How Do User Groups Cope with Delay in Real-Time Collaborative Note Taking

Claudia-Lavinia Ignat, Gérald Oster, Olivia Fox, Valerie L. Shalin and François Charoy

Abstract A property of general interest of real-time collaborative editors is delay. Delays exist between the execution of one user's modification and the visibility of this modification to the other users. Such delays are in part fundamental to the network, as well as arising from the consistency maintenance algorithms and underlying architecture of collaborative editors. Existing quantitative research on collaborative document editing does not examine either concern for delay or the efficacy of compensatory strategies. We studied an artificial note taking task in French where we introduced simulated delay. We found out a general effect of delay on performance related to the ability to manage redundancy and errors across the document. We interpret this finding as a compromised ability to maintain awareness of team member activity, and a reversion to independent work. Measures of common ground in accompanying chat indicate that groups with less experienced team members attempt to compensate for the effect of delay. In contrast, more experienced groups do not adjust their communication in response to delay, and their performance remains sensitive to the delay manipulation.

Introduction

Computer science work, including Ellis et al. (1991), Sun et al. (1998) and Ignat and Norrie (2008), provides the technical capability to distribute document editing among multiple users. Synchronous or real time collaborative editing allows

C.-L. Ignat (✉) · G. Oster · F. Charoy
Inria, 54600 Villers-lès-Nancy, France
e-mail: claudia.ignat@inria.fr

C.-L. Ignat · G. Oster · F. Charoy
LORIA, Université de Lorraine, CNRS, 54506 Vandoeuvre-lès-Nancy, France

O. Fox · V.L. Shalin
Department of Psychology, Wright State University, Dayton, OH, USA

N. Boulus-Rødje et al. (eds.), *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19–23 September 2015, Oslo, Norway*, DOI 10.1007/978-3-319-20499-4_12

a group of people to modify a shared document at the same time. One user's changes appear to other users almost immediately with very small time intervals of inconsistent document status. Real-time collaborative editing has gained in popularity due to the wide availability of free services such as Google Drive. Existing real-time collaborative editing tools are currently used in scenarios involving only a small number of people (e.g. up to 10) contributing to a shared document such as a research paper or project proposal or meeting notes. However, scenarios involving large number of users are currently emerging, such as group note taking during lectures or conferences. Existing tools are not currently designed to support this change completely in terms of the number of users, ultimately limiting the number of users that can simultaneously edit a document.

The requirements for group performance in the case of a large number of users are not established. One system property of general interest is delay. Delays exist between the execution of one user's modification and the visibility of this modification to the other users. This delay has many causes: network delay due to physical communication technology be it copper wire, optical fiber or radio transmission; time complexity of various algorithms for ensuring consistency, where most of them depend on the number of users and number of operations that users performed; the type of architecture such as thin or thick client. Understanding the requirements associated with delay informs the broader research community in the domain of collaborative editing, which continues to develop merging algorithms under the uniform assumption of high responsiveness requirements for real-time collaboration. Potentially, modest delay is well-tolerated and can suspend further optimisation research. Worse, high responsiveness could interfere with user productivity under certain circumstances.

Not all groupware applications appear sensitive to delay. For example, Dourish and Bly (1992) claim: "*We can tolerate a certain amount of delay*; *image updates may only occur every ten minutes*, *and so the user will not expect up-to-the-second information.*" Others argue that usability limitations of otherwise effective groupware may yield to adaptations in work practice (Olson and Olson 2000). Some designers even suggest the benefit of delay warnings, so that users can adjust their strategies if they are aware of system conditions (Vaghi et al. 1999; Gutwin et al. 2004). Some work has examined the effect of network delay on multi-player real-time games on the order of 1 s delay (Gutwin 2001). But, no study has been done in collaborative editing where much longer delays result from other factors than network delay. Such factors include consistency maintenance algorithms that may scale with the number of users and operations (Ahmed-Nacer et al. 2011).

In this paper we aim to evaluate the performance consequences of delay in real-time collaborative document editing. Setting up an experiment with numerous users that edit concurrently a shared document would not be possible with current tools. Existing tools restrict the number of users editing a document and most of them are not open-source in order to allow code instrumentation for an analysis of user behavior. We instead mapped the real-world setting to a laboratory task that permits the systematic manipulation of delay. First, we used a simulation with GoogleDocs to estimate the range of delays, taking into account the

number of users and their typing speed. Then we examined the effect of simulated delay within this range on a note taking task performed by a small group of users. As GoogleDocs code is not open source, we used another well-known editor, EtherpadLite, that we instrumented for introducing artificial delays. In particular, we analysed the effect of delay on the error rate and redundancy during the collaborative process. We also examined compensatory strategies for dealing with delay such as coordination.

We structure the paper as follows. We start by presenting our research questions and related work. We then describe our collaborative note taking task and design of artificial delays that we introduced in our experiment. We then present the experimental procedure we followed and the dependent measures. We next present results of our experimental design followed by a discussion. Concluding remarks are presented in the last section.

Research Questions and Related Quantitative Research

None of the field studies on collaborative writing tools such as the one presented by Tammaro et al. (1997) or usability studies such as the one presented by Noël and Robert (2004) provides quantitative behavioral evidence to define limits for collaborative editing technology. While delay certainly affects the performance of the individual, our interest lies in the consequence to real-time collaborative editing and the compensatory strategies at the team level that users adopt to overcome the negative effect of delay. Olson and Olson (2000) claim that coupling between sub-tasks influences tolerance for delay. Collaborative note-taking has the potential to maximize sub-task coupling between users, and provides an ideal task for identifying the range of delay tolerance. In the remainder of this section, we consider the existing research and its implications for our study along three dimensions: The likely range of effective delay, informative outcome measures, and informative collaborative measures.

