Assessing Group Interaction with Social Language Network Analysis

Andrew J. Scholand[1], Yla R. Tausczik[2], and James W. Pennebaker[2]

[1] Sandia National Laboratories*
Box 5800, Albuquerque, NM 87185
ajschol@sandia.gov
[2] Department of Psychology
University of Texas, Austin, TX 78712
{tausczik,pennebaker}@mail.utexas.edu

Abstract. In this paper we discuss a new methodology, social language network analysis (SLNA), that combines tools from social language processing and network analysis to assess socially situated working relationships within a group. Specifically, SLNA aims to identify and characterize the nature of working relationships by processing artifacts generated with computer-mediated communication systems, such as instant message texts or emails. Because social language processing is able to identify psychological, social, and emotional processes that individuals are not able to fully mask, social language network analysis can clarify and highlight complex interdependencies between group members, even when these relationships are latent or unrecognized.

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

1 Introduction

This research, addressing a technical means to make a socially informed group assessment, was motivated by interest in successful group interaction and collaboration. Much knowledge and expertise in successfully executing work quickly is not written down because this information is developed, shared, and acted upon in an operational context through informal conversations among and between groups of individuals. The fundamental research proposition is that digital records arising from interactions in the 'as-is' organization can be analyzed to create an approximate but meaningful representation of the work-centered social dynamics within the group. 'Meaningful' in this context implies facets of information relevant to interpersonal dynamics, aspects of distributed cognition and group work, and the development of organizational power and control.

* Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 248–255, 2010.

In earlier work, we used social language network analysis (SLNA) to diagram a hierarchy of professional respect and predict close personal friendships [6] and to identify dominant cognitive themes in work-related conversations [7]. In this paper, we turn the focus of SNLA to relations within a group (peer to peer) and between leadership and staff.

2 Background

This work leverages social language analysis performed by the Linguistic Inquiry and Word Count (LIWC) software program developed by University of Texas researchers James W. Pennebaker, Roger J. Booth, and Martha E. Francis [4]. LIWC is a program for quantitative text analysis that uses a word count strategy for both the analysis of content (what is being said) and style (how it is being said). Word count strategies are based on the assumption that the words people use convey psychological information over and above their literal meaning and independent of their semantic context. In this sense, they are "top down" in that they explore text within the context of previously defined psychological content dimensions or word categories. (In contrast, word pattern strategies such as latent semantic analysis mathematically detect "bottom-up" how words co-vary across large samples of text, typically to determine the degree to which two texts are similar in terms of their content.) LIWC searches for over 2300 words or word stems previously categorized by independent judges into over 80 linguistic dimensions. These dimensions include standard language categories (e.g., articles, prepositions, pronouns– including first person singular, first person plural, etc.), psychological processes (e.g., positive and negative emotion categories, cognitive processes such as use of causation words, self-discrepancies), relativity-related words (e.g., time, verb tense, motion, space), and traditional Freudian content dimensions.

3 Data Source

The National Infrastructure Simulation and Analysis Center (NISAC) developed a programmable collaboration library to facilitate secure collaborative interaction by geographically distributed decision-makers [3]. The collaboration framework offers the usual collaborative services (chat and file transfer) as well as the ability to publish multiple images for collaborative text and graphical annotation. These capabilities focus primarily on *synchronous* capabilities that allow the integration of multiple perspectives and quick convergence on a shared view of a problem to facilitate high-pressure, time-constrained analyses. This framework has been used since 2003 by the geographically distributed Computational Economics Group to plan, stage, execute, debug, and interpret high performance computing simulations of the national economy subject to regional disruptions. The group also used the tool to evaluate simulation initialization specifications derived from data fused across multiple government and commercial data sources. These work-related instant message conversations between 18 team members were collected for this analysis from September 2006 to November 2007.[1] In this period, there were 14,416 separate statements totaling 170,197 words. The participants included 7 females and 11 males, varying in age from 22 to 64 years old. Four other chat participants were excluded due to contributing less than 250 words in public chat during the period of the study. The number of words contributed per participant ranged from a maximum of 56464 to a minimum of 253, with a mean value of 9435 (standard deviation of 15755) and a median of 2298.

