Serious Games for Collaborative Dark Network Discovery

Matthew P. Daggett, Daniel J. Hannon, Michael B. Hurley, and John O. Nwagbaraocha

Illicit social networks, such as trafficking or terrorist organizations, are difficult to discover because their clandestine nature limits their observability to data collection. Technological advances in remote sensing and analytical software can reduce the time- and human-intensive nature of network data curation and analysis, if effective human-system integration is achieved. To better understand this integration, researchers at Lincoln Laboratory created a succession of serious games to investigate methodologies for developing user-centered tools and quantitative human-system instrumentation, with the goal of improving network discovery. These games were employed in a multiyear study of team analytical performance and collaborative decision making, encompassing more than 80 teams and upwards of 400 unique players.

» **For decades, governments, militaries,** researchers, and other organizations have focused significant resources toward the collection and analysis of information about illicit human social networks, such as gangs, cartels, traffickers, and terrorists. These networks, often referred to as dark networks, are difficult to study because their clandestine nature limits their observability to various data collection means and often precludes a full accounting of the network membership, structure, function, and dynamics [1–3]. Historically, the social sciences have provided the foundation for the study of dark networks, largely through the time- and human-intensive manual collection and curation of qualitative network data. However, this approach is not efficient, does not scale to large organizational studies, and generally only represents static points in time [4–6].

Over the past two decades as asymmetric conflicts and complex humanitarian crises have become more prevalent across the world, increased emphasis has been aimed at characterizing dark networks that operate in urban settings to perpetrate acts of violence, such as vehicular-borne explosive attacks, i.e., car bombings. The use of vehicles to facilitate explosive-laden attacks goes back to the 1920s and has been responsible for asymmetric attacks ranging from the Provisional Irish Republican Army's bombings during the Troubles in Northern Ireland in the 1960s to widespread explosive events by terrorist organizations during the conflicts in Iraq and Afghanistan in the last 15 years [7]. When a car bombing occurs, it can be extremely challenging for law enforcement to piece together information to determine

which vehicles, facilities, and people were involved in the attack (Figure 1). This challenge is compounded by urban settings that allow perpetrators to flee and meld back into the background populous. In the last decade, advances in airborne remote sensing and terrestrial surveillance have made it possible for military and police agencies to observe not only the execution of these types of attacks on urban areas but often the events and coordination directly before and after.

However, the ability to triage surveillance video and imagery along with other reporting data—such as news, law enforcement reports, or social media— immediately after an attack is laborious and often requires teams of individuals to sift through large amounts of data to discover pieces of relevant evidentiary information [8]. Additionally, the discovered information must then be deconflicted, analyzed, validated, and synthesized to make timely risk-informed decisions about potential follow-on courses of action. It is unclear how and in what ways these teams should organize and operate, and what roles analytic and decision support technology should play in making these operations more efficient and effective.

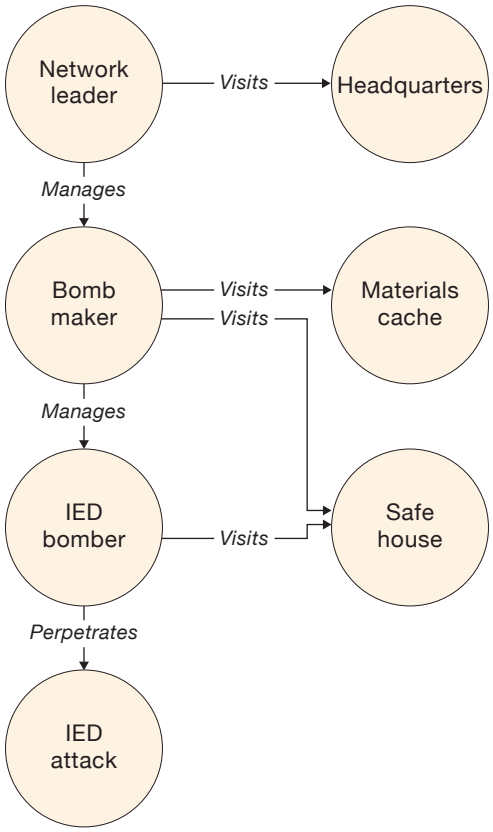


FIGURE 1. In this example of a small vehicular-borne explosives graph network, the circles (nodes) represent people, locations, and events, and the lines (edges) that connect them correspond to the nature of the association between the nodes. Arrows on the lines represent the directionality of the relationship.

Game Design

In 2009, we and other researchers at MIT Lincoln Laboratory developed a serious game to address some of the challenges regarding clandestine network discovery. We created a platform to better understand how a team of players uses multimodal geospatial data to discover information about a dark network and synthesizes those data to make decisions [9–11].

Serious Games for Research and Development

Since the 1950s, Lincoln Laboratory has performed applied research and development for national security missions on a foundation of rigorous systems analysis, full system prototyping, and development of long-term advanced technologies. As the discrete systems of earlier decades have been replaced with complex interconnected systems of systems, traditional modeling and simulation and systems analysis can be insufficient because these methods often fail to properly account for human dynamics. To overcome these limitations, researchers at the Laboratory developed a suite of methodologies and technologies to design serious games that can be used as tools to model, experiment with, and assess complex human-system dynamics that approximate those of realistic sociotechnical enterprises. In serious games, gameplay is used to achieve an explicit purpose other than amusement. We have used such games across a spectrum of the research and development process, including experiential learning, concept exploration, requirements analysis, tool development and evaluation, human performance assessment, and decision analysis.

Research Objectives

We identified four research objectives for this serious gaming work:

- Games as analysis tools. We wanted to demonstrate the value of using serious gameplay as a systems analysis tool for human-intensive workflows and applications.
- System requirements derived from decisions. In remote sensing research and development, the process often

starts with an understanding of the phenomenology of the sensing environment and observables of interest. This phase leads to the development of sensor hardware that is then integrated and fielded on the premise that the sensor capabilities are inherently useful; however, many sensor systems have not been jointly developed alongside the decision processes their data are meant to inform. In this work, we wanted to essentially invert this development and acquisition process by starting with an understanding of what information is needed to make decisions and work backward to build an end-to-end workflow that results in actionable information. Then, we could use the gaming process and simulation capabilities to determine what the technical and performance requirements should be for both the sensors and their data analysis systems.

- Effective game development scope. We wanted to learn how to build the right level of realism and fidelity into the game to create enough immersion and engagement to force players into an effective decision process, while limiting the scope and cost of development.
- Rapid tool and workflow utility assessment. Through the use of robust human-system instrumentation to collect quantitative human performance data, we wanted to develop an end-to-end process to assess the value and utility of tools early in their development cycle.

Scenario Development

During the design phase of the game, we spent a lot of effort on generating the requirements for the storyboard (hereafter referred to as the scenario) that drives the game data generation and game mechanics toward achieving the research goals. The most important requirement of the scenario design involved four elements of the geographic location of the game:

1. The game should be based in an area of future strategic importance to the U.S. government. At the time of design, many activities within the Middle East were within the purview of the U.S. Central Command, and we decided to focus instead on Africa because the U.S. Africa Command had just been established in October 2008.
2. The location should be in an area within Africa that is unfamiliar to most players, including potential players with a good understanding of global geopolitics or with prior military experience. This condition minimizes the effect of experiential knowledge and preconceptions about the scenario.
3. The area should have a history of instability and violence to build the scenario around, as well as a complex environment of actors composed of the indigenous population, foreign fighters from neighboring areas, an external coalition military presence, and multiple nongovernmental organizations (NGOs).
4. The scenario should be focused on a city with a compact, dense urban core that quickly fans out into a suburban and then rural expanse. This constraint limits the scale of the geographic area of regard for the game participants and aligns with the field of views of the sensor concepts to be used in the data simulation.

On the basis of these criteria, we chose a moderately large city in a landlocked country in Central Africa (hereafter referred to as the city). When this scenario was developed in 2009, the city had a recent history of instability. It had been briefly seized by insurgents in 2006, and in 2007 local rebels had declared war on foreigners and refugees from the surrounding region, requiring the deployment of thousands of international peacekeeping troops in 2008.

In the city, several prominent groups formed what we called the red, blue, gray, and white actors; this color naming convention is derived from military wargaming nomenclature. The red actors are those operating to incite violence in the city, such as the local rebel group, who is seen as anti-government and anti-foreigner and who has staged many recent attacks through car bombings and kidnappings. The blue actors work to counter red groups and include an international coalition of peacekeeping forces headquartered in the city and the game participants themselves. The gray actors are those who have an unclear affiliation with a side, such as the national army, who is undisciplined and believed to be heavily infiltrated by rebel groups. Lastly, the white actors consist of various NGOs and news media in the region.

From research into the city's historical events and groups, we constructed a timeline that laid out a sequence of activities that would take place in the scenario. Next, data from a geographic information system were analyzed to determine both public locations, such as the city's airport or the local army garrison, and private locations, such as previous weapons caches used

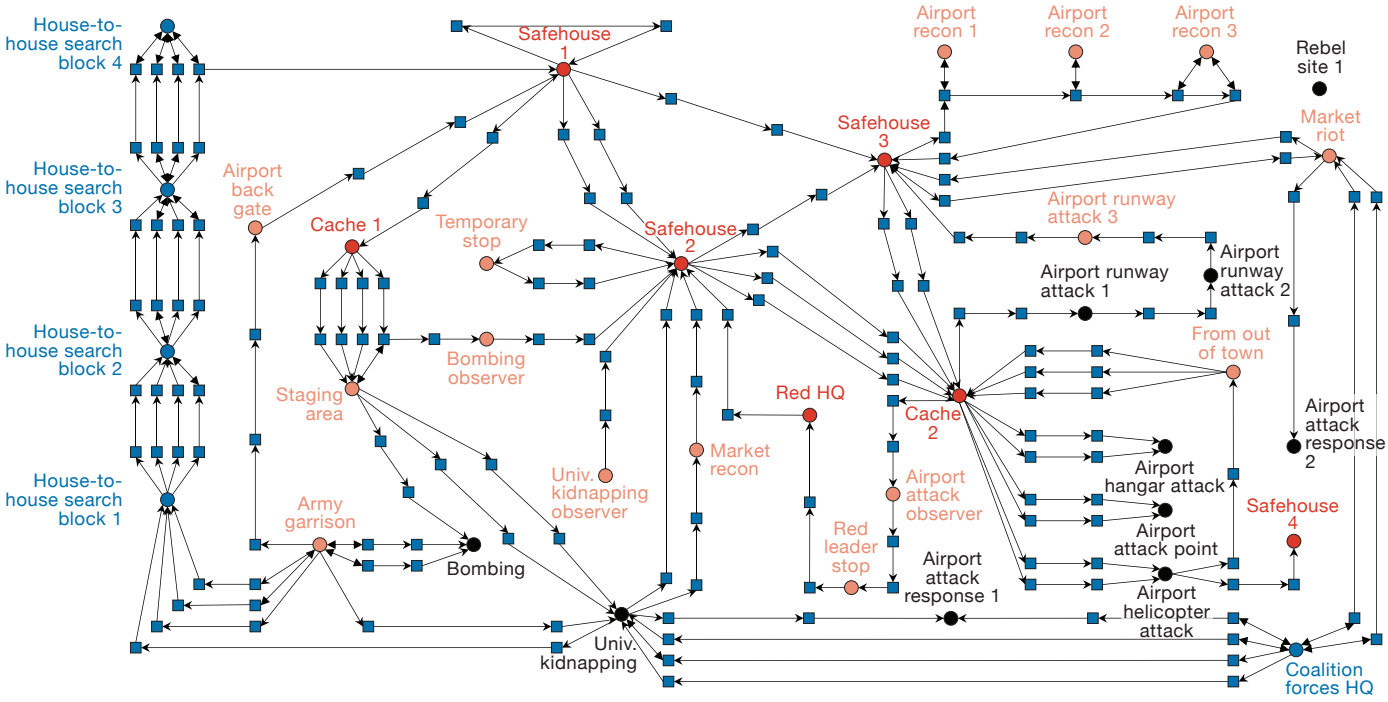


FIGURE 2. The graph of the scenario shows the network of facilities and the vehicle journeys, or tracks, that visit them. The circles (nodes) represent locations visited by a vehicle, such as a safe house or weapons cache, and the squares represent an intermediate stop of a vehicle. The lines (edges) that connect the circles and squares correspond to discrete vehicle tracks between two locations, and arrows represent the directionality of the tracks moving between the sites.

by the local rebel group and places that could be sites of interest within the scenario. Care was taken to make sure the locations chosen for these red actor sites had no known associations with public locations in any of the information sources examined.

With the locations of interest chosen, we scripted a series of activities that formed the scaffolding of the scenario, which broke down into three waves of activities. The first wave started with a truck bombing followed by a kidnapping at the local university. Next, the kidnapping prompted a neighborhood search by the national army and the discovery of a red safe house, necessitating movements of multiple red actors to other locations. In the final wave, certain groups staged a riot at the main city market to divert attention away from a coordinated attack on the airport that included the bombing of the runway and a nearby hangar. Next, we designed an intricate series of timed vehicle journeys, or tracks, between all of these event locations and other locations, such as staging areas or headquarter compounds; these tracks formed the basis of the networks of vehicles and facilities associated with the red actors. To add complexity to the scenario we gave many of these vehicle journeys intermediary stops and starts or circuitous routes between clandestine facilities, as these diversions are typical operational security principles. The final scenario consisted of nine hours of activity and comprised 37 networked sites, 27 of which were associated with the red network. Seven of the sites were high-value red facilities, eight sites were associated with clandestine red activities, seven were associated with overt red activities, and five were innocuous red vehicle stops. A graphical depiction of the network of these events, sites, and intermediary stops, which became the basis of the scenario truth, is shown in Figure 2. For simplicity, the figure does not show the times associated with each of the movement starts and stops. Hereafter, the terms scenario vehicles or scenario sites refer to those associated with the red network and not those of the background actors or their activity.

