Working apart together : using ICTs in research collaboration
Vasileiadou, E.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
4. Methodology

A. Introduction to the research design

In this chapter I explain the research design and methodology. First, in this introduction, I recapitulate the main research questions, and present the rationale behind the research design, and the choice of the two specific case studies. Second, in the following section (4B), I describe the data sources for each case study and the coding process. Finally, I explain the different analyses performed (4C).

The research design of this study was dictated by the research questions and the theoretical and methodological considerations which arose from the literature review on the topic. The central research questions of this study are:

- How do differences among media influence working practices in research collaboration?
- How do working practices in research collaboration influence the frequency of media use?

As discussed in the first chapter the differences among media relate to specific characteristics identified by medium theory. Moreover, following Nardi and O’Day (1999), media use has been operationalised as frequency of media use, on one hand, and as “media identities”, the routines and rules for the use of each medium, on the other hand. The second theoretical chapter elaborated processes of research collaboration, such as decision-making, conflicts, task allocation, output production and socialising. Finally, chapter 3 provided empirical findings in studies of e-science, which highlighted specific issues that still need to be resolved: e.g. Are ICTs used for socialising? Following the issues raised in the three previous chapters, the two overarching research questions are broken down to the following sub-questions:

1. What is the media configuration of a research team? In other words, what are the types of media used? What are the functions the media support? And what is the frequency of media use?
2. How do media configurations change over time? In other words, how does the type of media used change through time? How do the functions that media support change through time? How does the frequency of media use change through time?
3. How does the frequency of use of one medium (e.g. email) influence the frequency of use of another medium (e.g. blackboard)? In other words, is there a boosting effect between different media over time?
4. How do working processes of research collaboration influence the frequency of media use? Specifically the sub-questions:
   a) How are decisions taken? Is the use of ICTs different for peripheral and for central researchers? How do differences among media influence decision making processes of the research team? Does the use of ICTs increase the pace of decision-making processes?
b) How do conflicts emerge and get resolved? How do differences among media influence tensions and conflicts? Is the use of emails edgy therefore creating tensions?
c) How do socialising processes function? To what extent are different media used for socializing? Does socialising in a team take time to evolve? Do meetings boost the socialising activity?
d) How are tasks allocated? How do differences among media influence task allocation processes? Does the use of ICTs and especially shared databases change the allocation of work between collaborators, and bring communalization of research at an earlier stage?
e) How is knowledge produced? How do differences among media influence the way knowledge is produced in research teams?

5. Are there any other activities in research collaboration influencing the use of media? If so, which ones and how?

The nature of these questions, which focus on a contemporary issue, the use of ICTs in research collaborations, as well as the type of questions posed, which refer to an exploration of the ways in which one process influences another, make the case study design an interesting approach here. “A case study is an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident” (Yin, 2003; p 13). As explained in the theoretical section, the study of media configurations always needs to be situated in a given context (chapter 1), elements of which also influence the relationship between media configurations and collaborative working practices (chapter 2). Indeed, the case study approach can cope with “the technically distinctive situation in which there will be many more variables of interest than data points” (ibid. p. 13).

A two-case-study design was chosen, and the choice of the two specific case studies was based on the logic of deliberate but limited analytic replication. In other words, the research questions are addressed in two comparable contexts: two research teams working on EU-funded FP5 projects. The two teams studied are DELTA and ERICOM\(^\text{11}\). Both of them consisted of researchers from different European countries and different institutional and disciplinary contexts. Both teams worked under 5\(^\text{th}\) Framework Programme regulations and both of them studied the use of ICTs in some context. This, as explained already in the second chapter, means that elements of their organisation were the same. DELTA was investigating the use of email in organizations and its implications for trust and identity, for policy issues and participation in decision making processes. ERICOM was investigating the use of web and non web indicators for the science-technology-economy system. Finally, the duration of both projects was comparable (30 months for DELTA and 40 months for ERICOM).

To translate this in the variables of interest here, the level of *formalization* of the two research groups was the same (EU 5th Framework Programme projects), the level of

\(^{11}\) Fictitious names. I have followed Rasters (2004) in her naming of the DELTA team and its members.
internationalization and interdisciplinarity was comparable (both international and interdisciplinary), their duration was comparable, their intellectual focus was comparable (they were both studying the use of ICTs, DELTA from a social science perspective and ERICOM from an information science perspective). And most importantly, the level of awareness in terms of ICT use was expected to be similar, since both research teams were actually studying the use of ICTs.

