Computing Team Process Measures from the Structure and Content of Broadcast Collaborative Communications

Anup K. Kalia, Norbou Buchler, Arwen DeCostanza, and Munindar P. Singh, *Fellow, IEEE*

***Abstract*—Existing approaches to compute team process measures are primarily based on survey ratings, semantic classification of communications, and social network analyses. Although existing approaches reveal important information about team performance, they face specific limitations. Survey methodologies are in general unreliable, biased, and not dynamic; communication classifications are often a-theoretical; and social network analytics ignore the meanings of messages. Accordingly, we develop a better-defined, formal, empirical approach for computing team process measures.**

Our contribution builds on existing work in semantic classification of messages in broadcast communications and proposes a general set of meanings of messages for team processes. Using the meanings of messages, we propose formal approaches to compute team process measures. We evaluate these measures using a military dataset and find the following: (1) our text mining approach to infer meanings of messages significantly improves over the bag of words approach and yields macro and micro average F-measures of 70% and 80%, respectively, and (2) compared to baseline measures such as degree centrality, cognitive processes remain significantly stable with time whereas measures such as affective process significantly increase with time.

***Index Terms*—team performance, team processes, emergent states, text mining**

I. Introduction

In an organization, teams are composed of members that share one or more goals, display task interdependencies, perform within an organizational context, influence and motivate each other as well as members from other teams, and evolve as a cohesive group over time [23]. Often such teams work in collaborative environments where information and decision-requirements change quickly and are exacerbated by time-criticality. Example domains include stock trading, incident command, military operations, and critical healthcare. Advancements in our ability to assess the dynamic foundations of team performance in such situations is crucial to understanding, predicting, and augmenting team successes and failures.

Rather than focusing on performance outcomes, which do little to prescribe reasons for success or failure, the current research aims to advance assessment methodologies or measures by focusing on the group dynamics, or team processes, that are critical aspects of team performance across situations. Kozlowski and Ilgen [24] present a dynamic view of team processes based on the Input-Mediator-Output-Input (IMOI) framework [18]. According to the IMOI framework, team processes are of three types: affective, behavioral, and cognitive processes, described in more detail below.

Affective Processes. Affective processes capture motivational and affective relationships between team members. Affective processes are dynamic and include the constructs: trust, cohesion, confidence, and conflict. *Trust* refers to the willingness of a team member (as a truster) to be vulnerable to the actions of another team member (as a trustee) [29]. *Cohesion* refers to motivation among team members to identify with and remain within the group. Specifically, cohesion relates to member's attraction to their team (task commitments) [5]. *Confidence* refers to the shared belief that a team is capable of executing actions [26], [40]. *Team conflict* refers to the revelation of conflicting opinions among team members [2]. Dreu and Weingart [11] discover that task and relationship conflicts negatively correlate with team effectiveness.

Behavioral Processes. Behavioral processes refer to actions taken by team members toward goal accomplishments. Marks et al. [27] consider behavioral processes as a sequence of episodes containing three phases: action, transition, and interpersonal. During the *action phase*, a team monitors goal progress, provides back-up behavior, and supports coordination. In the *transition phase*, a team performs mission analyses, specifies goals, strategies, and plans. The *interpersonal phase* occurs in parallel with the action and transition phases. In this phase, team members are motivated to accomplish their goals. Kozlowski and Bell [23] consider behavioral processes as actions that involve coordination, cooperation, and communication among team members.

Cognitive Processes. Cognitive processes refer to the process of acquiring and organizing knowledge for decision-making. Knowledge is acquired using cognitive constructs such as team mental models [28] and team situational awareness [32]. Cognitive constructs guide team members in performing task-relevant interactions and accomplishments [22].

Approaches to Measure Team Processes. Existing techniques to compute team process measures employ rating or surveys, communication classification, and social network analysis. Each of these techniques has specific limitations, as described below.

Observer or Self Rating Based Approaches. In these approaches, external observers provide ratings or evaluations for

Anup K. Kalia is with IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA.

Norbou Buchler and Arwen Decostanza are with US Army Research Lab, Aberdeen Proving Ground, MD 21005, USA.

Munindar P. Singh is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA.

a member's behavior in a team, i.e., how he or she performed. Examples of observer ratings include (1) Teamwork Observation Measure (ATOM) [33], which provides quantitative information from observers (ratings) for team processes, and (2) Targeted Acceptable Responses to Generated Events or Tasks (TARGETS) [15], which provides evaluations from observers for generated time-ordered events or tasks. Examples of self-rating approaches include (1) Situational Awareness Rating (SART) [35], which captures subjective measures of situational awareness and (2) Group Environment Questionnaire (GEQ) captures team cohesion [8].

Limitations of Rating Based Approaches. In observer-ratings approaches, multiple observers provide ratings for qualitative evaluations which is time consuming, and can be unreliable, biased, or incomplete. Self-ratings approaches too are susceptible to bias. Also, interrupting team members to provide ratings on multiple measures at frequent intervals can influence their team processes and resulting team performance.

Communication Classification Strategies. These approaches identify types of illocutions [4] from team members' communications indicative of team processes. Entin and Entin [13] provide observers a communication matrix to code communications as *transfers*, *requests*, and *acknowledgments*. They classify *transfers* and *requests*, further, as *requests for information*, *actions*, and *coordination*. Fisher et al. [14] classify communications as *information sharing*, *problem solving*, and *team coordination*. Cooke et al. [9] provide an approach using latent semantic analysis (LSA) to infer meanings in team members' dialogs.

Limitations with Communication Classification Strategies. Communication classification strategies are limited to manual coding and techniques such as latent semantic analyses. Manual identification of classes or labels is labor intensive. The latent semantic analyses (LSA) approach is similar to the bag of words approach where the order of words is not important. Without considering the order of words the approach can produce a huge set of features that might lead to sparsity and over-fitting especially in small datasets. The approach is domain-specific since it is based on extracting specific word pairs instead of extracting generic features.

Social Network Analyses. In these approaches features such features such as *degree centrality*, *clique size*, *internal density*, *network constraint*, and *tie strength*, are extracted from a social network and related to team performance. Sparrow et al. [34] found that social network measures such as centrality and density relate to team performance. Ehrlich and Cataldo [12] determine that out-degree centrality is more related to productivity than in-degree centrality. Henttonent et al. [16] discover that network density relates to team performance.

Limitations with Social Network Analysis. In a social network, out-degree and in-degree centralities reflect outgoing and incoming edges that do not reveal their meanings. For example, an outgoing edge can be a message indicating a request for information or a directive, whereas an incoming edge can be a message indicating an informative or an acknowledgment. Thus, a social network analysis reveals little more than the structural patterns of collaborative interactions and says nothing about the information and content of the edges in the network. We assume that creating a model based on the information and content will relate team performance stronger than social network measures.

Contributions. The current paper addresses some of the limitations of existing approaches to compute team processes, to understand the root causes of team performance outcomes. Specifically, we contribute the following:

- We provide a general classification scheme based on meanings of messages by unifying existing schemes considering that existing contributions do not provide a generic scheme related to team process measures.
- We provide automated feature extraction and text mining approaches to classify general types of broadcast communications and compare our approach against the existing bag-of-words approach.
- We provide computational approaches to derive formal team process measures from the interactive patterns of communications between team members by mapping to specific team process. We compare our approaches with existing social network based measures that indicate team performance.

Finally, we provide a comparative evaluation of team process measures computed against structural baselines. Specifically, we ask the following research questions.

RQ$_1$. *How accurate is our feature extraction and text mining approach in automatically classifying messages in broadcast communications?*

RQ$_2$. *As an expression of team performance, how stable or variable are the proposed formal measures of team performance across developmental time?*

We evaluate the above research questions using a military dataset obtained from a division-level exercise focused on military planning and decision-making. The participants in the exercise were active duty soldiers and officers with operation staff experience. The participants operated at various levels such as Division, Brigade, Battalion, and supporting units. The units operated in a distributed fashion over a communication network using specialized millitary command and control hardware and software. Within each unit, participants carried out duties of different functional areas. Individual responses and responsibilities to a given scenario event in the training exercise depended upon adherence to established workflows and standard operating procedures both within the unit and in pursuing functional requirements. Workflows represent the responsibility for performing various tasks and sub-tasks necessary for mission success. The training scenario in the military exercise generates many overlapping series of event-driven tasks, the result of which requires a high degree of coordination among the participants.

II. Model for Team Process Measures

To compute team process measures, we provide an approach that takes the input as broadcast communication and outputs team process measures for each team member. In our approach, first, we infer meanings of messages, and then create social constructs based on these meanings. We use both meanings of messages and social constructs to compute team process measures.

Broadcast Communication. This type of communication contains messages that are publicly sent by a group member (sender). Such messages are visible to all group members and anyone (responder) in the group can respond to them.

Meanings of Messages. To compute team process measures from broadcast communications, we propose to understand meanings of messages sent or broadcast by team members. We consider meanings of messages as different classes (illocutions) for messages. Several existing works propose such classes to study their relationships to team performance. Entin and Entin [13] classify communications between team members into three types: transfer, requests, and acknowledgments. Fisher et al. [14] classify communications into three types: task-related, social, and responses. Kalia et al. [19] classify communications into three types: questions, directives, and informatives. A limitation with existing approaches is that none of them claim to provide a generic set of constructs for computing team process measures (affective, behavioral, and cognitive). Thus, we propose the following classes by synthesizing the existing approaches.

