2015-12-02

Topics in Complexity: Dynamical Patterns in the Cyberworld

Hong Qi
University of Miami, h.qi@umiami.edu

UNIVERSITY OF MIAMI

TOPICS IN COMPLEXITY: DYNAMICAL PATTERNS IN THE CYBERWORLD

By

Hong Qi

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

December 2015

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

TOPICS IN COMPLEXITY: DYNAMICAL PATTERNS IN THE CYBERWORLD

Hong Qi

Approved:

_____ Neil F. Johnson, Ph.D. Professor of Physics	_____ Joshua Cohn, Ph.D. Professor of Physics
_____ Chaoming Song, Ph.D. Assistant Professor of Physics	_____ Dean of the Graduate School
_____ Stefan Wuchty, Ph.D. Associate Professor of Computer Science	

QI, HONG (Ph.D., Physics)

Topics in Complexity: Dynamical Patterns in the Cyberworld (December 2015)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Neil F. Johnson.
No. of pages in text. (109)

Quantitative understanding of mechanism in complex systems is a common "difficult" problem across many fields such as physical, biological, social and economic sciences. Investigation on underlying dynamics of complex systems and building individual-based models have recently been fueled by big data resulted from advancing information technology. This thesis investigates complex systems in social science, focusing on civil unrests on streets and relevant activities online. Investigation consists of collecting data of unrests from open digital source, featuring dynamical patterns underlying, making predictions and constructing models. A simple law governing the progress of two-sided confrontations is proposed with data of activities at micro-level. Unraveling the connections between activity of organizing online and outburst of unrests on streets gives rise to a further meso-level pattern of human behavior, through which adversarial groups evolve online and hyper-escalate ahead of real-world uprisings. Based on the patterns found, noticeable improvement of prediction of civil unrests is achieved. Meanwhile, novel model created from combination of mobility dynamics in the cyberworld and a traditional contagion model can better capture the characteristics of modern civil unrests and other contagion-like phenomena than the original one.

To my beloved parents for their endless love,

understanding, support and encouragement

Acknowledgements

I would like to write down my sincere gratitude to Professor Neil F. Johnson, who introduced me into the field of complex systems. His advises, guidance and instructions are the key to my career to Ph.D. I am always inspired by his profound knowledge of physics and mathematics and his deep insights of hot topics is the essential direction of my work. During the past three years, he offered me invaluable mentorship, considerate support, intelligent suggestions. It is my most pleasant journey in my life in academia to work with him and learn from him.

I would also like to thank other professors in my research group, Dr. Chaoming Song and Dr. Stefan Wuchty. Their collaboration is also very important to my all works. I get a lot of instructions from them when constructing models and finding approaches to several important problems. Without them, I should not be as productive as I am now.

I would like to extend my gratitude to the Chairman of Physics Department, Dr. Joshua Cohn for his willing to become my committee member and for his help with my education.

And thanks to my other team members: Pedro Manrique, Nicolas Velasquez, Ana Morgenstern, Andrew Gabriel, Minzhang Zheng, Guannan Zhao and all other colleagues, for their warming kindness, valuable discussions and productive collaboration.

Hong Qi

University of Miami

December 2015

Table of Contents

LIST OF FIGURES

LIST OF PAPERS

1. A. P. Morgenstern, N. Velásquez, P. Manrique, Q. Hong, N. Johnson, and N. Johnson, Resource Letter MPCVW-1: Modeling Political Conflict, Violence, and Wars: A Survey, *Am. J. Phys.*, vol. 81, no. 11, p. 805 (2013)

2. P. Manrique, H. Qi, A. Morgenstern, N. Velasquez, T.-C. Lu, and N. Johnson, Context matters: Improving the Sses of Big Data for Forecasting Civil Unrest: Emerging Phenomena and Big Data, *IEEE International Conference on Intelligence and Security Informatics*, pp. 169–172 (2013)

3. N. F. Johnson, P. Medina, G. Zhao, D. S. Messinger, J. Horgan, P. Gill, J. C. Bohorquez, W. Mattson, D. Gangi, H. Qi, P. Manrique, N. Velasquez, A. Morgenstern, E. Restrepo, N. Johnson, M. Spagat, and R. Zarama, "Simple Mathematical Law Benchmarks Human Confrontations," *Sci. Rep*., vol. 3, p. 3463 (2013)

4. N. F. Johnson, G. Zhao, F. Caycedo, P. Manrique, H. Qi, Extreme Alien Light Allows Survival of Terrestrial Bacteria, *Scientific Reports*, vol. 3, p. 2198 (2013)

5. N. F. Johnson, G. Zhao, E. Hunsader, H. Qi, N. Johnson, Abrupt Rise of New Machine Ecology Beyond Human Response Time, *Scientific Reports* vol. 3, p. 2627 (2013). This paper was also featured on CNN, NPR and BBC among others.

6. H. Qi, N. Velasquez, P. Manrique, A. P. Morgenstern, D. Johnson, C. Song, S. Wuchty, and N. F. Johnson, New Online Ecology of Adversarial Aggregates: ISIS and beyond, *Science*, Reviewing 2015.

7. P. D. Manrique, H. Qi & N. Velasquez, Women Emerge as Dominant in Networks under Extreme Pressure, *Science,* Reviewing (2015).

8. P. D. Manrique, H. Qi, M. Zheng, C. Xu, P. M. Hui, and N. F. Johnson, "Anomalous Contagion and Renormalization in Dynamical Networks", Under Writing (2015)

CHAPTER 1 Introduction

1.1 Motivation

Complex Systems are becoming more and more researched in both the natural and social science [1]. And complexity refers to the dynamical patterns in complex systems, which pervades many natural and social fields. For instances, patterns in a sediment bed [2–4], emergence of crescent-shaped dunes [5], growth of river meander loops [6], spread of financial rumors [7], patterns in fatal human confrontations [8] and mass protests on streets [9,10]. The existing definitions of Complex System, however, are still vague, lacking specialty and conciseness. Nonetheless, it is still agreed by many researchers that complexity is a manifestation of a series of key characteristics [5], which imply that there are such things as complex systems distinguishing from simple systems. Though not as unique as those in hard core mathematical systems, the following serve as typical signals of complexity: feedback and adaptation by the exchange of energy, material or information at macroscopic/microscopic level [11], self-assembly among individuals [12], and strong coupling with the environments [13]. Quantitatively understanding the dynamical patterns in complex systems is said to be a remarkable revolution. Because it can lead to valuable predictions for those systems that have been proven difficult to forecast and control so far [1].

Like the many-body problems and subsequent collective phenomena in solid state physics [14], physicists in the Complex System field have emphasized the novel non-linear (e.g. fat-tail distribution) patterns observed in many real-world complex systems, such as animal aggregations (e.g. school of fish) [15], large-scale epidemics [16], massive civil unrests

[17] and global financial markets [18]. These non-conventional patterns can generally be divided into two stages. On the first stage, collective behavior emerges from self-organization that reduces the massive number of degree to a much smaller number of independent dynamical variables [19]. On the second, these variables interact and evolve to yield diverse, potentially emergent, phenomena [20,21]. Good examples are presented in chapter five and six of this thesis. In these two chapters, I show that individuals are self-assembled in online aggregators without evident leaders and the development of the organizations uncover several interesting results, including an innovative transition from unrests online to events on the streets.

Physicists are keen on looking for the complexity patterns and subsequent models of complex systems. In natural science, many theory and models have been successfully built to explain patterns of complex systems [19,22–25]. In social science, however, there are less achievements. The reason why it is so hard for previous investigations in sociology is that only collective statistical data was available before. Nowadays, fortunately, the technology advances bring the entire human into a cyberworld, where tracking individuals' activity and storage of tons of data become practical for common researchers. Good examples like Google trends and Facebook logs can provide very detailed time, location and contents of human activity at individual level [26].

Hence recently quantitative patterns for complex systems in social science and corresponding application have been explored a lot by researchers from physics, computer science, sociology and other disciplines [27–30]. What occur in all investigations, however, is that the such type of patterns is systematically complicated [13], with little resemblance to known elementary laws of dynamics. In comparison to physics, many

theory and models of complex systems (e.g. many-body and non-linear problems) have been successfully built to explain the behaviors of individual particles and macroscopic features, as a consequence of well-investigated microscopic law governing the particles and micro-structures. For example, spin glass [31], superconductivity [32], quantum mechanics [33] and high energy physics [34].

Therefore the work of our group is to investigate complex systems in social science. The focus is on finding some basic patterns which govern the individual behavior at microscopic or mesoscopic level. Furthermore, we try to develop a model with the help of these patterns to mimic specific non-conventional complex systems in social science. These patterns and models in social science are potential extremely beneficial to economy, military and society, just as indicated by works of other group members'. For example, finding dynamical patterns in guilds forming within massive multiplayer online (MMO) game [35] leads to better prediction of life of a game; Discovering underlying pattern in insurgent and terrorist activity is of importance in military to prevent fatality [36,37]; Unveiling the dynamical patterns characterizing the trajectories humans follow daily [38] can contribute a lot to public health [39], city planning [40], traffic engineering [41] and economic forecasting [42]. In addition, the work on civil unrests, introduced in this thesis, can give rise to better prediction of massive protests and thus less violence and loss resulting from the unrests [10,17,43]. The detailed contents and further discussion of the civil unrest work are represented in the following chapters.

1.2 Overview

This research lies on the basis of big datasets of civil unrests which unprecedentedly increase the accessibility to precise and reliable analysis on Complex System on the microscopic level. The content and history of civil unrests are described in chapter 2 that. Understanding the dynamics of groups of people in complex systems and the prediction on collective behavior becomes the focus of this research. At the beginning of this series of projects, our research group was perplexed on how the large systems evolve, how the development of groups is along with time, how rumors/information spread among the systems. As time, energy and curiosity are devoted to the projects, the profiles of the complex systems of civil unrests are gradually emerging to the surface. Finally, the vital question that "can collective behavior be predicted", is touched on and productively addressed by us. Several successful and practical predictions consistent with empirical data are accomplished which will be of great significance in preventing potential conflicts, fierce violence and economic loss.

