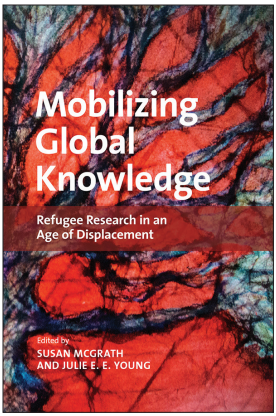

MOBILIZING GLOBAL KNOWLEDGE: REFUGEE RESEARCH IN AN AGE OF DISPLACEMENT

Edited by Susan McGrath and Julie E. E. Young

ISBN 978-1-77385-086-3

THIS BOOK IS AN OPEN ACCESS E-BOOK. It is an electronic version of a book that can be purchased in physical form through any bookseller or on-line retailer, or from our distributors. Please support this open access publication by requesting that your university purchase a print copy of this book, or by purchasing a copy yourself. If you have any questions, please contact us at ucpress@ucalgary.ca

Cover Art: The artwork on the cover of this book is not open access and falls under traditional copyright provisions; it cannot be reproduced in any way without written permission of the artists and their agents. The cover can be displayed as a complete cover image for the purposes of publicizing this work, but the artwork cannot be extracted from the context of the cover of this specific work without breaching the artist's copyright.

6

Big Data and Early Warning of Displacement

Susan F. Martin and Lisa Singh

Introduction

In recent years, the global population of people forcibly displaced by conflict and persecution has reached levels unprecedented since the Second World War: 68.5 million in 2017 (UNHCR 2018b). Acute natural hazards also lead to large-scale movement of people, some temporarily and others permanently. Between 2008 and 2016, the number of disaster-displaced populations averaged 21.5 million each year (IDMC 2016). In many cases, both human-made and natural factors precipitate large-scale displacement, as witnessed by recurrent famines in Somalia caused by the confluence of drought, conflict, and political instability that impede access to food relief.

Because much forced migration is unexpected, communities can be overwhelmed by refugees and displaced persons if they have insufficient warning. Even relatively wealthy countries may fall victim. The massive movements in 2015 of Syrian, Afghan, Iraqi, and other asylum seekers into Greece (with the hope of moving onwards to the rest of Europe) is a clear example of such chaos. In 2017, the concurrent outbreak of famine in northern Nigeria, Yemen, Somalia, and South Sudan also seriously

challenged capacities to respond to both the mass starvation and mass displacement that resulted, despite persistent drought and famine warnings.

Given the unprecedented levels of forced displacement, and recurrent problems in addressing large-scale movements, an urgent need exists to develop an evidence-based early warning system that can enable governments and international organizations to formulate contingency plans, establish appropriate policies, and pre-position shelter, food, medicines, and other supplies in areas that are likely to receive large numbers of refugees and displaced persons. This need for early warning, as part of comprehensive contingency planning, has been recognized in the recently adopted Global Compact on Refugees (2018a).

With the wealth of data available via social media, search engines, and more traditional data sources (see chapter 8), it is natural to begin discussing how this information can be used to make progress toward identifying and forecasting forced migration. The precedent is in place to forecast many of the drivers of displacement. For example, in recent decades, early warning systems alert the international community as well as national and local actors of impending humanitarian crises. Tsunami and famine early warning systems monitor and analyze data relevant in anticipating acute and slow onset crises, respectively, relying on scientific, technological, economic, social, and other indicators (see FEWS Net and NOAA National Tsunami Warning Center). Predicting crises in other domains, such as conflict and violence, has proven more difficult but organizations such as the International Crisis Group put out regular alerts of worsening conditions and ACLED, the Armed Conflict Location and Event Dataset, codes the actions of rebels, governments, and militias within unstable states, specifying the exact location and date of battle events, transfers of military control, headquarter establishment, civilian violence, and rioting.

Forecasting displacement during these situations, particularly when a complex mix of drivers are at work, as seen in places facing prolonged drought and conflict, has proven more elusive. This chapter identifies novel big data sources, methodologies, and challenges that need to be addressed in order to develop more robust, timely, and reliable evidence-based systems for detecting and forecasting forced migration in the context of humanitarian crises. The chapter also recognizes the immense challenges and barriers to establishing more reliable forecasting capabilities. Despite

the availability of data, patterns of forced migration in anticipation of, during, and following conflict and acute natural hazards are notoriously difficult to predict. What appear to be very similar pre-existing stressors and triggering events and processes can result in significantly different levels, forms, and destinations of displacement. The warning signs of displacement or significant changes in the nature of movement are often present but difficult to piece together in a coherent fashion.

Our research into these systems has identified a number of problems that must be solved to improve the effectiveness of early warning systems, particularly as they apply to displacement: 1) identifying and collecting masses of timely, reliable data on the complex factors that affect flight; 2) developing analytic capability to discover indicators of movement—specifically, leading indicators that displacement will occur rather than trailing indicators that confirm that movement has already taken place; 3) instituting mechanisms to allow policymakers and practitioners to test out scenarios to determine if actions will have positive or negative consequences in averting displacement or providing better assistance and protection; and 4) building the political will to act on the warnings. New technologies and analytic tools make it more likely that the first three problems can be tackled. The fourth problem is, of course, more difficult to solve but more effective early warning tools might challenge political leaders to act, at least in implementing more timely emergency relief operations. We focus our discussion on the lessons learned during a multi-disciplinary project on early warning of displacement, funded by the US National Science Foundation (NSF), Canadian Social Science and Humanities Research Council (SSHRC), the John D. and Catherine T. MacArthur Foundation, and Georgetown University's Massive Data Institute (MDI).