At the same time, there was one contrasting difference between the two teams of vital importance in this design, and that was the initial task allocation. In DELTA all participating groups were expected to contribute to all processes of the research design (literature review, data gathering and analysis, writing of results). This task allocation was unspecialized and was expected to lead to an integrative collaboration (Hara et al. 2003) with different dynamics in the working processes. In contrast, the task allocation in ERICOM was based on a specialized division of labour. The research design, the data gathering, data analysis and writing of reports were the tasks of different groups, which was expected to lead to a more complementary type of collaboration. The type of task allocation was also expected to influence all working processes (see chapter 2). Therefore, I would be able to study the influence of media configurations on different types of collaborations. So, the first reason behind my choice of the two case studies was theoretical, that is, the degree they varied on the type of original task allocation.

Furthermore, one of the aims of this study was to overcome a limitation of previous studies, which were based on self-reported data about ICT use, and study the use of communication media in real time (Matzat, 2004; Vasileiadou and van den Besselaar, 2004b). I wanted to avoid the weaknesses identified with self-reported data, that is, response bias, inaccuracy due to poor recall (Yin, 2003: p 86), as well as responses based on expectations and beliefs, rather than actual use. This would require the use of archival sources for intra-team media use, for example log-files or archived emails, which are generally very difficult to obtain for privacy reasons.

However, I was able to observe closely both projects, which gave me access to the archival systems of their emailing lists, and, for the case of ERICOM, the log-files of their website. I also had the opportunity to observe the interactions of the teams, attend meetings, and obtain external and internal documents. Therefore, the reason of data accessibility, and more importantly real-time internet data, which was an important aim of this study, was the second reason behind the choice of these two cases. It is emphasised that the members of both teams were aware and explicitly agreed to the observation of their communications and collaboration patterns during the period of the projects. Further, the coordinators of both teams granted explicit access to the public material for research purposes12.

After this introduction, I now turn to the presentation of the data collected for each case study and the coding process.

12 No archive of private emails of any member was used for the dissertation.
B. Data sources

The strength of a case study design is the multiple data sources which can be used in a triangulating fashion to answer the research questions. In the current design, different data sources are used to answer the research questions, together with different methods on the same data source. For each case study, I have used as data sources the following: documentation, interviews, archival records, observations. The variety of data sources in the current dissertation is intended to balance the weaknesses of each source of evidence, while capitalising on their strengths (see Yin, 2003: 86; Kelly and McGrath, 1988: 69-71). Hereunder I elaborate on the data sources I used.

Documentation refers to documents about the case under study and for both case studies I have collected and used 1) internal documents (draft papers, minutes of meetings etc) 2) external documents (deliverables, progress reports, cost statements, contracts, published articles addressed to peer scientists) 3) scientific publications as secondary sources (for the case of DELTA there was a scientific publication describing the collaboration process).

Archival records represented a major data source in the current dissertation, and they are traces of behaviour that the actors are aware of. The archival records used for both cases are the archives of the two emailing lists operated by both teams. These consisted of an archive of each message sent to the list, the time it was sent and the author, the subject line, and the content of the email as well as the attachments it may have included. To my knowledge, this type of data, namely the full content of emails and their attachments, is rarely accessible for research, which is one of the innovative points of the thesis.

The archive of the general emailing lists studied for the case of DELTA consists of all emails (2726 emails) from the start of the emailing list (August 2000), until November 2003, when all managerial tasks of DELTA had to be completed (one month after the formal completion of the project, end of October 2003). It also includes all 415 email messages sent to the managerial emailing list, which was intended for administrative and managerial issues of the project. For the case of ERICOM this archive consisted of 697 emails sent to the general list and 379 emails sent to the managerial list (from their establishment until April 2004). Moreover, the electronic archive of all activities of the ERICOM team in the password-protected online area (internal forum of the website) was studied (from 11/12/2001 until 31/4/2004).

Another source of evidence for both case studies was direct observations. For both case studies I have been an observer of the collaboration process and their communication patterns (24 months of the 30 working months for DELTA and 12 months of the 40 for

13 In the Appendix (p. 1) the complete list of data sources for the two projects can be found.
ERICOM). The observations were used as a secondary source, for the verification of facts and team events, for which there was no other archival data available e.g. whether a meeting went on for one or two days, etc. In this way, an attempt was made to reduce the influence of my own recollections, and their limitations, to the minimum. As a final data source, three focused interviews from two members of the ERICOM team (Jack and Hendrik) were taken, with questions focusing on missing factual details of the project work and history\textsuperscript{14}. Furthermore, personal discussions with most team members of the two projects were conducted during the observation periods.

