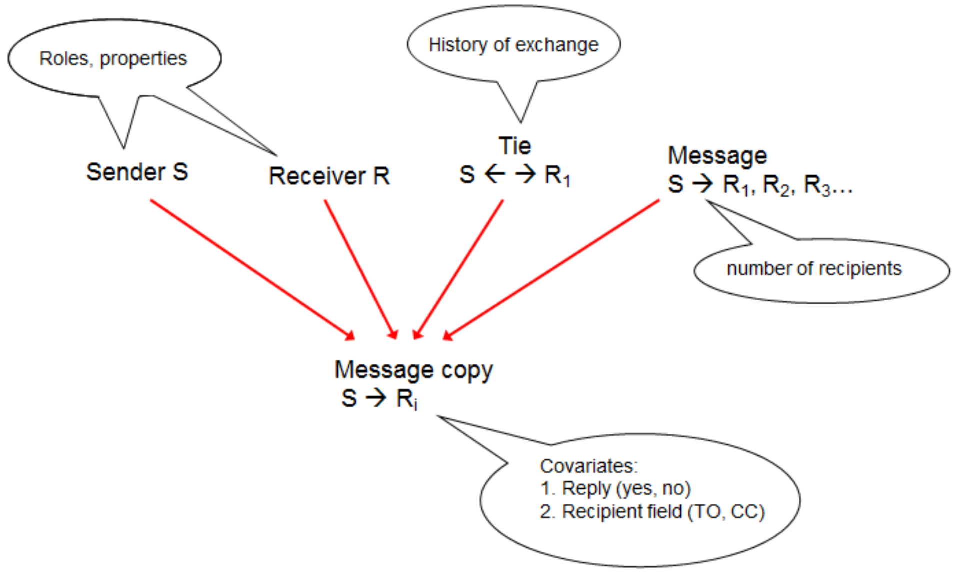


Figure 5.3: Crossed classified model, adapted from (Snijders & Kenny, 1999) to take into account Sender, Recipient, Dyad and Message. - Covariates include, at the level of the

5.3.1 The Data Set

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 (see section 3.4.2). A group of highly active and well connected email users were selected from the dataset in the following manner. Two periods were chosen, one for the stimuli and an overlapping, slightly longer period for the responses, according to the consideration mentioned above. After that, all the emails associated with these time intervals were searched for pairs of stimulus-reply according to the principle explained in section 5.2.[1] In most dyads, only a single stimulus-response pair was identified, and the overall ratio of stimulus-response pairs found was rather low. To raise the ratio, dyads were retained only if they were associated with ten or more stimulus-response pairs of emails. The result was a network that contained one large component and several smaller disconnected components. Only members of the large component were retained. This process

[1]The time interval between stimulus and response was set to a maximum of two weeks.

yielded a group of 71 individuals, within which 396 (unordered) dyads engaged in at least one email transaction (at least one of the actors sent an email to the other). Of these non-directed dyads, 207 are symmetric in the sense that they exchange emails in both directions. The other 189 dyads are asymmetric (i.e., messages flow in one direction only).

For a standard analysis of social networks, the description of nodes and related dyads would be sufficient for the construction of a social network model. But this model requires the description of the transactions, which play an important role in the explanatory model. The data set includes 2973 email messages sent within a period spanning the months of September to December 2001. The number of recipients in the group of emails range from one to eighteen, a large minority of which are single-recipient emails. Each multi-recipient email can now be broken down into sender-recipient pairs. In other words, we treat an email sent to J recipients as if J copies of the email were sent, each copy to a single recipient. However, the common identifier of the email retains the affiliation between the different copies to the original email.

If we break down each email that was sent to J recipients into J 'copies' of the message, we find 4194 'copies' in the data set. Each 'copy' is taken as the micro-level case. The number of replies identified in the dataset is 540, which means that the overall proportion of replies is is 12.9%.

In terms of the crossed factor multilevel model, our dataset consists of 4194 cases (micro-level entities) each representing a dyadic sender-recipient interaction. Each case is associated with the four crossed factors (macro-level entities): 71 senders, 71 recipients, 396 dyads and 2973 messages. Figure 5.4 compares the rate of reply for ten of the actors in the network. Take for example actor 71 and actor 46. We see that when actor 71 is the recipient of emails coming from actor 46, the rate of reply is rather high. But when 46 needs to reply to 71, he does a poor job. The same actor can therefore play in different roles, the role of recipient and the role of sender. Overall the matrix is more or less symmetric, indicating that within every pair, if one actor has a high rate of reply to the other, the other responds in kind.

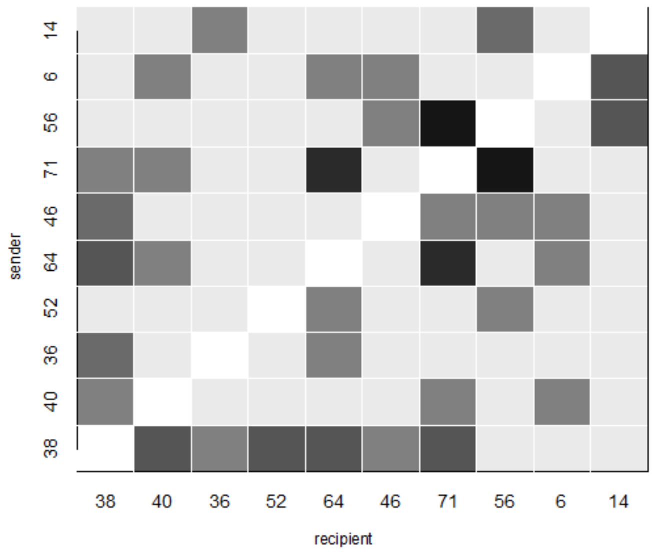


Figure 5.4: Sender/Recipient Rate of Reply - Aggregate rate of reply among a subset of the actors in the data set. The vertical axis denotes senders of stimuli emails, the horizontal axis denotes the recipients. The darker colors denote a higher rate of reply from recipients to senders. Note that the matrix has a weak but noticeable tendency to symmetry

5.3.2 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). The outcome variable y_{ij} is assumed to have a Bernoulli distribution:

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

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 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 $d[r,s] = d[s,r]$ [1].

[1]The model was designed to be as consistent as possible with (Snijders & Kenny, 1999)

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

To apply the multiple role model discussed above, the two residuals for any particular actor k correspond to the the two roles any actor can play (the role of sender and recipient), since u_k^{sender} and u_k^{recip} are random effects of the same actor, and therefore they are assumed to have a joint normal distribution with the exact composition of the dispersion matrix Σ of order 2 as presented below,

$$\begin{bmatrix} u_k^{sender} \\ u_k^{recip} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & \rho\sigma_s\sigma_r \\ \rho\sigma_s\sigma_r & \sigma_r^2 \end{bmatrix}\right)$$

Note the residuals for different actors are assumed to be *a priori* uncorrelated. The other effects are modelled analogously and are uncorrelated.

$$u_d^{dyad} \sim N(0, \sigma_{dyad}^2) u_j^{stim} \sim N(0, \sigma_{stim}^2)$$

For all models a uniform prior is assumed for the fixed effects and a flat prior (with lower bound of zero and upper bound of 100.) The priors used in the multivariate normal model assume a multivariate normal distribution for the two residuals (corresponding to the two factors) with inverse covariance matrix Σ^{-1} to which we assign the Wishart distribution with two degrees of freedom and the identity matrix for the scale matrix.

The first and second models estimate only the random effects, but the third model includes three fixed effects: first, we expect a greater probability for a reply from recipients addressed in the *to* field of an email than recipients addressed in the *cc* or *bcc* fields. A dummy variable denotes whether recipient i is assigned to the *to* field of the email j. It is assigned with values 1 if the recipient is in the *to* field, 0 otherwise. This is a micro-level effect unique to the email-recipient pair.

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 (see section 4.3) it is 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 characterized by frequent exchanges, and they are expected to yield a higher rate of reply from the actors involved according to Granovetter (1973).

5.4 Results

The three models were estimated using Markov chain Monte Carlo sampling approach. All models were fitted using JAGS version 2.1.0 (Plummer, 2004) running two parallel chains, discarding the first 3000 replicates and basing inference on the next 30000 for each chain. The findings presented in table 5.1 provide strong evidence that the four factors are important sources of variability in the effectiveness of emails to elicit replies, with a steady reduction of the Deviance Information Criterion (DIC), which is a goodness-of-fit index generalized from Akaike's Information Criterion (Spiegelhalter *et al.*, 2002)

	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)	0.0	0.0
σ_{dyad}		1.09 (0.14)	1.82 (0.31)
σ_{stim}			2.62 (0.52)
Deviance	2925	2833	2700

Table 5.1: 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 suggests that, at least in this specific dataset, actors who tend to reply to others also tend to elicit replies from others. The positive correlation between these two features can be seen in greater detail in figure 5.5. The figure also allows the identification of interesting or unusual actors. For example, actor 28 tends to have a strong sender effect (likely to get replies to her emails,) but only a modest recipient effect.

The second model suggests that adding the dyad effect reduces the variation explained by the sender and recipient effects. Also, adding the dyad factor necessitated the removal of the multiple-role model, and the correlation between sender and recipient random variables was set to zero.

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 broadcast emails are less likely to reply and (3) The frequency of prior email exchanges does not explain the likelihood for reply.

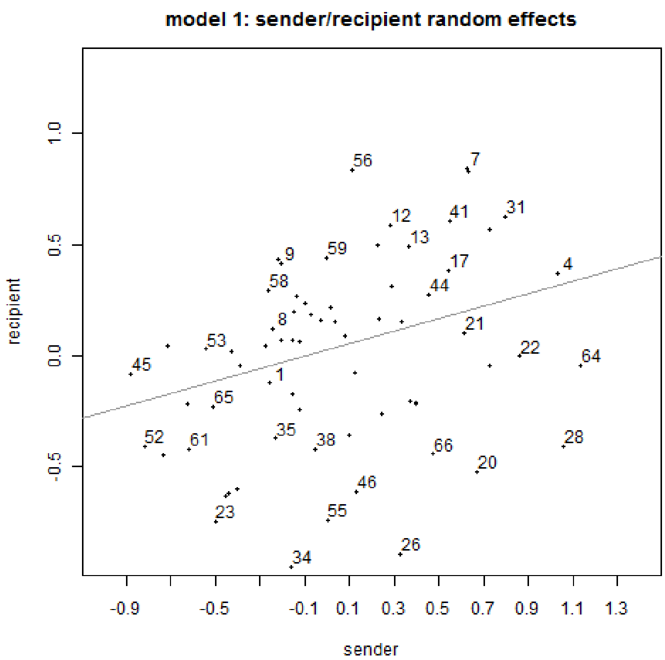


Figure 5.5: Comparing sender and recipient random effects. - Each email user is both a potential sender and receiver, and is thus associated with a variation from the estimated average effect

5.5 Summary and Reflections

This chapter 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 5.5), and between levels of abstraction (e.g., comparing the effect of ties, nodes and stimuli.) Substantially we see that ties are more important than nodes in explaining sources of variability. This suggests, rather than being a global property of individuals, replying to an email is rather like social capital, a resource that 'inheres in the structure of relations between actors' (Coleman, 1988). In other words, at least within this dataset, actor level effects are less important than dyad effects. This means that variations between actors are less important than variation between different ties of each actor. Put crudely, if you want to elicit a reply from your recipient, it matters less who you are or who your communication partner is, what matters is your relation to each other.

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 (Gibson, 2005; Butts, 2008). Second, it suggests a way to develop models of co-evolution processes that operate not only at the meso-level (nodes and dyads) (Snijders *et al.*, 2007), but also between the meso-level and the micro-level (transactions).

There are several insights gained from this study. First, it is interesting to note that sender and recipient variability are roughly the same. This means that throughout the three models, the variability in the outcome is shared fairly between the sender and recipient. We know from the literature that other types of social actions depend much more on the focal actor than on the target of the action (see for example (Snijders & Kenny, 1999)). Furthermore, it is likely that the positive correlation between sender and recipient effect is highly contingent on the context. In some contexts the correlation may be negative, if, for example, individuals occupying central roles in the organization have a high sender effect (the recipients of their emails tend to reply) but a low recipient effect (they are not necessarily responsive).

The method developed in this chapter can highlight various aspects about contingent to an organization. It can estimate variations in positions of actors and their relations. But most importantly it assesses the role and the consequences of social transactions. Multi-recipient emails are treated as a common background affecting the decision making of different recipients.

Probing the relationship between social transactions (e.g., a message, a rose or a wink (Oakeshott, 1991)) and social relationships (e.g., kinship, friendship or contractual relations) is an opportunity to highlight the differences between these concepts and develop conceptual tools to address their mutual influences. While remaining agnostic about the precise nature of ties, we can still argue that ties are not defined merely by their associated transactions, and that ties and transactions can be apprehended independently.

Arguably, this distinction is also at the very heart of the departure of the emerging field of 'computational social science' from traditional studies of social networks. Whereas the former focuses on networks of transactions, interaction and communication (Monge & Contractor, 2003), the latter focuses on social ties and the social capital that 'inheres' in their structure (Coleman, 1988). By combining dyads, actors and transaction into the same model, this chapter argues for the potential in bridging these two agendas

6

Discussion and Conclusion

Wandering through the frontiers of the sciences, and the arts, I have always trusted the eye while leaving aside the issues that elude it. It can mislead, of course, therefore I check endlessly and never rush to print.