The next section discusses the value of such a multi-disciplinary approach and describes our efforts to build a community of scholars and practitioners to make progress in this area. The following sections assess the directions of research that are necessary to harness potential benefits of big data for anticipating patterns of forced migration: the development and validation of a theoretical model of forced migration that captures the complexity and dynamism of the phenomenon; the identification and collection of relevant data related to the complex factors that affect flight; the need for methods that take disparate forms of data with varying degrees of reliability and completeness and extract meaningful, timely

evidence of movement; and the development of analytic tools that enable policymakers and practitioners to test different scenarios to respond to forecasted movements. Through this discussion, we describe initial efforts to use newspaper and social media data to begin generating direct and indirect indicators of movement. While big data has some important limitations, including the ratio of noise to signal that can distort the accuracy of forecasts and potential biases that exist because of incomplete data, these diverse data can be used by researchers to capture fragments of human behaviour at large scale, in real time; this glimpse into human behavior is not always available using traditional approaches. It is the combination of traditional survey data, available administrative data, and new structured data values extracted from big data sources that make early warning systems for forced migration plausible. While obstacles still exist for early warning tools in this area, the growing number of available sources and the advances in technologies makes this an area where significant progress can be made over the next decade.

Building the Community of Scholars and Practitioners

Developing an effective early warning system of population displacement requires collaboration and shared learning between subject matter experts who understand the factors that contribute to forced migration and technical experts who understand how to collect, store, mine, and analyze masses of data derived from international, national, and local sources. It also requires a close working relationship between these academic experts and practitioners who understand the intricacies of implementing an early warning system in the real-world context of mass displacement. In 2013, we began assembling such a team with funding from the National Science Foundation (NSF) in the United States. The team grew with additional funding from SSHRC. It has included scholars renowned in their respective fields, from Georgetown (US), York (Canada), Fairfield (US), Fordham (US), Kultur (Turkey), Sussex (UK) Universities, University of Toronto (Canada), and Lawrence Livermore National Laboratory (US). We have drawn on the advice of practitioners from the Jesuit Relief Services, Refugees International, Women's Refugee Commission, the Brookings-LSE Project on Internal Displacement, CARE Canada, Médecins Sans Frontières Canada, and the UN High Commissioner for Refugees (UNHCR).

The interdisciplinary approach has exposed social scientists to new modelling approaches for analyzing their subject matter. At the same time, computer scientists have benefited from domain expertise in the social sciences, enhancing the intellectual merit of our project. This expertise has provided insight for the development of beyond state of the art data mining and machine learning of very large, incomplete, and potentially biased open source databases for topic modelling, event detection, sequential mining, change detection, sentiment analysis, and dynamic graph mining to name a few. As social scientists on the team attempted to explain drivers of forced migration to computer scientists who were attempting to model movement, it became clear that the theoretical models needed improving and that data needed to be collected to test these models. Most theoretical work on migration has focused on labour movements and, to a lesser degree, conflict or environmental migration; relatively little has been done in building theoretical frameworks for understanding complex displacement driven by multiple factors.

The collaboration of researchers and practitioners contributed both to our scholarly knowledge of forced migration as well as to our understanding of the advantages and disadvantages of various potential early warning models. As we wanted the system we planned to develop to be timely, accurate, and user friendly, we determined from our practitioners that a simple alert system would not be particularly welcome in the field. Rather, our practitioners urged us to develop a system that would enable a field user to take the early warning information and explore a range of scenarios and options for response. The collaboration further helped us determine the extent to which information from the field is available and its utility for the purpose of early warning. Our approach to addressing these concerns is discussed in the following sections. See chapters 2, 7, 10 and 13 for further discussion of the benefits of interdisciplinary approaches.

Development of a Theoretical Model of Forced Migration

Our work on early warning focuses on displacement in the context of humanitarian crises—that is, any situation in which there is a widespread threat to life, physical safety, health, or basic subsistence that is beyond the

coping capacity of individuals and the communities in which they reside (Martin et al. 2014). We chose two principal case studies to use in testing our theoretical framework: Somalia (2006–07) and Iraq/Syria (2011–15). These cases were chosen because of the complexity of forces underway in displacing people from their homes as well as the familiarity of the study team with the drivers and their consequences for displacement. They also allowed the team to examine one retrospective though still pertinent case—Somalia—and one escalating and rapidly unfolding case—Iraq and Syria.

