SAND2009-6718C

Diagramming Workgroup Interaction via Social Language Network Analysis

Andrew J. Scholand
Sandia National Laboratories
Box 5800, Albuquerque NM 87185
ajschol@sandia.gov
(505) 284-9110

Yla R. Tausczik
Department of Psychology
University of Texas, Austin, TX 78712
tausczik@mail.utexas.edu

ABSTRACT

In this note we demonstrate how a new methodology that combines tools from social language processing and network analysis can be used to identify the nature of socially situated working relationships within a group. We call this approach social language network analysis (SLNA). We utilize this approach to create tree-like diagrams relating the linguistic categories of both long-term (15-month) and short-term (10 day) archives of discussions concerning massive high performance computing (HPC) simulations of the economic consequence of infrastructure disruptions. These example diagrams contrast the explicit mapping of short term pure technical interactions against the long term blending of social support and accomplishing work.

Author Keywords
social language processing, social network analysis, network structure, communication, content analysis, group.

ACM Classification Keywords
J.4 Social and Behavioral Sciences: Psychology.

General Terms
Algorithms, Human Factors, Management, Measurement

INTRODUCTION

As computational tools continue to grow in complexity and capability, multidisciplinary workgroups are increasingly needed to manage the information flowing in to and out from them. Workgroups can dramatically vary in outcome performance [1; 12], however, and to be most effective, managers need to monitor and evaluate the quality of workgroup interactions prior to the emergence of adverse outcomes. With the increasing pervasiveness of Computer Mediated Communication (CMC) technologies in the workplace, many workgroup interactions now create digital artifacts conducive to the creation of leading indicators of performance. Linguistic analysis in particular is an extremely powerful way to examine the foundational personality, cognitive, and biological processes underlying daily social interactions [2] and as such provides unique insight into how socially situated work is accomplished. For example, the Linguistic Inquiry and Word Count (LIWC) program measures word use in psychologically meaningful categories and has been successfully used to identify relationships between individuals in social interactions, including relative status (e.g. [6]), deception (e.g. [8]), and the quality of close relationships (e.g. [11]).

Linguistic metrics, however, can sometimes be difficult to interpret individually because similar language can be used in different ways. For example, the first-person plural pronoun ‘we’ can indicate either higher work role status [10] or social inclusiveness. We propose in this paper that a recently developed quantitative approach we call social language network analysis (SLNA) can aid in identifying the dominant mode of language use for these ambiguous situations by clustering intragroup linguistic measures. We show that the resultant pattern of linguistic category interrelationships is a signature that arises from the nature of the group discourse by comparing and contrasting the patterns from two alternate samples of text from the same group.

BACKGROUND

Computational Economics Discussions

This approach is demonstrated with two distinct but related archives of work-related conversations in a scientific research and development (R&D) organization. A group heterogeneous in academic discipline, age, gender, experience and geographic location used a custom-built synchronous collaboration framework to critically evaluate, discuss, and plan advanced high performance computing (HPC) simulations of regional and national economic activity. The group also used the framework to evaluate simulation initialization specifications derived from data fused across multiple government and commercial data sources. The first conversational archive, which we refer to as the Short Term Project, was collected in March 2005 as part of an explicit designed experiment, described in [7]. These conversations between six individuals, two females and four males, were collected in highly task-focused collaboration sessions lasting a total of 72 hours spread over the month. The second conversational archive, the Long Term Project, recorded public chat messages sent between these same six individuals and twelve additional team members for a period of 15 months, from September 2006 to November 2007. These participants included 7 females and 11 males, from 22 to 64 years old. This archive represents a long dwell observation of the extended group, and includes a mixture of both social interactions and explicit work conversations, ranging from technical troubleshooting to economic theorizing [14].

METHOD

SLNA consists of three interrelated processing steps. The first step assigns each unit of conversational data to one or more directed links, each from the speaker to listener(s). To convert the synchronous chat into relational conversations, conversations were defined as consecutive messages without more than a 5-minute delay between responses (as in [4]) and are assumed to be solely between those participants synchronously participating. The second step converts text associated with particular links to a quantitative metric. Since we are using the data to connect individuals to those they communicate with in a social network sense, the output of this step is a series of valued adjacency matrices, one for each metric computed. In these examples, we processed the language associated with each relational link using the LIWC program [9], resulting in adjacency matrices across 80 linguistic dimensions. We then normalized these metrics to sum to unity per originator (out-bound normalization). The third and final step uses one or more of these quantitative metric matrices in a graph-processing algorithm to compute an objective of interest.

