
Conversation Map: An Interface for Very-Large-Scale Conversations

WARREN SACK

WARREN SACK is Assistant Professor at the University of California, Berkeley, School of Information Management and Systems. Before coming to Berkeley, he was a research scientist at the Massachusetts Institute of Technology (MIT) Media Laboratory. He received a B.A. in Computer Science and Psychology from Yale University and an S.M. and Ph.D. in Media Arts and Sciences from MIT. His research interests include computer-mediated communication, online communities, architecture and design for online spaces, social networks, computational linguistics, and media studies.

ABSTRACT: Very large-scale conversation (VLSC) involves the exchange of thousands of electronic mail (e-mail) messages among hundreds or thousands of people. Usenet newsgroups are good examples (but not the only examples) of online sites where VLSCs take place. To facilitate understanding of the social and semantic structure of VLSCs, two tools from the social sciences—social networks and semantic networks—have been extended for the purposes of interface design. As interface devices, social and semantic networks need to be flexible, layered representations that are useful as a means for summarizing, exploring, and cross-indexing the large volumes of messages that constitute the archives of VLSCs. This paper discusses the design criteria necessary for transforming these social scientific representations into interface devices. The discussion is illustrated with the description of the Conversation Map system, an implemented system for browsing and navigating VLSCs.

KEY WORDS AND PHRASES: browsing, conversation analysis, discourse analysis, electronic mail, interface design, navigation, news groups, semantic networks, social networks, summarization, visualization.

THE RECENT GROWTH in the number of people with Internet access has made possible the unprecedented phenomenon of very-large-scale conversations (VLSCs) in which hundreds, even thousands, of people exchange thousands of messages in daily, often international, public, many-to-many communications. The most obvious manifestations of this phenomenon are Usenet newsgroups hosted on servers throughout the Internet and archived by a handful of large sites (e.g., www.deja.com).

Existing tools of social science (e.g., conversation analysis, discourse analysis, etc.) have been developed to analyze many-to-many, conversational interactions involving far fewer people than the number of participants now taking part in various online VLSCs. While it may be possible to scale up these tools and methodologies from linguistics, psychology, sociology, media studies, and anthropology, it is not a priori

obvious how this can be done. Consequently, the challenge before us is to use and improve upon the tools of social science to create new interfaces for VLSCs that provide a means for participants and interested observers to understand and critically reflect on them. In this paper, two tools from the social sciences—*social networks* (e.g., [46, 47]); and, *semantic networks* (e.g., [9, 35])—will be reexamined as possible interface devices for a VLSC browser.

In the social sciences, social and semantic networks can be used as representations of observed or hypothesized phenomena, that is, as scientific models. They are, in short, proposed as answers to outstanding scientific questions. For instance, in cognitive psychology, linguistics, and artificial intelligence, semantic networks have been understood to be an answer to the question, “What constitutes a reasonable view of how semantic information is organized within a person’s memory?” [35]. Within sociology, social networks are used to model and summarize empirical studies of interactions and relationships between people, groups, and institutions. In addition to these scientific uses, I propose that social and semantic networks be employed as interface devices, that is, as graphical devices for summarizing and browsing through large archives of e-mail collected over the course of a VLSC.

The more theoretical discussion of this paper is grounded with the presentation of an implemented VLSC browser, the *Conversation Map* system. In principle, the system can be used just like a conventional news or mail reader (e.g., Eudora, Netscape Messenger, RN, etc.). But the Conversation Map system offers several summaries of newsgroup messages that standard browsers do not provide. It performs a series of combined computational linguistic and sociological analyses on the messages of a newsgroup and presents the results of these analyses in the form of a graphical interface. The graphical interface can be used to browse the messages and explore connections between the messages.

After the system has been run on an archive of a few thousand messages from a newsgroup, one can use it to get an idea of who is “talking” with whom, the central players in the newsgroup, the important themes of discussion, and some of the emergent metaphors or definitions that might be peculiar to the language exchanged in the messages of the newsgroup. This summarization of a message archive is presented graphically as a number of social and semantic networks. The user of the Conversation Map system can manipulate these graphical networks with a mouse and keyboard in order to explore the messages and a layered set of summaries of the messages.

Having a set of machine-generated summaries of the archive (i.e., the social and semantic networks and several other representations, which will be discussed) provides one with a relatively quick way to decide if a VLSC is focused on what is expected or desired. The summaries provided also give some indication of how various subjects are being discussed and who is central to the discussion. In addition, there are ways of using the interface to steer around “spam” and focus in on threads of discussion that address specific themes. In short, the social science–derived tools of the interface give one a means of seeing the “forest” of the conversation before diving into the “trees.”

Here is an outline of the remainder of this paper: First, a longer description of semantic and social networks will be undertaken to explain why they are useful as

summaries of VLSCs. Second, example semantic and social networks used as interface elements are shown. These examples are computed and diagrammed by an implemented VLSC browser—the *Conversation Map* system—that allows one to explore large archives of newsgroup messages. A description of how social and semantic networks are automatically generated, diagrammed, and incorporated into the Conversation Map system is given. After some related work is reviewed, the paper concludes with a discussion of several design criteria for the use of social scientific representations as interface devices.

Summarizing Very-Large-Scale Conversations

TO EXPLAIN WHY A COMBINATION of social and semantic networks might serve as a summary of a VLSC, it is helpful to examine the claims of the linguist Michael Halliday. According to Halliday [19], language has at least three meta-functions: (1) *ideational*: language can represent ideas; (2) *interpersonal*: language functions as a medium of exchange between people; and (3) *textual*: language functions to organize, structure, and hold itself together; this function allows the various devices of cohesion, including citation, ellipsis, anaphoric reference, etc., to be used. Thus, for example, I can write here, in this present sentence, about the first sentence of the present paragraph and the reader can infer that I am referring to the sentence that begins like this: “To explain why . . .”

Any summary of a VLSC is incomplete if it does not incorporate all three of these meta-functions (ideational, interpersonal, and textual). I claim that a suitably improved implementation of social networks can represent, even if only very roughly, the interpersonal and textual aspects of a VLSC, and that semantic networks can be an approximation of the ideational content of a VLSC.