The archives of the emailing lists posed some challenges. For DELTA the archive of the list was in a password-protected area online. It was therefore simple to download the content of the archive. The archive was transferred in an excel file (which could then be inserted in an SPSS file) with the use of software, with each row representing an email message and the following data collected: subject line, author, time and date of message. The sheer number of emails would have made it difficult and time-consuming to order the data in this way manually and without mistakes. This automatic ordering and storing of online data is becoming increasingly important, and it is one of the skills that communication researchers need to master.

For ERICOM this posed additional problems, since there was no automatic saving of the email archive in the second period of the project, whereas in the first period, the archive of the general list was deleted. I used the combination of four different archives of the public lists from three participants, as well as the server archives for the managerial list in the first period\textsuperscript{15}. However, the different archives of the lists were in different formats, and transfer to one format proved impossible. The archives were inspected more than once to identify the missing emails, and also the exact time and date that each email was sent to the lists, which was a laborious process. The emails were put in an excel file semi-manually, which was very time-consuming.

**Databases and Coding**

The basic coding of the activities was performed against time, for theoretical reasons. Both media configurations and collaborative working practices were expected to change through time, as new media were introduced to the teams, communication routines were established and new ones introduced. Therefore, the duration of each project was divided in weeks and each communication activity was placed in this timeline. Weeks were selected as an intermediate time measurement: detailed enough in order not to lose information and to observe the interaction between collaborative working practices and media configurations, but not as specific as days, which would make the analysis of patterns more difficult, given the three years’ period that the teams were working. At the same time, weeks emerged from the observation of the communication patterns of the

\textsuperscript{14} All three interviews were tape-recorded and transcribed verbatim.

\textsuperscript{15} I remind that no archive of private emails was used for the analysis.
teams, since it was rare and idiosyncratic that an email was sent to the lists on Saturdays or Sundays, and there was never a meeting on Sunday.\(^{16}\)

This time coding was performed with the availability of data in mind. For DELTA, the data start before the formal beginning of the project, when the general list was established. There, week 1 was the first week of the start of the team-wide communication system, which coincided with the emergence of the team (as a communication system). For ERICOM, week 1 was when the kick-off meeting took place, which actually marked the formal start of the project, and it was when the establishment of the general and managerial lists was decided.

From then on, four databases for each project were constructed. The first was a database of all communication activities and collaborative practices per week: the content of emails, the minutes of the meetings, all documentation and interview excerpts were placed in time, which enabled me to track down all activities in the two projects each week and uncover time relationships. Qualitative analysis was performed on this database (see below).

Media configuration consists of three different dimensions: the type of media used, the functions that media supported, and the frequency of media use. These three were operationalised respectively as: distinct communication media used by the team; the functions as working processes that the media were used for, and the degree of formalisation (to reflect the practices and routines of use, suggested by Nardi and O’Day); and, finally, the frequency of use of team-wide media. Moreover, the changes in time of these three dimensions were also studied. The differences among media were not expected to change through time, as the distinct characteristics of face-to-face meetings in comparison to personal emails were not expected to change. However, the functions that the media supported through time, as well as the frequency of media use through time could change through time. Finally, the interconnections between the types of media used, the interconnections between the working processes and formalisation of each medium, and the interconnection between the frequencies of different media use were also investigated, according to the definition of media configuration (chapter 1).

Working processes are operationalised as public team-wide processes. Individual conflicts or decisions among members were not considered as such, unless they became visible to the team and could be traced through one of the team-wide means of communication (meetings or lists). In addition, some intra-group, local conflicts may not emerge at a team-level, but may nevertheless influence team-wide work, as for instance, the decision of an individual participant to quit working. These limitations will be further

\(^{16}\) For the DELTA case the percentage of the emails sent during weekends overall were: 5.8% of the emails in the general list, 6.3% of the emails in the managerial list. The weekend emails were written by managers and researchers alike. For the ERICOM case these were: 4.9% of the emails in the general list and 2.2% of the emails in the managerial list. In ERICOM more than six out of ten emails written in the weekends were by two managers of the team.
discussed in the conclusions of the dissertation. However, the focus of the dissertation is on the team-level processes (global dynamics), that is, how the differences among media use in a collaborative team influence team collaborative processes and how team collaborative processes may influence team media use. Local dynamics were studied insofar as they visibly influenced global dynamics.

Decision-making processes are operationalised as interactions through which one or more proposals or questions were communicated to the team, and then one or more team members offered their opinion about it. Sometimes this resulted in an actual decision being taken, and other times it did not lead to a decision. Conflicts and tensions were operationalised as interactions in which one or more members raised their voices (in a face-to-face meeting) to disagree with someone else, or use capital letters in an email. Moreover, the content of emails was scanned for phrases indicating tension e.g. “Listen, [name]”. In addition, disagreements that were conceptualised as such by the team members were also included. For instance, if one member referred through an email to “a fight” in a meeting, that means that the members, themselves, conceptualised some processes as conflicts.