Remote Sensing Concepts

Starting in the mid-2000s, large-format airborne imaging systems were being developed and fielded for military and other applications. These systems used multiple large-format optical focal planes to capture oblique panchromatic imagery of the ground from an airborne

platform in a circular orbit. Through sophisticated orthorectification algorithms and supercomputer-class processing hardware, the systems stitched all the raw data into large mosaiced images that appear as if they were collected from directly overhead. These early systems, which could produce imagery at approximately 0.5 meters per pixel at about two frames per second over small city-size fields of view, were termed wide-area motion imagery (WAMI) sensors [12]. While WAMI sensors were an amazing achievement in optical engineering and image processing, it was unclear at the time how best to make use of these nascent capabilities and the large volumes of data they produced.

When designing this game, we wanted to explicitly explore the applications of WAMI to the problems of network discovery and so made motion imagery the primary mechanism by which data about the scenario network were gathered and provided to players. We chose a sensing concept in which WAMI is collected from a hypothetical sensor over an area of interest that is 5 kilometers by 5 kilometers, which would have the majority of the roughly 8-kilometer-by-8-kilometer urban core of the city continuously within the field of view. A graphical depiction of this area is shown in Figure 3.

Several hypothetical collection concepts of operations were explored, including the real-time downlink of small chips of imagery that are a subset of the full sensor field of view and the traditional paradigm of offline data processing and the use of WAMI in a forensic capacity only. We also developed a companion sensing concept for an airborne ground moving target indicator radar that would provide coverage of up to a 20-kilometer-by-20-kilometer field of view in the suburban and rural areas surrounding the city. However, in early testing of this concept, users struggled to interpret and make sense of this nonliteral data modality, and it was later removed from the game to focus on the higher priority task of determining the best utility for WAMI.

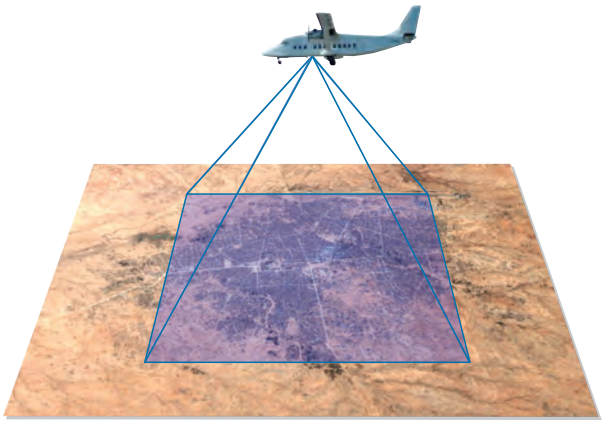


FIGURE 3. The illustration shows the sensing concept used in the game. The projected base image shows the urban core of the city and the superimposed blue box represents the instantaneous field of view of the wide-area motion imaging sensor on board the aircraft. The aircraft orbits around the perimeter of the city while the urban core remains persistently within the field of view.

Game Implementation

Data Generation

With the scenario and remote sensing concepts developed, we produced datasets that would become the primary sources of information used during gameplay, specifically a set of vehicle tracks, a multiresolution corpus of motion imagery, and a series of alert messages to cue teams to activities within the data.

The first step in the data generation process was to obtain a multispectral satellite image of the city from a commercial vendor to use as the basis for all other data products. The image was used to assign physical locations to the sites and events from the scenario, in congruence with the appearance of those locations within the imagery; for example, safe houses were chosen at locations of remote walled compounds. Next, we used geographic information system tools to develop a road network by tracing out all the primary, secondary, and tertiary roads.

With these data, we generated a vehicle track dataset by using a commercial vehicle-motion modeling and simulation tool that uses a road network, waypoints, and vehicle-motion models to generate track data through time. The scenario timeline and geographic locations were used to construct waypoints for the vehicles associated with the scenario activities, and the waypoints evolved through multiple runs of the modeling tool to match the scenario to the physics of the vehicle-motion simulator.

Next, background vehicle tracks representing the gray and white actors were embedded with the scenario tracks to create a realistic and more complex traffic environment. To create the background activities, we developed a statistical model to estimate a rough distribution of residences and workplaces across the city. Vehicles were

modeled as starting from randomly selected residential locations at a distributed set of times in the morning of the scenario with a destination randomly selected from the workplace distribution. A series of pauses and additional waypoints were then randomly selected for each vehicle to complete its waypoint list for the game duration. If a vehicle completed its waypoint list before the end of the scenario, it repeated the list until the scenario was over. This list of tracks and waypoints was then run through the same vehicle-motion modeling tool as used for the scenario tracks. To avoid confusing the game players and incurring possible false team decisions, the start, stop, and waypoint locations for background vehicles were filtered to reject areas that were at or near any of the static red scenario locations. Lastly, the track data were run through a process to apply noise to the true vehicle dynamics and to break tracks into multiple segments, thus mimicking the problems associated with optical multitarget tracking systems of the era.

Next, to generate the motion imagery dataset, we leveraged a technique from early video game graphics by which two-dimensional bitmap images, or sprites, are embedded into a larger image and then rendered as a single scene. To produce the base image, a graphic artist modified the original satellite image of the city to erase any vehicles visible on roadways and adjacent to sites associated with vehicle tracks in the scenario. Additionally, any people visible were also removed because the sensing concept used in the game instructs users that people are not visible at the resolution used. Next, exemplar vehicles, such as cars and trucks, were extracted from the unmodified satellite image and turned into sprites. To produce the simulated vehicle movements, the vehicle track positions at each time step in the scenario were turned into pixel locations in the modified base image, and the vehicle sprites were rotated to the direction of travel and inserted onto the base image. The resulting new image was rendered with vehicles embedded. This process was repeated for each time step of the scenario to generate a full-scale motion imagery dataset.

Lastly, we developed a dataset of text reports, or messages, to help give context to activities in the motion imagery data and to help keep the teams focused on the game objectives since teams will frequently get stuck on red herrings with the sensor data alone. In conjunction with the scenario creation, messages were written to tip the players to events of interest in the imagery, such as reporting of overt attacks. Each message contained information about the originating source—for example, regional news organizations, local law enforcement, NGOs, and coalition military forces—and about the time and location, with varying degrees of precision, that the text referenced. Some events generated multiple messages from multiple sources, requiring players to assess each message's relevance and veracity with respect to the objectives of the game.

Game Architecture

Because we wanted to employ a large degree of video data manipulation and collaborative tools and interfaces, we were unable to find an existing game development framework that met all the requirements, so we developed our own purpose-built game architecture. The approach was to push as much of the processing and display tasks to server-side components so that the game client could be made lightweight and responsive to players. Additionally, we wanted all game state information stored on the server so that if players accidentally closed their game client, it could restart right where players left off with no information loss (this feature is especially important in teamwork settings).

A game client named Bluestreak was developed in Java and built around NASA's WorldWind, an open-source software development kit for visualizing and hosting geospatial data in a 3D globe-like interface [13]. A description of the user-interface features and a screenshot are discussed in Figure 4. In addition to the individual client features described in the figure caption, another major capability was the ability to collaborate across Bluestreak clients; for example, when a user made a placemark, i.e., a geospatial bookmark, on one client, that object showed up on all other clients, greatly improving shared awareness that underpins effective collaboration. Also included was a set of interfaces that the teams could use to codify their final decisions to enable automated scoring of their answers. All user actions executed in the tool, such as user-interface state changes, and all polling events, such as the geospatial and temporal extents of the current data displayed in the map, were recorded with specialized software instrumentation.

The game server consisted of three major components: a specialized imagery and geospatial data server, a relational database, and a web service communication

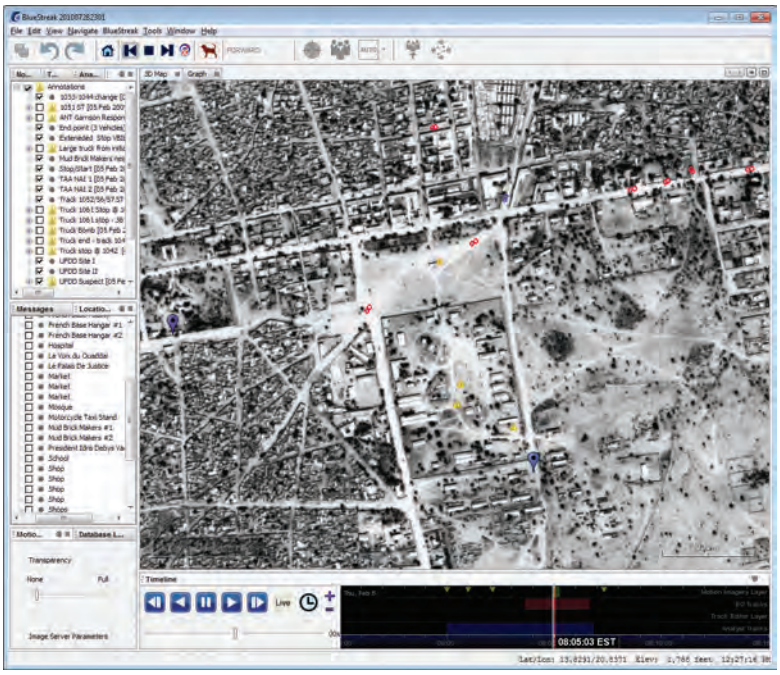


FIGURE 4. In the Bluestreak game client, the center map display fills the majority of the user interface and is flanked by configurable panels on the left and a custom timeline control on the bottom. The user-interface panels on the left are user configurable to enable viewing of additional layers of data on the map display, including data provided as part of the game and data generated by players. Provided data include geographic information system data, such as named areas and locations of interest relevant to the scenario, or text displays that show reports received as part of the scenario. User data can be geospatial polylines of vehicle movements, called tracks; geospatial bookmarks made by users, called placemarks; and other information. The timeline control allows users to manipulate the rendering of imagery, vehicle tracks, and other data by using a single temporal extent or selectable time range. This function, which gives users the ability to scrub forward and backward in time and see patterns in the data as they render on the screen, is especially useful for analyzing the movement behaviors of vehicles.

channel. The generated WAMI data was passed to Bluestreak through a custom-built JPEG2000 image server, designed to scale to multiple streams of imagery data sent to tens of clients. Hereafter, the term *video* will be used to refer to the viewing of these streams of motion imagery. The base satellite imagery and other geospatial data were served via an open-source web mapping system called MapServer [14] and translated into a pyramid of multiscale image tiles that can be efficiently passed to all the game clients for display. All game geospatial, message, and instrumentation data were read from or recorded to a PostgreSQL relational database, with PostGIS spatial database extensions. Lastly, publish and subscribe web service interfaces were used to transfer the data between the game server and game clients.

Human-System Instrumentation

From network operations control centers to expeditionary military detachments, teams of humans interoperate with complicated systems to create complex sociotechnical enterprises. Within these enterprises, the most critical component of overall performance is that of the humans, yet their contribution is often the least understood. Traditional measurement methodologies, such as human observation, are often subjective and anecdotal and can suffer from biases and differences in interpretation. Additionally, existing tools to measure human behavior can be qualitative and are insufficient in capturing intricate dynamics within an individual (intra-individual) and between individuals (inter-individual). Lastly, the time- and human-intensive collection of these data does not scale to large organizational studies. These limitations hinder the ability of researchers to draw objective conclusions and understand the parameters influencing team success.

Over several years, we have developed a data-driven research methodology and technical framework, Humatics, to address the aforementioned challenges by quantitatively measuring human behavior, rigorously assessing human analytical and cognitive performance, and providing data-driven ways to improve the effectiveness of individuals and teams. Humatics incorporates three major areas of research: system-level, physiological, and cognitive instrumentation; assessment methodology and metrics development; and performance feedback and behavioral recommendation. Figure 5 depicts our instantiation of this approach and its application to the study of teams' abilities to effectively discover data, make sense of those data, and make decisions in the context of a serious game.

The development of an instrumentation and data collection strategy for a given human-system research effort requires a careful consideration of the specific

learning objective for the process under study and the identification of observables to be measured to enable insight. A measurement strategy can then be based on which method and phenomenology are best suited to directly or indirectly measure those observables. For our research, specific instrumentation modalities were chosen to augment qualitative human observations with nearly continuous collection to enable the analysis of dynamic low-level behavioral signals.

The first element of the framework in Figure 5 is the instrumented analyst workstation, where both system-level and physiological instrumentation are used to characterize human-system interactions. System-level instrumentation is accomplished through the insertion or enabling of software code that logs graphical user interface interaction events, queries to and transactions with databases, the data visible to the user, and more. To add context to the data, screen recordings are continuously captured and a research-grade eye tracker detects the user's location of gaze on the screen. This physiological information is used for cross-referencing the system-level data.