Another way of explaining the usefulness of social and semantic networks as a description of VLSCs is in the terms of Paul Dourish and Matthew Chalmers, researchers in human–computer interaction and computer-supported cooperative work. Dourish and Chalmers claim there are at least three forms of navigation mechanism that can be combined in information systems: (1) spatial, (2) semantic, and (3) social navigation mechanisms [11]. Most graphical interfaces make use of spatial layout (and thus facilitate spatial navigation); some use semantic navigation (e.g., hypertexts); and some use social navigation (e.g., a variety of work in the sociology of social networks). However, very few interfaces combine all three sorts of navigation. The Conversation Map system provides the means to spatially navigate through social networks, semantic networks, and intersections of the two, thus implementing all three of Dourish and Chalmers’s forms of navigation.

Social Networks

One of the “results” of a VLSC is a social network. After a while, participants in an online discussion get to know one another and exhibit characteristic patterns of interaction with one another. Some participants post messages that often strike others as

interesting, evocative, or worthy of a reply, and so these people tend to garner more responses to their messages than others do. Some participants take pains to reply to the posts of newcomers and regulars alike and so build a reputation as virtual moderators for the discussion (even in groups with no officially designated moderators). Some people post what others consider to be “spam” and so, even though they may tend to post frequently, they are ignored (i.e., their posts do not earn replies from other participants). These collections of posting profiles for the participants of a group, when graphically assembled together, constitute a diagram that would be recognizable to social scientists as a social network.

The simplest social networks possible for VLSCs are those that can be computed through an analysis of references between messages—that is, an analysis of the “threading” of the messages. So, if participant A responds to a message posted by participant B, then a link can be drawn between A and B. Threading is straightforward to compute given the format of Usenet newsgroup messages, and the information computed can be graphically presented. An example of this can be found in [10]: Donath et al. describe a well-designed graphical presentation of Usenet newsgroup threading information in a system called *Loom*.

Obviously it would be useful to know something more than simply the number of times any two given participants exchanged messages. It would be nice to know, for instance, in their responses to each other whose messages were cited or quoted by whom. This sort of analysis has been extremely useful in science studies and is generally referred to as *citation indexing* [17]. However, while instances of citation are relatively easy to spot automatically in scientific papers because citations are required to appear in a standard format and must be listed in the references at the end of a paper, this is not the case for e-mail messages. In e-mail messages, citations often occur without mention of the original author, citations are often nested inside one another, and citations do not have a standard format, even though it is quite common (but not required) to precede each line of a quotation with some punctuation, like this:

```
>>> On 31 February 2003 Warren Sack
>>> <sack@sims.berkeley.edu> wrote:
>>> Hi guys! The future is really great!
```

These complicating factors make an automatic citation analysis procedure difficult to implement. An analysis procedure of this sort has been implemented in the Conversation Map system. (See Figure 1.) By automatically identifying who has either responded to or quoted from whom, the Conversation Map system calculates a social network given an archive of e-mail messages. The nodes in the network represent people—that is, participants in the online discussion—and the links represent *reciprocating* quotations or responses.

Thus, if participant A responds to or quotes a message from participant B and then, later in the discussion, participant B quotes from or responds to a message from participant A, a link is drawn between nodes labeled “A” and “B.” If A and B have reciprocated frequently, the link between them will be shorter than if they have only quoted from or responded to one another once or twice. Figure 2 shows the social



Figure 1. An example Conversation Map

network computed by the Conversation Map system after the system was run on two weeks' (16 April 1999–4 May 1999) worth of messages (over 1200 messages from about 260 participants) from a Usenet newsgroup devoted to a discussion of the situation in Kosovo (soc.culture.albanian). The archives of this newsgroup are a matter of public record and can be accessed at various publicly accessible Usenet archives, such as www.deja.com.

Note that there are certain “hubs” in the social network. These hubs represent participants who post many messages but who also receive many responses to their messages. They are virtual moderators of a sort for the newsgroup even though the group depicted has no official moderators. In Figure 2 the social network is displayed with all of the names turned off so that the overall shape of the network can be seen. If one wants to select a part of the network, a node (representing a particular participant) can be clicked on. When this is done, every other participant who has reciprocated replies or quotes with the selected participant is shown and the rest of the social network disappears.

Performing a citation analysis produces social networks (like the one shown above), but it also allows another interface feature to be implemented as well. Once quotations and their sources have been identified, messages in the archive can be hyperlinked so that clicking on a quote in a message opens a window containing the text of the cited message. Within the Conversation Map system one can move between messages by clicking on quotations.

While the metric who-cites-whom is more sensitive than the simpler metric of who-responds-to-whom, one can imagine an increasingly sophisticated series of metrics based upon more and more sensitive readings of message-to-message linkages. For