Socialising activities referred to interactions about non-work-related topics, such as exchange of personal information and emotional support, jokes and wishes, pleasantries, such as “Thank you very much for your email”, the use of irony or exclamations as “Dear Friends”. Task allocation processes were operationalised as interactions through which it was decided which team member was supposed to do what in which way within a time frame. It could be the scientific leader announcing that someone needed to improve the public website, or a common-decision through which all team members agreed that a deliverable needs to be written by someone and sent by a specific deadline. Initial task allocation was operationalised as the allocation of person-months in work packages between the different groups.

Output production was operationalised as all interactions involved in the production of a technical, scientific report either for the funding agency (internal output production) or for a wider audience, e.g. a conference presentation (external output production). This referred to the output production of the team, as opposed to individual members. For instance, a published article by one or two members, on the basis of their team work would be included in the analysis. However, a published article by a team member, on the basis of work other than the given project (e.g. another project, or supervision of students) was not taken into account. In short, output production was operationalised as interactions related to scientific and technical output and knowledge production tools that were produced on the basis of the project under investigation. Only reports based on data and their analysis were included: those that were or could be published in peer-reviewed technical journals.

\[17\] Of course the use of capital letters may be accidental. The cases coded as conflicts were when capital letters were used to point to a different opinion or argument during a disagreement.
The second database consisted of a quantitative account of all internal team-wide communication activities each week: the use of the two lists and their attachments, the use of the blackboard/ internal forum and the meetings. This was inserted in an SPSS file and time series analysis was performed (see below). This was deemed necessary as the frequency of media use is an important dimension of media configuration. Each row represented one week and the variables were: “the number of individual emails sent in a week” for the general and the managerial lists, and the attachments were coded as “number of unique attachments sent to the list in a week”. The meetings were coded as “the number of people participating in the meeting”. Here it should be noted that meetings were analysed together with the other communication media because of the theoretical dictation to examine all communication media together, which includes face-to-face meetings (see chapter 1). Moreover, deadlines were operationalised as “the number of formal documents expected to be delivered to the Commission that week, according to the contract”. I distinguished between managerial deadlines (deadlines for the texts referred to as project management in the contracts, progress reports and cost statements) and substantive deadlines (deadlines for other texts).

The use of the blackboard area was less straightforward. For the DELTA case, statistics and data were available from the blackboard platform that the team used, but a visual inspection of the data showed that it was unreliable (inconsistent in time and also inconsistent with each other). One of the team’s decisions was quite helpful in the coding. Since the introduction of the blackboard for file and data sharing, the team decided that each document uploaded would be announced to the team via an email to the general list, so that each member would know what is uploaded and in which folder. This proved especially useful: using as a data source the content of the emails in the general list I coded the use of blackboard as “the number of unique documents uploaded in a given week”. My observation indicated that this decision was followed as a practice.

For the case of ERICOM, the activity in the project website was stored as logfiles by the website server using Microsoft Internet Services. This meant that all activity in the internal forum (password-protected area) was recorded, but the data needed substantial cleaning and sorting. First, the format of the text files was processed, so that they could be inserted in Access. Then all the activity of the internal forum was isolated. Following this, the logfile entries with status code 200 were selected, with method “get” and a file

---

18 For instance, if a word document was sent in an email in two different formats (e.g. in .doc and in .rtf format) I coded this as one attachment.

19 The data for the deadlines were according to the contract of each project. For ERICOM, as there was an amendment of the first contract, I used the deadlines for the first contract until the time of validation of the second contract, and then I used the deadlines of the second one. Moreover there were no formal deadlines in the contract for the submission of management reports and cost statements. I used the dates of the evaluation meetings instead.

20 The website the blackboard and the managerial lists were not available from the beginning of the project. Also, during some weeks the blackboard or the website did not function due to technical reasons. In these weeks the media was coded as ‘missing’.
extension (***). In this way only the “successful downloading activity” of the members was coded as blackboard use\textsuperscript{21}. So for DELTA, the uploading activity was available for analysis, whereas for ERICOM the downloading activity.

Finally, this quantitative database also included the collaborative practices that were present each week. Each collaborative practice (decisions, conflicts, socialising, output production and coordination), with the operationalisation discussed above, was coded as 0 (indicating absence that week) or 1 (indicating presence that week).