Understanding not only why people are displaced but also when, where, and how they move is crucial to the effective prevention of, and response to, mass displacement. Leading indicators of forced migration range from macro-level political, security, economic, social, religious, cultural, and environmental indicators to micro-level material measures that determine whether individual households have the resources and motivation to leave their homes and meso-level factors that interfere with or facilitate movement (Government Office for Science 2011). One of the early efforts to understand underlying drivers of displacement was Schmeidl's (1997) work on root causes, proximate conditions, and intervening factors as potential determinants of refugee migration, with a particular emphasis on the role played by economic underdevelopment, human rights violations, ethnic and civil conflicts, external intervention, and interstate wars. In addition Schmeidl examined the impact of "flight facilitators," including migration networks and geographic proximity, and physical obstacles to movement (such as jungles or deserts). She found that underlying economic underdevelopment and population pressures have minimal impact on predicting displacement but instead it is the level and type of violence that determine the likelihood and size of refugee flows. While Schmeidl (1997) examined refugee flows, Naudé (2010) looked more broadly at patterns of international migration in sub-Saharan Africa and found that violent conflict and GDP growth differentials have the largest impacts on international migration in the region. He concluded that international migration from sub-Saharan Africa is both an adapting and mitigating strategy in the face of conflicts and economic stagnation (Naudé 2010, 350). Moore and Shellman (2004) find that the magnitude of genocide and politicide significantly increase both the likelihood and magnitude of forced migration. Melander and Oberg (2007) found that the intensity of armed conflict is not significantly related to the number of

forced migrants. Rather they suggest that "the threat perceived by potential forced migrants is more related to where the fighting is taking place, than to the overall intensity of the fighting" (2007, 157). Salehyan and Gleditsch (2006) find that the probability of violent conflict is more than three times higher in source countries than in receiving ones.

A number of studies have considered the role of intervening factors in explaining refugee flows. Massey (1988) and Moore and Shellman (2004) introduce the idea of migration networks in the destination countries as contributing to forced migration while Clark (1989) identified five intervening factors that may affect refugee outflows, including the existence of alternatives to international flight within the country, obstacles to international flight, expected reception in the asylum countries, patterns of decision-making among potential refugee groups, and seasonal factors. Schmeidl and Jenkins (1996) discuss some of the problems of timing where long-term or root causes may occur years before the exodus, while medium-term (or proximate) causes may occur only months beforehand. They argue that "triggering events are the most difficult to place. Theoretically, they would occur almost simultaneously with, or only days before, flight but most conventional methods are unable to evaluate the close timing of triggering events" (Schmeidl and Jenkins 1996, 6). They underscore the importance of triggering events: "for policy purposes, triggering events are critical in preparing for emergency relief" (Schmeidl and Jenkins 1996, 6). Davenport et al. (2003) posit that forced migrants make their decisions to flee when they observe threats to their personal integrity. Melander and Oberg (2006) look beyond the question of why people move to analyze the impact of forced migration flows on those that remain behind. Rather than finding that the departure of forced migrants leads to increased future flows, they find that the magnitude of flows declines over time (2006, 130).

There are also psychological and emotional reasons for flight. What is commonly referred to as dread threat theory (Slovic 1987, 2000; Slovic, Fischoff, and Lichtenstein 2000; Slovic, Kunruetheer, and White 2000; Starr 1969) identifies a heterogeneous list of "fright factors" to measure people's responses to safety questions. Because forced migration often occurs in situations of persistent threat, we add a dynamic element to dread threat theory—the menacing context that emerges when a dread threat persists and requires a community to reorganize its life to mitigate consequences of threat. The concept of menacing context has evident value in analyzing

the determinants of forced migration because it links situational factors to decision-making as well as macro, meso, and micro levels of analysis through local perceptions of, and responses to, dread threat (Collmann et al. 2016).

While these frameworks are useful in explaining the reasons that people stay or go in the context of conflict and other crises, they do not adequately capture the diversity in movements that occur in these situations. More effective early warning of displacement must provide greater perspective on when people move, where they go, with whom they move, what modes of transport they take, and other similar factors that determine mobility patterns in situations of conflict and repression. At this stage, no system of this type exists. One major obstacle has been access to relevant, timely local and regional data that can be incorporated into a flexible model of forced migration.

Using Big Data to Identify Determinants and Dynamics of Mass Displacement

This project gave us the opportunity to begin analysis of local print and social media, specifically Twitter, to identify the changing dynamics of events and perceptions that may directly or indirectly trigger displacement (see Payne and Millard chapter for other uses of social media). Our case example examined displacement in and from Syria and Iraq since 2011. We used an archive of more than 700 million publicly available open source media articles that has been actively compiled since 2006 (Singh and Pemmaraju 2017). News articles are added to this archive—the Expandable Open Source (EOS) database—at the rate of approximately 100,000 per day by automated scraping of Internet sources in forty-six languages across the globe. We also compiled a database containing over 1.5 billion tweets in English and Arabic from organizations and individuals that regularly post on developments in Iraq and Syria, and on relevant hashtags, including ones related to ISIL. Using newspaper and social media data begins to give us a glimpse into what people are talking about, what their perception of different events and conditions are, and whether commentary and concern about local conditions are increasing or decreasing.