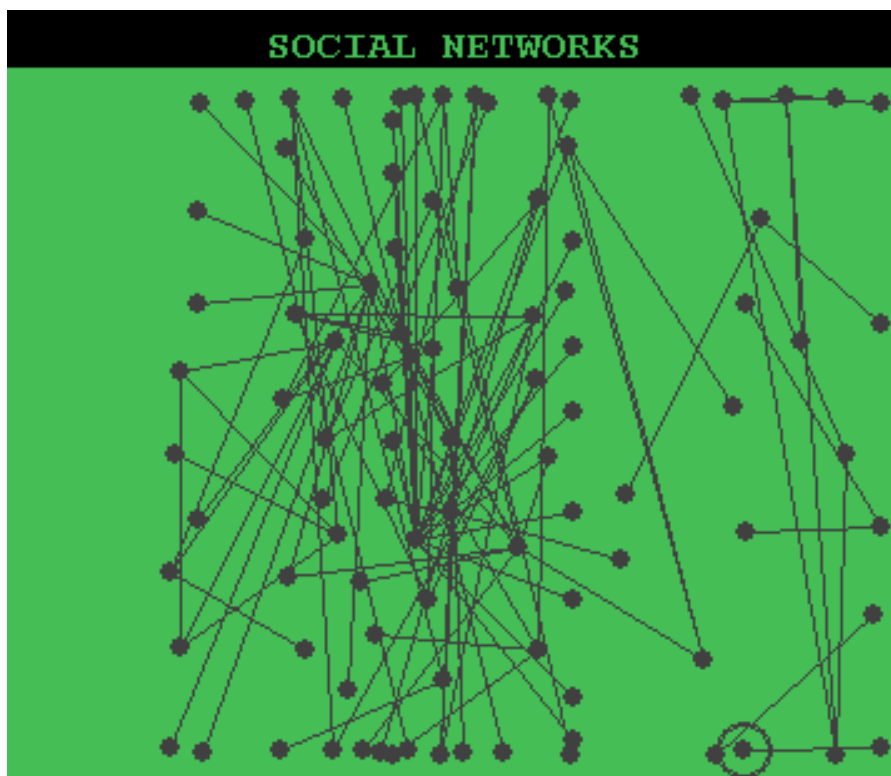


Figure 2. Social Networks Showing Who Has Reciprocally Replied to or Cited Whom

example, it is possible for me to implicitly cite the Shakespeare play about the two young lovers from feuding families without mentioning the name of the play. These sorts of more or less subtle linkages between texts and within texts are termed *ties of cohesion* in the terminology of *systemic functional linguistics* [20]. While there has been some progress in the area of computational linguistics techniques for intertextual cohesion analysis (e.g., [1]), it is necessary to merge such an analysis with a social network computation in order to be able to label the links of the computed social network with the sorts of ties that characterize the relationships manifest among participants in a VLSC.

As a first step in this direction a procedure for *social cohesion analysis* has been implemented in the Conversation Map system [39]. This procedure performs an analysis of lexical cohesion (cf., [24]) between messages and then overlays this information on top of the social network so that a rough approximation of the “theme” of the conversation that exists between posters can be detected. A step-by-step description of this procedure will be given later in the section entitled “The Computation.”

In the upper middle panel of Figure 1, the results of the cohesion analysis of the Conversation Map system can be seen: The automatic analysis of social cohesion produces a menu of “discussion themes.” If, for example, participant A mentioned the word “skiing” in a post that also quoted a part of a message from participant B wherein

B wrote about skating, and then, later in the conversation participant B wrote about cycling in response to a message by A concerning wrestling, then the link between A and B in the social network would be labeled with the term “sports,” since skiing, skating, cycling, and wrestling are all sports. This analysis requires, of course, the use of a thesaurus, specifically WordNet [15].

In the interface, when one clicks on the menu item “sports,” the link between A and B is highlighted (along with the links between any other pairs of posters who are connected through a discussion of sports). Figure 1 shows the same social network pictured in Figure 2, along with the menu of “discussion themes” (or, more accurately, the list of lexical ties) that link messages, and thus, people, together in conversation about the larger topic of Kosovo and Albanian culture in general. However, in Figure 1 most of the social network has been hidden so that the two pairs of posters who have exchanged messages concerning NATO are highlighted.

Labeling the links of a computed social network with the content of messages has some precedents in the literature of social network analysis (cf., [5]; see also, for example, [36]). However, no such computations of social cohesion have been implemented in a computer program as has been done for the Conversation Map system. The precedent research in social network analysis was performed by hand. Conversely, although procedures for automatic cohesion analysis have been implemented by researchers in computational linguistics (cf., [24, 43]), no such procedures have merged the results of a lexical cohesion analysis with a computed social network as has been done for the Conversation Map system. While the social cohesion analysis and display procedures in the Conversation Map system could be improved upon, even in their present form they effectively compute and interweave an analysis of the *interpersonal* and *textual* dimensions of the language of newsgroups (i.e., two of Halliday’s three meta-functions of language).

Semantic Networks

A second result of a VLSC is a semantic network. Over the course of many exchanges, participants in a VLSC coin new acronyms (e.g., IMHO—“in my humble opinion”), new punctuation (e.g., the (in)famous smiley faces :-)), and new or idiosyncratic ways of using existing vocabulary terms (e.g., “to flame” means something rather particular online). Semantic relations between new and existing lexical items (i.e., words and abbreviations) can be represented in a semantic network. In general, I will argue that, with a semantic network, it is possible to diagram some of the *ideational* meta-functions of a conversation (i.e., the third of Halliday’s meta-functions).

For reasons that will soon become apparent, it is possible to understand the semantic networks of a VLSC as a diagram of the emergent *metaphors* or *definitions* of participants’ discussion. To clarify this insight I review how the cognitive scientists George Lakoff and Mark Johnson [28] demonstrate the common usage of specific metaphors. Lakoff and Johnson offer the following sentences to support their claim that, in “our” (presumably U.S.) culture, the metaphor TIME IS MONEY is a common presupposition:

TIME IS MONEY:

You're wasting my time.

I don't have the time to give you.

How do you spend your time these days?

I've invested a lot of time in her.

I don't have enough time to spare for that.

You're running out of time.

You need to budget your time.

In these sentences the word “time” could meaningfully be replaced with the word “money.” The hypothesis is that if two words or concepts are thought about in the same way by a group of people, then they will be systematically used in the same way in the conversation of those people. Examples of language use can be used as symptoms for the diagnosis of semantically related words and phrases.

By mapping out which words in a VLSC are used like which other words, a diagram of some of the semantics—that is, the meaning—of the conversation can be displayed. The resulting semantic network for a given VLSC is a way to begin to investigate how the statements made by different participants in the conversation are similar to one another.

The semantic network pictured in Figure 3 (and also displayed in the upper-right-hand corner of Figure 1) was computed automatically by the Conversation Map system given the archive of Usenet newsgroup messages described above concerning Kosovo (from the Usenet newsgroup soc.culture.albanian). The network is a tree. The tree is plotted like a spider web, so that the child nodes of the root (in Figure 3 the root is the node labeled “people”) are drawn at a certain radius out from the root, the children of the children are drawn a bit further out in a ring around the children, and so forth until the edges of the available area are reached. If two nodes in the semantic network are connected, then empirically they have often been used in the same way in the archive. Examples of how to interpret these connections in the semantic network will be given below.