Meanwhile, for over fifty years, I have watched as some disciplines exhaust the 'top down' problems they know how to tackle. So they wander around seeking totally new patterns in a dark and deep mess, where an unlit lamp is of little help.

But the eye can continually be trained and, long ago, I have vowed to follow it, therefore work 'from the bottom up.' Like the Antæus of Greek myth, I gather strength and persist by often touching the earth.

A few of the truths the eye told me have been dis-proven. Let it be. Others have been confirmed by enormous and fruitful effort, and then blossomed, one being the four thirds conjecture in Brownian motion. Many others remain, one being the MLC conjecture about the Mandelbrot set, in which I believe for no other reason than trust in the eye.

Benoit Mandelbrot,
Response to the Edge Annual Question (2005)

6.1 Introduction

This work is driven by the following question: what are some of the mechanisms that link social transactions (at the micro-level) with social-ties and transaction-patterns (at the meso-level) and network structures (at the macro-level). The two empirical chapters demonstrate some of these mechanisms: Chapter 5 shows how the social action of replying to an email is influenced by entities at the micro-level (the email stimulus itself) and at the meso-level (properties of individuals, transaction patterns and ties.) Chapter 4 demonstrates how private and broadcast emails are stimuli (at the level of transactions) that prompt different types of responses (reply, reply-all or forward,) at the aggregate level contributing different types of structures to the network.

This final chapter discusses these findings in greater detail and concludes the thesis. It begins with a general critique of the entire work and an attempt to address it, and then organizes the empirical findings in light of the Coleman diagram (see section 2.2,). This is followed by a summary of the contributions, the limitations of this work and a critique of sorts. The chapter ends with a debate within the social sciences about the role and significance of digitally mediated transaction datasets.

6.2 Learning about social ties from transactions alone

The last empirical chapters have raised a general concern among readers of earlier drafts, a concern that I would like to address upfront. The problem arises from the gap between the concept of social ties and the type of data used in this dissertation. Throughout the dissertation a claim was made, that social ties cannot be simply reduced to a bunch of transactions (both in the introduction chapter and especially in section 2.3.3.) In other words, I argued for a distinction in kind between ties and transactions. However, the empirical chapters made use of data that refers to transactions exclusively, not to social ties. How then, could I make any reasonable claim about ties if I have no (independent) data about them?

This is a valid concern, and the lack of independent data sources for social ties is indeed a limitation, compared, for example, to a study authored by Quintane

& Kleinbaum (2011). However, there are two ways of replying to this concern, as described in the following two subsections. Each of these two subsections appeals to a different audience. The first appeals to nominalists and the second to realists; recall from the discussion in section 2.3, that nominalists insist there is no different in kind between social ties and patterns of transactions (see the discussion about network reductionists in section 2.3.3). The realists believe that there is a difference in kind between social ties and patterns of social transactions. Both schools of thought deserve an answer, as detailed below.

6.2.1 An appeal to nominalists: ties are equivalent to patterns of transactions

As described in section 2.3.3, nominalists reject the distinction in kind between ties and transaction-patterns, insisting on defining the latter in terms of the former. For nominalists there is no problem, because all the information is already in the transactions. Recall that in chapter 4, all the network models were in fact networks of transaction patterns, because there was no independent way of measuring the association between individuals in the dataset. There is nothing outlandish about the claim that one can define the meso-level object in terms completely reducible to micro entities (this is the definitional macro-micro link in section 2.2.) On the contrary, this claim is closely related to claims made by Max Weber, George Caspar Homans and Monge & Contractor (2003) (as discussed in section 2.3.3.)

Arguing that there is nothing interesting to say about transaction-patterns just because the data is all about transactions, is equivalent to the argument that there is nothing interesting to say about networks just because the available data about individuals and ties. Such a claim would pull the rug from under the feet of much of the body of network studies that rests on survey based, traditional network datasets, since these are nothing but reports of people's personal ties. It would be rather puzzling to claim that you cannot study networks just because you do not have an independent source of information on the network as a whole (such as an organigram in formal organizational settings, or by following the paradigm of Cognitive Social Structures Krackhardt, 1987, for example). But in practice, most empirical studies of networks do not rely on anything but a set of personal ties, those reported by individuals in surveys or interviews. Those reported ties are aggregated to form the network, without taking into account any independent measure at the (macro) level of the network. And although they have only data about ties, researchers can still say a whole lot about the aggregate networks as a whole.

Are networks anything different than a collection of individuals, their properties and the ties connecting them? Realists would argue that the answer is probably yes, referring to properties such as group identity, solidarity and perhaps a set of formal or informal rules or goals of the group qua distinct unit. But nominalists would accuse realists of engaging in mysticism.[1] There is little hope that this ontological question about the link between the whole and its parts will be solved any time soon. But I don't think we have to reach a resolution of this debate in order to recognize that whatever one's conviction on the matter is, those who treat networks as completely reducible to individuals and their connections still have interesting things to say about networks. By the same token, even if we accept that ties are something 'over and above' transactions, those who define ties in terms of transactions might still have something interesting to say about ties.

How do transactions (sending an email for example) compare to network events (tie formation, dissolution etc. see section 2.3.2)? Some authors (de Nooy, 2011; Brandes, Lerner & Snijders, 2009; Butts, 2008) do not find anything very interesting in the analytical distinction between these two types of events, transaction-events at the micro level and network events at the meso-level of ties and nodes. For them, when actors engage in a social transaction, they are merely 'selecting' a tie, an equivalent to the process of tie-formation. For example, de Nooy (2011) considers the act of publishing a review on a book equivalent to a 'selection' event in which a new tie is formed between the critic and the reviewed author. Brandes, Lerner & Snijders (2009) write that the only difference between surveying people about their social ties and collecting data about transactions (email, phone-calls etc.) is a technical one, namely that the former are 'panel data' and the latter 'event data.' The consequence is that event data includes the precise moment in

[1]See Tarde's quote opening the Chapter 2.

which the event occured, and panel data only includes the information that it occurred at some unkown moment between two panel waves. But this is just a technical difference, one with implications on the statistical method and not one which is of any theoretical significance. The authors consider political events such as ‘visits, agreements, and provision of military aid, accusations, threats and military actions’ as tokens of tie-formation. Most explicitly, (Butts, 2008) claims that ’relational events are temporally local phenomena, and thus represent the opposite end of the temporal continuum from the (relatively) long-term structures that have formed the primary subject matter of classical network analysis.’

Those authors do have a point: in some situations the distinction between network events at the meso-level of ties and transactions at the micro-level may seem redundant, indeed a hair-splitting exercise. They are also right to point out the similarities between events on both levels of abstraction. For one thing, note the principle of interdependency between events, one event triggering the next in sequence. Sending an email to two recipients makes it likely that the two will start contacting one another, an example of one social action triggering the next in a stimulus-response chain of transactions. By the same token, befriending two strangers makes it likely for these to become friendly one day. But whereas the first example is clearly a stimulus-response link between two transactions, would we be comfortable to say that the second example is of the same kind?

6.2.2 An appeal to realists: ties are more than transaction-patterns

There is a second, stronger and perhaps more controversial (Vromen, 2010; Abell *et al.*, 2010) answer to the concern. We could maintain the claim that ties and transaction-patterns are different in kind, but qualify the claim by adding that they are not completely orthogonal to one another, both levels of analysis influencing, shaping and leaving their traces in one another (in other words, there exists a process of co-evolution between ties and the transaction patterns with which they are associated.) Thus, even if we do not have access to independent data about the tie as a whole, we have access to something that covaries with it, namely the transactions. In other words, by virtue of the mutual influences between ties and transactions, one could impute properties of the ties by observing transaction data alone, just as one could look at a person's traces in the sand and say something reasonable about the person's properties without actually having independent data about the person herself. Indeed, this is exactly the approach taken in Chapter 5, and to a lesser extent in Chapter 4.

Having no direct or independent access to a phenomenon does not mean that we cannot gain knowledge of it. This is really the basic assumption of epistemic realism. In simple terms, consider the way latent variable analysis works. We accept the existence of a latent construct, say intelligence, a feature of an individual to which we have no direct access and no independent measure. It would not do to ask a person how intelligent she is, because she might not know. Instead of observing intelligence directly, we gain access to measurable indicators that covary with intelligence, so-called manifest variables that evaluate measurable properties such as mathematical and verbal performance. By measuring what is easy to access and making reasonable assumptions about the distributions of a property in the population, we can say a whole lot about that which is unknown and not directly accessible to independent measurement. There would be few people who would claim that intelligence is defined by, or reducible to mathematical and verbal performance. And yet the consistent correlation between these skills suggests that there is some third confounding variable, such as intelligence, that drives the correlation between these two indicators.

This relationship between manifest transactions and the latent social ties is relevant to the findings in both empirical chapters. Chapter 5 demonstrated that even after controlling for various effects at the level of the sender, the recipient and the transaction, there was still a residual effect that was associated with a specific pair of individuals. Whatever effect depends on the dyad of email users and after controlling for the individuals and the transactions, must be attributed to the latent variable at the level of the tie. And since a significant dyad effect was found, we must conclude that it has to do with an unobservable property, related to a specific sender-recipient pair. This can only be an effect of the tie, which is (arguably) irreducible to the transactions themselves.

How does that relate to the link between transactions and ties? The previous subsection introduces an argument made by scholars who reject the distinction in kind between social transactions and ties. But there are counter-arguments as well, some of which were discussed in section 2.3.3. It seems to me that the arguments for a difference in kind between transactions and ties are more convincing, but my point is that whether one agrees or not, transaction data is a sufficient resource to learn about processes of co-evolution between transactions and social ties, independently of how one might define the latter.

6.3 The empirical findings in light of micro-foundations

Recall that the research question asked us to identify mechanisms of co-evolution between communication transactions and network structures in the context of email communication. In more detail, we would like to know how we might use (email) transaction data to account for the links between the macro-level of the group, the meso-level of ties and transaction-patterns (properties such as strength etc.) and the micro-level of social transactions (sending an email, replying to one etc.) Special attention was given to the distinctive properties of emails as a communication medium, specifically the notion that emails have the property of being assigned to multiple recipients. We shall now review the findings and try to fit them into the Coleman diagram depicted in figure 6.1.[1]

6.3.1 Chapter 5 Results and micro-foundations

Let us start with the findings in Chapter 5. An incoming email (stimulus) lands in the inbox of one of its recipients, and the recipient has to make a decision between two courses of action: to reply or not to reply. The situation is similar to the one facing the doctor in Coleman's hospital, as she faces the decision whether to adopt the new medicine or not (Coleman *et al.*, 1957). We must imagine the recipient (or the doctor) at the lower right hand of the Coleman diagram, two types of 'forces' or considerations impinging on her. Arrow number 2 in the diagram represents the unique properties of the transaction preceding the moment

[1]A note of caution: the diagram is a loose and general framework. It works well for diffusion processes as discussed in section 2.2 and I suppose it was built with this kind of process in mind, but it is applicable in many other organizational contexts as well (Abell *et al.*, 2010). But I think there are cases where the assignment of explanation to an arrow in the diagram could be somewhat loose.

of decision and influencing its outcome. In the recipient's case it was the stimulus email, in the doctor's case it was the consoltations she had with colleagues.

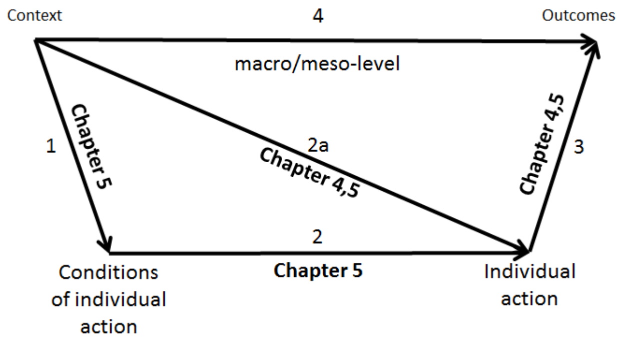


Figure 6.1: Organizing the contributions according to Coleman's diagram. - Main contribution of Chapter 5 along arrow number 2 and 2*a*, main contributions of Chapter 4 along arrow number 3

The results in table 5.1 indicate significant effects that include but are not limited to the number of co-recipients and whether or not the recipient was assigned to the ***to*** field on the message. Confirming the findings from Chapter 4, we find that the more recipients associated with the email, the less likely the reply. In contrast to the findings from Chapter 4, here we have direct evidence that the relation between the incoming transaction and the outgoing one is of a stimulus-response type, and not independent messages flowing both ways. In addition, we find that the assignment to the *to* field increases the chances of a reply - both effects are significant. In addition, there is a significant effect of variability that is due to the message itself. That is, variation in reply-likelihood between the decisions made by recipients of the same message is significantly less than the overall variation in reply-likelihood. There are some unknown factors unique to the message itself that affect the rate of reply, most probably due to its content or the moment in which it was sent. Something specific in the message and not the identity of the sender nor the receiver nor the combination of the two is responsible for this, since those were controlled for.

