

# Stories and Social Networks

Warren Sack

## Abstract\*

A computational, social network-based approach to story understanding is proposed and implemented in the *Conversation Map* system. Analyses of audiences' online discussions following the airing of two episodes of a well-known television show are presented.

## Introduction

**What's so important about stories?** The Internet has engendered a myriad of new social relations. These social relations, or "social networks"<sup>1</sup> are forged by individuals through electronic mail and Internet-based chat. Some of the very active interchanges focus on movies, television programs, and news stories. In other words, a non-trivial portion of these social networks are based on discussions of widely circulated stories. Virtual, on-line communities are a result of these net-mediated, story-based relations.

To imagine that these new social relations (and the resultant virtual communities) are important, one must also believe that stories are important. It matters which stories people know, which stories they tell, how they tell them, and how they are referred to. Narration, methods of citation and quotation, specific narratives, and general narrative forms constitute a kind of common sense<sup>2</sup> upon which virtual and imaginary communities,<sup>3</sup> have been built. These presuppositions are the presuppositions of media studies<sup>4</sup> and have also been integrated into some artificial intelligence (AI) research projects.<sup>5</sup>

---

\* Appears in the *Proceedings of the Workshop on Narrative Intelligence*, Michael Mateas and Phoebe Sengers (editors), Cape Cod, MA: American Association of Artificial Intelligence, November 1999.

<sup>1</sup> In this paper "social network" means a set of interrelated people. The phrase comes from social science. See, for example, Stanley Wasserman and Joseph Galaskiewicz (editors) *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences* (Sage Pub.: Thousand Oaks, CA, 1994).

<sup>2</sup> "...common sense is our storehouse of narrative structures, and it remains the source of intelligibility and certainty in human affairs." Roy Schafer. "Narration in the Psychoanalytic Dialogue" In W.J.T. Mitchell (editor) *On Narrative* (Chicago: University of Chicago Press, 1981)

<sup>3</sup> Benedict Anderson. *Imagined communities: Reflections on the origin and spread of Nationalism* (London: Verso, 1983).

<sup>4</sup> Stuart Hall. "The rediscovery of 'ideology': return of the repressed in media studies" In M. Gurevitch, T. Bennett, J. Curran, J. Woollacott (editors) *Culture, Society and the Media* (New York: Routledge, 1982).

<sup>5</sup> The work of Roger Schank, Robert Abelson and their students is

A rather blurry line separates the Internet-based practices of relating and retelling widely-circulated stories authored by mass-media producers (e.g., Hollywood, CNN, etc.) from the practices of independently producing stories for Internet distribution. It is the former sort of practice that is the concern of this paper. Quotation, citation, and fragmentary repetition of stories are the life-blood of audience discussions and analysis of mass-produced stories.<sup>6</sup> Audience members recirculate famous lines from movies (e.g., "Frankly my dear I don't give a damn," "I'll be back," "Make my day," etc.), comment on the plots and characters of known stories, summarize and retell pieces of stories for one another. The technology presented here is a first step towards a better understanding of story quotations, citations, and repetitions as the "threads" that weave people together into social networks.

A social network-based approach to story understanding differs from the standard approaches to "story understanding" that have been pursued by researchers in symbolic AI. Rather than examining stories as cognitive structures internal to individuals, the social network perspective is to see stories as shared ties that gather people into communities or social networks.<sup>7</sup> Moreover, unlike various media studies content analyses and structuralist analyses of narrative and film, it assumes the existence of an active, creative audience and uses audience activity (e.g., their discussion about a story) as the focus for gaining an understanding of stories.<sup>8</sup> This alternative

---

notable in this regard. Its close affinities with certain questions of media studies is unsurprising given the genealogy of the work. Robert Abelson did political analysis with media studies colleagues before his work in AI. For example, Ithiel de Sola Pool, Robert P. Abelson and Samuel L. Popkin. *Candidates, issues and strategies; a computer simulation of the 1960 and 1964 Presidential elections* (Cambridge, MA: MIT Press, 1965).

<sup>6</sup> Henry Jenkins discusses these audience practices as tactics of "poaching." Henry Jenkins. *Textual Poachers: Television Fans and Participatory Culture* (New York: Routledge, 1992).

<sup>7</sup> An analogous difference in approaches to narrative theory was described by Mikhail Bakhtin in his critique of Formalist approaches to literature and his advocacy of a sociolinguistic method. See, for example, Pavel Nikolaevich Medvedev [Mikhail Mikhailovich Bakhtin] *The formal method in literary scholarship: A critical introduction to sociological poetics* (Baltimore: Johns Hopkins University Press, 1978). Bakhtin's "dialogical" approach to language and literature has been widely employed in literary theory, sociology, and media studies. See, for instance, Henry Jenkins, *Op. Cit.*

<sup>8</sup> This distinction between research approaches in media studies (i.e., "content analysis" versus ethnographic approaches to the "active audience") has been recently explained in books such as

perspective shares some affinities with AI collaborative filtering techniques. Outside of AI, in the field of sociology, social network-based approaches to story understanding are not unusual, but the techniques of sociology can be improved through the use and development of an array of tools from natural language processing/computational linguistics. The research described here folds together insights from computational linguistics and the sociology of social networks to support the design of a new kind of story understanding technology; a technology predicated on the existence of verbally active story audiences.