To determine which ideas, types of events, and topics are correlates of movement or correlates to direct indicators that may not be available during different crisis situations, we also used statistics compiled by the United Nations High Commissioner for Refugees, the International Organization for Migration, the Office for the Coordination of Humanitarian Affairs, and the Internal Displacement Monitoring Centre. Demographic data and economic indicators can also be drawn from standard sources, such as the UN Human Development Index and the World Bank. These data can serve as direct and indirect indicators/variables in the context of migration. Therefore, we need to understand the relationship between these known variables and the variables extracted from noisy, partial, open source big data. Are they well correlated or is there a limited relationship between them? We also plan to correlate big data variables to interview data collected in different volatile regions around the world. Because interviewing is not scalable, if we can find strong correlates between big data variables and interview variables we can use them as proxies for traditional interview variables that may be difficult to obtain in certain unstable regions of the world.

So the primary question becomes: how do we identify meaningful forced migration-related variables from big data sources? An important secondary question regards how we assess reliability and bias of output variables generated from noisy, big data streams? In the previous section, we highlighted a number of factors that influence an individual's decision to migrate or not during conflict. Our approach hinges on understanding 1) the changing dynamics of each factor in a particular location, and 2) the importance of each factor within a particular location or community. We measure both of these by analyzing the changing newspaper and social media content related to these factors in different regions. To accomplish this, the process begins by identifying relevant documents using state of the art information retrieval techniques, extracting useful structured data representations, i.e., sketches of the data, and then using these data representations to construct variables for use in a dynamic forced migration model (Wei et al. 2014). For example, creating a vector of words about violence and computing the frequency of these words across newspaper articles each day can be used as the basis for a time series variable that captures the changing dynamics of violence in a particular location. These changing dynamics may be a strong indirect indicator of movement in

certain conflict areas. Another type of data sketch may be a semantic graph that contains words and phrases as nodes and relationships based on co-occurrence of these words and phrases in articles or tweets. This type of graph can be useful for identifying frequently occurring groups or clusters of words/discussions of local and regional topics of interest. Finally, a third type of data sketch translates words to a mathematical vector space where the weights in the vector space are based on the context in which words are used. Words that are used in similar ways have similar vectors in this word embedding vector space, i.e., similar relationships to other words in the vector space. Of course, many other types of data sketches exist. We highlight these three because they are particularly well suited for tasks involving textual data of varying lengths and *speech quality*.

While it is also possible to generate administrative variables from these data sketches from big data sources, we believe that researchers need to explore big data in new ways and produce new types of variables to gain insight that differs from values that can be determined in other ways. Here we describe interesting variables that we hypothesize will help our understanding of movement in general, and have found important in the context of our analysis of movement in Iraq (Singh et al. forthcoming).

Topic Buzz: Discussions revolve around different themes or topics. Determining the topic(s) being discussed in an article or post is central to understanding how its dynamics are changing through time. Is discussion about political violence increasing or decreasing? Are people talking more or less about weather conditions in a particular town? While different approaches have been proposed for extracting topics from text (Blei et al. 2003; Teh et al. 2006; Wang and McCallum 2006; Blei and Lafferty 2006; Churchill et al., 2018), these models have been designed for longer textual documents that are more coherent than social media, e.g., research articles. New algorithms that adequately handle the noise of social media text streams and the short length of these posts are still in their infancy. Even without automated methods for topic identification, words that are representative of topics can be manually determined by experts. While time consuming, manual annotation is always a reasonable option. If we associate factors of movement to topics of conversation, we can see the prevalence of these topics through time.

Buzz represents the amount of interest in a topic through time. Topic buzz may be popular and trending one week, e.g., discussion about

violence, and have low values the next week. What is interesting is the variation of buzz strength of a topic over time. This buzz strength is based on the frequency of occurrence of relevant words and word embeddings in articles and social media posts for a particular location. One can imagine using a heat map to see the buzz of different topics (indirect indicators of different factors) in different locations. This can give us immediate insight into the distribution of the factors that may be more relevant in a particular region. This distribution is vital for understanding the specific factors that may be more important in different parts of the world.

Our initial focus was on the topics of *violence* and *migration* for computing buzz variables from newspaper data in both English and Arabic (Hockett et al. 2018; King 2016). While we had some expert seed words, we also wanted to determine if using different strategies for augmenting those words would improve topic quality. After evaluating the strengths and weaknesses of different methods for computing topic buzz, Hockett et al. (2018) found that using expert seed words, their synonyms, and similar words to the seed words from a word embedding space, led to the highest topic quality. Hockett et al. (2018) then correlated buzz values for these two topics from over 1 million newspaper articles related to Iraq in 2016 to data from the United Nations International Organization for Migration (IOM) that tracks the number of internally displaced persons (IDPs) in Iraq. This was done to see whether either of these two topics were indirect indicators of possible movement. The research found there was indeed a high correlation between buzz and movement data (a Pearson correlation of 0.76). This high correlation means that buzz has potential to be a reasonable proxy variable for movement. It is a strong indication that using buzz as a leading indirect indicator with other big data generated variables is an important direction for future research.