The next element is cognitive instrumentation, which is used to measure behaviors associated with the cognitive processing of information. To quantify the comprehension and situational understanding of teams during scenario-based training or serious games, knowledge elicitation techniques are employed [15, 16]. Measuring a player's or team's understanding requires explicit elicitation of information from individuals through a series of free-response and targeted multiple-choice or Likert-scale questions that are focused on the concepts of comprehension and inference development. Comprehension is a measurement of the facts presented in the data (e.g., who, what, when, and where), such as the location and time of an attack, and an inference is a measure of a player's interpretation of the data (e.g., how and why), such as who a player believes facilitated the attack and the attacker's possible motive. In addition to its use for gaze tracking on the screen, the eye tracker is used to perform pupillometry (precise measurement of the pupil's diameter) to noninvasively estimate human cognitive load [17], another facet of cognitive instrumentation.

The last framework instrumentation modality uses wearable sensors called sociometric badges [18] to record nonlinguistic metadata of speech behaviors, body movement, and other data. Originally developed by the MIT Media Laboratory, the badges have often been employed in longitudinal studies of the communication patterns of large organizations. We used badges with modified firmware and custom post-processing software

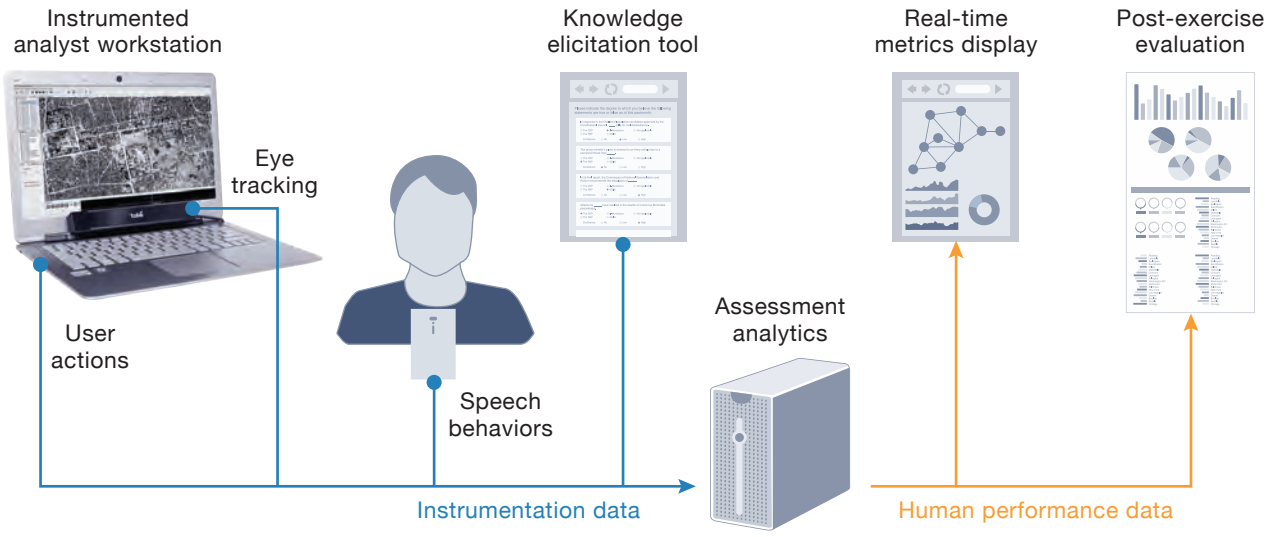


FIGURE 5. This diagram depicts the Humatics framework—a platform to measure and make sense of human analytical performance data. System-level, physiological, and cognitive sensors and instrumentation are used to produce rich quantitative data of human-human and human-system interactions. Instrumentation data are jointly processed with advanced metrics and turned into measures of human performance that are visualized in custom displays to provide performance feedback and pinpoint areas for behavioral recommendation.

to increase granularity for small group dynamics within hierarchical teams.

Our collected instrumentation data were processed with specialized metrics and used for real-time diagnostic displays or post-experiment assessment. Real-time displays allow for immediate team evaluation to enable behavioral redirection, while offline post-processing supports in-depth analysis and process improvement. Our team assessments are an example of the latter.

Mechanics and Gameplay

White Team

The role of the white team (not to be confused with the white scenario actors) is to ensure the smooth, effective operation of the game. This oversight includes monitoring the physical setup and tear down of the gaming facilities, preparing and presenting all materials, conducting briefings and training, and generally facilitating the overall event. As facilitators, the white team answered questions about tool use and reminded teams about overall objectives, but they did not give away information about the scenario or provide feedback during gameplay about the relative effectiveness of different strategies. The white team provided real-time assessment at the end of the game and briefings of the results to the different team members. White team members included many of the original game developers and other staff who have extensive experience with the game.

Game Event Timeline

After the initial test versions of the game were employed, we honed in on a game event process that would allow for four to 12 competitive teams per day to play through the exercise, depending on available hardware infrastructure and white team members, with a game event lasting one or two days. More than 80 teams and 400 participants have played this game over the life cycle of this research effort, and we have analyzed in detail a large subset of these teams.

We began the game event process by obtaining informed consent from the participants in accordance with approved human-subject research protocols. During different phases of this research, we recruited subjects from a wide population that included college students, scientists, engineers, professional military analysts, military instructors, and senior government officials. Next, the participants received introductory briefings that highlighted the research purpose and goals, and provided background knowledge, such as a primer on social network analysis. The network analysis primer is critical to success in the exercise because it introduces concepts about how people and facilities are associated in a network, what differences exist between static and transient location types, how leadership is often isolated within dark networks, and how to build and interpret graph network diagrams. After the background presentations, participants received a live plenary tool demonstration, followed by a mission briefing that oriented them to the scenario and tasks they would be required to perform. This presentation was designed with the look and feel of a military-style mission briefing, with fragmentary operational orders defining the rules of engagement, an overview of the city and its destabilization, an overview of the remote sensing and other data capabilities available to teams, and a review of possible end-state courses of action and recommended decision criteria.

Next, individuals were assigned to teams, known as the blue teams, through a process that used limited demographic data collected during orientation to attempt to balance the team members' backgrounds, skill sets, seniority, and organizational affiliations. Teams then moved to separate rooms where they could play the game and deliberate in private, and where individualized training on the game tools would take place. The white team used training checklists to ensure that each participant had a minimum proficiency with the game software. Next, a team strategy session took place, and teams prepared for a practice scenario. The purpose of the practice scenario was three-fold: to try out the plans of action that teams developed in the strategy session, to identify any areas of training that needed reviewing, and to be familiarized with each facet of the gameplay. A second team strategy session allowed teams to discuss what went well and what went wrong during practice, and regroup before the start of the main exercise.

During a short break after the main exercise, the blue teams moved back to the plenary room while the white team scored and analyzed the teams' performance. In a following "hotwash," a representative from each team explained to all participants what that person's team believed happened in the scenario and what approach that team took. Then, the white team gave the scenario

reveal, which described step by step all the activities within the scenario and the information that teams should have found and the decisions that they should have made. Finally, the scoring and performance assessment results were presented and winners received an inexpensive trophy. In practice, we found that teams compete fiercely for the chance to win even an inexpensive trophy, and organizational affiliation and pride also significantly affect team competitiveness and engagement.

The usual game block lasted four hours, with the main exercise taking up about one and a half hours; generally, two game blocks were performed per day with as many as six concurrent teams per block. Some of the earliest games required eight hours of gameplay per team, but we later switched to a shorter, simplified game format to focus on specific teamwork and decision-making facets of the game and to yield more games played per game event.

Blue Team Strategy

One of the challenges of collaborative games is that team members often do not know one another or have not worked together previously. Because this arrangement can lead to ineffective team dynamics, one of the purposes of the two team strategy sessions is to force a dialog between the individual players to get them to think about team structure and roles. During these sessions, we instructed the teams to consider these five major facets:

- Approach. Teams should think about the scenario briefing and decide on an initial concept of operations, which they can later refine in the second strategy session once they've tried it in practice. Members should also discuss their assumptions about the scenario, their risk tolerance, and other factors so that there is less potential for conflicting ideals later in the game. They should also decide if they want to use some of the automated tools provided in the game software or stick with a more manual tradecraft.
- Resource allocation. Teams are provided one less game workstation than the number of team members, so they need to decide how to allocate their human and compute resources. In early testing, we found that if we gave every player a workstation, the members failed to organize into a team, and by having one less game client than players, hierarchies formed with one player taking a leadership and integration role and the rest taking on the discovery tasks. Teams also have the option of not using all of the workstations, and some opt for a pair programming model with two players collaborating around a single workstation.
- Team roles. Teams need to decide who plays what roles, generally leader and worker roles. Leaders solicit workers for the latest information to synthesize into higher-level meaning and also often serve a scribe function by categorizing this information on the provided whiteboard or other means. The worker roles break out into a multitude of possible tasks, such as tracking vehicles from source to destination, watching for new messages to alert the team, and building the network, either on a whiteboard or in the graph tool in the game client. Multiple players may take on any of the leadership or worker positions, and it's up to the team to self-organize their gameplay.
- Collaboration. Teams must ask how they will function and collaborate on the tasks that need to be performed. For example, who will assign tasks and track their status, and who will monitor work that has been done? Because the game client provides a number of ways to annotate with text and color the information discovered and input it into the software, teams should discuss naming and color conventions, such as putting player initials on information or using the color of annotation to label potential decision criteria.
- Decision making. Teams must decide how to select a course of action related to the sites they have discovered in the game. They should discuss if they want to make decisions as they discover new information or wait until the end to take stock of all available information. They should also determine how aggressively or conservatively to play, judging how their decisions and the ensuing risks and rewards impact scoring and game performance.

Gameplay

After all the training, practice, and strategy sessions, gameplay on the main scenario began with two to six concurrent competitive teams. Teams contained between three and eight players, with the standard configuration being five—four players on computers and one team leader. The task given to the teams during the mission briefing was to uncover as many of the sites (locations) used by the red network to perpetrate the attacks in the scenario, and then to make recommendations on a course of action for each discovered location by the end of gameplay. Within

the game were two main phases: the discovery phase, in which players analyzed the video, track, and message data to discover red activities and their associated locations, and the decision phase, in which players synthesized their collected information and adjudicated their uncertainty and risk to choose courses of action. How teams moved between the two phases was one of preference: some teams spent the first 80 percent of gameplay discovering information and the last 20 percent making decisions, and other teams assigned potential courses of actions to sites as they discovered them and continually adjudicated those decisions throughout gameplay. A visual depiction of the game workflow, broken down by the two phases, is shown in Figure 6.

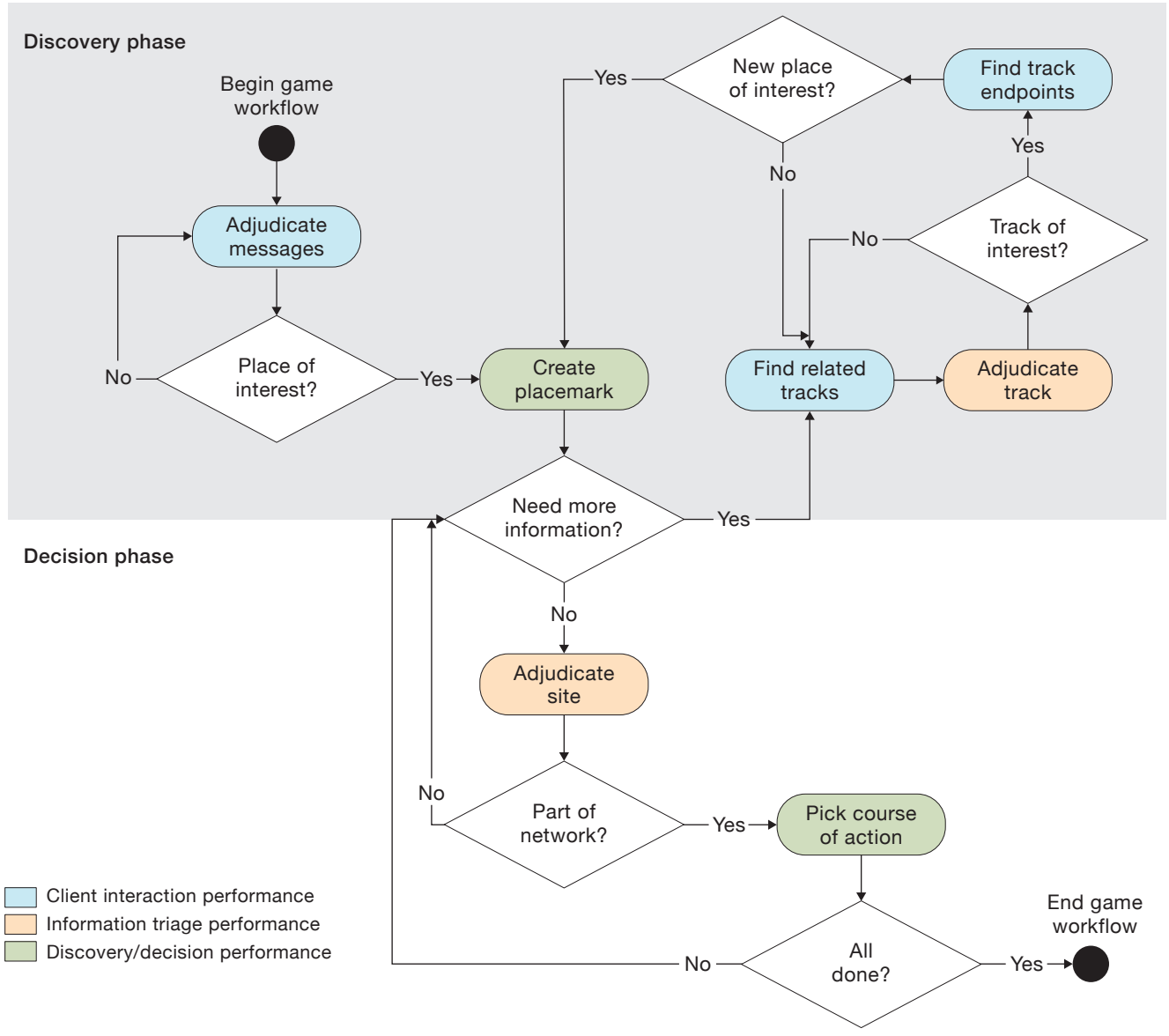


FIGURE 6. A canonical workflow diagram of gameplay provides a visual of which steps in the workflow map to specific measures of human performance in the game. The top half of the diagram shows the discovery phase of the game, during which players triage and make sense of game data to discover the network of actors and facilities they are trying to uncover. The lower half of diagram represents the decision phase of the game, during which players adjudicate the information they discovered and make risk-informed decisions regarding which locations they believe are part of the scenario network and how strong a course of action should be taken against those locations. Each of the three different types (colors) of game performance was the focus of a large human-subject experiment and assessment.

