2012 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
SYSTEMS ENGINEERING AND INTEGRATION (SE) MINI SYMPOSIUM
AUGUST 14-16, MICHIGAN

CROWDSOURCING: A PRIMER AND ITS IMPLICATIONS FOR SYSTEMS ENGINEERING

Richard J. Gerth, Ph.D.
National Automotive Center
TARDEC - NAC
Warren, MI

Alex Burnap
Panos Papalambros, Ph.D.
Design Science
University of Michigan
Ann Arbor, MI

ABSTRACT

Crowdsourcing is an overarching term that denotes a number of ways to use the web as means to enlist a large number of individuals to perform a particular task. The tasks can range from simply providing an opinion, to contributing material, to solving a problem. Because the term crowdsourcing is used to denote a variety of activities in many different contexts, strong opinions have formed in many minds. This paper is an attempt to inform the reader of the complexity that underlies the simple term "crowdsourcing." We then describe the connection between the DARPA Adaptive Vehicle Make program with the potential limitations of crowdsourcing complex tasks using examples from industry. Using these examples, we present a research motivation detailing areas to be improved within current crowdsourcing frameworks. Finally, an agent-based simulation using machine learning techniques is defined, preliminary results are presented, and future research directions are described.

INTRODUCTION

"Crowdsourcing," "Open Innovation," "The Wisdom of Crowds," "Democratized Design," "Grand Challenges" are all phrases related to the term crowdsourcing. Various forms of crowdsourcing have been tested and evaluated by different government agencies and support for its broader use has grown to the point that The White House has encouraged the use of "challenges" as a means of more cost effectively accomplishing the government's mission.

Recent applications of crowdsourcing include Apps for the Army (A4A) championed by General Chiarelli, where soldiers were invited to compete on developing smart phone applications for the Army (Valpolini, et al., 2011). NASA posted 100 difficult challenges through innocentive.com and was very pleased with the results having solved some of the most difficult problems they faced (Davis, 2010). Most recently Local Motors completed under the DARPA Advanced Vehicle Make (AVM) program an eXperimental Crowd derived Combat support Vehicle (XC2V), which has been garnering a lot of attention within and without the Army.

Crowdsourcing to solve problems is not something new per se. One can argue that the Longitude Prize established in 1714 by the British Government to develop a simple and practical method for the precise determination of a ship's longitude was a form of crowdsourcing (Jeppesen and Lakhani, 2010). What is different today is the ability through web-based social media technology to simply, quickly, and inexpensively reach out to a global crowd of potential participants.

To some crowdsourcing is a magic bullet, to others a fad, and to still others something to be researched. This paper provides an introductory description of crowdsourcing with a focus on how it might relate to the design of ground vehicles from a systems engineering standpoint. It also presents some recent research results of applying machine learning techniques to crowdsourced simulations.

CROWDSOURCING ELEMENTS

"Crowdsourcing is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community (a "crowd"), through an open call (wikipedia.org, 2012)."

Crowdsourcing is used to accomplish a variety of tasks, ranging from simple evaluations to full systems design and prototype demonstrations. The manner in which the task is formulated and the crowd is managed varies depending on the task.

The general structure of crowdsourcing contains the following elements:

1. Task Statement: this is a statement of what the crowd is expected to do. This is also called the challenge or problem statement. Sometimes the task is obvious or the task statement is a simple sentence, such as "please rate the value of the product on scale of 1 (worst) to 5 (best)." Other times the problem statement may involve many pages of requirements to be fulfilled. The content of the task statement is very important and can determine the success of the crowdsourcing effort (Chaordix, 2010; Innocentive, 2011). The organization that desires the task to be completed is called the sponsor.
2. Task Submission: this is the output of the crowd requested from the task statement, i.e., the work product. Depending on the task, the work product may or may not be evaluated relative to the task statement and / or relative to other submissions.
3. Open Period: this is the time from which the task statement is available to the time the task is to be completed, or the product(s) from the task are to be delivered. The open period can vary from a few days (idea generation, algorithm development), to several months (difficult problem solving, grand challenges), to practically infinite (product evaluation). Sometimes the task statement is made available in advance and the open period begins when the sponsor begins accepting task submissions. Typically one wants the open period to be long enough for the crowd to complete the task, yet significantly shorter than the time it would take to complete the task using traditional processes.
4. Reward: Some task statements may have associated rewards ranging from monetary to recognition rewards. Others have no reward other than the intrinsic reward to the individual for having contributed. The type and magnitude of the reward relative to the task statement will obviously affect the motivation and response of the crowd.
5. Crowd: The crowd is any group of people. It can be a restricted set of people, such as only individuals with TOP SECRET clearance, or it can be the whole world. Innocentive.com, a company that conducts challenges on the behalf of its clients, including NASA, restricts its crowd to any individual that has signed the legal documents signing over their intellectual property (IP) rights prior to viewing the Task Statement. Their crowd is reported to consist of over 250,000 individuals from nearly 200 countries (innocentive.com, 2011). The size of the crowd is important because the larger the crowd the greater the chances that someone is willing and capable of completing the task statement. The individuals in the crowd are called members or participants. An important note is that participants are self-selective, i.e., the participant chooses whether or not he or she will participate in the task.
6. Crowd Management: Crowd management refers to tools and processes for crowd recruitment and retention, encouraging involvement, ensuring fair and secure participation, enabling communication amongst the participants, and facilitating communication between the crowd and the crowd manager or task sponsor (legal, IP, challenge clarification, site technical support, etc.). This is a very important aspect of crowdsourcing as it is one of the few explicitly controllable variables. Since the crowd is self-selecting, it is possible that no one in the crowd chooses to take on the task or the task is too difficult to accomplish correctly. This may be a sign that the crowd is too small, not properly managed, or does not collectively contain enough expertise. Regardless, the properties of a crowd cultivated through any crowdsourcing engagement represents an asset that should be nurtured for future efforts.