Events: An *event* is something that happens at a particular time and location, e.g., a bombing in Anbar on 10 January 2015. A targeted event is an event in a particular location that is associated with a particular theme or topic of interest to the user, e.g., politics, violence, football, etc. (Wei et al. 2016). Tracking the frequency of targeted events allows us to compute a time series containing the number of targeted events related to topics correlated to forced migration each day. The frequency of different types of discovered events and the topics associated with these events can themselves also be used as indirect indicators of forced migration. Because of

this, we are also interested in mapping the detected events and their topics to different factors associated with forced migration. This approach allows us to integrate knowledge from interviews with knowledge from open source text data—e.g., newspapers (Wei et al. 2016; Zhao et al. 2017; Wei et al. 2018) and Twitter data (Wei and Singh 2017a; Wei and Singh 2017b)—to gain a more accurate picture of the situation.

Perception: In order to understand whether people will choose to migrate, it is important to understand their perceptions about relevant direct and indirect indicators, e.g., wages, schools, etc. Perceptions can be measured in different ways. Three that are important in the context of migration are tone (sentiment), stance (position), and emotion. An important research direction is to learn to identify tone, stance, and emotion from social media and newspaper content so that perception can be more accurately captured. While a rich body of literature exists for identifying these variables from text, the accuracies for detection still need improvement. Sentiment or tone indicates a global measure of the overall positivity or negativity associated with *how* a document or tweet is written. Tone can be positive, negative, or neutral (see Ribeiro et al. 2016 for a survey of current methods). Our preliminary work suggests that sentiment related to groups that impact migration—e.g., ISIL—changed as different events occurred. What is also evident is that the sentiment related to a similar topic is not always the same in different languages and/or locations (Singh et al. forthcoming).

A variant of sentiment that provides a different form of perception information is stance. Stance is specific to a topic and describes whether the text contains a negative ("anti") or positive ("pro") position towards that topic. In general, there is no guarantee that tone and stance will be exactly correlated. A positive overall tone does not guarantee a positive stance on all topics in the text, and a positive stance towards some topics does not preclude a negative or pessimistic tone. This is why it is important to capture both forms of perception. Work on determining stance from text is in its infancy. Current methods are very similar to those used to determine sentiment (Sobhani et al. 2016; Mohammad, Sobhani, et al. 2016; Mohammad, Kiritchenko et al. 2016). New methods that are able to compute stance with high accuracy using a very small amount of labelled data are needed for dynamic domains like social media.

Finally, emotion considers whether the tweet or article contains emotional content. Researchers are working on identifying a number of different emotions, including happy, sad, relaxed, stressed, and depressed (Canales et al. 2014; Hasan et al. 2014). Similar to sentiment and stance, lexicons containing emotion words and basic machine learning algorithms are currently considered the state of the art. Our team has been able to capture emotion from newspapers and relate it to movement (Agrawal et al. 2016).

An important future direction is to use perception determined from open source data to further investigate dread threat variables on a broader scale. For example, if other sources of information suggest an increase in dread threat levels in Iraq over time, we can determine if that same increase occurs on Twitter. If we are able to map variable values obtained from other sources to variables extracted from tweets, we may be able to further our understanding of the drivers and triggers of forced migration and see the escalation of dread threat levels before large-scale displacement occurs.

Tools and Analytics

As mentioned in the introduction, we must have tools to help policymakers understand the impact of not acting in certain situations. Our research has focused on two interconnected tools. The first provides early warning and the second allows policymakers and practitioners to analyze the evidence and simulate scenarios (see chapter 7 for other online networks that are useful for knowledge dissemination).

Early Warning Tool: An early warning tool should be capable of using indicators drawn from different data sources (many real time sources) within a dynamic theoretical model to alert decision-makers to likely changes in patterns of displacement. In some cases, the displacement will be new, but in many situations, the alert will mark potential shifts in movements. The alert system should go well beyond the binary decision to move or stay. It should seek to provide information to decision makers on who will move (i.e., what are their demographic and socio-economic characteristics), in what numbers, from where, to where. It should also present policymakers with the evidence used to generate the alert and a way for the policymaker to input the strength, reliability, and timeliness

of the evidence, thereby allowing the tool to learn from human analysis of the evidence.

Simulation Tool: Simulation tools can provide decision makers the capacity to test responses to patterns of movements under varying scenarios. For example, if displacement is related to increasingly more severe food insecurity, decision makers could test various scenarios involving the delivery of food to at-risk populations—including purchases of food in neighbouring countries, vouchers to enable people to buy available food, shipment of food from more distant countries, food drops, food distribution in camps, etc. We see two purposes for such scenario testing. First, it helps determine the likely results of a humanitarian action, e.g., what if food relief is dropped at a particular location? Second, it gives insight into determining the likely results of a third-party action, e.g., what if the Jordanian government closes its border with Syria?

One type of analytic tool that can be particularly helpful is a computational simulation that gives practitioners an opportunity to posit a scenario via a web user interface, run the simulation, and view the simulated results though a geographic visualization. A simulation could forecast seven to twenty days ahead, based on what is known and what can be inferred. We anticipate that practitioners who had access to such a system would run many such scenarios each day in order to better understand the scope of what is possible.