The Conversation Map system parses all of the messages of the archive and for every noun in the archive it builds a “profile.” A noun’s profile consists of a weighted vector, including verbs (with which it appeared as a subject, a direct object, or an indirect object), adjectives (which modify it in the language of the messages of the group), and nouns (which have been used to modify it in noun–noun phrases). All of the nouns’ profiles are compared to one another and a nearest neighbor is computed for each noun.

A semantic tree, like the one shown above, is formed when one term—for example, “people”—is the computed nearest neighbor for several other nouns (i.e., in this case, for the nouns Serbs, Albanians, country, and refugee).

The procedure described produces many semantic networks for every archive of messages analyzed. All of the unique nouns encountered in the corpus of messages are included in a semantic network, but not all of the computed semantic networks are displayed in the interface. The semantic networks that are picked for

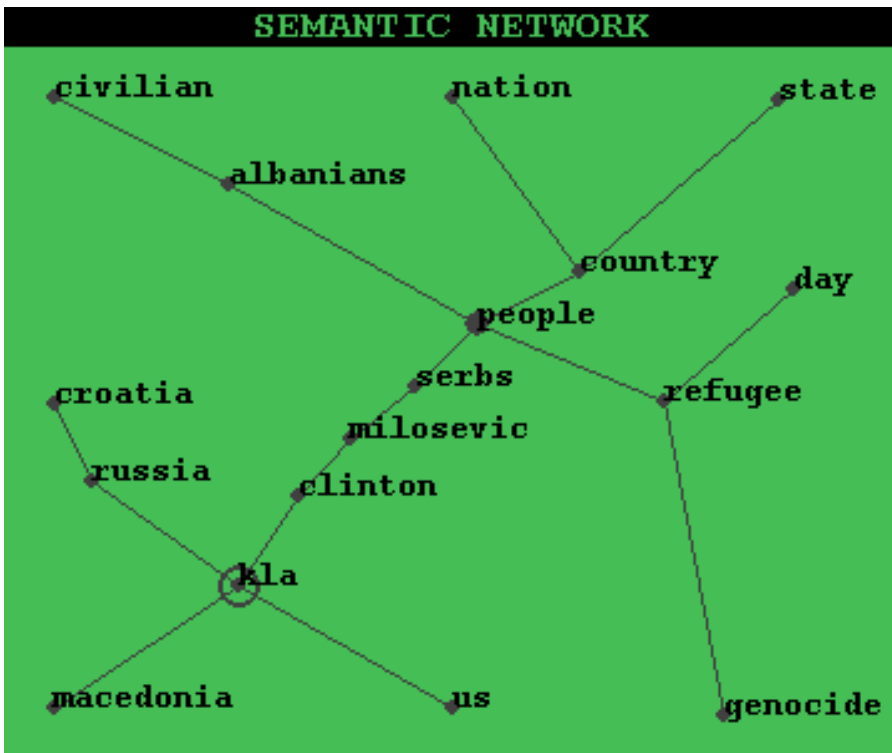


Figure 3. Semantic Network Showing Possible Metaphors or Definitions

display in the Conversation Map interface contain terms that are used frequently as themes of discussion. This ordering criterion on the semantic networks ensures that the networks seen first are those that are formed by group discussion rather than, for example, the verbose postings of a single spammer. In short, information from the social networks and social cohesion computation is used to sort the set of semantic networks. So, the ordering criteria knit together the ideational, textual, and interpersonal dimensions of the VLSC interface rendered by the Conversation Map system.

The profiles of the terms that appear in the semantic networks can be examined and compared by selecting one or more terms in the semantic network. For instance, if the terms “Serbs” and “people” are selected, a list of statements can be constructed like the one Lakoff and Johnson wrote for the TIME IS MONEY metaphor. If the list is restricted to only those verbs for which both “people” and “Serbs” appeared as a subject of the verb, then the resultant list looks like this:

SERBS ARE PEOPLE (both terms appear one or more times as the subject of each of the following verbs): allow, be, destroy, die, do, drive, exist, flee, get, give, have, keep, know, lay, leave, live, make, need, pay, remember, tell, think, turn

In other words, by looking at the archive of messages one can find many places where, for instance, both “people” and “Serbs” appear as subjects of the same verb. From the intersected lists of verbs one can see that in the archive of soc.culture.albanian messages “Serbs” and “people” are discussed in similar terms because there exist one or more statements in the archive for both “Serbs” and “people” where they are described separately as agents that allow, destroy, die, do, drive, exist, etc.

The verb “to need” is one of these shared verbs found in the intersection of the “Serbs” and “people” profiles. Clicking on a verb in the intersected profiles (not shown here, but displayed by the Conversation Map interface when two terms in the semantic network are selected) reveals the following two example sentences, which partially underpin the link between “Serbs” and “people” in the diagram: (1) “You have to realize that Greeks and *Serbs need* a just solution, . . .” (2) “It is not enough to be alive, *people need* normal life.”

Verb lists are displayed on demand by the Conversation Map system for any pair of terms in the semantic network and, if desired, example sentences of the terms in use can also be viewed.

ALBANIANS ARE PEOPLE (both terms appear one or more times as the subject of each of the following verbs): cross, displace, do, flee, have, hate, hide, leave, lose, say, suffer, think, walk

SERBS ARE ALBANIANS (both terms appear one or more times as the subject of each of the following verbs): do, flee, found, have, insist, leave, shoot, think, want

One optimistic way of reading the semantic network computed by the Conversation Map system for the soc.culture.albanian group is this: “People” is a neutral term and both “Serbs” and “Albanians” are people. This is a sort of thin humanism (after all, we are all people), which begins to explain why any sort of exchange can happen in this VLSC concerning Kosovo, even though, admittedly, the exchange is a very heated and argumentative one.

The approach described above for the computation and display of semantic networks is related to a variety of contemporary, corpus-based techniques of computational linguistics (cf., [27]). These techniques are, in turn, related to a variety of older work in linguistics, including the lexical fields approach of Trier [44] the work done on collocation by Firth [16], and the distributional approach to discourse analysis introduced by Harris [21] and pursued in a restricted but computational manner by, for example, Pêcheux [34]. See also the discussion by Lyons [29] concerning the longer history of these approaches to semantics in linguistics. More recent work by Callon et al. [6] is in a similar vein and has been carried out as a project in the sociology of science in order to summarize and analyze the discourse of science.