- **Questions (Q)**. An inquiry made by a sender. Examples are “What is the phone # for the 344 S3?”, “Has anyone gained visual contact with enemy elements?”, and “B, whats your status on personnel and equipment?”
- **Directives (D)**. An order by a sender. Examples are “COs, send all reports up to BN over this net”, “give me the grid and your grid”, and “send up locations with the contact reports.”
- **Requests (R)**. The sender asks for resources. Examples are “requesting CAS time now”, “i just requested a majic kill for 6 bad dudes”, and “1 kia request EVAC.”
- **Commissives (C)**. A commitment or willingness of the sender to perform a task. Examples are “will breech momentarily”, “I’m also going to try and figure out what killed them”, and “I’ll just send any further guidance myself over transverse.”
- **Informatives (I)**. A piece of information or report provided by the sender, e.g., about an action or activity being performed. Examples are “no casualties, have all equipment except for Ravens… not sure what”, “Sir, our platoon is currently moving to our AA north of the airfield”, “UAZ-4690. Information can be found in event on my tree viewer”, and “engaging two EN PAX and EN VIC IVO 12SWG 67673 90834.”
- **Acknowledgments (A)**. An assertion or an appreciation sent by a sender. Examples are “rgr”, “B Co, Roger”, “ack, keep me posted on the fight”, “roger, let us know if you need CAS”, “I think were making good progress”, and “rgr, keep doing your best to make it work.”

Elements of Communication. Elements of communication include dyadic pair-wise interactions that form the basis for understanding communication events. Thus, we create the following dyadic elements of communication based on the above classes.

- *Responses to Questions* (RQ)
- *Responses to Directives* (RD)
- *Responses to Requests* (RR)
- *Responses to Commissives* (RC)
- *Responses to Informatives* (RI)

Here, the responses can be questions, directives, requests, commissives, informatives, acknowledgments, or others.

We represent dyadic pair-wise interactions using an adjacency matrix $\mathcal{S}o_{ij}$ where i and j represent actors $\mathcal{A}_i$ and $\mathcal{A}_j$, respectively, and a directed edge ij from $\mathcal{A}_i$ (responder) toward $\mathcal{A}_j$ (sender) represents $\mathcal{A}_i$’s response to $\mathcal{A}_j$’s broadcast message. $\mathcal{S}o_{ij}$ = 1 if an edge exists else $\mathcal{S}o_{ij}$ = 0. We mark $\mathcal{S}o_{ij}$ = 0 if i equals j since we assume $\mathcal{A}_i$ does not respond itself. Examples of dyadic pair-wise interactions in broadcast communications are given in Table I.

Affective Processes. According to Marks et al. [27], affective processes are constructed based on the outputs of behavioral processes and subsequently become new inputs to the next behavioral processes. We consider three affective processes: trust, confidence, and cohesion as indicators of team performance. We represent such affective processes using dyadic pair-wise interactions. We compute trust of one team member (truster) toward another (trustee) as the probability of the positive evidence gained by the truster from its interactions with the trustee. We compute confidence of one team member toward another as how accurate is the estimation of the team member’s trust for another. Confidence reflects the knowledge of a team member about a colleague’s past performances. We compute cohesion as the average response time delay between messages sent and received between team members.

a. Trust. Mayer et al. [29] define trust as the willingness of a truster to be vulnerable to the actions of a trustee. Trust can be of two kinds [30]: (1) *cognitive trust*—based on evidence-based reasons as perceived by the truster from the trustee; (2) *affective trust*—based on emotional ties perceived by the truster with the trustee. We consider only cognitive trust since the data we evaluate focuses more on actions or tasks executed and less on emotions expressed. To compute trust from interactions in broadcast messages, we assume that trust increases between a truster and trustee if the trustee responds to the truster’s directives and requests by taking the required actions to fulfill them. For brevity, we refer to “cognitive trust” as “trust.”

We consider trust as directed from a truster toward a trustee. We represent trust as the evidence pair $\langle r,\ s\rangle$ [39], where r and s represent positive and negative evidence as perceived by the truster from its interactions with the trustee, respectively. Also, r and s are real numbers where $(r{+}s) \geq 0$. If $r + s = 0$, this means the truster did not send directives, requests, informatives, or commissives to the trustee. Hence, we treat the trust as absent and set $\alpha = 0$ as defined below.

For computing trust from interactions, Kalia et al. [21] include neutral evidence in addition to positive and negative evidence since neutral evidence occurs more frequently than other evidence. Including neutral evidence, we represent evidence for trust as $\langle r{+}0.5{*}n,\ s{+}0.5{*}n\rangle$ where we equally increment positive and negative evidence by 0.5*n. Here, n represents the neutral evidence. We choose 0.5*n since Kalia et al. [21] shows the configuration yields higher accuracy of

TABLE I
EXAMPLES OF DYADIC PAIR-WISE INTERACTIONS FROM A MILITARY CHAT DATA.

Sender	Messages	Constructs
S3	A TRP can you destroy that enemy arty?	Question
ACDR	N of NAI 2	Response (Informative)
S3	B TRP send sitrep	Directive
BCDR	1/2/3 PLTs reached LOA with 70% of PLT	Response (Informative)
BCDR	requesting CAS time now	Request
C_2	give me your grid	Response (Directive)
ACDR	A TRP UAV in air will figure-8 IVO NAI 3	Informative
C_2	Roger	Response (Acknowledgment)
ACDR	they revive 12 of my guys so I have 24 pax now	Informative
C_2	glad to hear it	Response (Appreciation)

predicting trust than choosing any other configuration. Finally, trust is computed as

$$\alpha = \frac{r + 0.5 * n}{r + s + n} \tag{1}$$

Based on the trust model and social constructs, we propose trust as the following

$$\alpha_{ij} = \frac{RD_{ji} + RR_{ji} + 0.5 * (RI_{ji} + RC_{ji} + RQ_{ji})}{RD_{ji} + RR_{ji} + RI_{ji} + RC_{ji} + RQ_{ji}} \tag{2}$$

In the above equation, α_{ij} indicates the trust of a truster i toward a trustee j and is computed based on positive and neutral evidence experienced by i from j. For example, if j responds to i's directives (RD$_{ji}$) or requests (RR$_{ji}$), we assume i gains positive evidence from j and hence set the positive evidence as r = RD$_{ji}$ + RR$_{ji}$. If j responds to i's informatives, questions, and commissives, we assume i gains neutral evidence and hence set neutral evidence as n = RI$_{ji}$ + RQ$_{ji}$ + RC$_{ji}$.

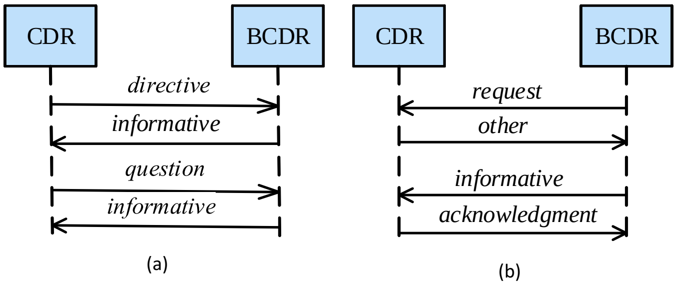


Fig. 1. Examples of measures that indicate trust.

From Figure 1, we compute CDR's trust for BCDR ($\alpha_{\text{CDR,BCDR}}$) considering that CDR perceives a positive evidence when it receives a response to its directive from BCDR and perceives a neutral evidence when it receives a response to its question from BCDR. Thus, we compute $\alpha_{\text{CDR,BCDR}}$ as $\frac{1+0.5*1}{1+1}$ = 0.75. We compute BCDR's trust for CDR ($\alpha_{\text{BCDR,CDR}}$) considering that BCDR perceives a positive evidence when it receives a response to its requests from CDR and perceives a neutral evidence when it receives a response to its informatives from CDR. Thus, we compute $\alpha_{\text{BCDR,CDR}}$ as $\frac{1+0.5*1}{1+1}$ = 0.75. To compute the overall trust score for a team member, we consider the average of all other team members' trust assignments as trusters toward the team member as the trustee.

b. Confidence. Confidence refers to the shared belief among team members that they are capable of executing actions toward the accomplishments of goals [26], [40]. To represent the idea of shared belief among team members, we consider confidence as a social construct where the confidence increases as a function of the number of interactions between any two team members. To computationally represent this idea of confidence, we use the certainty function proposed by Wang et al. [37]. Consider two team members i and j who work with each other to accomplish a mission. Thus, according to Wang et al., i's certainty about j represents the strength of belief of i that j can bring about a positive outcome. Clearly, i's certainty is dependent on i's knowledge about j's past performances (positive or negative outcomes). As i's knowledge about j increases, it should become more certain about j. Based on this idea, Wang et al.'s [37] formulate i's certainty about j as the probability of the probability of a positive outcome.

