Proposal White Paper

FOA# : (DoDGARS) 32 CFR 22.315(a)

ONR Announcement #N00014-17-S-F006

Proposed Title:

Developing top-down data driven models to predict population flows during crises.

Proposer's technical point of contact:

Itai Cohen

Professor of Physics at Cornell

Cornell University

508 Clark Hall

Ithaca NY, 14850

Tel: 607 255 0815

Email: itai.cohen@cornell.edu

Topic 8 (ARO): Modeling Interdependence among Natural Systems and Human Population Dynamics

Developing top-down data driven models to predict population flows during crises.

Predictive, quantitative models of the mass flow of human beings migrating in response to environmental and political pressures would greatly improve responses to developing crises. One need only look at the European response to they Syrian Refugee Crisis in the summer of 2015 to understand that as a society we are poorly positioned to predict and respond to sudden, large-scale population dynamics in response to emerging crises (Hynie 2016; Goodman 2016; Willekens et al. 2016). The challenge for making such predictions is formidable. Recently however, more data sets are becoming available that provide spatio-temporal detail of population dynamics on different scales (Lu et al. 2016), detailed socio-demographic characteristics, data on the permanence of the moves (Bohra-Mishra et al 2014), and information on how these population flows relate to environmental and political events (Hauer 2017). Coupled with this extraordinarily detailed data, is the development of new top-down statistical techniques by members of our team that enable simplification of complicated agent based models and, more importantly, enable predictions about future events. Using these statistically motivated top-down approaches in combination with data sets that cover spatiotemporal scales ranging from crowd dynamics to national migration, we propose to develop mathematical models that make predictions about the future response of populations to crises. Such predictions would provide crucial information to governments preparing for future disasters in hot spots such as the coastlines of Bangladesh, the Mekong Delta, and the American Gulf Coast where migration crises involving many millions of people could emerge in the 21st century (Clark et al. 2016).

The Research Need: A key hurdle that must be overcome is that agent-based modeling, the predominant modeling approach for predicting mass migration, requires complete specification of an essentially arbitrary set of rules for interactions between agents and their environment. These rules are difficult to quantify and validate, even when sufficiently detailed data sets are accessible. In fact, because the space of possible rules is so large, many permutations of the rules are consistent with the observed data. Thus, such models are prone to statistical overfitting, which hampers their interpretability. Moreover, while agent-based models can reproduce behaviors observed in training data, they often fall short when predicting behaviors outside the training data because different sets of rules apply. Since it is not possible to simulate all the possible scenarios associated with all the possible rules a priori, such agent-based models often fail to provide useful information under the time and resource limits created by a crisis.

Proposed Technical Approach: In contrast to such bottom up approaches, we propose to use tools from statistical physics and mathematics to infer rules for mass behaviors directly from observations of local population densities and to quantitatively predict mass behavior under new circumstances. A central tenant of statistical physics is that generic thermodynamic or system scale behaviors emerge from sometimes complicated interaction rules among large numbers of particles. This reduction in complexity allows us to describe the thermodynamic state of a gas with a small number of variables such as pressure, density, volume, and temperature, rather than the $\sim 10^{23}$ equations of motion for all the atoms. Remarkably, these emergent behaviors are often insensitive to the detailed nature of the underlying interactions. Crucially, the vast reduction in the number of variables makes it much easier to map out large regions of the thermodynamic landscape, simulate the state of the system, as well as interpret and quantify uncertainties of the collective behavior. Here, we pursue the hypothesis that a similar scenario emerges in the study of large populations, so that behaviors arising from generic agent-based models can be predicted using simpler top-down approaches. As such, we will adapt canonical frameworks from classical density-functional theory, systems theory, and information theoretic approaches to determine the minimal set of variables necessary to describe the overwhelming majority of the data.