To examine the effect of delay experimentally we require a range of delay values to study. Studies of the effect of delay in gaming environments such as the ones presented by Gutwin et al. (2004) and Vaghi et al. (1999) examine tasks with time constants (or turns) on the order of 700 ms. Results suggest performance decrements with delays as small as 200 ms (Gutwin et al. 2004). However, 200 ms delays are much smaller than delays in collaborative editing that are in the order of magnitude of several seconds. As shown by Karat et al. (1999), the average typing frequency is around 2 characters/s. We therefore expect that the task-time constant of collaborative editing is proportional to the time to type a word, or 2.5 s for an average of 5 characters per word. We should not expect delay effects in such tasks below this level of delay. However, the absence of a clear precedence for our paradigm suggested the need for a supplementary simulation study to determine a likely effective range of delay.

For studying the effect of delay in real-time collaborative editing we also require an outcome metric that quantifies group performance in terms of the quality of the document and a process metric for quantifying the compensatory communication in response to limitations of the collaboration technology. In what follows we review quantitative research related to these two metrics and we define our research questions.

Need for an Outcome Metric Olson et al. (1993) exemplifies the need for an outcome metric to evaluate the quality of the work produced with groupware. While Erkens et al. (2005) developed an outcome metric for prose quality, it was not sensitive to the experimental manipulations, permitting conclusions only about post hoc covariates and process. Some researchers such as Birnholtz et al. (2013) focus only on process measures, with unclear implications for outcome.

Candidate metrics for text quality appear in research that investigates writing skill apart from the technology. The skill models resulting from this research identify the facets of composition at multiple levels of analysis, including goals, processes and cognitive demand (Hayes 2012). Two points from Hayes (2012) concern us here. First, he notes the cognitive demand of transcription, including spelling and writing. Second, he relies on quantified topics to score written essays. In fact topics and topic transitions define different levels of writing proficiency.

Latent semantic analysis (LSA) by Landauer and Dumais (1997) provides methods for comparing the content of two documents. LSA uses a reference lexicon to describe the frequency of lexical terms in a document. Similar to factor analysis, an approximation of the full frequency matrix merges similar terms and represents the document in question, which among other applications, permits comparison against other documents. One of the limitations of LSA is that it does not account for grammar or word order. However, emphasis on lexical items in Landauer and Dumais (1997) converges with emphasis on topic in Hayes (2012) rather than detailed propositional analysis.

Our research questions concerning the outcome metrics related to document quality follow:

RQ1 How does delay influence the quality of the final document in terms of the number of grammar errors?

RQ2 How does delay influence the quality of the final document in terms of the amount of redundancy?

RQ3 How does delay influence the quality of the final document in terms of the number of keywords from the transcript?

These research questions assume an independence of quality metrics. This is not necessarily the case. For example, a redundant text with increased length might very well be responsible for an increase in grammatical errors, as participants become unable to monitor and correct each other.

Compensatory Communication Communication provides a backup for local uncertainties in the coordination of coupled tasks (Tremblay et al. 2012). Participants might discuss alternatives to articulating the task, or manage their own communication. Collaborative editing often includes a chat capability, allowing

direct type-written communication between team members. Whether or not chat is task-related, chat establishes links between team members and tells the group something about what the writer is doing, even if only to indicate that the writer is not doing task-related work. An analysis of chat language can inform us about the collaborative process.

Other work has examined chat content during collaborative editing. For example, Birnholtz et al. (2013) used a categorization scheme driven by an interest in group sentiment. While some of our dimensions are similar, we chose individual lexical measures that are inspired by psycholinguistics, and avoid the challenge of quantifying units of analysis (such as phrases or sentences), multiple categorizations of the same phrase and inter-rater reliability. Our approach has precedence. For example, Gibson (2010) examined the prevalence of first words in the examination of turn taking behavior in face-to-face interaction. Both Birnholtz et al. (2013) and Gibson (2010) base their analyses on Schiffrin (1987), who also relies on single word measures.

Similar to Birnholtz et al. (2013), we examined accord language (both agreement and objection terminology) as an attempt to manage both the dialogue between participants as well as the task itself. In spoken language, accord participates in dialogue management by providing feedback to the present speaker, and controlling turn-taking (Clark and Krych 2004). Accord language also facilitates the management of the task, including transitions within and between sub-tasks (Bangerter and Clark 2003).

Classic work on the English language suggests that orderly discourse attaches new information to understood (given) information (Chafe 1976). Drijkoningen and Kampers-Manhe (2012), Dimroth et al. (2010) confirm that French, the language used in our empirical work, respects the given-new convention. Clark and Haviland (1977) demonstrated that language produced in this way facilitates comprehension, by allowing recipients to attach new information to previously activated old information. Thus, we say "*Pierre ate a banana*" to introduce the idea that Pierre ate an unspecified banana and "*Pierre ate the banana*" to describe what happened to a previously identified, specific banana. In the first case the referent is indefinite, and in the second case the referent is definite. As the previous example illustrates, one device for marking given and new information is the determiner. The definite determiner "*the*" appears with an established, specific referent, while an indefinite determiner "*a*" appears to introduce a new referent. Participants may be introducing new information in an orderly fashion, or they are aware of what others are doing and know. In either case, respect for the given-new convention suggests the presence of shared context and common ground.

Our research question regarding the effect of delay on compensatory communication follows:

RQ4 How does delay influence the use of accord language and definite determiners as an indication of common ground in the chat?

Delay and compensatory communication may interact with a third variable, collaborative experience. Experienced users may make assumptions regarding tool functionality. Delay could catch them by surprise. Inexperienced users may be

more unsure about how the process works. Therefore our final research question follows:

RQ5 How do delay, experience, and compensatory collaboration effort interact to affect task performance?

Experimental Editing Task

We selected a collaborative note taking task where delay is likely problematic. A group of four participants had 15 min to: (i) listen to a 12 min audio taped interview (ii) take notes for assigned topics (iii) consolidate the notes to reduce redundancy (iv) eliminate grammatical and spelling errors.

The task has at least several methodological advantages. First, the source interview provides a content-based performance standard. A performance standard is more challenging for more open-ended collaborative editing tasks. Second, the distribution of task assignments promoted interactivity and dependency. Third, the note taking task loads on transcription, with known cognitive demand. This task should therefore bound the tolerance for delay in collaborative editing.