Besides the effect of the message, there are effects of the identity of the sender, the recipient and the tie between them. These are captured by arrows number 1 and number 2*a*. Both arrows reflect the effects that macro/meso conditions have on the decision of the recipient, and they include the organizational roles of the sender and the recipient, their individual properties that might influence them to deviate from the average likelihood to reply, and the properties of the relationship between them.

Of all these, the most relevant for the debate about the nature of the social tie, is the variation in the likelihood of reply that is associated with the specific sender-recipient dyad, after controlling for the overall properties of the sender and the recipient. First, on the dyad level, the frequency of email exchange prior to the email in question is not significant, after controlling for the other variables. This is rather surprising, because it means that the likelihood to reply is not significantly correlated with the frequency of messages exchanged after controlling for the other factors.[1]

But I think that conceptually, the most interesting finding is the significant variability that is attributed to the dyad, after controlling for all other issues. What is at stake here is that between two individuals A and B, A could have a lousy sender effect (virtually nobody replies to her emails) and B could have a lousy recipient effect (virtually never replying to her emails.) However, when A sends B a message, the rate of reply is exceptionally high. There is something latent in the dyad, the relationship between the two actors that influences the likelihood for reply, in a manner that cannot be explained by looking at the overall properties of the two individuals separately.

This can be seen as a novel method to measure a property of the tie, a method that does not depend on the frequency of exchanges. It is the property of the dyad that influences the likelihood of replying to an email, and there is substantial variation between one dyad and another in respect to this value. This attribute is closely related to Granovetter's definition of strength of ties (Granovetter, 1973), when he defines it as 'reciprocal services.'

Although the empirical investigation models a micro-level outcome, figure 5.5 describes a pattern at the macro-level of the group as a whole, a slight but significant correlation between people's sender effects and recipient effects. The pos-

[1] Recall that from Granovetter's definition of the strength of ties Granovetter (1973), frequency of exchange and reciprocity are both associated with the strength of the tie, so we might expect a correlation here. In this context the lack of correlation between reciprocity and exchange rate was also found in mobile data by Kovanen *et al.* (2010).

itive correlation suggests that in this group of people, those who have a higher (lower) than average sender effect also have a higher (lower) than average recipient effects. Those who tend to receive more (less) replies to their emails tend to reply more (less) freqeuntly. Putting it a bit extravagantly, in this group, one might conclude, there is an overall sense of justice when it comes to replying to emails. This macro-level pattern is a result of the aggregation of many patterns of reply at the micro-level, and is therefore associated with arrow number 3 in the Coleman diagram.

6.3.2 Chapter 4 Results and micro-foundations

Let us now turn to the results from Chapter 4. The key result is that private emails contribute more to the reciprocity of the network and less to its transitivity, as compared with broadcast emails. We do not have any hard evidence to suggest why this is the case, but there are some reasonable explanations for this, all of which have to do with the way people use the email medium to carry out their communicative transactions. Thus, these results are associated with arrow number 3 in the Coleman diagram, micro-level transactions give rise to network level patterns. I shall now suggest several explanations that could explain why we might see diminishing reciprocity[1] with increasing recipient number.

1. The simplest explanation is that this is a spurious correlation and there is a third confounding variable which affects both reciprocity and recipient number. The confounding variable is the reason (read: purpose, motivation, expectation) of the sender for sending the stimulus email. When individuals make an announcement or want to disseminate information, they may want to act on a large group of people and make them aware of the content of the message without necessarily expecting them to act upon it.

[1] Note that I use here 'reciprocity' to denote stimulus-response type patterns, where the incoming email is the stimulus and the response of the recipient is in the form of a reply. Reciprocity thus defined has not been tested directly in Chapter 4. What was measured was merely two-way communication exchanges between pairs of actors, or symmetry. To test stimulus-response transactions in the data, it would be necessary to identify and associate between stimulus messages and response messages, for example by comparing between subject line fields, as was done in Chapter 5.

To achieve their purpose, senders dispatch emails to a large group of individuals (hence 'broadcast message',) signaling to recipients that a reply is not expected. According to this mechanism, recipients do not refrain from replying because of the long recipient list per-se, but because they judge that the sender does not expect them to reply. Hence the correlation between recipient number and reciprocity is confounded by the reason (expectations, purpose) for sending the email. In terms of the Coleman diagram, the decision of the recipient not to reply is the consequence of the transaction, therefore this mechanism is associated with arrow number 2.

2. Composing a single-recipient email involves a certain amount of time, attention, effort and resources, all concentrated on the relationship with a single contact. In contrast, when sending broadcast emails, the effort is distributed among its numerous recipients. From a theoretical point of view, one could apply exchange theory (Emerson, 1976) to explain why an incoming private email (especially a long one) might elicit a sense of obligation and duty to reply back, more so than an incoming broadcast email. Consequently, an act of reply to a private email would be driven either from the need to conform to a norm (arrow number 2*a*,) or from a strategic decision (arrow number 2,) to pay off one's 'social debt' to the sender.

3. A related explanation would be that because of the greater effort per recipient inolved in private emails, sending them may be associated with a stronger tie. Recall that Granovetter (1973) defines the strength of a tie in terms of 'reciprocal services' and the 'amount of time' associated with the tie. In this latter explanation, the strength of the tie operates as a confounding variable that drives both higher reciprocity and fewer recipients. Since the strength of the tie is a meso level property acting on both sender and recipients from the height of the meso-level, it would act on the sender via arrow 1 (driving her to send private emails to her strongest contacts) and on the recipient via arrow 2*a*, driving her to reply to her strongest contacts by virtue of the strength of the tie connecting them.

4. Another mechanism could be related to the theory of collective responsibility or social loafing Karau & Williams (1993). According to this argument, the 'responsibility' to reply is 'distributed' among the recipients. As their number increases, a higher level of 'defection' (read: no reply) is expected. Here the recipient is acting strategically along arrow 2, her decision interacting with what she hopes would be the decision of other co-recipients.

5. Finally, consider a mechanism that operates in the opposite direction to those mentioned above. The success of spam (unsolicited bulk email) to elicit a response among recipients relys on being distributed to numerous recipents (Cranor & LaMacchia, 1998). The more recipients, the more likely to elicit a response. This is an interesting exception to the mechanisms mentioned above because flooding the system with emails is a macro-to-micro process acting on individuals through arrow number 2*a*. Arguably, the act of sending unsolicited mail is no longer at the micro-level of transactions but at the top left corner of the Coleman diagram, impinging on a large collective. Notice the difficulty to adjudicate whether a social-phenomenon (such as sending out broadcast emails,) should be taken at the micro-level or the macro-level. I suspect this marks the beginning of the cracks in Coleman's theoretical framework, an issue I shall return to in section 6.4.3.

Of course these mechanisms should mark the beginning, not the end, of an entire empirical investigation, one that attempts to tease out which if any of these mechanisms operates, what are their relative strengths in driving a reply to an email, under what conditions and in what contexts does each of them operate.

The second important finding in chapter 4 suggests that broadcast emails contribute to the transitivity of the network. This means that they tend to be sent to people who exchange email messages. There could be one of two explanations for this; either the broadcast emails acts as a stimulus, triggering direct exchange among co-recipients (micro-micro link associated with arrow number 2,) or the email was sent to these recipients together because the sender knows they are already connected to one another (macro-micro link associated with arrow 1 or 2*a*). It is also possible that multiple recipient emails are part of discussion threads among a group of people, and the discussions proceeds every time one of the recipients uses the 'group-reply' feature, thus increasing transitivity in the network. For example, given a broadcast email sent to n recipients, the first to use the 'group-reply' feature adds $n-1$ triangles to the network, possibly increasing transitivity in her region substantially. The transitivity finding is related to two additional empirical results.

1. There is a distinction between sending one broadcast email and sending multiple private emails, the former being a better indicator of contact between recipients than the latter. But this type of distinction is meaningless when we look for the equivalent at the meso-level of social ties. Transitivity at this level only tells us that when strangers share a common friend, they are likely to become friendly themselves. When applied to the macro-level of tie configuration, transitivity is more abstract and simple. Taken down to the micro-level of transactions, we need to be more precise about the properties of the transactions taking place.

2. Another interesting result is seen in figure 4.4. As the number of email recipients increases, and against the background of steadily declining reciprocity, the level of transitivity first increases, reaches a maximum and then begins to decrease. If the assumption is true, that transitivity levels reflect group discussions, perhaps the peak of the curve designates a socially significant value, such as an optimal number of participants for collaboration or discussions in emails. However, at this point we cannot be sure. To further explore this proposition one needs to study the chains of related messages.

Taking both findings together, the recipient number correlated with reciprocity and transitivity, we reach a rather surprising conclusion which I will explain below. The conclusion is this: if we adopt the definition of the strength of ties from Granovetter (1973), we see a violation of the strength of weak ties hypothesis in the context of email mediated transactions. Though it would be hasty to claim that this is a general law, the explanations I offered above, to the extent that they are the ones that account for the observed pattern, are not context dependent. They suggest that the violation has something to do with the way people communciate with one another through the medium of emails in general.

Let me explain where the violation comes from. Recall that the hypothesis states that stronger ties are embedded in network regions of higher density. But this cannot work in the context of emails if we accept the key finding that private emails contribute more to the reciprocity of the network and less to its transitivity, as compared with broadcast emails. On the contrary, private emails are associated with more reciprocity and more frequent exchanges. Thus, private emails are consistent with the stronger ties. However, group level discussions and collaboration tasks require broadcast emails, which also contribute more to transitivity, and the 'reply-all' feature increases the number of closed triangles substantially, even if only a few of the recipients use it. Thus it is the stronger ties that are mediated by private messages, which are not embedded in dense regions. Weaker ties that are mediated by broadcast messages are embedded in denser regions. As long as these two mechanisms are at work, they push the network to violate the strength of weak tie hypothesis.

This is probably the most interesting finding in the Chapter, a mechanism associated with arrow number 3 in Coleman's diagram: under certain conditions and thanks to the unique features of the email artefact, email communication networks violate Granovetter's strength of weak ties hypothesis.

6.4 Review of the contributions and critique of sorts

This section consists of a brief overview of the main contributions of the thesis, organized into three categories as follows.

6.4.1 Distinctive Properties of Networks of Transactions

> Insanity in individuals is something rare - but in groups, parties, nations and epochs, it is the rule
>
> Friedrich Nietzsche, Beyond Good and Evil

The micro-macro distinction is a fundamental theoretical foundation, well known both in the social and physical sciences, even marking an institutionalized division of intellectual labour between microeconomics and macroeconomics. Whereas micro-economics addresses individual entities (such as buyers, sellers, households, firms) as they behave within an exogenously given environment, macroeconomics addresses the aggregate effects of economic activity (such as inflation, unemployment, productivity.) These and other intellectual projects demonstrate time and again, that even if we define the collective in terms of individuals and nothing 'over and above,' we find system effects. By this I refer to phenomena in which the properties of the aggregate differ from the properties of its members taken one by one. Here are some examples of famous system effects:

1. ***Condorcet's Paradox.*** Transitivity does not scale, so that even if each individual in a group has transitive preferences, aggregate preferences of the group are intransitive (Gehrlein, 1983) .

2. ***The Doctrinal Paradox.*** Logical consistency does not scale, so that even if each individual in a group is logically consistent, aggregate judgement of the group is logically inconsistent (List & Pettit, 2002).

3. ***The Prisoner's Dilemma.*** Utilitarianism does not scale, so that even if each individual in a group is rational and utilitarian, their decisions interact with each other making all concerned worse off.

The above examples are taken from economics, game theory and political science, and there are plenty of additional examples from physics (Anderson, 1972), law Vermeuele (2009) philosophy and sociology (Elster, 1989b; Jervis, 1997). It would be incorrect to say that this dissertation addresses system effects, but perhaps one could say that it handles a problem that has a weak form of family resemblance with these effects. I mean this in the sense that theories and concepts that work at the level of traditional social networks, seem to work differently at the level of networks of email transactions.

Obviously, such differences are known in the literature and some have been reviewed in previous chapters. Recall, for example, the study by Quintane & Kleinbaum (2011), comparing between two social network models associated with the same group of 23 individuals, one model based on reported ties and the other based on email mediated transaction data. The study finds different network mechanisms operating in the two network models. Kovanen *et al.* (2010)

find a complicated relation between the reciprocity of calls, and their frequency and duration in the context of mobile phone communication networks, which is not what one expects to find in social networks. Finally, Liben-Nowell & Kleinberg (2008) find that chain-mail networks exhibit very different properties from the small-world network of email users that produced it. All these studies indicate that when considering the 'nuts and bolts' of digitally mediated transactions, aggregate patterns have very different patterns to what one might expect from traditional network studies. This dissertation contributes to this line of thinking in a couple of ways.

1. ***Mechanisms.*** Mechanisms operate in a different manner in traditional networks of friendship, say, and email transaction networks. For example, although the strength of weak ties has been confirmed in numerous studies of traditional social networks, in this thesis I argued why there might be a mechanism that operates in the opposite direction in the context of email networks. Here is a mechanism that works one way at the level of social ties, and another at the level of email-transactions. There might well be other such mechanisms.

