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, 78712
tausczik@mail.utexas.edu

James W. Pennebaker
Department of Psychology
University of Texas, Austin, 78712
Pennebaker@mail.utexas.edu
(512) 232-2781

ABSTRACT

In this note we introduce a new methodology that combines tools from social language processing and network analysis to identify socially situated relationships between individuals, even when these relationships are latent or unrecognized. We call this approach social language network analysis (SLNA). We describe the philosophical antecedents of SLNA, the mechanics of preprocessing, processing, and post-processing stages, and the results of applying this approach to a 15-month corporate discussion archive. These example results include an explicit mapping of both the perceived expertise hierarchy and the social support / friendship network within this group.

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 communicative social beings, humans are profoundly influenced by activities, attitudes, beliefs, and behaviors expressed at a communal level. We actively leverage relationships to both make sense of the world and select optimally among our available choices in a socially situated way, with the salience of various groups waxing and waning in different contexts. In any given group, however, the informal organization that structures and defines processes such as sensemaking is often not explicit or even consciously recognized by participants. Within organizations, similar amorphous behavior and decision shaping concepts (such as 'culture' and 'norms') are recognized and discussed, but do not have computable formulations. Driven by research questions that seek to bring to the fore these intangible yet powerful influences, we have developed a new quantitative approach that leverages the ability of social language processing to identify psychological, social, and emotional undercurrents in interpersonal communication with the structural insights of network analysis. We call this approach social language network analysis (SLNA). We believe the understanding provided by application of SLNA has immediate application for organizations trying to create efficient group structures that facilitate performance or improve employee retention. SLNA results also have potential theoretical value by providing a means to address questions such as the role of social relationships in reinforcing work relationships, and the emergence of coordination in groups.

CSCW 2010, February 6–10, 2010, Savannah, Georgia, USA.

BACKGROUND

Social Network Analysis

Social network analysis (SNA) measures and represents the regularities in the patterns of relations among entities. SNA is predicated on the concept of the relational tie as an essential building block, focusing on social structure via a collection of methods. Three decades ago, Tichy [20] pointed to the stable patterns of interaction within the social groupings of an organization as especially suitable for analysis of the causes and consequences of these relationships.

Social network analysis has been an important methodology in quantifying informal structure and group processes. In studying small group work, researchers have used network analysis to study the effect of relationships on performance in both academic [1] and business [19] settings. Hossain et. al. [5] showed a statistically significant relationship between network centrality in Enron email and project coordination. The strength of a knowledge transmission network between divisions in a company predicts time to complete a project [4]. Finally centrality in an advice network, not job rank, predicts obtaining high status privileges such as acceptance, the ability to take risk, and information access [6].

SNA researchers have also constructed networks from actual communication data. Tyler and colleagues looked at email messages sent between employees in a large company to confirm working relationships [21]. Mutton [12] showed a new technique to create a communication network between speakers based on references and collocated responses in conversations. Characteristic of SNA, these applications focus primarily on link existence, as opposed to SLNA's

concentration on the rich, detailed structure of the communicative content.

Social Language Processing

Social language processing is built on the idea that language conveys information beyond the literal meaning of the words used. Empirical studies have shown that the way in which people use language can reveal information about their thoughts and emotions [3]. Linguistic Inquiry and Word Count (LIWC) was designed to measure word use in psychologically meaningful categories. LIWC has been successfully used to identify relationships between individuals in social interactions, including relative status (e.g. [17]), deception (e.g. [13]), and the quality of close relationships (e.g. [18]). Certain word categories are relevant in demonstrating relationships between individuals. Pronoun use provides information about how people are referencing each other. Social and affective words can reveal whether someone is socially focused and their degree of emotionality. Discourse markers, such as punctuation, can show how formal or informal the language being used is.

In social language processing the goal is to infer traits of individuals, such sex, age, relative status, or mental health based on language use [15]. The focus is on the individual (e.g. mental health) or loosely the role of the individual within the group (e.g. relative status) but without any complex articulation of the group structure. There is also a tradition within the related field of sociolinguistics of combining social network analysis and linguistic style to understand linguistic variation with respect to social position [11]. In these studies, a network of individuals is first derived using traditional social network analysis methods and language is superimposed upon this network. For example, Paollilo [14] constructed a network of individual Internet Relay Chat room (IRC) users based on frequency of interaction between participants, and showed more central members used more obscenities and less code switching. In contrast, SLNA uses language variation to elucidate internal structure rather than using an external definition.