Wang et al. [37] consider evidence as the pair of positive evidence (r) and negative evidence (s). We consider such evidence as outcome of interactions between i and j. Based on evidence $\langle r, s\rangle$, we compute x that represents the probability of the positive outcome ($x \in$ [0,1]). Thus, the probability density function of x given $\langle r, s\rangle$ can be represented as

$$f(x|\langle r, s\rangle) = \frac{x^r(1-x)^s}{\int_0^1 x^r(1-x)^s dx} \tag{3}$$

Since, $f(x)$ represents the probability density function, the mean of $f(x)$ can be computed as $\frac{\int_0^1 f(x)dx}{1-0}$=1 considering f: [0, 1] $\mapsto$ [0, ∞]. If i knows nothing about j, f(x)=0 else if i knows something, f(x) > 1 or f(x) < 1 for $x \in$ [0,1] (above or below the mean 1). i's overall knowledge or confidence about j as can be represented as the mean absolute deviation (MAD) from the uniform distribution. Wang et al.'s [37] consider MAD as certainty and compute it as $\frac{1}{2}\int_0^1$|f(x) - 1|dx. Thus, based on Equation 3, Wang et al. represent certainty (or confidence) based on evidence $\langle r, s\rangle$ as

$$\mathbf{c}(r, s) = \frac{1}{2}\int_0^1 |\frac{x^r(1-x)^s}{\int_0^1 x^r(1-x)^s dx} - 1| dx \tag{4}$$

We differentiate between trust and confidence by considering α as the trust and **c**(r, s) as the confidence. To compute confidence from communications, similar to trust, we consider r as the positive evidence and s as the negative evidence experienced by i from j. We compute r as the sum of responses made by j to directives and requests made by i (r = RD_{ji} + RR_{ji}). We we include neutral interactions by considering r and s as r+0.5*n and s+0.5*n respectively. We compute neutral interactions as the sum of responses made by j to the informatives, questions, and commissives made by i (n = RI_{ji} + RQ_{ji} + RC_{ji}).

Wang and Singh [38] state two properties of certainty considering the amount of conflict, which is a key indicator of low levels of certainty. They consider conflict as the ratio of positive evidence (r) to negative evidence (s). If r equals s, conflict is high. If r>>s or s>>r, conflict is low. Two properties of certainty are: first, certainty increases with increase in amount of evidence (r + s) provided the amount of conflict is fixed, and second, certainty decreases with increase in conflict provided the amount of evidence is fixed. Thus, representing confidence as certainty is natural since with increase in conflict, confidence between team members decreases and vice versa.

To illustrate confidence, consider an example where 11 interactions occurred between i and j out of which 8 were positive (r), 2 were negative (s), and 1 was neutral (n) as perceived by i. Thus, the trust of the i toward j is α= $\frac{8+0.5}{8+2+1}$= 0.77 and the confidence is **c**(8+0.5, 2+0.5) = 0.52. Now, consider another example where 22 interactions occurred between i and j out of which 16 were positive, 4 were negative, and 2 were neutral. The trust of i for j remains the same, i.e., α= $\frac{16+1}{16+4+2}$= 0.77, whereas the confidence of i for j increases to **c**(16+1, 4+1) = 0.62.

c. Cohesion Cannon-Bowers et al. [6] consider team cohesion as the willingness to take input from team members and a belief that team is more important than individual members. Beal et al. [5] propose that cohesion among team members can be of three types: interpersonal attraction, commitment to tasks, and group pride. Interpersonal attraction refers to shared liking or admiration between team members. Commitment to tasks refers to the degree to which team members are mutually committed to accomplishing their goals. Group pride refers to the liking or support shown by group (or team) members to ideologies that the group supports.

To represent the idea of willingness to contribute, we propose cohesion as the average response time delay in milliseconds between messages sent and responses to such messages between team members. We limit such messages to questions, directives, commissives, informatives, and requests since they indicate time-sensitive and mission-focused communications. We assume that when a team member j makes a quicker response to a request made by its colleague i, i can estimate how willing the team member j is to contribute toward the team. Thus, we compute cohesion (co_{ij}) as follows, where i and j represent team members and art_{ji} the average response time delay in j responding to requests from i:

$$co_{ij} = \frac{1}{1 + art_{ji}} \tag{5}$$

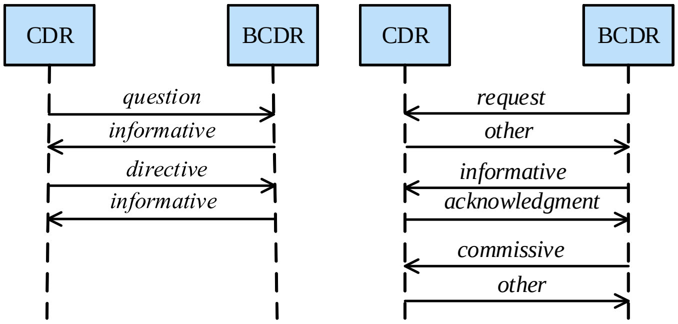


Fig. 2. Team cohesion represented as the inverse of the average response time delay.

In Figure 2, to compute CDR's cohesion toward BCDR ($co_{\text{CDR,BCDR}}$), we compute BCDR's average response time delay toward CDR ($art_{\text{BCDR,CDR}}$) by averaging response time delay made by BCDR to respond to CDR's question and directive. To compute BCDR's cohesion toward CDR ($co_{\text{BCDR,CDR}}$), we compute CDR's average response time delay toward BCDR ($art_{\text{CDR,BCDR}}$) by averaging response time delay made by CDR to respond to BCDR's request, informative, and commissive. For computing the average response time delay, we define the epoch length as the duration over which interactions happen between any two team members.

Behavioral Processes. Marks et al. [27] define behavioral processes as members' independent actions—taken on the basis of team members cognitive and motivational states and directed toward goal accomplishments—that convert team inputs to outcomes. To compute behavioral processes we consider messages broadcast by team members. Note that team members' messages are reflective of goals accomplished, actions performed, and information requirements. In some instances, the actions performed and information requirements are explicit, whereas goals accomplished are often implied. Thus, based on such instances of messages among team members, we compute behavioral process for each team member. We represent the behavioral process as the sum of five types of communication interactions—questions, directives, requests, commissives, and informatives—that are aggregated for each team member. The behavioral process measure is normalized for each team member as the ratio of such communications (questions, directives, requests, commissives, and informatives) sent by the member to the overall total number of messages sent.

$$bp = \frac{\#Q + \#C + \#D + \#R + \#I}{\#AllMessages} \tag{6}$$

Some messages may not indicate any of the above meanings. We refer such messages as *others*. We avoid such messages in the numerator for computing behavioral processes since they may not be directly relevant to the accomplishment of goals.

Whereas affective processes focus on multiple aspects of interactions, behavioral processes focus only on actions, and lead

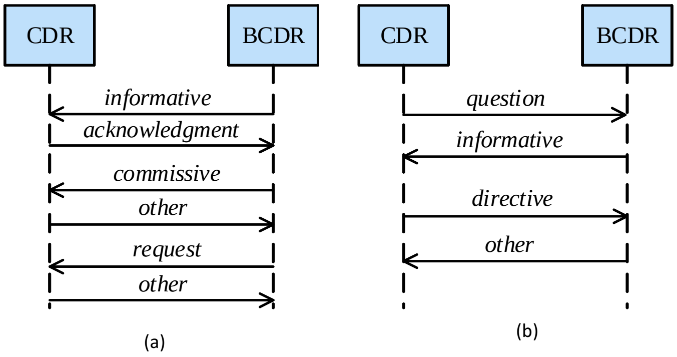


Fig. 3. Examples of interactions that indicates team processes.

to only one measure, the normalized behavioral process bp. In Figure 3, CDR broadcasts a question, directive, acknowledgment, and two other messages. Thus, bp for CDR is $\frac{1+1}{1+1+1+2}$ = 0.4. BCDR sends two informatives, one commissive, one request, and one other. Thus, bp for BCDR is $\frac{2+1+1}{2+1+1+1}$ = 0.8.

Cognitive Processes We consider cognitive process as the process of acquiring and organizing knowledge. Knowledge is represented as team mental models and team situational awareness. A team's mental model represents how knowledge is distributed across individual team members' minds [28] whereas a team's situational awareness represents team members' shared understanding about the current situation and its implications [32]. To measure cognitive processes, Cook et al. [9] suggest eliciting knowledge at (1) the individual or (2) the team level. They elicit knowledge mainly by asking questions to team members at regular intervals and tracing their actions during the team process.

a. Clarity. To determine cognitive processes from broadcast messages, we adopt Cook et al.'s [9] approach for eliciting knowledge. Specifically, we consider how team members make decisions versus questions they ask while making such decisions. Clearly, the decisions taken by team members represent the projection of their mental model and situational awareness. Questions indicate confusion or the need for knowledge to carry out certain tasks. Thus, we propose clarity, a new measure that captures a team member's decisions (directives, informatives, commissives, and requests) normalized with respect to questions. We define clarity of a team member as:

$$\mathbf{d} = \frac{\#D + \#C + \#R + \#I}{\#D + \#C + \#R + \#I + \#Q} \tag{7}$$

In Figure 4 the clarity of C_2 can be computed as $\frac{1}{1+1}$ = 0.5 (one directive and one question) and the clarity of BCDR can be computed as $\frac{2}{2+1}$ = 0.67 (two informatives and one question).

We consider clarity as one of the measure for cognitive process since the process focus on acquisition and organization of knowledge among team members.

Baselines (In-degree and Out-degree).

Since traditional approaches relate structural measures (indegree and out-degree) with team performance [12], [16], [34], [41], we adopt them as our baselines. To this end, we define a network where each team member is a node and directed edges from i to j represent messages sent from i toward j. We normalize the in-degree and out-degree, i.e., x to [0, 1] as x' = $\frac{x-x_{min}}{x_{max}-x_{min}}$.

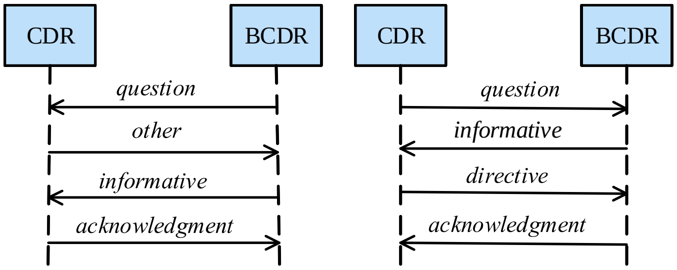


Fig. 4. Example of measures indicating clarity.