2. ***Concepts.*** Various well known concepts from traditional network literature seem to work differently in the context of email mediated transactions. Traditional network literature takes notions such as reciprocity and transitivity, abstracting them away from the medium of communication and interaction in which they are preformed (this critiqe very similar to the one raised in Feld, 1981; Feld & Elmore, 1982). More specifically, it is completely innocent of stimulus-response chains that are the defining property of human transactions (Gibson, 2005). This innocence is not due to oversight but is intended, part of the Durkheimian strategy inspired by the founding fathers of the field (see section 2.3.3.) But chains of transactions have macro-level consequences. It is what drives the autocorrolation and burstiness of human transactions (Barabási, 2005), and it is what drives reciprocity in Chapter 5 and most probably in Chapter 4 as well. Let us consider a couple of typical network concepts in this light.

(a) ***Reciprocity.*** In traditional network research, the term simply denotes matter, services or information flowing between two individuals in both directions. At the level of transactions it means that what flows in one direction is linked to what flows in the other via a stimulus-response process. Reciprocity at the micro-level is thus understood as a stronger version (or narrower definition) of reciprocity at the macro-level.

(b) ***Transitivity.*** The abstract and general mechanism associated with the term in the context of traditional networks (strangers with common friends tend to become friendly) is qualified at the level of transactions. In the context of emails we need to make a distinction between, say, sending multiple private emails or sending one broadcast email, two transactions that have different consequences in terms of the likelihood of a mutual contact to induce a connection between strangers.

(c) ***Degree Distribution.*** In traditional networks this term denotes the distribution of the number of friends. One could distinguish between in-degrees (number of nominations recieved in a survey method) and out-degree (number of nominations given.) But that's about it. The non-random distribution of degrees has attracted debate that spanned many decades (Moreno & Jennings, 1938; Barabási & Albert, 1999), partly because it is precisely the non-random distribution that is believed to conceal a peculiarly 'social' mechanism. However, when we start looking at transactions, we see many more non-random distributions, each demanding a social explanation. Chapter 4 introduces three such distributions, the production, consumption and dissemination of emails, each of the distributions non-random for reasons that may be of some consequence.

3. ***Interdependencies.*** Over fifty years ago, Siegfried F. Nadel (2007, p 227) described what he found most interesting in the concept of 'networks' with the following visionary words: '…I do not merely wish to indicate the 'links' between persons; this is adequately done by the word relationship. Rather,

I wish to indicate the further linkage of the links themselves and the important consequence that, what happens so-to-speak between one pair of [nodes], must affect what happens between other, adjacent ones.' Nadel was looking mechanisms that govern tie-interdependency. In traditional networks, tie interdependeny works mainly through mechanisms such as popularity, transitivity and homophily. But in email transactions the logic is much more complicated, and it is noted not only by the observing scientist but also by the actors themselves. Multiple recipient emails are a powerful example of how connections between people are instantiated or reinforced in tandem. Recipients become aware of others' communicative transactions and study them in detail, informing themselves of the existence of other co-recipients, looking at whether or not they were assigned to the *to* field, assembling the evidence to judge the expectations of email sender and co-recipients, finally reaching a decision about the most adequate response. The type of interdependency at the level of transactions involve intention, purpose and meaning for the individuals themselves. At the level of traditional networks, they are much less concrete and immediate.

4. ***The technological medium.*** The medium through which people communicate is all but irrelevant for traditional work on social networks.[1] But the emphasis on transactions bring the distinctive properties of the technology into relief. Emails can be sent to a single recipient or to multiple ones. Recipients can be assigned to the *to* field or to the *cc* field. These are all small choices on the part of the sender, with consequences on the action of recipients. Unlike traditional networks where we find individuals connected to one another, here one person interacts with others via a medium, and the properties of the medium are involved in the transaction and influence its outcome. Furthermore, investigating transactions challenges the notion that technology is merely a great tool for researchers to elicit data (Lazer *et al.*, 2009b), and encourages the notion that it is an active participant in the situation, steering the network's unfolding structures.

[1]But see one exception in Licoppe & Smoreda (2005).

The first contribution is the distinction between networks of transactions and traditional networks of ties: different mechanisms are at work, and some of the concepts might need to be revised or qualified in order to capture more precisely the structures of transaction chains. Finally, the notion of interdependency becomes more qualified and the specific properties of the technological artifact play a more important role.

6.4.2 Methods and Limitations

The methodological chapter (Chapter 3) problematizes the process of constructing network models from transaction data, highlighting information that exists in the dataset but through the process of its construction, disappears from the network model. It argues that this information is relevant for the exploration of social mechanisms that are involved in shaping the network. One example of lost information is the differences between private and broadcast emails. Two methods were developed in Chapters 4 and Chapter 5, that seek to address this issue to some extent.

The literature demonstrates that decisions made by modellers (e.g., thresholds, noise filtering techniques) have an impact on the properties of the resulting network (De Choudhury *et al.*, 2010; Grannis, 2010), but it is not clear what are the structural implications, say, of filtering out broadcast messages. For example, Kossinets & Watts (2006) filter out emails with more than four recipients. The findings in Chapter 4 suggest that filtering out broadcast emails would potentially underestimate transitivity and overestimate reciprocity in the network.

The Chapter also developed a method to incorporate the number of email recipients into the strength of the tie, in a similar way carried out by Newman (2001a) for networks of authors of academic papers. Regression analysis confirmed the hypothesized correlation between levels of reciprocity and the strength of the tie thus calculated. In Chapter 5, a different measure of a tie was devised, a value that indicates the likelihood that either of the individuals associated with the tie would reply to an email, controlling for their overall reply pattern and for the properties of the message. Both these methods could be used in future research for constructing networks.

In addition, Chapter 5 devised a statistical method to tease out the various factors that influence a recipients' decision to reply to an incoming message, and to compare between the relative importance of each factor. This method may have some practical applications. For example when launching marketing campaigns, companies may like to know why some campaigns are successful and others fail, what factors are involved in determining success/failure, and what is the relative importance of those factors. Factors to consider include the identity of the company launching the campaign (sender,) the identity of the customer (recipient,) the campaign itself (message) Or the special organizational relationships between a company and a customer (tie.) The method developed in Chapter 5 could suggest where to start looking for answers to these questions.

There are several limitations to this work, and because of its exploratory nature it may raise quite a few questions and opportunities for further research. Some of the arguments could be presented more formally, so they could be more amenable to further study and investigation. Theoretically, the previous sections introduced a whole range of mechanisms that could potentially explain the phenomena of reciprocity and transitivity in email network datasets. These could be rigourosly tested in order to understand how they work and how they interact. Most importantly in the context of Chapter 4, these mechanisms all assume that observed structures of reciprocity and transitivity are due to stimulus-response type patterns at the micro-level of transactions. This assumption was only assumed, and it should be directly measured to verify that these are indeed the mechanisms at work.

Empirically there is a potential to expand on the current work. first, there is a need to reproduce the findings in a different email datasets, ideally backed up with survey based network data. Additional datasets can rule out the explanation that the findings are unique to this set of empirical data. That said, I think the explanations are compelling and that the number of recipients is indeed a confounding variable that leads to a (spurious) negative correlation between reciprocity and transitivity. If such a violation is not captured in other datasets, one may want to investigate whether there is a different mechanism that works in the opposite direction. In that case there would be a need to control for that mechanism in order to disentangle the two.

It may also be interesting to look for independent ways to confirm the violation of the strength of weak ties hypothesis. In fact, one attempt is made in Chapter 4 by regressing the strength of each tie against the ratio of mutual contacts of the two individuals associated with it. Unfortunately, no significance was found so it is hard to interpret the result, except saying that if the effect is there, this method of regression is either not adequate or does not have the power to capture it. In addition, a more systematic way is needed to compare between private and broadcast email mediated transaction networks, controlling for issues such as network density. A natural candidate for such a method would be ERGM,[1]but the limitation with ERGM is that large data models often degenerate,(Snijders *et al.*, 2006) and the method requires a rather steep learning curve to run it and interpret its results. In fact, not only does it fail to converge with large datasets, the assumptions that justify using it are incorrect when using it on large networks.(Snijders *et al.*, 2006) In any case, it would be reassuring to find an independent way to confirm the violation of the hypothesis, both on another dataset and by finding an independent method to test the correlation between strength and local clustering.

Also in Chapter 5 it would also be interesting to reproduce the findings in another dataset and consider adding more covariates into the model. These might include properties of individuals (their position in the organizational hierarchy for example) or even attributes of the tie such as reporting relations in organizational settings. In case there is independent survey data, one could add a covariate to flag whether or not a dyad was reported as a tie in the survey data. These covariates should control for some of the variation associated with the nodes and ties, as well as controlling for standard network mechanisms such as homophily. One could also consider grouping the emails according to their content and adding dummy variables to see if they control for the variation attributed by the emails, or better still, one could even group the emails into stimulus-response chains, and take the emails to be nested within the chains. One drawback of the model is that it is very complicated, consisting of crossed factors, multiple roles and a binary outcome variable. All this makes convergence difficult to attain in larger datasets, putting a limitation on the utility of this method.

[1]For abbreviations and nomenclature see the glossary on page xii.

Recall that the findings in Chapter 4 explained network patterns of reciprocity and transitivity based on stimulus-response type chains. But these chains were only assumed, not directly observed. The objective of Chapter 5 was to measure these chains directly, but work has been limited only to the mechanism of reciprocity. It would be interesting to complement this work with a measure of stimulus-response chains that account for transitivity. A question to investigate would be this: what is the likelihood for incoming broadcast emails to trigger email exchange between co-recipients? Though applied to transitivity, the objective here is the same, to link the network topological structures to sequences of stimulus-response type chains.

6.4.3 The Coleman Diagram

From a theoretical point of view, this dissertation can be seen as an attempt to apply the Coleman diagram to the a study of email mediated networks, and now we are more or less in a position to judge how fruitful this attempt has been. I think that as a general theoretical, or rather meta-theoretical framework, the diagram has merits in guiding the process of investigation and aiding the interpretation of the results. It organizes the findings, mitigating the danger of macro-to-macro explanations and acting as a sensitization device to the role of micro-level transactions as a locus of structural change. It helps thinking about the necessity to tease out the effects into those that operate via micro-level transactions (arrow number 2) and those that operate uniformly on all individuals via macro-level entities (arrows number 1 and 2*a*.) Furthermore, it creates a common vocabulary for communicating and thinking about the findings and the way they complement one other. Associating the mechanisms with the various arrows produces a sense of completeness, ensuring a more or less exhaustive account of all the types of mechanisms involved in the process of change, lending the exploration a sense of progress in which one assembles different pieces of a puzzle into one theoretical whole.

Theoretically, the Coleman diagram facilitates the discussion of several ideas. The lower part of the diagram is often said to be associated with the micro-level of interactions. This raises the question, should network events be considered equivalent to social transactions? The literature on the formation of friendship asserts that friendship is formed as a culmination of a process that involves multiple transactions (Hartup & Stevens, 1997; Lazarsfeld & Merton, 1954). This suggests that to reach rock-bottom explanations, one would need to dig deeper into transactions.

The Coleman diagram does present challenges, though. For one thing, at times it is not clear whether the mechanism operating is arrow number 1 or arrow number 2*a*. A more serious issue is the difficulty in classifying entities. Take an email transaction for example. Transactions are invariably classified at the micro-level of analysis. But it is not clear whether one should classify a broadcast email at the group level (macro-entity) or at the level of transaction (micro-entity.) Furthermore, when a broadcast email is sent, it is associated with several ties. But ties are at a higher level of analysis, so what does it mean to say that a micro-level entity (email) is associated with multiple meso-level entities (ties?) Is this like saying that individuals (micro-level) are influenced by multiple norms (macro-level)? I think that the best way to use the diagram is to consider it as a general guide for the principles of micro-foundations, rather than to stick to every arrow religiously.

A modest proposal

In one of the closing scenes of the film Barton Fink by the Coen Brothers, Charlie, a murderer who feigns to be a door-to-door insurance salesmen ('you might say I sell peace of mind,') gives a rectangular parcel to Barton Fink, a scriptwriter struggling with a writer's block. We never find out for certain what is in the parcel, but there are indications that it contains the severed head of one of Charlie's victims. Upon giving Barton the parcel, Charlie mutters: 'Funny, huh, when everything that's important to a guy, everything he wants to keep from a lifetime - when he can fit it into a little box like that. I guess . . . I guess it's kind of pathetic.'

Charlie ponders for a moment over the micro-macro mystery. He is holding a box that contains a head of an individual with a vast array of identities, experiences, memories, thoughts and emotions, a product of an entire lifetime, numerous contexts and situations which have all left their marks in an object that fits into a surprisingly small parcel. This individual is what Dennis H Wrong (1961) calls 'the socialized man.' If Wrong is right saying that the individual is in some sense the product of so many macro-entities, it is almost tempting to allow the socialized man to enter Coleman's framework in the upper left corner, where macro entities dwell. But that corner is reserved for anything macro except individuals. All individuals belong to the micro-level, at the bottom of the diagram.

This raises a puzzle: beyond being able to fit his severed head into a small parcel, in what sense is the 'socialized man' a micro entity? How is he different from other macro entities, such as the price of a product at equilibrium, norms or institutions? There is no question that these two types of entities (individuals and the prices at equilibrium) influence one another, but in what sense is one micro and the other macro? Human interaction and transactions are typically considered to be micro events. Is the action of sending an email to multiple recipients a micro affair or a macro one? Some types of emails, unsolicited spam messages for example, are sent to thousands, perhaps hundreds of thousands of people - are these micro or macro entities? Where exactly do they fit in the Coleman diagram?