A large amount of AI research is justified or motivated by pragmatic goals and there may in fact be pragmatic goals that would justify why we need a new technology of story understanding. In contrast, the poetics of AI have almost always been articulated around the need to get to know ourselves better. This poetics of the design and construction of intelligent, non-human entities has long been a theme of science fiction and science fantasy (not to mention its importance in philosophy since at least the time of Socrates when it was expressed as the ethical imperative “Know yourself.”) Sherry Turkle nicely illustrates the ways in which AI programs can function as a “second self.”<sup>9</sup> It is within this tradition of poetics – what the philosopher Michel Foucault has described as “technologies of the self”<sup>10</sup> – that I would argue that we need a new technology of story understanding. As new narrative forms are developed and new media proliferate, we need to invent new means for understanding how and where we are located in the emerging social networks.

## Methodology

### **Methodology = Computational Sociolinguistics = Computational Linguistics + Quantitative Sociology**

Within the field of sociology a number of computational approaches to understanding the social significance of literatures have been developed. Most prominently these methods have been applied to the literatures of science. For example, the methods of co-citation analysis<sup>11</sup> are routinely applied to determine the relative importance of a scientific article: its significance is thought to be a function of the number of other articles that cite it.<sup>12</sup> The methods

---

Virginia Nightingale. *Studying Audiences: The Shock of the Real* (New York: Routledge, 1996).

<sup>9</sup> Sherry Turkle. *The Second Self: Computers and the Human Spirit* (New York: Simon and Schuster, 1984).

<sup>10</sup> Michel Foucault. “Technologies of the Self” in *Ethics: Subjectivity and Truth (Essential Works of Foucault 1954-1984), Volume One*. Edited by Paul Rabinow. Translated by Robert Hurley and others (New York: The New Press, 1997).

<sup>11</sup> E. Garfield. *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities* (New York: John Wiley, 1979).

<sup>12</sup> AI elaborations of the techniques of co-citation analysis include

of social network theory<sup>13</sup> and actor-network theory<sup>14</sup> provide technologies akin to co-citation analysis, but have their own particular strengths and weaknesses.

These sorts of sociological “story understanding” technologies are very different from the story understanding technologies of an older, symbolic AI, but they have some affinities with techniques of newer AI work in agent-based architectures for information filtering and recommendation. Thus, for example, the “meaning” of a movie or television show for a system like Firefly<sup>15</sup> is the set of ratings members of a user community have assigned to it. Users of such a system can be said to form a group to the extent that they have given similar ratings to the same items.<sup>16</sup> For the most part these newer technologies (from sociology and from AI collaborative filtering research) for understanding stories as locations in and/or producers of social networks pay scant attention to the form and content of the stories: from this perspective stories are mostly “black boxes.”<sup>17</sup>

While the sociologists and AI, collaborative filtering researchers “black box” the form and content of stories, the corpus-based, computational linguistics and information retrieval researchers “black box” the social context of the stories they index. Corpus-based computational linguistics is most often performed on large corpora described as, for instance, “10 million words from several volumes of the Wall Street Journal,” or “1 million words from a wide variety of text genres.” How the authors of the texts included in the corpora interact with one another or are

---

Wendy Lehnert, Claire Cardie, and Ellen Riloff. “Analyzing research papers using citation sentences. In *Proceedings of the 12<sup>th</sup> Annual Conference on Cognitive Science*, 1990

<sup>13</sup> See, Stanley Wasserman, *Op. Cit.*

<sup>14</sup> Michel Callon, John Law, Arie Rip (editors) *Mapping the Dynamics of Science: Sociology in the Real World* (London: Macmillan Press, Ltd., 1986). See also Bruno Latour and Geneviève Teil “The Hume Machine: Can association networks do more than formal rules” *Stanford Humanities Review (special issue on artificial intelligence)* 4.2 (1995): 47-65. The technique of actor-network analysis is basically the calculation of mutual probabilities between nouns in scientific abstracts and so this technique probably has more affinities with techniques in computational linguistics than with those developed by other sociologists.

<sup>15</sup> Formerly at [www.firefly.com](http://www.firefly.com). See also, [agents.www.media.mit.edu/groups/agents/projects/](http://agents.www.media.mit.edu/groups/agents/projects/)

<sup>16</sup> Yezdezard Lashkari, “Feature guided automated collaborative filtering,” MIT Media Laboratory, Master’s Thesis, 1995.

<sup>17</sup> This is not to say that the content of the stories is necessarily completely ignored by these technologies. Lashkari, for example, describes an algorithm for collaborative filtering that takes into account the “content” of texts rated by the system’s users. However, the content analyses performed in practice by the system he describes were only simple, keyword-based information retrieval techniques that, for instance, do not take the order of words into account much less anything resembling the narrative or discourse structure of the texts.

related to one another is not factored into the analysis of the corpus. The one exception to this anonymity of authors is the use of corpus-based techniques for author identification purposes. But, even in these cases, the task is usually to determine who is the most likely author of a given text of a small set of possible candidates. The social network that incorporates (or the fact that no known social network incorporates) the set of candidate authors is not something that is taken into account in the design of the corpus-based, computational linguistic methods of analysis.

The techniques of corpus-based, computational linguistics are oftentimes technically related to the techniques employed by sociologists since both sets of techniques can depend upon similar tools from statistics and information theory (e.g., measures of mutual information and entropy). But the techniques are essentially inverses of one another due to the fact that what the sociologists black-box in their analyses is almost exactly what the corpus-based linguistics and information technology researchers do not black-box in their own research, and vice versa.

Any significantly new methodology for the development of a technology of story understanding should involve the combination of these two approaches. To understand a story as embedded in and (re)productive of both a network of related stories and other forms of discourse *and* as a facilitator or inhibitor of social networks, it is necessary to explore how social and semantic networks overlap.<sup>18</sup>

## Technology

**System Design and Implementation** I have been analyzing Usenet newsgroup, audience discussions of popular television programs in an attempt to understand how the stories of television are pulled apart, reiterated, quoted, summarized, and – in general – appropriated into and used for the social networks of television viewers.

