BUILDING COMPUTATIONAL SOCIAL SCIENCE MODELS FROM CROWD INSIGHT

Alicia Ruvinsky, Alden Roberts

Lockheed Martin Advanced Technology Labs
alicia.i.ruvinsky@lmco.com, alden.b.roberts@lmco.com

ABSTRACT

Accurate models of social processes are invaluable tools for understanding, exploring, explaining, and predicting phenomena that impact decisions about policies and courses of action. However, computational social science (CSS) models today take months to create and can only be generated by a fraction of modeling experts. Building models in this way is an "Ivory Tower" approach where highly specialized modelers spend an impractical amount of time building a model, creating "heavy" models that are slow to develop, hard to maintain, and may not reflect rapidly changing factors "on the ground." CSS models must be developed more easily and in a more timely fashion to support the dynamicity of the social space.

This work describes a vision and proof of concept in which observations of average people are harnessed by a system capable of generating adaptive models that leverage individual and collective insights and experiences. Rather than relying on few experts with similar experiences, our approach relies on a larger crowd of individuals with diverse knowledge. This approach augments the Ivory Tower with the wisdom of crowds to enable speedier development and seamless update and maintenance of models.

Index Terms— computational social science, crowd sourcing, adaptive modeling

1. INTRODUCTION

Computational social science modeling of an operational environment is an invaluable tool for assessing significantly relevant, impactful and critical aspects of military endeavors into foreign regions. However, the process to create these models is slow and expensive. The delay between the commissioning of a model and its completion means that models are an educated guess as to what will be needed in an unknown future. Modelers (social scientists) are a rare-breed, already in high demand, as their models prove increasingly valuable and integral in the day-to-day activities of our soldiers. Demand for computational social science models will only continue to rise.

We propose *Crowd Sourced Models* (CSM), an innovation that at its core seeks to spur a disruptive paradigm shift in the way that models are generated and used. CSM aims to transform models into a commodity that are fast and cheap to build, can be built quickly to new or evolving situations, and can be rapidly and continuously adapted to changing factors on the ground. CSMs can be generated in an on-demand fashion to solve immediate and pressing problems, and if current events render the model obsolete, it may be replaced easily and quickly to provide persistent support to decision and policy makers.

Our approach to solving this ambitious goal is to decouple the two primary roles of the modeler, that of (1) creation of social science theory (and related artifacts) upon which models are built and (2) awareness of the current social specialization such as geographic familiarity or topical knowledge. The former will always be the focus and area of interest of the social scientist, however the latter is a problem better suited for a crowd made up of individuals either specializing in the topical knowledge, familiar with the geography in question, living in the area of interest, having recently traveled or been deployed to the area of interest, etc. The crowd's power comes from its diversity of opinion, potential expertise, experience and adjacency to the problem or area being modeled. The very nature of the crowd makes it unlikely to fall victim to "group think" and more likely to explore a wider swath of possibility discovering new maxima that might have been overlooked by the academic or operational community [1]. By separating modeling we are allowing each participant to focus on the part of the problem they do best: modelers can focus on laying the theoretical foundation upon which models can be generated, while members of the crowd focus on providing both their direct observations and unique perspective and insights in a form that can be applied to the theoretical foundation provided.

Modelers developing operational social science models spend a large amount of time tuning their models to new data [2] [3]. The model's theoretical constructs do not change, but the most effective operationalizations for those theories change as new data emerges over time. CSM overcomes the process in which modelers chase the data searching for the ever transient best fit for operationalization of a theoretical concept.

The CSM approach ultimately hinges on three factors:

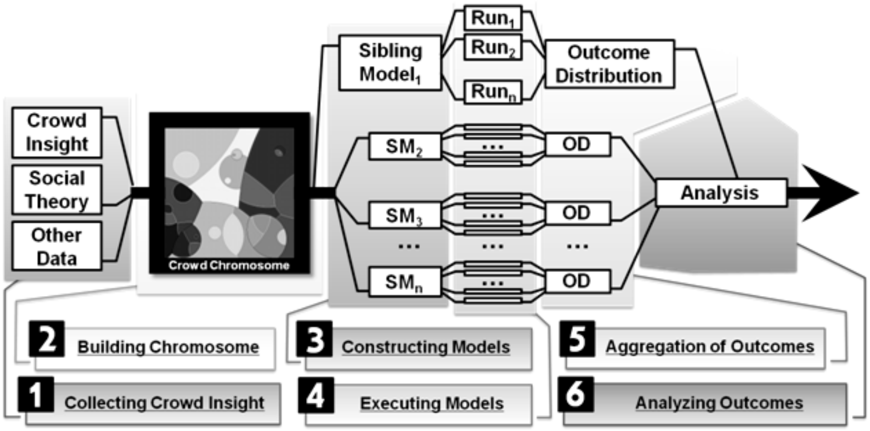

Figure 1: The Crowd Sourced Model building process

- Reimagining the role and products of the social scientist as that of creating theoretical scaffolding.
- Harnessing the consensus and dissonance of the crowd into an actionable data structure representative of its perspective.
- Viewing the models generated through this process in a new way that leans heavily on permutations, combinations and aggregate analysis.