Stated more generally, what do we actually mean when we use the terms micro and macro? There's an intuitive answer, that the macro is large and the micro is small. But when conjuring different micro and macro entities, many of them are abstract and their size is not one of their properties, let alone a feature that distinguishes between them. Another answer is that every macro entity is associated with multiple micro entities. That is of course true, but the reverse is also true - every micro entity is associated with multiple macro entities. Think of the relationships between any two entities in the empirical chapters, they all seem to be of a many-to-many type: every individual is associated with multiple ties and vice versa, every email is associated with multiple individuals and vice versa, every tie is associated with multiple emails and vice versa.

Even the relationship between networks and individuals is of a many-to-many character. Obviously, every network is associated with multiple individuals, but the reverse is also true. A single email network can be mapped into different sub-networks of activity, private emails and broadcast emails produce networks with strikingly different structural properties, both existing in parallel on top of the same group of people. This image is analogous to the distinction between anatomic and functional networks known(Shalizi *et al.*, 2006) from the study of neural networks, where one anatomic network of neuron-fibers can be mapped into multiple task-related functional networks (see the discussion in section 4.2). The same individual could be central in one such functional-network and marginal in the other, so the relation between individuals and networks is many-to-many and not one-to-many.

If we dare to suggest a modest modification to the Coleman diagram, it would restore the empirical intuition that hierarchical structures of one-to-many relationships are not as common as network structures of many-to-many. This involves a sense of symmetry between individuals and any other entities one might think about (see figure 6.2).

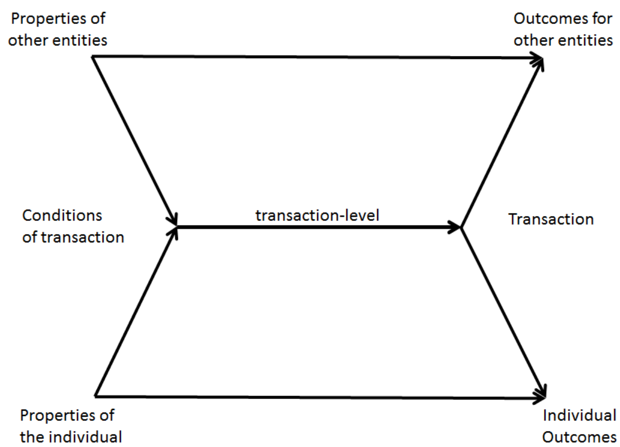


Figure 6.2: The Coleman butterfly - Attempting to account for a symmetry between the individual and the group

In this figure, individuals and other entities are each a product of many different elements. When they interact, only a subset of their properties play a role in the interaction, some of these properties changing as a consequence, possibly resulting in lasting cumulative effects that can carry on to other situations. Recall the example of Romeo and Juliet from section 2.2. They are both affiliated with rival families, but they are also affiliated with the young, those who are prone to fall in love. Each of them is a 'macro' product of multiple networks, but in the context of a Capulet's masquerade ball, certain dimensions become more salient and others fade into the background, allowing for the tragic story to unfold. Consider another example, namely the notion of faultlines in groups (Lau & Murnighan, 1998). Group faultlines are lines that may split a group into subgroups based on the different affiliations of each individual. For example, a group could divide into subgroups based on age, sex or personal values. Thus different tasks or contexts could increase the potential for friction and subgrouping along one of the many possible faultlines. Talk about affirmative-action may generate racial tension, retirement or pension issues may activate faultlines based on age, and a discussion about a perceived glass-ceiling in the organization may activate sex-related divisions, etc. Moreover, in some groups faultlines can align, for example when a group includes five young, white male interns and five senior black females, the group's faultline becomes 'stronger,' since many topics can bring to a confrontation, dividing the group along the same lines.

The theoretical consequences of this proposal must be thought through more carefully, but I think it retains the most important element of Coleman's diagram and the essence of micro-foundations, namely the rejection of Durkheimian style, macro-to-macro transitions that do not involve transactions or interactions. This proposed extension does not amount to the adoption of holism. But it does introduce the possibility of a symmetry between entities, placing them all on the same plane and replacing the one-to-many relationship that is typical to micro-foundations with a many-to-many one that, I think, is a closer approximation to the empirical world.

6.5 The Rebirth of Social Physics

> 'It is generally in these ill demarcated domains, that the urgent problem lies'
>
> Marcel Mauss (1973)

> No one descends with such fury and in so great a number as a pack of hungry physicists, adrenalized by the scent of a new problem.
>
> Duncan Watts (2004b, p 62)

> The idea that science and philosophy are different disciplines meant to complement each other ... arouses the desire and also imposes on us the duty to proceed to a confrontation.
>
> Henry Bergson (in the context of a debate with Albert Einstein, quoted in Canales, 2010, p 183)

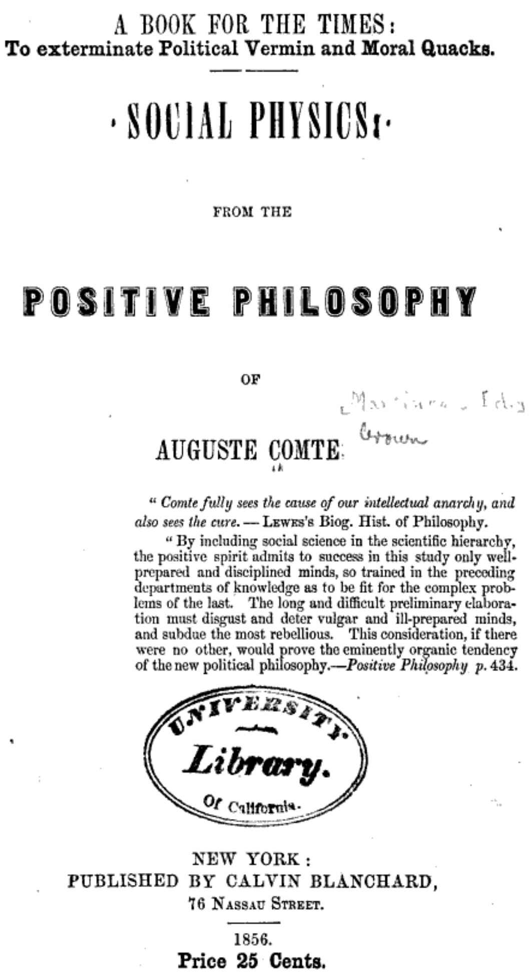
The dissertation demonstrated some of the mechanisms that are involved in the co-evolution between micro-level email transactions and meso- and macro- level ties and networks, broadly defined. This concluding chapter began with a reply to a common concern, about the possibility to infer network level constructs from transaction level data.

The reply depends on the way one wishes to define the social tie, and its relation to social transactions. If one wishes to define the tie in terms of the transactions (in the tradition of Tarde, Homans and others,) there is no problem and no need for further data. If the tie is contingent on the transactions, we can treat it as a latent variable and identify some of its properties through its interaction with the manifest variables of email transactions.

After addressing this concern the findings were presented, organized according to the different mechanisms in the Coleman diagram. The most interesting findings in Chapter 4 are associated with the aggregation link from micro-to-macro on the diagram, and they involved the process by which email based interactions at the micro-level give rise to network patterns at the macro level. The most interesting findings in Chapter 5 are associated with the (macro- and meso-) conditions for action. The model was designed to disentangle between micro-level conditions (the incoming email and its properties) and macro-level ones (general etiquette for email communication, sender/recipient's role in the organization etc.) The third section reviewed the contributions critically, demonstrating some of the empirical and conceptual elements that distinguish between network models based on email mediated transactions and social models based on surveys and questionnaires. This was followed by a section discussing some of the methodological contributions and some general remarks about the theoretical framework of micro-foundations, the way the framework facilitates thinking about and communicating the empirical investigation and some of the puzzles it raises.

As the academic and systematic study of digitially mediated transactions is entering its second decade (Lazer *et al.*, 2009b), it is worth thinking of its consequences, in the context of social network studies. Granted, digitally mediated transaction datasets have triggered a heated controversy among social scientists, particularly among those who specialize in the field of social networks. For one thing, the large datasets attracted physicists and mathematicians who are now publishing surprisingly influential papers (Lazer *et al.*, 2009a), surprising given that some of them are rightly accused of claiming ownership on discoveries that were already known for decades (Scott, 2011; Freeman, 2011).

Others maintain that traditional network scholars are reluctant to adopt digitally mediated transaction datasets for substantive research because of the data's 'theoretical and empirical ambiguity,' and because it is unclear how these datasets stand in relation to traditional network data (Quintane & Kleinbaum, 2011). But there is another, slightly more optimistic view regarding these new developments. Latour, Jensen, Venturini, Grauwin & Boullier (2012) suggest that we may be witnessing a gestalt switch (in the Kuhnian sense) in the field, a shift from a Durkheimian paradigm back to a Tardian one. According to these authors, the Durkheimian strategy organizes the world into layers of analysis, each two layers linked in a one-to-many relationship. Whether the macro-micro link is contingent or definitional, system effects such as Condorcet's paradox and the prisoner's dilemma (see section 6.4.1) give the sense that macro- and micro-level entities are situated on different ontological plains, entities that are different in kind. On the other hand, the Tardian perspective, according to Latour *et al.* (2012), is a departure from this hierarchical view, in that it adopts a symmetric approach,


A BOOK FOR THE TIMES:
To exterminate Political Vermin and Moral Quacks.

SOCIAL PHYSICS

FROM THE

POSITIVE PHILOSOPHY

OF

AUGUSTE COMTE

" *Comte fully sees the cause of our intellectual anarchy, and also sees the cure.* — LEWES'S Biog. Hist. of Philosophy.

" By including social science in the scientific hierarchy, the positive spirit admits to success in this study only well-prepared and disciplined minds, so trained in the preceding departments of knowledge as to be fit for the complex problems of the last. The long and difficult preliminary elaboration must disgust and deter vulgar and ill-prepared minds, and subdue the most rebellious. This consideration, if there were no other, would prove the eminently organic tendency of the new political philosophy.—*Positive Philosophy* p. 434.

UNIVERSITY
Library.
OF California.

NEW YORK:
PUBLISHED BY CALVIN BLANCHARD,
76 NASSAU STREET.

1856.
Price 25 Cents.

Figure 6.3: The Cover of Auguste Comte's book 'Social Physics' - 'A book for the times - to exterminate political vermin and moral quacks'

bringing all entities, both individuals and 'social facts' onto one ontological plane, each type of entity associated with others in a many-to-many relationship, each changing by interacting with the others. It is too soon to tell whether this rather vague prognosis will be realized, but I think that it would be rather exciting if new types of data challenge us to consider more carefully the properties of transactions and their consequences, before aggregating them into networks.

This research was a quest for mechanisms that link immediate social actions with the structural properties of networks. It is inspired by a non-Durkheimian approach, one that replaces the 'oversocialized man' (Wrong, 1961), with a sensitivity to momentary decisions, fleeting events and impulsive (trans)actions. It is driven by the notion that the question 'to reply or not to reply' is of greater consequence than the question 'to be or not to be.' Many great men and women accompanied us along this journey, reminding us that the controversies are far from contemporary and most probably far from over:[1] Is the macro-micro link definitional (Homans, Weber, Monge & Contractor (2003), Margaret Thatcher and the reductionists) or contingent (Simmel, Mauss, Coleman, Lévi-Strauss and the structuralists)? Are macro-level entities and micro-level ones different in kind (Durkheim, Nietzsche, structuralists and those studying system effects) or not (Gabriel Tarde, Bruno Latour)? What is the nature of social action (Simmel, Parsons, Mauss, Weber, Coleman)? Is the individual really a micro entity (Romeo, Juliet and the Coen brothers)?

Another common thread throughout this work is the way social ties interact with transactions, acts of communication and exchanges, meetings and social events. This issue is crucial to those who construct models of social networks from digitally mediated transaction datasets. It also problematizes the role of the technological artefact that mediates communication transactions, for its properties shape the way transactions and network structures interact. Surprisingly perhaps, even this question of action and ties is not limited to our time. The dissertation opened with a poem describing the process of tie formation, beginning with innocent and playful acts of courtship, climbing the hysteresis curve with passion and swept away by a storm of emotion. The same process, only in reverse, is described by Virginia Wolf almost a century ago. Following a rather dull

[1] These names are just a small and non-representative sample, far from exhaustive.

social gathering, the acquaintances of lady Bruton part, each going their own way, and as the memory of the event fades, the emotions associated with the social ties dissipate in a manner not unlike the one described in figure 3.2. In contrast to the process of tie formation, here there is neither joy nor anguish, neither benefits nor costs, neither animal spirits nor social facts, but only the dwindling of spirit, the drifting away in silence, a sense of lethargy, resignation and indifference.