Photo: U.S. Navy (a) Photo: U.S. Navy (b)

FIGURE 7. Participants engaged in an exercise with the Naval Special Warfare Command. The analysis discovery phase of the exercise is picured in (a), and (b) shows the later decision-making phase.

As the scenario began, teams were alerted in real-time to events unfolding in the scenario via messages that arrived and were cued to the place and time in the video associated with the messages. Players observed the events in the video and adjudicated the relevance and veracity of the associated messages since messages can be factual or ambiguous depending on the message source. If the location of the activity in the message was of interest, players made a placemark there and then queried for tracks that either originated or arrived from that location. They then determined if any of the tracks were associated with the event through spatio-temporal analysis of the video. Players followed tracks associated with the previous red activity to their source or destination and entered placemarks at those locations to indicate potential association with red activity. As the scenario evolved, more messages came in, cueing players to other locations of both relevant and nonrelevant activities. Through the association of video vehicle tracks with their user placemarks, players built out the network of red sites. Teams could catalog their understanding of the network as it evolved by using tools within Bluestreak or on the provided whiteboards and large-format notepads.

As the teams entered the final decision phase, they went through each of the placemarks believed to be associated with the red network, discussed the evidence they had accrued about that site and the courses of action they should take, and then chose from three potential actions in the placemark menu:

- Assault. Sites that should be assaulted are those that have a static association with the red network and that, if law enforcement or military were sent to interdict these facilities, would certainly reveal red personnel or material. Examples of sites to assault are safe houses, weapons caches, and the red headquarters.
- Surveil. Sites for which the team cannot determine if they should be assaulted or regarded as transient sites associated with temporary red activity, and should be nominated for continued surveillance because they may be associated with red activities in the future. Examples of surveil sites are attack staging areas, the garrison of the local army, and long-duration stops by red vehicles.
- No action. No action should be chosen for all placemarks that are not associated with the red network and are innocuous.

This process continued until all placemarks were adjudicated and courses of action chosen, with no action being the default action. As teams approached the expiration of game time, team dynamics became very animated, often with heated discussion and a frenetic pace of locking in and checking all course-of-action choices. An example of gameplay can be seen in Figure 7.

Team Scoring and Evaluation

Depending on the game event, teams are evaluated across multiple performance factors, including decision making, information discovery, and verbal communication, with team decision performance as the primary mechanism for declaring a winner. After the teams

Blue action	Site class: Red facility	Site class: Red activity	Site class: Gray sites
Assault	+4	−2	−2
		Risk, loss of good will	
Surveil	+2	+1	−1
			Resource waste
No action	−1	0	0
	Opportunity loss	Null situation	

FIGURE 8. In the scoring matrix used to adjudicate the decision-making performance of teams during the game, the columns represent three classes of locations, or sites, and the rows correspond to three levels of courses of action the teams can assign to each instance and class of sites discovered during the game. Cells shaded in green indicate that the team's chosen course of action was appropriate for the respective class of site, resulting in a gain of points, and red cells represent a course of action that was inappropriate, resulting in the loss of points. Gray cells represent action and class mappings for which points were neither gained nor lost.

finished the game and their decisions were stored in Bluestreak, a server-side scoring script was run to take into account several factors, such as geospatial closeness, to determine which teams correctly identified the location and value of each of the red sites in the scenario. A scoring matrix was used to award points to each correctly identified site and points were subtracted for incorrect decisions, as detailed in Figure 8.

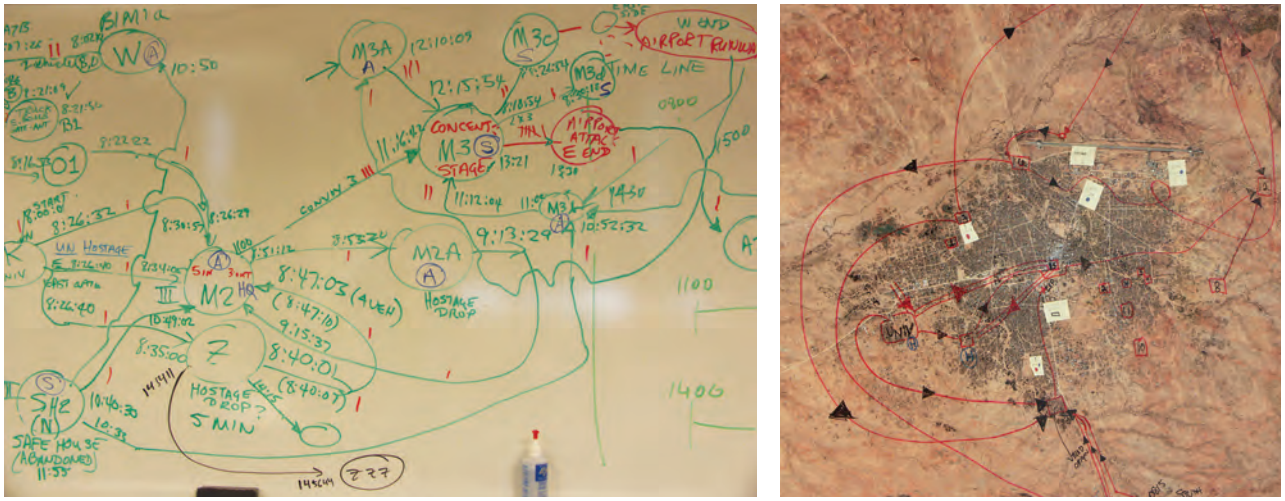
For red facilities, the correct decision in the game was to assault, and it earned the most points. If the facility was surveilled instead, then half the point value was awarded because some information was gained, and if no action was chosen, then points were deducted because the opportunity for some discovery was lost. For red activities, the correct action was to surveil them and points were awarded accordingly. If a red activity was assaulted, points were deducted because this action added risk to the interdicting force and lost good will with the local population when an innocuous location was assaulted. If no action was chosen for red activities, then points were neither awarded nor deducted. For gray sites, or those involving the background populous, points were deducted for an incorrect assault and for a surveil because these actions lost good will and wasted surveillance resources. The correct action for all gray sites was no action.

The weights of the points between the levels of the courses of action and their correct and incorrect value were constructed to match the concept of the scenario while also prohibiting teams from trying to "game" the game. Point values for the red facilities and red activities were totaled into a single score for each team, and the team with the highest score of the game event won. Often, scores could be negative if teams were aggressive in their approach, and if a tie occurred, additional performance measures were used to break the tie.

Experimentation Phase 1: User-Centered Tool Development

Considering the work involved in the development of the sensor and traffic simulation models and the complex scenario, we knew that completing this game would be challenging and that some tooling and automation would be required, especially with respect to information organization and knowledge management, for the game to be effective. However, rather than building those capabilities into the initial iteration of the game software, we wanted to use this opportunity to learn new methods for designing effective human-system tools.

In general, users are ineffectual at explaining to others what is hard for them and what types of capabilities would improve their work process. Frameworks like user-centered design have gone a long way toward analyzing and envisioning how users are likely to use technology, and then validating those user behavior assumptions with real-world tests and evaluation [19]. In our case, because we were working with a new type of data, WAMI, with no established workflows and best practices, explicitly studying the intended user was not straightforward. Instead, we wanted to see if gameplay could be used to implicitly learn what tasks were hard for users and where in the process there was friction. Our approach was to study user solutions to the game in the absence of the needed tooling and then turn our observations and user artifacts into a requirements specification for developing new user-centered capabilities. Once those new capabilities were developed, we could use the same methods to deploy the capabilities, measure their utility, and retool them to be more effective.



(a) (b)

FIGURE 9. Teams used a whiteboard and map to manually organize information during gameplay. The image in (a) shows a node-and-link network diagram representing different sites (circles) discovered in the game data and the vehicle tracks between them (lines). Also annotated on this diagram are names given to each of the sites and tracks by the teams, and the start and stop times of the vehicle journeys. The image in (b) shows a geospatial network view of similar information using markers and sticky notes on a laminated map.

Experiment Design for Requirements Generation

In this phase of research, we wanted to better understand how users might best use real-time video information to aid network discovery during an unfolding event. We designed an experiment in which users had a static base satellite image on their map display and the ability to overlay streams of up to eight real-time 100-meter-by-100-meter video chips from the airborne WAMI sensor's field of view. Users could slave those chips to follow a specific vehicle or persistently stare at a location on the ground. In this construct, players not only had to manage their human and compute resources but also their sensor resources. While video was only available within the eight available chips, track positions of vehicles were available across the entire sensor field of view. However, when vehicles stopped within the scene, the track broke and started with a new track identifier as the vehicle started moving again, thus requiring teams to devise methods for how best to mentally stitch all these tracks back into a single vehicle journey. We knew bookkeeping was going to be a challenge in this experiment but wanted to see the methods that teams came up with through gameplay.

A series of team games was deployed, and we used both photographic and room video recordings to track how teams discussed and captured information via the whiteboard and hardcopy maps. Among the many different approaches to capturing and coding the game network information were the two examples of this instrumentation seen in Figure 9.

By studying how teams solved various problems through different methods on the whiteboard and paper map, we determined the requirements for a set of tools that users would have liked to have had during the exercise. Figure 10 shows how a team's map suggests ideas for a new tool. In this example, a player could benefit from a network visualization tool that is integrated with the map and track paradigms within the game client. The requirements for the tool fall into three groups of network information representation:

- Node information. Users would like the ability to customize the names of sites (nodes) with their own annotations and to represent track metadata, such as the duration of a vehicle stop, as attributes of a particular node.
- Link information. Users would like the ability to visualize track metadata along links (tracks) between nodes (sites); such metadata could include name annotations, departure and arrival times, autogenerated track identifiers associated with a track, and the number of track (vehicle) counts between two nodes.
- Graph layout. Users would like to represent

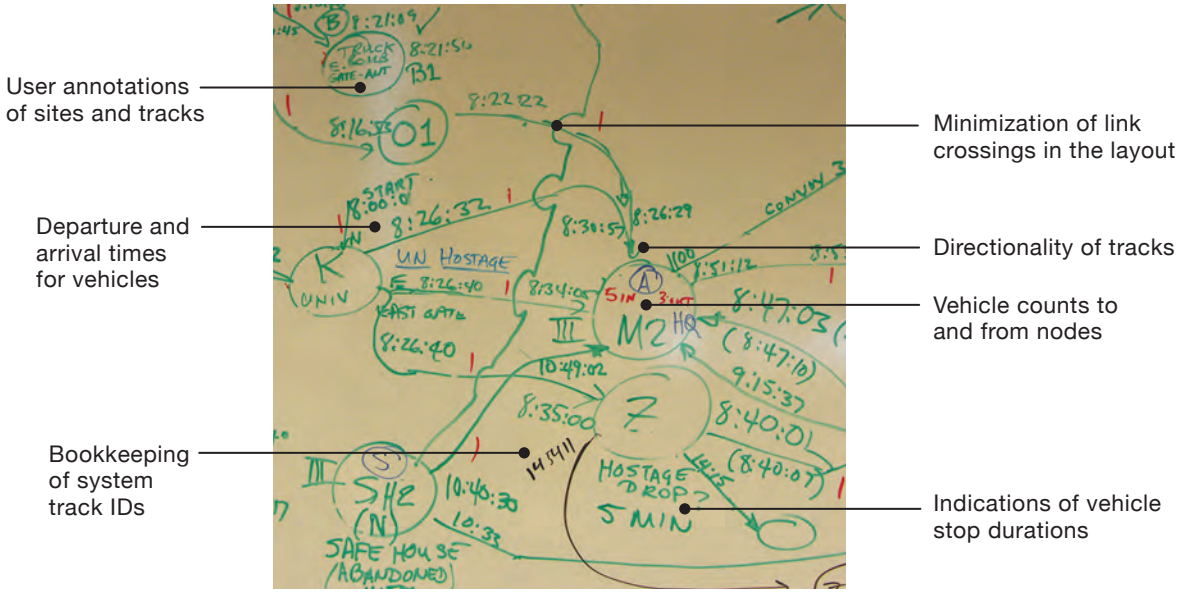


FIGURE 10. Studying how teams manually organized their information can provide insight on ways to improve information management through new tool development and optimization of existing capabilities. In this example, the callouts detail software requirements or features that would address some of the information management and visual layout needs of building a network diagram of sites of interest and the vehicle tracks that transit between them.

source-to-destination directionality of tracks and to minimize the crossing of links in the graph representation.