TYPES OF CROWDSOURCING TASKS

Crowdsourcing is a term used to denote a means by accomplishing a variety of tasks. These tasks are so varied that a short description of some of the major task types developed to date is warranted.

Grand Challenges

Grand Challenges or simply Challenges have been popularized by the XPrize Foundation, which has run several grand challenges ranging from developing a commercial space plane (ANSARI XPrize) to developing a 100 mpg automobile (Progressive XPrize). The DARPA Grand Challenge on developing an autonomous ground vehicle was another widely publicized program. Grand Challenges are characterized by having very difficult problem statements requiring a demonstration prototype that is evaluated based on performance criteria, a long open period, and large rewards (generally \$0.5 million to \$5 million). The challenges are answered by teams with significant technical capability and, often, significant financial backing.

Grand Challenges are akin to standard competitive prototyping acquisition strategies, except only the "winner" is paid. Thus, they are often considered a "cost effective" means of developing functional prototypes. However, since only the winner(s) receive prizes, traditional corporations generally will not invest the significant R&D funding required to develop the full prototype. Thus, most challenge participants are organizations that have significant financial backing to create the prototypes without significant business risk if they do not win the challenge prize. Examples of such organizations are university led teams (with or without traditional industry backing) and start-up companies (who are submitting their venture funded prototypes). Though Grand Challenges can be a cost effective means to develop and establish technology demonstrations of complex systems from external sources, they can be difficult to manage if requirements or constraints change during the course of the challenge. They are also generally limited in terms of the sponsoring organization's technical involvement and learning. The sponsoring organization usually cannot be technically involved during the challenge phase with any of the participants for fear of favoritism and introducing bias. And while IP agreements may state the sponsoring organization owns the IP, the challenge will not have established the internal capability to work with the IP, as the organization could not be involved with its development.

Grand Challenges are not considered by these authors to be the best suited form of crowdsourcing for improving the ground vehicle system design process. Grand Challenges outsource the entire process. They are not a change or improvement of the current process. Accordingly, this paper focuses on tasks that can be crowdsourced to individuals with modest or minimal financial investment. It is of interest to understand how the systems engineering process can be decomposed into tasks that can be completed by individuals.

Evaluation

Product reviews are a type of crowdsourced evaluation. An example is amazon.com, where the crowd is asked to rate the value of a product on a 5 star scale. Similar types of systems have been used to evaluate everything from marketing campaigns to proposed vehicle designs. The evaluation methodology can range from simple to complex involving multiple levels of evaluation by different groups, including internal expert panels.

The evaluation task is one of the simplest, and yet highest valued added tasks one can ask the crowd to do. For example, engineers are constantly asked to evaluate between alternatives on a variety of factors. Surveys often take the form of a series of evaluation tasks. Thus, the author's selected to model the evaluation task for research purposes (see **Error! Reference source not found.** below).

Content Repositories

YouTube.com and eBay.com are examples of websites that have built a business around crowd submitted content. YouTube sells advertising, and eBay sells the content. Some sites specialize in specific content such as stock photos for journalists and marketing professionals, where the quality of the content is closely monitored and the web interface is streamlined for the associated set of customers (Howe, 2006).

The above examples are simply content repositories: the content is developed by the crowd and simply uploaded to the repository. However, sites such as wikepedia.org use the crowd to collaboratively create content through self-organizing teams. This same concept is used in some cases of product development as well, with perhaps the largest impact being the open source software movement. Repositories such as github.com provide processes and tools such that individuals, sometimes in the tens of thousands, can collaborate on a single codebase.

An individual begins a particular new content area, and other interested parties are allowed through a structured process to contribute in various ways. Some will write a piece of code or develop a new algorithm, others will debug the code, and still others will improve on the basic code by "forking" the code to their own project with the possibility to merge it back in with the original project. This example shows that these tasks are actually composed of subtasks that can be performed by other participants than the individual who originally started the new content. Many of the greatest software successes spanning the entire programming stack have come from this crowdsourced model. Examples across this spectrum include compilers such as g++/gcc; operating systems such as the Linux kernel; programming languages such as C, Ruby, and Python; and server/client architectures such as Javascript and HTML/CSS.