Summarising, for DELTA, this database consisted of: number of weeks in each row (N = 171); number of emails sent to the general list each week (total = 2726 emails); the number of emails sent to the managerial list each week (total = 415); the amount of unique attachments sent to the general list each week (total = 472); the amount of attachments sent to the managerial list each week (total = 70); the amount of documents uploaded in the blackboard area (total = 96); the amount of people participating in meetings (total amount of meetings = 11); the amount of substantive texts under deadline that week (total number of deadlines = 14); and the amount of managerial texts under deadline that week (total amount of managerial deadlines = 15).

For ERICOM, this database consisted of: number of weeks in each row (N = 172); number of emails sent to the general list each week (total = 697 emails); the number of emails sent to the managerial list each week (total = 379); the amount of unique attachments sent to the general list each week (total = 232); the amount of attachments sent to the managerial list each week (total = 98); the amount of documents downloaded from the online area (total = 3922); the amount of people participating in meetings (total amount of meetings = 25); the amount of substantive texts under deadline that week (total number of substantive deadlines = 16); and the amount of managerial texts under deadline that week (total amount of managerial deadlines = 5).

The third and fourth databases of each project consisted of all email messages sent to the general list and the managerial list (each row was one email message). For DELTA the general list database had N = 2726 observations (emails) and the managerial list N = 415 observations (emails). For ERICOM the general list had N = 697 observations and the managerial list N = 379 observations. The data coded there was, the date, week (total 171 weeks) and month (total = 40 months) that each message was sent, the author of the email, which local group he/she belonged to, the subject of each email, the number of attachments it contained, and whether the email contained socialising elements. This was coded as “the use of jokes, irony, wishes, pleasantries, emoticons, emotional support, or information about private life” in the email\textsuperscript{22}. Moreover, the status of each author was coded as to whether he/she was a “researcher” in the team or a “scientific coordinator”:

\textsuperscript{21} For more information on the technical details required for the cleaning of logfile data see the online manual of Microsoft Internet Information Services and the related technical support webpages at www.microsoft.com
\textsuperscript{22} This was a dichotomous variable: absence coded as 0 and presence coded as 1.
the latter were the people responsible for the project in each local group. This was how peripheral and central members were operationalised.

Finally, the subject header was used to identify whether it was a reply to a previous message (when it started with RE:/ Re:/ Rep./ Antw.) or an original contribution. In this way the initiation of communication through the list (direction of information flow) was operationalised. For DELTA, the general list contained 936 original contributions in total and the managerial list 115 original contributions. In ERICOM, the general list contained 337 original contributions in total and the managerial list contained 213 original contributions in total. On these two databases the methods of analysis performed were cross-tabulations, to identify whether peripheral or central researchers used the two lists in a different way, also whether peripheral or central researchers socialised more through the general list.

---

23 This was also a dichotomous variable: researchers coded as 0 and managers coded as 1.
C. Methods of analysis

The design of this study is based on the mixed methods approach, which focuses on collecting and analyzing both quantitative and qualitative data in a single study (Creswell, 2003: 210). The reason behind this is not only to expand our understanding from one method to another, but also to be able to study media configurations and collaborative working practices in a systematic and quantifiable way (quantitative methods of analysis), as well as to understand their interaction in depth (qualitative methods of analysis).

Moreover, in chapter 1, the idea of different levels of analysis for the directionality of influence was introduced. There it was argued that the influence of research collaboration practices on the frequency of media use can be identified through studying at a low level of analysis: the level of everyday activities. As explained above, instead of using everyday activities, which for a period of three and more years would be time-consuming, I used the weekly activities. So, how do processes of research collaboration influence the frequency of media use? The answer to this question was answered with quantitative analyses (ARIMA models, visual inspection of graphs) focusing on the frequency of media use, and is presented in the first chapter of each case study (chapters 5 and 7).

At the same time Meyrowitz (1994) suggests that the influence of media on social processes, that is, “what difference do different media make?”, can be identified by historical analysis and large-scale pattern identification. In the context of the two research teams, historical analysis can be understood as longitudinal analysis, and large-scale pattern identification was performed through qualitative analysis, which seeks to identify common patterns at a more aggregated level of analysis. The answer to these questions was answered using mainly qualitative analysis, and as a secondary method quantitative analysis, and is presented in the second chapter of each case study (chapters 6 and 8).

As mentioned above, the types of analyses performed were: qualitative analysis, statistical analyses and time series analysis. In what follows, I briefly describe them, concentrating more on ARIMA models, since it is an uncommon method in communication science and sociology of science.

1. Qualitative analysis

The qualitative analysis was performed both for the identification of the functions of media and for collaborative working practices. Qualitative analysis is often described as a creation process, where the analyst discovers and develops categories in an iterative bottom-up approach (Strauss and Corbin, 1990).