2. CROWD SOURCED MODELS APPROACH

The CSM process is a combination of process, theory and software that investigates the creation of models through crowd sourcing. This work is still in the early stages; the initial effort focused on the instantiation of Agent Based Models based on simulated crowd insights and the results were encouraging. The experimentation showed that the CSM process was able to create models similar to an academic modeler's "Ivory Tower" model, indicating initial feasibility.

CSM is built upon the theoretical foundation developed by the social scientists contributing to the model construction. This foundation includes the theoretical building blocks of the model and the artifacts necessary to effectively harness the crowd's contribution in the context of those building blocks (perhaps including questions and metadata that help understand the modeler's contribution). To put it another way, the modeler's responsibilities in a post CSM world include not only the theoretical work but also a meta-level task of building the surrounding artifacts that enable the crowd to do their portion of the job. Making the modelers (and by extension the foundation) an enabling force for the rapid modeling process.

From the theoretical foundation provided by the social scientist, the crowd input is used to orient or calibrate the theory to the current state of the domain space of interest. The process to connect the sociologist's theory with the crowd's operationalization is based on an epistemological decomposition of a model as described in [4]. The epistemological decomposition of a model is essentially a resume of a model that specifies the conceptual, theoretical, and operational content of a model. This decomposition details the relationship between a theory and its various possible operationalizations (i.e., instantiations in a real-world domain). By using mechanisms and techniques in line with those defined under [4], the social scientist can generate supporting content from the crowd's current and appropriate operationalizations of the concepts and relationships being modeled. For example, if the social scientist generates a theoretical foundation describing the relationship between government hostility and rebellion, the operationalization of this relationship will need to capture a real-world, current, and domain-relevant operationalization for government hostility. To do this, the social scientist may generate a question or set of questions designed to query a crowd member about their perspective on the nature of government hostility in the domain of interest. In this way, as the nature of government hostility changes and evolves (e.g., government hostility towards insurgents transforms into generalized government hostility in the form of human rights violations), the crowd insights that inform the model may reflect these changes. CSM will be able to generate and manage mappings from theoretical foundations to crowd operationalizations.

The experimentation and results will be presented herein, yet we begin by considering the defined approach to building a CSM as illustrated in Figure 1. The CSM approach consists of 6 steps: (1) Collecting crowd insights, (2) Building the crowd chromosome, a novel data structure for storing and leveraging crowd insights, (3) Constructing models based on the crowd chromosome, (4) Executing the crowd simulations generated from the crowd chromosome, (5) Aggregating the results, and (6) Analyzing outcomes.

2.1. Collecting Crowd Insights

Collecting insights from a crowd about a specific domain is an activity that can happen at or before model creation time.

The gathering of crowd insights is guided by two factors, (1) the needs and guidance of the model commissioner, and (2) the theoretical foundation laid by social scientists. At its core, collection is a crowd survey or a crowd querying process with the intention of answering the kinds of questions that a modeler would normally answer with domain research. The queries may be made of a live crowd in the form of crowd surveys, or the queries may be made of data accumulated from resources reflecting crowd views such as social media data, news reports, financial data, or other sources of socio-cultural content describing a crowd's perspective. By querying the local population (or data sources reflecting the perspective of the local population), better domain content may be extracted [1]. Ultimately, this domain familiarity has the potential to shape answers to difficult questions and even highlight times when the wrong question was asked (sometimes the mark of a true expert is not answering a question but knowing you're asking the wrong one). This can be accomplished through the previously mentioned foundation. The foundational artifacts have the potential to not only inform the way models are created but also what models and how they are targeted.

The crowd may be composed of analysts responsible for a specific region, members of the Intelligence community, boots on the ground in a particular theater, the citizens of a country or simply open to the public (or a combination of these). As a result it is possible to gather a wide range of perspectives and first hand information from people immersed in the domain of the model being generated.

2.2. Building the Chromosome

CSM aims to capture a crowd's diverse insights, and translate this information into operationalizations that computational social science models may exploit. The data structure that captures the crowd's nuanced and diverse perspectives is called the Crowd Chromosome.