III. Extracting Meanings of Messages from Broadcast Communications

A. Extracting Features

To identify meanings of messages from broadcast messages automatically, we extract relevant features from these messages to train a classifier (e.g., Support Vector Machine). We consider meanings of messages proposed in Section II as labels or classes. The classes are questions, directives, commissives, requests, informatives, acknowledgments, and others. To train the classifier, we propose the following features.

1) **N-Grams (Baseline)**. We extract *unigrams*, *bigrams*, and *trigrams* from the training data. To extract n-grams, first, we remove stop words from sentences and then lemmatize remaining words in sentences to avoid redundant features. We employ n-grams as the baseline and the remaining features as enhancement over n-grams.
2) **Modal Verbs**. We extract modal verbs from sentences such as *must*, *shall*, *should*, *could*, *may*, and *can*. Modal verbs indicate a sentence as commissives ("will breech momentarily") or directives ("can you engage DHY?") [20], [25], [31].
3) **Action Verbs**. We check if a sentence contains an action verb since a sentence with an action verb may indicate a commissive ("I'll send further guidance"), a directive ("give me your grid"), a request ("I sent up a resupply request"), a question ("has anyone gained visual contact?"), or an informative ("engaging with SAF now"). We use lexicons prepared from the military chat dataset and the Enron email dataset to verify if a verb in a sentence is an action verb [19]. In addition to checking action verbs, we check if the verb is in the *present tense* or in some *other tenses* such as *past*, *future*, and *present progressive* [20], [31]. Present tense verbs indicate a sentence as a commissive or a directive whereas past, future, and present progressive tense verbs indicate a sentence as an informative or a request. We also check if a sentence begins with an action verb. For example, several examples of directives begin with action verbs [31] ("provide locations", "send all reports up to BN").
4) **Personal Pronouns** We check if a sentence contains personal pronouns ("I", "We", "you") [20], [31], which

feature in sentences indicating commissives ("I'll send further guidance") and requests ("I just requested a magic kill") whereas sentences indicating directives ("can you please engage DHY?") contain second person personal pronouns.

5) **Question Words**. We check if a sentence contains a question word such as *when*, *what*, *why*, *how*, *where*, *who*, *whom*, *which*, and *whose*, which often indicate questions ("what is the phone number for 344 S3?", "how many buildings do you still have to clear?") [31]. We also check if a sentence ends with a question mark [20], [31].
6) **Acknowledgment Words**. We check if a sentence contains acknowledgment words such as *ok*, *okay*, *ack*, *roger*, *rgr*, *yes*, *yup*, *thanks*, *thx*, *copy*, and so on [19]. Examples of such sentences are: "B Co, Roger", "rgr, ack", and "ack, let us know when it actually moves".
7) **Request Words**. We check if a sentence contains request words such as *request*, *requesting*, *requested*, and *requests*. Examples of such sentences are: "requesting CAS specific time now" and "1 kia request EVAC".
8) **Specific Bigrams**. We check if a sentence contains the following bigrams: (1) modal verb + second person personal pronoun, (2) first person personal pronoun + modal verb, (3) first person personal pronoun + "need", (4) "planning to", and (5) "going to". The first bigram indicates a directive ("can you") [20]. The second bigram indicates a sentence as a commissive ("I will"). The third bigram indicates a sentence as a directive ("wpns, I need a ftl") [20]. The fourth and the fifth bigrams indicate a sentence as commissive ("I am going to fire MIRs") [31].
9) **Sentence Length**. We check if a sentence is a longer or a shorter sentence. Longer sentences indicate informatives whereas shorter sentences indicate directives, questions, and acknowledgments [25]. We define a long sentence as one that contains more than nine words—the mean length of sentences in the training data—and a short sentence otherwise.
10) **Word Properties** We check for various word properties in sentences. First, we check if sentences contain numeric words [25]. These sentences may indicate informatives ("B Co ELT 12SWG 63166 89812" and "5.56 X 2700 rds 7.62 X 350 rds 60MM HE X 3 rds").
Second, we check if the average word length in a sentence is greater than four—the mean length of words in the training data. This feature can be useful since team leaders tend to use longer words whereas subordinates tend to use shorter words [36].
11) **Filtered N-Gram Features**. To improve the classifier's prediction accuracy, we obtain a set of filtered n-gram features by taking the following steps. One, we trained the SVM classifier with all n-gram features. Two, after training, for each label or class, we obtain specific n-gram features and their SVM weights (can be negative or positive). Three, for each class, we obtain the lowest negative and highest positive weights of their features. We store negative and positive weights for all classes in variable A and B respectively. Fourth, we compute their means, i.e., X = mean(A) and Y = mean(B). Fifth, we select or filter features for each class either below X or above Y. Sixth, we combine all selected or filtered features and remove duplicates to obtain the final list of filtered n-gram features.

IV. Evaluation

We perform two evaluations. First, we evaluate how accurate and robust our text mining approach is to predict meanings of messages from broadcast communications. Second, we evaluate how the team process measures (affective, behavioral, and cognitive) computed using meanings of messages from broadcast communications change with respect to time. Recall that Kozlowski and Ilgen [24] emphasize that team performance evolves over time. For evaluations we receive the ground truth for team performance from team members, however, we find that the ground truth do not have a large variance (σ^2), i.e., ground truth provided by the members were same as the mean. Thus, we consider team process measures as surrogates to understand change. We propose the following hypotheses to research questions RQ_1 and RQ_2. For each Hypothesis H_i, we assume a corresponding null hypothesis, written $\overline{H_i}$, indicating that the respective measures are equal.

- $\mathbf{H}_1$. *Identifying meanings of messages considering all features yields higher accuracy than considering n-grams.*
- $\mathbf{H}_{2a}$. *Affective processes computed from broadcast messages are more stable than baselines.*
- $\mathbf{H}_{2b}$. *Affective processes computed from broadcast messages are more variable than baselines.*
- $\mathbf{H}_{3a}$. *Behavioral processes computed from broadcast messages are more stable than baselines.*
- $\mathbf{H}_{3b}$. *Behavioral processes computed from broadcast messages are more variable than baselines.*
- $\mathbf{H}_{4a}$. *Cognitive processes computed from broadcast messages are more stable baselines.*
- $\mathbf{H}_{4b}$. *Cognitive processes computed from broadcast messages are more variable than baselines.*

A. Data Description

We obtained a military dataset from a division level command exercise. The dataset was prepared from a simulation experiment (SIMEX). The dataset contains 20 chat rooms on average with 15 team members each and 6,998 messages. We select four chat rooms, C_1, C_2, C_3, and C_4, based on the facts that they contain more intrateam messages than interteam messages, which suggests that members are strongly connected with each other. Below are some details of the chat rooms with their numbers of messages and participants.

TABLE II
Chat room details for C_1, C_2, C_3, and C_4.

Chat Room	#Messages	#Team Members
C_1	506	10
C_2	407	12
C_3	153	27
C_4	155	18

B. Evaluating H_1

We label messages in the four chat rooms as questions (Q), directives (D), commissives (C), requests (R), informatives (I), and acknowledgments (A). If a message does not fall under Q, D, C, R, I, or A label, we label it as other (O). We label messages via two raters. We obtained a high raters' inter-agreement (kappa score [7]) as 0.93. Thus, we arbitrarily choose one of the rater's assigned labels as the ground truth since we cannot take the average. Based on the labels, Figure 5 shows the distribution of labels in each chat room.

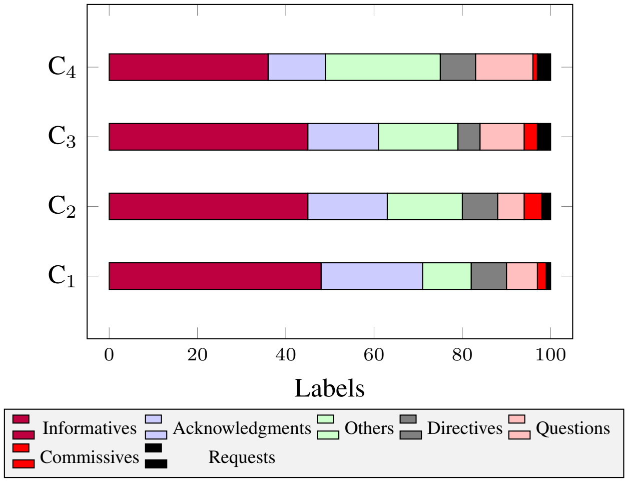


Fig. 5. Percentage distribution of labels across four chat rooms.

Next, we prepare two training and test datasets. We prepare the first training dataset by combining messages and labels from the C_1, C_3, and C_4 chat rooms. For testing, we consider messages in C_2. We prepare the second training dataset by combining messages and labels from C_1, C_2, and C_3. For testing, we consider messages in C_4. From the training datasets, we extract *features* that help identify individual labels: questions (Q), directives (D), commissives (C), requests (R), informatives (I), acknowledgments (A), and others (O). For each training data, we apply 10-fold cross validation to train the Support Vector Machine (SVM) classifiers. We use the trained SVM classifier to predict labels for the test data.