To analyze these and other newsgroups the *Conversation Map* system has been designed and implemented. The input to the system is an archive of thousands of messages from a newsgroup. The output of the system is four-fold and is pictured in the figure below.

<sup>18</sup> While this intersection of social network and content analysis has been envisioned in sociology attempts to design and implement computer programs that combine sophisticated computational linguistic analysis with social network analysis are as yet unrealized.

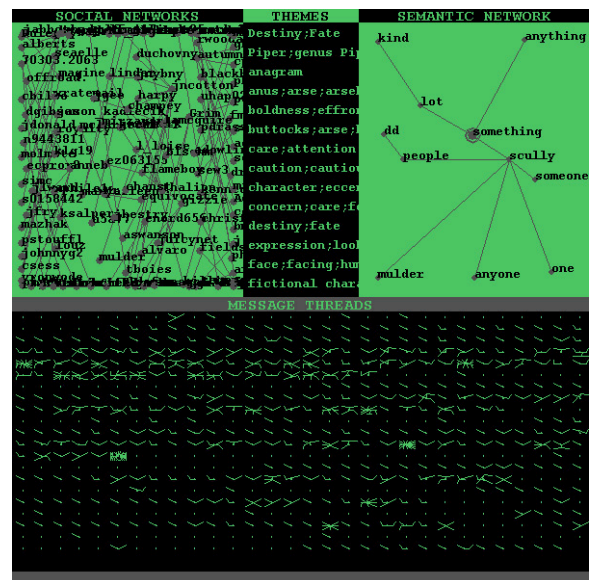


Figure 1: The Conversation Map interface

- (1) *Social Networks*: The upper left-hand panel displays a social network showing who is in conversation with whom. The nodes of the network are labeled with the names of the participants in the newsgroup conversation. If two names are connected and close to one another, then the two participants have been responding to or quoting from each other more frequently than if they are connected but far apart from one another. Two names are connected if both participants have responded to or quoted from the other. In other words, the social network diagrams *reciprocity*. If someone in the conversation posts a lot of messages, but no one responds to those messages, then that someone will not show up in the social network.
- (2) *Themes*: The upper middle panel is a menu of discussion themes. Themes listed at the top of the menu are those themes that are most commonly used in the conversation. The list of discussion themes is extracted from the archives by examining the words and synonyms of words in quotations and replies to previous messages. In linguistics, this analysis is properly described as an analysis of *lexical cohesion* between messages. The links between participants in the social network are labeled with the discussion themes from the menu of themes.
- (3) *Semantic Network*: The upper right-hand panel displays a semantic network. If two terms in the semantic network are connected together, then those two terms have been found to be synonyms -- or terms that may have similar meanings -- in the conversation. The semantic network is produced through the

application of corpus-based linguistics techniques<sup>19</sup> referred to in the literature as techniques of “semantic extraction” and “automatic thesaurus construction.

- (4) *Message Threads*: The panel that occupies the lower half of the window is a graphical representation of all of the messages that have been exchanged in the newsgroup conversation over a given period of time. The messages are organized into “threads,” i.e., groups of messages that are responses, responses to responses, etc. of some given initial message. The threads are organized chronologically, from upper-left to lower-right. The oldest messages can be found in the upper left-hand corner.

For a newsgroup which concerns a television program, the computed themes and terms in the semantic network often include names of characters and episodes from the television show, thus, these are the pieces of the television story that one can empirically observe as being appropriated into and employed by the audience’s discussions of the story. Obviously, with a more sophisticated set of computational linguistic analysis tools one might observe larger portions of the narrative structure being woven into the audience’s discussion. However, the set of computational linguistic procedures we employ and have developed expressly for our system are more sophisticated than any others compared to contemporary, computational work on the social and linguistic analysis of Usenet newsgroup discussions.<sup>20</sup>

<sup>19</sup> Cf., D. Hindle. “Noun classification from predicate-argument structures” In *Proceedings of the 27<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, pp. 118-125, 1990. Marti A. Hearst. “Automatic extraction of hyponyms from large text corpora” In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pp. 183-191, 1992.

<sup>20</sup> Many of the computational techniques developed for the analysis of Usenet 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, Marc Smith. “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, California, February 13-16, 1997 (see [www.sscnet.ucla.edu/soc/csoc/papers/sunbelt97/](http://www.sscnet.ucla.edu/soc/csoc/papers/sunbelt97/)). 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, Judith Donath, Karrie Karahalios, and Fernanda Viegas. “Visualizing Conversations” *Proceedings of HICSS-32*, Maui, HI, January 5-8, 1999.) 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 quotations and citations of one message in another (e.g., M.L. Best. “Corporal ecologies and population fitness on the net.” *Journal of Artificial Life*, 3(4), 1998). Research that has combined content analysis with an analysis of co-referencing of messages and discussion

The analysis engine of the Conversation Map system performs the following steps on an archive of Usenet newsgroup messages in order to compute the four outputs described above:

- (a) Messages are threaded.
- (b) Quotations are identified and their sources (in other messages) are found.
- (c) A table of posters (i.e., newsgroup participants) to messages is built.
- (d) For every poster, the set of all other posters who replied to the poster is recorded. Posters who reciprocally reply to one another’s messages are linked together in the social network.
- (e) The “signatures” of posters are identified and distinguished from the rest of the contents of each message.
- (f) The words in the messages are divided into sentences.<sup>21</sup>
- (g) Discourse markers (e.g., connecting words like “if”, “therefore”, “consequently”, etc.) are tagged in the messages.<sup>22</sup>
- (h) Every word of every message is tagged according to its part-of-speech (e.g., “noun”, “verb” “adjective”, etc.)<sup>23</sup>
- (i) Every word is morphologically analyzed and its root is