In order to test the effect of delay we introduced artificial delays between a user's modification and its appearance to other users. In order to determine potentially effective values of delay we performed some measurements using simulations with GoogleDocs. We used Selenium WebDriver Java 2.44.02 to simulate users that type simultaneously on the same shared document with different typing speeds. One user simulated a reader that only reads the document. Another user simulated a writer that writes special strings that the reader will read. Other users were simulated as dummy writers that write some non-meaningful text. The role of dummy writers was to simulate concurrent access to the document. Writers can insert or delete text. We measured the delay as the difference of time between the writer inserts a particular string and the time when the reader reads this string. To eliminate clock synchronisation issues, both writer and reader were executed on a same computer. We analysed how the delay depends on the number of users and the typing speed of the users. We varied the number of users from 1 to 50 (the maximum number of users that can simultaneously edit a document in GoogleDocs) and the frequency speed from 1 to 10 characters/s.

Karat et al. (1999) mentions that the average rate for transcription is 33 words per minute (wpm), and 19 words per minute for composition. The same study reports that, when the group was divided into "fast", "moderate" and "slow" groups, the average speeds were 40, 35, and 23 wpm respectively. As the common conversion factor between words per minute and characters per minute is 5, we chose to report here on an average frequency of typing of 2 characters/s. Figure 1 shows the results we obtained for repeated measures of the delay in terms of the number of users that concurrently modify the document with a typing speed of 2 characters/s. For this typing speed, when the number of clients exceeded 38 we

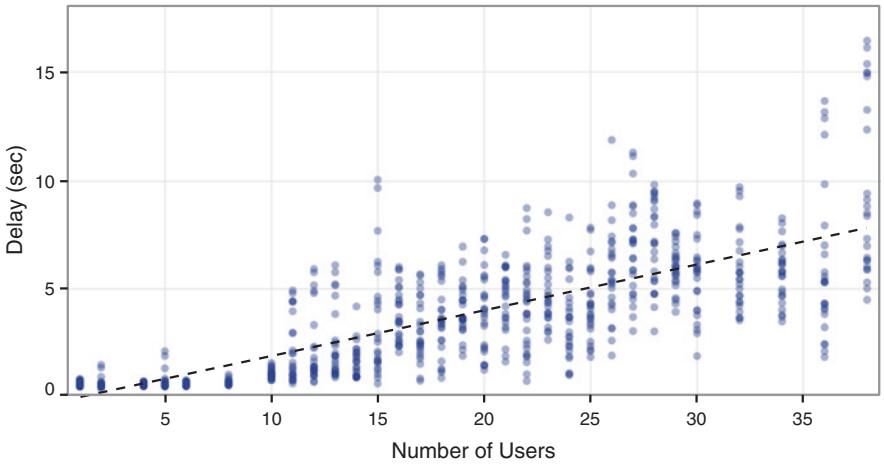

Fig. 1 Delay measurements in GoogleDocs according to the number of clients for a typing speed of 2 characters/s

faced substantial client disconnections and the collaborative editing process was stopped. We therefore reported for the results obtained for number of clients varying from 1 to 38.

Even a small number of users incurs delay between 1 and 2 s. Moreover, 4 s delays are a common result for more than 11 users. Based on this analysis we introduced artificial delays of 0, 4, 6, 8 and 10 s into the experimental design.

Methods

Participants

Eighty students affiliated with a French university participated in this experiment, in mixed gender groups of 4. Due to a change in instructions (see below) we dropped three groups of initial participants. The analysis below includes the remaining sixty eight participants.

The participants ranged in age from 21 to 27. All participants used French in their daily activities. An electronic announcement solicited participation. One of the researchers organized interested participants into sets of 4 and scheduled the session. All participants received a 10 Euro gift certificate for their participation.

Apparatus

The experiment was conducted using four GNU/Linux desktop computers in a classroom setting. Participants were separated by partitions and could not directly

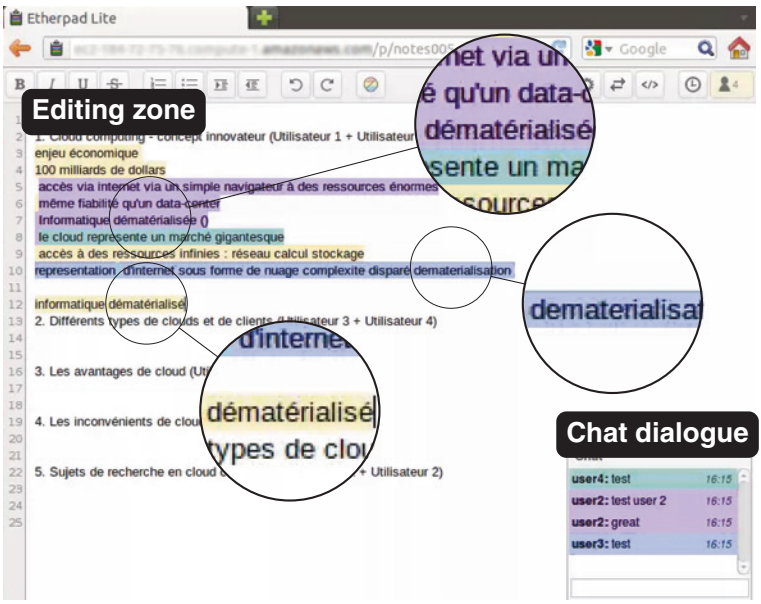

Fig. 2 Etherpad editor—each modification is highlighted with a color corresponding to the user who performed it

observe other team members while they worked, although typing activity was audible. The server running the Etherpad application was hosted on an Amazon Elastic Compute Cloud (EC2) instance located in the US East (Northern Virginia) Region. Each desktop ran the Mozilla Firefox web browser executing the Etherpad web client application. Etherpad hosted the task stimuli and a Chat dialogue facility (see Fig. 2). User operations appeared color-coded in both the text and chat. Etherpad relies on a client-server architecture where each client/user edits a copy of the shared document. When a user performed a modification it was immediately displayed on the local copy of the document and then sent to the server. The server merged the change received from the user with other user changes and then transmitted the updates to the other users. When a user edited a sequence of characters, the first change on the character was immediately sent to the server, while the other changes were sent at once only upon reception of an acknowledgement from the server. With each change sent to the server, it created a new version of the document. Gstreamer software enabled the video recording of user activity. We also instrumented Etherpad to register all user keyboard inputs on the client side and to introduce delays on the server-side. The editor window displayed 50 lines of text. Users editing above the field of view of a collaborator could cause the lines within the collaborators' view to "jump" inexplicably. Such a property is consistent with the inability to view an entire document as it undergoes modification from multiple team members.