> And Lady Bruton went ponderously, majestically, up to her room, lay, one arm extended, on the sofa. She sighed, she snored ... And they went further and further from her, being attached to her by a thin thread (since they had lunched with her) which would stretch and stretch, get thinner and thinner as they walked across London; as if one's friends were attached to one's body, after lunching with them, by a thin thread, which (as she dozed there) became hazy with the sounds of bells striking the hour or ringing to service, as a single spider's thread is blotted with rain-drops, and, burdened, sags down. So she slept [...] lying on the sofa, let the thread snap; snored.
>
> Mrs. Dalloway, (Woolf, 2012)

References

ABELL, P. (2003a). On the prospects for a unified social science: economics and sociology. *Socio-Economic Review*, **1**, 1–26.

ABELL, P. (2003b). The role of rational choice and narrative action theories in sociological theory. *Revue francaise de sociologie*, **Volume 44**, 255–273.

ABELL, P., FELIN, T. & FOSS, N. (2008). Building micro-foundations for the routines, capabilities, and performance links. *Managerial and Decision Economics*, **29**, 489–502.

ABELL, P., FELIN, T. & FOSS, N. (2010). Causal and Constitutive Relations, and the Squaring of Coleman's Diagram: Reply to Vromen. *Erkenntnis*, **73**.

ADAMIC, L. & ADAR, E. (2005). How to search a social network. *Social Networks*, **27**, 187–203.

ADAMIC, L.A. & HUBERMAN, B.A. (2002). Zipf's law and the internet. *Glottometrics*, **3**, 143–150.

ADAMIC, L.A. *et al.* (2000). Power-law distribution of the world wide web. *Science*, **287**, 2115–2115.

ADAMS, J. (2010). Distant friends, close strangers? inferring friendships from behavior. *Proceedings of the National Academy of Sciences*, **107**.

ADAR, E. & ADAMIC, L.A. (2005). Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, 207–214, IEEE Computer Society, Washington, DC, USA.

ANDERSON, B.S., BUTTS, C. & CARLEY, K. (1999). The interaction of size and density with graph-level indices. *Social Networks*, **21**, 239–267.

ANDERSON, P.W. (1972). More is different. *Science*, **177**, 393–396.

BACHELARD, G. (2000). *The Dialectic of Duration (Philosophy of Science)*. Clinamen Press Ltd, 1st edn.

BALKWELL, J.W. (1991). From expectations to behavior: An improved postulate for expectation states theory. *American Sociological Review*, **56**, 355–369.

BARABÁSI, A.L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, **435**, 207–211.

BARABASI, A.L. (2009). Scale-free networks: A decade and beyond. *Science*, **325**, 412–413.

BARABÁSI, A.L. (2010). *Bursts: the hidden pattern behind everything we do*. Dutton.

BARABÁSI, A.L. & ALBERT, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509.

BARNES, J.A. (1954). Class and committees in a Norwegian island parish. *Human Relations*, **7**, 39–58.

BARNES, J.A. (2008). *Humping my drum: a memoir*. Lulu.com.

BARNEY, J.B. (2001). Is the resource-based 'view' a useful perspective for strategic management research? yes. *Academy of management review*, **26**, 41–56.

BARRAT, A., BARTHELEMY, M., PASTOR-SATORRAS, R. & VESPIGNANI, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 3747–3752.

BARRAT, A., BARTHELEMY, M. & VESPIGNANI, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge University Press, 1st edn.

BAZERMAN, M.H. & MOORE, D.A. (2008). *Judgment in managerial decision making*. Wiley.

BEARMAN, P. & PARIGI, P. (2004). Cloning headless frogs and other important matters: Conversation topics and network structure. *Social Forces*, **83**, 535–557.

BEARMAN, P.S., MOODY, J. & STOVEL, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, **110**, 44–91.

BELLOTTI, E. (2008). What are friends for? elective communities of single people. *Social Networks*, **30**, 318–329.

BERLIN, I. (2013). *The Hedgehog and the Fox: An Essay on Tolstoy's View of History*. Princeton University Press, 2nd edn.

BERNARD, H.R., KILLWORTH, P.D. & SAILER, L. (1980). Informant accuracy in social network data iv: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks*, **2**, 191–218.

BERNARD, H.R., KILLWORTH, P. & SAILER, L. (1981). Summary of research on informant accuracy in network data and the reverse small world problem. *Connections*, **4**, 11–25.

BERNARD, H.R., KILLWORTH, P., KRONENFELD, D. & SAILER, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annu. Rev. Anthropol.*, **13**, 495–517.

BERNARD, H.R., JOHNSEN, E.C., KILLWORTH, P.D., MCCARTY, C., SHELLEY, G.A. & ROBINSON, S. (1990). Comparing four different methods for measuring personal social networks. *Social Networks*, **12**, 179–215.

BLAU, P.M. (1972). Interdependence and hierarchy in organizations. *Social Science Research*, **1**, 1–24.

BLOOR, D. (1982). Durkheim and Mauss Revisited: Classification and the Sociology of Knowledge in Special Issue. The Sociology of Knowledge: Cases and Debates. *Studies in History and Philosophy of Science London*, **13**, 267–297.

BOND, C.F., JONES, R.L. & WEINTRAUB, D.L. (1985). On the unconstrained recall of acquaintances: a samplingâĂŢtraversal model. *Journal of personality and social psychology*, **49**, 327–337.

BORGATTI, S.P. & EVERETT, M.G. (1997). Network analysis of 2-mode data. *Social Networks*, **19**, 243–269.

BORGATTI, S.P., EVERETT, M.G. & FREEMAN, L.C. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

BORGATTI, S.P., MEHRA, A., BRASS, D.J. & LABIANCA, G. (2009). Network analysis in the social sciences. *science*, **323**, 892–892.

BORGE-HOLTHOEFER, J., BAÑOS, R.A., GONZÁLEZ-BAILÓN, S. & MORENO, Y. (2013). Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, **1**, 3–24.

BÖRNER, K., DALL'ASTA, L., KE, W. & VESPIGNANI, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, **10**, 57–67.

BOUDON, R. (1986). *Theories of social change: a critical appraisal*. University of California Press, Berkeley and Los-Angeles, California, USA.

BOXER, D. (1993). Social distance and speech behavior: The case of indirect complaints. *Journal of Pragmatics*, **19**, 103–125.

BRAHA, D. & BAR-YAM, Y. (2006). From centrality to temporary fame: Dynamic centrality in complex networks. *Complexity*, **12**, 59–63.

BRANDES, U., LERNER, J. & SNIJDERS, T.A.B. (2009). Networks evolving step by step: Statistical analysis of dyadic event data. In *International Conference on Advances in Social Networks Analysis and Mining*, ASONAM, Athens, Greece.

BRANDT, A.M. (1985). *No magic bullet: a social history of venereal disease in the United States since 1880*. Oxford University Press, New-York, New-York, USA.

BURT, R.S. (2001). Structural holes versus network closure as social capital. In R.S. Burt, K. Cook & L. N, eds., *Social Capital: Theory and Research*, Aldine Transaction.

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

BUTTS, C.T. (2009). Revisiting the foundations of network analysis. *Science*, **325**, 414–416.

BYRON, K. (2008). Carrying too heavy a load? The communication and miscommunication of emotion by email. *The Academy of Management Review*, **33**, 309–327.

CALVINO, I. (1978). *Invisible cities*. Harcourt Brace Jovanovich, San-Diego, CA.

CANALES, J. (2010). *A tenth of a second: A history*. University of Chicago Press.

CARENINI, G., NG, R.T. & ZHOU, X. (2005). Scalable discovery of hidden emails from large folders. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 544–549, ACM.

CARLIN, W. & SOSKICE, D. (2006). *Macroeconomics: Imperfections, Institutions and Policies*. Oxford University Press, USA.

CARRASCO, J.A., HOGAN, B., WEILMAN, B. & MILLER, E.J. (2008). Collecting social network data to study social activity-travel behavior: an egocentric approach. *Environment and Planning B: Planning and Design*, **35**, 961–980.

CAULKINS, D. (1981). The Norwegian connection: Eilert Sundt and the idea of social networks in 19^{th} century ethnology. *Connections*, **4**.

CHAPANOND, A., KRISHNAMOORTHY, M. & YENER, B. (2005). Graph theoretic and spectral analysis of enron email data. *Computational & Mathematical Organization Theory*, **11**, 265–281.

COASE, R.H. (1937). The nature of the firm. *Economica*, **4**, 386–405.

COFF, R.W. (1999). When competitive advantage doesn't lead to performance: The resource-based view and stakeholder bargaining power. *Organization Science*, **10**, 119–133.

COLANDER, D.C. (1993). The macrofoundations of micro. *Eastern Economic Journal*, **19**, 447–457.

COLEMAN, J. (1988). Social capital in the creation of human capital. *The American Journal of Sociology*, **94 [Supplement]**, S95–S120.

COLEMAN, J., KATZ, E. & MENZEL, H. (1957). The diffusion of an innovation among physicians. *Sociometry*, **20**, 253–270.

COLEMAN, J.S. (1990). *Foundations of social theory*. Belknap Press of Harvard University Press, Cambridge, MA.

COLLINGSWORTH, B. & MENEZES, R. (2009). Identification of social tension in organizational networks. In *Complex Networks*, 209–223, Springer.

COLLINS, R. (2004). *Interaction ritual chains*. Princeton University Press.

COURGEAU, D. (2003). *Methodology and Epistemology of Multilevel Analysis: Approaches from Different Social Sciences*. Springer.

CRANOR, L.F. & LAMACCHIA, B.A. (1998). Spam! *Communications of the ACM*, **41**, 74–83.

CRAWFORD, A.B. (1982). Corporate electronic mail-A communication-intensive application of information technology. *MIS quarterly*, 1–13.

DABBISH, L. & KRAUT, R. (2006). Email overload at work: An analysis of factors associated with email strain. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 431–440.

DABBISH, L.A., KRAUT, R.E., FUSSELL, S. & KIESLER, S. (2005). Understanding email use: predicting action on a message. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 691–700, ACM, Portland, Oregon, USA.

DAFT, R.L. & LENGEL, R.. (1986). Organizational information requirements, media richness and structural design. *Management Science*, **32**, 554–571.

DANTO, A.C. (1973). *Analytical philosophy of action*. Cambridge [Eng.]: University Press.

DAVIS, A., GARDNER, B.B. & GARDNER, M.R. (1941). *Deep south*. University of Chicago Press Chicago.

DAVIS, J., HOSSAIN, L. & MURSHED, S.H. (2007). Social network analysis and organizational disintegration: The case of enron corporation. In *ICIS 2007 Proceedings*.

DE CHOUDHURY, M., MASON, W.A., HOFMAN, J.M. & WATTS, D.J. (2010). Inferring relevant social networks from interpersonal communication. In *WWW '10: Proceedings of the 19th international conference on World wide web.*.

DE NOOY, W. (2011). Networks of action and events over time. a multilevel discrete-time event history model for longitudinal network data. *Social Networks*, **33**.

DEMEULENAERE, P. (2011). *Analytical sociology and social mechanisms*. Cambridge University Press.

DIESNER, J. & CARLEY, K.M. (2005). Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*, Citeseer.

DIESNER, J., FRANTZ, T.L. & CARLEY, K.M. (2005). Communication networks from the enron email corpus 'it's always about the people. enron is no different'. *Computational & Mathematical Organization Theory*, **11**, 201–228.

DOREIAN, P. (2002). Event sequences as generators of social network evolution. *Social Networks*, **24**, 93–119.

DUIJN, M.A., SNIJDERS, T.A. & ZIJLSTRA, B.J. (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, **58**, 234–254.

DUNBAR, R.I. (1998). The social brain hypothesis. *brain*, **9**, 10.

DUNBAR, R.I. & SHULTZ, S. (2007). Understanding primate brain evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**, 649–658.

DURKHEIM, E. (1997 [1893]). *The division of labor in society*. The Free Press.

EBEL, H., MIELSCH, L.I. & BORNHOLDT, S. (2002). Scale-free topology of e-mail networks. *Phys. Rev. E*, **66**, 035103–035103.

ECKMANN, J.P., MOSES, E. & SERGI, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14333–14337.

ELSTER, J. (1976). A note on hysteresis in the social sciences. *Synthese*, **33**, 371–391.

ELSTER, J. (1989a). *The cement of society: A survey of social order*. Cambridge University Press, Cambridge UK.

ELSTER, J. (1989b). *Nuts and bolts for the social sciences*. Cambridge University Press.

ELSTER, J. (1989c). Social norms and economic theory. *The Journal of Economic Perspectives*, **3**, 99–117.

ELSTER, J. (1998). A plea for mechanisms. In P. Hedström & R. Swedberg, eds., *Social mechanisms: An analytical approach to social theory*, Cambridge University Press.

EMERSON, R.M. (1976). Social exchange theory. *Annual review of sociology*, **2**, 335–362.

ENGEL, O. (2011). Clusters, recipients and reciprocity: Extracting more value from email communication networks. *Procedia - Social and Behavioral Sciences*, **10**, 172–182.

EVELAND, J.D. & BIKSON, T.K. (1986). Evolving electronic communication networks: an empirical assessment. In *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*, 91–101, ACM, Austin, Texas.

FARARO, T., SKVORETZ, J. & KOSAKA, K. (1994). Advances in e-state structuralism: further studies in dominance structure formation. *Social Networks*, **16**, 233–265.

FARARO, T.J. & DOREIAN, P. (1984). Tripartite structural analysis: Generalizing the Breiger-Wilson formalism. *Social Networks*, **6**, 141–175.

