Spatial Surrogates to Forecast Social Mobilization and Civil Unrests[1]

Feng Chen[1,2], Jaime Arredondo[5], Rupinder Paul Khandpur[1,2],
Chang-Tien Lu[1,2], David Mares[4], Dipak Gupta[3], and Naren Ramakrishnan[1,2]

[1]Department of Computer Science, Virginia Tech
[2]Discovery Analytics Center, Virginia Tech
[3]Department of Political Science, San Diego State University
[4]Department of Political Science, University of California San Diego
[5]Latin American Studies Program, University of California San Diego

Introduction

Spatial computing is now widely pervasive in the engineering and science disciplines but we argue that there is an even bigger revolution happening in our ability to comprehend human behavior. Modern geo-tagged communication forms such as social media and microblogs are rapidly advancing the methods by which we can comprehend, and even influence, the progression of events as they unfold. The rise of "massive passive" data (e.g., tweets), in particular, has given significant impetus to being able to understand events across the globe.

Two key trends are manifest in the above developments. First, it has become possible to use public-domain, seemingly innocuous, aggregated data, to infer quantitative indicators of population level change. At the same time, as the scope of such inference enlarges, novel computational methods are becoming imperative for fusing data from such high-throughput sources. This position paper argues for a concerted effort to use "spatial surrogates" as an enabling mechanism to model and forecast social mobilization across the world.

Spatial surrogates are data reductions that we can exploit to aid in understanding population-level phenomena. As the name indicates, surrogates are cheap, easy-to-compute, statistics that are correlated with or that precede phenomena of interest. Surrogate modeling is an established practice in numerous domains such as multidisciplinary optimization and economic forecasting, and here we argue for its use in modeling key societal events. For instance, the idea of tracking flu activity geographically using search query data (in Google's FluTrends) is a modern example of knowledge discovery using surrogates. A second example is using spatial luminosity data to quantify economic output of countries [Chen and Nordhaus, 2010]. A final example is using Landscan data as a surrogate for population density.

Social Mobilization

Our domain of interest is social mobilization, i.e., how civilian populations mobilize to raise awareness of key issues or to demand changes in governing or other organizational structures. Protests, strikes and "occupy" events are part of such mobilizations. Such events occur in a variety of political systems, even authoritarian ones. Most of the time governing institutions can ignore, repress, or respond to social mobilization in ways that do not fundamentally challenge public policy and the political institutions that generated those policies. Nevertheless, at times

[1] Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract D12PC00337. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

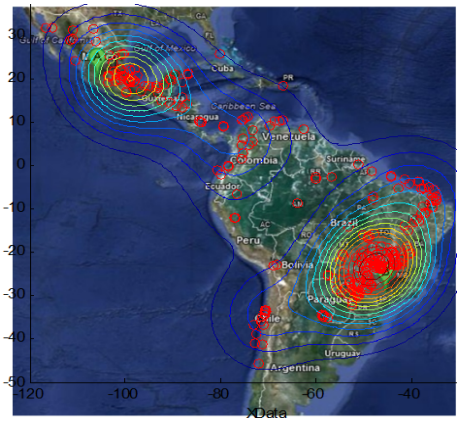
governing institutions, even democratic ones, are overwhelmed by civil unrest generated by significant and repeated protests (e.g., Chile 1970s, Poland 1980s, Bolivia 2000-2005, Arab Spring 2010-present).

Scholars of democracy, policymakers and most social activists are aware that significant levels of civil unrest make politics, economics and social relations difficult and unstable. Any efficient solution to civil unrest has to find ways to channel political power in the streets back into stable, institutionalized channels of interest representation where bargains can be negotiated. This requires understanding when significant social mobilization may occur and when that civil unrest will be of significant size and its likelihood of becoming violent.

Decades of intense social science research have shed important light into our understanding of the root causes of social mobilization and political violence. However, challenge for the scholars and policymakers alike, is the ability to forecast these events. Traditional empirical academic research develops forecasting models based on past data, collected and published by various governmental and non-governmental agencies. Yet, these efforts are time consuming and the resulting inquiries are akin to the astronomers looking at distant galaxies, which contain information about how they used to be rather than what they are now. The monitoring of the Internet and the social media has opened up a brand new area of social inquiry that is unique not only in it aspiration but also its inherent need to bring together scholars from all areas of academia like never before.