Development of Network Analysis Tools

Our requirements development process led to three new major features that were added into Bluestreak and the back-end game architecture:

- Joint space-time queries. In the initial iterations of the game, if players had interest in vehicles that may depart or arrive from a site of interest, they would have to scrub through all of the temporal extents of the data to find tracks. To increase the efficiency of this operation, we created a feature called Nomination to allow players to choose a point on the map, a temporal extent, such as 30 minutes before and 30 minutes after the current time step, and a geographic radius, such as 50 meters around the selected point. The game server would retrieve all tracks that matched that joint space-time query and display those to players.
- Track repair tool. As mentioned in the section on data generation, track breaks were introduced to mimic the real-world performance of optical multitarget tracking systems of the day. With those systems, tracks would manifest as single source-to-destination journeys and comprise multiple track segments, requiring players to monitor which track identifications corresponded to which vehicle journey. To improve this process, we developed a tool named Bloodhound to allow players to use the video data to positively identify when the same vehicle is responsible for the end of one track and the start of another. Bloodhound then lets players stitch those two system tracks into an analyst track, greatly simplifying the information management and network representation.
- Integrated network visualization tools. As shown in Figures 9 and 10, organizing and visualizing all of the information related to the sites and tracks that form the scenario network requires a lot of effort and bookkeeping to be useful for unraveling the game scenario. The new Nomination and Bloodhound features enabled the players to quickly find tracks associated with points of interest and quickly repair them from source to destination, allowing the network to be rapidly built out and effectively visualized. We developed two graph visualization tools, one to produce abstract node-and-link diagrams and one to produce a geospatial node-and-link diagram showing the spatial representation of the sites and tracks in network. An example of both representations can be

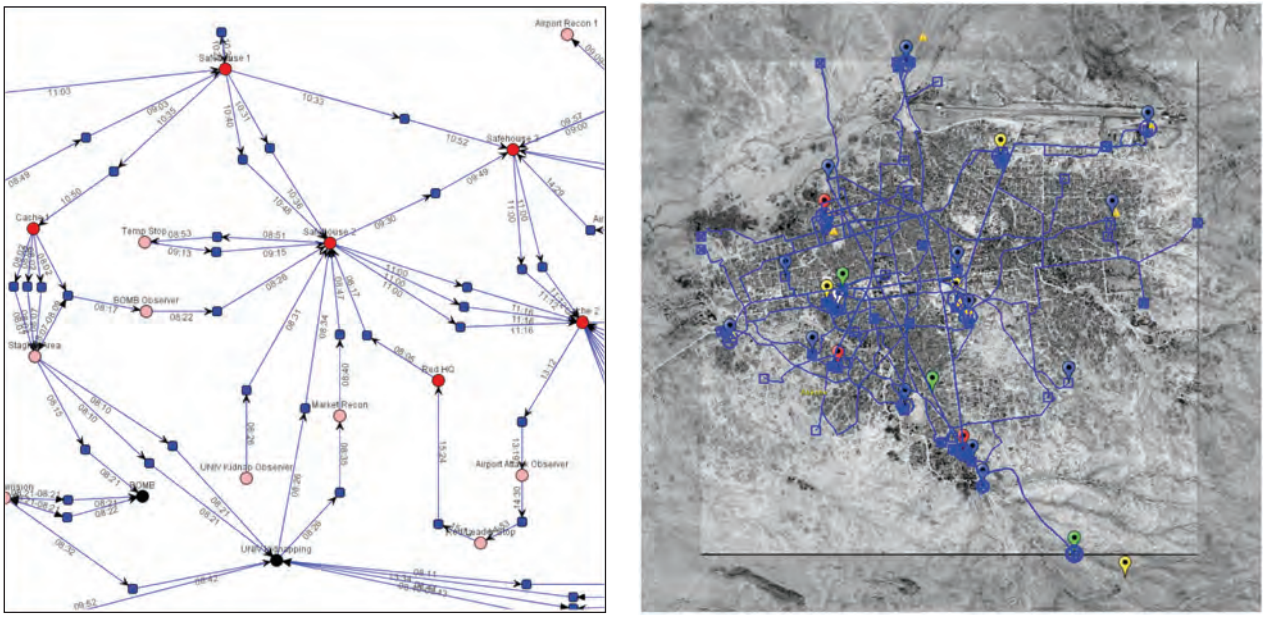


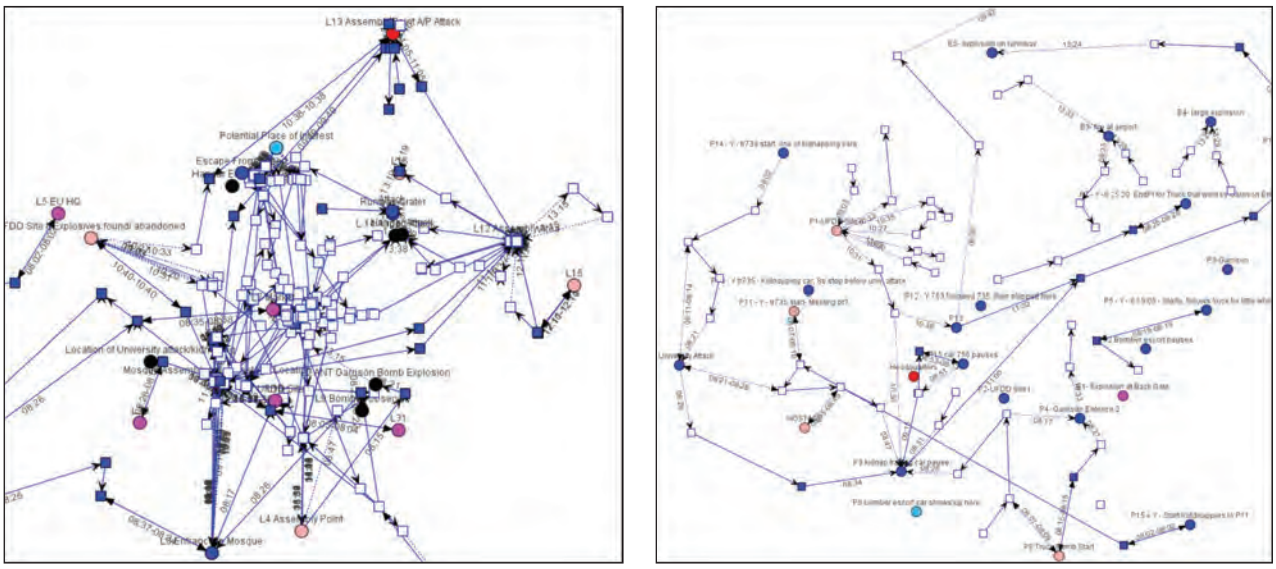
FIGURE 11. Abstract and geospatial graph representation tools are developed through a user-centered requirements process. In the abstract graph view (a), the circles (nodes) represent locations visited by a vehicle, and the squares represent an intermediate stop of a vehicle. The lines (edges) that connect the circles and squares correspond to discrete vehicle tracks between two locations, and arrows represent the directionality track moving between the sites. Similarly, (b) is the geospatial graph view. The blue circles, squares, and lines have the same connotations as the symbols in the abstract view; however, the edges now follow the full geospatial extent of the tracks they represent. Additionally, the multicolor pushpin icons represent placemarks of interest to the team.

seen in Figure 11. While abstract node-link diagrams have been used for a long time, the geospatial graph was entirely novel at the time of development. Lastly, one additional key feature of the abstract graph is that it was built to be fully collaborative with the other game clients, so when one player moved a node on a client, the node also moved on all the other game clients, allowing teams to have true shared representations.

Utility Assessment

After some initial user testing of the new tools, a series of game exercises was employed to assess the utility of these tools in improving the abilities of teams and reducing some of the human-intensive aspects of the network discovery and information management. During the debriefings from these exercises, we found that in general the players really liked the Nomination feature to find tracks associated with a site of interest. However, the judicious use of this feature had an unintended consequence. The tool developers thought that after a Nomination was executed and the results returned, it would be convenient for the player to have the site and tracks associated with the Nomination automatically placed on the graph. But this automation ended up cluttering the graph displays with both user-placed and system-placed information, with no clear distinction between the two. Once this clutter occurred, the team stopped using the graph tool and went back to using the whiteboard because that was a representation over which they had full control. One player described the automated placement function as similar to using the top of a desk to store documents that need to be read, without realizing that other people would constantly place other documents on the desktop, rendering it useless as an organizational mechanism.

To fix the placement problem, we added a step that asks users after they make a Nomination query if they also want the results added to the graph. The graph tools then started to provide great utility for network organization, and several winning teams in this testing phase used it exclusively. A comparison of a graph cluttered by the system and one built solely by players is depicted in Figure 12. This example of gameplay forcing users to



(a) (b)

FIGURE 12. Network graphs are generated by players during utility testing of new tools. The circles (nodes) represent locations visited by a vehicle, and the squares represent an intermediate stop of a vehicle. The lines (edges) that connect the circles and squares correspond to discrete vehicle tracks between two locations. The diagram in (a) is from a game in which automation added information to the user's graph, inadvertently cluttering the workspace with information of unknown provenance and limiting the utility of the tool. The diagram in (b) is a user graph from a subsequent exercise in which players were given the option to accept or reject automated information, leading to much more effective use of the tool because of a greater understanding and trust of the automation.

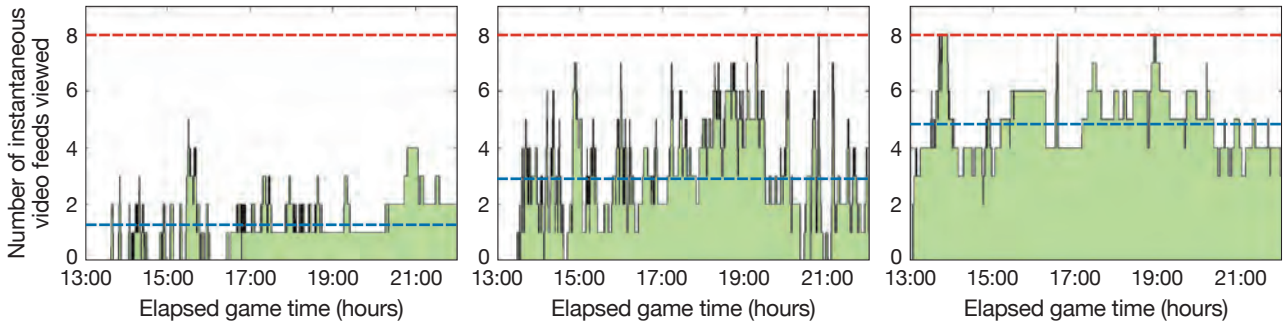


FIGURE 13. The plots provide an analysis of video feed utilization by game players. The *x*-axis represents the elapsed time during gameplay. The *y*-axis represents the number of instantaneous video feeds being viewed by a team at a given time step, with 0 representing no feeds in use and 8 representing all the feeds being used. The maximum number of feeds is represented by the dashed red line and the average number of feeds used is represented by the dashed blue line. Case (a) shows video utilization with the baseline tooling that precluded effective use of the video feeds, with an instantaneous average of 1.3 video chips. Case (b) shows improved video usage after new tools were deployed in a subsequent game to better integrate the video into the workflow and reduce the human-intensive nature of using the video, with an instantaneous average of 2.9 chips. Case (c) shows increased video usage after refinements were made to the new tools in response to targeted user feedback, with an instantaneous average of 4.8 chips.

evaluate what parts of new features have utility and which need refinement or reimagining was much more efficient and effective than simply asking users how they might make use of a particular tool or feature.

In addition to qualitative analysis of user experience, we leveraged the game client instrumentation to characterize how well players used the eight available video chips described in the experiment design. In the initial games without the improved tools, the imagery analysis and information management tasks dominated the teams' time, precluding their ability to make best use of the eight possible video feeds, as shown in Figure 13a.

The plot in Figure 13b details increased video utilization after the deployment of the new network analysis tools (Nomination and Bloodhound) that improved the integration of the video feeds into the workflow. The plot in Figure 13c shows both additional increased video usage after the user-feedback process led to the refinement of the new tools and more effective use of the graph to organize and prioritize work.

Besides analyzing video utilization, we looked at how well teams kept up with real-time information as the scenario evolved because this ability is often associated with stronger game performance and decision making. Instrumentation was built into the Bluestreak game client that logs both the elapsed game time when players are looking at data and the point in the scenario timeline that the data are referencing. An example of this instrumentation data is shown in Figure 14. As the scenario began, an abundance of activity caused teams to spend a lot of time looking forensically at older data to orient themselves before they felt comfortable reviewing new data arriving in near-real time. Through our analyses, we found that the addition of the three network analysis tools shortened the amount of game time teams spent observing data forensically before they transitioned to real-time operations.