In this diagramming application, we carry out this third step by first computing the pair-wise correlation between LIWC category adjacency matrices. As an example, Figure 1 uses shades of gray to illustrate the adjacency matrices for the Long Term group LIWC categories ‘six letter words’ (all words longer than six letters), ‘relativity’ (638 words such as area, bend, exit, stop), ‘you’ (20 second person pronouns, including you, your, and thou), and ‘assent words’ (30 words such as agree, OK, yes). The top adjacency matrices, ‘six letter words’ and ‘relativity’, look very similar because these categories are used in very similar ways across the group – they occur in conversations between the same group members – resulting in a correlation of 0.8645. In contrast, the categories ‘you’ and ‘assent words’ have quite different patterns of use across the group, as indicated by both the two distinct images in the bottom row of Figure 1 and a correlation of only 0.1184. These category-to-category correlations are then combined into a symmetric matrix representing the group’s use of linguistic categories. Treating the correlations as similarity measures allows the calculation of a dendrogram by agglomerative hierarchical clustering with weighted average linking.

We select *a priori* the LIWC categories to examine on a theoretical basis. This is especially important for categories with multiple modes of use. For example, inclusive language is used both in a cognitive sense and in a social sense. Limiting the available clustering to only social categories, we highlight the social uses of this language even if that sense is less frequently used than the cognitive uses.

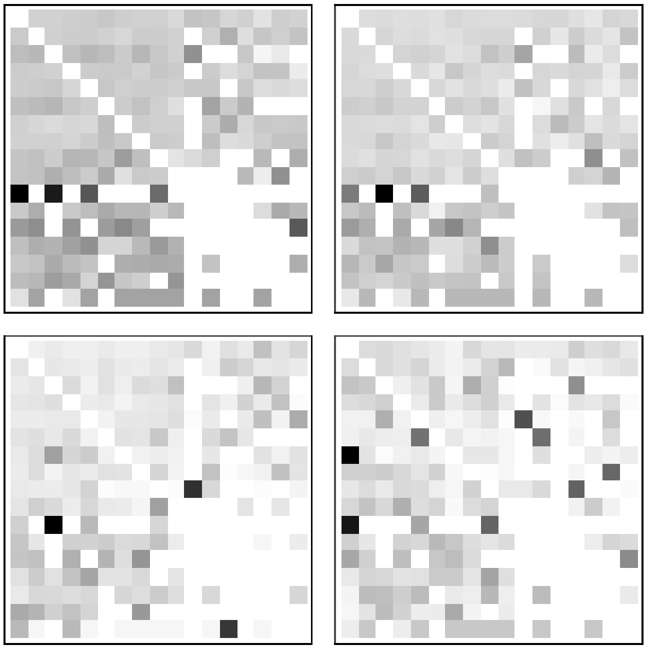

Figure 1. Language Use Across Long-Term Group: Six Letter Words and Relativity (top); You and Assent Words (bottom)

ANALYSIS

We selected the following LIWC categories for clustering, preceded by the motivating reason for examining each set:

cognitive mechanisms: causation, insight, discrepancy, certainty, tentative, inclusive, exclusive

describing and visualizing: relativity, hear, see

work related: quantifiers, numbers, six letter words

conversation feedback: negations, assent words, question marks

pronouns: you, I, she/he, they, we

LIWC is partially hierarchically organized, with super-categories comprised of a number of categories, which may in turn be comprised of sub-categories. For example, the super-category ‘pronouns’ is comprised of ‘personal pronouns’ and ‘impersonal pronouns’. ‘Personal pronouns’ is further divided into the sub-categories ‘I’, ‘we’, ‘you’, ‘she/he’, and ‘they’. The categories listed above were specifically selected to be independent, to avoid a phenomenon we term ‘composition bias’, whereby a sub-category is correlated to the category it belongs to simply because of the accounting mechanics: counting a word in a lower category also increments the word count for categories further up the hierarchy. (Shifts in the specific sub-category/category pair most strongly associated do indicate changes in the dominant mode of the encompassing category and can be informative in longitudinal studies, but the example presented here is aggregative and cross-sectional.)

Dendrograms created by pvclust, an R package for hierarchical clustering with p-values [13], for the two example archives are shown in Figures 2 and 3. These dendrograms are assembled in a 'bottom-up' fashion via the growth of associated terms in clusters, so that relations closer to the bottom edge of the figures are more significant. The pvclust package estimates an approximately unbiased p-value indicating the degree of support in the data for the indicated clusters via multiscale bootstrap resampling. Clusters for which we can reject the hypothesis that "the cluster does not exist" with a significance level of 0.10 are outlined with a thin frame.

Examining Figures 2 and 3 at a high level, we see that in both archives two clusters emerge at 0.90 confidence intervals and the content of each is roughly comparable. The smaller cluster entails questions and insight; broadly speaking those who ask the questions are able to verbally express insights. The larger cluster is twice as big in both archives, and represents complex (six or more lettered words), hedged (tentative and discrepancy), and precise (exclusion, quantifiers, certainty) reasoning about causality. We believe this cluster is an artifact of distributed cognition as the groups' experts converse to arrive at answers for the subgroup represented by the smaller cluster. At the coarsest level, then, this diagramming technique illustrates a question-and-answer type dialog and inquiry-led knowledge creation within the (related) groups of both archives.