The Support Vector Machine (SVM) [10] is a discriminative classifier that categorizes the data into classes by generating an optimal separating hyperplane. Consider a Cartesian plane, where data points belonging to two classes are present. SVM based on the these data points generates a separating hyperplane that provides the largest minimum distance to the training examples. For training, SVM involves the minimization of the following error function:

$$\frac{1}{2}w^T w + C\sum_{i}^{N}\xi_i \tag{8}$$

The function is subjected to the following constraints: $y_i(w^T\phi(x_i)$+$b)$≥1-ξ_i and ξ_i≥0, i=1, . . ., N. In the function, C is the regularization parameter, w is the vector of coefficient, ξ_i represents the parameter for handling non-separable data, i represents the labels for N training cases, y_i represents the classes, and ϕ is the kernel that transforms the data from input to different feature space, e.g., linear, polynomial, radial basis function (RBF), and sigmoid. For tuning the function, we consider the kernel ϕ and the regularization parameter C. The higher value of C leads to a narrow margin and thus, makes it difficult to ignore constraints. The lower value of C leads to a larger margin, and thus, makes it easy to ignore constraints. For our work we considered linear kernel for training the data. In linear kernel, $\phi(x_i)$ takes the form of $x^T x'$. For training the data we consider one-versus-rest classifier for SVM where a class is assigned based on the highest output.

C. Evaluating H_2, H_3, and H_4

We compute our team process measures: affective (trust, confidence, and cohesion), behavioral, and cognitive (clarity) using message labels. In addition, we compute our baselines: in-degree and out-degree.

To evaluate how team process measures evolve over time, we create four cumulative time periods, as shown in Figure 6.

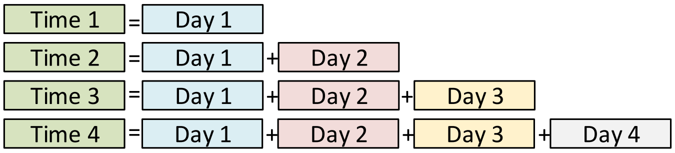

Time 1 = Day 1
Time 2 = Day 1 + Day 2
Time 3 = Day 1 + Day 2 + Day 3
Time 4 = Day 1 + Day 2 + Day 3 + Day 4

Fig. 6. For each chat room, four time periods are constructed from messages broadcast by team members on four successive days.

Based on the time periods, we present team members and messages broadcast by them in Table III.

TABLE III
SESSION DETAILS FOR C_1, C_2, C_3, AND C_4 INCLUDING TEAM MEMBERS AND MESSAGES BROADCAST PER SESSION.

Sessions	**C_1**		**C_2**	
	Members	*Messages*	*Members*	*Messages*
Time 1	7	46	6	62
Time 2	8	240	11	155
Time 3	10	363	12	322
Time 4	10	506	12	407
Sessions	**C_3**		**C_4**	
	Members	*Messages*	*Members*	*Messages*
Time 1	12	27	3	9
Time 2	22	85	8	31
Time 3	23	127	11	85
Time 4	27	153	18	155

To evaluate hypotheses H_2, H_3, and H_4, we compute the team process measures for each team member in a chat room for each time period. We normalize them to [0,1] and obtain a mean for each team performance measure for a time period in a chat room [1]. We eliminate any team member who does

[1] The weights for each team member are by default set to '1'. We considered ranks of team members to compute the mean but the approach did not add clarity to the overall results.

not broadcast a message that indicates a meaning. Algorithm 1 describes our approach. Since we have four time periods in a chat room, we obtain four values for each team process measure computed as the mean in Algorithm 1.

```
Algorithm 1: compute a measure(inputs) return mvalue;
for $i \leftarrow 1$ to CHAT ROOM do
    for $j \leftarrow 1$ to TIME PERIOD do
        for $k \leftarrow 1$ to MEMBER do
            val (k) $\leftarrow$ compute (messages (i, j, k));
        mvalue (i, j) = mean (normalize (val (1:k)));
```

Based on team process measures obtained for four time periods, we perform a regression analysis. In general, the regression analysis is done to analyze relationships between a dependent variable and one or more independent variables. The regression analysis help us to infer how the dependent variable changes when anyone of the independent variables is varied while other independent variables are fixed. The regression analysis is mostly used for prediction and forecasting. Thus, for our problem, we us quadratic regression analysis since we want a stable and a variable component for the evaluation of our hypotheses.

A quadratic equation can be represented as follows:

$$f(x) = \alpha + \beta.x + \gamma.x^2 \tag{9}$$

where x is the input to the function, α is the constant or the free term also referred as the *intercept*, β is the *linear* coefficient (slope), and γ is the *quadratic* coefficient. A high α indicates greater stability whereas a high β indicates greater change or variability. To compute α, β, and γ for each measure, we provide four values for each such measure across four time periods. We use MATLAB's polynomial regression (quadratic regression) to compute α, β, and γ.

We use one-tailed t-tests to evaluate the hypotheses. For one-tailed t tests, we create samples corresponding to each team process measure. Each sample contains either α or β values computed from each chat room. For comparison, we considered 95% confidence interval.

To verify robustness of our method, we create additional chat rooms by combining messages in primary chat rooms C_1, C_2, C_3, and C_4. For each such additional chat room, we compute team process measures across four time periods using Algorithm 1 and perform quadratic regressions to obtain α, β, and γ, for each chat room, respectively.

V. Results

We now provide results based on our evaluations.

A. Evaluating H_1

We show the F-Measures for predicted labels using C_2 and C_4 as test data in Tables IV and V, respectively. The F-Measure is the harmonic mean of precision and recall. Precision is given by $\frac{\text{true\_positives}}{\text{true\_positives+false\_positives}}$. Recall is given by $\frac{\text{true\_positives}}{\text{true\_positives+false\_negatives}}$.

TABLE IV
F-Measures for the C_2 chat room.

Features	Ques	Dir	Req	Comm	Info	Ack
1. N-Grams (N)	0.53	0.29	0	0.1	0.75	0.69
2. 1+Mod Verbs (MV)	0.53	0.3	0	0.1	0.75	0.69
3. 2+Act Verbs (AV)	0.56	0.34	0	0.12	0.74	0.71
4. 3+Pers Prons (PP)	0.57	0.37	0	0.11	0.74	0.71
5. 4+Ques Words (QW)	0.72	0.34	0	0.11	0.75	0.71
6. 5+Ack Words (AW)	0.7	0.31	0	0.11	0.73	0.85
7. 6+Req Words (RW)	0.7	0.34	0.4	0.11	0.74	0.85
8. 7+Bigrams (SB)	0.7	0.34	0.4	0.11	0.74	0.85
9. 8+Sent Len (SL)	0.72	0.34	0.4	0.11	0.73	0.86
10. 9+Word Props (WP)	0.76	0.34	0.4	0.12	0.76	0.84
11. 10+Filt N-grams (FN)	0.75	0.46	0.57	0.2	0.75	0.84

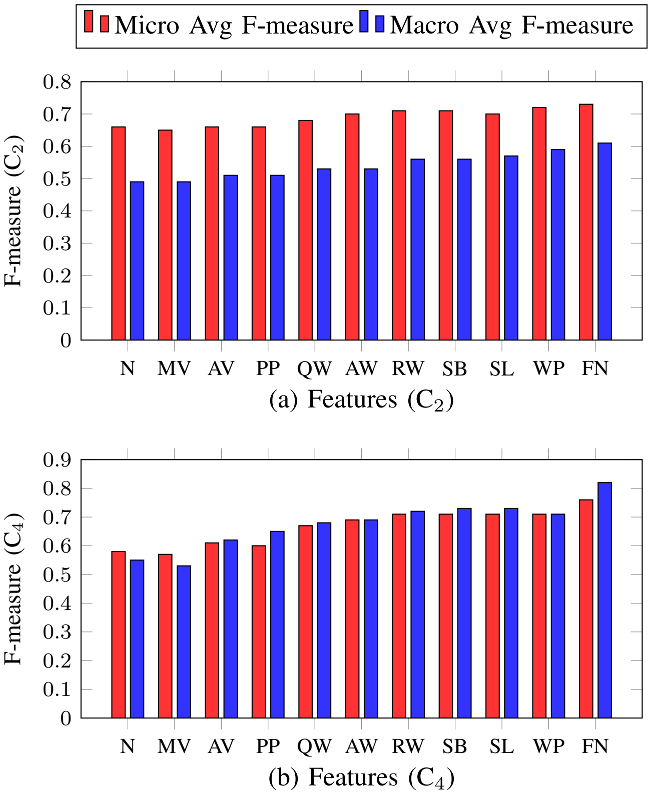
TABLE V
F-Measures for the C_4 chat room.

Features	Ques	Dir	Req	Comm	Info	Ack
1. N-Grams (N)	0.38	0.25	0	0.5	0.62	0.74
2. 1+Mod Verbs (MV)	0.38	0.24	0	0.33	0.62	0.75
3. 2+Act Verbs (AV)	0.4	0.71	0	0.5	0.59	0.75
4. 3+Pers Prons (PP)	0.75	0.75	0	0.4	0.59	0.75
5. 4+Ques Words (QW)	0.86	0.74	0	0.4	0.59	0.75
6. 5+Ack Words (AW)	0.86	0.74	0	0.33	0.61	0.8
7. 6+Req Words (RW)	0.86	0.74	0.57	0.33	0.62	0.8
8. 7+Bigrams (SB)	0.86	0.74	0.57	0.4	0.63	0.8
9. 8+Sent Len (SL)	0.86	0.74	0.57	0.33	0.64	0.82
10. 9+Word Props (WP)	0.83	0.74	0.5	0.5	0.63	0.83
11. 10+Filt N-grams (FN)	0.9	0.88	0.67	1	0.67	0.77

F-Measures. In Tables IV and V, the first column represents incremental features described in Section III-A. In both the tables, we find that F-Measures for each class obtained are maximum. For example, considering feature set 11, we obtain an average F-Measures of C_2 and C_4 for Questions: 0.83 (0.75 and 0.9); Directives: 0.67 (0.46 and 0.88); Requests: 0.62 (0.57 and 0.67); Commissives: 0.6 (0.2 and 1); Informatives: 0.71 (0.75 and 0.67); and Acknowledgments: 0.81 (0.84 and 0.77). Clearly, Questions and Acknowledgments are the most easily predictable followed by Informatives, Directives, Requests, and Commissives. For evaluating Hypothesis H_1, we check if considering all features improves F-Measure above the baseline (n-grams). Using the one-tailed t-test at the significance level of 5%, we find that the improvement in F-Measures is significant for both C_2 (p = 0.03) and C_4 (p = 0.001), thereby rejecting the null hypothesis, $\overline{H_1}$.