Outsourcing

Various websites have been developed to outsource well defined units of work to the lowest bidder, such as writing a specific piece of software code. Howe (2006) describes an example where Java coders were hired to code a script describing simple repair flows for airline maintenance. Going from dedicated companies to crowdsourcing the price of the task dropped from $2,000 per script to $5 per script.

Broadcast Search

Broadcast Search is the other end of the problem solving spectrum from Grand Challenges. In Broadcast Search the problem is sufficiently decomposed and defined that individuals rather than teams are able to solve the problems.

A variety of organizations, such as innocentive.com and chaordix.com, use the web to harness the knowledge of the crowd to solve difficult problems through challenges. While the crowd contributes potentially many solutions, generally only one of the solutions is selected as having solved the problem.

Broadcast search companies have cultivated a crowd of problem solvers who have registered to see the problems posted on the company's website. The companies have developed tools and processes for crowd recruitment and management; problem definition and classification; reward definition, collection and distribution; solution submission and evaluation; and IP management and transfer.

A 2010 study by Jeppesen and Lakhani of 166 problems in a broadcast search by Innocentive.com between 2001 and

2004 found that 30% of the problems were solved, with a third having more than 1 solution. Approximately 16% of their 80,000 member crowd read the problem statements. 7.8% of those who read the problem statements submitted solutions, and only 6% of the submitters (or 0.07% of the total crowd) were winners Clearly one needs a large crowd to solve difficult problems.

Since that time, Innocentive.com (2011) reports they have

- Grown from a crowd of 80,000 to 250,000
- Posted a total of more than 1,300 problems over a 10 year period (versus 166 over a 4 year period)
- Increased their problem solution success rate from 30% to 50%. This may be a direct result of the increase in their crowd size.

Another finding from the Jeppeson and Lakhani study was that women and non-experts in the field submitted statistically significantly more winning solutions relative to the proportion of women and non-experts who submitted solutions. This is seen as further evidence that diversity trumps expertise in broadcast search as discovered by Page (2007).

The above are only some of the type of tasks that have used crowdsourcing. There are other tasks and models that also arguably use crowdsourcing, such as prediction markets (reference), voting (reference), and certain types of games (reference). However, the authors selected those tasks and aspects of crowdsourcing deemed most relevant to the design of complex ground vehicle systems. Ultimately the breadth and depth of possible crowdsourcing tasks is limited only be the limits of human imagination and crowd capability.

THE ADVANTAGES AND CONCERNS OF CROWDSOURCING

Proponents of crowdsourcing cite several advantages of crowdsourcing.

The quality of the task product is often higher than non-crowd based processes. This is based on the "wisdom of crowds," which basically states that under the assumption that there do not exist perfect experts for a complicated problem, the crowd is on average more likely to correctly understand and solve a task than any given individual. This is applicable to evaluation and problem solving tasks, although the mechanisms by which this is true are different. The aggregate crowd evaluation is more likely to be "correct" than any randomly selected individual based on statistical arguments (Page, 2007). By crowdsourcing the problem solving task, individuals with expertise in nontraditional subject domains can often provide an innovative approach to solving a problem (Jeppesen and Lakhani, 2010)

The speed at which the task is performed is often faster than non-crowd based processes (Howe, 2006). One can limit the time in which a task should be performed and assume someone in the crowd is likely to be able to perform it in that time. Perhaps more useful from a system engineering standpoint, one is leveraging the fact that many people can work in parallel, thus shorting the task completion times.

The cost of completing the task is less, either because one is sourcing to the lowest bidder, or one only has to pay the best performer a prearranged fee, assuming they complete the task statement satisfactorily. Some tasks require no remuneration at all (Howe, 2006).

Crowdsourcing Concerns

In discussing crowdsourcing as a potential method for solving problems or designing vehicles, people have raised various concerns, some of which are addressed here.

If the crowd is evaluating the submissions, I will submit last because I don't want anyone to steal my ideas

This is an understandable concern that turns out to be partially unfounded and partially manageable. First, it is possible to structure the crowd management system to clearly identify who has submitted which ideas first. Second, the crowd is self-policing. There will be individuals who will quickly report offenders that violate the established practices of the crowd to the crown manager (Rogers, 2010).

Local Motors' experience (Rogers, 2010) has been that early submitters often fared better than late submitters, because they could incorporate the input of the crowd ("there's critic in every crowd") and submit a second, better proposal prior to the end of the open period. This also led in some instances to the formation of self-selective teams, as individuals with different aspects contributed to a common design.

Please note that the above system implicitly allows the crowd to communicate and link the submission to the particular participant. This is one type of crowd management, which may or may not be suitable for all task types. But, it does point to the need to protect the process from participants who would "game the system."

You need to have experts to solve the problem – and we already know all the experts.

Crowdsourcing research (Jeppesen and Lakhani, 2010) shows this to be false. In 166 broadcast searches conducted on innocentive.com, non-experts accounted for over 50% of the winning solutions. This is explained by two effects. First, people who come from other areas have other perspectives and experiences and can see the problem in new ways and have access to solutions that the experts do not.

