Psycholinguistic Network Analysis

First Author Name (Blank if Blind Review)
Affiliation (Blank if Blind Review)
Address (Blank if Blind Review)
e-mail address (Blank if Blind Review)
Optional phone number (Blank if Blind Review)

Second Author Name (Blank if Blind Review)
Affiliation (Blank if Blind Review)
Address (Blank if Blind Review)
e-mail address (Blank if Blind Review)
Optional phone number (Blank if Blind Review)

ABSTRACT

In this note we introduce a new methodology that combines tools from psycholinguistics and network analysis to identify socially situated relationships between individuals, even when these relationships are latent or unrecognized. We call this approach psycholinguistic network analysis (PNA). We describe the philosophical roots of PNA, 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.

Author Keywords

Guides, instructions, author's kit, conference publications.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

As communicative social beings, humans are profoundly influenced by activities, attitudes, beliefs, and behaviors expressed at a communal level. We actively leverage social 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. Within 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. We have developed a new quantitative approach that leverages the ability of psycholinguistics to identify psychological, social, and emotional undercurrents in interpersonal communication with the structural insights of network analysis. We call this approach psycholinguistic network analysis (PNA) to both recognize and distinguish it from the disciplines of Social Network Analysis (SNA) and Psycholinguistics. We believe the understanding provided by application of PNA has immediate application for organizations trying to create efficient group structures that facilitate performance or improve employee retention. PNA 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.

BACKGROUND

Social Network Analysis

Social network analysis (SNA) measures and represents the regularities in the patterns of relations among entities (Knoke, 2008). SNA builds on the concept of the relational tie as an atomic entity, focusing on social structure via a collection of methods (Scott, 2000). Three decades ago, Tichy (1979) 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. Although these early studies typically relied on letters, memos, organizational charts, meeting minutes, survey data, interviews, and direct observation to provide data on the social networks of interest, these materials were not computationally processed.

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 friendship relationships on group performance (Baldwin, Bedell, & Johnson, 1997). Among business students assigned to groups to complete class projects, groups with more friendship ties outside the group had lower group grades on projects. Higher centrality in advice networks and lower centrality in adversarial networks in a workplace predicted higher individual ratings of performance by outside leaders (Sparrowe, Liden, Wayne, & Kraimer, 2001, Hossain 2006). In addition, higher density of adversarial network predicted lower group performance. The strength of knowledge transmission, measured using network analysis, between divisions in a company predicts time to complete a project (Hansen, 2002). 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 (Ibarra & Andrews, 1993).

SNA researchers have also constructed networks from actual communication data in a workplace. Tyler and colleagues looked at email messages sent between employees in a large company to confirm working relationships (Tyler et al., 2003). Mutton (2004) showed a new technique to create a communication network between speakers based on references and collocated responses in conversations.

Psycholinguistics

Psycholinguistics is built on the idea that language conveys information beyond the literal meaning of the words used. Empirical studies have shown that they way in which people use language can reveal information about their thoughts and emotions (Chung & Pennebaker, 2006). 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. Sexton & Helmreich, 2000), deception (e.g. Newman, Pennebaker, Berry, & Richards, 2003), and the quality of close relationships (e.g. Slatcher & Pennebaker, 2006). 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.

Psycholinguistics 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 believe that by using the linguistic content of communication we can discover relationship subtleties missed by content-agnostic SNA analyses.

METHOD

PNA consists of three interrelated processing steps. The first step, preprocessing, involves preparing communication data for psycholinguistic analysis. Since subsequent analysis steps assume a network of dyadic ties, each atomic unit of data must be assigned as linking one or more dyadic pairs in the group. For example, for email data, the newly authored portion of each email body forms the atomic data unit, and it is assigned to dyadic links from the author to each of the recipients. Once all such atomic data units have been assigned to appropriate links between the participants, the preprocessing step is complete. The second step, processing, involves converting link-specific text to a quantitative metric. Typically the quantitative metric is constructed according to a particular psychological, social, or emotional theory or stylized fact, such as the observation that the use of the first person plural pronoun 'we' is often used as a marker of in-group belonging, while the use of the pronoun 'they' also is used by groups as defining out-group individuals. Metrics may need to be normalized in some fashion (for example, per recipient if attention is conserved, or per originator if theory suggests energy is a more binding constraint). Ratio metrics are typically computed per atomic data unit, and then averaged as opposed to aggregating the text data first then computing a metric; metric averaging provides results that do not statistically correlate with the link data set size. 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 the 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. Because psycholinguistics are able to identify psychological, social, and emotional processes that individuals are not able to fully mask, this graph processing can clarify and highlight complex interdependencies between group members which may be latent and unrecognized.

EXAMPLE APPLICATIONS

This approach has been applied to an archive of work-related conversations in a scientific research and development organization. Twenty-two individuals used a Jabber-based chat client to evaluate, discuss, and plan advanced high performance computing modeling and simulation. Messages sent using the public chat program were recorded for a period of 15 months, from September 20, 2006 to November 15, 2007. Four individuals were excluded from the study because they had spoken fewer than 250 words during the study period. The remaining 18 participants included 7 females and 11 males, from 22 to 64 years old. At the time the messages were sent to the public chat forum all participants were aware that their comments were being recorded.