Task and Stimuli

Participants listened to a 12 min, 1862 word interview on cloud computing,[1] divisible into five main sections.

Procedure

The entire procedure was approved by a US University institutional review board. Participants began the session with informed consent for three different experimental tasks and a survey conducted in the same sequence:

- A proofreading task, in which participants corrected a short text, containing several grammatical and spelling errors
- A sorting task, in which participants located the release dates of an alphabetized list of movies, and sorted them accordingly and
- A note taking task, in which participants listened to a 12 min interview on the topic of cloud computing, and provided an integrated set of notes on the interview
- All participants completed a follow-up questionnaire at the completion of the three task series.

The second task was analysed by Ignat et al. (2014). The task that we present in this paper is the third task, on note taking. Scripted instructions (translated here into English) for this task follow: "*Researchers will provide you and your team with an audio lecture. Your task is to take notes on this lecture using the editing tool and assemble a unified report for your team. After the end of the audio you will have three additional minutes. Please work as accurately as you can while still being efficient. You are free to coordinate your efforts with your team mates as you like at the beginning and throughout the task, using the chat interface at the right side of the screen.*" The task took 15 min. In the first 12 min participants listened to the audio tape and took notes on the shared document by using the Etherpad editor. After the end of the audio, participants were allowed 3 additional minutes to revise their notes and generate a reconciled summary of their notes by continuing to use the collaborative editor.

Initial review of data from the first three groups showed that participants took their own notes separately for the whole interview. That is, the shared document was not structured according to the main parts of the audio interview. This resulted in redundancy across the entire audio content, which was replicated four times. Furthermore, the quality of the summary suffered. As each participant wrote only the most important information, significant detail was missing.

[1]The 12 min interview is available online at the following url: https://interstices.info/jcms/i_60795/calculer-dans-les-nuages.

We decided to drop the first three groups of participants and change the instructions. We divided the shared document into five sections corresponding to the five main parts of the audio interview. For each section of the document two participants were assigned the role of taking notes of the main content of the corresponding audio part. The other two participants were assigned the role of revising the notes taken by the first two participants. The roles were inverted for each section of the document. This is consistent with real world collaborative note taking tasks during meetings where the discussion subtopics are usually known before the meeting. We added the following phrases to the above presented instructions for the task: "*In order to help you coordinate on this task we divided the document into five sections corresponding to the five main parts of the audio lecture. For each section we assigned two among you to take the main role on taking notes. These two participants are identified by their identity* (*User1*, *User2*, *User3 or User4*) *right after the title of each section. Each participant knows his/her identity from the previous tasks. The other two participants not mentioned after the section title have the role of revising the notes taken for that section. Your roles turn for each section.*"

Design

The note taking task was conducted with teams of 4 participants for each level of the continuous independent variable Delay, tested at 0, 4, 6, 8 and 10 s in addition to the 100 ms delay inherent in the EC2. Three teams experienced 0 s condition (i.e. no delay was introduced), three teams experienced 4 s delay condition, four teams experienced 6 s delay condition, three teams experienced 8 s delay condition and four teams experienced 10 s delay condition. While participants viewed their own document changes in real-time, they viewed other participants' changes according to delay condition. Chat was implemented in real time for all conditions. Delay conditions were tested in random order, and all groups experienced a single level of delay across the three-task session.

Dependent Measures

Number of words is computed by the number of words in the text base. For each group in the experiment we examined recorded versions of the shared document at every minute. For each document version we computed its total number of words by using a script written in Python.

Keywords is one measure of document content and quality. Keywords is computed as the number of main keywords present in the final version of the document provided by each group of users. We identified 121 keywords or short phrases distributed over the document sections. Keywords included nouns, (e.g., "*services*",

"*clients*"), verbs, (e.g., "*payer*", "*consommer*") and adjectives, e.g., ("*cohérence*"). Crucially, we included misspellings as corresponding to the presence of keywords. For each section in the final document we automatically identified the number of keywords corresponding for that section as present or missing. We examined the number of keywords divided by the number of words as well as an arcsin transformation of this ratio measure. These give consistent results and we report only the transformed metric here.

Redundancy is a second measure of document content and quality. Redundancy is computed as the sum of redundancies of each section in the document. Redundancy of a section was measured by analysing the recorded videos of the collaborative editing session. Redundancy of a section represents the maximum number of occurrences in that section of any topic present in the audio. The topic contained one or more keywords belonging to that section. This measurement was performed on the document version that corresponded to the end of that section in the audio. Redundancy of a section can be equal to 0, 1, 2, 3 or 4, as a topic can be replicated maximum 4 times corresponding to the maximum number of participants. The redundancy sum at 12 min corresponds to the end of the audio, prior to the 3 min proof-reading opportunity. A binary redundancy metric captured the redundancy that remained at the end of the proofreading period, that is, whether the redundancy was caught and repaired. For example, we can notice in Fig. 2 that redundancy of Sect. 1 of the document is 3 as three users marked down the idea of "*dematerialised computer science*": two users wrote "*informatique dématérialisé*", while the third one wrote "*dématerialisation*" as shown by the three zoomed zones depicted in the figure. We therefore find a triple repetition of the base of the keyword "*dematerialisation*".

Error Rate serves as a third measure of document quality. Error rate is computed using *Reverso* tool.[2] Reverso checks misspellings and grammar of a text in any language. For each group we generated the versions of the shared document at every minute during the experiment and we computed the number of errors for every such version by using reverso. The number of errors was computed automatically using a script written in Python. We examined the number of errors divided by the number of words as well as an arcsin transformation of this ratio measure. These give consistent results and we report only the transformed metric here.