One basic premise behind the study of complex systems is that local non-trivial interactions yields the complex collective behavior. However, at the beginning stage of projects, the research group had not got a good understanding of characteristics of the local interaction and thus led the study to direct analyzing the macroscopic signals, such as Google Trends on relevant words of protests and category of current signals by machine learning of history. As work proceeding, it was discovered that this direction cannot lead to insights on the underlying dynamics. As a consequence, the group failed to generate reliable predictions that can beat null model of history replication.

The unsuccessful attempts to directly analyze macroscopic features enlightened our team the importance of investigating the fundamental dynamical law governing the behavior at microscopic level. Meanwhile in another project, data of insurgent and terrorists' activity in Mid-East was studied and a simple mathematical law under the base of human confrontations was surprisingly found. With the discovery of this mathematical law, the group proceeded to research the civil unrests in Latin America. A covert development stage of civil unrests, which takes place online and has a non-stationary escalation, are then uncovered. In order to generalize the finding, an innovative two-phase scheme of the civil unrest in cyberworld is provided, with a phase transition governed by an equation identical to the divergence in percolation theory. Subsequently, we have generated successful prediction on the time of possible outburst in Brazil and Venezuela.

Finally, a new model is proposed, capturing the mobility characteristics of individuals from online to offline. It involves a common space of information sharing and a private space where individuals are isolated, mimicking modern people's alternation of online media and offline life in digitally driven world. Therefore, this model can be considered a new type of "contagion" model in which motivated people are the infectives and ideas spreading online are the virus. Additionally, the consistence with empirical data from online aggregators of Islamic State and civil unrests in Latin America provides big support for the reliability of the newly-discovered model.

CHAPTER 2 Modeling Civil Unrests: A General Survey

2.1 Background Overview

The hard core of civil unrests: political violence, civil wars, and terrorism are all complex social problems, which involve dynamical interactions among individuals, groups or communities and evolving process in exogenous factors (terrorism, inter-state wars) as well as endogenous factors (civil wars, social unrest) [9]. In standard social science, the conditions under which protest, wars, terrorism, and other forms of conflicts occur are persistent themes in the field. Even an unequivocal definition of this field still remains elusive, as there are widely different opinions among academics as to which aspects of civil unrest should be the main focus. Many researchers focus on the explanations of the reason and consequences of unrests, while others are focused on the strategies of key leaders. Generally speaking, theoretical approaches emphasize the explanation or prediction to causes of violence by addressing its underlying micro/macro achievements.

Our interest is ignited by the complex nature of eruptions of civil unrests, and the novel approaches from a physics perspective are made practical by growing availability of event-level datasets. Actually, the first ever work aimed at quantitative understanding across conflicts was done about 100 years ago by a physicist, Lewis Fry Richardson [9]. Richardson failed to construct a beautiful quantitative scheme of social unrests in his years. Because he suffered a lot from lacks of availability of event-level datasets, which is now accessible to many investigators.

These big datasets include manually-coded data on civil unrests, such as protests, political conflict and inter-state wars. In existing studies, the data adopted usually come from non-academic sources like newspapers, televisions, blogs and other social media open to public. A considerable amount of work has been done to employ these datasets with technology from agent-based modeling, logistic regression and machine learning [44–48]. However, these research failed to clarify quantitative algorithm or pattern inside the complex civil unrests, and consequently, subsequent predictions have no good performance. As perceived by this research group, an important reason leading to such results is that the big datasets used in previous research contain only collective statistics of the civil unrests, thus the insight on micro-structure of civil unrests are failed to be generated.

Fortunately, updated event-level datasets with higher precision become available to researchers since the onset of bursts of information technology and devices. The advanced information technology (e.g. aggregated data from use of the Internet) provide larger volume of storage, longer time of monitoring, precise geolocations and boosted calculation/analysis. Meanwhile the spread of usage of smart devices, such as cellphone, tablet and laptop, and the access to internet makes it possible to record individual-level activity. Therefore, an increasing number of physicists are being drawn to this field, together with strikingly growing multidisciplinary teams investigating political conflicts.

2.2 Quantitative Approaches to Social Unrests

We focus on the methods and theoretical tools associated with different instances of civil unrest and political conflict. The field is vast and has evolved from many disciplines ranging from political science and economics to sociology—and more recently ecology and biology. As a result, there are several definitions of political conflict; from the field of contentious politics and social movements focusing on protests and strikes, to conflict defined as violence. In addition, the field offers a wide variety of methods to account for collective action; its occurrence, how it spreads, and how it might even be predicted. The sections in this review comprise a discussion of some relevant works on social movements and the studies of diffusion or contagion processes of social movements; studies on illicit organizations such as insurgencies and terrorism; studies on political violence; and finally the analysis of civil wars. In all cases, there are typically extreme asymmetries of power between social actors and the state—however, figuring out the reasons of these asymmetries and bridging the differences is still to be accomplished.

For those studying social movements, the aim of actors is usually policy change or political inclusion and the means are more often than not peaceful [49]. On occasion, the target of claims for social movements is not government but private enterprises or other organizations such as universities. Nonetheless, these interactions for the most part can also be characterized as peaceful; the goals and tactics are also similar. On the other hand, for insurgents and terrorism organizations, the goals are regime change or territorial control and the means by which these goals are to be accomplished are extremely violent. However, in these cases, one still observes a dynamic relationship between the state, social actors, and civil society.

2.3 Diffusion and Contagion of Social Unrests

An interesting feature of social unrest has been its contagion or diffusion. As a group is able to mobilize, episodes of contention tend to spread throughout geographical areas—not inevitably bound by the nation-state— or in temporal spaces. Social movement scholars have long observed that mobilization occurs in waves but the mechanisms by which these waves unfold or develop remains a promising research topic. Analysis of these mechanisms is still a field of interest to many researchers [50,51]. As scholars are concerned to the dynamics of contention, mechanistic explanations are the one that reached general consensus [52,53].

Recently, the use of power laws has become more popular in this field, as many researchers have uncovered that seldom observed events may not be random but in fact follow a power-law distribution. This is an important contribution as it provides grounds for the non-trivial prediction of rarely observed events. In parallel, many of the social science network approaches inspired by Granovetter (Ref.[54]), are now moving towards a dynamic picture of networks. Hence social science is moving towards computational field of agent-based modeling, with the goal of addressing how nodes and links evolve in time and so ultimately unveiling the connection between network structure and function

CHAPTER 3 Context Matters: Improving Forecast Civil Unrest with Big Data

3.1 Perspectives of Civil Unrests Prediction

Civil unrest is a key problem of social science and has complex dynamical patterns in development. Various events of civil unrest have led to significant political, cultural, economic and societal changes to the entire world. Social unrests in recent years, like the Arabic Spring in Mid-East and the rebellion in Syria, have rattled the world. The contagion of civil unrest occurs when stress of the society accumulate gradually, then the outbreak of stress comes out spontaneously in the form of social unrest within a short time scale, spreading from nearest to long-range neighborhood that are also susceptible to the stress [55]. Although they have been a characteristic of state-society relations [56] for a long time, the organization methods, spreading mode and their impact are boosted by technological advances in such a cyber world [57,58]. Despite the exceptional favorable conditions created by information technology for the unrest, open source of data offer a new way to investigate how information flows generate human behavior in such events. Exploring how these data source help predict behavior leading to civil unrest, is currently one of the core questions in exploiting the potential power of Big Data. The analysis in this chapter is based on the data from Google Trends (GT) and the predictions produced by our model were compared to a database of civil unrest events collected by an independent agency authority from several major news sources. Our work mainly emphasizes the advantage of introducing big data to the traditional sociology methods. The database of events were human-coded by the agency authority and we refer to it in the following text as a Gold Standard Report (GSR), which establishes an independent

version of the ground truth or null model. Our prediction from the model will be checked against the GSR.

The current challenge of stepping into the explored terrain "big data" is now not accessing and collecting data/information, but figuring out effective ways to make use of it [59]. Our prediction model is to target the type of events to be forecast within a time-frame in a specific geolocation based on the big data. For this purpose, we firstly classify the data by their source of countries and tune them for idiomatic differences and contextual information. The key algorithm of our model for prediction relies on a search in volume of event-related terms and corresponding momenta.

3.2 Data Characteristics and Processing

The data used in this chapter is from GT aggregated on a weekly basis and the database of events from GSR. As there is a big variation in the volume of cases of complex civil unrest [60], we then feed our model with the data specifically located in Spanish-speaking countries in Latin America.

The GT data which is provided by an independent agency authority contains the volume of search of different words serving as the features of target. Although all countries are Spanish-speaking, there are idiomatic variations regarding to uses and spelling of some words. The GSR provides a sequence of events cases that are regarded as the ground truth of our prediction. Each GSR event has the following key features that will be adopted in this study:

i. category number (e.g. 011 for unrests related to employment)
ii. geolocation of the event, comprising of (country, state, city)
iii. Date of the events

Despite we have the GSR data with a resolution to city, the model of this study runs country by country as the lowest resolution of geolocation. For each country, a list of event-related words are selected manually and the volume of search of enlisted words throughout that country are extracted. The selection of words was conducted by two sociologist: Nicolas Velesquez and Ana Morgenstern. For example, the list of event-related words for Argentina is introduced as "Protesta, Huelga, Manifestacion, Marcha and Paro." As iterated before, although Latin American countries share many common features, the word lists for different countries have necessarily nuances and salience of issues is various. For example, Venezuela were fraught with mobilizations in recent electoral periods, while in Chile the most salient issue was resulted from education reform.

3.3 Warning Generating

3.3.1 Binary Decision Tree: Optimization of Combination of Words

Binary decision tree method is a simple but popular method in binary classification. For each input feature, the decision tree assigns a value threshold to that feature. If the value of specific feature exceeds the threshold, it yields a prediction of +1, otherwise -1. In our case, binary label (+1, -1) is "events will occur" and "events will not occur", while features are the search volume of relevant words in Google. With pre-selected word lists and related data, the model of this study reads the volume of search of all the words, aggregated biweekly. Afterwards, the momenta of each words is calculated. Say if there are n words, $2n$ features will be presented in the model. It is necessary to give two definitions of prediction score, which are commonly used in other fields: term and

threshold. Term refers to the combination of volume of search, whereas threshold is the value above which the terms in model will generate a warning of events regardless the types of them.