4 Relational Analysis Results

The pattern of conversations in chat illustrate a dichotomous structure underlying the group interaction that in turn affects language use. Figure 1 shows a tree constructed by clustering (Johnson's hierarchical clustering [1] with weighted average clustering) individuals based on the number of conversations they had with each other as a measure of similarity. Individuals who are connected to each other near the left hand side of the dendogram shown in Figure 1 were involved in a higher number of conversations with each other. The gray rectangle superimposed on the dendogram divides the group into two subgroups of equal size. The upper subgroup represents the highly connected 'core' of the group. The lower subgroup represents (a portion of) the more loosely interconnected periphery. Although there is some pairwise structure in the peripheral subgroup, these ties mostly occur at weaker levels than the ties in the core subgroup.

The first person plural pronoun group includes the pronouns 'we,' 'us,' and 'our' as well as various plural and possessive variants. The LIWC program computes the relative ratios of these words to all words spoken in the recorded conversations. Due to variations in speech patterns by age and gender, these metrics are normalized by computing the average of non-zero 'We' pronoun usage percentages for each speaker and deducting this average from those values. Zero values for various individuals indicate no conversations occurred between the speaker and that individual, and so these zeros were left unmodified. The resulting pattern of strongly positive links, indicating conversations between individuals with normalized 'We' pronoun usage values above 2%, were all observed to originate in the core group and link to the peripheral group. Conversely, examining arcs with percentages of -0.75% or lower (relative to each speaker's average usage of 'We' pronouns), 17 of the 22 arcs (77%) both originate and terminate within the core group. The use of 'We' pronouns appears to be substantially less within the subgroup of individuals who comprise the core of this group.

The essence of this finding, then, is that 'We' pronoun usage is inversely related to the degree to which members belong to the group. Those individuals engaging in the most conversations within the group use pronouns from the 'We' group most infrequently when chatting with other frequent conversation partners. To test this association statistically, the Quadratic Assignment Procedure (QAP) [2] can be

[1] The use of these data has been reviewed and approved by Sandia's Human Studies Board in Research Protocol SNL0806.

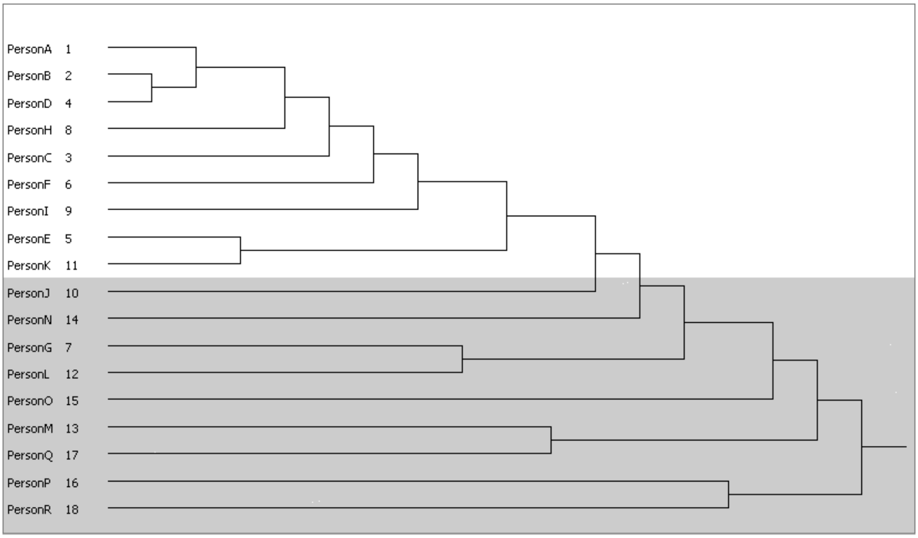

Fig. 1. Johnson Hierarchical Clustering of Conversation Count

Table 1. QAP Correlation of 'We' use and Conversation Count data

Statistic	Value
Pearson Correlation:	-0.314
Significance:	0.000
Permutation Average (50000 permutations):	0.000
Permutation Standard Deviation:	0.086
Minimum Permuted Value:	-0.291
Maximum Permuted Value:	0.306

applied to measure the degree of correlation between the normalized 'We' pronoun use matrix and the conversation count matrix.[2] First, the corresponding cells of the two adjacency matrices are correlated using ordinary Pearson correlation. Second, a large number (50,000) of randomly re-arranged matrices are correlated to assess if the observed match is likely by pure chance. If the proportion of random trials that would generate a coefficient as small as the statistic actually observed is small enough, typically below 0.05, the hypothesis of no association is rejected. Table 1 shows that randomly permuted matrices on average have no correlation whatsoever (Pearson Correlation of 0.000), and therefore the observed inverse correlation of -0.314 is highly significant statistically.

[2] The conversation count data was zero-meaned before being processed, so that the Pearson's coefficient is computed for centered data.