We also examined covariates obtained from the chat behavior and survey responses.

Chat Behavior included the number of words, accord language, and definite determiners. For accord language we tallied all versions of "*oui*" ("*ouai*", "*ouis*", "*ouaip*"), negation ("*ne*", "*not*", "*naan*"), "*OK*" ("*ok*", "*k*", "*d'accord*") and objection ("*sinon*", "*objection*", "*contre*"). We did not tally the words paired with "*ne*" such as "*pas*", "*rien*" etc., to avoid double counting. We did not tally "*si*", which is a version of "*yes*" used in response to negation, because it also means "*if*". For definite determiners we tallied "*le*", "*la*", "*les*", "*au*" and "*aux*", but adjusted the count of "*la*" to exclude the case of "*de la*".

[2]Reverso tool is available on-line at http://www.reverso.net/.

Survey responses examined here include: (a) Which exercise did you find most difficult? Why? (b) Did anything annoy you about the text editor? If, yes, why? (c) What was the impact of the collaborative editing tool for note taking task? (Using a 10 point Likert scale) Explain. (d) Have you previously used collaborative tools? We split the groups by the consistency of experience. In the high experience groups, all members had previous collaborative editing experience. In the low experience groups, one or more members lacked collaborative editing experience.

Results

We used regression modeling to describe the quantitative consequences of delay condition to performance measures. We show the consequences of delay to document content and errors, and suggest the role of document redundancy as a mediator of these relationships. Subsequent analysis of redundancy shows that the more experienced groups manage redundancy less purposefully and are hence subject to the effect of delay. Low experienced groups attempt to manage redundancy as revealed by chat metrics for common ground.

Performance Measures

We examined both document content and errors as performance measures. For the purposes of contrast, we also examine subjective ratings.

Document Content The text base is larger for the high delay groups at 15 min, $F(1, 15) = 5.198$, $p = 0.0377$, $\beta = 0.5073$, adjusted $R^2 = 0.2078$. We characterized document quality as the ratio of keywords to number of words in the text base (or version of the shared document) at 15 min. Proportion of keywords is negatively related to delay condition, $F(1, 15) = 7.8610$, $p = 0.0134$, $\beta = -0.5864$, adjusted $R^2 = 0.3001$. Quality content decreases with delay condition. Finally, document redundancy at 12 min is a function of delay condition, $F(1, 15) = 14.66$, $p = 0.0016$, $\beta = 0.7030$, adjusted $R^2 = 0.4605$. Figure 3 illustrates the relationship between delay condition and proportion of keywords and word count. In summary, delay increases the text base, decreases the proportion of keywords and increases the redundancy.

Error Proportions at 15 min Error rate is a function of condition, $F(1, 15) = 15.94$, $p = 0.0012$, $\beta = 0.7178$, adjusted $R^2 = 0.4829$. The error proportion metric is negatively correlated with the proportion of keywords, $F(1, 15) = 26.98$, $p = 0.0001$, $\beta = -0.8017$, adjusted $R^2 = 0.6188$.

Redundancy and error rate are correlated, $F(1, 15) = 27.17$, $p = 0.0001$, $\beta = -0.8027$, adjusted $R^2 = 0.6206$. Figure 4 illustrates this relationship between

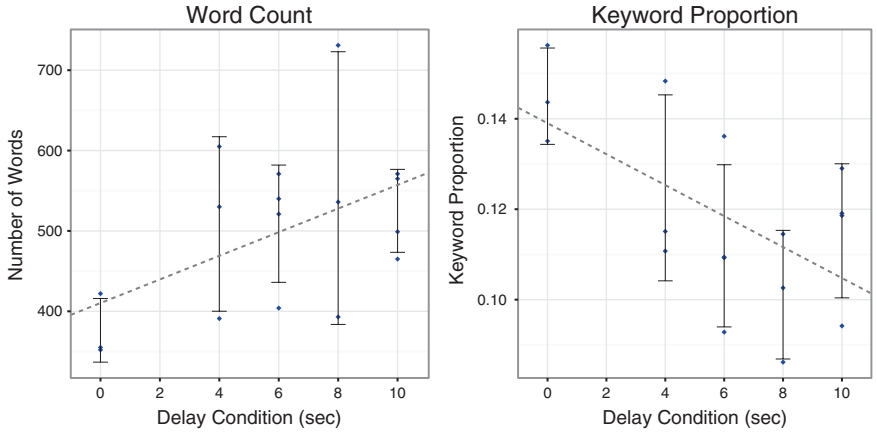

Fig. 3 Number of words (*left*) and proportion of keywords (*right*) as a function of delay condition

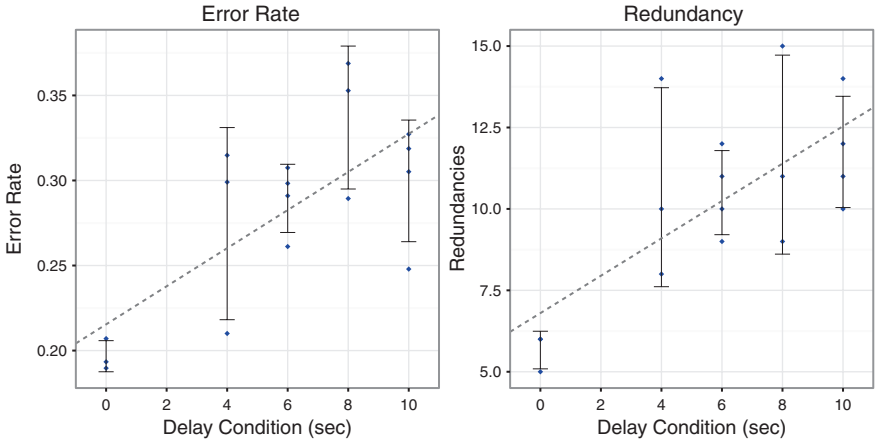

Fig. 4 Error rate (*left*) and redundancy (*right*) as a function of delay condition

delay condition and error rate and redundancy. Thus, error rates, like document content measures, appear sensitive to delay condition.