The criterion of the prediction is based on the three following parameters:

$$B1 = \frac{TP + TN}{TP + TN + FP + FN} \quad Accuracy$$

$$P1 = \frac{TP}{TP + FN} \quad Sensitivity$$

$$P0 = \frac{TN}{TN + FP} \quad Specificity$$

$$S0 = B1 + P1 + P0 \tag{3.1}$$

In the above equations, *TP, TN, FP* and *FN* are abbreviations for True Positive, True Negative, False Positive and False Negative, respectively. True Positive value has an increment of one under the condition that one warning from model of this research and a real event happened at the same time step, bi-weekly in our case. False Positive means invalid warning, i.e. no event occurring at that time range of the warning. In addition, True Negative refers to the number of false warnings, and the False Negative represents times that event is failed to be predicted.

With selected word lists and data, our model will try out all the combinations of key terms and thresholds for changes in volume for two consecutive time steps above a specific threshold and word pairings. Detailed process are as follows. First, the system runs all combinations, i.e. one key term and one value of threshold, that maximize the score of accuracy. Secondly, the model drops out combinations that are unable to yield maximum sensitivity from the subset. Finally, the subset resulted from the previous steps are concentrated by maximum specificity. In cases the model generates more than one

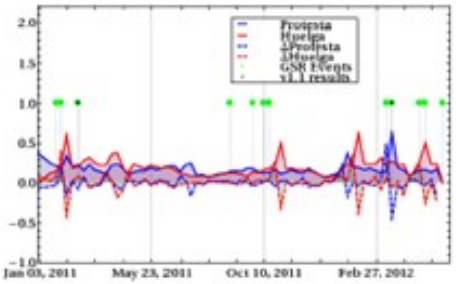
combinations, the first one will be selected. In practical process, we generate parameters pair ($B1$, $P1$, $P0$) for each combination and sort the combination by their $B1$, $P1$ and $P0$ values in literal order.

Additionally, it is found during the process of experimenting that pairing one term with another unrest terms will yield the best results. The pairing means that now there are two terms and two value of thresholds and will yield warnings if and only if the momenta of both terms exceed their thresholds. The pairing will obviously reduce the number of warnings including wrong warnings. Thus, it has a tendency to reduce *TP* but also *FP* and *FN*. As a result, the numerator and denominator of $B1$ are both reduced. The final $B1$, however, has a better results from the pairing according to test.

The Fig.3.1 illustrates the terms volume, real events and warnings from our model along the time axis for Mexico and Venezuela. The warnings, manifested as black dots in the figure, are optimized among a list of words and pairing. Meanwhile, the green dots in the figure stand for GSR events. As explained in the previous subsection, the TP, TN, FP and FN value can be counted in Fig. 3.1. When a black dot appears at the same position where a green dot locates, the TP is increased by one. Similar process can be applied to the other three variables.

The binary decision tree is not static but evolving with time, reading latest events report and yielding prediction daily. Because the optimizing process depends on the "training" of previous events sequence, e.g. the events occurred one year before "today". On a specific date, the tool reads all events one year before that date, optimizes the thresholds, and generate prediction in the next day based on the search volume of words and momenta occurring on this specific date. Hence the model evolves as time flows.

Mexico

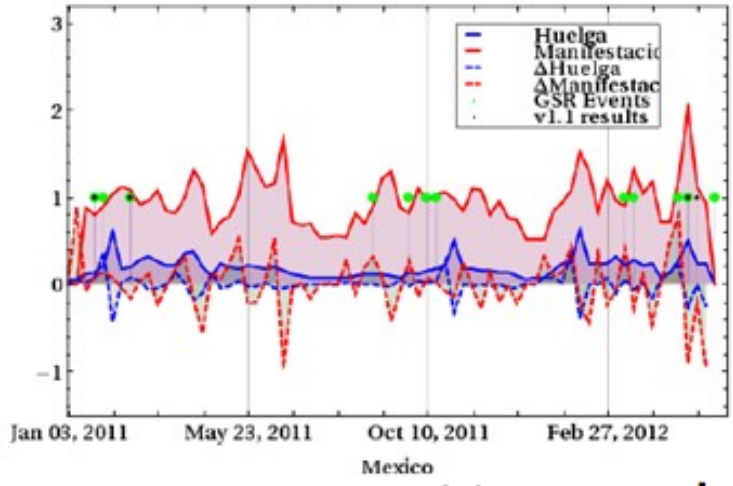


12 GSR events

TP = 2
TN = 59
FP = 0
FN = 10

B1=0.8591
P1=0.1666
P0=1

S0 = B1+P1+P0 = **2.0258**

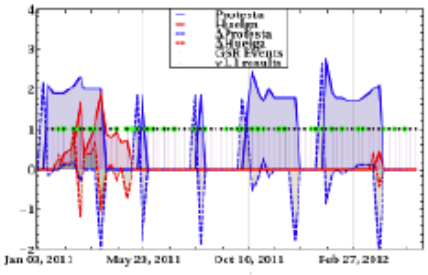


TP = 3
TN = 58
FP = 1
FN = 9

B1 = 0.8591
P1 = 0.25
P0 = 0.9830

S0 = B1+P1+P0 = **2.0922**

Venezuela

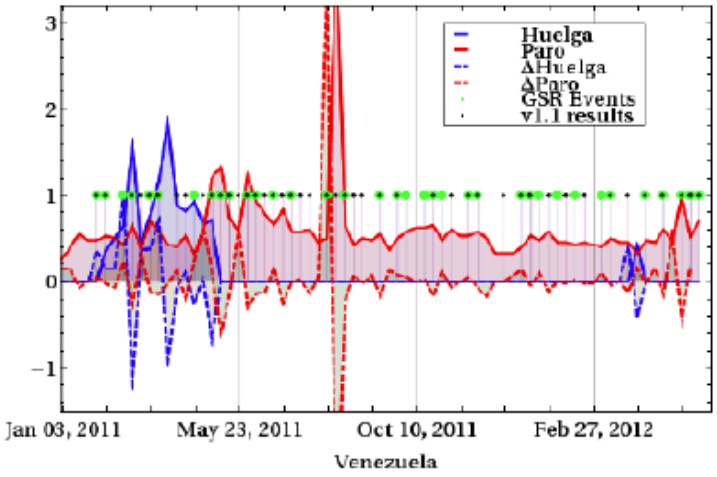


36 GSR events

TP=35
TN=3
FP=32
FN=1

B1=0.5352
P1=0.9722
P0=0.0857

S0 = B1+P1+P0 = **1.5931**



TP = 27
TN = 17
FP = 18
FN = 9

B1 = 0.6197
P1 = 0.75
P0 = 0.4857

S0 = B1+P1+P0 = **1.8554**

Figure 3. 1 Civil Unrest occurrences for Mexico and Venezuela. The cases shown here covers period from January, 2011 to March, 2012. Green dots mean varied types of civil unrest occurrences, and black dots show predictions. A dot both in green and black shows an accurate prediction.

Fig.3.2 shows the result of pairing of terms. Five words (Protesta, Huelga, Manifestacion, Marcha and Paro) for each country has two out of five (equals ten) combinations of words from itself. Each panel of Fig.3.2 shows the performance of our model for one country. Underneath the x-axis shows marks for pair selection. The light blue bars inside Fig. 3.2 are already the results optimized from threshold, meanwhile, the three sub-bars represent the accuracy, sensitivity and specificity respectively. The number inside the yellow box is the $S0$ score of each combination and the maximum of which is marked by green. At last, the pair of optimized terms are written in green and located at the left head of the panel. The optimized pair may not generate the $B1$, $P1$ and $P0$ all at maximum, but it has the best performance in comprehensive score $S0$, as the green number marked.

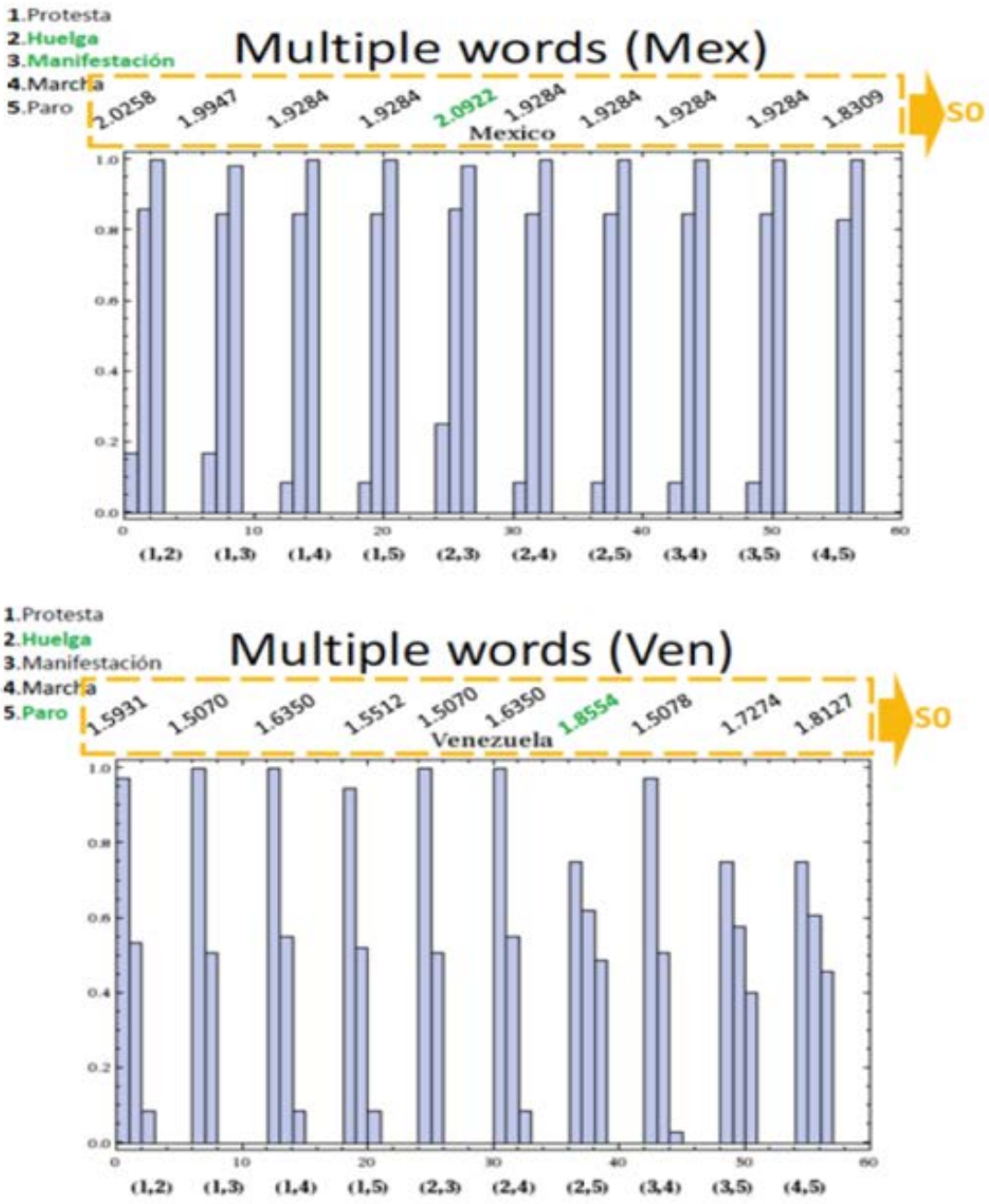


Figure 3. 2 Evaluation of measure's thresholds. Ten groups of bars correspond to all possible combinations. The number in parenthesis underneath x-axis show the index of the pair. And three sub-bars inside each group represent the accuracy, sensitivity and specificity. Green texts show the words used and green number indicates the maximum value of score.