The computational techniques sketched out above for computing semantic networks in the Conversation Map system largely follow the techniques developed by Grefenstette [18]. Technically, the main advance over Grefenstette’s techniques discussed here is the use of computed social networks and lexical cohesion to automatically select a subset of the calculated semantic networks for display.

Computing and Displaying the Social and Semantic Networks

IN THE FOLLOWING SUBSECTIONS the procedure for calculating social and semantic networks is sketched out and then a short overview of the Conversation Map system interface is presented.

The Computation

The analysis procedure of the Conversation Map system performs the following steps on an archive of messages from a VLSC in order to compute the social and semantic networks described above:

1. Messages are “threaded.”
2. Quotations in the messages are identified and their sources (in other messages) are found.
3. The “signatures” of posters are identified and distinguished from the rest of the contents of each message.
4. An index of posters (i.e., newsgroup participants) to messages is built.
5. For every poster, the set of all other posters who replied to the poster—or quoted from messages authored by the poster—is recorded. Posters who reply to or quote from one another are linked together in the social network. Reciprocity is therefore highlighted in the computed social network.
6. The words in the messages are divided into sentences, tagged with part-of-speech information, and their roots are identified. To divide the words into sentences, a tool built at the University of Pennsylvania is used [38]. To accomplish the part-of-speech tagging, a simple trigram-based tagger has been constructed (cf., [8]). The morphological analyzer built for the Conversation Map system uses a freely available morphology and syntax database [26].
7. Discourse markers (e.g., connecting words like “if,” “therefore,” “consequently,” etc.) are tagged in the messages. The Conversation Map system employs a list of discourse markers compiled by Marcu [31].
8. The words of the messages are parsed into sentences using a partial parser. The partial parser is a reimplementaion of the parser described in [18].
9. An analysis of lexical cohesion is performed on every pair of messages, where a pair consists of one message of a thread and another message that either *immediately* follows the first message in the thread (i.e., is a reply to the first message) or follows the first message in the thread and contains a quotation from the first message. In other words, a message will be paired with a previous message even if it is not an immediate reply to the previous message but, instead, only quotes from the previous message. This analysis produces a series of lexical ties between messages that can be understood as a crude approximation to the theme of the conversation in a sequence of messages. The lexical database WordNet [15] is used in the lexical cohesion procedure. See [20] for a definition of lexical cohesion. See [24] for an example implementation of a lexical cohesion routine.

10. Using the index created in step 4 with the results of step 9, a set of lexical ties are computed for every pair of posters who have replied to or quoted from one another over the course of time represented by the VLSC archive under analysis. These aggregated lexical ties are layered on top of the social network computed in step 5. The result is that most of the links between pairs of posters are labeled with one or more lexical ties (i.e., one or more “discussion themes”). The combination of social networks and lexical cohesion results is here called *social cohesion*. The social cohesion analysis procedure developed for the Conversation Map system is partially described in [39].
11. The lexicosyntactic context of every noun in the archive is compared to the lexicosyntactic context of every other noun in the archive. Nouns that are used or discussed in the same manner are calculated to be similar and are placed close to one another in the semantic network. An algorithm similar to the one described in [18] is used. Once all of the noun–noun pairs have been compared and a nearest neighbor for each noun computed, a subset of the semantic networks computed is selected for display by ranking the semantic networks. The top-ranked semantic network contains a set of terms (used as “discussion themes”) that connect the greatest number of poster pairs linked in step 10.

The Graphical Display

After analyzing an archive of e-mail messages, the Conversation Map system generates a display of the archive and the social and semantic networks that constitute the analysis of the archive. An example of such a display is shown in Figure 1.

The top portion of the screen displays the social and semantic networks. The bottom portion is a graphical representation of the analyzed messages sorted by subject thread. Each thread is allotted a small rectangle of screen space and the threads are laid out chronologically from upper left to lower right. The threads are graphically plotted with the same tree-as-spider-web algorithm used to lay out the semantic network. The first message of a thread is plotted in the middle of the allotted rectangle. Responses to the first message are plotted around it and further toward the edges of the rectangle, and so on for the responses to the responses.

A small window containing an enlargement of one of the threads in the archive is created if one double-clicks on one of the threads displayed in the lower half of the screen (Figure 4). Each node in the thread display represents one message in the thread. Double-clicking on any of the nodes of the thread causes the text of the message to be displayed in a third window.

The graphically displayed social and semantic networks can be used to explore the archive of messages. For instance, clicking on one node of the social network (i.e., clicking on the name of a poster in the network) highlights all of the threads in the archive to which the poster has contributed one or more messages.

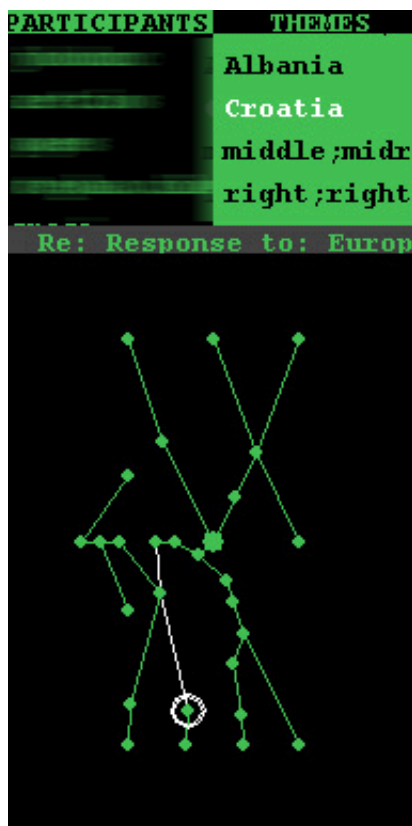


Figure 4. A Message Thread

In the screenshot shown in Figure 1, the menu item labeled “North Atlantic Treaty Organization” in the menu of discussion themes has been selected. Links between two pairs of posters in the social network have been highlighted because those posters have reciprocally exchanged messages concerning NATO. Also, many threads in the archive are highlighted with a white border because NATO has appeared as a discussion theme in them. Note that a discussion theme might appear throughout the archive, but only link one or two pairs of posters. Linking two posters is a more involved requirement than simply linking two messages in a thread. The link joining two posters, A and B, in the social network is labeled with a discussion theme if and only if poster A has responded to poster B concerning the theme and vice versa.