Another high-level observation can be made concerning the perceptual categories 'hear' and 'see'. The 'hear' category – comprised of words and stems such as say*, sound*, said, noise, heard, and hear – is excluded from the significant clusters in the long-term project archive. Conversely, the 'see' category – with words like see, look, looking, looks, red, screen, blue*, view, yellow*, and green* – is significant in both archives. We believe the continued importance of 'see' language in both diagrams reflects the unique features of the collaborative framework, which was custom developed specifically to allow the sharing and annotation of images. The environment specifically supports visual pattern recognition and we see vision emerge as a primary perceptual mode. Interestingly, there is a change in how this visual mode supports group cognition across the two different archives. In the initial short-term project archive, 'see' words are associated with certainty words (e.g. all, sure*, real, correct*, exact*, every, never, always, and true). The high-level implications of the simulations at this early stage of development were not fully trusted and required critical evaluation, but the collaborating group members could be certain of the concrete results they saw. In the later long-term project, these visual terms are more associated with questions and in particular with the pronoun 'we'. In this example quote, the collaborators are attempting to coordinate shared attention. (Words from the 'see' and 'we' categories are in bold font.)

> PersonA: Which run are **we looking** at first and what should **we** be **looking** at initially?
>
> PersonD: **we** will be **looking** at CogEcon runs... but only those that do not have REMOVE status
>
> PersonD: **We** are **looking** at this first to get a feeling for what the Epi model alone introduces into N-ABLE.... which is basically just that agents die and are removed from the sim at some time

The migration of visual references from the concrete to abstract questions is one way these diagrams reveal evolution in the group's knowledge over time.

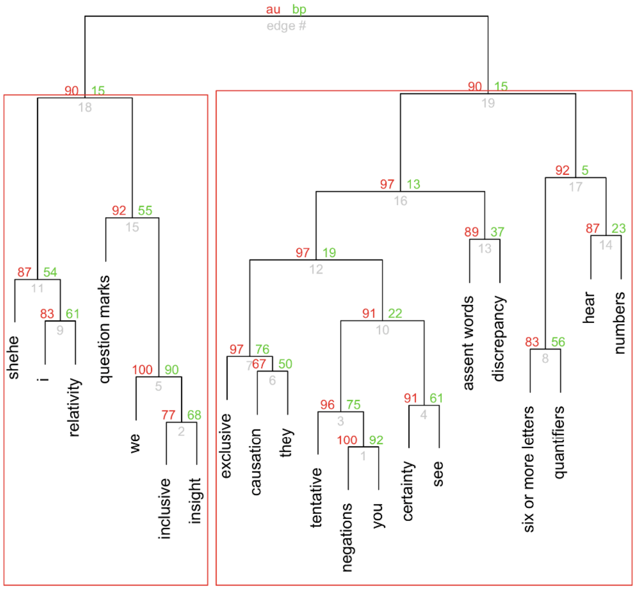

Figure 2. Short Term Group Linguistic Categories Diagram

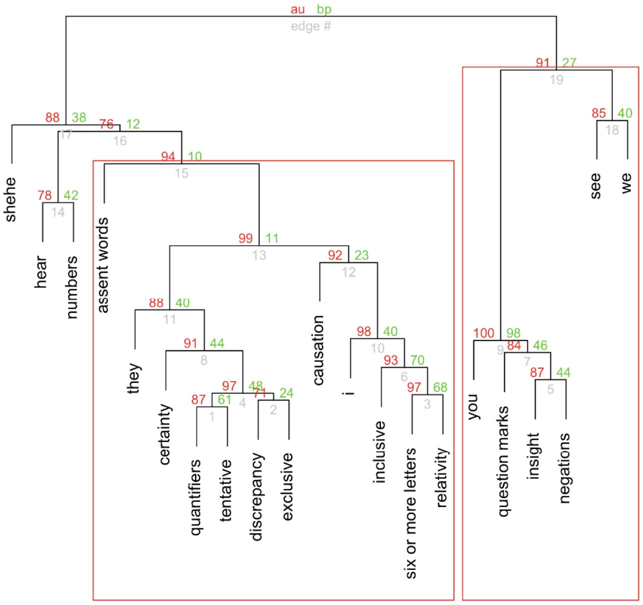

Figure 3. Long Term Group Linguistic Categories Diagram

This knowledge shift is also evident in the movement of the 'causation' branch in Figures 2 and 3. Because group knowledge was more tenuous in the earlier short-term archive, reasoning about the simulations needed to be more concrete. The 'causation' category was therefore strongly associated with the pronoun category 'they', referring specifically to the agents in the simulation. In the later long-term archive, a more general association of 'causation' to an abstract 'relativity' category cluster in addition to the agent-specific 'they' category documents the advances in the group's understanding of the motive forces driving the simulations.