For the collaborative working practices a database was constructed which contained all decisions separately, all conflicts separately, and so forth. This data was taken from the content of the emails available, the operation of the blackboards each week, the minutes
of the meetings, the draft documents sent by the team members. I went through these
different databases with the following questions in mind: Which medium is used in this
interaction? Who participates in this interaction? What is their relative power? What are
the formal and informal rules of the process? How does it relate to other team processes?
What types of interactions are they (face-to-face, mediated, quasi-mediated)? Who, if
anyone, controls the circulation of the content? What type of visibility or publicness
emerges and how does this visibility influence the process? These questions come from
the discussion in chapter 1 about how differences among communication media may
influence working processes. The answers to these questions led to the emergence of
patterns throughout some or all working processes.

For the identification of the media functions a similar theme analysis was performed. A
database containing a description of the uses and functions of the media per week was
used and the questions guiding the identification of the themes were: For which working
processes is each medium used? Who communicates through this medium? Are media
used for internal or external communication? Are media used for team-level
communication or personal communication? These two questions relates to the possible
emergence and development of team identity, which is based on the circulation of
material internal to the team, but not to outsiders (Meyrowitz, 1994). What are the formal
and informal rules for the use of each medium? This question relates to how Nardi and
O’Day (1999) identify the “media identities”. Going through the database with these
questions in mind I identified relevant patterns for the function of each medium.

For both the identification of working processes and of the media functions, the process
was the same, even though the questions were somewhat different. I went through each
database identifying the answers to these questions for every interaction noted down, and
organised the answers to these questions in patterns, as some of them were linked. For
instance, very often when there was someone controlling the circulation of the content
through the general list, it was the scientific coordinator. This meant that in most
interactions the answer to one question (Who participates in this interaction? What is
their relative power?) was linked to the answer to another question (Who, if anyone
controls the circulation of the content?). The linkages between answers of these
questions, when they emerged often, formed a pattern. Then, in the next interaction under
analysis I controlled whether these patterns applied and could describe and organise the
interaction (Boyatzis, 1998: p. 4). Sometimes the patterns were adjusted in the process,
and sometimes new patterns emerged, as the function of a medium changed through time.
Sometimes, these patterns were refined as more interactions were studied, and other times
they were analytically divided into two different patterns. This was performed twice: I
first went through the databases organising the answers into patterns and identifying, and
elaborating the patterns. After this process, I went through the databases for a second
time to confirm that these patterns described and organised the data.

So, this process followed an inductive logic insofar as the patterns emerged from the data
(Patton, 1987: 150) but at the same time the questions leading the analysis were based on
the theoretical framework and the elaboration of the concepts in use.
2. Time series analysis/ ARIMA models

Often in social sciences we have observations of a variable over a period of time. In this dissertation, emails were sent each week, for a period of more than 3 years. What makes time series analysis stand out, compared to conventional analytical methods, is that it takes into account the dynamic process and potential change of a variable over a period of time. This allows for a different set of research questions to be asked. For instance in econometrics, one is interested in identifying how the GDP of a country fluctuates over a period of 20 years. Similarly, here I am interested in understanding how the communication patterns of the team develop over a period of three years. Further, the advantage of time series is that it may enable causal relationships to be established strongly. It is generally assumed that a cause (independent variable) has to precede the consequence (dependent variable). Time series analysis enables the identification of such time lagged relationships.

The second database of each case study (described in p. 69-71) contained the use of media per week, for 171 observations (DELTA) and 172 observations (ERICOM). It is reminded that weeks were selected as the lowest aggregation level possible without losing any information of the variable (see also p. 69). These variables are time series variables, as they were obtained by measuring a single variable (e.g. number of emails) regularly over a period of time. Conventional statistical methods cannot be used for such time series variables, because these methods assume that the observations are independent from each other (Yanovitzky and VanLear, 2007; 90). The observations of these variables, however, are usually not independent: one would expect for example that the amount of emails sent in week 1 influences the amount of emails the following week (as e.g. response). Therefore, specific time series analysis was used, in order to fully understand the dynamics of the variables in time, and to treat them in a statistically appropriate manner.

Even though there are a number of time series analytical techniques, here ARIMA models were used. ARIMA stands for Autoregressive Integrated Moving Average, and was introduced by Box and Jenkins (1970). ARIMA is often used in econometrics when the aim is to understand the dynamics of a time series variable (with more than 50 observations) endogenously. That is, the idea behind ARIMA is that the past of a variable contains information which can help us understand the variable, and only at a second stage are exogenous explanatory variables considered. With ARIMA, one is basically building a model of the variable based on its past. In econometrics, ARIMA models have been used mainly for forecasting.