A similar functionality allows one to explore the archive by clicking on nodes in the semantic network. Moreover, as discussed above, the semantic networks can be “unfolded” to reveal the features (i.e., verbs, adjectives, etc.) shared by two or more terms. When two nodes in the semantic network are highlighted, a window appears containing lists like the profile intersections for “Serbs” and “Albanians” discussed above.

Related Work

Of the many different computer-mediated communication (CMC) systems that have been implemented to allow large groups of people to converse with one another, some have been designed to make existing groups communicate more efficiently or effectively (e.g., [30, 49]). Hiltz and Turoff describe the design philosophy behind systems of this kind: “The design philosophy is to start with the existing communication forms and functions of a user group and to build a system that accommodates or replicates such communication patterns at an overall increase in speed or efficiency and decrease in cost” [23, p. 18].

However, CMC systems engender changes in existing communication forms and functions even if they are initially designed to reinforce or speed up old patterns of communication. Palme points out that “The introduction of electronic mail changes communication patterns, so that people communicate with different people more often, and about other subjects than before [i.e., before the installation and use of a CMC system] [32, p. 9]. In many cases, groups originate online, so no prior group communication patterns exist to replicate or accommodate.

Consequently, it is necessary to formulate a complementary design philosophy for CMC systems in which the point is to help participants and observers spot emerging groups and changing patterns of communication. This paper argues for a complementary design philosophy of this sort. In other words, it is here advocated that CMC systems should not be built to just reinforce old patterns of communication, but rather CMC systems should be designed to help participants make and identify new patterns of communication.

The Conversation Map system has been designed to visually summarize emerging social and semantic networks that arise out of online communications. It is not designed to categorize a group’s communications into a set of predefined patterns. In short—unlike many previous CMC systems—the Conversation Map is a technology of clustering (i.e., a means of finding new groups and new patterns) and is not a technology of categorization (i.e., an attempt to fit a group’s communications into a predefined set of patterns).

In this vein, two types of prior work exist: (1) tools and browsers that incorporate automatic means for analyzing online discussions, and, (2) “hand” analyses of online conversations performed by linguists, sociologists, media studies researchers, and others.

Its novel, automatic text analysis procedures, in combination with a graphical interface, make the Conversation Map system an example of a new sort of content-based browser. Earlier examples of content-based browsers (e.g., [37]) used simpler text analysis procedures akin those employed in information retrieval systems. New content-based browsers, clients, and readers (like the Conversation Map system) will incorporate more sophisticated text analysis (and probably, eventually, image analysis) techniques. Several new analysis techniques—if combined with a graphical interface—could also serve as content-based browsers for spotting emerging social and linguistic patterns of online groups (see, especially, [4] and [48]).

Some content-based newsgroup browsers have been built with text analysis procedures simpler than those incorporated into the Conversation Map system. For example, Isahara and Ozaku [25] describe an intelligent network news reader that performs a sort of example-based, relevance feedback procedure to select small collections of messages from an archive, given an example message. The intelligent network news reader also contains a method for identifying subthreads within larger threads by analyzing the content of the messages in a thread [45]. However, systems of this sort (see also [41]) are mostly concerned with filtering messages rather than with one of the problems addressed by the Conversation Map system: How can all of the messages in an archive be graphically displayed and organized according to content of the messages to allow one to explore the emerging social structure and communication patterns of an online group?

Many of the computational techniques previously developed for the analysis of newsgroups do not take the linguistic content of the messages into account at all, using, instead, exclusively information that can be garnered from the headers of the messages (see, for example, [42]). Other work does employ some keyword spotting techniques to identify and sort the messages into categories but does not involve the analysis of grammatical or discourse structures (see, for instance, [10]). Work that does use the contents of the messages for analysis often does not take the threading of the messages into account, or, if it does, does not pay attention to the social network produced by newsgroup participants (e.g., [3]). Or, if the work does take the threading and citation information into account, it does not necessarily use any of the linguistic contents of the messages to compute the graphical display (e.g., [7]). Work by Paolillo [33] is an exception and an important precedent to the Conversation Map because it does combine linguistic and social network analyses together.

Research that has combined content analysis with an analysis of coreferencing of messages and discussion participants has often been done by hand (e.g., [2]). Some of the most interesting work that analyzes message threading, participant interaction, and the form and content of messages is often ethnographically oriented sociolinguistic analysis of newsgroup interactions that is done without the assistance of computers and is so, necessarily, based on a reading of only a small handful of messages (e.g., [22]). Ideally one could program the computer to emulate the latter sort of analysis, but that would require many advances in the field of computational linguistics. What is unique to the text analysis procedures of the Conversation Map system is the automatic construction and combination of social and semantic networks, which, together, provide a means for exploring both the social and semantic structure of a very-large-scale conversation.

The Conversation Map cannot be used to replace careful hand analyses of message archives. Rather, it can complement such work. Since a conversation or discourse analyst rarely has the time or resources to examine more than a few thousand messages, there exists a need for a tool like the Conversation Map to automatically chart out those (often extremely large) unexamined sections of an archive. Such a tool can both allow one to find interesting areas of an archive worthy of a closer reading and also provide a summary of the larger context of those messages that can be examined by hand.

Conclusions

There are several design criteria that have been used to construct and display the social and semantic networks of VLSCs with the Conversation Map system. These criteria negotiate a divergence between a scientific use of the social and semantic networks of the Conversation Map and a possible, popular, non-scientific use of the same.

Ideally, for the sake of science, the system-generated social and semantic networks would be constructed as carefully and rigorously as the “hand-built” analyses of, for instance, ethnographic or sociolinguistic studies of online discussions (e.g., [22]). However, clearly, the system-generated networks will never be as precise as analyses accomplished by hand. Thus, while from a scientist’s viewpoint it might at first appear to be a good idea to attempt to automate much of the process of online conversation analysis, a closer look at the pragmatics of such a design goal shows it to not be such a good idea.