Social language processing assesses behavior (speech patterns) that individuals are not consciously aware of and therefore may reflect more accurately than surveys or other self-reporting mechanisms the processes underpinning interpersonal communications. We argue that by using the linguistic content of communication in SLNA's quantitative models we can discover relationship subtleties missed by content-agnostic SNA analyses.

METHOD

SLNA consists of three interrelated processing steps. The first step, preprocessing, involves preparing communication data for social language analysis. Since subsequent analysis steps assume a network of dyadic ties, each unit of text data must be assigned as linking one or more dyadic pairs in the group. For example, for email, the newly authored portion of each email body forms the data unit, and it is assigned to a series of dyadic links, each from the author to an individual recipient. Once all such data have been assigned to appropriate directed links between the participants, the preprocessing step is complete. The second step, processing, involves converting text associated with particular links to a quantitative metric. Typically the quantitative metric is constructed according to a particular psychological, social, or emotional theory, such as the observation that the use of the first person plural pronoun 'we' is often used as a marker of in-group belonging. Metrics may need to be normalized in some fashion. For example, metrics may be normalized to sum to unity either per recipient (in-bound normalization) or per originator (out-bound normalization). Ratio metrics are typically computed per data unit, and then averaged as opposed to aggregating the text data first then computing a metric; metric averaging provides results that are more robust to variations in sizes of the data sets associated with each directed link. Because we are using the data to connect individuals to those they communicate with in a graph-based framework, the output of this step is a series of valued adjacency matrices, one for each metric computed. The third and final step, post-processing, uses one or more of these quantitative metric matrices (see the friendship example below) in a graph-processing algorithm to compute an objective of interest. For reasonably sized graphs, visualization of the results may be helpful.

EXAMPLE APPLICATIONS

This approach has been applied to an archive of work-related conversations in a scientific research and development (R&D) organization. Twenty-two individuals used a Jabber-based chat client to evaluate, discuss, and plan advanced high performance computing. Messages sent using the public chat program were recorded for a period of 15 months, from September 2006 to November 2007. Four individuals were excluded from the study because they had typed fewer than 250 words during the study period. The remaining 18 participants included 7 females and 11 males, from 22 to 64 years old.

These data were preprocessed into relational conversations based on natural time sequences in the data. Conversations were defined as consecutive messages without more than a 5-minute delay between responses (see [7]). We selected for further analysis only those conversations in which at least two individuals interacted; this was a subset of 517 conversations. Conversations are assumed to be solely between those participants synchronously participating. This is a simplification, since the chat room persisted up to the last 100 lines of chat history for absent clients, but it accurately describes the majority of conversations.

The language associated with each relational link was then processed using the LIWC program, resulting in valued adjacency matrices across 80 linguistic dimensions. Post processing in SLNA is application specific, and so is discussed further in the following two examples.

Socially Constructed Group Status Hierarchy

In group work, effective task decomposition, delegation, and result integration depend on shared perceptions of expertise, competence, and engagement [2]. Particularly in knowledge work where the total scope of the problem exceeds any individual's knowledge, socially constructed beliefs about relative expertise define how problems are tackled collaboratively.

To assess the group-level attitude toward the expertise of its members, we used a normalized adjacency matrix measuring first person singular pronoun (e.g. "I", "I've", "me", "mine") usage in chat conversations. Out-bound normalization converted the raw LIWC counts to the proportion of personal pronouns used with each conversant. Previous studies [9] have shown that usage of this class of pronouns (unconsciously) increases as a speaker interacts with a person of higher status. Thus the relative value on each arc between team members measures the extent to which the originator of the arc views the receiver of the message as being of higher class. We then post-processed this matrix with the Google PageRank™ algorithm, effectively using each team member's language to 'vote' for the individuals with the highest status. The results of this analysis suggested that the status hierarchy, in terms of roles, is: Group Leads, Programmers, Analysts, Manager, Students and Matrixed Staff. This hierarchy corresponds exactly to previous ethnographic findings about the culture of the R&D organization, where technical skill-based roles are prized above the compliance-centric role of management, and working within one's own organizational out ranks cross-organizational work-for-hire roles. This technology-centric hierarchy suggests that status is at least in part a function of expertise for this work-based group.