Question marks are associated with the pronoun category 'you' in the long term project, as the core experts are joined by newer and more peripheral staff who need to direct their questions to those with this expertise. The earlier short-term conversations, in contrast, were between an egalitarian group, as shown by the question and 'we'/'inclusive' language association. In addition to this knowledge asymmetry across people in the long-term archive, the knowledge itself has become less amorphous. Negations (such as not, no, don't, can't, didn't, doesn't, never, isn't, haven't, and won't) are hence more useful in excluding certain known possibilities as the knowledgeable experts offer them.

CONCLUSION

Hinds and McGrath [3] argue for group-level measures in attempting to understand the structural relations of groups (teams in their language), noting "team-level network measures highlight the dynamics in the team as a whole, enabling us to examine the overall social and work structures within these teams."

We believe an important and useful set of measures is made possible through social language processing's ability to access information about the relative social, psychological, and emotional connections that situate us within a community. Social language network analysis (SLNA) brings these attributional and dyadic data up to the group level. In this note, we demonstrate how the application of SLNA to two examples of real world knowledge-intensive collaborative work communication [5] archives highlights and makes explicit important components of group functioning.

Despite significant differences in group size, topics of conversation, and duration of observation, the underlying domain and intellectual activity create similar relations in both the SLNA diagrams, as does the visual-verbal integration of the communication medium. The increasing evolution of knowledge is also evident in the permutation of several correlations between cognitive categories. We believe these characteristics make these SLNA diagrams useful for identifying the nature of socially situated working relationships within a group, and will ultimately lead to the measures advocated by Hinds and McGrath.

ACKNOWLEDGMENTS

Thanks to the Army Research Institute (W91WAW-07-C-0029) and the Sandia Laboratory Directed Research & Development Seniors' Council for our funding.

REFERENCES

1. Baldwin, T.T., Bedell, M.D., and Johnson, J.L., The Social Fabric of a Team-Based M.B.A. Program, *Acad. of Management J.*, 40, 6 (1997), pp. 1369-1397.
2. Chung, C.K. and Pennebaker, J.W., The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology*, Psychology Press (2007), pp. 343-359.
3. Hinds, P., and McGrath, C., Structures that Work: Social Structure, Work Structure and Coordination, *CSCW'06*, ACM Press (2006), pp. 343- 352.
4. Issacs, E., Walendowski, A., Whittaker, S., Schiano, D.J., and Kamm, C., The Character, Functions, and Styles of Instant Messaging in the Workplace, *Proc. CSCW'02*, ACM Press (2002), pp. 11- 20.
5. Julsrud, T.E, Core/periphery Structures and Trust in Distributed Work Groups, *Structure and Dynamics*, 2, 2 (2007) pp. 1-28.
6. Kacewicz, E., Pennebaker, J.W., Davis, M., Jeon, M., and Graesser, A.C. (under review). The language of status hierarchies.
7. Linebarger, J.L., Scholand, A.J., and Ehlen, M.A., Representations and Metaphors for the Structure of Synchronous Multimedia Collaboration, *HICCS '06*, IEEE Computer Society (2006), pp. 58-68.
8. Newman, M.L., Pennebaker, J.W., Berry, D.S., and Richards, J.M., Lying words, *Personality and Social Psychology Bulletin*, 29 (2003), pp. 665-675.
9. Pennebaker, J.W., Booth, R.J., and Francis, M.E., Linguistic Inquiry and Word Count: A computerized text analysis program. Austin, TX (2007). LIWC.net
10. Sexton, J.B. and Helmreich, R.L., Analyzing cockpit communications, *Human Performance in Extreme Environments*, 5, 1 (2000), pp. 63-68.
11. Slatcher, R.B. and Pennebaker, J.W., How do I love thee? *Psychological Science*, 17 (2006), pp. 660-664.
12. Sparrowe, R.T., Liden, R.C., Wayne, S.J., and Kraimer, M.L., Social Networks and the Performance of Individuals and Groups, *Academy of Management Journal*, 44, 2 (2001), pp. 316-325.
13. Suzuki, R. and Shimodaira, H., Pvclust: an R package for assessing the uncertainty in hierarchical clustering, *Bioinformatics*, 22, 12 (2206), pp. 1540-1542.
14. Tausczik, Y.R., *Linguistic Analysis of Workplace Computer-Mediated Communication*, Masters Thesis, The University of Texas at Austin, August 2009.