This data was 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 Issacs, Walendowski, Whittaker, Schiano, & Kamm, 2002). There were a total of 1013 conversations, the majority of which were announcements by a single speaker. 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 320 valued adjacency matrices across 80 linguistic dimensions.

Post processing in PNA is application specific, and so is discussed further in the following two examples.

Socially Constructed Group Expertise Hierarchy

In group work, effective task decomposition, delegation, and result integration depend on shared perceptions of expertise, competence, and engagement. (Cross & Parker? Hinds?). 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 matrices measuring first person personal pronoun (e.g. "I", "I've", "me", "mine") usage in chat conversations. Normalization converted the raw LIWC counts to the proportion of personal pronouns used with each conversant. Previous studies [Pennebaker] 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 allowing each team member to 'vote' for the individuals with the highest status. In a work-based group, we contend that status is a function of expertise. 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 'stylized facts' 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.

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. This approach is more than just a direct evaluation of who in the group the individual is directly deferential to, as the opinion of the most respected individuals also factors into the final ranking. Evaluating the perspective of the group's manager against that of the entire group (see Figure 1) revealed two interesting insights. One, there is 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. Two, there were two anomalously low rankings of members of the team (Person G and Person I in Figure 1), again relative to the group norm. Both these individuals experienced value-of-contribution recognition problems with this manager after the period of this study.

Group Support

Groups are known to be a source of social support to their members. We applied PNA to identify friendship within this group, as these results could be shared with group members for evaluation without substantial risk. We first hand coded (4 coders, Cronbach's alpha 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 selected LIWC categories as the predictors. With an alpha level for removal of 0.01, we arrive at a model for combining the values of the Number, Dash, and Apostrophe adjacency matrices:

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

where e^x represents the exponential function and subscripts 'i' and 'j' represent the position in the adjacency matrixes. A relative ranking of strength friendship to individuals in the network for each person can then be computed by a weighted number of independent paths algorithm (White and Smyth, 2003) across this combined model graph. This approach is surprisingly good at identifying the relative strength of friend ties. In a survey-based evaluation (82% response rate), 61% of respondents agreed ranking provided by this algorithm was accurate, double the rate of a ranking based solely on frequency of conversation.

DISCUSSION

Psycholinguistics is an acknowledged probabilistic approach (Pennebaker), 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 (MIT, 2009). Both example PNA 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 psycholinguistics 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 these PNA metrics is non-trivial. The size of the set of networks created by the combination of psycholinguistic metrics is constrained only by the imagination and creativity of the researcher. Even when the number of candidate networks can be constrained by theoretical considerations or data fitting as described above, the plethora of network algorithms provides another source of combinatoric explosion in solution space. The discovery of new explanatory PNA methods is an inherently explorative process.

CONCLUSION

As Weick (1997) noted with the quote, "How can I know what I mean until I see what I say?," communication is central to negotiating meaning out of the events around us. Lave and Wagner (1994) interpret on-going dialog across and within a spectrum of expertise as central to legitimate participation in communities of practice. Psycholinguistics suggests, however, that this same communication is also richly layered with information about the relative social, psychological, and emotional connections that connect and 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, Psycholinguistic Network Analysis (PNA), and demonstrate how the application of PNA to a real world knowledge-intensive collaborative work communication corpus (Julsrud, 2007) highlights important components of organizational functioning, such as information exchange and evaluation (a function of perceived expertise) and social support can be made explicit.

ACKNOWLEDGMENTS

We thank the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories for providing funding through the Seniors LDRD Council.

REFERENCES

1. Adobe Acrobat Reader 7. http://www.adobe.com/products/acrobat/.
2. Anderson, R.E. Social impacts of computing: Codes of professional ethics. *Social Science Computing Review 10*, 2 (1992), 453-469.
3. How to Classify Works Using ACM's Computing Classification System. http://www.acm.org/class/how_to_use.html.
4. Klemmer, R.S., Thomsen, M., Phelps-Goodman, E., Lee, R. and Landay, J.A. Where do web sites come from? Capturing and interacting with design history. In *Proc. CHI 2002*, ACM Press (2002), 1-8.
5. Mather, B.D. Making up titles for conference papers. *Ext. Abstracts CHI 2000*, ACM Press (2000), 1-2.
6. Schwartz, M. *Guidelines for Bias-Free Writing*. Indiana University Press, Bloomington, IN, USA, 1995.
7. Zellweger, P.T., Bouvin, N.O., Jehøj, H., and Mackinlay, J.D. Fluid Annotations in an Open World. *Proc. Hypertext 2001*, ACM Press (2001), 9-18.
8. White, S. and Smyth, P., Algorithms For Estimating Relative Importance In Networks, Ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Washington, D.C., August 24 - 27, 2003). KDD '03. ACM, New York, NY, 266-275.