In general, context-specific adjustment will enhance uses of big data for prediction of events. We demonstrate a trial of applying relevant words to the mining of data (GT), and obtain a relatively good score for prediction of events. However, the method in this subsection can only give a binary prediction of events: whether there is an event tomorrow. In order to obtain a prediction of event with probability (e.g. 50% chances to have event tomorrow), a new prediction tool involving the optimizing result of word combination is introduced.

3.3.2 Prediction with Support Vector Machine

In previous subsection 3.3.1, we demonstrate the optimized combination of words as well as corresponding momenta of them. After Support vector machine (SVM) is a machine-learning mechanism of solving pattern recognition problems [61], [62]. In the method, the machine maps the features of data into a high dimensional input space (dimensions equals number of features) and constructs an optimal hyperplane separating those input points in this space [63]. Different optimizing parameters generate various type of core function of SVM. The basic version of SVM deals only with binary label of points in high dimensional space. For example, if we want to recognize the sexuality of one person from his/her other biotic data, features input into the machine can be like height of brow ridge and waist-to-hip ratio, while the label of each point is binary: female or male.

To be more general, the simple binary SVM is a problem in mathematic of finding an optimized p-1 dimensional hyperplane in a p dimensional hyperspace. The criteria of

optima is achieving a minimum sum of geometric margins of the points to that hyperplane. Mathematic form of the optimization algorithm is as followings [63], [64]:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{3.1}$$

In equation 3.1, (x_i) is the vector of feature values and y_i is the binary value for label (+1, or -1). Function $\phi(x)$, which is usually linear function, rescales the value of features and yield the coordinates in the hyperspace. Vector w and γ define the hyperplane, while w is called weighting vector. Variable ξ_i is called slack variable and is able to eliminate the influence of outliers collaborating with constant C. Slack variable and constant C is tuned to fit the data systematically well, reducing influence of several outliers. As indicated in figure 3.3, points in a hyper-space are labelled by two color: black and blue and many hyperplane in this space can divide the points into two sets by three hyperplanes L_1, L_2 and L_3, among which L_3 is the one that SVM seeks for. That is why the optimized hyperplane is usually called boundary. Among all points, nearest points from the boundary are referred to as support vector.

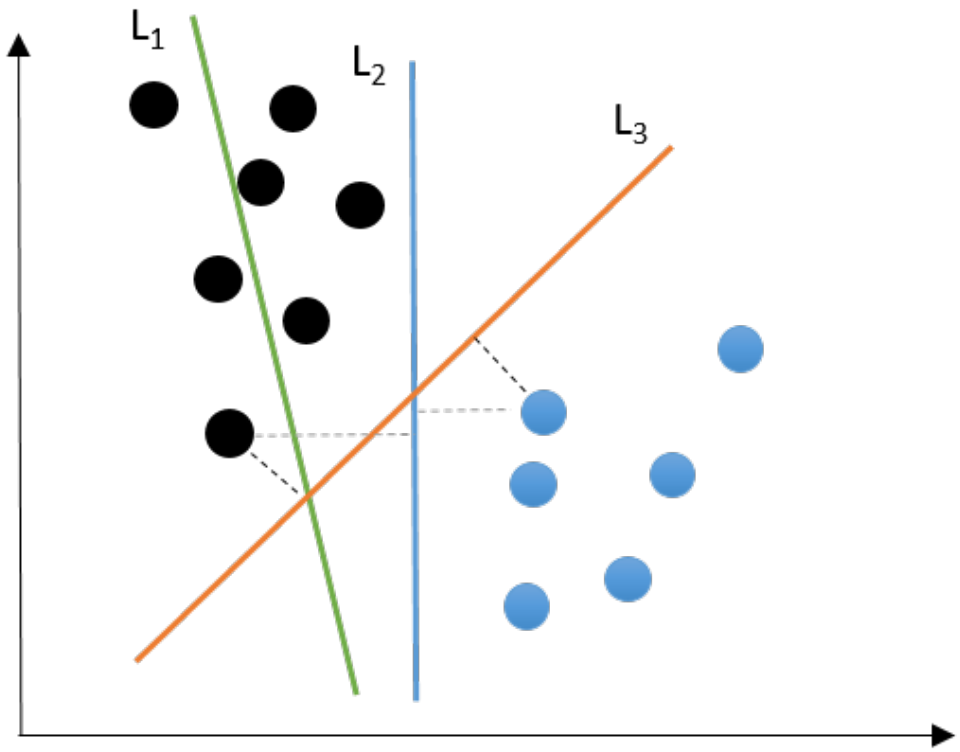


Figure 3. 3 Scheme of Classification in SVM. Hyperplane L_1 does not separate the black and blue points correctly. Hyperplanes L_2 and L_3 both separate the points, but sum of square of distance from the points to L_3 (black dashed line) is less than the one to L_2 (grey dashed line)

For the results below in this chapter, LIBSVM, developed by Prof. Chih-Chung Chang and Chil-Jen Lin, is employed. Its detailed mathematic process including slack variables tuning, weight vector calculation and probability assessment can be found in Ref. [64]. The LIBSVM 2.2 with interface in Python was downloaded at "https://www.csie.ntu.edu.tw/~cjlin/libsvm/#download", and used to predict the probability of warnings for those civil unrest in Latin America contained in GSR.

To find the optimized hyperplane yielding least sum of margin, different tools of SVM have their own mathematic algorithm of parameters tuning to have best performance. In our work, slack variable and constant *C* are tuned automatically by the software

LIBSVM, details in Ref.[65]. After calculation of the hyperplane, the model is trained by all input data of features. Then the model is able to give out whether a new point in the hyperspace fall into domain of either +1 or -1. Also the model can produce a probability associated with this classification, based on the margin from the new point to the hyperplane. When the margin of new point is larger than the one of support vector, the probability is usually one. The assessment methods of probability on the margin are different for different realization of SVM.

Below is the flow diagram of SVM prediction on civil unrest events in GSR. Firstly features of point – status of social unrest on each date – are determined by the result from previous subsection 3.3.1. For example, when training with events history in Venezuela, combination of search volume of “Huelga” and “Paro” (Green words in Fig. 3.2) as well as their momenta serves as the coordinates of point in hyperspace, i.e. all components in vector X. The subscript i denotes the time of the input vector in unit of one day. Sequence $\{y_i\}$ is binary variable: +1 for events occurring on that day, otherwise -1.

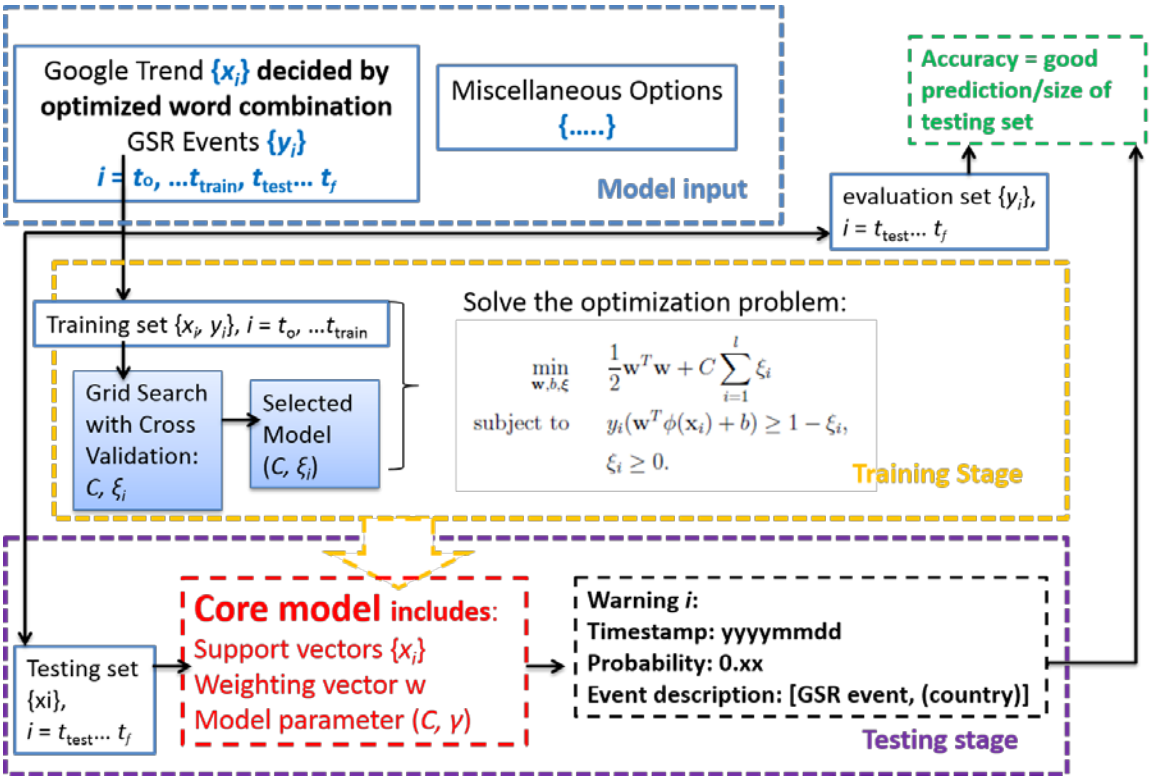


Figure 3. 4 Flow Diagram of Prediction with SVM. Input google trend $\{x_i\}$ is provided by the optimized combination of terms. Miscellaneous Options include the range of parameters searching, upper-limit of trial time, resolution of grid search and etc.