---

participants has often employed non-computational means to categorize the contents of messages (e.g., Michael Berthold, Fay Sudweeks, Sid Newton, Richard Coyne. “It makes sense: Using an autoassociative neural network to explore typicality in computer mediated discussions” In F. Sudweeks, M. McLaughlin, and S. Rafaeli (editors) *Network and Netplay: Virtual Groups on the Internet* (Cambridge, MA: AAAI/MIT Press, 1998). Some of the most interesting work that analyzes message threading, participant interaction, and the form and content of messages is often ethnographically-oriented, sociolinguistic analyses 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., Susan Herring, Deborah A. Johnson, Tamra DiBenedetto. “‘This discussion is going too far!’: Male resistance to female participation on the Internet” In K. Hall and M. Bucholtz (editors) *Gender Articulated: Language and the Socially Constructed Self* (New York: Routledge, 1995). Ideally one could program the computer to emulate the latter sort of analysis, but that will require many advances in the field of computational linguistics.

<sup>21</sup> The tool described in the following paper is used: Jeffrey C. Reynar and Adwait Ratnaparkhi. “A Maximum Entropy Approach to Identifying Sentence Boundaries.” In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, March 31-April 3, 1997. Washington, D.C.

<sup>22</sup> We use a list of discourse markers compiled by Daniel Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph.D. Thesis (Toronto: Department of Computer Science, University of Toronto, December 1997)

<sup>23</sup> A simple trigram based tagger is used to accomplish the part-of-speech tagging.

recorded.<sup>24</sup>

- (j) The words of the messages are parsed into sentences using a partial parser.<sup>25</sup>
- (k) An analysis of lexical cohesion<sup>26</sup> is performed on every pair of messages where a pair consists of one message of a thread followed by a message that follows the message in the thread by either referencing it or quoting a passage from it. This analysis produces the themes of discussion. The themes of the discussion label the arcs of the calculated social network. This allows one to see, for any given pair of posters, the theme of the posters' discussion.
- (l) The lexical and syntactic context of every noun in the archive is compared to the lexical and syntactic context of every other noun in the archive.<sup>27</sup> 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 networks. One can understand this semantic network as a crude approximation to the sorts of metaphors of discourse identified by linguists like George Lakoff.<sup>28</sup> Thus, for example, if the noun "economy" and the noun "plant" are often associated with the same verbs and adjectives (e.g., "plants grow", "the economy grows", "plants have roots", "the economy has root", "we have a healthy economy", "we have a healthy plant" etc.) the two words will be closely coupled in the word associations network and one can read that network as stating something like "the economy is like a plant."

Three parts of the fourfold output of the system (social networks, themes, and semantic networks) correspond to the three *metafunctions* of language defined by the linguist Michael Halliday:<sup>29</sup> the *interpersonal* (language connects

---

<sup>24</sup> The database containing morphological and syntactic information comes from the University of Pennsylvania: Daniel Karp, Yves Schabes, Martin Zaidel, and Dania Egedi. "A Freely Available Wide Coverage Morphological Analyzer for English" COLING-92.

<sup>25</sup> The partial parser is a re-implementation and revision of the parser described here: Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery* (Kluwer Academic Publishers: Boston, 1994).

<sup>26</sup> See Michael A.K. Halliday and Ruqaiya Hasan *Cohesion in English* (Longman: New York, 1976). The lexical cohesion analysis procedure we have developed is akin to, but different than, the one described here: Graeme Hirst and David St-Onge. "Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms" In Christiane Fellbaum (editor) *WordNet: An Electronic Lexical Database* (MIT Press, Cambridge, MA, 1998).

<sup>27</sup> An algorithm similar to the one described in Gregory Grefenstette, *Op. Cit.* is used.

<sup>28</sup> George Lakoff and Mark Johnson. *Metaphors We Live By* (University of Chicago Press: Chicago, 1980).

<sup>29</sup> Michael A.K. Halliday. *An Introduction to Functional Grammar, Second Edition* (Edward Arnold: London, 1994).

people together), the *textual* (language connects itself together by referencing other pieces of language through practices like quotation), and the *ideational* (language contains or carries ideas in it that are associated with other ideas).<sup>30</sup> The vast amount of research that has been done in sociolinguistics within a Hallidayean framework illustrates ways in which the current system could be improved if -- for the kinds of work sociolinguists have been doing by hand -- analogous computational linguistic techniques can be developed.

A user's manual for the Conversation Map system and interfaces for several archives (including the two message archives discussed in this paper) can be found on the web at this address: <http://www.media.mit.edu/~wsack/CM/index.html>. With the Conversation Map interface, the interested reader can explore the example messages, social and semantic networks, and themes discussed in the following section.

## Message Archives

Two message archives will be discussed. Both archives contain messages posted to the Usenet newsgroup alt.tv.x-files, a group devoted to discussion of the internationally broadcast television show entitled *The X-files*. The Usenet newsgroup discussion is archived and publicly available at a variety of websites including, [www.dejanews.com](http://www.dejanews.com). The staff at DejaNews was kind enough to provide us with the two archives discussed here.

*The X-files* is a weekly show produced by Twentieth Century Television in association with Fox Broadcasting Company. The show has two main characters, FBI Agents Dana Scully and Fox Mulder (played by actors Gillian Anderson and David Duchovny, respectively), who investigate cases reported to involve extraterrestrials, paranormal phenomena, and government conspiracy. It is an award winning television show now in its sixth season. More information about the show and short descriptions of the episodes can be found at the official *X-files* website: <http://www.thex-files.com/>.