Feature Comparisons. Comparing the individual features we find that modal verbs (MV) improve the F-Measure for directives (C_2). Adding action verbs (AV) improves the F-Measure for questions (C_2, C_4), directives (C_2, C_4), and commissives (C_2, C_4). Adding personal pronouns (PP) improves the F-Measure for questions (C_2, C_4) and directives (C_2, C_4). Adding question words (QW) improves the F-Measure for questions (C_2, C_4). Adding acknowledgment words (AW) improves the F-Measure for acknowledgments (C_2, C_4). Adding Request words (RW) improves the F-Measure for requests

(C_2, C_4). Adding specific bigrams (SB) slightly improves the F-Measure for commissives (C_4). Adding sentence lengths (SL) improves the F-Measure for questions (C_2), informatives (C_4), and acknowledgments (C_2, C_4). Adding word properties (WP) slightly improves F-Measures for questions (C_2), informatives (C_2), commissives (C_2, C_4), and acknowledgments (C_4). Adding filtered n-grams (FN) improve the F-Measures for questions (C_2, C_4), directives (C_2, C_4), requests (C_2, C_4), commissives (C_2, C_4), and informatives (C_4). Since the combination of all features (10 + Filtered n-gram) improves the overall f-measure scores for all classes, we adopt it as the final set of features. We describe below the overall F-Measure further in terms of macro and micro F-Measure.



(a) Features (C_2)

(b) Features (C_4)

Fig. 7. Micro and macro average F-Measures for identifying labels for messages in the C_2 and C_4 chat rooms.

Overall Performance. To determine the overall performance of our approach in terms of F-Measures, we compute micro and macro average F-Measures, which capture different incremental features in Figure 7. Micro average F-Measure is a useful predictor when the distribution of classes varies across the dataset. Macro average F-Measure gives a picture of overall performance. Micro average F-Measure is computed as the harmonic mean of micro average precision and recall. Micro average precision is given by $\frac{\sum_i^N \mathrm{true\_positives_i}}{\sum_i^N \mathrm{true\_positives_i}+\sum_i^N \mathrm{false\_positives_i}}$. Micro average recall is given by $\frac{\sum_i^N \mathrm{true\_positives_i}}{\sum_i^N \mathrm{true\_positives_i}+\sum_i^N \mathrm{false\_negatives_i}}$. Here, N represents the number of classes. Macro average f-measure is computed as the harmonic mean of macro average precision and recall. Macro average precision is given by $\frac{\sum_i^N \mathrm{precision_i}}{N}$. Macro average recall is given by $\frac{\sum_i^N \mathrm{recall_i}}{N}$.

Considering all the features, we find that micro and macro average F-Measures for C_2 are 0.73 and 0.62, respectively, and for C_4, 0.73 and 0.82, respectively. Overall, we find that the SVM classifier trained using messages from C_1, C_2, and C_3 performs better than one trained using messages from C_1, C_3, and C_4. The difference is likely due to the first training dataset being larger. This suggest the F-Measure can be improved with additional training data.

B. Evaluating H_2, H_3, and H_4

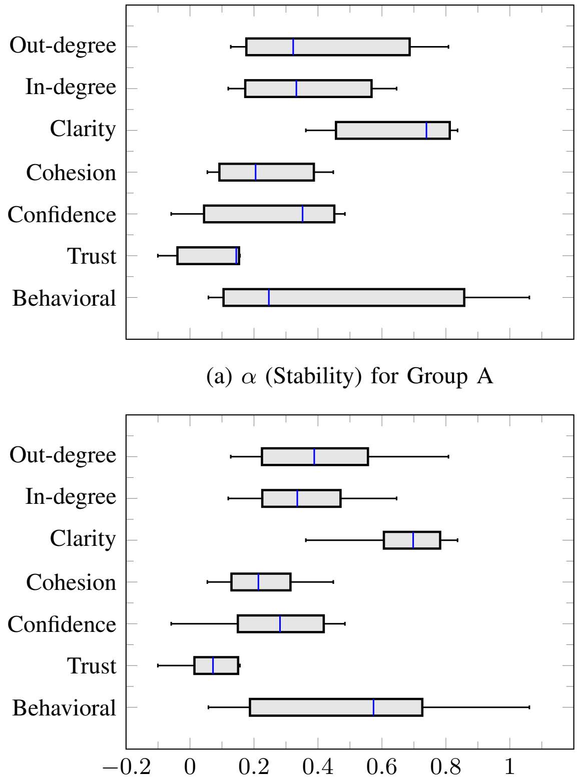
To verify H_2, H_3, and H_4 for each chat room, we compute team process measures (affective, behavior, cognitive, and baselines) across four different time periods. For the evaluation, the following constraints were necessary for team members in the different chat rooms:

- We monitor team process measures for team members who are present in the first time period (Time 1). If a team member shows up in other time periods other then Time 1, we ignore the team member.
- We ignore a team member if one or more team process measures for the team member are zero in the first time period (Time 1). This mean the team member did not broadcast a message that was meaningful in terms of the mission.

Based on the above constraints, we find that C_1 has seven, C_2 has five, C_3 has ten, and C_4 has two valid team members. Since C_4 has only two valid team members, we omit it from our evaluation. For robustness, we create datasets by combining team members and messages across chat rooms to produce an additional four datasets: (1) C_1+ C_2, (2) C_1 + C_3, (3) C_2 + C_3, and (4) C_1 + C_2 + C_3. Thus, we create two groups for evaluations: Group A of the three original chat rooms and Group B of the three original and four combined chat rooms.

Stability. To evaluate H_{2a}, H_{3a}, and H_{4a}, we obtain intercepts α (stability) for different team process measures computed from chat rooms in Groups A and B, respectively. Figure 8(a) and (b) show boxplots of intercepts for Groups A and B, respectively. These boxplots show that in both groups, the mean intercept for clarity is higher than mean intercepts for baselines. In Group B, in addition to clarity, the mean intercept for behavioral processes is higher than the mean intercept of baselines. To verify if the mean intercept for clarity and behavioral process is significantly higher than mean intercepts for baselines, we performed one-tailed t-tests at the significance level of 5%. Table VI show the results: a p-value greater than 0.05 indicates that we fail to reject the null hypotheses.

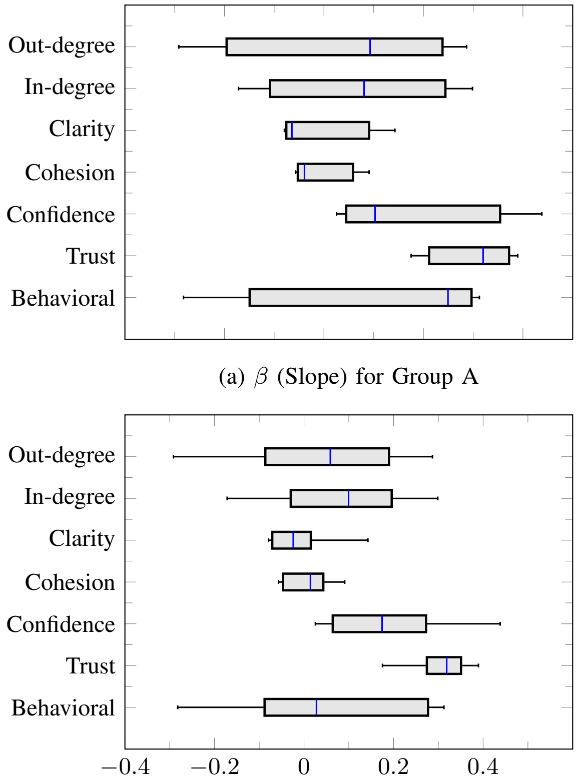
From the t-test results for clarity, we find that clarity is significantly more stable than baselines considering communication interactions in Group B than Group A. The results indicate that the improvements might be due to a larger dataset in Group B. Based on the results, we reject the null hypothesis $\overline{H_{4a}}$. This suggest that clarity remain stable with time, hereby, indicating that knowledge and information requirements to accomplish mission remain stable. Thus, we consider clarity



(a) α (Stability) for Group A

(a) α (Stability) for Group B

Fig. 8. (a) Boxplots comparing team process measures in terms of stability for chat rooms in (a) Group A and (b) Group B.



(a) β (Slope) for Group A

(a) β (Slope) for Group B

Fig. 9. Boxplots comparing different team process measures in terms of slope (rate of change) (a) for chat rooms in Group A and (b) for chat rooms in Group B.

TABLE VI
STATISTICALLY COMPARING THE MEAN INTERCEPT OF CLARITY AND BEHAVIORAL PROCESS WITH MEAN INTERCEPTS OF BASELINES.

Mean Comparison	**Group A (p-val)**	**Group B (p-val)**
Clarity > In-degree	0.17	0.00
Clarity > Out-degree	0.23	0.03
Behavioral > In-degree	0.43	0.23
Behavioral > Out-degree	0.47	0.32

as a good indicator of overall team performance, indicating lower confusion among team members in accomplishing their tasks, thereby, improving their performance.