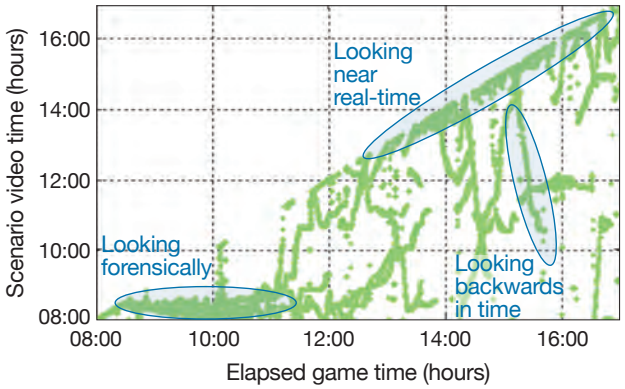


FIGURE 14. The plot depicts an analysis of how timely teams were at keeping up with real-time game data as the scenario evolved. The *x*-axis represents the elapsed game time from the beginning to the end. The *y*-axis represents the time in the scenario that the data references. Each green dot represents a record of these two timestamps, achieved through the software instrumentation within the game client. A team analyzing the data in real time would create green dots across the diagonal, and any dots below the diagonal represent a forensic examination of data. As teams oriented to the initial set of activities in the scenario, they began to view the arriving data and often did deep backward dives in time to assess all activities at a particular location.

Experimentation Phase 2: Assessing Teamwork and Decision-Making Performance

A major finding from our first phase of experiments was that team dynamics played a critical role in the outcome of the game, and anecdotally we could often predict just by observing the strategy sessions and gameplay which teams would do well at decision making. We had teams that were introverted and precise in their coordination and communication, and we had teams that were verbose and constantly challenging each other's assumptions; both of these team dynamics were found to be successful. The success of two almost opposite styles of teamwork made us want to understand on a granular level the underlying factors that influenced success. We designed a set of experiments to study teamwork and its effect on decision making. For these experiments, we modified the game format and employed the Humatics human-system instrumentation framework to augment our qualitative human observations with quantitative, persistent, and objective measurements of human-system behavior. By jointly processing the collected multimodal instrumentation data, we could make holistic characterizations of human-human and human-system interaction. The fidelity and granularity of these data were informative and, in some instances, could predict performance in the activities being measured [20].

Experiment Design for Instrumenting the Analysis and Decision Processes

To implement the second phase of experiments, we modified the format of the game to emphasize the team collaboration and decision-making components and to reduce human-intensive data analysis aspects of the original game. In this second format, we made the following primary modifications:

- Shortened the length of the scenario by half so that gameplay and the overall game event would be shorter.
- Changed the sensing concept to make all motion imagery available to players across the entire field of regard at the start of the game; having all the video data rather than only eight small time-based video chips would increase information discovery.
- Replaced the track dataset, including its track breaks and sensor ambiguities, with the ground truth track

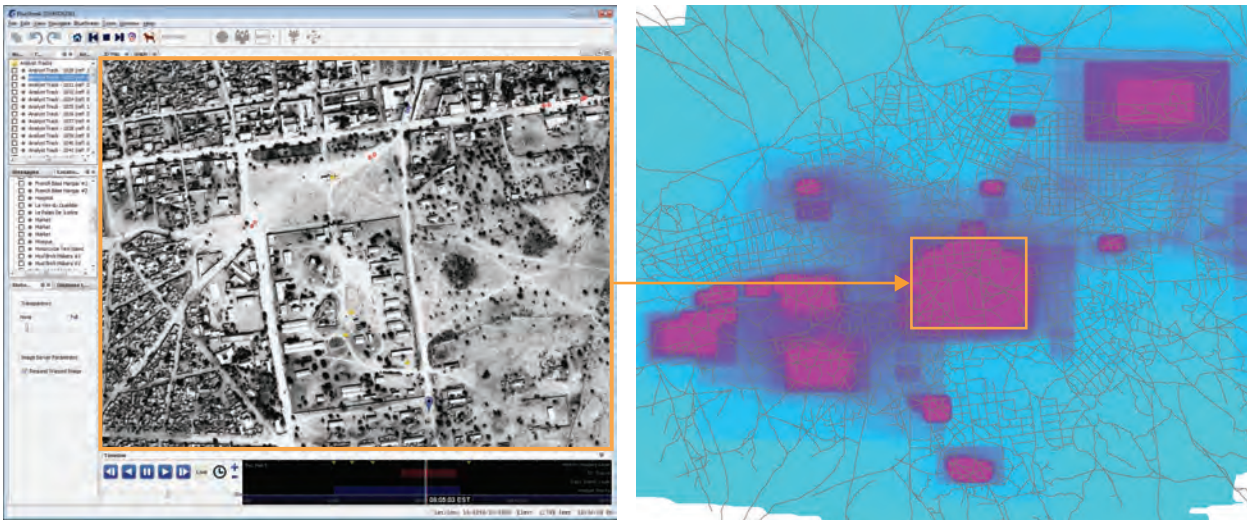


FIGURE 15. In this illustration of viewport instrumentation, the game client (a) is viewing a portion of the video data, whose viewport, or geospatial extents, are represented by the orange box on the game client and heat map (b), as indicated by the orange arrow. In the viewport heat map, areas of magenta represent areas in which the teams viewed a large amount of video data, whereas the cyan areas indicate areas looked at infrequently. These heat maps can be used to understand a team's geospatial analysis strategy.

data to allow players to focus on determining the connections between source and destination locations rather than spending a lot of time stitching together broken tracks.

- Increased the number and relevance of the game messages to ensure teams focused on the game objectives.

We also made improvements to the staging of the game event by standardizing training processes, materials given to teams, and types of information provided by game docents to players. The intent of this standardization was to reduce as much variability in the game events as possible so that the game could be run many times with different teams to produce a dataset for follow-on human-subject research.

Team Assessment Case Study

To assess human performance during the game, we decomposed each major step of the workflow and mapped it to instrumentation data and performance metrics that characterize players' behaviors. As noted in Figure 6, three major facets of performance emerged: client interaction, information triage, and discovery and decision. Additionally, the performance of this entire workflow is underpinned by a team's ability to effectively organize and collaborate through face-to-face communication. Our case study of four five-member teams illustrates how system-level and physiological instrumentation can be used to better characterize a team's performance during gameplay.

Game Client Interaction Performance

Software instrumentation built into Bluestreak recorded various user interactions both on demand and at specific intervals. The recorded data can be used to understand macro behaviors, such as the volume or rate of interactions with specific tools in the client. For example, by recording placemark creation and modification attributes, we can quantify team analytical behaviors in the workflow as a function of time. These data can also be used to analyze micro behaviors, such as a user's current look at geospatial data, known as the viewport [9]. Viewport data are recorded each second and include the current time of gameplay, the time in the scenario being displayed, and the geospatial bounding box of the video footprint in the map section of Bluestreak. An example of viewport instrumentation is shown in Figure 15.

Scenario Information Triage Performance

After the viewport data were logged, they were correlated with the scenario ground truth and processed using specialized information theoretic metrics [9, 21] to

determine which relevant (scenario network) and irrelevant (background population) tracks or sites were being viewed at each scenario time step.

A graphical representation of the scenario and background track information, shown in Figure 16a, was used to assess a team's ability to effectively triage vehicle track data. If players were properly interpreting the information in the report messages, they should have focused only on the red scenario vehicle tracks and not the yellow background population tracks. As shown in Figure 16b, the performance of Team 3 and Team 4 plateaued as the scenario evolved, whereas Team 1 and Team 2 continued to find and analyze more relevant (red) scenario tracks throughout gameplay.

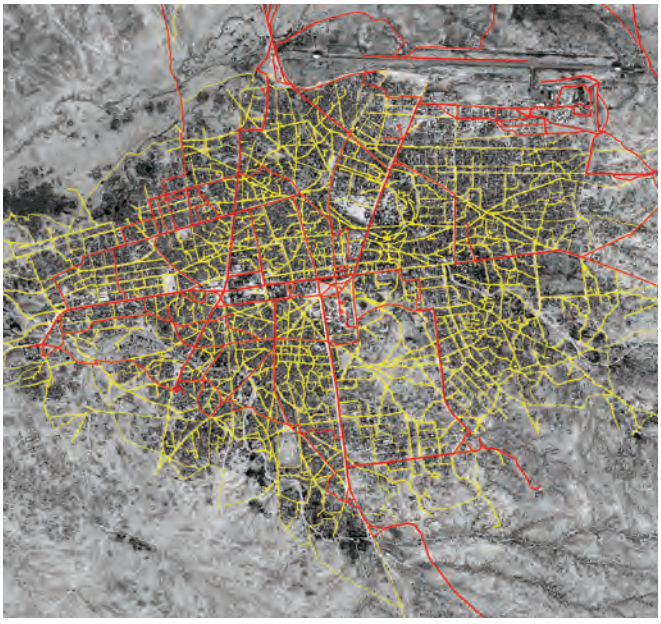
A graphical representation of the scenario red sites is shown in Figure 17a. If players are properly interpreting the information in the report messages, they should focus only on vehicle behaviors and activities around the sites with the red icons and not on the ones denoted with yellow dots that indicate locations of the background population not associated with the red network.

Similarly, Figure 17b illustrates teams' ability to effectively triage video of site-related activities. As the figure shows, Team 1 and Team 2 spent substantially more effort observing scenario site information compared to Team 3 and Team 4. In many cases, teams spent a lot of time analyzing sites but ultimately chose an incorrect action or took no action at all.

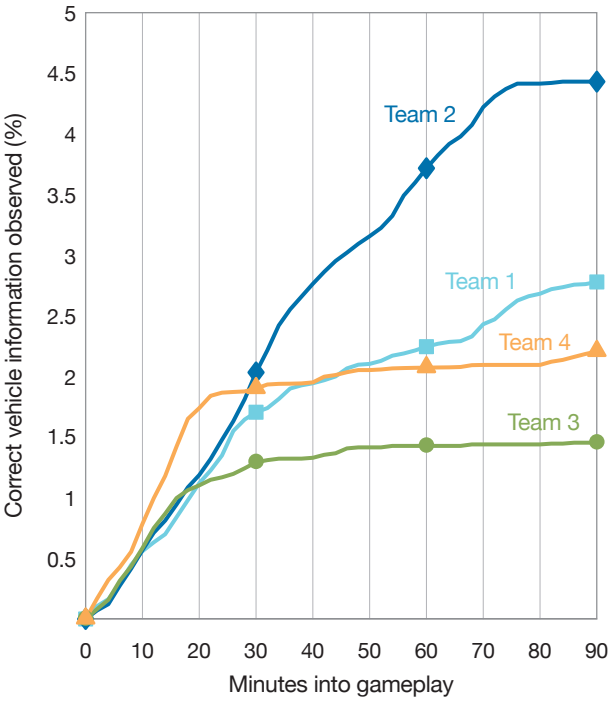
Team Discovery and Decision-Making Performance

Because the scenario was constructed to have the scenario activities completely separated from the background activity, the game can be analyzed from the perspective of signal detection theory. Essentially, teams can be considered detectors of scenario network activity in that they are attempting to extract these signals from the noise of the normal activities of the rest of the population [10]. The receiver operating characteristics (ROC) measurements of detection theory can be used to assess the teams' performance (Figure 18).

Results from two different tasks are plotted: the discovery of scenario sites, which is measured by team placemarks at those sites, and the declaration of scenario sites, which is the subset of the total placemarks that are assigned a course-of-action decision. Decision actions are



(a)



(b)

FIGURE 16. Team vehicle triage performance is depicted in the two plots. The plot in (a) shows the extents of all vehicle track data in the game, with the red lines denoting tracks associated with the scenario network vehicles and the yellow denoting tracks of background population tracks. The plot in (b) shows the team triage performance, with the *y*-axis representing the percentage of total red tracks observed in the video and the *x*-axis representing the number of minutes elapsed since the start of the game.

directly related to the teams' comprehension of the scenario and their confidence in that understanding. For example, it can be seen that Team 2 had placemarks on 100 percent of the scenario sites but only had the confidence to declare 30 percent of those sites. They also declared sites not part of the network, resulting in a 0.2 percent probability of false declaration. Team 4 had discovery performance similar to that of Team 1 and zero probability of false declaration. Team 1 had the highest detection probability but at the expense of more false declarations.

Team Verbal Communication Performance

Face-to-face communication is a key factor in overall team performance for highly cooperative tasks [22–25].