At Georgetown, we built a prototype of how local perception of threat in the locality drives actions to mitigate that threat, including both planned migration and unplanned flight. The simulation already developed is based on system dynamics, defined as a "computer-aided approach to policy analysis and design" that "applies to dynamic problems arising in complex social, managerial, economic, or ecological systems" (Systems Dynamics Society n.d.). Simulations based on systems dynamics have several advantages, including ease of development and computational tractability, but also come with limitations on modelling the inherent economic and social diversity of human populations. In effect, systems dynamics models do not necessarily capture decision-making at the household and individual level.

By contrast, agent-based simulation of forced migration allows for modelling each individual household, where a household decides whether and when to migrate, based on its unique assets, location, social

connections, time-varying perception of threat, and other factors (Edwards 2008; Kniveton et al. 2011; Kuznar and Sedlmeyer 2005; Smith 2012). Often lost in this type of analysis, however, are the systems that may facilitate or impede the household from taking certain actions. Finding ways to leverage both models could be valuable for simulating different types of interactions. To date, no full-scale alert system or simulation platform that incorporates either analytic model exists for forced migration.

Challenges and Limitations in Using Big Data

For all the benefits of big data, a number of challenges exist. First, most of these data are noisy and partial. The signal to noise ratio for most topics is very low. Second, the reliability of different sources, and even authors of articles/social media posts, is not clear since real and false information can be shared using these mediums. These data may also have significant biases. Systematic bias is very different from random error and may be hard to identify, much less compensate for. In order to effectively use big data, we must develop methods and tools to quantify and adjust for the variability in reliability and the potential high levels of bias. Third, big data population coverage varies considerably in terms of demographic and movement data. As technology continues to get cheaper and more pervasive, the utility of big data will continue to grow. Next, there is a lack of reliable ground truth data to compare algorithm output to. While there is some knowledge about where and when people move, it is inconsistent, noisy, and not timely. In order to calibrate algorithms and understand their strengths and weaknesses, having ground truth data is important. Finally, it is difficult to integrate large numbers of sources of data that have varying temporal and spatial resolutions. Using time and GPS coordinates is the most straightforward way to combine these data, but using semantic similarity is an important future direction. A large public and/or private initiative that promotes standardization and interoperability across different distributed platforms and entities is an important direction for making traction on these large-scale challenges.

Reliable, accurate, detailed data are fundamental to making progress on this problem. We need as granular and dynamic data as possible in order to identify relevant indicators of forced migration. The scale of migration can significantly redistribute a population, within and across borders,

in very short periods. As "big" as our data sources are now, they do not include information in all of the language groups needed to forecast displacement, nor are the sources sufficiently local (meaning to the community, and even at the household level) to allow us to get at the meso- and micro-level factors influencing movement, particularly in areas where social media penetration is low. Data availability will be vital for making significant progress in this area. We also need to use these data with care, considering anonymization strategies to ensure privacy and developing guidelines for the ethical uses of these personal data (see chapter 13 for more on the ethical dimensions of research on forced migration). While these data can be used for social good, their availability also allows for disruptive forces (Singh 2016). We are particularly concerned that such information could be used to target people, as was done with census data during the Rwandan genocide, or to deter flight even if it is the only way for people to achieve safety. Efforts need to be undertaken to ensure that does not happen.

Conclusion

Making progress on understanding the drivers of forced migration, and developing tools to forecast when, where, how, and who will be displaced, will have a potentially profound impact on understanding and coping with future movements. Early warning holds the potential to save lives and to make humanitarian responses more effective. It would improve planning as well as directly aid potential refugees before, during, and after their exodus. Such planning can lead to action to try to avert mass displacement by addressing the causes of movement, help divert forced migrants from risky modes of movement (e.g., via unseaworthy boats or across landmine infested borders), and enable governments and international organizations to pre-position shelter, food, medicines, and other supplies in areas that are likely to receive large numbers of refugees and displaced persons. Although governments will not always act benevolently in the face of early warning of displacement—such warnings can also give governments more time to stop refugees from crossing onto their territory—the alternative is often chaos, with the displaced and the communities they enter left without adequate assistance or protection. Big data, if integrated responsibly and combined with available administrative data, can be the catalyst for a

timely, reliable early warning system and a mobility simulation platform that identifies likely movement patterns given different policy options.

It is unlikely that further progress will be made in the absence of the two types of collaborations described in this chapter. First, a multidisciplinary approach is essential to answering the core questions arising in the context of early warning—why are people forced to move (or become trapped), what triggers the actual movements, who is likely to become displaced or trapped, and when, where, and how will those who move arrive at their destinations. These decisions are based on a complex mix of political, social, economic, environmental, psychological, and other factors, necessitating the involvement of multiple social science disciplines. Effective early warning requires that computational scientists work closely with their social science colleagues to mine, analyze, and present the data in a practical way. Ensuring that the resulting system is effective requires the active engagement of practitioners throughout the process. As mass displacement is unlikely to reduce significantly in the near future, it is important for humanitarian agencies, researchers, and governments to work together to improve the situation of those forced to migrate through more effective early warning.