Firstly, the time sequence is divided into two parts of "train" and "test", the prior of which is input into the LIBSVM for training the model. LIBSVM employs a method calling grid search with cross validation to search the parameter of C and ξ_i [65]. There are also miscellaneous options include the range of parameters searching, upper-limit of trial time, resolution of grid search and etc. SVM generates a series of probability that there is an event happening on each date. However, a threshold of probability should be defined to decide whether a warning of civil unrest will be sent out, required by the funding agency IARPA. The threshold, of course, is optimized to achieve best performance score $B0$ from training part. That is all the process of training part.

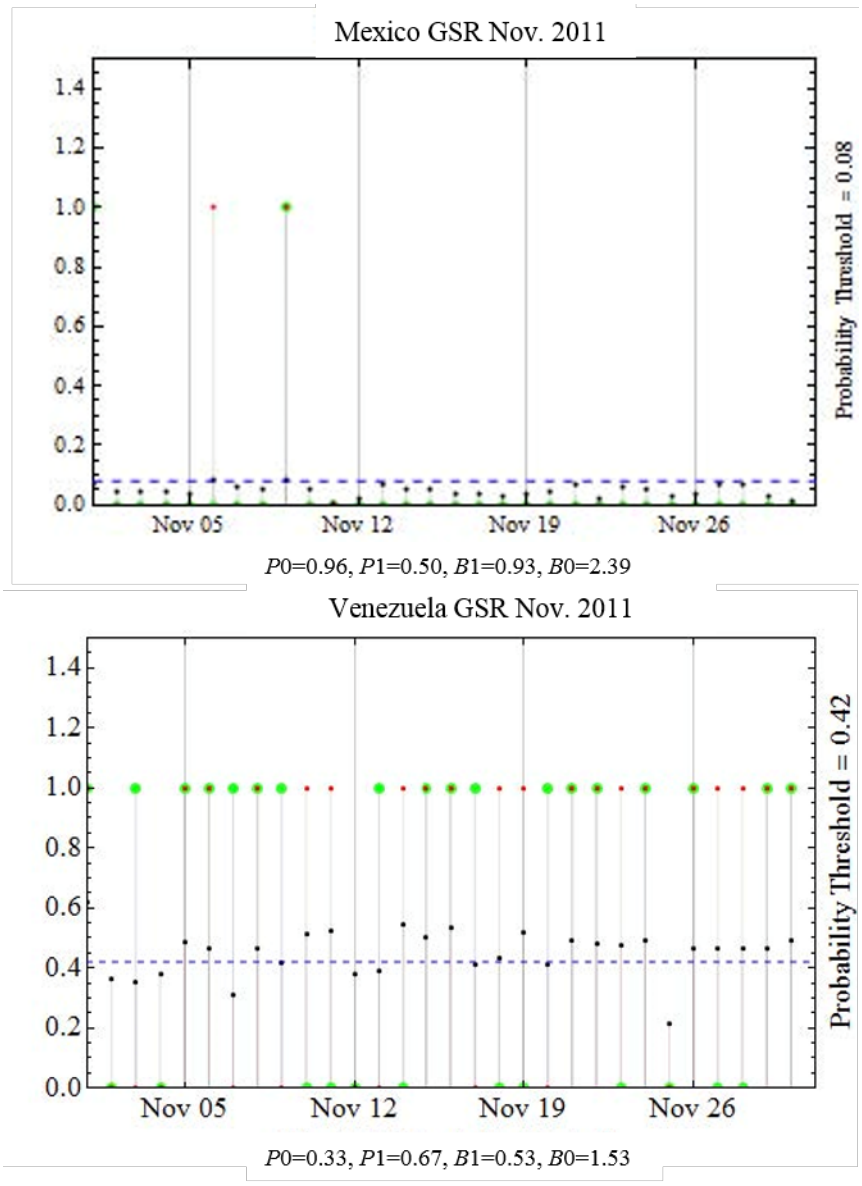


Figure 3. 5 Prediction from SVM. Different from method in subsection 3.3.1, SVM gives probability of events happening and a threshold can be made. Terms for Mexico are "manifestacion" and "huelga" as well as their momenta. Terms for Venezuela are "huelga" and "paro" as well as their momenta. Warning is omitted only when the probability is higher than the threshold. Black dots mark the probability of events happening on that day, green dots mark real event and red dots mark the warning we send out.

Secondly, the test part is input into the trained SVM model to generate warnings and corresponding probability, as illustrated by figure 3.5. The data used are the events in GSR and Google Trend of optimized combination of words in November 2011. We assess the prediction with identical score criteria in previous subsection 3.3.1 (accuracy, sensitivity and specificity). Additionally, a probability of events happening is assigned to that warning.

3.4 Results and Summary

In this chapter, a trial to forecast civil unrest events is made by two methods: binary decision tree and support vector machine (SVM). The prior method tests all the combination of relevant words to achieve a relatively good score for predictions. Thus the binary decision tree method provides the optimized words combination for each country. Furthermore, a SVM model is built in the hyperspace of the optimized words combination in order to generate probability for each warning. According to the score of $B0$, these two methods have similar output performance.

However, the key question is still unsolved that when there will be an outburst of unrests, because the two methods both generate warnings only one day ahead. With a deeper insight into the algorithm of both methods, it is the nature of Google Trend that impedes predictions earlier than several days. The search volume of words reflects only the activity of people at present time. It provides little information about the development stage of those events. Therefore, a new project is conducted with different sight of angle, seeking for the dynamical pattern of development of those civil unrests.

CHAPTER 4 Human Confrontations Benchmarked by Simple Mathematic Law

4.1 One Type of Human Confrontations: Red-Blue Conflict

Human confrontations, causing many high-profile societal problems, involve an individual or group repeatedly attacking another – from one-to-one fights [39,61,62] through to collective civil unrests [63], mass violence [64] and even online aggregations [65]. These detrimental phenomena currently has caught the interest of many sociologists; however, the understanding of the dynamics at the event-by-event level remains uncertain [66]. For example, in the scenarios introduced in last chapter: the protestors' repeated attacks against a government or a child's incessant cry-attack against the parent [62,67,68], each side, usually referred as "Red" and "Blue", is engaged in a complex conflict of adaptation and counter-adaptation.

While 'big data' has shed a light to approaches to non-confrontational human activity [53,69–72], one characteristic of the confrontations obstructs copying those methods to confrontations: agility and secrecy can boost the ability to launch attacks [73,74]. And as indicated by Ref.[75], incomplete or biased analysis of event records for a particular confrontation can hardly be avoided, due to the presence of aggression and danger. These consideration makes us motivated to collect and analyze a broad spectrum of heterogeneous events data, from independent sources across multiple disciplines [76–78]. Our data sources are listed in the Appendix 4.1, ranging from local to global geolocations and in both real and cyber world. The distribution of the severity of events and the trend in the timing of events, are characterized by heavy-tailed distributions, which can be found in many complex systems. For instance, Zipf's law in describing the frequency and occurrence of words [79], Pareto's law in describing the wealth of nations [80] and more

recent scale-free topologies distributions of complex networks [81]. Such distributions is typically attributed to the rich-gets-richer principle according to the latest opinions, since models growing with some degree of preference are usually the first approach to simulate heavy-tailed distributions [82,83].

In accordance to the rich-gets-richer principle, the expected growth speed of a firm, group or social activity is proportional to its size by Gibrat's law [42,84–86]. However, in spite of some exceptions [26,53,87], existing research generally paid less attention to the time evolution of complex systems. Analysis in this study uncovers a simple mathematical law as a good benchmark of the rich-gets-richer principle, specifically on the intervals of events in developing stage. This simple law can contribute to the need to quantify the likely severity and timing of human confrontations and shed light on likely perpetrators [88].

In this chapter, three major contents are presented by the research group: the combined analysis of multiple datasets, the simple mathematical law that can benchmark them all, as well as the derivation of this benchmark and its interpretation. Quantitative predictions of future attacks can be provided by this simple mathematical law of progress curve, thus it is helpful to develop a tool detecting common perpetrators and abnormal behaviors [88]. Further work on this mathematical law is still undergoing within this research group led by Dr. Johnson, and hopefully it will be the starting point for cross-disciplinary theorized aggression of human confrontations at the individual and group level in both real and cyber worlds.

4.2 Data Characteristics and Preprocess

4.2.1 Data for Infant-Parent Interaction

Data for the infant-parent dyads comes from the Department of Psychology in University of Miami, obtained by Dr. Daniel Messinger's group from their laboratory experiments. A brief description of the data set is provided here. The data consists of several subjects and each of them is performed separately with different time and prohibited inter-participant communication to assure the independence of data derived from subjects. An event is taken into account as the infant presents a cry-face and regarded as ending when the infant stops that cry-face. Each subject has one unique infant-parent dyad where the parent is instructed to temporarily keep still-face. the cry-face is identified by the modern recognition technology of human facial expression, while the intervals between successive infant cry-face attacks is recorded by digital monitoring in a unit of millisecond. For details of the experiment, including the recognition technology, information of dyad, structure of the data and etc. Details please consult Cohn and Ekas's articles in Ref. [89],[67]

In this experiment, the baby side is referred as "Red" while the parent side as "Blue". Baby attacks the parent by sporadically presenting a cry-face to obtain the care of its parent, in other words, to protest the lack of interaction with its parent under the circumstance that parent-baby interaction is refrained by researchers. One point worth mentioning here is that the attacking baby may not be looking at the parent's face nor knowing the reaction of the parent. Thus it is called unconscious attacks that the Red is unable to know the reaction of its counterpart, i.e. Blue. In the study introduced in this

chapter, both types of attacks are treated and recorded as events equally. In further projects, attacks evidently targeting the Blue are filtered and only this type of events are recorded for the analysis in the next steps.