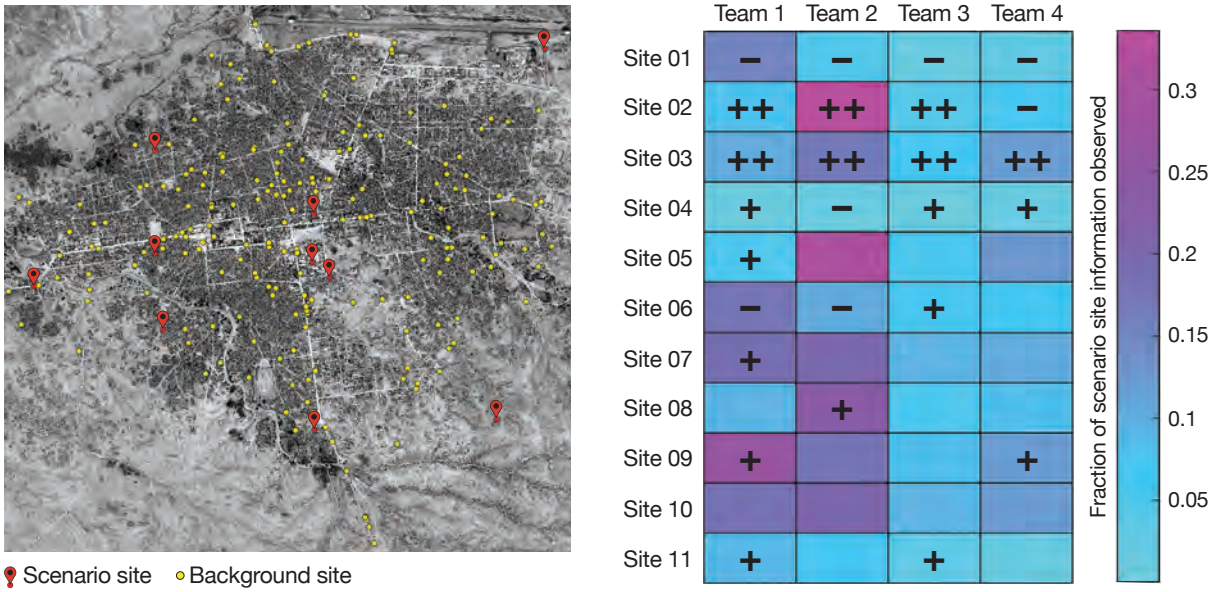


FIGURE 17. Team site triage performance is depicted. The plot in (a) shows the scenario sites to be discovered, annotated with red icons, and the background sites, denoted with yellow dots. The table in (b) shows teams' performance at accumulating information at each of the scenario sites, indicated by the fill color of each box and color bar scale. Decision outcomes for sites are also plotted in (b), with a + or – representing a correct or incorrect decision, respectively.

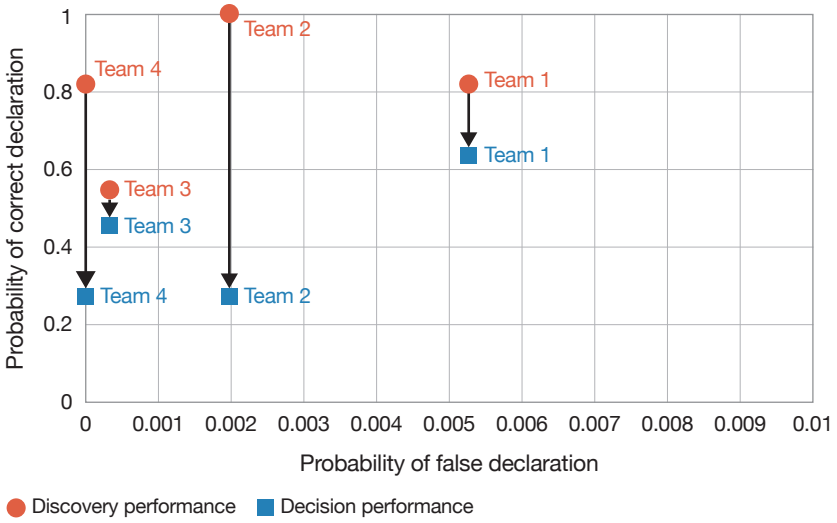


FIGURE 18. In this receiver operating characteristics plot, the *y*-axis represents the probability of correct declaration, or the fraction of correct sites found and acted upon by the teams, and the *x*-axis represents the probability of false declaration, or the ratio of incorrect sites declared divided by the total possible discoverable sites. The blue squares represent decision performance for sites that were declared to be associated with the network. The red circles show the fraction of all sites that were correctly discovered before the course-of-action selection process. The black arrows show the amount of performance lost moving from the information discovery process to the decision process, with the amount of performance loss a factor of each team's certainty about their understanding of the scenario, their risk tolerance, and their approach to making decisions.

Traditional methods to characterize these communications have largely focused on speech content; however, more recent methods center on the collection of nonlinguistic speech features that enable the characterization of team dynamics without having to analyze the linguistic content of a team's utterances [18, 23].

To collect speech metadata, we gave sociometric badges to each player during gameplay (Figure 19). The badges continuously recorded the time, duration, and identity of each player's speech, and post-processing software provided measurements of when a player spoke alone, when speech overlapped with another player, which players were listening, and when players were silent. These data naturally formed a directed graph of communication between players (Figure 20). For simplicity, graphs for only Team 1 and Team 2 are provided here.

Previous studies of face-to-face communication behaviors of small teams in a collaborative setting have found that balanced participation and speaking time along with increased turn-taking are associated with better team performance [26]. In Figure 20, Team 1 players A and C are dominating the conversation, as seen by their edge thicknesses, while the rest of the players are

FIGURE 19. A U.S. Navy service member wears a sensor called a sociometric badge (a) that can record nonlinguistic metadata of speech behaviors, body movement, and other data. The battery-powered badge (b) incorporates a number of sensors, including a microphone, wireless and infrared transceivers, and a three-axis accelerometer. Microphones combined with specialized filters and signal processing characterize when the wearer is speaking. Wireless and infrared transceivers allow the badges to identify other badges proximal to them, and when data from the nearby badges are combined with the speech data, the communication patterns of who speaks to whom within a team can be determined. This directed speaking data can be used to measure team-based speech behaviors, such as turn-taking and interruptions. The accelerometer data can determine features associated with excitement and engagement.

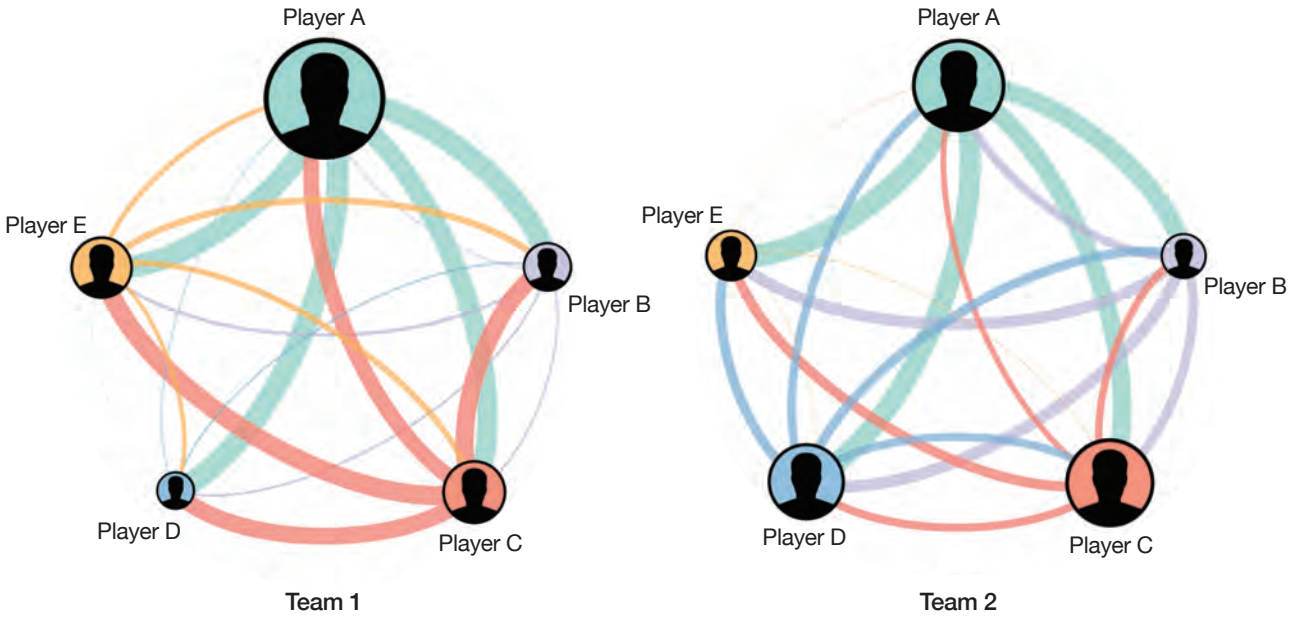


FIGURE 20. The graphs illustrate face-to-face communication networks. Vertices (circles) represent players and edges (lines) represent directed communication from one player to another. Vertex size is proportional to total participation for a player, edge thickness is proportional to directed speech time to each teammate, and edge color indicates directionality by matching the source vertex color.

less engaged with lower participation (smaller vertices) and less speaking time (thinner edges). Conversely, Team 2 has a much more balanced distribution of both speaking time and participation than Team 1, with player A acting in the role of team leader. Analysis that uses such graph data is a promising area of active research into group influencers and team role estimation [27].

For deeper insight into the communication network, we explored a social network analysis approach to characterizing player interaction. By computing the directed, normalized closeness centrality of each player [28], we can derive an estimate of the connectedness of players. Larger centrality magnitudes indicate a player's graph closeness to all other players. One useful application of this measure is to inspect the time-varying behavior of player centrality [29] during gameplay (Figure 21). In the figure, the visual representation of Team 1's and Team 2's closeness centrality can be useful for identifying team dynamics, such as the emergence of a leader. In Team 1, we see the same communication dominance exhibited by players A and C as seen in Figure 20. In Team 2, player A clearly emerges as the leader during the discovery phase of the game, with A's centrality decreasing toward the end when the team moved into the collective decision-making phase of the game.

In addition to performing a social network analysis, we did a recurrent pattern analysis that used the data collected by the sociometric badges. First, speech patterns were coded into symbols according to various speech behaviors and then analyzed as a time series [30]. The strength of the recurrent structure within these code sequences is called determinism (DET). In a strict turn-taking situation, DET will be high (near 100 percent) as the conversation is highly structured. In a situation with random speech intervals, DET will be low (close to 0 percent), indicating that the conversation is highly unstructured. DET scores were comparable for the four teams, with local maxima near 60 percent and local minima near 30 percent. Fluctuations in the values occurred over time, indicating that the structure of the communication ebbed and flowed throughout gameplay. Further analysis showed a high correlation between DET magnitude and the percentage of time an individual spoke while all others listened, suggesting that structure occurs, even in a complex team setting with five participants, when individuals speak and others listen.

Total Team Performance

We quantitatively measured team performance at several points in the overall game workflow. However, combining these metrics into a single total performance measure

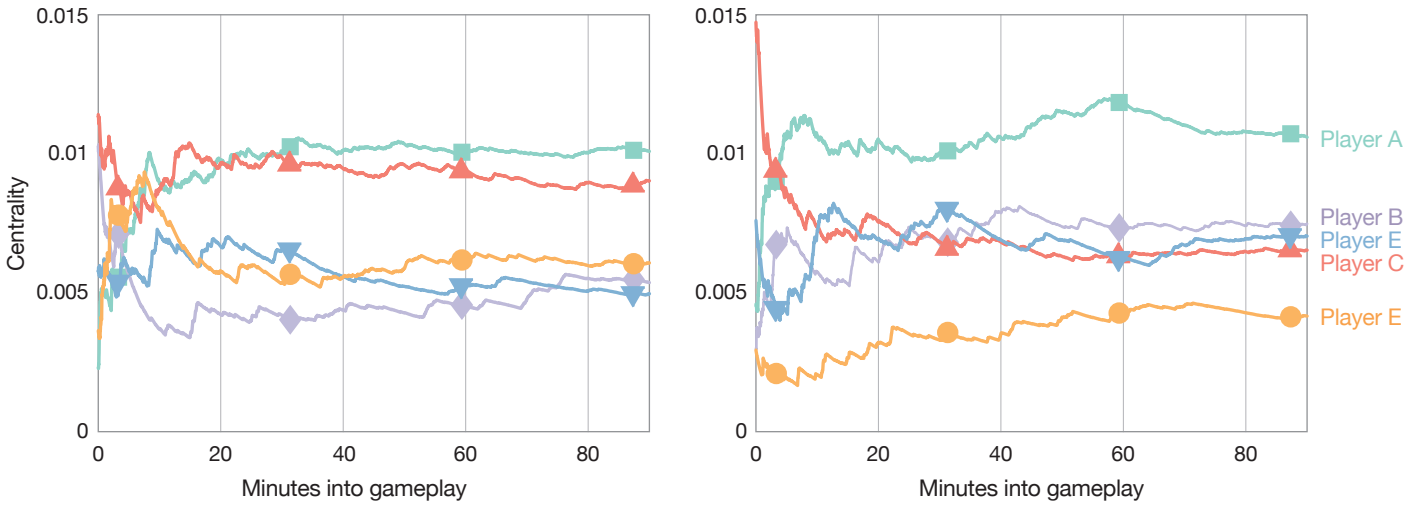


FIGURE 21. The plots depict the time-varying player communication centrality. Centrality is a social network analysis measure that can be used to identify the most important vertices in a network. The directed normalized closeness centrality of each player is an estimate of the connectedness of players in a network. The *x*-axis represents elapsed time during gameplay, and the *y*-axis represents the centrality of a player. Larger centrality magnitudes indicate a player's graph closeness to all other players. In both teams, player A is considered the leader and transitions to gain the highest centrality midway through the game. Qualitative observations during gameplay supported these findings.

warrants careful consideration. Qualitatively, Team 1 and Team 2 excelled at communication, triage, and site discovery, but they had more false declarations than Team 3 or Team 4. Conversely, although Team 3 and Team 4 did not observe as much information or discover as many sites as did Team 1 and Team 2, they were very accurate in adjudicating what they found. Team 2 ultimately won the four-team competition with the best overall performance and game scores.