The Crowd Chromosome is synthesized to encapsulate the diversity of opinion of the crowd and create a structure that accurately represents the views of the crowd. It does not simply capture the consensus, but also surfaces the diversity of the crowd's perspective. The Crowd Chromosome visualized in Figure 2 is an illustration of the diversity found in a crowd. The crowd's insights are highly dimensional with many perspectives describing the domain. That dimensionality is reduced through clustering techniques. This produces a more manageable set of clusters that are arrayed using a potential field approach (repellant and attractive forces that position similar vs. dissimilar crowd insights – e.g., attracting opinions of varying levels of government hostility and repelling opinions of government cooperation). The potential fields approach positions the foci of each cluster across a two dimensional place such that similar clusters are close together while dissimilar clusters are farther apart. The cluster foci were then used to generate a nested Multiplicatively Weighted Voronoi Tessellation, where the magnitudes assigned to the foci are based on the popularity of the represented clusters (such that more popular opinions have a greater weight and thus a larger representation in the tessellation). The result is a structure representing the distribution of the diversity of a crowd's insights, arraying clusters of consensus in ways that group based on similarity.

The Crowd Chromosome is analogous in character to genetic material because the chromosome is the source from which a set of Crowd Sourced Models may be instantiated such that various perspectives garnered from the crowd may be incorporated into the models and ultimately manifested in the results. This instantiation process is covered in more depth in the next section, but because these models are all derived from the same "genetic material," (i.e., Crowd Chromosome) they will be referred to as siblings. Each sibling will be instantiated as a derivative of the Crowd Chromosome. The instantiated siblings are meant to encapsulate and highlight different permutations and combinations found within the Crowd Chromosome (and by extension the observations of the crowd).

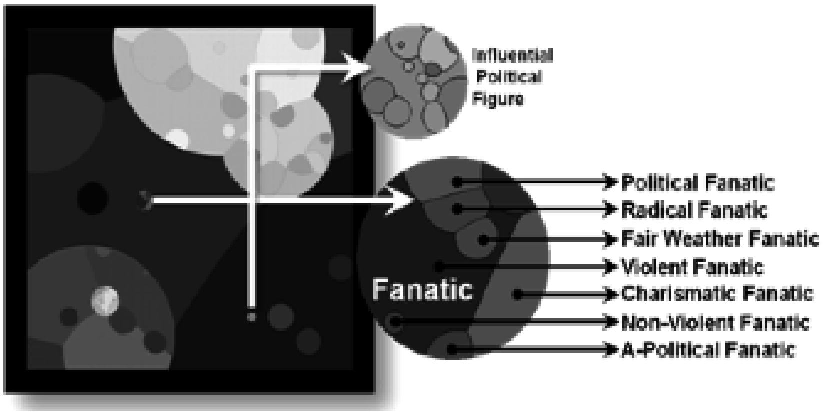

Figure 2: A crowd chromosome deconstructed

2.3. Constructing Models

The Crowd Chromosome is designed as a genetic code from which all the sibling models are created. The Crowd Chromosome captures all of the diversity exhibited by the crowd, diversity that can be combined and recombined in a near infinite number of ways. For each model, a collection of observations of the crowd are extracted. The high level concepts making up the theoretical foundation (created by social scientists) are then operationalized as observations of the crowd.

While the foundation provides the path for creating the model, it is crucial that backward traceability is maintained, allowing the user to trace backwards from an interesting model result to the specific combination of observations/approaches that yielded the result. This transparency is essential to aggregate and analyze the results (see Sections 2.5 and 2.6).

2.4. Executing Models

Depending on the nature of the models, execution may be as simple as running the model once (for statistical models or other deterministic models) or a matter of doing many runs (for instance with Agent Based Models). In either case this is an easily streamlined problem. As both the individual siblings and individual runs provide very natural parallelization points (easily distributed across hardware with no cross dependencies), these processes are embarrassingly parallel.

2.5. Aggregation of Outcomes

A variety of methodologies for dealing with the aggregation of multiple runs (or data points, in general) of non-deterministic models have been investigated [5]. However, with the CSM approach we're looking at an order of magnitude more because we are aggregating the aggregation of many models (and many runs), models that are potentially heterogeneous. To accomplish this, we use a two-step process. The first step is to analyze each individual sibling models based on a measure of interest, essentially capturing each sibling's contribution to the aggregate set of results. (For the models investigated here, the measures were those of probability of the occurrence of an event of interest such as rebellion.) In the second step, we begin clustering model results to identify trends manifesting among the siblings. The aggregation of aggregations generates a distribution of possible outcomes as indicated by a set of siblings generated from a crowd chromosome, a representation of the crowd itself. Hence, the distribution of possible outcomes is a manifestation of the crowd insights.