4.2.2 Data for Flash Trading

Flash trading refers to the events in stock market that predatory high-frequency traders (Red) abruptly offend slower market participants (Blue), the durations of which are far beyond the normal reaction time of human (<650ms for price dips, <950ms for spikes of price) [18]. Each subject of the data comes from a particular U.S. financial institution's stock. The time range of this dataset is in the period prior to the 2008 global financial crash, which evokes a lot of interest of people. Terms as "dip" or "spike" has various definitions since it is not a completely objective feature. In economics and finance field, popular definition of such term is the value of stock price with Z-score larger than 3.

This dataset can be explicitly found in charts on NANEX website: www.nanex.net. And we are extremely grateful to Eric Hunsader of NANEX for his help with this data. And the steps to access the data manually on NANEX website is:

1. Go to http://www.nanex.net
2. Click on "Nanex Research"
3. In column "Research" on left hand side, click "Micro Flash Crashes"
4. This web page contains all the events implemented in the study. http://www.nanex.net/FlashCrashEquities/FlashCrashAnalysis_Equitites.html
5. Simply download the files, unzip and start viewing

4.2.3 Analysis and Results

For wide range of human confrontations, the time consumed to finish a challenging project decreases with successive repetitions, following an special heavy-tailed distribution – power law progress curve [90–94]. Thus, analyzing dataset of this study in a similar way is highly inspired. The scheme of analysis method and the results for the interval of individual events involving face-to-face confrontations are illustrated in Figure 4.1 and 4.2, respectively. The successive interval points from each subject, i.e. an infant-parent dyad or a financial institution, are fitted into a power-law equation as shown in Fig.4.1 (B)

$$\tau_n = \tau_1 n^{-\beta} \tag{4.1}$$

where the τ_n is the interval between the (n-1)th and nth events, as illustrated in Figure 4.1 (A). As briefly mentioned in subsection 4.2.2 here: the events for infant-parent data are defined as the baby showing a cry-face recognized by software; meanwhile, for ultra-fast trading data, the meaning of events are the operationalized by price dip/spikes whose durations are far beyond the normal reaction time of human (<650ms for dips, <950ms for spikes) [18]. With respect to the interval τ_n between two successive events, it is counted from the end of previous event and the beginning time of the next event, and n is its ordinal index, e.g. interval between first and second events is called τ_1 . Parameter β is the escalation coefficient, which is the results from the power-law curve fitting tool in software *OpenOffice.org Calc* manufactured by *Apache Inc* [8], with logarithm base as 10.

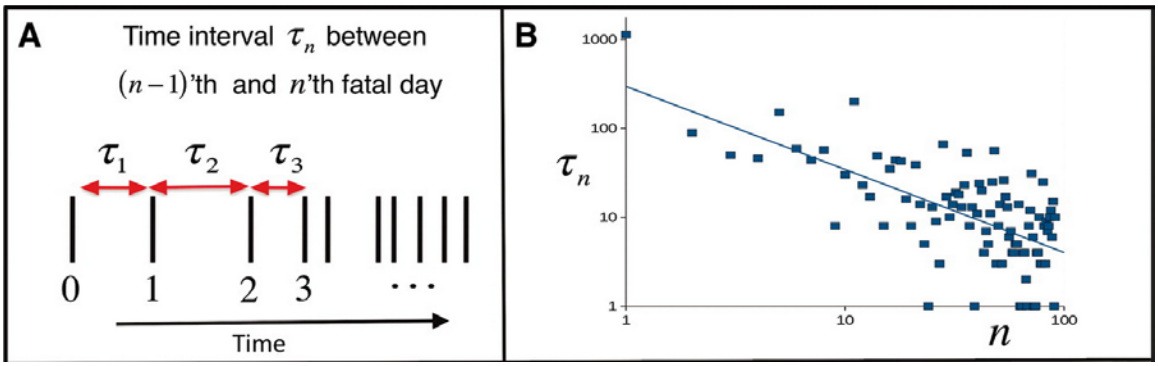


Figure 4. 1 from Ref. [8] (A) Scheme of the interval between successive events in time sequence. (B) Scheme of the power law fit of τ_n vs. n in log-log scale, the intercept of the fitted straight line is the value of τ_1 and slope of it is negative β**.**

For each timeline of the events, a pair value of τ_1 and β can be captured from fitting trend in inter-event times. Since there is a set of subjects in dataset of infant-parent and the ultra-fast trading, multiple feature points ($\log\tau_1$, β) for either set is able to be obtained. The points from infant-parent and ultra-fast trading are plotted in Figure 4.2 (A) and Figure 4.2 (B), respectively. Each point corresponds to one timeline of events from one subject, most of which show escalation ($\beta>0$) with some de-escalation ($\beta<0$). Non-stationary behavior, rather than how, why or when each individual confrontations ends, is the concentration of this research. Therefore, the relationship between the escalation coefficient (non-stationary process) and the initial performance (time from first event to

second event) is carefully investigated. This gives rise to the simple mathematical law that benchmark the pattern of human confrontations, as indicated in the chapter name.

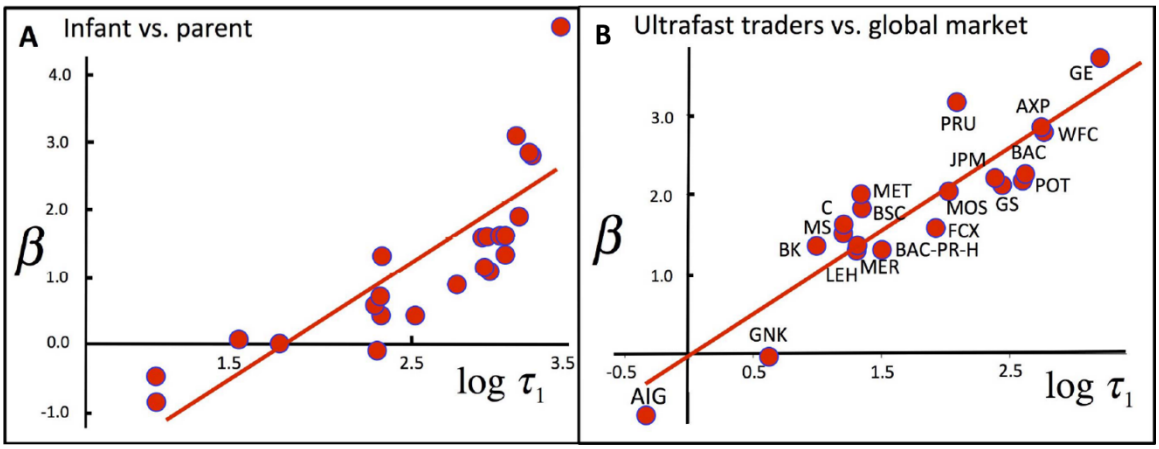


Figure 4. 2 Event-timing benchmark across domains from Ref. [88] (A) Each point denotes a unique dyad of infant-parent, obtained from the progress curve features of itself. Underlying events are infant (Red) attacks parent (Blue) with cry-face. The details of experiment is described in Ref.[67]. (B) Each point denotes a particular U.S. financial institution stock, obtained from NANEX. Underlying events are ultrafast predatory traders (Red) attacks the market of slower investors (Blue). Both panels show a good linear relationship among the points, with R^2 larger than 0.8

In Figure 4.2, simple linear regression is applied to estimate the coefficients and goodness of fit between the variables of β and log τ_1. The R^2 of them are 0.74 and 0.80, respectively. Moreover, the progress curves of non-stationary confrontations are from different subject in various regions, e.g. dyad of infant-parent are randomly selected and financial institutions is throughout the U.S. Therefore, characteristics of their progress curve can be assumed as independent from each other. However, in Figure 4.2, the feature points from the same dataset forms an approximate linear relationship between the escalation coefficient β and τ_1. This fact implies that within each domain the confrontations with initial events tend to slow down their development speed with a high

frequency over time (β even negative when τ_1 is small), whereas the events sequence will accelerate if initial events started slowly (β large positive when τ_1 is large).

Based on the fact that the subjects in either dataset are relatively independent from each other, the simple linear relationship between the features of progress curve β and log (τ_1) is particularly remarkable, especially for the infant-parent dyad. Due to possible cooperation and communications among the financial institutions, the simple mathematical law may be attributed to some incorporating strategy in business. The good fit of linear regression among dyads, however, highlights the existence of a universal simple law within human confrontations as the dyad is comprised of different and random infant and parent. In addition, the parents in experiments come from various locations and races [67], which reduces possible interference among them. Therefore a null model hypothesis is produced in the next subsection, involving a Monte-Carlo process in which the interval between Red attacks are in Poisson distribution.

4.2.4 Null Model and Hypothesis Test

In this subsection, the null hypothesis test of the argument of that simple mathematical law is shown in 4.2.3. The null statement is that the linear relationship between points in Figure 4.2 (A) and (B) are yielded by randomness. Hence stochasticity (randomness) to the event times is to be added and the probability of a good linear-dependence happening from a Poisson distribution will be calculated:

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{4.2}$$

The steps are as follows:

a) For each dyad i, calculate the total times T_i as the time spanning the whole events sequence.

b) For that dyad i, record the number of events n_i and divide it by T_i . This is the average time of intervals between two successive events. According the statistics knowledge, the average of intervals is an unbiased estimator of the expected value of interval for Poisson Process, which equals the scale parameter λ in Poisson distribution.

c) Generate a random series of intervals $\{\tau\}$ by a random number generator in Matlab 2013a that are in an exponential distribution. The scale parameter λ is in the same value as the one in part (b). And the number of intervals for each dyad should be the same with n_i. Hence the synthetic time-series produced by the random generators are obtained, with restriction of having the identical number of events and also the correct duration on average.

d) For each sequence of intervals, fit them with the same software tool to the same power law equations in section 4.2.3. And each sequence thus gives rise to a point of (β, log τ_1). Repeat part a), b), c) and d) for all the dyads in order to get a bunch of simulated points of (β, log τ_1). And plot them down.

e) Calculate the R^2 of the best linear regression of the points in part d).

f) Part (a) – (e) comprise a round of simulation in Monte Carlo process. We need to generate a lot of rounds with different seed of the random machine, to have a statistical view of the value of R^2

g) Calculate the mean and standard deviation for the R^2 distribution obtained in last step. Empirically, the distribution is approximately Gaussian in our test.

h) Assuming that the distribution of R^2 as Gaussian, we can estimate the *p*-value of obtaining a value of R^2 not less than the empirical value given by real data of infant-parent dyads.