However, since the system-generated results are quicker and easier to attain than comparable results compiled by hand, the results of the Conversation Map system could serve different needs for scientists and non-scientists. For the scientist—for example, for someone who is trained as a discourse or conversation analyst or an ethnographer—the networks computed by the Conversation Map system can be understood as a rough sketch of where one might begin to explore an archive of messages. (See [40] for one example of a social scientific use of the Conversation Map system.) With such a “sketch” in hand, the scientist can begin a set of more rigorous close readings of the archive of messages.

For the non-scientist, the computed social and semantic networks probably represent a much more detailed analysis of a VLSC than anything the non-scientist would ever produce. However, there are pitfalls associated with non-scientific usages of scientific-looking images. Scientific images have always been open to interpretation and put to alternative uses by non-scientists—for example, journalists, lawyers, politicians, interested laypersons, and non-specialists (i.e., even scientists or doctors who are not from the discipline directly responsible for the production of a set of scientific images). For example, the anthropologist of science and technology, Joseph Dumit, has examined how PET scans of the brain are used and understood both within science and “outside” in medicine, law, journalism, and popular culture [12].

The common pitfall associated with many vernacular presentations of scientific images concerns the manner in which the images are often “untethered” from the data used to produce them. Untethered scientific images—that is, images that have been unlinked from supporting data—sometimes become too easy to manipulate because they are no longer manipulated within the rigorous constraints of science—for example, when a popular magazine recolors an image received from a biologist to make the image easier to print, or more colorful. At other times, these untethered images become too hard to manipulate because a layperson has no access to the phenomenon pictured. For example, how would a non-scientist redraw the traces of subatomic particle collisions produced by physicists with an instrumented particle accelerator?

One crucial design question is therefore this: How can the images of science be used for interface devices without untethering the images from the supporting data? Specifically, how can the social scientific images of semantic and social networks be rendered as interface devices that cross-index and provide access to thousands of e-mail messages from the archives of very-large-scale conversations? The approach to this design problem taken in the construction of the Conversation Map system has been this: Make all of the original data (i.e., the e-mail messages) accessible through the graphical network displays (i.e., through the act of double-clicking on the social and semantic networks).

A second tension that exists between possible scientific versus possible popular uses of the Conversation Map system is this: To automatically generate the social and semantic networks, the Conversation Map produces an enormous amount of quantitative data on the messages and the participants of the newsgroup analyzed. Social scientists often want to see the numbers produced. For example, the Conversation Map system generates a set of statistics that could be useful to provide at least preliminary answers to the following questions: What was the mean number of responses sent to a message? How many other participants did a given participant respond to? How many times did a given theme show up as a theme of discussion? What proportion of the population of participants contributed toward the discussion on a given topic? Is there a specific subset of participants who started most of the threads concerning a given theme of discussion?

To negotiate this tension between the need for numbers and the larger scientific and non-scientific need to better understand the linguistic and social structures of VLSCs, the quantitative results calculated by the Conversation Map system are displayed not as final results, but rather as aids to begin more careful, reflective readings of an archive of e-mail messages. The displayed social and semantic networks can be moved and redrawn by the participant or interested observer, but they do not directly yield definitive, summary statistics on the group or individual participants of the group.

As interfaces for VLSCs come to be more and more representative of the social structure of the conversing group, many in the group may feel that an interface display impinges on their privacy by rendering transparent the history of their interactions with the group. Displaying all of the statistics calculated by the Conversation Map system would probably render the participants' profiles too transparent for comfort. Even today, the poster profiles computed and indexed at sites like www.deja.com incite these worries for some posters. Preferably, then, the construction of interfaces for VLSC will follow an aesthetic of social translucence [14] and encourage an ethic of social reflection, rather than aiming at an aesthetic of realism and transparency that would make all of us feel we are under a microscope of surveillance.

Acknowledgments: Parts of this paper appeared as Warren Sack, "Discourse Diagrams: Interface Design for Very Large-Scale Conversations," in the *Proceedings of the Hawaii International Conference on Systems Sciences*, Maui, HI: IEEE Computer Society (January 2000). I would like to thank Tom Erickson and Susan Herring for organizing and inviting me to participate in the *Persistent Conversations* section of that conference. I would also like to thank the

four anonymous referees who reviewed my conference paper and three other anonymous reviewers from the *JMIS*, all of whom gave thorough and thought-provoking critiques. I designed and implemented the Conversation Map system as a Ph.D. student and research assistant at the MIT Media Laboratory funded by the News in the Future Consortium. I thank my sponsors and my thesis advisors at MIT (Walter Bender, Ken Haase, John Rajchman, and Krzysztof Wodiczko).