2.6. Analyzing Outcomes

Once the individual sibling's results are clustered into forecasted crowd trends, a set of possible projections or futures may be generated. These clusters can be traced back to the crowd's particular observations that generated them, and can be compared based on the portions of the genetic materials that they are derived from. For example, if aggregation detects a trend of rebellion in a group of sibling models, we are able to trace each sibling back to the observations in the Crowd Chromosome that were used in instantiating the sibling. This analysis may show that many of the siblings are based on various crowd observations dealing with government aggression. This process of traceability allows the system to provide insights based on the aspects of the crowd's contribution that were the lynch pin in a particular cluster of results, ultimately enabling the traceability of potential future trends observed in the model results back to elements of the current state of the world that were used in generating the model's internal state. This possibly causal mapping (as supported by the sibling models) between what may happen in the future and what is currently happening now teases out possible levers or foci for mitigating the potential for future events.

3. PROOF OF CONCEPT EXPERIMENT

The CSM process fundamentally changes the way we think about and approach model creation. It also introduces some layers of indirection between creators' perception of the world and the way that hypothesis manifests in the actual model. Our experimental approach was designed to isolate those new layers of indirection and verify that they faithfully preserve those hypotheses, and by extension create models that can be strongly correlated to the crowd observations that generated them, ultimately proving useful.

The primary layers of indirection that we sought to explore were: (1) the crowd survey methodology; (2) the clustering and encoding of those responses into the crowd chromosome; and (3) the creation of models using the resultant crowd chromosome.

In order to get quantitative results, our approach was to seed the synthetic crowd in a way that their opinion on average would be similar to an existing vetted model. This left significant opportunity for individuals within the synthetic crowd to dissent, resulting in a crowd chromosome that contained a great many observations / opinions not held by the creator of the baseline domain model. By comparing the models created by the academic modeler to those generated by the CSM process (seeded with a similar world view), we were able to measure the deviation introduced by the layers of indirection. Finally both the academically vetted model and the models created using the CSM approach will be compared to randomly generated models, giving us a baseline.

The model comparisons are designed to predict five distinct events of interest (EOI) so our quantitative model comparison will be based on the similarity in prediction across those EOIs.

3.1. Design

The synthetic crowd is the starting point for our experimentation, as it provides the raw data that becomes the Crowd Chromosome and ultimately creates the models we'll be comparing. Because these models are solely derived from the observations and perceptions of the world it was crucial that individuals within the synthetic crowd have some semi-informed notion of the world. To accomplish this, the crowd was seeded with the baseline academic model's parameters and then individuals within the crowd were perturbed using those as a baseline. As a result no synthetic individual perfectly matched the academic baseline's parameters and some (~3%) strongly disagreed with that baseline, but on average the crowd held a similar view of the world to the academic model (allowing them to be compared objectively).

Once the synthetic crowd has created a response (composed of ~60,000 observations spread across the eight individual types within the model) the system then clusters those responses and generates the nested multiplicatively weighted Voronoi tessellation (MWVT) that is the Crowd Chromosome. Random walks across the Crowd Chromosome will create the agents that populate the set of sibling models. This is the next major question the experiment is meant to address: Does the Crowd Chromosome and random walk approach yield models that are fundamentally different from the academic baseline, but also don't introduce so much variance it questions the faithful stewardship of the crowd's observations?

In the end, the hypothesis we designed validates that models generated by the CSM process should be starkly different from the random baseline and similar but not identical to the models created by academics. These results indicate that the layers of indirection introduced by the CSM process faithfully extract both the diversity and consensus of the crowd (shown by the similarity and deviation from the academic model) and also that the Crowd Chromosome and random walk approach is capable of creating agent populations that are consistent with those expected by the EOI's applied to them.

3.2. Framework

The PS-I (Political Science – Identity) framework developed under Dr. Ian Lustick of University of Pennsylvania was used for the development and execution of the sibling models [6] [7]. The current version of this software is being maintained by Lustick Consulting.[1]

3.3. Execution

For each sibling model born of the Crowd Chromosome, the model is executed 100 times, resulting in 100 unique data files containing output from the model run. These data files contain values for various variables generated by the model and stored at each time step of execution. PS-I modelers can configure the output to store their own defined variables at designated time intervals. For the CSM model runs, we captured the same variables captured by the baseline model at the same time intervals in order to appropriately compare the CSM model results to those of the baseline. Specifically, the model variables measure events of interest such as religious and ethnic conflict and were captured at every time step.