FELD, S. (1981). The Focused Organization of Social Ties. *The American Journal of Sociology*, **86**, 1015–1035.

FELD, S.L. & ELMORE, R. (1982). Patterns of sociometric choices: Transitivity reconsidered. *Social Psychology Quarterly*, **45**, 77–85.

FELIN, T. & FOSS, N.J. (2006). Thoughts on a micro-foundations project for strategic management and organizational analysis. `http://hdl.handle.net/10398/7470`, accessed: 26/03/2013.

FOSS, K. & FOSS, N.J. (2005). Resources and transaction costs: how property rights economics furthers the resource-based view. *Strategic Management Journal*, **26**, 541–553.

FOSS, N.J. (2010). Micro-foundations for management research: what, why, and whither? *Cuadernos de Economía y Dirección de la Empresa*, **13**, 11–34.

FRANK, O. & CARRINGTON, P.J. (2007). Estimation of offending and co-offending using available data with model support. *Journal of Mathematical Sociology*, **31**, 1–46.

FRANTZ, T.L. & CARLEY, K.M. (2008). Transforming raw-email data into social-network information. In *Intelligence and Security Informatics*, 413–420, Springer.

FREEMAN, L.C. (1982). Centered graphs and the structure of ego networks. *Mathematical Social Sciences*, **3**, 291–304.

FREEMAN, L.C. (2011). The development of Social Network Analysis - with an Emphasis on Recent Events. In J. Scott & P.J. Carrington, eds., *The SAGE handbook of social network analysis*, SAGE Publications Limited.

FREEMAN, L.C., ROMNEY, A.K. & FREEMAN, S.C. (1987). Cognitive structure and informant accuracy. *American anthropologist*, **89**, 310–325.

FRIEDKIN, N. (1980). A test of structural features of granovetters strength of weak ties theory. *Social Networks*, **2**, 411–422.

FRIEDKIN, N.E. (2004). Social cohesion. *Annual Review of Sociology*, 409–425.

FRIEDKIN, N.E. & JOHNSEN, E.C. (1999). Social influence networks and opinion change. *Advances in Group Processes*, **16**, 1–29.

GAVETTI, G. (2005). Cognition and hierarchy: Rethinking the microfoundations of capabilities development. *Organization Science*, **16**, 599–617.

GEHRLEIN, W.V. (1983). Condorcet's paradox. *Theory and Decision*, **15**, 161–197.

GIBSON, D.R. (2000). Seizing the moment: The problem of conversational agency. *Sociological Theory*, **18**, 368–382.

GIBSON, D.R. (2005). Taking turns and talking ties: Networks and conversational interaction. *American Journal of Sociology*, **110**, 1561–1597.

GIBSON, D.R. (2008). How the outside gets in: Modeling conversational permeation. *Sociology*, **34**, 359–384.

GIDDENS, A. (1979). *Central problems in social theory: Action, structure, and contradiction in social analysis*, vol. 241. University of California press.

GLOOR, P.A., LAUBACHER, R., DYNES, S.B.C. & ZHAO, Y. (2003). Visualization of communication patterns in collaborative innovation networks - analysis of some w3c working groups. In *Proceedings of the twelfth international conference on Information and knowledge management*, 56–60, ACM, New Orleans, LA, USA.

GOFFMAN, E. (1967). *Interaction Ritual - Essays on Face-to-Face Behavior*. Doubleday.

GOFFMAN, E. (1976). Replies and responses. *Language in Society*, **5**, 257–313.

GOFFMAN, E. (1983). The interaction order. *American Sociological Review*, **48**, 1–17.

GOLDSTEIN, H. (2011). *Multilevel statistical models*, vol. 922. Wiley.

GRANNIS, R. (2010). Six degrees of 'who cares?'. *American Journal of Sociology*, **115**, 991–1017.

GRANOVETTER, M. (1973). The strength of weak ties. *The American Journal of Sociology*, **78**, 1360–1380.

GUILLAUME, J.L. & LATAPY, M. (2004). Bipartite structure of all complex networks. *Information Processing Letters*, **90**, 215–221.

GUIMERÁ, R., DANON, L., DÍAZ-GUILERA, A., GIRALT, F. & ARENAS, A. (2003). Self-similar community structure in a network of human interactions. *Physical Review E*, **68**, 65103.

HACKING, I. (1990). *The taming of chance*. Cambridge University Press.

HACKING, I. (1995). The looping effects of human kinds. *Causal cognition: A multidisciplinary approach*, 351–83.

HARTUP, W.W. & STEVENS, N. (1997). Friendships and adaptation in the life course. *Psychological bulletin*, **121**, 355.

HEDSTRÖM, P. (2005). *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge University Press.

HEDSTRÖM, P. & BEARMAN, P. (2009). What is Analytical Sociology all about? An introductory essay. In *The Oxford Handbook of Analytical Sociology*, Oxford University Press.

HEDSTRÖM, P. & SWEDBERG, R. (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge University Press.

HICKS, J.W. (2002). The physiological and evolutionary significance of cardiovascular shunting patterns in reptiles. *Physiology*, **17**, 241–245.

HOGAN, B., CARRASCO, J.A. & WELLMAN, B. (2007). Visualizing personal networks: Working with participant-aided sociograms. *Field Methods*, **19**, 116–144.

HOLLAND, P.W. & LEINHARDT, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology*, 492–513.

HOLLAND, P.W. & LEINHARDT, S. (1977). Social structure as a network process. *Zeitschrift für Soziologie*, **6**, 386–402.

HOLME, P. & SARAMÄKI, J. (2011). Temporal networks. *arXiv:1108.1780*.

HOMANS, G.C. (1951). *The human group*. Transaction Publishers, New-Brunswick,NJ, 8th edn.

IBN HAZM ALI IBN AHMAD (1981). *The ring of the dove*. AMS Press Inc, New-York.

JACKMAN, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons, Ltd.

JAVA, A., SONG, X., FININ, T. & TSENG, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, 56 – 65, ACM, New York, NY, USA.

JERVIS, R. (1997). *System effects: Complexity in political and social life*. Princeton University Press.

JOHNSON, C. (1994). Gender, legitimate authority, and leader-subordinate conversations. *American Sociological Review*, **59**, 122–135.

KAPLAN, S. & HENDERSON, R. (2005). Inertia and incentives: Bridging organizational economics and organizational theory. *Organization Science*, **16**, 509–521.

KARAU, S.J. & WILLIAMS, K.D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, **65**, 681–681.

KARSAI, M., KIVELÄ, M., PAN, R., KASKI, K., KERTÉSZ, J., BARABÁSI, A.L. & SARAMÄKI, J. (2011). Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, **83**, 025102.

KARSENTI, B. (2010). Imitation: Returning to the Tarde-Durkheim debate. In *The Social after Gabriel Tarde: Debates and Assessments*, Routledge, Oxon, UK.

KAY, J. (2012). *Obliquity: Why Our Goals Are Best Achieved Indirectly*. Penguin Books, reprint edn.

KEMPE, D., KLEINBERG, J. & KUMAR, A. (2000). Connectivity and inference problems for temporal networks. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 504–513, ACM.

KILLWORTH, P.D. & BERNARD, H.R. (1980). Informant accuracy in social network data iii: A comparison of triadic structure in behavioral and cognitive data. *Social Networks*, **2**, 19–46.

KLEINBAUM, A.M., STUART, T. & TUSHMAN, M. (2008). *Communication (and coordination?) in a modern, complex organization*. Harvard Business School Boston, MA.

KLIMT, B. & YANG, Y. (2004). The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, 217–226, Springer.

KOSKINEN, J. & EDLING, C. (2012). Modelling the evolution of a bipartite network - Peer referral in interlocking directorates. *Social Networks*, **34**, 309 – 322.

KOSSINETS, G. & WATTS, D.J. (2006). Empirical analysis of an evolving social network. *Science*, **311**, 88–90.

KOSTAKOS, V. (2009). Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, **388**, 1007–1023.

KOVANEN, L., SARAMÄKI, J. & KASKI, K. (2010). Reciprocity of mobile phone calls. *Dynamics of Socio-Economic Systems*, –.

KOVANEN, L., KASKI, K., KERTÉSZ, J. & SARAMÄKI, J. (2013). Temporal motifs reveal homophily, gender-specific patterns and group talk in mobile communication networks. *arXiv preprint arXiv:1302.2563*.

Krackardt, D. (1987). QAP partialling as a test of spuriousness. *Social networks*, **9**, 171–186.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, **9**, 109–134.

Krieger, N. (1994). Epidemiology and the web of causation: has anyone seen the spider? *Social science & medicine*, **39**, 887–903.

Krings, G., Karsai, M., Bernhardsson, S., Blondel, V.D. & Saramäki, J. (2012). Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, **1**, 1–16.

Kumar, R., Novak, J., Raghavan, P. & Tomkins, A. (2005). On the Bursty Evolution of Blogspace. *World Wide Web*, **8**, 159âĂŞ178.

Kumar, R., Tomkins, A. & Vee, E. (2008). Connectivity structure of bipartite graphs via the knc-plot. In *Proceedings of the international conference on Web search and web data mining*, 129–138, ACM.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, 591âĂŞ600, ACM, New York, NY, USA.

Latour, B. (1988). Irréductions. In *The Pasteurisation of France*, Harvard, Cambridge, MA, USA.

Latour, B., Jensen, P., Venturini, T., Grauwin, S. & Boullier, D. (2012). t̀he whole is always smaller than its part´-a digital test of gabriel tardes' monads. *The British journal of sociology*, **63**, 590–615.

Lau, D.C. & Murnighan, J.K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review*, **23**, 325–340.

Lazarsfeld, P.F. & Menzel (1961). On the relation between individual and collective properties. *Complex organizations*, **1**.

LAZARSFELD, P.F. & MERTON, R.K. (1954). Friendship as a social process: A substantive and methodological analysis. In *Freedom and control in modern society*, 18–66, Van Nostrand, New York.

LAZEGA, E., JOURDA, M.T., MOUNIER, L. & STOFER, R. (2008). Catching up with big fish in the big pond? multi-level network analysis through linked design. *Social Networks*, **30**, 159–176.

LAZER, D., MERGEL, I. & FRIEDMAN, A. (2009a). Co-Citation of Prominent Social Network Articles in Sociology Journals: The Evolving Canon. *Connections*, **29**.

LAZER, D., PENTLAND, A., ADAMIC, L., ARAL, S., BARABASI, A.L., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., GUTMANN, M., JEBARA, T., KING, G., MACY, M., ROY, D. & VAN ALSTYNE, M. (2009b). Computational social science. *Science*, **323**, 721–723.

LEIJONHUFVUD, A. (1967). Keynes and the Keynesians: A suggested interpretation. *The American Economic Review*, **57**, 401–410.

LIBEN-NOWELL, D. & KLEINBERG, J. (2008). Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, **105**, 4633–4638.

LICOPPE, C. & SMOREDA, Z. (2005). Are social networks technologically embedded?: How networks are changing today with changes in communication technology. *Social Networks*, **27**, 317–335.

LIPPMAN, S.A. & RUMELT, R.P. (2003a). A bargaining perspective on resource advantage. *Strategic Management Journal*, **24**, 1069–1086.

LIPPMAN, S.A. & RUMELT, R.P. (2003b). The payments perspective: microfoundations of resource analysis. *Strategic management journal*, **24**, 903–927.

LIST, C. & PETTIT, P. (2002). Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, **18**, 89–110.

LOTKA, A.J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences. Vol*, **16**, 317–323.

LUHMANN, N. (1998). *Love as Passion: The Codification of Intimacy*. Stanford University Press, 1st edn.

MAHONEY, J. (2000). Path dependence in historical sociology. *Theory and society*, **29**, 507–548.

MAKADOK, R. & BARNEY, J.B. (2001). Strategic factor market intelligence: An application of information economics to strategy formulation and competitor intelligence. *Management Science*, **47**, 1621–1638.

MANDELBROT, B. (2005). Response to the Edge Annual Question, 2005. `http://www.edge.org/documents/mandelbrot2010/mandelbrot2010_index.html`, accessed: 26/03/2013.

MANSKI, C.F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, **60**, 531–542.

MANSKI, C.F. (1995). *Identification problems in the social sciences*. Harvard University Press.

MARSDEN, P.V. (1990). Network data and measurement. *Annual Review of Sociology*, **16**, 435–463.

MARSDEN, P.V. (2005). Recent developments in network measurement. In *Models and methods in social network analysis*, Cambridge University Press.

MARSDEN, P.V. (2011). Survey Methods for Network Data. In J. Scott & P.J. Carrington, eds., *The SAGE handbook of social network analysis*, SAGE Publications Limited.

MAUSS, M. (1973). Techniques of the body. *Economy and society*, **2**, 70–88.

MAUSS, M. (2000). *The gift: Forms and functions of exchange in archaic societies*. W. W. Norton and Company, New-York, New-York, USA.

McPherson, M., Smith-Lovin, L. & Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.*, **27**, 415–444.

Merton, R.K. (1936). The unanticipated consequences of purposive social action. *American sociological review*, **1**, 894–904.

Miller, R.W. (1978). Methodological individualism and social explanation. *Philosophy of Science*, 387–414.

Miritello, G., Moro, E. & Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E*, **83**, 045102.