Second, broadcast search provides access to marginalized groups, such as women, who may have been excluded from the expert community for social reasons. Ultimately broadcast search problems are not solved by novice, but rather by highly knowledgeable individuals in adjacent disciplines that are otherwise not exposed to the particular problem domain.

I don't want everyone to know my problem

This can be guarded against in a number of ways. The problem sponsor can be anonymous, and it is often possible to craft the task / problem statement such that the exact application domain is not identifiable. Another alternative is to limit the crowd to those individuals who have the clearance to know the specific problem domain and sponsor.

You have to guard against saboteurs

There is a valid concern that someone will submit a solution that will fail or have some defect or malicious functionality built into the solution. This can be mitigated in a number of ways. First, one should examine what the potential risks are. For example, if the risk lies in the design if it should be adopted, then the probability that the winning design is also the one that contains malicious content may be relatively low. It also allows one to focus one's risk mitigation strategies on only the winning design. Another possibility is to allow the crowd to evaluate solutions, thus giving many individuals the opportunity to find defective or malicious solutions.

IP is going to be headache. I will never get it past the lawyers

This is something that can and has been addressed in several ways as evidenced by commercial models as well as the Apps for the Army program. The specific process that would be used for complex vehicle design would have to investigate IP issues, and that may or may not have an impact on the process developed.

Everyone is a contributor and you get a mess – too much work to sift through all the "stupid" contributions

First, not everyone contributes. Depending on the crowd and the task structure, only a subset contributes. Other members of crowd may be evaluating solutions, policing the crowd, solution critics, solution improvers, and so on. There are various roles required by the crowd depending on the tasks requested and the particular crowdsourcing architecture. The crowd may be called upon to evaluate the contributions, if the number of submitted task products is large. The crowd is generally very adept at sorting through and identifying the truly infeasible or unworthy solutions.

It is not manageable

Crowds are manageable with the proper processes and tools. However, it is not clear for complex vehicle development what the appropriate processes and tools would be. This is an area that requires further investigation and is the main impetus to the research described herein.

VEHICLE SYSTEM ENGINEERING DESIGN

There exist many previous works on system decomposition and task assignment from a systems engineering standpoint. The commonality of many of these works is a hierarchical flow of information from the largest complex level to more simple levels of task complexity.

One company that has used crowdsourcing to design a ground vehicle is Local Motors Inc. in Chandler, Arizona. Their approach is to develop a market and emotional ownership to the vehicle by having a crowd of potential users design the vehicle they would like to own and evaluate the designs of others within the crowd.

Table 1. Local Motors Model of using Crowd Sourcing for Complex Vehicle Design.

System Level	Architecture	Interface Definition
Vehicle	Crowdsourced	Internally Developed
Major Subsystems	Crowd or Outsourced	Internally Developed
Subsystems	Crowd or Outsourced	Internally Developed
Components	Crowd or Outsourced	Internal Verification

The basic form of the design process is shown in **Error! Reference source not found.**. It is a coordinated series of broadcast search coupled with crowd evaluations and combined with internal design processes conducted by expert staff. The starting point (not shown in Figure 1) is the vehicle requirements. These can be as broadly or narrowly defined as desired, such as a street legal light duty vehicle for passenger use. These requirements then form the basis for the first crowdsourced task statement, namely to design the basic vehicle styling and packaging. Arguably, these two elements represent the vehicle architecture, since once the external styling and general internal packaging is known, many of the requirements for the other subsystems, such as the instruments, powertrain, and chassis / suspension can be derived.

Deriving the requirements for the subsystems and designing the interfaces between the subsystems is done by internal company experts, depicted as the arrows in Figure 1. This is how the organization can retain control over the development process and ensures compliance with laws, reliability, profitability, and manufacturability. This is also the point at which decisions are made as to which sub-systems will be crowdsourced or simply outsourced by purchasing from vendors. For example, Local Motors decided for their Rally Fighter to purchase the BMW L6 diesel engine and mate it to a ZF transmission (Anderson, 2006).

If the internal company experts choose to crowdsource a particular subsystem, the requirements for the subsystem are defined (interfaces, packaging constraints, etc.) and stated as a challenge / task statement. The crowd then designs and evaluates the architecture of the seats: shape, styling, materials, structure, etc. The company then takes the designs, determines manufacturability, and if needed, identifies vendors to manufacture and supply the crowd designed subsystems.

In this manner, the company sequentially utilizes crowdsourcing to help design a vehicle that will appeal to a particular customer base.

It is important to note that design evaluation by the crowd is an important aspect of their process. This has advantages and disadvantages over selecting a review board of "specialists" to evaluate the suitability of designs to meeting the task statement. Advantages are:

1. Designs criticized by the crowd in the early stages enable the submitter to resubmit an improved design that addresses the crowd's inputs.
2. The crowd can evaluate many more designs more quickly than a dedicated review board.
3. The selection method is transparent.
4. The crowd develops ownership in the winning design, which may later boost adoption of the product.