The hypothesis test results for the two domains we discuss in this chapter are as follows:

Infant-Parent	$p = 0.0089$	$R^2_{real} = 0.74$	Mean $R^2_{rand} = 0.36$
Ultrafast Trading	$p = 0.0087$	$R^2_{real} = 0.80$	Mean $R^2_{rand} = 0.45$

The comparison between the R^2 from real world and Monte Carlo Simulation shows that the linear-dependence that benchmark the human confrontations can only be resulted from random process with a probability not larger than 0.01 (*p*-value). Hence we can reject the null hypothesis at the threshold of 1%, which is good enough to prove the existence of the universal mathematical law underlying human confrontations.

Besides the infant-parent and ultra-fast trading, other types of confrontations are also been investigate in this research, such as protests versus government in Colombia, fatal death of insurgence in Mid-East, and social movements in Ireland etc. All of them represent a similar mathematical law benchmarking all timing of events from independent groups, institutions, members and so on. More examples and details can be found in the original paper *Simple mathematical laws benchmarks human confrontations* [88].

4.2.5 Linear Dependence for Timing is Non-trivial

In this subsection, it is demonstrated that the benchmark of linear-dependence for timing of events is only valid in confrontations where face-to-face attacks (Red to Blue) occur. This mathematical law is particularly non-applicable to benchmarking data from real world experiments in which active Blue opponent is not available.

It is necessary to recall the main characteristics of human confrontations discussed in the above subsections to have a comparison. All of them should contain the conflict between two sides: Red and Blue and the conflict involves adaption process in both sides. There is, however, other kinds of human tasks in which participants are facing an effectively static Blue side or even a one-sided task. Thus it is argued that there is no need to deal with the counterattacks from the opponents (Blue), or an evolving Red side when engaged in such activity, such as proof-reading, solving a puzzle, or purchasing something online.

As shown in figure 4.3 below, it is recalled in panel (A) the scheme of power law fit for escalation rate and intercept. Figure 4.3 (B) below shows a table summarizing the classic results of escalation rate (β) and intercept ($\log\tau_1$) when completing given type of task addressed by Crossman [95]. The weak dependence between these two parameters is possibly due to the heterogeneity of human behaviors (Red). In figure 4.3 (C), it is presented that this lack of linear dependence is also true for other kinds of human behaviors, like passing cyber tasks, the navigation of different websites. The data in panel (C) is from Ref.[96], and the entire compiled figure is from the Supporting Information of Ref. [88]. Additionally, the escalation rate is referred to by its common alias – escalation parameter in following chapters.

The picture indicates that human (Red) completing one-sided tasks without an active/adaptive opponent (Blue) tend not to have systematic linear dependence between escalation rate and the intercept, compared to confrontations described in figure 4.1 and 4.2. Hence it becomes evident that benchmark is not applicable to a general category of completing tasks, but more likely to be valid to human confrontation with two active and competing sides. Human activity involving static opponent or one-sided tasks is prone to lack this systematic linear dependence.

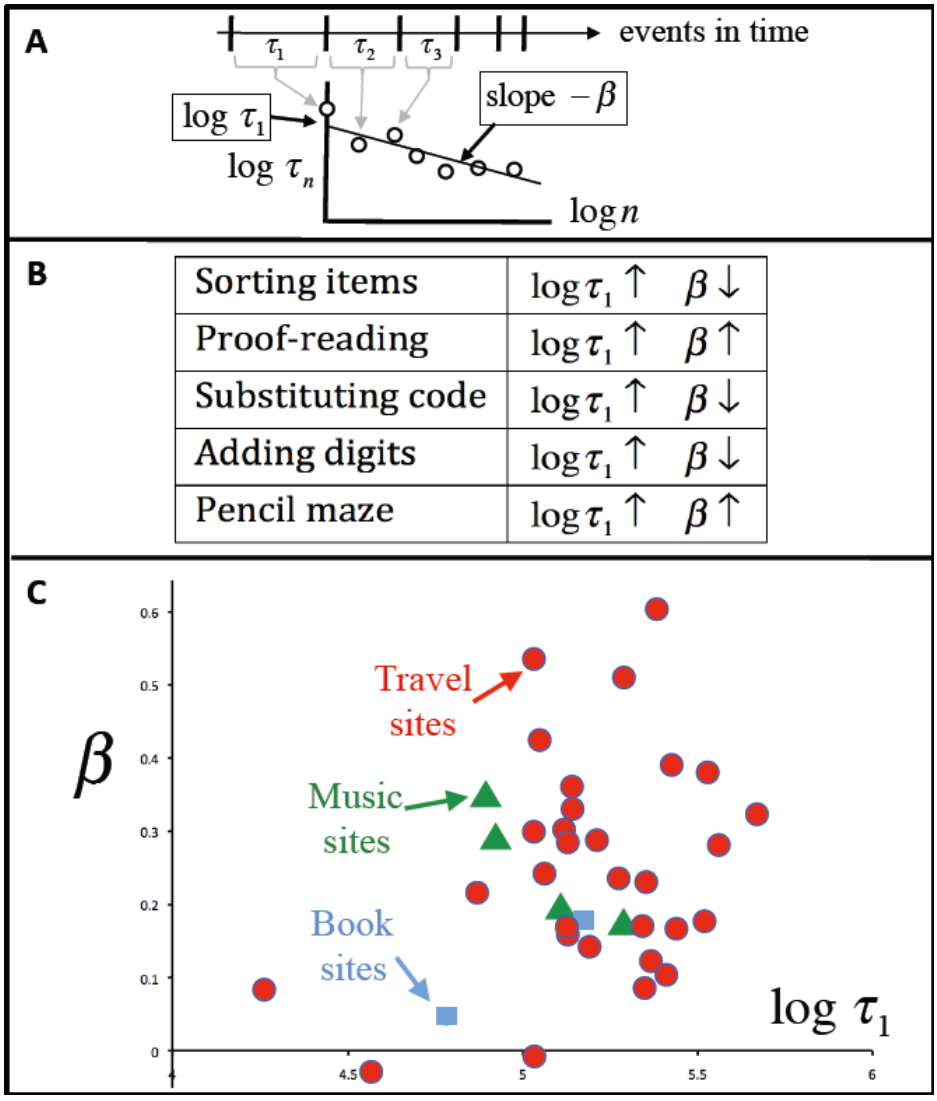


Figure 4. 3 from SI of Ref.[88] (A) Scheme of the fitting procedure described in section 4.2.3. The slope is the negative value of escalation rate. (B) Existing empirical results in

the literature for passive tasks. There is no systematic linear relationship found for two sided confrontation, in contrast to figure 4.2. (C) Results for one-sided activity of searching internet sites. Still there is no evident linear dependence, in contrast to Figs 4.2. Data sources are both given in above texts.

4.2.6 Demonstration of Prediction of Future Events

In this subsection, the accuracy of out-of-sample predictions will be introduced with the timings benchmark in this chapter for future fatal attacks on Blue, i.e. coalition military in Afghanistan. This part is majorly done by Dr. Johnson and Dr. Guannan Zhao, but it is important to verify the significance of the mathematical law.

The sole input is the time intervals data τ_1 between the first two attacks recorded in a given region X (e.g. Kabul) that was previously peaceful. The previous stability in tested region is an important control variable to exclude the interference from previous attacks and to assure recorded τ_1 is the real first interval. In the absence of any benchmark, it would be impossible to forecast the future time intervals of attacks since one point is insufficient to prediction with regression. However, simple linear relationship between β and $\log\tau_1$ throughout all regions that have had attacks can be applied, as long as assuming that the Red underlying the attacks in these regions are the homogenous. Obviously, the data of region X is excluded in the linear regression for intercept and slope since it is the target of prediction test.

The next step is to estimate the linear relationship among the points of (β,$\log\tau_1$) from other regions, especially the slope ($-\beta$) and intercept ($\log\tau_1$), just like in Figure 4.3 (A). So we then use the actual interval between first and second events as an estimate of the intercept, and then calculate the escalation rate β from the linear relationship. Afterwards, the interval time of future events can be predicted using the progress curve (Equation.

4.1), as illustrated in Figure 4.4 (b). And hence a predicted time line of future attacks can be obtained, shown as orange dots and line in Fig.4.4 (a), (c) and (d). The comparison between real timeline and predicted timeline for three regions is illustrated in figure 4.4 (from SI of Ref.[88]) below.