Subjective Difficulty Ratings Editor difficulty ratings are not related to delay condition $F(1, 15) = 3.487$, $p = 0.0815$. Editor difficulty ratings do not correlate with any of the performance measures: Error rate, $F(1, 15) = 1.87$, $p = 0.1916$, Redundancy at 12 min, $F(1, 15) = 0.1343$, $p = 0.7191$, Proportion of keywords $F(1, 15) = 0.377$, $p = 0.5484$ and Word count $F(1, 15) = 0.0067$, $p = 0.9359$.

Mediation Analyses

A model of grammatical error rate with both delay and redundancy suppresses the relationship between delay condition and error rate. A corresponding graphic for this mediation analysis and beta weights appear in Fig. 5.

	R	*Adj.* R^2	β
Analysis 1: error rate = delay condition	0.6949	0.4829	
Delay condition			0.7178**
Analysis 2: redundancy = delay condition	0.6786	0.4605	
Delay condition			0.7030**
Analysis 3: error rate = redundancy	0.7878	0.6206	
Redundancy			0.8027***
Analysis 4: error rate = delay condition + redundancy	0.8042	0.6467	
Delay condition			0.3035
Redundancy			0.5893*

* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

A similar mediation analysis suggests a similar, albeit non significant mediation of redundancy, disrupting the relationship between delay condition and proportion of keywords.

These suggest that managing redundancy is the process that contributes to the observed effects of delay on outcome.

Redundancy Management Analyses

We examined covariates recovered from both the questionnaire and chat behavior to better understand the factors that influence the management of redundancy. We

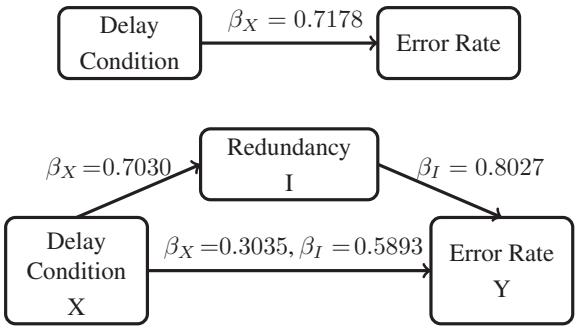

Fig. 5 Redundancy at 12 min mediates the relationship between delay condition and grammatical error rate at 15 min

divide the exposition into two sections: Redundancy awareness as indicated in the questionnaire and common ground as indicated in the chat.

Redundancy Awareness Some groups did complain about the difficulty in managing redundancy. Redundancy Awareness appears in two places in the post hoc questionnaire. Those who found the note taking task most difficult sometimes referred to redundancy. Participants also sometimes explained their ratings for the editor in terms of the ability to manage redundant text. Examples illustrating these explanations are provided in English: "*We are lacking time to organise*, *therefore we write the same things, taking notes on the same thing at the same time is generally complex when we do not know in advance what will be said*" (a group in condition 8) and "*Difficult to divide the tasks*, *we obtain a lot of redundant text (multiple participants write almost the same thing)*" (a group in condition 4).

In all groups except two, at least one group member complained about the management of redundancy. However, this awareness metric was unrelated to the measurement of redundancy at twelve minutes, $F(1, 15) = 0.1182$, $p = 0.7358$, or the resolution of redundancy $F(1, 15) = 0.1572$, $p = 0.6973$.

Experience A model of redundancy with delay, experience and delay × experience suggested an interaction between the effect of delay and experience, $t(13) = 2.287$, $p = 0.0396$. To pursue this interaction, we split the data by experience level. Groups in which all participants were experienced with collaborative editing show a persisting effect of delay $F(1, 6) = 18.1$, $p = 0.0054$, $\beta = 0.8666$, adjusted $R^2 = 0.7096$. Groups in which some of the participants were less experienced do not show the same sensitivity to delay $F(1, 7) = 1.815$, $p = 0.2199$, adjusted $R^2 = 0.09244$. Figure 6 illustrates the relation between redundancy

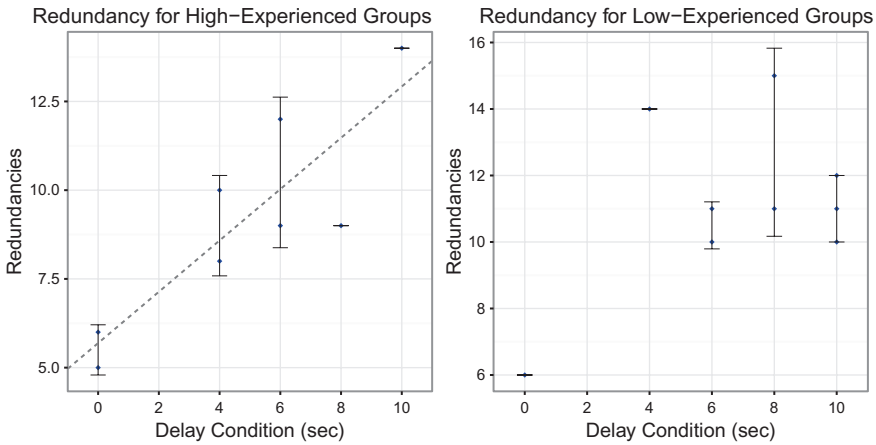

Fig. 6 Delay condition predicts redundancy for high collaborative experience (*left*) groups, but not for low experience groups (*right*)

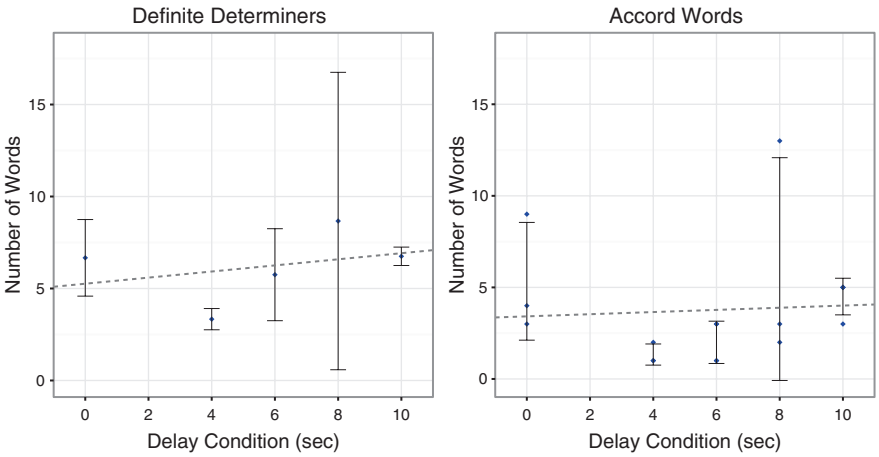

Fig. 7 Definite determiners (*left*) and agreement (*right*) by delay condition

and delay condition for groups with high and respectively low collaborative experience.