Note

We are fortunate to have a large team of contributors. We would like to acknowledge the work of Jeff Collmann, especially for his perspectives on dread threat, as well as Lara Kinne, Nili Yossinger, Abbie Taylor, Yifang Wei, and Chris Kirov at Georgetown University and Susan McGrath and her team at York University. This work was supported in part by the National Science Foundation (NSF) Grant SMA-1338507, the Georgetown University Mass Data Institute (MDI), the John D. and Catherine T. MacArthur Foundation, and the Canadian Social Science and Humanities Research Council (SSHRC). Any opinions, findings, conclusions, and recommendations expressed in this work are those of the authors and do not necessarily reflect the views of NSF, MDI, the MacArthur Foundation, or SSHRC.

References

Agrawal, A., and A. An. 2012. "Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations." *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*—Vol. 1. IEEE Computer Society, Washington, DC.

Armed Conflict Location and Event Dataset. Available at http://www.acleddata.com/ (10/12/2017).

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Blei, D. M., and J. D. Lafferty. "Dynamic Topic Models." 2006. *IEEE International Conference on Machine Learning (ICML).*

Canales, Lea, and Patricio Martínez-Barco. 2014. "Emotion Detection from Text: A Survey." In *Processing in the 5th Information Systems Research Working Days* (JISIC), 37–43.

Churchill, R., L. Singh, and C. Kirov. 2018. "A Temporal Topic Model for Noisy Mediums." *Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD).*

Clark, Lance. 1989. *Early Warning of Refugee Flows.* Washington, DC: Refugee Policy Group.

Collmann, Jeff, Jane Blake, David Bridgeland, Lara Kinne, Nili S. Yossinger, Robin Dillon, Susan Martin, and Kai Zou. 2016. "Measuring the Potential for Mass Displacement in Menacing Contexts." *Journal of Refugee Studies* 29, no. 3: 273–94.

Davenport, Christina A., Will H. Moore, and Steven C. Poe. 2003. "Sometimes You Just Have to Leave: Domestic Threats and Forced Migration, 1964–1989." *International Interactions* 29, no. 1: 27–55.

Edwards, Scott. 2008. "Computational Tools in Predicting and Assessing Forced Migration." *Journal of Refugee Studies* 21, no. 3: 347–59.

FEWS Net: Famine Early Warning System. Available at http://www.fews.net/ (10/12/2017).

Frías-Martínez, Enrique, Graham Williamson, and Vanessa Frías-Martínez. 2011. "An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information." *Proceedings of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust*, 57–64, Boston, MA, Oct 2011.

Government Office for Science. 2011. "Foresight: Migration and Global Environmental Change." Final Project Report. London: Government Office for Science.

Hasan, Maryam, Elke Rundensteiner, and Emmanuel Agu. 2014. "Emotex: Detecting Emotions in Twitter Messages." Paper presented at ASE Big Data/SocialCom/ Cybersecurity Conference, Stanford University, May 2014.

IDMC. 2016. *Global Report on Internal Displacement*. Accessed 12 Oct 2017. http://www.internal-displacement.org/globalreport2016/.

International Crisis Group. Available at https://www.crisisgroup.org/crisiswatch (10/12/2017).

Hockett, J., Y. Lui, Y. Wei, L. Singh, N. Schneider. 2018. "Detecting and Using Buzz from Newspapers to Understand Patterns of Movement."

King, Jordan. 2016. *Methods to Overcome Challenges When Learning Arabic Word Embeddings for Text Mining Tasks*. Undergraduate thesis in computer science, Georgetown University.

Kniveton, Dominic, Chris Smith, and Sheila Wood. 2011. "Agent-Based Model Simulations of Future Changes in Migration Flows for Burkina Faso." Global Environmental Change 21 (Supplement 1): S34–S40.

Kuznar, Lawrence A., and Robert Sedlmeyer. 2005. "Collective Violence in Darfur: An Agent-Based Model." *Mathematical Anthropology and Cultural Theory: An International Journal* 1, no. 4: 1–22.

Martin, Susan, Sanjula Weerasinghe, and Abbie Taylor. 2014. *Migration and Humanitarian Crises: Causes, Consequences and Responses*. New York and London: Routledge.

Massey, Douglas. 1988. "Economic Development and International Migration in Comparative Perspective." *Population and Development Review* 14: 383–413.

Melander, Erik, and Magnus Oberg. 2006. "Time to Go? Duration Dependence in Forced Migration." *International Interactions* 32, no. 2: 129–52.

———. 2007. "The Threat of Violence and Forced Migration: Geographical Scope Trumps Intensity of Fighting." *Civil Wars* 9, no. 2: 156–73.

Menkhaus, Ken. 2010. "Stabilisation and Humanitarian Access in a Collapsed State: The Somali Case." *Disasters* 34: S320–S341.

Mohammad, Saif M., Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. "Semeval-2016 Task 6: Detecting Stance in Tweets." *Proceedings of International Workshop on Semantic Evaluation*, 31–41, San Diego CA, June 2016.

Mohammad, Saif M., Parinaz Sobhani, and Svetlana Kiritchenko. 2016. "Stance and Sentiment in Tweets." ACM Transactions on Embedded Computing Systems, arXiv preprint arXiv:1605.01655.