Once the models were run and the data was collected, an analysis was conducted to calculate aggregated forecasts for events of interest including rebellion, domestic political crisis, insurgency, ethnic conflict, and religious conflict. To generate the aggregation, first, each of the 100 model runs was assessed for the event of interest. These individual model run measures were binary values indicating whether an event occurred in the model run or not. The measure was derived for each time step of a run such that each run generated 60 values. (There were 60-time step per model execution.) These 60 time steps were aggregated into 12-month intervals where five time steps represented a month. The aggregation mechanism was a logical OR-ing of each of the five time steps composing a month. If an event occurred in any of the five time steps, then the event was considered as having occurred in the aggregated month measure.

For the analysis performed here, each run was aggregated into 12 event measures, one for each month of the year. For each month, we had 100 measures generated (one measure per model execution). From these 100 values representing the measures for a single month, a probability was generated based on the ratio specified in equation 1.

$$\frac{\textit{Number of model runs generating an event}}{\textit{Number of model runs}} \tag{1}$$

Using this calculation, 12 aggregate measures were derived, one per month of the year. These 12 measures were then compared to those of two other models: (1) the baseline model output and (2) a control output based on a randomized population.

3.4. Outcomes, Aggregation and Analysis

Our initial findings were encouraging with a few caveats. First, the level of deviation and variance we saw between the academic models and those created using the CSM process show a reasonable but not excessive level of variance proportional to the variance introduced into the population. Second, with some notable exceptions (no longer compatible EOIs) the models created show an ability to be useful as predictive tools in the same way as the academic models.

This inapplicability of certain EOI's is to be expected, as the disruptive paradigm shift means that certain baseline

[1] http://lustickconsulting.com/

assumptions encoded into those EOIs are no longer applicable. For example, the CSM process for building our ABMs resulted in a group of agents within a single sibling model that was highly varied from each other within that single sibling model. The ethnic conflict measure is based on a level of ethnic diversity in an area (among other factors). The level of diversity imbued to an ABM society by way of the CSM process was triggering our existing definition of ethnic conflict. For this reason, the CSM model was showing significantly higher levels of ethnic conflict than the baseline model. However, in future research there is an opportunity to develop new and potentially more accurate measures that embrace the diversity and nuance of CSM models, in addition to using tried and true measures.

4. CONCLUSION

The Crowd Sourced Modeling (CSM) approach relies on a large crowd of individuals with diverse knowledge to capture insights and experiences that may be leveraged to achieve the capability to model without requiring trained, elite, and rare expert modelers. Instead, the CSM approach makes use of the experience and insight of individuals such as soldiers or relief workers in the field. By “listening” to the individual interacting more immediately with the environment being modeled, the CSM approach may adapt to and incorporate rapidly changing contextual factors more quickly.

The research presented here is the first step of a feasibility analysis. The experimentation described is rudimentary, but provides initial feasibility of mechanisms for leveraging a crowd and translating crowd insights into computational models with initial results indicating viability. Future work will continue this feasibility analysis by addressing issues regarding (1) the manifestation of crowd diversity in the generated ABM; (2) comparing and aggregating across sibling model results; and (3) assessing how Crowd Sourced models compare to what modelers create today.

5. REFERENCES

[1] S. Page, The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies, Princeton: Princeton UP, 2007.

[2] S. O'Brien, "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research," *International Studies Review,* pp. 87-104, 2010.

[3] B. Kettler and M. Hoffman, "Lessons Learned in Instability Modeling, Forecasting, and Mitigation from the DARPA Integrated Crisis Early Warning System (ICEWS) Program," in *2nd International Conference on Cross-Cultural Decision Making: Focus 2012*, San Francisco, 2012.

[4] A. Ruvinsky, J. Wedgwood and J. Welsh, "Establishing Bounds of Responsible Operational Use of Social Science Models Via Innovations in Verification and Validation," in *2nd International Conference on Cross-Cultural Decision Making: Focus 2012*, San Francisco, 2012.

[5] D. Thomas and W. Luk, "Estimation of Sample Mean and Variance for Monte-Carlo Simulations," in *International Conference on Electrical and Computer Engineering: ICECE Technology 2008*, 2008.

[6] I. Lustick, "PS-I: A User-Friendly Agent-Based Modeling Platform for Testing Theories of Political Identity and Political Stability," *JASSS,* 2002.

[7] I. Lustick and D. Miodownik, "Deliberative Democracy and Public Discourse: The agent-based argument repertoire model," *Complexity,* pp. 13-30, 2000.