Predicting Team Performance

When we assessed teams' analytical and decision-making performance, common questions arose regarding how performance in one facet of a decision process affects the performance of either subsequent processes or the aggregate overall process. The previous sections illustrate that the collected measurements enabled detailed insight about individual facets of performance; however, we wanted to take this a step further to determine whether behaviors in specific facets of the intra-game workflow were predictive of analytical performance of players or the outcomes of games. To approach this investigation, we processed data collected over several years of gameplay, encompassing 71 different teams and more than 350 unique players. For all 71 teams, system instrumentation data were recorded. For a subset of 15 teams, face-to-face communication data were also collected.

Robust linear regression analyses were used to statistically estimate how predictive were the various facets of intra-game performance with respect to workflow processes. For each model, residual analysis, significance testing, and other regression diagnostics were performed, and were evaluated for each prediction finding. In undertaking this analysis, we wanted to address three overarching research propositions:

- Client interaction effectiveness. The first proposition asked whether more effective interaction with the game software client led to better game performance. From our analysis, we found that teams who had higher usage across all analytic functions of the game client discovered more total sites and had a higher probability of correct site discovery. The effect was even more pronounced for the functions of the game client associated with the frequency with which players submitted Nomination space-time queries for track data and its correlation with increased site discovery. Higher total game client interaction was also associated with more effective observations of scenario site information and track information. Essentially, teams who were more effective at interacting with the functions of the game client observed more relevant scenario information and found more correct sites.
- Information triage effectiveness. The second proposition asked whether discovery of more scenario information led to better game outcomes. From our analysis, we found that teams who observed more relevant scenario site information and track information also scored higher in game outcome. The overall game score takes into account several aspects of how well the players perform, but it also encapsulates the confidence of players' decisions (course–of-action strength) and reflects their overall strategy for the game (aggressive to risk averse).
- Team communication effectiveness. The third proposition asked whether teams who communicate more effectively have higher game performance. Our analysis found that teams who communicated more (total time) throughout the exercise also observed more relevant scenario site information and track information. Additionally, teams who had higher participation (frequency of communication) from all members throughout the game also observed more relevant scenario site information and track information. Lastly, teams who communicated more (total time) throughout the exercise also made better decisions on the most challenging sites to adjudicate. These findings about total team engagement and participation agree with our qualitative observations of teams during the decision-making process. Team centrality metrics did not have a significant association with other aspects of team performance and warrant further investigation.

Follow-on Work

The concepts explored during this work and the lessons learned yielded two major accomplishments. The first included the expansion of the Humatics instrumentation framework to take in additional sources and types of data, the development of new methods for real-time and post-exercises metrics and assessment visualizations, and a series of research efforts focused on a better understanding of analytical performance.

The second major consequence was the production of two serious games that followed a set of development and employment mechanisms that were similar to those we used in our work. One game focused on an airport security scenario in which teams who had access to actual closed-circuit video from a major U.S. airport monitored the video and other data feeds to discover suspicious activities being performed by scripted actors. The second game involved all-source information analysis during which participants analyzed documents, answered questions, and made recommendations regarding a complex geopolitical event while intricate human-system interaction data were collected with a high-frame-rate, near-infrared, eye-tracking system and a custom instrumented instance of the Palantir Technologies data analysis platform. This latter game focused on a detailed user-workflow decomposition and metrics development to characterize individuals' reading behaviors, estimate their cognitive load, and objectively assess their performance at information discovery, factual recall, inference development, and decision making.

Acknowledgments

The authors would like to thank Christian Anderson, Joaquin Avellan, Peggy Chernoff, John Collins, Gary Condon, Jason Hepp, Benjamin Landon, Rebecca Madsen, Joseph Marini, Kyle O'Brien, Travis Riley, Timothy Schreiner, Kenneth Senne, Bryan Tipton, Andrew Wang, and James Won for their contributions to this research. ■

References

1. D.A. Bright, C.E. Hughes, and J. Chalmers, "Illuminating Dark Networks: A Social Network Analysis of an Australian Drug Trafficking Syndicate," *Crime, Law and Social Change*, vol. 57, no. 2, 2012, pp. 151–176.
2. F. Calderoni, "The Structure of Drug Trafficking Mafias: The 'Ndrangheta and Cocaine," *Crime, Law and Social Change*, vol. 58, no. 3, 2012, pp. 321–349.
3. R.M. Bakker, J. Raab, and H.B. Milward, "A Preliminary Theory of Dark Network Resilience," *Journal of Policy Analysis and Management*, vol. 31, no. 1, 2012, pp. 33–62.
4. P.K. Davis and K. Cragin, eds., *Social Science for Counterterrorism: Putting the Pieces Together*. Santa Monica, Calif.: Rand Corporation, 2009.
5. J.N. Shapiro, *The Terrorist's Dilemma: Managing Violent Covert Organizations*. Princeton, N.J.: Princeton University Press, 2013.
6. S. Helfstein and D. Wright, "Covert or Convenient? Evolution of Terror Attack Networks," *Journal of Conflict Resolution*, vol. 55, no. 5, 2011, pp. 785– 813.
7. "Car Bomb," Wikipedia, accessed online 17 July 2018, https://en.wikipedia.org/wiki/Car bomb.
8. A.H. Tapia, N.J. LaLone, and H.-W. Kim, "Run Amok: Group Crowd Participation in Identifying the Bomb and Bomber from the Boston Marathon Bombing," in *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management*, 2014, pp. 265–274.
9. M. Daggett, K. O'Brien, and M. Hurley, "An Information Theoretic Approach for Measuring Data Discovery and Utilization during Analytical and Decision-Making Processes," in A. De Gloria and R. Veltkamp, eds. *Games and Learning Alliance, Lecture Notes in Computer Science*, vol. 9599. Basel, Switzerland: Springer, 2015.
10. J.C. Won, "Influence of Resource Allocation on Teamwork and Performance in an Intelligence, Surveillance, and Reconnaissance (ISR) Red/Blue Exercise within Self-Organizing Teams," PhD dissertation, Tufts University, 2012.
11. J.C. Won, G.R. Condon, B.R. Landon, A.R. Wang, and D.J. Hannon, "Assessing Team Workload and Situational Awareness in an Intelligence, Surveillance, and Reconnaissance (ISR) Simulation Exercise," in *Proceedings of the IEEE First International Multi-disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, 2011, pp. 163–167.
12. R. Porter, A.M. Fraser, and D. Hush, "Wide-Area Motion Imagery," *IEEE Signal Processing Magazine*, vol. 27, no. 5, 2010, pp. 56–65.
13. D.G. Bell, F. Kuehnel, C. Maxwell, R. Kim, K. Kasraie, T. Gaskins, P. Hogan, and J. Coughlan, "NASA World Wind: Opensource GIS for Mission Operations," in *Proceedings of the 2007 IEEE Aerospace Conference*, 2007, pp. 1–9.
14. R.R. Vatsavai, S. Shekhar, T.E. Burk, and S. Lime, "UMN-Mapserver: A High-Performance, Interoperable, and Open Source Web Mapping and Geo-spatial Analysis System," in M. Raubal, H.J Miller, A.U. Frank, and M.F. Goodchild, eds. *Geographic Information Science. Lecture Notes in Computer Science*, vol. 4197. Berlin & Heidelberg, Germany: Springer, 2006, pp. 400–417.
15. M.R. Endsley, "Situation Awareness Misconceptions and Misunderstandings," *Journal of Cognitive Engineering and Decision Making*, vol. 9, no. 1, 2015, pp. 4–32.
16. P. Salmon, N. Stanton, G. Walker, and D. Green, "Situation Awareness Measurement: A Review of Applicability for C4i Environments," *Applied Ergonomics*, vol. 37, no. 2, 2006, pp. 225–238.
17. J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker," in *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications*, 2008, pp. 69–72.
18. D. Olguín-Olguín and A. Pentland, "Sensor-Based Organizational Design and Engineering," *International Journal of Organisational Design and Engineering*, vol. 1, no. 1–2, 2010, pp. 69–97.

19. J.J. Garrett, *The Elements of User Experience: User-Centered Design for the Web and Beyond*. Berkeley, Calif.: Pearson Education, 2010.
20. M. Daggett, K. O'Brien, M. Hurley, and D. Hannon, "Predicting Team Performance Through Human Behavioral Sensing and Quantitative Workflow Instrumentation," in I. Nunes, ed., *Advances in Human Factors and System Interactions. Advances in Intelligent Systems and Computing*, vol. 497. Basel, Switzerland: Springer, 2017, pp. 245–258.
21. E.K. Kao, M.P. Daggett, and M.B. Hurley, "An Information Theoretic Approach for Tracker Performance Evaluation," in *Proceedings of the IEEE 12th International Conference on Computer Vision*, 2009, pp. 1523–1529.
22. X.S. Apedoe, K.V. Mattis, B. Rowden-Quince, and C.D. Schunn, "Examining the Role of Verbal Interaction in Team Success on a Design Challenge," in *Proceedings of the 9th International Conference of the Learning Sciences*, vol. 1, 2010, pp. 596–603.
23. A.J. Strang, S. Horwood, C. Best, G.J. Funke, B.A. Knott, and S.M. Russell, "Examining Temporal Regularity in Categorical Team Communication Using Sample Entropy," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1, 2012, pp. 473–477.
24. L.A. Whitaker, S.L. Fox, and L.J. Peters, "Communication Between Crews: The Effects of Speech Intelligibility on Team Performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 37, no. 9, 1993, pp. 630–634.
25. H.P. Andres, "The Impact of Communication Medium on Virtual Team Group Process," *Information Resources Management Journal*, vol. 19, no. 2, 2006, pp. 1–17.
26. W. Dong, B. Lepri, T. Kim, F. Pianesi, and A.S. Pentland, "Modeling Conversational Dynamics and Performance in a Social Dilemma Task," in *Proceedings of the 5th International Symposium on Communications, Control and Signal Processing*, 2012, pp. 1–4.
27. W. Dong, B. Lepri, A. Cappelletti, A.S. Pentland, F. Pianesi, and M. Zancanaro, "Using the Influence Model to Recognize Functional Roles in Meetings," in *Proceedings of the 9th International Conference on Multimodal Interfaces*, 2007, pp. 271–278.
28. L.C. Freeman, D. Roeder, and R.R. Mulholland, "Centrality in Social Networks: II. Experimental Results," *Social Networks*, vol. 2, no. 2, 1979–1980, pp. 119–141.
29. K. Ara, N. Kanehira, D. Olguín-Olguín, B.N. Waber, T. Kim, A. Mohan, et al., "Sensible Organizations: Changing Our Businesses and Work Styles Through Sensor Data," *Journal of Information Processing*, vol. 16, 2008, pp. 1–12.
30. J.C. Gorman, N.J. Cooke, P.G. Amazeen, and S. Fouse, "Measuring Patterns in Team Interaction Sequences Using a Discrete Recurrence Approach," *Human Factors*, vol. 54, no. 4, 2012, pp. 503–517.

About the Authors

Matthew P. Daggett is a member of the technical staff in the Humanitarian Assistance and Disaster Relief Systems Group. He joined Lincoln Laboratory in 2005, and his research focuses on using operations research methodologies and quantitative human-system instrumentation to design and measure the effectiveness of analytic technologies and processes for complex sociotechnical systems. He has expertise in remote sensing optimization, social network analysis, natural-language processing, data visualization, and the study of team dynamics and decision making. He holds a bachelor's degree in electrical engineering from Virginia Polytechnic Institute and State University.

Daniel J. Hannon is a research psychologist and a member of the technical staff in Lincoln Laboratory's Bioengineering Systems and Technology Group. His work spans interests in psychological health, cognitive science, teamwork, and human factors engineering. In addition to his research career, he is an active clinician, working in emergency psychological care and also teaches in the Mechanical Engineering Department at Tufts University. Prior to joining the Laboratory, he was the director of the Human Factors Engineering Program at Tufts University and a program manager in Aviation Human Factors at the U.S. Department of Transportation, Volpe Center. He holds a bachelor's degree in psychology from Nazareth College and master's and doctoral degrees in experimental psychology from Brown University.

Michael B. Hurley is a member of the technical staff in the Intelligence and Decision Technologies Group at Lincoln Laboratory. Over his career at the Laboratory, he has worked on projects involving real-time servo control systems, multitarget tracking, multisensor data fusion, and probabilistic and information theoretic methods for assessing performance. His current research interest is the use of Bayesian probability theory and information theory to design decision support algorithms. He holds a bachelor's degree in physics from Carnegie-Mellon University and a doctorate in physics from the University of Pennsylvania, where his graduate work was in neutrino physics.

John O. Nwagbaraocha is an assistant group leader of the Embedded and Open Systems Group, where he focuses on service-oriented architectures to address challenging national security problems. In his previous role as a technical staff member, he led teams in the development of reference architectures and prototypes for the Air Force and the Intelligence Community. He joined the Laboratory in 2007 and worked on synthetic aperture radar processing, activity-based analytics for moving target indicator data, and user interface design for serious games. He holds a bachelor's degree in computer engineering from the Rochester Institute of Technology and a master's degree in electrical engineering from Northeastern University.