Table 2. Pearson Correlation of LIWC Categories and Conversation Count data

Category	Examples	Value	Signif.	Avg.	S. D.	Min.
biological processes	breakfast, cafeteria, pizza	-0.330	0.000	0.000	0.082	-0.289
smileys	:-)	-0.325	0.000	0.000	0.082	-0.269
body	eye*, face, sleep*	-0.319	0.000	0.000	0.075	-0.252
first person plural	we, us, our	-0.314	0.000	0.000	0.086	-0.291
health	ache*, exercis*, pills	-0.312	0.000	0.000	0.071	-0.249
inclusive	both, come, inclu*	-0.285	0.000	0.000	0.077	-0.282
ingestion	ate, chew*, coffee	-0.269	0.000	-0.001	0.088	-0.305
family	family, husband, wife*	-0.264	0.000	0.000	0.061	-0.272
third person plural	their*, them, they've	-0.244	0.000	-0.001	0.074	-0.240
discrepancy	besides, if, problem*	-0.237	0.000	0.000	0.077	-0.258

To put these findings in context, Table 2 lists this category with the other LIWC categories that are most strongly associated with group structure. All of these associations are negative, suggesting that the core subgroup focuses on these categories primarily in communications with the peripheral subgroup rather than among themselves. The topics suggest attention to health and wellness (health, body, ingestion, biological processes), non-work issues (family), and minimizing communication misunderstandings (smileys) in these communication channels. As discussed above, there are multiple components suggesting outreach and perhaps attempts to verbally assimilate the periphery into the core, including the 'we,' 'they,' and 'inclusive' categories. The 'discrepancy' group is the only category suggesting a specific work-related focus; discrepancy words are used to differentiate concepts.

5 In-Group and Management Relationships

Participants completed a questionnaire rated their attitudes toward each of their collegues by evaluating nine different statements on a seven point Likert scale. Eleven of the 18 participants whose messages were recorded in the public chat forum responded to our request to complete surveys. These participants, 4 female and 7 male, ranged in age from 22 to 64. The ratings each participant gave to each of the other group members can be considered that person's conceptualization of their dyadic relationship. Chats with those individuals is then an instance of the class of conversations that occurs at that degree of relation. For example, assume Person A rates Person B as a close personal friend at Likert scale degree 7, 'a great deal.' When Person A chats to Person B, then, this language can be considered to be representative of chat between very close friends. If Person E similarly rates Person K as a very close friend, chat from Person E to Person K would also be categorized as being between very close friends. Aggregating all of the language from individuals who rated their conversational partner at each Likert scale level yields a sample of language across the spectrum of sentiment for a given question. We subjected this partitioning of the chat language sample to LIWC content analysis and correlated the results to the scale level.

A common observation in the organizational studies literature is that negative information is increasingly filtered as it moves up the management chain [5]. Figure 2 provides empirical support for this assertion. The graph on the left side of Figure 2 shows a negative correlation between the LIWC category 'Negemo' and higher social status. People spoke with fewer negative terms when conversing with individuals they perceived to be of higher status. The graph on the right shows a reinforcing effect, namely that people in this group tended to use more positive terms ('Posemo') the less well they knew the person. Hence, management, having both higher social status[3] and being less well known by most staff members than their peers, receive both less negative information and more positive information from staff.

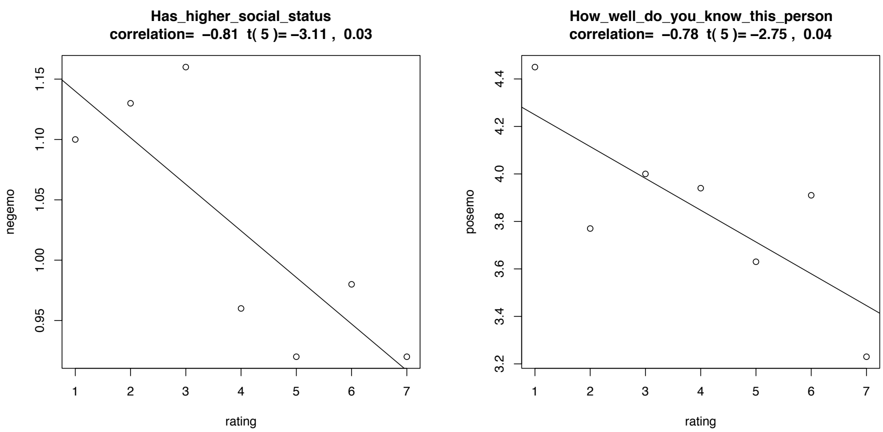

Fig. 2. Emotional Filtering as a Function of Status and Familiarity

How group members rated the communication skills of their peers strongly predicted the types of conversations held with those peers. The graph on the left side of Figure 3 shows a link between how easy to talk to a person is perceived to be and the extent to which conversations with that person contain tentative words. People are willing to share thoughts and interpretations they are less sure of when their conversational partner is easy to talk to. Conversely, the graph on the right side of Figure 3 shows that more difficult to approach individuals are met with more "causative" language. Conversations are initiated with these individuals predominantly when there is a reason to approach them.