Chat Behavior The metrics for definite determiners and agreement are highly correlated $F(1, 15) = 10.09$, $p = 0.0063$, adjusted $R^2 = 0.3622$ (see Fig. 7). In the spirit of factor analysis, we added the two metrics to create an aggregated measure, Common Ground. A model of redundancy with delay and common ground reveals significant effects for both delay condition, $t(14) = 4.587$, $\beta = 0.7514$, $p = 0.0004$ and common ground $t(14) = -2.274$, $\beta = -0.3725$, $p = 0.0392$. Common ground opposes the effect of delay condition on redundancy. A model of redundancy with delay and total chat word count is not significant for total chat word count, $t(14) = -1.549$, $p = 0.1436$, confirming that the common ground findings are not an artifact of word count.

Finally, we examined the effect of common ground and delay condition for both the high and low experience groups. In general, the average amount of common ground behavior does not differ between the high and low experience groups ($M = 10.25$ words, $SE = 2.93$; $M = 9.78$ words, $SE = 1.20$). However, for the high experience groups, a model of redundancy with delay and common ground reveals that delay condition is significant $t(5) = 5.307$, $p = 0.0032$ but common ground is not $t(5) = -1.811$, $p = 0.1300$. In contrast for the low experience group, delay misses significance $t(6) = 2.336$, $p = 0.0582$ but common ground is significant $t(6) = -5.142$, $p = 0.0021$. The negative value suggests that common ground comprises an effort to decrease redundancy. Thus, the general effect of common ground suggested by the overall analysis is localized to the low experience groups. High experience groups do not use common ground in the same way.

Discussion

We examined the effect of delay on collaborative note-taking task, using levels of realistic delay consistent with an independent simulation. The collaborative note-taking task creates dependency and interactivity in collaborative editing, and permits the measurement of task outcome with reference to transcribed audio tape. Here we return to our original questions regarding the relationship between delay and collaboration effort before turning to implications for design.

Delay

We demonstrated a general hinderance of delay on four performance measures, such that delay increases grammatical errors and redundancy and decreases the proportion of key words relative to the text base. We also showed that text redundancy mediates the relationship between delay condition and grammatical errors. Given the increase in word count that we also observed, the effect of delay is to increase redundancy and create a larger, more erroneous and less manageable text base to be corrected after the audio tape is completed. A similar, albeit non-significant relationship between redundancy, delay condition and proportion of key words is consistent with the role of redundancy as a mediator between delay and performance.

We suggest that delay interferes with the ability to monitor team members' activities and adjust ones behavior accordingly. In effect, delay forces independent, redundant work.

Delay, Collaboration Effort and Experience

A complete account of the delay effect must also consider the effect of collaboration effort and experience. In general common ground opposes the effect of delay on redundancy. However experience interacted with delay, such that redundancy increased with delay for the high experience groups, but not the low experience groups. While both groups appeared to exercise the same amount of collaboration effort overall, the low experience groups adjusted their collaboration effort to manage the redundancy, while the high experience groups did not. Thus, the high experience groups appear to be caught off-guard when the editor did not operate as expected, and they did not attempt compensatory collaboration effort by means of communication through chat. This is the case although the task in question was third in a three-task series, and thus participants had previous opportunity to discover the delay and adjust accordingly.

Implications for Design

As our primary purpose was to demonstrate the effect of delay in collaborative editing, and motivate continuing work on the optimization of collaborative editing algorithms, we chose delay values that were highly likely to disrupt performance. This work is the first to study effect of delay in collaborative note taking. Results of both the simulation and experimental studies suggest refinement of the limits of delay in the range of 0–4 s in order to analyse the limit of user tolerance to delay. Testing small levels of delay will establish the shape of the delay-performance function. This function needs not be linear, and knowledge of a critical point will help further constrain design.

The study we performed shows that reducing delay influences the efficiency of the group and the quality of note taking. This finding is important because the choice of the underlying architecture of the collaborative editor has an impact on the delay or feedthrough time, which measures the time from a user performing an action to other users seeing the result (Graham et al. 2006). An architecture with a thick client, where computations are executed on the client side rather than on the server side is more suitable for minimizing feedthrough time. Popular collaborative editing systems such as GoogleDrive, relying on the Jupiter algorithm for synchronizing changes, do not rely on a thick client architecture. As a result, transformations among operations necessary for synchronizing changes are performed not only on the client side but also on the server side. Executing transformations on the server side introduces additional delay. Choosing synchronisation algorithms where computation is done only on the client side such as SOCT2 by Suleiman et al. (1998) or LogootSplit by André et al. (2013) reduces the delay. Among these families of algorithms, CRDTs support a better scalability in terms of number of users and feature better time complexities for integration of remote operations (Ahmed-Nacer et al. 2011).

Conclusions

In this study we evaluated the performance consequences of simulated network delay in real-time collaborative note taking. We designed an artificial note taking task where groups of four participants must take notes on an audio lecture and revise their notes in a limited period of time. Results of our study show that the general effect of delay on this task is to encourage independent work. We showed that delay increases grammatical errors and redundancy, resulting in a decreased quality of the task content. Measures of accompanying chat indicate that less experienced groups attempt to compensate for the effect of delay. In contrast, more experienced groups do not adjust their communication in response to delay, and their performance remains sensitive to the delay manipulation.