Importantly, often the reduced variables in these top down approaches can be split into those that depend on interactions between different agents and those that depend on interactions between agents and their environments. When this splitting is achievable, *it is possible to mix and match such variable dependencies to predict responses in entirely new situations*. In fact, we already demonstrated that density-functional techniques can indeed separate between agent-agent and agent-environment interactions and make *quantitatively accurate predictio*ns for population distributions in a system comprised of hundreds of flies, schools of fish, and human pedestrians (**Arias** & **Cohen**). We hypothesize these ideas will carry over even more naturally to larger populations including mass migration. Specifically, data on the interaction between people and a new migration route during a period of calm, will be combined with data on how people previously responded to the local density of migrants during periods of distress in other migration routes, to predict population flows in new migration route during periods of distress. *The ability to use population measurements during times of normal migration, combine them with previous data on distressed migration in other regions, to predict population flows during times of crisis in a new region of interest would constitute a profound advance for the field.*

To ensure development of a predictive capability, we shall develop our models using state of the art data sets encapsulating three relevant spatiotemporal scales: (a) hundred to thousands of person crowds over time scales of minutes to hours, (b) tens to hundreds of thousands of persons over days, weeks, and months (c) hundreds of thousands to millions of persons over years. These scales are chosen as those relevant to the key stages in population response in crises, specifically the immediate dispersal pattern as people escape the initial threat, the initial displacement into temporary shelter and housing, and, finally, changes in permanent residence.

(a) **Video of Crowds:** On the smallest spatiotemporal scale, our data will consist of motion tracked trajectories of people from films of crowds at museums, concerts, and political rallies (Silverberg et al. 2013) (**Cohen).** For these data automated image analysis methods will be used to determine each individual's positions and velocities in time. Such data are directly analogous to the studies **Arias** and **Cohen** conducted on calm versus excited flies walking in 2D arenas. For example, crowds at classical versus heavy metal concerts can be used to develop models for calm versus excited states of people. A very useful aspect of these data sets is that the ease of data collection enables further development of the top down modeling capabilities.

(b) **Cell Phone Tracking:** At the intermediate our data will consist of the activity of mobile phone users on cellular network, known as Call Detail Records (Lu et al. 2016) that indicate the changing location of a phone through time. These data will be aggregated to enable study of migration behaviors at very high spatiotemporal resolution over regional/county scales and time scales of days to months (Wesolowski et al. 2013, Lu et al. 2016). For the current work we will use ONaTel mobile network data collected in Senegal between 1 January and 31 December 2013 to evaluate mobility around a tropical storm that struck in August 2013 resulting in catastrophic flooding across Senegal, including the capital city, Dakar (**Wrathall**).

(c) **Population Census Tracking:** At the largest scale our data will consist of county-to-county scale population density maps over large stretches of time. One large data set will come from Japan's Statistics Bureau of the Ministry of Internal Affairs and Communication and the Nationwide Evacuee Information Exchange System to examine migration patterns in the wake of the Fukushima Daiichi nuclear disaster (**Hauer**). A second data set will come from the Internal Revenue Service data on county-to-county migration (1990-2015), American Community Surveys on detailed demographic information of these migrants (2005-2015), and Spatial Hazard Events and Losses Database to determine how environmental crises alter normal migration (**Brown** & **Hauer**). These are the most comprehensive available US migration data covering 97% of all tax filers as well as their dependents and reports on their locations, demographic makeup, and how environmental crises alter normal county-to-county migrations.

Management Plan: The group will work in three teams, each of which will tackle a specific scale. Teams will be headed by postdoctoral fellows who will lead the graduate students working to gather the data and develop the theoretical models. While covering distinct spatio-temporal scales, these data share many common characteristics. For example, at each scale it is possible to define a normal versus excited or crisis mode of behavior. In addition, in each data sets we can define analogous interaction mechanisms between people or people and their environment. Thus, each month the three teams will get together to share best practices and the various techniques they learned and developed as well as hurdles they overcame to analyze the data. For example, in all three cases it will be necessary to design our methodology to be robust to incomplete data. When data on some behaviors is missing we will describe and transfer techniques such as Bayesian inference used by a particular team to the other teams (**Arias** & **Frazier**).

The data sets for each team will be split into a training data set that will be used to develop the top down modeling approach and a testing data set that will be used as a testing ground for the teams developed models. For example, the data set on long term migration will contain data on migration in the Sothern United States during periods of relative calm as well as data on the response to Hurricane Katrina. Therefore, a subset of the data that addresses this region can be used to develop our top down model. Once such a model is developed we can apply it to other regions such as Florida and try to use the migration data during periods of calm to predict the response to Hurricane Andrew or Superstorm Sandy.