[3] The manager and team leads in this work group were rated as having high social status in the questionnaire.

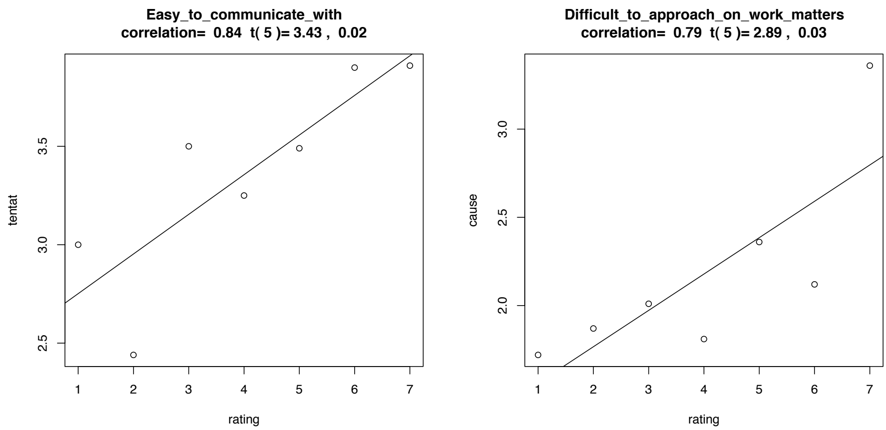

Fig. 3. Type of Conversation as Function of Communication Skills

6 Conclusions

This work combines the high fidelity assessment of relationships between entities made possible by language analysis with the contextual framework of social network processing to both predict underlying structural relations and retrospectively describe patterns of group interaction. By selectively extracting, combining, and processing different psychological, social, and emotional linguistic markers it is possible to map the rich relationships within and across groups, making difficult tasks such as managing organizational change, organizational design, and interorganizational relationships easier.

A study of language use between the well-connected core and periphery of a group of 18 scientific collaborators highlights efforts at outreach and assimilation, including exceptional use of pronouns such as ‘we’ and ‘they,’ and other ‘inclusive’ language. Only one work-related LIWC category (‘discrepancy’ words; used to differentiate concepts) was used in a similar statistically significant manner, suggesting work is primarily accomplished in subgroups rather than across the group as a whole. Correlating LIWC measures with survey elicited evaluations of co-workers revealed people in this organization are more willing to share tentative interpretations when their conversational partner is easy to talk to. In contrast, conversations are typically initiated with less socially skilled individuals only when there is a specific reason to approach them. Vertical information flow within the group is also shown to be influenced by both status and social distance, with high status and aloof leaders receiving both decreases in negative information and increases in positive information from staff.

Acknowledgments. Thanks to the Army Research Institute (W91WAW-07-C-0029) and the Sandia Laboratory Directed Research & Development Seniors Council for the funding that made this publication possible.

References

1. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika 32(3), 241–254 (1967)
2. Krackhardt, D.: QAP Partialling as a Test of Spuriousness. Social Networks 9(2), 171–186 (1987)
3. Linebarger, J.M., Scholand, A.J., Ehlen, M.A., Procopio, M.J.: Benefits of Synchronous Collaboration Support for an Application-centered Analysis Team Working on Complex Problems: a Case Study. In: GROUP 2005, pp. 51–60. ACM Press, New York (2005)
4. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic Inquiry and Word Count: A computerized text analysis program, Austin, TX (2007), `www.LIWC.net`
5. Reason, J.: Managing the Risks of Organizational Accidents. Aldershot, Ashgate (1997)
6. Scholand, A.J., Tausczik, Y.R., Pennebaker, J.W.: Social Language Network Analysis. In: CSCW 2010. ACM Press, New York (2010)
7. Scholand, A.J., Tausczik, Y.R.: Diagramming Workgroup Interaction via Social Language Network Analysis. In: CSCW 2010 Workshop on The Changing Dynamics of Scientific Collaborations (2010)