Considering results for behavioral process, we fail to reject the null hypotheses $\overline{H_{3a}}$ for both the groups. This suggest that in terms of stability, behavioral process may not be a good indicator of team performance. We omit significance results for other measures: trust, confidence, and cohesion with respect to stability since the means of these measures are lower than the means of the baselines.

Variability. To evaluate H_{2b}, H_{3b}, and H_{4b}, we obtain β (slopes) for the team process measures from Groups A and B, as shown in Figure 9(a) and (b), respectively. In both groups, we find that the mean slope for trust and confidence is higher than the mean slopes for baselines. To verify if the mean slope for trust and confidence is significantly higher than the mean slopes of baselines, we performed using one-tailed t-tests at the significance level of 5%, as shown in Table VII. A p-value greater than 0.05 indicates that we fail to reject the null hypotheses.

TABLE VII
STATISTICALLY COMPARING MEAN SLOPE OF TRUST AND CONFIDENCE WITH MEAN SLOPES OF BASELINES.

Mean Comparison	**Group A (p-val)**	**Group B (p-val)**
Trust > In-degree	0.12	0.00
Trust > Out-degree	0.13	0.01
Confidence > In-degree	0.35	0.20
Confidence > Out-degree	0.32	0.16

The t-test results show that mean slope for trust is significantly higher than mean slopes of baselines. The significance results for trust improve by considering additional communication interactions in Group B. This suggest that trust is a good indicator of team performance since since they indicate increasing mutual reliance among team members, which is what we would expect from well-performing teams. Thus, we fail to reject null hypothesis $\overline{H_{2b}}$ for trust.

Considering results for confidence, we fail to reject the null hypotheses $\overline{H_{2b}}$ for both the groups. This suggest that in terms of variability, confidence may not be a good indicator of team performance. We omit significance results for other measures: clarity, behavioral, and cohesion with respect to variability since the means of these measures are lower than the means of baselines.

C. Evaluation Summary

We summarize the results of our hypotheses. Based on our evaluation we found the following.

- H_1. We find that identifying meanings of messages considering all features yields significantly higher accuracy than considering n-grams or existing bag-of-words approach, thereby, successfully rejecting the null hypothesis $\overline{H_1}$.
- H_{2a}. Considering the stability for affective processes, we find that means of trust, confidence, and cohesion are lower than baselines, thereby, failing to reject the null hypothesis $\overline{H_{2a}}$ for all three measures.
- H_{2b}. Considering the variability for affective processes, we find that means of the of trust are significantly higher than baselines, thereby, rejecting the null hypothesis $\overline{H_{2b}}$for trust. For confidence and cohesion, we could not reject the null hypothesis.
- H_{3a}. Considering the stability for behavioral processes, we find that mean is higher than baselines but not significantly high enough to reject the null hypothesis $\overline{H_{3a}}$.
- H_{3b}. Considering the variability for behavioral processes, we find that the mean is lower than baselines, thereby, failing to reject the null hypothesis $\overline{H_{3b}}$.
- H_{4a}. Considering the stability for cognitive processes, we find that the mean is significantly higher than baselines, thereby, successfully rejecting the null hypothesis $\overline{H_{4a}}$.
- H_{4b}. Considering the variability for cognitive processes, we find that the mean is lower than baselines, thereby, failing to reject the null hypothesis $\overline{H_{4b}}$.

VI. Discussion

We make the following contributions. First, we propose meanings of messages to compute team process measures. We propose six meanings: questions, directives, requests, commissives, informatives, and acknowledgments. We obtain several of these meanings from the literature on classifying communication types.

Second, we provide a computational approach to compute team process measures from broadcast communications using meanings of messages. We address a variety of team process measures, including affective (trust, confidence, and cohesion), behavioral, and cognitive (clarity).

Third, we provide a text-mining approach to extract meanings of messages from broadcast communications. We propose eleven interesting features to use as indicators. Our approach yields micro and macro average F-Measures for our approach of 0.78 (≈80%) and 0.68 (≈70%), respectively. Comparing the accuracy of predicting different classes we find that the accuracy for predicting questions (0.83) is highest followed by the accuracy for predicting acknowledgments (0.81), informatives (0.71), directives (0.67), requests (0.62), and commissives (0.6).

Fourth, we compute stability and variability for team process measures. In terms of stability, we find that cognitive processes (clarity) are most stable than other measures. In terms of variability, the affective processes (trust) have higher mean slopes than other measures.

Since our approach is based on the general framework of team processes [24] and relies on extracting the meanings of messages, our approach can be generalized to other domains such as education, healthcare, IT, and so on. Some of the approaches to extract meanings of messages such as Kalia et al. [19] do apply the classification scheme to datasets from different domains such as military and corporations (Enron).

VII. Limitations and Future Work

Our contribution has the following limitations. First, given the complex nature of military field exercises, objective measures of team performance exist at the unit level and are globally relevant to the achievement of mission objectives, training outcomes, and reflect performance of the adversary (enemy forces). However, post-event survey measures can be used to establish qualitative measures of team performance that can be leveraged as ground-truth in future work.

Second, identifying meanings of messages manually is tedious and does not scale to large datasets. To address this concern, we are exploring unsupervised approaches for identifying meanings of messages.

Third, we limit our evaluations to a military dataset. It would be extremely interesting to expand our contributions to datasets collected from enterprise settings such as the IBM Small Blue dataset [17] that uses money as a process measure of matrixed teams.

Fourth, among the team process measures we studied, cohesion did not indicate a significant contribution to team performance. Considering that cohesion has been extensively cited in the literature as an important factor, e.g., [6], [5], we will investigate it further.

VIII. Related Work

We now describe related work on computing team process measures from social network analysis, semantic classification of communications, and ratings based approaches.

A. Social Network Analysis

Sparrowe et al. [34] conduct a field study with to find that individual job performance is positively related to centrality in advice networks and negatively related to centrality in hindrance networks. In addition, they find that density of a hindrance network is significantly related to the group performance. Sparrowe et al.'s work has two limitations. One, they collect data based on responses to questionnaires that can be biased since employees may not reveal truthful information. Two, they do not consider these responses to compute team performance measures such as team processes and emergent states. We address Sparrowe et al.'s limitations by considering communications between team members and computing team process measures.

Zhang et al. [41] extract structure indicators such as degree centrality and content indicators such as sentiments from emails. Zhang et al. find that as teams move into a more mature stage, group density decreases since more members get involved and members' emotional attachment increases that pushes them to work more creatively. Compared to Zhang et al.'s contribution our work takes a step further by considering the content of communication to determine team process measures.

Ehrlich and Cataldo [12] find that when team leaders share more information (out-degree centrality) than they gather or receive (in-degree centrality), there is a significant improvement in productivity and quality of software created. to determine team process measures. Compared to Ehrlich and Cataldo's [12] contribution, we compute various team measures considering the content of broadcast communications and also show that such measures (except cohesion) perform better than in-degree and out-degree centrality with time.

Adalı et al. [1] identify indicators of social behavior, which combine text and network structure to predict social relationships in a dataset of tweets. Specifically, they find that certain linguistic features from text correlate well with certain networking features, e.g., that emotional words are more likely to occur in intimate conversations than in professional conversations. In contrast, we are not concerned with conversations on Twitter but in connection with a mission and with preexisting teams. We consider all n-grams, so as not to manually limit the dimensions, and several other features. We can compute nuanced team process measures from observations of communications.

Henttonen et al. [16] find that dense and fragmented instrumental network has a positive affect on team performance whereas fragmented expressive network has a negative effect on team performance. Henttonen et al. advance the contribution made by Sparrowe et al. [34] by considering networks related to task and emotional relationships. However, their contribution has the same limitations as Sparrowe et al.'s, namely, that they construct such networks based on responses to questionnaires.

B. Semantic Classification of Communications

Entin and Entin [13] capture both the semantic and quantitative aspects of communication stream to compute team process measures. However, their contribution is limited to providing basic measures instead of specifying how to compute team process measures from different classes of communications. Entin and Entin's contribution does not include automatic extraction of classes from communications.

Fischer et al. [14] extract task-related and social dimensions from a team's communications to find that the team performance is significantly related with team member's task-related communications. Their approach is limited to considering word counts using LIWC. In contrast, we provide domain-independent features to classify text and use the class labels to compute measures for team processes and emergent states, which are missing in Fischer et al.'s approach.

Cooke et al. [9] propose to analyze the content of communication both manually and using Latent Semantic Analyses (LSA) to automatically compute measures of team cognition. In contrast, we introduce domain-independent features to train the SVM classifier. One major limitation of Cooke et al.'s work is that it appears to be mainly a proposal for research: they do not describe any data, experiments, or results.

C. Survey Rating Based Approaches

Annett et al. [3] provide a procedure to first identify team skills and then relate the measure to the team performance. To identify team skills, Annett et al. apply Targeted Acceptable Responses to Generated Events or Tasks (TARGET) [15], a survey methodology, to collect team responses to key events generated during team coordination. In contrast, we avoid surveys and identify such message meaning and use such meanings to compute team process measures.

IX. Acknowledgment

This research is primarily supported by a US Army Research Laboratory (ARL) Oak Ridge Institute for Science and Education (ORISE) Fellowship and partially supported by the Army Research Laboratory in its Network Sciences Collaborative Technology Alliance (NS-CTA).