The analysis of each variable starts by employing visual techniques to assess the distribution shape (e.g. if there is a trend over time), the possible outliers, and whether the variable changes behaviour over time (Warner, 1998; Parsonson and Baer, 1978)\(^{24}\).

\(^{24}\) Instead of using the residuals of OLS regression to perform ARIMA, which some authors recommend as a strategy (Warner, 1998; Brockwell and Davis, 2002), I have used the variables themselves to perform the ARIMA, as recommended by McCleary et al. (1980).
First, it needs to be ensured that the variable is stationary. This means that the variance of the variable, and the mean of the variable are stable through time and they do not depend on the time of measurement (Gottman, 1981: 61). The visual inspection of the variable indicates whether the mean is not stable (e.g. drift of the mean, or linear trend) or whether the variance is not stable (more turbulent in higher values, more tranquil in lower values). However, there are more formal stationarity tests. Here I used the augmented Dickey-Fuller test (ADF), which tests for stationarity of the mean. If the ADF test indicates non-stationarity, that means that the variable is integrated of order one I(1). This means that it is necessary to take the difference of the series and build an ARIMA model on the differences. For the differenced series it is again required to be stationary, and if it not, we need to difference the differenced series.

The next step is autocorrelation analysis (ACF) and partial autocorrelation (PACF). This analysis helps one identify whether each variable has memory, that is, the extent to which the present values of a series are predictable from its past values (Gottman, 1981: 33). ACF gives the correlation of the variable between two lags, and PACF gives the correlation of the variable between two lags, controlling for the effect of earlier lags.

The visual inspection of the ACF and PACF graphs leads to a tentative identification of a suitable ARIMA model (McCleary et al., 1980; SPSS Inc., 1987: E9-E11; Hollanders and Vliegenthart, 2008). Of importance is how many spikes the two graphs have (which shows whether it is a first or second- order model), and how smoothly they decline. If the ACF graph declines smoothly and the PACF abruptly, the variable can be modelled with Autoregression. If the PACF graph declines smoothly and the ACF abruptly, the variable can be modelled with Moving Average. In practice, it is often not easy to detect these differences, so there are a number of additional criteria for the best-fit model.

The first, most important, criterion is that all parameters of the model are statistically significant. Second, the indicators AIC (Akaike Information Criterion) and SBC (Schwarz Bayesian Criterion) are used to identify which model fit the data best. The lowest AIC and SBC indicate the most parsimonious model. Finally, each model needs to have residuals (error) normally distributed, that is, without any autocorrelation. For this reason, for each model the residuals were saved and an ACF/ PACF function was run on those residuals, to confirm that they had no memory.

The ARIMA analysis is based on the concept of random disturbances or shocks in the process of each variable. That means that between two observations, a disturbance

---

26 For autocorrelation to exist in a series the value of the ACF/PACF needs to exceed two standard errors and the Box-Ljung test needs to be statistically significant.
27 In the Appendix p. 2 exemplary ACF and PACF models are shown for each type of ARIMA model.
28 Random shocks or disturbances are the same, and the terms will be used interchangeably throughout the dissertation.
occurs that affects the level of the series. This disturbance, or random shock, may be any factor which varies across time and interacts with the variable under study (McCleary et al., 1980). In effect, the ARIMA analysis creates a model of the variable in terms of its past and a current disturbance. There are three types of processes that can be modelled with ARIMA: autoregression (AR), moving averages (MA), integrated models (I).

The AR e.g. ARIMA (1,0,0) models show that the value of the variable is a linear function of the preceding 1 or 2 (in models 2,0,0) values. Conceptually the model is one with “memory” in the sense that each value is correlated with all preceding values. Thus each shock or disturbance to the system has a diminishing effect on all subsequent time periods. The φ coefficient indicates how strongly each value depends on the preceding value. The formula of the model is:

\[ \text{Value}(t) = \text{disturbance}(t) + \varphi \times \text{Value}(t-1) \]

The MA models e.g. ARIMA (0,0,1) indicate that each value of the variable is determined by the average of the current disturbance and the previous disturbance. So in an ARIMA (0,0,1) model a disturbance affects the process for the current week and the week after that, and then it abruptly ceases to affect it. For instance, negative MA values at short lags indicate a process that is influenced by shocks, and then moves back to its mean (Hollanders and Vliegenthart, 2008). In this sense, “the AR process is said to have a longer memory than the MA process” (Romer, 2006: 173, italics in the original). The coefficient of the process is called θ and shows how strongly each value is correlated with the previous disturbance. The formula of the model is:

\[ \text{Value}(t) = \text{disturbance}(t) + \theta \times \text{disturbance}(t-1) \]