Disadvantages are:

1. Evaluations may be unique to that crowd. A different crowd could develop different evaluations.
2. The crowd may not be knowledgeable or sufficiently diverse to fully evaluate the solutions, depending on the complexity of the evaluation task and the diversity of solutions submitted.
3. Crowd evaluation may not be desirable or possible, if there are other factors that must be considered that could not be revealed to the crowd. If they cannot be revealed, then the crowd will not know all the critiera for task quality and will not be able to perform the task correctly or well.

Relevance to DARPA AVM Program

DARPA initiated the Adaptive Vehicle Make (AVM) program in 2010 built upon the tenets of open source democratized design and geographically distributed networked manufacturing (Eremenko, 2010). As part of the pilot portion of the program, DARPA contracted Local Motors Inc. to develop the XC2V (eXperimentally Crowd designed Combat Vehicle) using their crowdsourcing framework. Level 1 of Figure 1 was crowdsourced, and the remaining steps were completed by in-house staff. 163 designs were submitted in 36 days. The winning design was selected, and the manufactured vehicle was delivered 138 days after the design challenge was launched (beginning of Open Period).

The DARPA AVM includes 3 additional complex challenges: powertrain challenge, structure challenge, and entire vehicle challenge (Eremenko, 2010). The application domain is an amphibious marine vehicle, adding complexity over a standard ground vehicle. The design framework and supporting tools is currently being developed by a consortium of over 20 universities and companies, including Ricardo, Penn State ARL, Carnegie Mellon, Georgia Tech, MIT, Vanderbilt, and General Electric. The general layout of the system is to be a collaborative web environment (VehicleForge.mil) with a dedicated CAD component / modelica model library, design verification tools, and manufacturing process feasibility checks. The challenges will be run more like grand challenges, in that DARPA will not manage the subsystem interface definitions nor run a series of challenges for each of the subsystems. The individual teams are expected to complete all aspects of the design on their own.

RESEARCH MOTIVATION

Assuming that the interfaces between major subsystems can be sufficiently defined on design architectures, then the systems engineering framework for crowdsourcing shown in Figure 1 could be useful for the DARPA AVM project. This would include broadcast search of designs and evaluation of designs like the XC2V project, but additionally allow the possibility for subsystem tasks to be crowdsourced effectively as well. In addition to decomposition based system engineering with crowdsourcing, the current proposal for the DARPA AVM crowdsourcing framework can benefit from research derived implementations of tools to manage the multitude of relationships between members of the crowd and tasks. The ways such research could be beneficial may be better understood by looking at what is problematic within actual testing grounds, namely crowdsourcing companies within industry.

Over 500 relatively nascent companies exist whose products are crowdsourcing for a particular market niche (Directory of Sites, 2012). Examples include 99designs.com for design of graphic design, threadless.com for t-shirts, and quirky.com for simple product design. In addition, over 50 companies exist whose product is to act as a crowdsourcing platform for other companies to build off of. Most famous is Amazon's Mechanical Turk, but many others exist such as crowdflower.com.

What may be problematic with many of these companies is they do not have sophisticated mechanisms to increase the effectiveness of their crowds. Many of these companies have an all-to-all relationship between the crowd and the tasks, i.e. every member in the crowd is able to see every task. While filters, where individuals are able to filter by categories like "biomedical" may exist, no advanced mechanisms exist to ensure the optimal subset of the crowd is seeing the optimal subset of tasks with the optimal task delivery process.

Without more advanced crowdsourcing mechanisms, there is a fundamental limit to the complexity of a problem solving task. One of the best examples is when Fiat performed the first fully crowdsourced vehicle, the Fiat Mio. The crowd submitted over 11,000 designs, to which Fiat had to respond by turning its engineering and design teams outwards and into design evaluators. The level of aggregate expertise needed to give correct evaluations on complex facets such as safety considerations was not imbued within the crowd, thus necessitating tapping into the expertise within Fiat itself. This is similar to the case of Local Motors being able to crowdsource the shell for the XC2V, but needing to outsource the components requiring expertise knowledge such as the engine and transmission.

The question then arises of how could one increase the complex problem solving ability of a crowd? One possibility is to use expert. But a crowd of experts, where an expert is defined in the orthodox manner, i.e., an individual who traditionally solves problems within the application domain, is simply an outsourced engineering house within the industry. We still want to reap the advantages of crowdsourcing, such as the potential for innovation stemming from the diversity of crowd input. Another important distinction from outsourcing is that we still want self-selection of tasks by the individual within a crowd. The argument is individuals perform better on self selected tasks than assigned tasks. ti

RESEARCH METHODOLOGY

It is hypothesized there exists a relationship between the characteristics of a task, the expertise of an individual, and the quality of a task submission. The hypothesis will be tested by creating a multi-agent based simulation for data generation, and machine learning techniques for data analysis. The justification for an agent-based simulation is that it is much more reasonable to model a crowdsourcing interaction between an individual, or agent, and a given task, as opposed to trying to encode macroscopic system level phenomenon (Jennings, 2000).