REFERENCES

1. Bagga, A., and Baldwin, B. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of ACL-COLING '98*. Montreal, Canada, June 1998.
2. Berthold, M.; Sudweeks, F.; Newton, S.; and Coyne, R. It makes sense: using an autoassociative neural network to explore typicality in computer mediated discussions. In F. Sudweeks, M. McLaughlin, and S. Rafaeli (eds.) *Network and Netplay: Virtual Groups on the Internet*. Cambridge, MA: MIT Press/AAAI Press, 1998.
3. Best, M. Corporal ecologies and population fitness on the net. *Journal of Artificial Life*, 3, 4 (1998).
4. Best, M. Microevolutionary language theory. Ph.D. Thesis. Cambridge, MA: MIT Media Laboratory, 2000.
5. Black, S.; Levin, J.; Mehan, H., and Quinn, C. Real and non-real time interaction: unraveling multiple threads of discourse. *Discourse Processes*, 6 (1983), 59–75.
6. Callon, M.; Law, J.; and Rip, A. (eds.) *Mapping the Dynamics of Science: Sociology in the Real World*. London: Macmillan Press, Ltd., 1986.
7. Cannon, S., and Szeto, G. *Parasite*, <http://parasite.io360.com/index.html> and <http://www.cybergeography.org/atlas/topology.html>, 1998.
8. Charniak, E. *Statistical Language Learning*. Cambridge, MA: MIT Press, 1993, pp. 39–40.
9. Collins, A.M., and Loftus. E.F. A spreading activation theory of semantic processing. *Psychological Review*, 82 (1975), 407–428.
10. Donath, J.; Karahalios, K.; and Viegas, F. Visualizing Conversations. *Proceedings of HICSS-32*, Maui, HI: IEEE Computer Society, January 5–8, 1999.
11. Dourish, P., and Chalmers, M. Running out of space: models of information navigation. Presented at *HCI '94*. Glasgow, UK, 1994.
12. Dumit, J. *Whose Brain Is This? PET Scans and Personhood in Biomedical America*. Forthcoming; see also <http://symptom.mit.edu>.
13. Eisenman, P. Diagram: an original scene of writing. *Architecture New York*, 23 (1998), 27.
14. Erickson, T.; Smith, D.N.; Kellogg, W.A.; Laff, M.; Richards, J.T.; and Bradner, Erin. Socially translucent systems: social proxies, persistent conversation, and the design of “babble.” In *Proceedings of CHI '99*. Pittsburgh, PA, May 1999.
15. Fellbaum, C. (ed.). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.
16. Firth, J.R. *Papers in Linguistics, 1934–1951*. London: Oxford University Press, 1957.
17. Garfield, E. *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities*. New York: John Wiley & Sons, 1979.
18. Grefenstette, G. *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic Publishers, 1994.
19. Halliday, M.A.K. *An Introduction to Functional Grammar*, 2d ed. London: Edward Arnold, 1994, p. 179.
20. Halliday, M.A.K., and Hasan, R. *Cohesion in English*. New York: Longman, 1976.
21. Harris, Z. Discourse analysis. *Language*, 28 (1952), 1–30 and 474–494.
22. Herring, S.; Johnson, D.A.; and DiBenedetto, T. “This discussion is going too far!”: male resistance to female participation on the Internet. In K. Hall and M. Bucholtz (eds.), *Gender Articulated: Language and the Socially Constructed Self*. New York: Routledge, 1995.
23. Hiltz, S.R., and Turoff, M. *The Network Nation: Human Communication via Computer*, rev. ed. Cambridge, MA: MIT Press, 1993, p. 18.

24. Hirst, G., and St.-Onge, D. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.
25. Isahara, H., and Ozaku, H. Intelligent Network News Reader. In Proceedings of IUI '97, Orlando, FL, 1997.
26. Karp, D.; Schabes, Y.; Zaidel, M.; and Egedi, D. A freely available wide coverage morphological analyzer for English. In *Proceedings of COLING-92*, 1992.
27. Klavens, J., and Resnik, P. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: MIT Press, 1996.
28. Lakoff, G., and Johnson, M. *Metaphors We Live By*. Chicago: University of Chicago Press, 1980, pp. 7–8.
29. Lyons, J. *Semantics, Volume 2*. Cambridge: Cambridge University Press, 1977.
30. Malone, T.W.; Grant, K.R.; Lai, K.-Y.; Rao, R.; and Rosenblitt, D. Semistructured messages are surprisingly useful for computer-supported coordination. *ACM Transactions on Office Information Systems*, 5, 2 (April 1987), 115–131.
31. Marcu, D. The rhetorical parsing, summarization, and generation of natural language texts, Ph.D. Thesis. Toronto: Department of Computer Science, University of Toronto, December 1997.
32. Palme, J. *Electronic Mail*. Boston: Artech House, 1995.
33. Paolillo, J. The virtual speech community: social network and language variation on IRC. *Journal of Computer-Mediated Communication*, 4, 4 (June 1999), see <http://www.ascusc.org/jcmc/vol4/issue4/paolillo.html>.
34. Pêcheux, M. *Automatic Discourse Analysis*. Amsterdam: Editions Rodopi, 1995.
35. Quillian, M.R. Semantic memory. In M. Minsky (ed.), *Semantic Information Processing*. Cambridge, MA: MIT Press, 1968, p. 80.
36. Rafaeli, S., and Sudweeks, F. Interactivity on the nets. In F. Sudweeks; M. McLaughlin; and S. Rafaeli (eds.), *Network and Netplay: Virtual Groups on the Internet*. Cambridge, MA: MIT Press/AAAI Press, 1998.
37. Rennison, E. Galaxies of news: an approach to visualizing and understanding expansive news landscapes. In Proceedings of UIST '94, 1994.
38. Reynar, J.C., and Ratnaparkhi, A. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC, March 31–April 3, 1997.
39. Sack, W. Diagrams of social cohesion. In *Descriptions of Demonstrated Systems, ACL '99*. College Park, MD: University of Maryland, Association for Computational Linguistics, June 1999.
40. Sack, W., and Dumit, J. Very large-scale conversations and illness-based social movements. Presented at *Media in Transition*. Cambridge, MA: MIT Comparative Media Program, October 1999.
41. Sheth, B. NEWT: a learning approach to personalized information filtering. Master's thesis. Cambridge, MA: MIT Media Laboratory, 1993.
42. Smith, M. Netscan: measuring and mapping the social structure of Usenet. Presented at the *17th Annual International Sunbelt Social Network Conference*, Bahia Resort Hotel, Mission Bay, San Diego, CA, February 13–16, 1997.
43. Stairmand, M.A. Textual context analysis for information retrieval. In the *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, August 1997.
44. Trier, J. Das sprachliche Feld. Eine Auseinandersetzung. In *Neue Jahrbücher für Wissenschaft und Jugendbildung*, 10 (1934), 428–449.
45. Uchimoto, K.; Ozaku, H.; and Isahara, H. A method for identifying topic-changing articles in discussion-type newsgroups within the intelligent network news reader HISHO. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, Phuket, Thailand, December 2–4, 1997.
46. Wasserman, S., and Galaskiewicz, J. (eds.) *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*. Thousand Oaks, CA: Sage Publications, 1994.
47. Wellman, B. Living in a wired world. *IEEE Intelligent Systems* (January/February 1999), 15–17.

48. Whittaker, S.; Terveen, L.; Hill, W.; and Cherny, L. The dynamics of mass interaction. In *Proceedings of Conference on Computer Supported Cooperative Work*, pp. 257–264.

49. Winograd, T., and Flores, F. *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex Publishing Corporation, 1986.