Because the PageRank™ algorithm is a Markov-chain analysis, we can also impose a prior distribution upon it, and evaluate an individual's perception of the expertise hierarchy. Evaluating the perspective of the group's manager against that of the entire group (see Figure 1) reveals some interesting insights. As noted above, the group values reflect a bimodal distribution, with a higher status group (Persons A-J) and a lower status group (Persons K-R). The manager's perspective (shown by the darker bars) is largely characterized by a 'retention bias' – the manager actually overvalues the team's top talent and undervalues the lesser performers, relative to the group. In other words, the manager is more concerned about losing a 'star performer' than rank-and-file members of the group. Person G and Person I, however, have anomalously low rankings from the manager's perspective despite being members of the high (internal) status group. Both these individuals experienced value-of-contribution recognition problems with this manager after the period of this study.

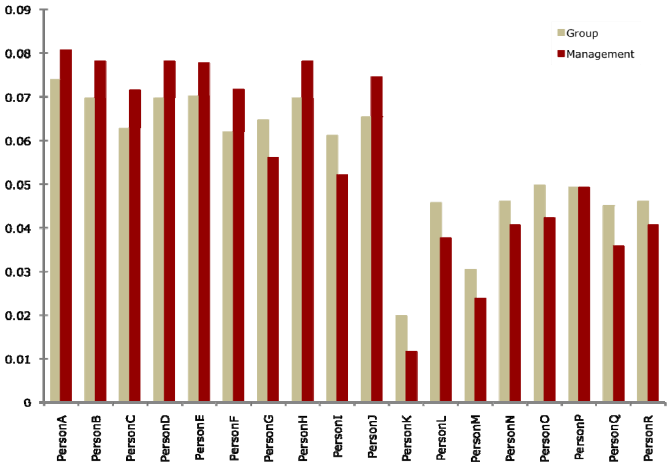

Figure 1. Group and Management Perceptions of Status.

Group Support

Groups are known to be a source of social support to their members. We applied SLNA to identify friendship within this group using a statistically derived model instead of a model from the literature. We first manually coded (4 coders, Cronbach's alpha for inter-coder reliability 0.821) each two-person conversation in the chat data as overtly friendly or not. We then ran a logistic regression using the coded response as the binary outcome variable and back selected LIWC categories as the predictors. With an alpha level for removal of 0.01, we derive a model for combining the values of the Number, Dash, and Apostrophe adjacency matrices to predict friendliness between individuals:

$$A_{ij} = e^{0.358 \bullet \text{Number}_{ij}} * e^{0.129 \bullet \text{Dash}_{ij}} * e^{0.219 \bullet \text{Apostrophe}_{ij}} \quad \text{(Eq. 1)}$$

where e^x represents the exponential function and subscripts 'i' and 'j' represent individuals in the adjacency matrixes. A relative ranking of each person's friendship to everyone else in the network is then computed by a weighted number of independent paths algorithm [23] across this combined model graph. In a survey-based evaluation (82% response rate) comparing this algorithm to three other algorithms and a 'none of these' option, 61% of respondents agreed this friendship ranking was accurate, compared to 31% for a ranking based solely on frequency of conversation.

DISCUSSION

Social language processing is an acknowledged probabilistic approach [3], and recent work suggests that any given communication medium – email, phone, instant message, videoconferencing, face-to-face meetings – carries only a portion of the total discourse on any given topic [16]. Both example SLNA applications discussed above, however, were able to reconstitute a sufficiently holistic approximation of the underlying processes to match external accuracy measures by leveraging the network. In other words, the use of the whole network reconstitutes sampling gaps at an individual level precisely because social networks are not random networks. Clustering, transitive closure, shared

perceptions and views among close friends and other well-known group-based social phenomena provide redundant information that appropriate algorithms can leverage. This means that the ability of social language processing to access information only partially under the conscious control of the speaker gives insights into whole group and organizational dynamics not otherwise obtainable. We must add the caveat, however, that development of explanatory SLNA metrics is a non-trivial and inherently explorative process due to both the open-ended possibilities of combining social language metrics and the plethora of network algorithms available to process the resulting networks.