Modeling human behavior is difficult. Thus it was decided to model a relatively simple task: evaluation. Evaluation tasks are intrinsically easier to model compared with other tasks, such as problem solving. Also many crowdsourcing systems used today administer evaluation tasks, providing potential access to large data sets.

Machine learning can be used to analyze these data sets. Machine learning methods offer a "black-box" method for determining models that may not be explicitly defined. Accordingly, one does not have to model the human cognition of an agent making an evaluation of a design along all relevant dimensions. Instead, one may assume certain known psychological phenomena account for some variability in the evaluation, while the rest of the variability may be "machine learned", i.e., deduced from statistical patterns. For example: suppose an agent has to choose the best design between 26 designs labeled Design A to Design Z. A particular behavioral model might predict given the experiences, demographics, and expertise of the agent, it should choose Design T. However, the data might show it chooses Design T only 40% of the time, meaning the behavioral model does not account for 60% of the variability. Given enough data on similar agents, machine leaning can create models that can explain some of the previously unexplained 60% variability. Much of the time, these learned models are non-intuitive, because they uncover hidden relationships or the underlying models may simply contain too many variables (high dimensionality) (Vapnik, 1998).

The preliminary simulation environment uses a simple agent model executing a ranking evaluation task. Rather than accounting for complex behavioral models of the evaluation task, the focus is on understanding if under certain assumptions the "wisdom of the crowd" is able to outperform a smaller group of "experts." Accordingly, the simulation is setup as a crowd of agents conducting pair-wise evaluations on a set of designs. The simulation will show that a machine was able to learn the evaluation processes of crowd members, and use that information to predict the best vehicle design given a set of mission requirements.

RESEARCH MODEL

Definitions

We define a design $\boldsymbol{x_i} \in D$, where $\boldsymbol{x_i} = \{x_1, \dots, x_R\}$ for N designs, and a single agent $\boldsymbol{w_a} \in A$ where each $\boldsymbol{w_a} = \{w_1, \dots, w_R\}$ for M agents. In our case $\boldsymbol{x_d}$ is a set of basis functions spanning the space D corresponding to various mission requirements of a ground vehicle (ability to accelerate out of harm's way, blast survivability, vehicle cost, etc.). Next, the elements of $\boldsymbol{w_a}$ represent the evaluation heuristic of that agent. This corresponds to the weighting a particular agent puts on each of the mission requirements.

For example, a powertrain engineer may have a lot of expertise on whether the vehicle design would be able to accelerate out of harm's way, but likely little expertise on blast survivability. For that reason, the agent may have a good chance of knowing the correct weight for one mission requirement, while bring completely incorrect on another. However, having no expertise on a dimension is equivalent to treating it as a latent variable; thus, an agent will not consider that dimension in an evaluation. This is understandable as oftentimes human decision making is based on only a few aspects of a problem (Payne, 1976). These two cases are termed evaluation expertise and existence expertise, where evaluation expertise represents the deviation from the optimal evaluation, and existence expertise represents the proportion of evaluated dimensions to latent dimensions.

Ranking Evaluation Process

It is assumed there exists an optimal tradeoff between mission requirements, denoted $\boldsymbol{w}_*$. The following equation then describes the true utility of a vehicle design given the mission requirements.

$$y_{d*} = \boldsymbol{w}_*{}^T \boldsymbol{x_d}$$

This is can be viewed as a hyperplane, which acts as a ranking mechanism. Since we want to understand how close our crowd gets to the optimal hyperplane, the machine must learn the $\boldsymbol{w_a}$ of each member. But similar to many crowdsourcing systems, the crowd is initially unknown and only the items to be evaluated, the vehicle designs in this case, are known.

To learn the evaluation heuristics of the crowd, one can use a technique from machine learning literature called the ranking support vector machine (Joachims, 2002). This technique learns the $\boldsymbol{w_a}$ of an agent by watching a series of pair-wise evaluations between vehicle designs. This technique was selected because it provides good predictive performance versus other methods particularly on smaller datasets (Vapnik, 1998), which can occur in crowd sourced evaluations. Often, an individual performing an evaluation task stops the task after just a small number of evaluations due to fatigue, resulting in a sparse dataset.

The ranking support vector machine is posed as a unconstrained optimization problem (Chapelle and Keerthi, 2010):

$$\min\ |\boldsymbol{w}_a{}^2| + \widetilde{\boldsymbol{C}} \sum_{i,j=1}^{i,j=N} \max\,[0, t_{i,j}\left(\boldsymbol{w}_a{}^T \boldsymbol{x}_i - \boldsymbol{w}_a{}^T \boldsymbol{x}_j\right) - 1 + \xi_{i,j}]$$

where $\widetilde{\boldsymbol{C}}$ is a matrix controlling the regularization between training data and model complexity, $t_{i,j} = \{-1, 1\}$ represents whether the agent $\boldsymbol{w}_a$ evaluated $\boldsymbol{x}_i$ higher than $\boldsymbol{x}_j$, and $\xi_{i,j}$ represents how well our agent's evaluation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_i$ is trusted.