Message Archive 1: These messages were exchanged during the week following the airing of the episode entitled "Quagmire" (4 May 1996 - 10 May 1996). In the "Quagmire" episode a Loch Ness monster-like creature is suspected of killing several people. About 700 participants posted over 1900 messages to the Usenet newsgroup alt.tv.x-files during this week after this episode was shown. A sketch of the analyzed messages from this archive can be

---

<sup>30</sup> A Hallidayean framework is also being applied by other researchers working on similar corpora, but with simpler text analysis procedures; see, for example, Simeon J. Yates "Oral and written linguistic aspects of computer conferencing" in Susan C. Herring (editor) *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives* (John Benjamins Pub. Co.: Philadelphia, 1996).

seen in Figure 1.

**Message Archive 2:** These messages were exchanged during the week following the airing of the episode entitled “Hell Money” (30 March 1996 – 5 April 1996). The “Hell Money” episode concerns a high-stakes gambling game in which the players risk their own organs (e.g., their eyes and kidneys). Approximately 900 participants posted 2400 messages to the Usenet newsgroup after this episode. Figure 2 below shows the Conversation Map automatically generated from the analysis of messages posted that week.



Figure 2: Conversation Map interface for Archive 2

## Preliminary Discussion

Before proceeding to a closer examination of the Conversation Maps, two points need to be made.

Firstly, in many structuralist, formalist, and/or older Marxist-inspired analyses of narrative and media audiences, the audience member is often assumed to be a “cultural dupe.” That is to say, it is assumed that a story delivered through the media (e.g., radio, television, the Internet, etc.) is not really open to interpretation and/or appropriation and means, more or less, one -- and only one -- thing. Moreover, the one and only meaning of the story is exactly what the audience member receives and, in this reception, is seen to be “programmed” by the story to behave or think in a certain manner by the story. This description is an over simplification, but it underlies the heat generated in arguments over which stories should or should not be taught in schools (i.e., the debate over the so-called “canon”) and also is a preferred viewpoint for many writers of non-fiction as well as that of past builders of AI technologies for “story understanding” who believed a machine could be built to understand “the point” of a story.

On the other end of the realist-to-relativist spectrum are

many post-structuralist and cultural studies-inspired media scholars who have tended to emphasize the extraordinary creativity of audience members. Stories, and media productions in general, are seen as raw materials for audience members to rewrite, reinterpret, and recreate in novel and undetermined ways.

By spending some time with the Conversation Maps of audiences’ online conversations, it should become clear that neither of these idealisms is empirically supported. On the one hand, the range of responses to the television stories is very diverse both in content and in genre. The “genres” of response include these: some responses are close intertextual analyses of the plot and characters of the episode, others are simple questions (e.g., “What’s your favorite X-files episode?”), others are wildly tangential (e.g., “I have two kittens, one named Mulder, the other Scully, and I’m looking for someone to adopt them...”). On the other hand, only someone who is very easily amused will be likely to see the messages contained in these archives as wildly creative.

Thus, as a first point, I maintain that a machine-assisted, empirical examination of audience conversation makes it quite easy to resolve an issue that is often a point of debate in narrative theory and media studies: audience members are not “cultural dupes,” but, neither are they more likely than any of the rest of us to be wildly creative with the “raw material” of the stories seen, heard, or read.

The second point also concerns the computational form of the analyses presented here. It has often been the case that audience studies have been formulated and written in a specialist’s language (e.g., the vocabulary of academic media studies) and presented in a medium unlike the medium of the story and unlike the media used by the audience members to communicate amongst themselves (e.g., studies of television audiences are oftentimes written up as academic books). For Internet-based audiences, it is now possible to build technologies that are designed to be accessible to the audience members and specialists alike. The Conversation Map system has been designed to be available online. To use the Conversation Map interface as a newsgroup browser for any of the messages discussed here use a Java 1.2 enabled web browser to explore this URL: <http://www.media.mit.edu/~wsack/CM/index.html>.

My second preliminary point is this: audience-accessible, networked, media studies cannot – as previous work repeatedly has – treat audiences as commodities or scientific objects because the network provides a means for the audience members to dispute the interpretations offered by the specialists. Consequently, what is presented below can best be understood as one place to begin an examination of the audiences’ understandings of the two X-files episodes, and not as a definite, final discovery of those understandings.

## Two Analyses

In what follows, the social networks, themes, and semantic networks displayed in the Conversation Maps of the two message archives will be more closely examined.

**Social Networks:** Figures 3 and 4 are enlargements of the social networks visible in Figures 1 and 2 respectively. In Figures 3 and 4 the names of the newsgroup participants have been turned off to allow one to see the topology of the networks more clearly.

What should be clear in Figures 3 and 4 is that participants are grouped into many small networks. The small networks are not connected to one another although it can be seen that the social networks shown in Figure 3 are more highly connected than the networks shown in Figure 4. In Figure 3, for example, the circled participant is a “lynchpin” of sorts holding together several smaller networks.

The lack of connections in the social networks is interesting because a quick glance at them makes it immediately apparent that the newsgroup is a space in which many different, probably unrelated, conversations are happening. Obviously the “effects” of a television story do not include the straightforward production of a cohesive social order.

It is interesting to compare the interconnections of these social networks with the social networks of other types of online discussions. Some of these can be seen at [www.media.mit.edu/~wsack/CM/index.html](http://www.media.mit.edu/~wsack/CM/index.html).