References

[1] Adalı, S., Sisenda, F., Magdon-Ismail, M.: Actions speak as loud as words: Predicting relationships from social behavior data. In: Proceedings of the 21st International Conference on World Wide Web. pp. 689–698. ACM, Lyon (2012)

[2] Amason, A.C.: Distinguishing the effects of functional and dysfunctional conflict on strategic decision making: Resolving a paradox for top management teams. The Academy of Management Journal 39(1), 123–148 (1996)

[3] Annett, J., Cunningham, D., Mathias-Jones, P.: A method for measuring team skills. Ergonomics 43(8), 1076–1094 (2010)

[4] Austin, J.L.: How to Do Things with Words. Clarendon Press, Oxford (1962)

[5] Beal, D.J., Cohen, R.R.: Cohesion and performance in groups: A meta-analytic clarification of construct relations. Journal of Applied Psychology 88(6), 989–1004 (2003)

[6] Cannon-Bowers, J., Tannenbaum, S.I., Salas, E., Volpe, C.E.: Defining competencies and establishing team training requirements. In: Guzzo, R.A., Salas, E. (eds.) Team Effectiveness and Decision Making in Organizations, pp. 333–380. Jossey-Bass, San Francisco (1995)

[7] Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics 22(2), 249–254 (Jun 1996)

[8] Carron, A.V., Widmeyer, W.N., Brawley, L.R.: The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire. Journal of Sports and Exercise Psychology 7(3), 244–266 (1985)

[9] Cooke, N.J., Salas, E., Kiekel, P.A., Bell, B.: Advances in measuring team cognition. In: Bell, B., Salas, E., Fiore, S.M. (eds.) Team Cognition: Understanding the Factors that Drive Process and Performance, pp. 83–106. American Psychological Association, Washington, DC (2004)

[10] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)

[11] Dreu, C.K.W.D., Weingart, L.R.: Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. Journal of Applied Psychology 88(4), 741–749 (2003)

[12] Ehrlich, K., Cataldo, M.: The communication patterns of technical leaders: Impact on product development team performance. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. pp. 733–744. ACM, Baltimore (2014)

[13] Entin, E.E., Entin, E.B.: Measures for evaluation of team processes and performance in experiments and exercises (2001), http://www.aptima.net/publications/2001_EntinEE_EntinEB.pdf

[14] Fischer, U., McDonnell, L., Orasanu, J.: Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. Aviation, Space, and Environmental Medicine 78(5), B86–95 (2007)

[15] Fowlkes, J., Dwyer, D.J., Oser, R.L., Salas, E.: Event-based approach to training (EBAT). The International Journal of Aviation Psychology 8(3), 209–221 (1998)

[16] Henttonen, K., Janhonen, M., Johanson, J.E.: Internal social networks in work teams: Structure, knowledge sharing, and performance. International Journal of Manpower 34(6), 616–634 (2013)

[17] IBM: SmallBlue (2013), http://smallblue.research.ibm.com/solution-smallblue.html

[18] Ilgen, D.R., Hollenbeck, J.R., Johnson, M., Jundt, D.: Teams in organizations: From Input-Process-Output Models to IMOI models. Annual Review of Psychology 56, 517–543 (2005)

[19] Kalia, A.K., Buchler, N., Ungvarsky, D., Govindan, R., Singh, M.P.: Determining team hierarchy from broadcast communications. In: Proceedings of the 6th International Conference on Social Informatics. LNCS, vol. 8851, pp. 493–507. Springer, Barcelona (2014)

[20] Kalia, A.K., Nezhad, H.R.M., Bartolini, C., Singh, M.P.: Monitoring commitments in people-driven service engagements. In: Proceedings of the 10th IEEE International Conference on Services Computing. pp. 160–167. IEEE, Santa Clara (2013)

[21] Kalia, A.K., Zhang, Z., Singh, M.P.: Güven: Estimating trust from communication. Journal of Trust Management 3(1), 1–19 (2016)

[22] Klimoski, R., Mohammed, S.: Team mental model: Construct or metaphor? Journal of Management 20(2), 403–437 (1994)

[23] Kozlowski, S.W.J., Bell, B.S.: Work Groups and Teams in Organizations, vol. 12, pp. 333–375. John Wiley & Sons, Inc. (2003)

[24] Kozlowski, S.W., Ilgen, D.R.: Enhancing the effectiveness of work groups and teams. Psychological Science in the Public Interest 7(3), 77–124 (2006)

[25] Lampert, A., Dale, R., Paris, C.: Detecting emails containing requests for action. In: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 984–992. Los Angeles, California (2010)

[26] Lindsley, D.H., Brass, D.J., Thomas, J.B.: Efficacy-performing spirals: A multilevel perspective. The Academy of Management Review 20(3), 645–678 (1995)

[27] Marks, M.A., Mathieu, J.E., Zaccaro, S.J.: A temporally based framework and taxonomy of team processes. The Academy of Management Review 26(3), 356–376 (2001)

[28] Mathieu, J.E., Heffner, T.S., Goodwin, G.F., Cannon-Bowers, J.A., Salas, E.: Scaling the quality of teammates' mental models: Equifinality and normative comparisons. Journal of Organizational Behavior 26(1), 37–56 (2005)

[29] Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. The Academy of Management Review 20(3), 709–734 (1995)

[30] McAllister, D.J.: Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. The Academy of Management Journal 38(1), 24–59 (1995)

[31] Qadir, A., Riloff, E.: Classifying sentences as speech acts in message board posts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 748–758. Association for Computational Linguistics, Edinburgh (2011)

[32] Salas, E., Prince, C., Baker, D.P., Shrestha, L.: Situation awareness in team performance: Implications for measurement and training. Human Factors 37(1), 123–136 (1995)

[33] Smith-Jentsch, K.A., Johnston, J.H., Payne, S.C.: Making decisions under stress. In: Implications for Individual and Team Training, pp. 61–87. American Psychological Association (1998)

[34] Sparrowe, R.T., Liden, R.C., Wayne, S.J., Kraimer, M.L.: Social networks and the performance of individuals and groups. The Academy of Management Journal 44(2), 316–325 (2001)

[35] Taylor, R.M.: Situational awareness rating technique (SART): The development of a toll for aircrew systems design. In: Proceedings of the AGARD Aerospace Medical Panel Symposium on Situational Awareness in Aerospace Operations. Neuilly Sur Seine (1990)

[36] Tchokni, S.E., Séaghdha, D.Ó., Quercia, D.: Emoticons and phrases: Status symbols in social media. In: Proceedings of the Eighth International Conference on Weblogs and Social Media. pp. 485–494. AAAI, Ann Arbor (2014)

[37] Wang, Y., Hang, C.W., Singh, M.P.: A probabilistic approach for maintaining trust based on evidence. Journal of Artificial Intelligence Research 40, 221–267 (Jan 2011)

[38] Wang, Y., Singh, M.P.: Formal trust model for multiagent systems. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI). pp. 1551–1556. IJCAI, Hyderabad (Jan 2007)

[39] Wang, Y., Singh, M.P.: Evidence-based trust: A mathematical model geared for multiagent systems. ACM Transactions on Autonomous and Adaptive Systems (TAAS) 5(4), 14:1–14:28 (Nov 2010)

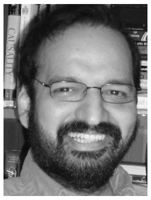
[40] Zaccaro, S.J., Blair, V., Peterson, C., Zazanis, M.: Collective efficacy. In: Self-Efficacy, Adaptation, and Adjustment, pp. 305–328. The Plenum Series in Social/Clinical Psychology, Springer (1995)

[41] Zhang, X., Gloor, P.A., Grippa, F.: Measuring creative performance of teams through dynamic semantic social network analysis. International Journal of Organisational Design and Engineering 3(2), 165–184 (2013)

Anup K. kalia is a Research Staff Member at IBM Thomas J. Watson Research Center, NY in cloud services automation and analytics team. His research interests are service computing, multiagent systems, cognitive science, and software engineering. Anup received his M.S. and PhD in computer science from North Carolina State University.

Norbou Buchler is a cognitive scientist in the Human Research & Engineering Directorate of the Army Research Laboratory. His basic research interests lie in understanding human cognition and collaborative decision-making at network levels of interaction using multidisciplinary approaches including social network analysis, cognitive modeling, multiagent simulation, behavioral laboratory experimentation, and field studies. His applied research focuses on human system integration and developing agent-based decision-support technologies for application in both cybersecurity and Mission Command environments. Norbou received his Ph.D. in experimental psychology from Syracuse University.

Arwen Decostanza is a Research Psychologist at the U.S. Army Research Laboratory. Dr. DeCostanza leads a program of research to enhance performance in complex, networked teams with a focus on unobtrusive, systems-based measurement of emergent states and processes. As an applied psychologist, her research focuses on the application of laboratory-based research findings in applied settings, field research, and the transition of measurement and feedback technologies to the training community and operational Army for enhanced training effectiveness and unit performance. Dr. DeCostanza received her Ph.D. in Industrial and Organizational Psychology from The George Washington University in 2008.



Munindar P. Singh is a Professor in the Department of Computer Science at North Carolina State University, Raleigh. His research interests include multiagent systems and social computing with a special interest in the challenges and techniques relating to norms, security, trust, privacy, governance, and interaction modeling in open environments. Munindar's books include the coauthored *Service-Oriented Computing* (Wiley, 2005). Singh is the Editor-in-Chief of the *ACM Transactions on Internet Technology* and was the Editor-in-Chief of *IEEE Internet Computing* from 1999 to 2002. He is a member of the editorial boards of *IEEE Internet Computing*, *Autonomous Agents and Multiagent Systems*, the *IEEE Transactions on Services Computing*, and the *ACM Transactions on Intelligent Systems and Technology*.