To simulate the pair-wise evaluations that each agent makes, a temporary $\widetilde{\boldsymbol{w}_a}$ was created by providing a gaussion blur to $\boldsymbol{w}_*$. Two forms of expertise arise from this. The first is the degree to which $\widetilde{\boldsymbol{w}_a}$ is blurred, which represents the aforementioned evaluation expertise for that mission requirement. The second is the "zeroing out" of particular elements of $\widetilde{\boldsymbol{w}_a}$, which essentially turns the dimension into a latent variable. The ratio of non-zero dimensions to latent variable is the existence expertise. The $\widetilde{\boldsymbol{w}_a}$ is then used to create a set of pair-wise evaluations between all vehicle designs $\boldsymbol{x}_i$

Performance Metrics.

The goal is to measure the difference between the simulation results and the optimal tradeoff of mission requirements. If one assumes the design space is convex, then by using the primal linear formulation of the ranking support vector machine, tone can use the Euclidean distance metric to describe the ranking error:

$$Ranking\ Error = \sum_{i=1}^{N} (Rank_i - Optimal\ Rank_i)^2$$

where ***Rank*** is an ordered set of $\boldsymbol{x}_i$ for all N vehicle designs, and ***Optimal Rank*** is an ordered set of $\boldsymbol{x}_i$ derived from the optimal tradeoff between mission requirements $\boldsymbol{w}_*$.

Defining the error metric thusly allows one to use a result from crowd diversity research, namely the Diversity Prediction Theorem (Page, 2007; Bommarito, et al., 2011). This theorem states the diversity of input from a crowd results in the "wisdom of the crowd", expressed as.

$$Avg.\ Individual\ Error = Crowd\ Error + Crowd\ Diversity$$

where the average individual error is given by:

$$\frac{1}{M} \sum_{a=1}^{M} (\boldsymbol{Rank}_a - \boldsymbol{Optimal\ Rank})^2$$

the crowd error is given by:

$$\left(\frac{1}{M} \sum_{a=1}^{M} \boldsymbol{Rank}_a - \boldsymbol{Optimal\ Rank} \right)^2$$

and the crowd diversity is given by:

$$\frac{1}{M} \sum_{i=1}^{M} \left(\boldsymbol{Rank}_a - \frac{1}{M} \sum_{a=1}^{M} \boldsymbol{Rank}_a \right)^2$$

PRELIMINARY RESULTS

The ratio of crowd error to average individual error, analogous to the performance of a crowd versus performance of any of its individual members (see Figure 1). For this metric, the *lower* the value the better the relative performance of the crowd.

Increasing crowd size and decreasing existence expertise were combined. As previously defined, existence expertise is a measure to which an agent considers all mission requirements, not just the ones it knows best. The same plot shows the difference due to evaluation expertise. Essentially the evaluation expertise is the ability of an agent to correctly weigh a vehicle design's mission requirements against each other in an effort to evaluate the optimal level of tradeoffs.

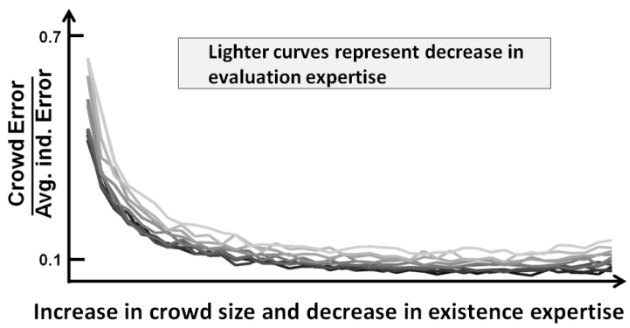

Figure 1. Plot of the Ratio of Crowd Error to Average Individual Error as a Function of Crowd Size and Decreasing Existence Expertise.

As stated above, the *lower* the ratio of crowd error to average individual error, the more apparent the effect of the "wisdom of the crowd". One can see under the simulation assumptions, a large crowd of non-experts performs at the same level as a smaller group of experts. It is most apparent when taking the limits of size. A crowd of one expert makes an error of about 0.7, while a large crowd has an error around 0.1. Additionally, evaluation error decreases the effectiveness of the crowd, i.e. incorrect evaluations favor neither the crowd nor the small group of experts.

FURTHER RESEARCH

The initial simulation model presented here includes a large number of assumptions. For example, it assumes convexity of the mission requirement space, which may not be a good assumption. For a given mission, a fast and lightly armored ground vehicle may be just as useful as a slower heavily armored vehicle, but a cross between the two may be disastrous. It is planned to circumvent this assumption through nonlinear kernel based models. This will also enable the machine learning of the covariance relationship between mission requirements, as well as those between different agent evaluation heuristics.

Future research will also imbue the agents with individual characteristics, such as fatigue and bias. These models will be partially fit to data from both Local Motors Inc. and DARPA.