MISQ (2013). Management information systems quarterly: About page. `http://www.misq.org/about/`, accessed: 26/03/2013.

Mitchell, J.C. (1969). The concept and use of social networks. In *Social Networks in Urban Relations: Analyses of Personal Relationships in Central African Towns*, 1âĂŞ–50, Manchester University Press.

Mitchell, T. (2009). Carbon democracy. *Economy and Society*, **38**, 399–432.

Mizruchi, M.S. (1996). What do interlocks do? an analysis, critique, and assessment of research on interlocking directorates. *Annual review of sociology*, 271–298.

Monge, P.R. & Contractor, N.S. (2003). *Theories of communication networks*. Oxford University Press, New-York, USA.

Moody, J. (2002). The importance of relationship timing for diffusion. *Social Forces*, **81**, 25–56.

Moody, J. (2009). Network dynamics. In *The Oxford Handbook of Analytical Sociology*, Oxford University Press, Oxford, UK.

Moody, J., McFarland, D. & Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology*, **110**, 1206–41.

Moreno, J.L. & Jennings, H.H. (1938). Statistics of social configurations. *Sociometry*, 342–374.

MORGAN, M.S. & MORRISON, M. (1999). *Models as mediators*. Cambridge University Press.

MORRIS, M. & KRETZSCHMAR, M. (1995). Concurrent partnerships and transmission dynamics in networks. *Social Networks*, **17**, 299–318.

NADEL, S.F. (2007). On Social Structure. In R.A. Manners & D. Kaplan, eds., *Theory in Anthropology: A Sourcebook*, Routledge & Kegan Paul.

NEWBY, G.B., GREENBERG, J. & JONES, P. (2003). Open source software development and lotka's law: Bibliometric patterns in programming. *Journal of the American Society for Information Science and Technology*, **54**, 169–178.

NEWMAN, M.E. & PARK, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, **68**, 036122.

NEWMAN, M.E.J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, **64**.

NEWMAN, M.E.J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, **64**.

NEWMAN, M.E.J. (2003). The structure and function of complex networks. *SIAM Review*, **45**, 167–256.

NEWMAN, M.E.J. (2005). Power laws, pareto distributions and zipfś law. *Contemporary physics*, **46**, 323–351.

OAKESHOTT, M. (1991). On human conduct.

ONNELA, J.P., SARAMÄKI, J., HYVÖNEN, J., SZABÓ, G., DE MENEZES, M.A., KASKI, K., BARABÁSI, A.L. & KERTÉSZ, J. (2007a). Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, **9**, 179.

ONNELA, J.P., SARAMÄKI, J., HYVONEN, J., SZABO, G., LAZER, D., KASKI, K., KERTESZ, J. & BARABASI, A.L. (2007b). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, **104**, 7332–7336.

OPSAHL, T. (2011). Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*.

OPSAHL, T. & PANZARASA, P. (2009). Clustering in weighted networks. *Social Networks*, **31**, 155–163.

ORLIKOWSKI, W.J. & IACONO, S. (2001). Research commentary: desperately seeking the it in it research âĂŞ a call to theorize the it artifact. *Information Systems Research*, **12**, 13–13.

PALLA, G., BARABÁSI, A.L. & VICSEK, T. (2007). Quantifying social group evolution. *Nature*, **446**, 664–664.

PAN, R.K. & SARAMÄKI, J. (2011). Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, **84**, 016105.

PASTOR-SATORRAS, R. & VESPIGNANI, A. (2010). Complex networks: Patterns of complexity. *Nature Physics*, **6**, 480–481.

PISANI, E. (2010). Has the internet changed science? *Prospect Magazine*, *November 17th 2010*.

PLUMMER, M. (2004). Jags: Just another gibbs sampler.

PODOLNY, J.M. (2001). Networks as the pipes and prisms of the market. *American Journal of Sociology*, **107**, 33–60.

PODOLNY, J.M. & BARON, J.N. (1997). Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review*, **62**, 673–693.

QUINTANE, E. & KLEINBAUM, A.M. (2011). Matter over mind? e-mail data and the measurement of social networks. *Connections*, **31**, 22 – 46.

RANSMAYR, C. (1990). *The Last World*. Grove Press, New-York, NY.

RAUB, W. & WEESIE, J. (1990). Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology*, 626–654.

RICE, R.E. & BAIR, J.H. (1984). New organizational media and productivity. *The new media*, 185–216.

RICOEUR, P. (1984). *Time and Narrative, Volume 1*. University of Chicago Press, Chicago,IL.

RIEKE, F., WARLAND, D., DE RUYTER VAN STEVENINCK, R. & BIALEK, W. (1997). *Exploring the neural code*. The MIT press.

RIOLO, C.S., KOOPMAN, J.S. & CHICK, S.E. (2001). Methods and measures for the description of epidemiologic contact networks. *Journal of Urban Health*, **78**, 446–457.

ROBINS, G., ELLIOTT, P. & PATTISON, P. (2001a). Network models for social selection processes. *Social Networks*, **23**, 1–30.

ROBINS, G., PATTISON, P. & ELLIOTT, P. (2001b). Network models for social influence processes. *Psychometrika*, **66**, 161–189.

ROCHA, L.E.C., LILJEROS, F. & HOLME, P. (2010). Information dynamics shape the sexual networks of internet-mediated prostitution. *Proceedings of the National Academy of Sciences*, **107**, 5706–5711.

ROWE, R., CREAMER, G., HERSHKOP, S. & STOLFO, S.J. (2007). Automated social hierarchy detection through email network analysis. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 109–117, ACM.

RYLE, G. (1971). The thinking of thoughts: what is le penseur doing. *Collected papers*, **2**, 480–96.

SARAMAKI, J., LEICHT, E., LOPEZ, E., ROBERTS, S., REED-TSOCHAS, F. & DUNBAR, R. (2012). The persistence of social signatures in human communication. *arXiv preprint arXiv:1204.5602*.

SCHULTZ, U. & VANDENBOSCH, B. (1998). Information overload in a groupware environment: Now you see it, now you don't. *Journal of Organizational Computing and Electronic Commerce*, **8**, 127–148.

SCOTT, J. (2011). Social Physics and Social Networks. In J. Scott & P.J. Carrington, eds., *The SAGE handbook of social network analysis*, SAGE Publications Limited.

SCOTT, J. & CARRINGTON, P.J. (2011). *The SAGE handbook of social network analysis*. SAGE Publications Limited.

SCOTT, J.K. (1991). *Social network analysis: a handbook*. SAGE.

SEDLEY, D.N. (2003). *Plato's Cratylus*. Cambridge University Press.

SERRANO, M.A., BOGUÑÁ, M. & VESPIGNANI, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, **106**, 6483 –6488.

SHALIZI, C.R., CAMPERI, M.F. & KLINKNER, K.L. (2006). Discovering functional communities in dynamical networks. *q-bio/0609008*.

SHELLY, R.K. & TROYER, L. (2001). Speech duration and dependencies in initially structured and unstructured task groups. *Sociological Perspectives*, **44**, 419–444.

SHETTY, J. & ADIBI, J. (2004). The enron email dataset database schema and Brief Statistical Report.

SIMMEL, G. (1908). Die Kreuzung sozialer Kreise. In *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*, Verlag von Dunker und Humbolt, München und Leipzig.

SIMMEL, G. (1964). The triad. In *The Sociology Of Georg Simmel*, The Free Press., twelfth edn.

SKVORETZ, J. & FARARO, T. (1996). Status and participation in task groups: A dynamic network model. *American Journal of Sociology*, **101**, 1366–1414.

SMALL, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, **24**, 265–269.

SMITH, S.S. (2005). 'Don't put my name on it'? Social capital activation and job-finding assistance among the black urban poor. *American Journal of Sociology*, **111**, 1–57.

SNIJDERS, T.A. & BOSKER, R.J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications Limited.

SNIJDERS, T.A.B. & BAERVELDT, C. (2003). A multilevel network study of the effects of delinquent behavior on friendship evolution. *Journal of mathematical sociology*, **27**, 123–151.

SNIJDERS, T.A.B. & KENNY, D.A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, **6**, 471–486.

SNIJDERS, T.A.B., PATTISON, P.E., ROBINS, G.L. & HANDCOCK, M.S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, **36**, 99–153.

SNIJDERS, T.A.B., STEGLICH, C.E.G. & SCHWEINBERGER, M. (2007). Modeling the co-evolution of networks and behavior. *Longitudinal models in the behavioral and related sciences*, 41–71.

SNIJDERS, T.A.B., VAN DE BUNT, G. & STEGLICH, C.E.G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, **32**, 44–60.

SONG, Y., ZHANG, C. & WU, M. (2010). The study of human behavior dynamics based on blogosphere. In *Web Information Systems and Mining (WISM), 2010 International Conference on*, vol. 1, 87–91, IEEE.

SPIEGELHALTER, D.J., BEST, N.G., CARLIN, B.P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

SPROULL, L. & KIESLER, S. (1986). Reducing social context cues: Electronic mail in organizational communications. *Management Science*, **32**, 1492–1512.

STRITE, A. (2013). Making the most of sentiment scores with IKANOW and R. "http://www.ikanow.com/blog/05/22/making-the-most-of-sentiment-scores-with-ikanow-and-r/", accessed: 26/05/2013.

SZELL, M. & THURNER, S. (2010). Measuring social dynamics in a massive multiplayer online game. *Social Networks*, **32**, 313–329.

TANG, L., BARBIER, G., LIU, H. & ZHANG, J. (2010). A social network analysis approach to detecting suspicious online financial activities. In *Advances in Social Computing*, 390–397, Springer.

TARDE, G. (2010). *Gabriel Tarde on communication and social influence: Selected papers*. The University of Chicago Press, Chicago, IL, USA.

TYLER, J.R., WILKINSON, D.M. & HUBERMAN, B.A. (2005). E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, **21**, 143–153.

UDDIN, S., MURSHED, S.T.H. & HOSSAIN, L. (2011). Power-law behavior in complex organizational communication networks during crisis. *Physica A: Statistical Mechanics and its Applications*, **390**, 2845–2853.

UDEHN, L. (2001). *Methodological individualism: Background, history and meaning*. Psychology Press.

UDEHN, L. (2002). The changing face of methodological individualism. *Annual Review of Sociology*, 479–507.

VAN DUIJN, M.A.J., VAN BUSSCHBACH, J.T. & SNIJDERS, T.A.B. (1999). Multilevel analysis of personal networks as dependent variables. *Social Networks*, **21**, 187–210.

VARGAS, E.V., LATOUR, B., KARSENTI, B., AÏT-TOUATI, F. & SALMON, L. (2008). The debate between Tarde and Durkheim. *Environment and Planning D: Society and Space*, **26**, 761–777.

VERMEUELE, A. (2009). System effects and the constitution. *Harvard Law School John M. Olin Center for Law, Economics and Business Discussion Paper Series*, 633–633.

VESPIGNANI, A. (2009). Predicting the behavior of techno-social systems. *Science*, **325**, 425–425.

VROMEN, J. (2010). Micro-foundations for strategic management: Squaring Coleman's diagram. *Erkenntnis*, **73**, 365–383.

WANG, P., ROBINS, G. & PATTISON, P. (2006). XPnet: PNet for for Multivariate Networks. *The University of Melbourne - School of Behavioural Science*.

WANG, P., PATTISON, P. & ROBINS, G. (2012). Exponential random graph model specifications for bipartite networksâĂŤa dependence hierarchy. *Social Networks*.

WASSERMAN, S. & FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1st edn.

WATKINS, J.W. (1957). Historical explanation in the social sciences. *The British Journal for the Philosophy of Science*, **8**, 104–117.

WATTS, D.J. (2004a). The 'new' science of networks. *Annual review of sociology*, **30**, 243–270.

WATTS, D.J. (2004b). *Six degrees: The science of a connected age*. WW Norton, London.

WATTS, D.J. (2007). A twenty-first century science. *Nature*, **445**, 489–489.

WATTS, D.J. & STROGATZ, S.H. (1998). Collective dynamics of 'small-world' networks. *nature*, **393**, 440–442.

WEBER, M. (1978). The nature of social action. In *Max Weber: Selections in Translation*, 7–32, Cambridge University Press.

WHITE, H.C. (1992). *Identity and control: A structural theory of social action*. Princeton University Press, Princeton, NJ.

WIMMER, A. & LEWIS, K. (2010). Beyond and below racial homophily: ERG models of a friendship network documented on facebook. *American Journal of Sociology*, **116**, 583–642.

WOOLF, V. (2012). *Mrs. Dalloway*. Broadview Press.

WRONG, D.H. (1961). The oversocialized conception of man in modern sociology. *American Sociological Review*, 183–193.

ZHOU, D., SONG, Y., ZHA, H. & ZHANG, Y. (2005). Towards discovering organizational structure from email corpus. In *Machine Learning and Applications, 2005. Proceedings. Fourth International Conference on Machine Learning and Applications*.

ZIJLSTRA, B.J.H., VAN DUIJN, M.A.J. & SNIJDERS, T.A.B. (2006). The Multilevel p2 Model. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **2**, 42–47.

ZWIJZE-KONING, K.H. & DE JONG, M.D.T. (2005). Auditing information structures in organizations: a review of data collection techniques for network analysis. *Organizational Research Methods*, **8**, 429–429.