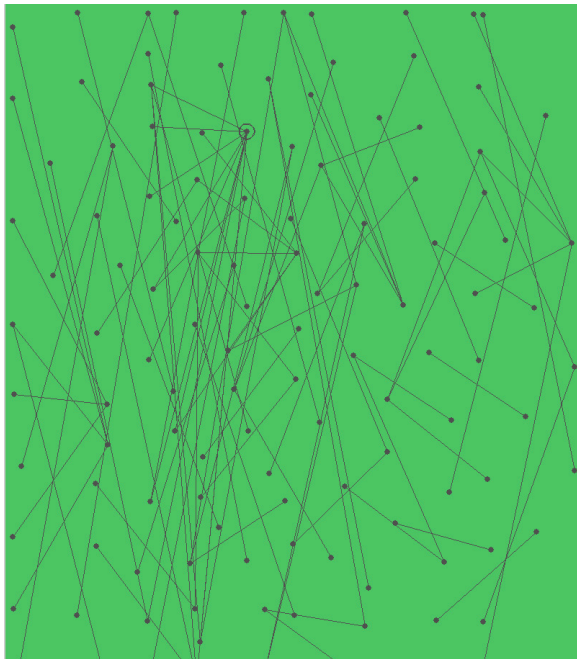


Figure 3: Social Network for Archive 1

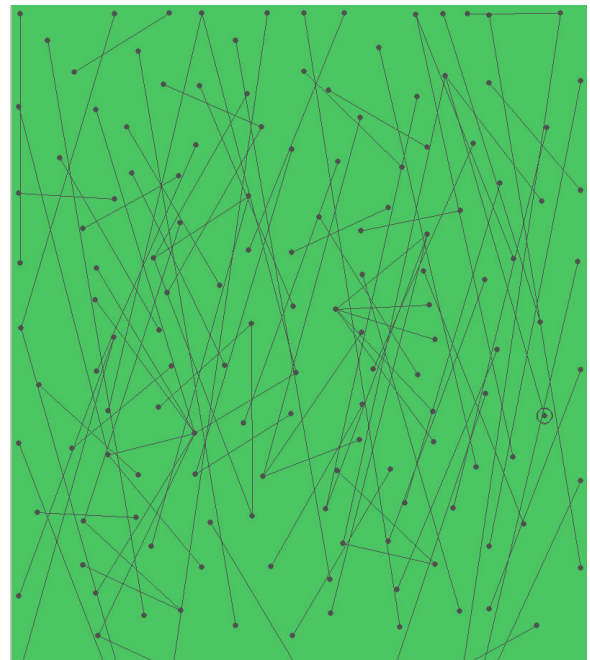


Figure 4: Social Network for Archive 2

**Themes:** Another measure of the diversity of conversation in a newsgroup is provided by the menu of computed “discussion themes” (i.e., what in linguistics would more properly be described as the *lexical ties* between messages). Figures 5 and 6 list the tops of the theme menus for message archives 1 and 2 respectively.

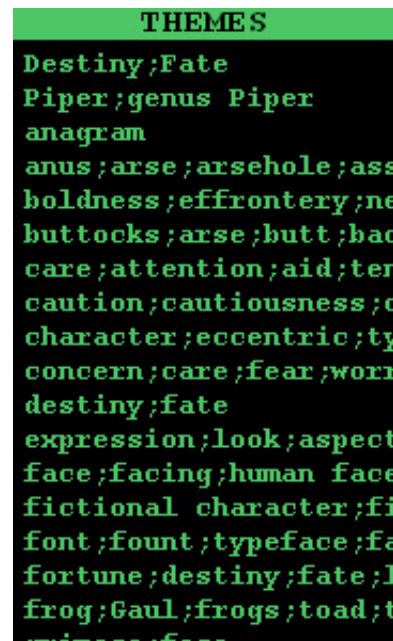


Figure 5: Themes Menu for Archive 1



Figure 6: Themes Menu for Archive 2

Themes in the menus of themes are ordered according to the number of arcs in the social network that they label.

Remember that an arc in the social network connects two newsgroup participants if and only if those two participants have replied to each other or cited from one another's messages. Thus, for example, A and B are connected in the social network and the arc between A and B is labeled with a theme – e.g., “sports” – if and only if A and B have had at least one interchange like the following: A posts a message about baseball, B replies with a post about football, B posts a message about swimming, and A cites or responds to B's message with one about skiing. Since baseball, football, swimming, and skiing are all sports, the link between A and B might be labeled with the more abstract term “sports” (computed by the Conversation Map system using the WordNet version 1.6 thesaurus). So, the themes listed in the menus are only there if there has been one or more reciprocated responses in which the theme (or a semantically similar) term was mentioned in each of the exchanged messages.

Figure 6, showing the reciprocated discussion themes in the messages of archive 2, is a rather surprisingly short list. Usually the menu of themes lists many items. Clicking on the items to highlight the parts of the social network that they label shows even more clearly how fragmented the discussion of archive 2 is. All of the themes listed connect only one pair of posters. In short, only a small handful of the interchanges concerning the “Hell Money” episode are focused around a specific theme of discussion.

Figure 5, showing the reciprocated discussion themes in the messages of archive 2, shows again that the social interchange visible in the message archives is more cohesive in the first archive than it is in the second archive. This can be interpreted from the longer list of reciprocated themes for archive 1.

**Semantic Networks:** The semantic networks shown in Figures 7 and 8 show that the conversations after both episodes are concerned with the main characters (Scully and Mulder) and, moreover, it is interesting to see the computed similarities between the main characters and the more generic terms of “you,” “me,” “someone,” “anyone” etc. These calculations provide a way of seeing how the audience members talk about themselves in ways comparable to the way they talk about the main characters. This calculation might be compared to analyses of character “identification” discussed in the literatures of film theory and other media studies.