To our knowledge this study is the first to evaluate the effect of delay on performance of collaborative note taking and efficacy of compensatory strategies on this task. This initial study is fundamental for the refinement of the limits of tolerable delay in real-time collaborative editing. Establishing the shape of the delay performance function places fundamental constraints on the choice of collaborative editing architecture and underlying synchronisation mechanisms.

Acknowledgments The authors are grateful for financial support of the USCoast Inria associate team, the Inria internships programme and the research program ANR-10-SEGI-010 and for sabbatical support from the Department of Psychology, Wright State University. The authors thank Vinh Quang Dang for his help on simulations with GoogleDocs.

References

Ahmed-Nacer, M., Ignat, C.-L., Oster, G., Roh, H.-G., & Urso, P. (2011). Evaluating CRDTs for real-time document editing. In: *Proceedings of the Eleventh ACM Symposium on Document Engineering—DocEng 2011* (pp. 103–112). Mountain View, CA, USA: ACM Press.

André, L., Martin, S., Oster, G., & Ignat. C.-L. (2013). Supporting adaptable granularity of changes for massive-scale collaborative editing. In: *International Conference on Collaborative Computing: Networking, Applications and Worksharing—CollaborateCom 2013* (pp. 50–59). Austin, TX, USA: IEEE Computer Society.

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science, 27*(2), 195–225.

Birnholtz, J. P., Steinhardt, S. B., & Pavese, A. (2013). Write here, write now!: An experimental study of group maintenance in collaborative writing. In: *ACM SIGCHI Conference on Human Factors in Computing Systems—CHI 2013* (pp. 961–970). Paris, France: ACM.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and points of view. In: *Subject and Topic* (pp. 25–55). USA: Academic Press.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In: *Discourse Production and Comprehension* (pp. 1–40). New York: Ablex Publishing.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addresses for understanding. *Journal of Memory and Language, 50*, 62–81.

Dimroth, C., Andorno, C., Benazzo, S., & Verhagen, J. (2010). Given claims about new topics: How romance and germanic speakers link changed and maintained information in narrative discourse. *Journal of Pragmatics, 42*(12), 3328–3344.

Dourish, P., & Bly, S. (1992). Portholes: Supporting awareness in a distributed work group. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI 1992* (pp. 541–547). Monterey, CA, USA: ACM.

Drijkoningen, F., & Kampers-Manhe, B. (2012). Word order in French and the influence of topic and focus. *Linguistics, 50*(1), 65–104.

Ellis, C. A., Gibbs, S. J., & Rein, G. (1991). Groupware: Some issues and experiences. *Communications of ACM, 34*(1), 39–58.

Erkens, G., Jaspers, J., Prangsma, M., & Kanselaar, G. (2005). Coordination processes in computer supported collaborative writing. *Computers in Human Behavior, 21*(3), 463–486.

Gibson, D. R. (2010). Marking the turn: Obligation, engagement, and alienation in group discussions. *Social Psychology Quarterly, 73*(2), 132–151.

Graham, T. N., Phillips, W. G., & Wolfe, C. (2006). Quality analysis of distribution architectures for synchronous groupware. In: *International Conference on Collaborative Computing: Networking, Applications and Worksharing—CollaborateCom2006.* Atlanta, GA, USA: IEEE Computer Society.

Gutwin, C. (2001). The effects of network delays on group work in real-time groupware. In: *Proceedings of the Seventh Conference on European Conference on Computer Supported Cooperative Work—ECSCW 2001* (pp. 299–318). Bonn, Germany: Kluwer Academic Publishers.

Gutwin, C., Benford, S., Dyck, J., Fraser, M., Vaghi, I., & Greenhalgh, C. (2004). Revealing delay in collaborative environments. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI 2004* (pp. 503–510). Vienna, Austria: ACM.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388.

Ignat, C.-L., & Norrie, M. C. (2008). Multi-level editing of hierarchical documents. *Journal of Computer Supported Cooperative Work, 17*(5–6), 423–468.

Ignat, C.-L., Oster, G., Newman, M., Shalin, V., & Charoy, F. (2014). Studying the effect of delay on group performance in collaborative editing. In: *Proceedings of International Conference on Cooperative Design*, *Visualization and Engineering—CDVE 2014* (pp. 191–198). Mallorca, Spain: Springer International Publishing.

Karat, C.-M., Halverson, C., Horn, D., & Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI 1999* (pp. 568–575). Pittsburgh, PA, USA: ACM.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Noël, S., & Robert, J.-M. (2004). Empirical study on collaborative writing: What do co-authors do, use, and like? *Computer Supported Cooperative Work, 13*(1), 63–89.

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction, 15*(2), 139–178.

Olson, J. S., Olson, G. M., Storrøsten, M., & Carter, M. (1993). Groupwork close up: A comparison of the group design process with and without a simple group editor. *ACM Transactions on Information Systems, 11*(4), 321–348.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Suleiman, M., Cart, M., & Ferrié. J. (1998). Concurrent operations in a distributed and mobile collaborative environment. In: *Proceedings of the International Conference on Data Engineering—ICDE 1998* (pp. 36–45). Orlando, FL, USA: IEEE Computer Society.

Sun, C., Jia, X., Zhang, Y., Yang, Y., & Chen, D. (1998). Achieving convergence, causality preservation, and intention preservation in real-time cooperative editing systems. *ACM Transactions on Computer-Human Interaction, 5*(1), 63–108.

Tammaro, S. G., Mosier, J. N., Goodwin, N. C., & Spitz, G. (1997). Collaborative writing is hard to support: A field study of collaborative writing. *Computer-Supported Cooperative Work, 6*(1), 19–51.

Tremblay, S., Vachon, F., Lafond, D., & Kramer, C. (2012). Dealing with task interruptions in complex dynamic environments: Are two heads better than one? *Human Factors, 54*(1), 70–83.

Vaghi, I., Greenhalgh, C., & Benford, S. (1999). Coping with inconsistency due to network delays in collaborative virtual environments. In: *Proceedings of the ACM symposium on Virtual reality software and technology—VRST 1999* (pp. 42–49). London, United Kingdom: ACM.