Studying and Forecasting Civil Unrests

In general, computational modeling of civil unrests is in its infancy. Recent research [Gonzalez-Bailon, 2011] has focused on how protest recruitment happens through an online network but comparatively little attention has been paid to forecasting civil unrests in society through information gleaned from online, geo-tagged, media.



We have taken some initial steps along this direction by first organizing a dictionary of 726 terms related to protests, and redescribing a geo-tagged tweet stream in terms of frequencies of terms from this dictionary. The objective is to identify anomalous spatial regions based on Poisson mixtures. A linear time subset scan [Neill, 2012] is applied to identify anomalous spatial regions which are then scored using p-values computed by Monte Carlo simulation.

Our work focuses on countries in the Latin American region; the displayed map from June 30, 2012 provides an illustrative example where each (geo-tagged) tweet containing at least one term from our dictionary is plotted as a small red circle. Two anomalous spatial clusters are detected, as shown. One cluster is located in Mexico with 'país', 'trabajador', 'trabaj', 'president', and 'protest' as the top five frequent terms. These refer to the student-led protests that happened during the Mexican election held on July 1, 2012. The second cluster, located in Brazil, involves the high frequency terms: 'país', 'protest', 'empres', 'ciudad', and 'gobiern'. This cluster is related to the situation where approximately 2,500 people closed the Friendship Bridge linking Ciudad del Este (Paraguay) and Foz de Iguazu (Brazil), a demonstration held in support of Paraguay's president Fernando Lugo. Thus, initial results are encouraging.

Research Issues for Discussion

Spatial Surrogates: There are now a significant number of spatial data sources available, especially through the advent of location-based social networks such as Facebook, Twitter, and Foursquare. How can we leverage such a multiplicity of data sources to design accurate spatial surrogates? Although each data source by itself is unlikely to provide the desired specificity, it is possible that combinations of them will yield the desired quality of forecasting.

Machine Learning Models of Spatio-temporal Phenomena: Traditional models of spatial and spatio-temporal phenomena have been prohibitively expensive, e.g., involving the estimation of non-stationary covariance matrices. Modern methods such as the linear time spatial scan [Neill, 2012] promise to usher in significantly more efficient methods for detection. Can we establish an emphasis on both efficient and expressive algorithms for machine learning research?

Spatial Event Forecasting: Most current research focuses on event detection in the form of spatial or temporal bursts or clusters whereas the forecasting of events has not been well studied. However, significant domain knowledge can be harnessed in the form of how mobilization occurs on a spatial or temporal scale. How can machine learning algorithms exploit such prior knowledge effectively for spatial forecasting of civil unrests?

Geolocation: Only a small percentage of communications data harvested from social media are geo-tagged natively but it is possible to envision semi-supervised and transfer learning paradigms that enable a greater variety of data sources to be geo-tagged. What data sources provide corrobative and complementary evidence for geotagging purposes?

Integration of Crowdsourcing and Machine Learning: Concomitant with better geotagging capabilities, it is instructive to examine how a modicum of "active" crowdsourcing can augment "passive" data assimilation, and how such a data gathering loop can be integrated with a machine learning loop. This can yield systems that can systematically increase specificity of modeling by crowdsourcing data gathering in regions of most uncertainty.

Integrated Crisis Management: Finally, integrating the methodologies above for civil unrest modeling can lead to a powerful system for integrated crisis management, one that can quickly disseminate information spatially in the most efficient manner and reduce congestion and overload both in physical and in communication infrastructures.

References

X. Chen and W.D. Nordhaus, Using Luminosity Data as a Proxy for Economic Statistics, *PNAS,* May 16, 2011.

S. Gonzalez-Bailon, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, The Dynamics of Protest Recruitment through an Online Network, *Nature Scientific Reports,* Vol. 1, Article 197, Dec 2011.

D.B. Neill. Fast Subset Scan for Spatial Pattern Detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360, 2012.