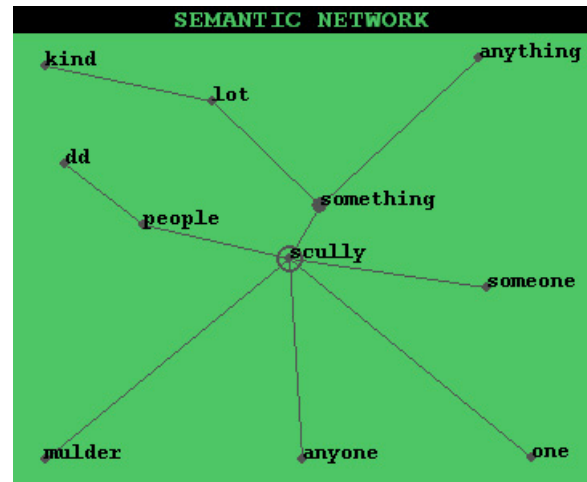


Figure 7: Semantic Network from Archive 1

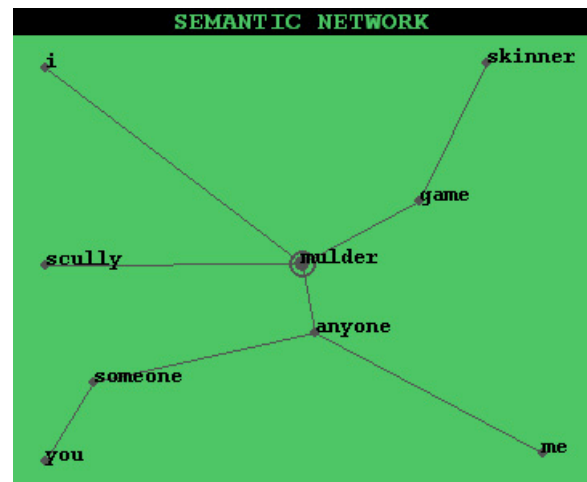


Figure 8: Semantic Network from Archive 2

## Conclusions

A computational sociolinguistic analysis of stories has been proposed and implemented in the *Conversation Map* system. The significance of a story is seen as a function of the social network that it engenders and/or inflects. The proposed analysis method was compared to related work in AI collaborative filtering, sociology, and computational corpus-based linguistics. It was also briefly compared to the relatively unrelated work in story understanding done within the symbolic AI tradition. The Conversation Map system has been designed and implemented to perform a sociolinguistic analysis of Usenet newsgroup analysis postings and it has been employed in the analysis of television audiences' newsgroup discussions of stories from a popular television show. The output of the implemented system illustrates sociolinguistic analyses of the television stories as they are visible in the social

networks and language of the television audiences' newsgroup postings.

## References

<sup>1</sup> In this paper "social network" means a set of interrelated people. The phrase comes from social science. See, for example, Stanley Wasserman and Joseph Galaskiewicz (editors) *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences* (Sage Pub.: Thousand Oaks, CA, 1994).

<sup>1</sup> "...common sense is our storehouse of narrative structures, and it remains the source of intelligibility and certainty in human affairs." Roy Schafer. "Narration in the Psychoanalytic Dialogue" In W.J.T. Mitchell (editor) *On Narrative* (Chicago: University of Chicago Press, 1981)

<sup>1</sup> Benedict Anderson. *Imagined communities: Reflections on the origin and spread of Nationalism* (London: Verso, 1983).

<sup>1</sup> Stuart Hall. "The rediscovery of 'ideology': return of the repressed in media studies" In M. Gurevitch, T. Bennett, J. Curran, J. Woollacott (editors) *Culture, Society and the Media* (New York: Routledge, 1982).

<sup>1</sup> The work of Roger Schank, Robert Abelson and their students is notable in this regard. Its close affinities with certain questions of media studies is unsurprising given the genealogy of the work. Robert Abelson did political analysis with media studies colleagues before his work in AI. For example, Ithiel de Sola Pool, Robert P. Abelson and Samuel L. Popkin. *Candidates, issues and strategies; a computer simulation of the 1960 and 1964 Presidential elections* (Cambridge, MA: MIT Press, 1965).

<sup>1</sup> Henry Jenkins discusses these audience practices as tactics of "poaching." Henry Jenkins. *Textual Poachers: Television Fans and Participatory Culture* (New York: Routledge, 1992).

<sup>1</sup> An analogous difference in approaches to narrative theory was described by Mikhail Bakhtin in his critique of Formalist approaches to literature and his advocacy of a sociolinguistic method. See, for example, Pavel Nikolaevich Medvedev [Mikhail Mikhailovich Bakhtin] *The formal method in literary scholarship: A critical introduction to sociological poetics* (Baltimore: Johns Hopkins University Press, 1978). Bakhtin's "dialogical" approach to language and literature has been widely employed in literary theory, sociology, and media studies. See, for instance, Henry Jenkins, *Op. Cit.*

<sup>1</sup> This distinction between research approaches in media studies (i.e., "content analysis" versus ethnographic approaches to the "active audience") has been recently explained in books such as Virginia Nightingale. *Studying Audiences: The Shock of the Real* (New York: Routledge, 1996).

<sup>1</sup> Sherry Turkle. *The Second Self: Computers and the Human Spirit* (New York: Simon and Schuster, 1984).

<sup>1</sup> Michel Foucault. "Technologies of the Self" in *Ethics: Subjectivity and Truth (Essential Works of Foucault 1954-1984), Volume One*. Edited by Paul Rabinow. Translated by Robert Hurley and others (New York: The New Press, 1997).

<sup>1</sup> E. Garfield. *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities* (New York: John Wiley, 1979).