The Integrated models - e.g. ARIMA (0,1,0) - were discussed above since they are models of non-stationary variables. Substantially they reflect the cumulative sum of the shocks in each previous time period. Each value equals the previous value (which is the cumulative sum of changes/differences in the previous stages) and some random fluctuation (disturbance). This type of process is the sum of all past disturbances, and in this sense the integrated models have ‘perfect memory’: they are the most sensitive to disturbances, since any shock (event) has a “permanent effect” (Hollanders and Vliegenthart, 2008; 53), and not a diminishing effect (as in the AR-model). As discussed already, in a (0,1,0) model the variable is differenced, and in practice, this means that the mean of the variable drifts, and the variable is not stationary in its mean. The formula of the model is:

\[ \text{Value}(t) = \text{value}(0) + \text{disturbance}(1) + \text{disturbance}(2)+\ldots+\text{disturbance}(t) \text{ or simpler:} \]

\[ \text{Value}(t) = \text{value}(t-1) + \text{disturbance} \]

Apart from differencing, a further transformation that a variable might need is the log transformation. This is necessary when the higher the mean of the variable, the higher its variance. This is called autoregressive conditional heteroscedasticity (ARCH) and results in residual variance that is not stable over time (Vliegenthart, 2007). ARCH is also a specific modelling technique, but it was not used here. Instead I performed log transformation, to deal with non-stationarity in the variance. Whether or not a log transformation is required can be identified during the visual inspection: if the variable follows a cubic or other non-linear trend, a log transformation is required.
A variable that needs differencing (Integrated model) or log transformation is not stationary. In contrast, variables modelled by MA or AR have stationary underlying processes.

For the analysis to become clearer, the distinction is drawn between the observed time series (the variable) and the process underlying this observation (Gottman; 1981: p. 61). In the context of this dissertation, for instance, the number of emails through the general list is the actual variable, and the everyday coordination of the team is the process underlying this variable. Therefore, the variable may fluctuate in time (e.g. 15 emails in week 1, 20 emails in week 2), but the underlying process may be stable in historical time throughout the period of observation (stationary).

Even though ARIMA models have been criticised of being a-theoretical, each element of the analysis can help us understand the behaviour and dynamics of the variable. Especially given the importance of the time dynamics in complex dynamic systems, ARIMA models can be used in a fruitful way. A variable described with an Integrated model (0,1,0) is considered the most sensitive to random shocks, as each value is a sum of all past disturbances, and therefore contains “perfect” memory of all past disturbances. This can also be understood as path dependence of the initial conditions (shocks). The AR-model is less sensitive to random shocks, since each shock has a diminishing effect on the value of the variable over time. It has longer memory than a variable described by a MA-model, since in a MA-model, a random shock influences the variable at time t and time t+1 (if it is a 0,0,1 model) and then it abruptly ceases to influence it. So, a MA-model has temporary short-term sensitivity to random shocks. When a variable shows a linear trend (that is, when it needs differencing), any shock/disturbance has a permanent effect, whereas variables described with an AR-model, or a MA-model are mean reverting: how big the influence of the shocks may be at time t, the variable comes back to its mean (Hollanders and Vliegenthart, 2008).

Finally, the ARIMA analysis may provide a model with more than one process. For instance, an ARIMA (0,1,1) may be identified as the correct model of the variable. This is the case when the differences of the values of the variable still have memory described by an ARIMA (0,0,1) model (McCleary et al., 1980). This means that the variable is differenced and the MA values indicate how random shocks in previous change scores influence current change scores.

Once the variable has been modelled, this model can be expanded to include exogenous, explanatory variables. In practice, the SPSS ARIMA routine allows one to add explanatory variable(s) in the same model used to describe the dependent variable. In this way, a model is created for each dependent variable that describes it in terms of its own past, and the influence of an independent variable. Often, a change in the independent variable precedes (or is expected to precede) in time the change in the dependent variable. In these cases, it is necessary to include lagged values of the independent variable in the equation. To identify the appropriate lag length, as well as to identify
whether lagged (or contemporaneous) values should be included, the cross-correlation function can be used.

Cross-correlations (CCF) perform the same function as correlations but are used for the identification of associations between two variables at different time lags: e.g. when we want to see whether a variable at time $t=1$ is associated with another variable at time $t=2$ (or any other time lag). I used cross-correlations to establish whether lagged values of the independent explanatory variables should be added in a model of a dependent one, and what was the appropriate lag length. Further, cross-correlations between media variables were also conducted. A statistically significant negative cross-correlation between two media variables would mean that the more one medium is used, the less the other one is used, which may suggest a substitution effect, at least in short term. A statistically significant positive cross-correlation would mean that the more one medium is used, the more the other one is used as well, indicating a clear boosting effect.