A much longer range plan would be to create a system that may be trained "online", as in use models that adapt in real-time to crowd input for a given crowdsourcing system. Finally, as with much of recent crowdsourcing research, crowd communication is an incredibly powerful mechanism to increase performance. The simple model presented herein assumes independence of crowd members, which may be understandable for certain crowdsourcing applications, but for many others does not reflect reality. Although it is difficult to compare a crowd of independent agents with an outsourced team from the results presented, future models may have greater refinement by building on recent results within the social dynamics (Castellano, et al., 2009) the collective intelligence community (Mavrodiev, et al., 2012).

CONCLUSIONS

Crowdsourcing is an overarching term that really encompasses many different tasks and is far more complex that the simple term might indicate. The effectiveness of crowd based techniques to aid in the design of a complex ground vehicle is not well understood. However, research and experience both in the commercial and military domain have shown indications that crowdsourcing can provide value to the warfighter.

It is recommended that a serious effort be made to better understand the appropriate use of crowdsourcing techniques within systems engineering for ground vehicle design, and to develop a set of processes and tools that can utilize these techniques for maximum cost, speed, and performance benefit. Though the implementation of the DARPA Adaptive Vehicle Make platform may help further catalyze research in this area, cases from industry, such as Local Motors and Fiat, have already illustrated initial limitations of crowdsourcing complex tasks. Further study into the various relationships that exist between crowd members, tasks, and system level objectives is needed. It is hoped that the expanded development of the agent-based simulation system presented here may lead to future crowdsourcing frameworks capable of adaptively improving themselves for complex tasks.

REFERENCES

[1] Anderson, C (2011) Atoms are the New Bits, Wired Magazine, February 2010, http://www.wired.com/magazine/2010/01/ff_newrevolution/all/1 Last accessed 18JUN2012.

[2] Bommarito, M., Franzese, R., Gong, A., Page, S.E. (2011) "Technical Report: Crowdsourcing Design and Evaluation Analysis of DARPA's XC2V Challenge,"

[3] Castellano, C., Fortunato, S. and Loreto, V. (2009) "Statistical physics of social dynamics," Reviews of Modern Physics, vol. 81, no. 2, pp. 591-646, May 2009.

[4] Chaordix (2010) 8 Principles of Successful Crowdsourcing, Chaordix Whitepaper.

[5] Chapelle, O. and Keerthi, S. S. (2010) “Efficient algorithms for ranking with SVMs,” Information retrieval, vol. 13, no. 3, pp. 201–215.

[6] Davis, Jeffrey; Director, Space Life Sciences, NASA, phone interview, 15NOV2010.

[7] Directory of Sites." *Directory of Sites*. Web. 18 June 2012. <http://www.crowdsourcing.org/directory>.

[8] Eremenko, P (2010) DARPA Advanced Vehicle Make, Proposers Day Briefing, Oct 2010.

[9] Howe, Jeff; The Rise of Crowdsourcing, Wired Magazine, June 2006, http://www.wired.com/wired/archive/14.06/crowds_pr.html. Last accessed 19AUG2011.

[10] http://en.wikipedia.org/wiki/Crowdsourcing. Last accessed 05AUG2011.

[11] http://en.wikipedia.org/wiki/Longitude_prize. Last accessed 19AUG2011.

[12] https://www.innocentive.com/about-innocentive/facts-stats. Last accessed 18AUG2011

[13] Innocentive (2011) Harnessing The Power of Challenges to Accelerate Innovation Outcomes and Improve Business Performance, Innocentive White Paper

[14] Jennings, N. R. (2000). On agent-based software engineering, *117*(September 1999), 277-296.

[15] Jeppesen, L.B., and Lakhani, K.R (2010) Marginality and Problem-Solving Effectiveness in Broadcast Search, Organization Science, Articles in Advance, pp. 1–18.

[16] Joachims, T. (2002) “Optimizing search engines using clickthrough data,” in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 133–142.

[17] Mavrodiev, P., Tessone, C. J. and Schweitzer, F. (2012) “Effects of Social Influence on the Wisdom of Crowds,” in Proceedings of the 2012 Collective Intelligence Conference, pp. 1-7.

[18] Page, Scott E. (2007) The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press, Princeton, NJ.

[19] Payne, John W. (1976) Task complexity and contingent processing in decision making: An information search and protocol analysis, Organizational Behavior and Human Performance, Volume 16, Issue 2.

[20] Rogers, John B., CEO, Local-Motors, phone interview, 24JUN2010.

[21] Valpolini, P., Biass, E.H., Keggler, J. A (2011) New Breed, Compendium New Soldier Equipment, Armada International, 35(2):1-3.

[22] Vapnik, V. (1998) Statistical Learning Theory. John Wiley and Sons, New York

Disclaimer: Reference herein to any specific commercial company, product. process, or service by trade name, trademark manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the Department of the Army (DoA) The opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or the DoA, and shall not be used for advertising or product endorsement purposes.