<sup>1</sup> AI elaborations of the techniques of co-citation analysis include Wendy Lehnert, Claire Cardie, and Ellen Riloff. "Analyzing research papers using citation sentences. In *Proceedings of the 12<sup>th</sup> Annual Conference on Cognitive Science*, 1990

<sup>1</sup> See, Stanley Wasserman, *Op. Cit.*

<sup>1</sup> Michel Callon, John Law, Arie Rip (editors) *Mapping the Dynamics of Science: Sociology in the Real World* (London: Macmillan Press, Ltd., 1986). See also Bruno Latour and Geneviève Teil "The Hume Machine: Can association networks do more than formal rules" *Stanford Humanities Review (special issue on artificial intelligence)* 4.2 (1995): 47-65. The technique of actor-network analysis is basically the calculation of mutual probabilities between nouns in scientific abstracts and so this technique probably has more affinities with techniques in computational linguistics than with those developed by other sociologists.

<sup>1</sup> Formerly at [www.firefly.com](http://www.firefly.com). See also, [agents.media.mit.edu/groups/agents/projects/](http://agents.media.mit.edu/groups/agents/projects/)

<sup>1</sup> Yezdezdard Lashkari, "Feature guided automated collaborative filtering," MIT Media Laboratory, Master's Thesis, 1995.

<sup>1</sup> This is not to say that the content of the stories is necessarily completely ignored by these technologies. Lashkari, for example, describes an algorithm for collaborative filtering that takes into account the "content" of texts rated by the system's users. However, the content analyses performed in practice by the system he describes were only simple, keyword-based information retrieval techniques that, for instance, do not take the order of words into account much less anything resembling the narrative or discourse structure of the texts.

<sup>1</sup> While this intersection of social network and content analysis has been envisioned in sociology attempts to design and implement computer programs that combine sophisticated computational linguistic analysis with social network analysis are as yet unrealized.

<sup>1</sup> Cf., D. Hindle. "Noun classification from predicate-argument structures" In *Proceedings of the 27<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, pp. 118-125, 1990. Marti A. Hearst. "Automatic extraction of hyponyms from large text corpora" In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pp. 183-191, 1992.

<sup>1</sup> Many of the computational techniques developed for the analysis of Usenet 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, Marc Smith. "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, California, February 13-16, 1997 (see [www.sscnet.ucla.edu/soc/csoc/papers/sunbelt97/](http://www.sscnet.ucla.edu/soc/csoc/papers/sunbelt97/)). 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, Judith Donath, Karrie Karahalios, and Fernanda Viegas. "Visualizing Conversations" *Proceedings of HICSS-32*, Maui, HI, January 5-8, 1999.) 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 quotations and citations of one message in another (e.g., M.L. Best. "Corporal ecologies and population fitness on the net." *Journal of Artificial Life*, 3(4), 1998). Research that has combined content analysis with an analysis of co-referencing of messages and discussion participants has often employed non-computational means to categorize the contents of messages (e.g., Michael Berthold, Fay Sudweeks, Sid Newton, Richard Coyne. "It makes sense: Using an autoassociative neural network to explore typicality in

computer mediated discussions” In F. Sudweeks, M. McLaughlin, and S. Rafaeli (editors) *Network and Netplay: Virtual Groups on the Internet* (Cambridge, MA: AAAI/MIT Press, 1998). Some of the most interesting work that analyzes message threading, participant interaction, and the form and content of messages is often ethnographically-oriented, sociolinguistic analyses 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., Susan Herring, Deborah A. Johnson, Tamra DiBenedetto. “‘This discussion is going too far!’: Male resistance to female participation on the Internet” In K. Hall and M. Bucholtz (editors) *Gender Articulated: Language and the Socially Constructed Self* (New York: Routledge, 1995). Ideally one could program the computer to emulate the latter sort of analysis, but that will require many advances in the field of computational linguistics.

<sup>1</sup> The tool described in the following paper is used: Jeffrey C. Reynar and Adwait Ratnaparkhi. “A Maximum Entropy Approach to Identifying Sentence Boundaries.” In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, March 31-April 3, 1997. Washington, D.C.

<sup>1</sup> We use a list of discourse markers compiled by Daniel Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph.D. Thesis (Toronto: Department of Computer Science, University of Toronto, December 1997)

<sup>1</sup> A simple trigram based tagger is used to accomplish the part-of-speech tagging.

<sup>1</sup> The database containing morphological and syntactic information comes from the University of Pennsylvania: Daniel Karp, Yves Schabes, Martin Zaidel, and Dania Egedi. "A Freely Available Wide Coverage Morphological Analyzer for English" COLING-92.

<sup>1</sup> The partial parser is a re-implementation and revision of the parser described here: Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery* (Kluwer Academic Publishers: Boston, 1994).

<sup>1</sup> See Michael A.K. Halliday and Ruqaiya Hasan *Cohesion in English* (Longman: New York, 1976). The lexical cohesion analysis procedure we have developed is akin to, but different than, the one described here: Graeme Hirst and David St-Onge. “Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms” In Christiane Fellbaum (editor) *WordNet: An Electronic Lexical Database* (MIT Press, Cambridge, MA, 1998).

<sup>1</sup> An algorithm similar to the one described in Gregory Grefenstette, *Op. Cit.* is used.

<sup>1</sup> George Lakoff and Mark Johnson. *Metaphors We Live By* (University of Chicago Press: Chicago, 1980).

<sup>1</sup> Michael A.K. Halliday. *An Introduction to Functional Grammar, Second Edition* (Edward Arnold: London, 1994).

<sup>1</sup> A Hallidayean framework is also being applied by other researchers working on similar corpora, but with simpler text analysis procedures; see, for example, Simeon J. Yates “Oral and written linguistic aspects of computer conferencing” in Susan C. Herring (editor) *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives* (John Benjamins Pub. Co.: Philadelphia, 1996).