Impact on Department of Defense Capabilities: As we near conclusion of the grant, we will use our methodology to create a computational planning tool for assessing risk of sudden, large-scale migrations arising from future natural disasters. This program will use an ever growing training data set to inform its predictions, and will calculate probability distributions for potential migration patterns for the short- (hours), medium- (days and weeks), and long-term (years) resulting from a disaster. For example, to determine how dense crowds will behave in response to an explosion at a venue, our tool will take the relevant population dynamics data gathered from videos of crowds in excited states such as mosh pits and stampedes and synthesize them with data gathered on how calm crowds exit the venue of interest. Alternatively, to assess the probability distributions for migration patterns in response to a typhoon hitting south east Asia, our tool will synthesize the population dynamics in the wake of disasters ranging from Fukushima to Hurricane Andrew, tease out the relevant portion of the population response and combine it with information on current migration patterns in south east Asia. Using this approach we will, for the first time, provide risk assessments of sudden large-scale migration as well as visualize the predicted population dynamics that would result under new disasters. Such tools could be used to assess plans for new concert venues, determine supply lines during crises, and enable governments of states and municipalities to make long term preparation plans in the event of a catastrophic environmental or political disaster. Thus, the impact of such a tool on the ability of national and international agencies to plan for impending disasters would be profound.

The Team: Our team is uniquely poised to address each aspect of the proposed work. **Arias** (Physics, Cornell) and **Frazier** (School of Operations Research at Cornell and Uber) are leading experts in statistical physics and mathematics techniques including density functional theory, systems theory, and information theoretic approaches. **Cohen** (Physics, Cornell) is an expert at developing image analysis techniques for tracking particle motions as well as statistical approaches for analyzing their collective behaviors. **Brown** (Applied Demographics, Cornell) is a member of the National Academy of Sciences Standing Committee for Reengineering Census Operations, runs the Program on Applied Demographics (PAD), and is a Senior Research Associate in the Cornell Institute for Social and Economic Research (CISER). **Hauer** (Applied Demographics, UGA), recipient of the Terry Award in applied Demography leads the Institute of Government's Applied Demography Program. Finally, **Wrathall** (Geography, Environmental Sciences) is one of the pioneers using cell phone data to study migration patterns.

Summary of Estimated Costs: $1.1 million per year for five years will pay for six graduate students, three postdocs, and access to experimental and data resources we propose to utilize.

Organizational Conflicts of Interest: None.

Anticipated Human or Animal Subject Research: None.

References:

- Hynie, Michaela. "Life on the edge." (2016): 444-445.
- Goodman, Allan. "Lifeline for refugee scholars." (2016): 1207-1207.
- Willekens, Frans, Douglas Massey, James Raymer, and Cris Beauchemin. "International migration under the microscope." *Science* 352, no. 6288 (2016): 897-899.
- Lu, Xin, David J. Wrathall, Pål Roe Sundsøy, Md Nadiruzzaman, Erik Wetter, Asif Iqbal, Taimur Qureshi et al. "Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh." *Global Environmental Change* 38 (2016): 1-7.
- Bohra-Mishra, Pratikshya, Michael Oppenheimer, and Solomon M. Hsiang. "Nonlinear permanent migration response to climatic variations but minimal response to disasters." *Proceedings of the National Academy of Sciences*111, no. 27 (2014): 9780-9785.
- Hauer, Mathew E. "Migration induced by sea-level rise could reshape the US population landscape." *Nature Climate Change* (2017).
- Clark, Peter U., Jeremy D. Shakun, Shaun A. Marcott, Alan C. Mix, Michael Eby, Scott Kulp, Anders Levermann et al. "Consequences of twenty-first-century policy for multi-millennial climate and sea-level change." *Nature climate change* 6, no. 4 (2016): 360-369.
- Silverberg, Jesse L., Matthew Bierbaum, James P. Sethna, and Itai Cohen. "Collective motion of humans in mosh and circle pits at heavy metal concerts." *Physical Review Letters* 110, no. 22 (2013): 228701.
- Wesolowski, Amy, Nathan Eagle, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. "The impact of biases in mobile phone ownership on estimates of human mobility." *Journal of the Royal Society Interface* 10, no. 81 (2013): 20120986.