CONCLUSION

As Weick [22] noted with the quote, "How can I know what I mean until I see what I say?," communication negotiates meaning out of the events around us. Lave and Wagner [10] interpret on-going dialog across a spectrum of expertise as central to participation in communities of practice. Social language processing suggests, however, that this same communication is also richly layered with information about the relative social, psychological, and emotional connections that situate us within a community. Social network approaches can construct higher order structures from these attributional and dyadic data. In this note, we argue for the importance of fusing these theories into a new methodology, social language network analysis (SLNA), and demonstrate how the application of SLNA to a real world knowledge-intensive collaborative work communication corpus [8] highlights and makes explicit important components of organizational functioning, such as information exchange and evaluation (a function of perceived expertise) and social support.

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. Baldwin, T.T., Bedell, M.D., and Johnson, J.L., The Social Fabric of a Team-Based M.B.A. Program, *Acad. of Management Journal*, 40, 6 (1997), pp. 1369-1397.
2. Borgatti, S.P. and Cross, R., A Relational View of Information Seeking and Learning in Social Networks, *Management Science*, 49, 4 (2003), pp. 432-445.
3. 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.
4. Hansen, M.T., Knowledge Networks, *Organization Science*, 13, 3 (2002), pp. 232-248.
5. Hossain, L., Wu, A., and Chung, K.K.S, Actor Centrality Correlates to Project Based Coordination, *Proc. CSCW'06*, ACM Press (2006), pp. 363-372.
6. Ibarra, H. and Andrews, S. B., Power, Social Influence, and Sense Making, *Administrative Science Quarterly*, 38 (1993), pp. 277-303.
7. 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.
8. Julsrud, T.E, Core/periphery Structures and Trust in Distributed Work Groups, *Structure and Dynamics*, 2, 2 (2007) pp. 1-28.
9. Kacewicz, E., Pennebaker, J.W., Davis, M., Jeon, M., and Graesser, A.C. (under review). The language of status hierarchies.
10. Lave, J. and Wagner, E., *Situated Learning: Legitimate Peripheral Participation*, Cambridge Press (1994).
11. Milroy, L. and Milroy, J., Social network and social class: towards an integrated sociolinguistic model. *Language in Society*, 21, 1 (1992), pp. 1-26.
12. Mutton, P., Inferring and Visualizing Social Networks on Internet Relay Chat, *IV 2004*, (2004), pp. 35-43.
13. 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.
14. Paollilo, J.C., Language variation on Internet Relay Chat: A social network approach, *Journal of Sociolinguistics*, 5, 2 (2001), pp.180 – 213.
15. Pennebaker, J.W., Mehl, M.R., and Niederhoffer, K., Psychological aspects of natural language use, *Annual Review of Psychology*, 54 (2003), pp. 547-577.
16. Pentland, A. *Honest Signals*, MIT Press (2008).
17. Sexton, J.B. and Helmreich, R.L., Analyzing cockpit communications, *Human Performance in Extreme Environments*, 5, 1 (2000), pp. 63-68.
18. Slatcher, R.B. and Pennebaker, J.W., How do I love thee? *Psychological Science*, 17 (2006), pp. 660-664.
19. 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.
20. Tichy, N.M., Tushman, M.L., Fombrun, C., Social Network Analysis for Organizations, *The Academy of Management Review*, 4, 4 (1979), pp. 507-519.
21. Tyler, J., Wilkinson, D., and Huberman, B., Email as Spectroscopy, (2003), http://www.hpl.hp.com/shl/papers/email/index.html
22. Weick, K.E., *Sensemaking in organizations*, Sage Publications (1997).
23. White, S. and Smyth, P., Algorithms For Estimating Relative Importance In Networks, *KDD '03*, ACM Press (2003), pp. 266-275.