Moore, Will H., and Stephen M. Shellman. 2004. "Fear of Persecution: A Global Study of Forced Migration, 1952–1995." *Journal of Conflict Resolution* 48, no. 5: 723–45.

Naudé, Wim. 2010. "The Determinants of Migration from Sub-Saharan African Countries." *Journal of African Economies* 19, no. 3: 330–56.

NOAA: National Oceanic and Atmospheric Administration National Tsunami Warning Center. Available at http://ntwc.arh.noaa.gov/ (10/12/2017).

Ribeiro, Filipe N., Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. "SentiBench: A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods." *EPJ Data Science* 5, no. 1: 1–29.

Salehyan, Idean, and Kristian S. Gleditsch. 2006. "Refugees and the Spread of Civil War." *International Organization* 60, no. 2: 335–66.

Schmeidl, Susanne, and J. Craig Jenkins. 1996. "Issues in Quantitative Modelling in the Early Warning of Refugee Migration." *Refuge* 15, no. 4: 4–7.

Schmeidl, Susanne. 1997. "Exploring the Causes of Forced Migration: A Pooled Time-Series Analysis, 1971–1990." *Social Science Quarterly* 78, no. 2: 284–308.

Singh, Lisa. 2016. "Data Ethics: Attaining Personal Privacy on the Web." In *Ethical Reasoning in Big Data: An Exploratory Analysis*, edited by Jeff Collmann and Sorin Adam Matei, 81–90. New York: Springer.

Singh, Lisa, and Raghu Pemmaraju. 2017. "EOS: A Multilingual Text Archive of International Newspapers and Blog Articles." IEEE International Conference on Big Data (BigData), 4835–7. Boston, MA.

Slovic, Paul. 1987. "Perception of Risk." *Science* 236 (April): 280–5.

———. 2000. *The Perception of Risk*. Sterling, VA: Earthscan Publications.

Slovic, Paul, Baruch Fischoff, and Sarah Lichtenstein. 2000. "Facts and Fears: Understanding Perceived Risk." In *The Perception of Risk*, edited by Paul Slovic, 137–53. Sterling, VA: Earthscan Publications.

Slovic, Paul, Howard Kunreuther, and Gilbert F. White. 2000. "Decision Processes, Rationality and Adjustment." In *The Perception of Risk*, edited by Paul Slovic, 1–31. Sterling, VA: Earthscan Publications.

Smith, Christopher D. 2012. "Assessing the Impact of Climate Change upon Migration in Burkina Faso: An Agent-Based Modeling Approach." PhD diss., University of Sussex.

Sobhani, Parinaz, Saif M. Mohammad, and Svetlana Kiritchenko. 2016. "Detecting Stance in Tweets and Analyzing Its Interaction with Sentiment." *Proceedings of *SEM 2016: Fifth Joint Conference on Lexical and Computational Semantics*, 159–69, Berlin, Aug 2016.

Starr, Chauncey. 1969. "Social Benefit Versus Technological Risk." *Science* 165: 1232–8.

Systems Dynamic Society. n.d. *Introduction to System Dynamics*. Accessed 12 Oct 2017. http://www.systemdynamics.org/what-is-s/.

Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101, no. 476: 1566–81.

United Nations High Commissioner for Refugees (UNHCR). 2017. *Figures at a Glance*. Accessed 12 Oct 2017. http://www.unhcr.org/en-us/figures-at-a-glance.html.

UNHCR. 2018a. Global Compact on Refugees. Geneva: UNHCR

———. 2018b. *Global Trends: Forced Displacement in 2017*. Geneva: UNHCR.

Wang, X. and A. McCallum. 2006. "Topics over Time: A Non-Markov Continuous-time Model of Topical Trends." ACM International Conference on Knowledge Discovery and Data Mining (KDD).

Wei, Yifang, Abbie Taylor, Nili S. Yossinger, Eleanor Swingewood, Christopher Cronbaugh, Dennis Quinn, Lisa Singh, Susan F. Martin, Sidney Berkowitz, Jeff Collmann, and Susan McGrath. 2014. "Using Large-Scale Open Source Data to Identify Potential Forced Migration." Presented at and published for the 2014 KDD Workshop on Data Science for Social Good, University of Chicago, Aug 2014.

Wei, Yifang, Lisa Singh, Brian Gallagher, and David Butler. 2016. "Overlapping Target Event and Storyline Detection of Online Newspaper Articles." IEEE International Conference on Data Science and Advanced Analytics, Montreal.

Wei, Y., L. Singh, B. Gallagher, and D. Butler. 2018. "Using Semantic Graphs to Detect Overlapping Target Events and Story Lines from Newspaper Articles. *International Journal of Data Science and Analytics* 5, no. 1: 41–60.

Wei, Y., and L. Singh. 2017. "Location-Based Event Detection Using Geotagged Semantic Graphs." *ACM International Workshop on Mining and Learning with Graphs (MLG) at KDD*, Nova Scotia, Canada.

———. 2017. "Understanding the Impact of Sampling and Noise on Detecting Events Using Twitter." *IEEE International Conference on Big Data (BigData)*, Boston, MA.