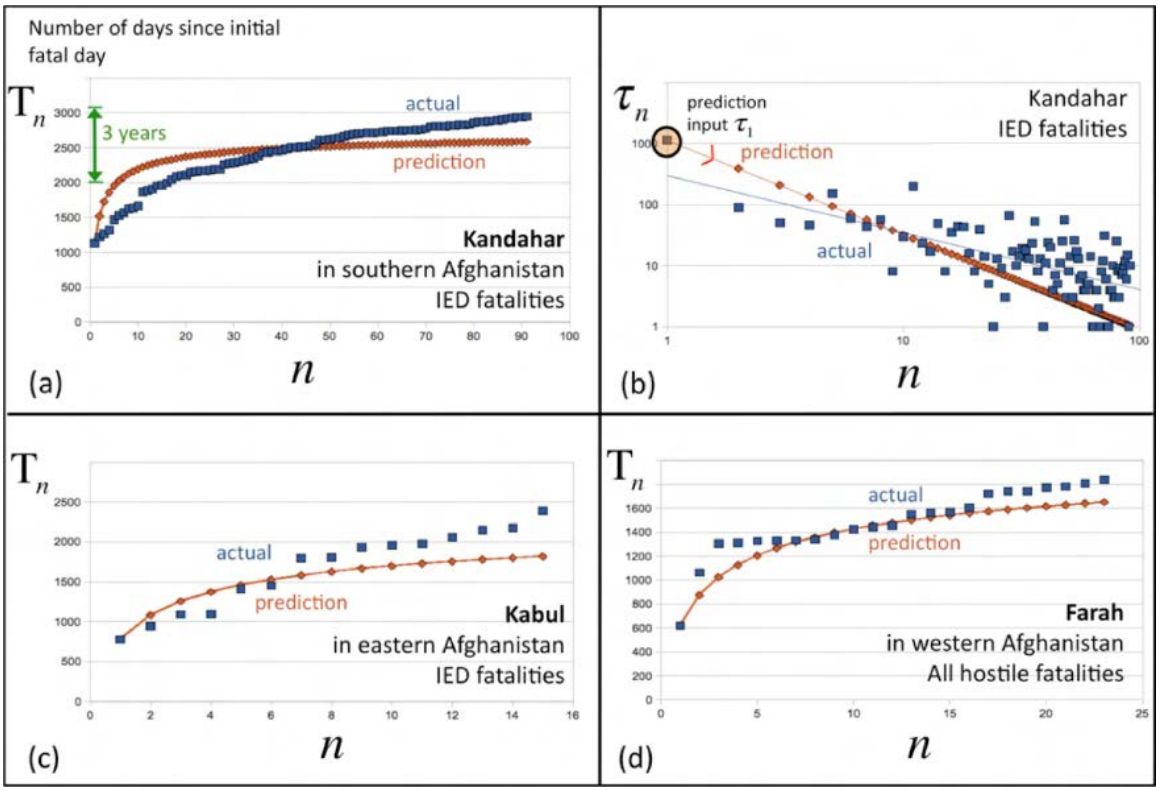


Figure 4. 4 from SI of Ref.[88] Results for prediction of time of attacks in the future, using only one point as input (i.e. initial time-interval). (a) Successive attacks in Kandahar against coalition forces (Blue dots and line). Symbol n is the number of events and T_n is the total time counted from first event. Orange dots and line are from prediction. (b) Plot of intervals between successive events with n, in log-log scale. Blue dots are from real events captured in all regions except the test region X, and the blue straight line is the fitted line. The Orange dots are predicted τ_n from input intercept $\log\tau_1$ and calculated slope as negative β. (c) Similar results for Kabul. (d) Similar results for Farah.

4.3 Summary and Discussion

In this chapter, the focus of discussion is on the confrontations that follow the benchmark behavior featured by a linear dependence between escalation rate and logarithm of initial interval of events. They generally consist of a side of Red actors (e.g. ultra-fast predatory traders, infant) who is in principle faint in strength compared to their Blue counterpart (respectively, global stock holders, parents). Nonetheless, the Red side manages to launch up a series of escalating (escalation rate β) sporadic attacks. Taking dyad of infant-parent and ultra-fast trading as examples, it is shown that both the severities and the timings of events follow a power-law functional form. The model developed in this chapter provides an explanation of the underlying dependence of escalation rate on the state of initial events, which is hardly found in conflict with static opponent (Blue) or one-sided actions.

Next, consider there are several events of a confrontation that is known sporadic and involving active Red-Blue sides, as well as a series of events initiated by the same Red side in other places, institutions or times. Fitting the events series with power law equation (4.1), a bunch of escalation rate (β) and intercept of logarithm of first interval (τ_1) can be obtained. The linear regression of those points, together with the interval of real events, can generate the predicted time of next events in the future. Applying it to the insurgencies data in Afghanistan, a relatively good prediction result is generated and the forecast application of the model is well-evident.

CHAPTER 5 Facebook Activity Leading to Real Unrest in Brazil

5.1 Modern Civil Unrests: Relying More on the Cyberworld

In Chapter 4, the mathematical law that benchmark the human confrontations with two active sides (Red and Blue) is elaborated, as well as its preliminary function of prediction. In this chapter, the concentration is on a further development of the benchmark law, which can handle dynamics of escalation and forecast the phase transition stage of evolving. After a through interpretation of the benchmark law, this paper will illustrate its remarkable application using the example of a real social unrest event occurred in Brazil, 2013.

5.1.1 Digitally Driven Uprisings Rattling the World

As widely-recognized, numerous unexpected social unrests have rattled the world in recent years, some of which lead to tremendous society, political, humanism and economical disasters [97–99]. Among the worst of them are the Arabic Spring, rebellion in Syria and Libya, and uprisings in Latin America. Researchers suspected that advanced information technology (e.g. modern social media) helps accelerate and amplify the idea spreading, protests planning and mass mobilization of people during the entire period of unrest [57,58]. For instance, in 2009 a new phrase 'Twitter revolution' was created to describe mass mobilizations driven by social media in Iran [58] and Moldova [57] in response to the scandal of fraudulent election.

From 2009 and later on, the method of aggregating people using online approaches such as social media has widely been adopted by various organizations. In 2010, the

revolutionary movement called Arab spring spreading in the Arab world involved a huge volume of online streaming of information. During that event, people sent messages of demonstrations, protests, riots and civil wars via Facebook, Twitter, BBS and other digital social media. Similarly, social media played a key role in uprisings in Egypt and Tunisia as well, where 90% claimed that they organized protests and spread awareness via Facebook [100]. Meanwhile, 30% indicated that blocking Facebook by government significantly damped or disrupted their communications and activities [100].

5.1.2 Data Mining of Social Media

The key feature of digitally enabled social unrests is the unprecedented huge stream of information. News from mainstream sources can now be accessed and analyzed globally. Many governments offer free online publications. Social media also present an inclination of booming, for example, there are more than 500 million tweets posted on Twitter every day, with the volume notably increasing [101]. Similarly, people can easily share messages, have discussion, spread video and speeches on Blogs and online forums. With respect to this tremendous stream of information, researchers who are interested in these unrest started to develop statistical and machine learning methods, aiming at mining information stream and thus predicting future events as well as the collective behavior of people [101–104]. Despite the important role the social media played in uprisings, the mechanisms underlying are unexplored, especially how it helped protestors to mobilize and leaded to a critical mass of outburst. Although results are not yet fully demonstrated, most researchers, including this research team, believe that those digitally driven events are in a way endogenous to the system and a prediction method can be achieved [104].

5.1.3 Prevalent Phase Transitions in Real Complex Systems

Phase transition stage is another intriguing aspect among mining of the digital information. It is becoming increasingly clear that lots of complex systems have a phenomenon of so-called tipping points, a critical threshold where the system shifts from one state to another (static or non-static) [105], such as abrupt shifts in global/local climate and ocean circulation in the Earth system [106], catastrophic shifts in fish populations in ecosystem [107,108]. This phenomenon is also applicable in global finance market to the worldwide concern about flashing and systematic crashes arises after several crisis in recent years [109,110]. In biology and medicine, there are spontaneous systemic failures like asthma attacks [111] or epileptic seizures [112,113]. Even in some cyclic and chaotic systems, some transitions usually associated with different classes of bifurcation can still be found [114]. Notably, it is difficult to predict such critical transitions because the model of complex systems are usually not accurate enough to forecast reliably where and when critical thresholds may occur [105]. However, as shown in many other researchers, the dynamics of complex systems near a tipping point have generic properties, regardless of the details of each system [115]. Therefore, it is highly encouraging to find a criterion of the sharp transitions in a range of complex systems, i.e. the early warning signal about a catastrophic transition.

5.1.4 Precursors of Transitions from Dynamics

The capability to forecast the abrupt shift, i.e. phase transition in dynamics, is particularly intriguing to scientists in many fields. As a result, there have been numerous

investigations about the precursors of transitions, most of which are based on the analysis of dynamics of complex system. For instance, critical slowing down' is the most important indicator of whether a system is approaching a critical threshold [116]. The phenomenon of it usually results in three possible warnings in the dynamics of a system which is getting close to the tipping point: slower recovery from perturbations, increased autocorrelations and increased variance [105]. In addition, the asymmetry of fluctuations may occur and become increasingly larger as getting close to a catastrophic transition, which leads to the skewness and flickering before transitions [117]. The skewness of the distribution of states, however, is not expected to happen only approaching transition point, but also occur when the system is driven closer to the basin boundary by an increasing perturbation [117]. Besides, particular spatial patterns can also serve as early-warning signals [118,119], especially in the field of sociology.

These new observations of precursors of transition are all notably meaningful, however, this chapter presented an innovative two-phase paradigm for those uprisings driven by digital technology, featuring an online gestational phase that precedes any uprisings on street by several months. To prepare for the practical basis of investigation of the underlying mechanism, detailed data of uprising events on the streets of Brazil is collected and obtained through a systematic press review of Brazil's main media outlets. Two group members, Nicolas Velasquez and Ana Morgenstern, monitored the mainstream press's coverage of date, geolocation and approximate size of significant demonstrations, and unrests and/or protests in Brazilian main cities from 2011 to 2013. The details of peer review and information extraction will be shown in next sections. Remarkably, within this time range this paradigm managed to observe an outburst of civil

unrest events that came out in many cities of Brazil in June 2013 simultaneously. These outburst is called “Brazil Winter” by researchers, and dynamics of online groups’ development in this series of event comprise the main body of the first phase where organizations evolve online.

5.2 Data Collection and Preprocess

5.2.1 Social Unrest Data and Weakness of Epidemic Model

In order to investigate the underlying mechanisms that lead to transition from online unrest to social uprisings, an independent source of data set of social unrest events is employed from MITRE, a non-profit research centers sponsored by the federal government. The organizations of MITRE collect most of the reports about civil unrest events in Latin America from mainstream press reports and list the earliest appearance of an unrest event in their “Gold Standard Report” (GSR) [26,120]. In particular, some reliable news media outlets such as O’Globo, BBC and other national/ international sources were harnessed to identify time and location of protests that were manually double-checked by subject matter experts (mainly by Nicolas and Ana) [17]. The report is provided via the third party of HRL Laboratories, LLC, located in Malibu, California. The whole data is originally the part of a project aiming at forecasting future unrests by data of online media, funded by an independent agency authorities IARPA. This major project is called open source indicator (OSI), aiming at establishing a series of forecasting tools based on open source information online.

Furthermore, unrest data with the number of protesters, as reported by local press, is augmented by this research group via manually checking. Meanwhile, more details are added to the GSR report as shown in Figure 5.1:

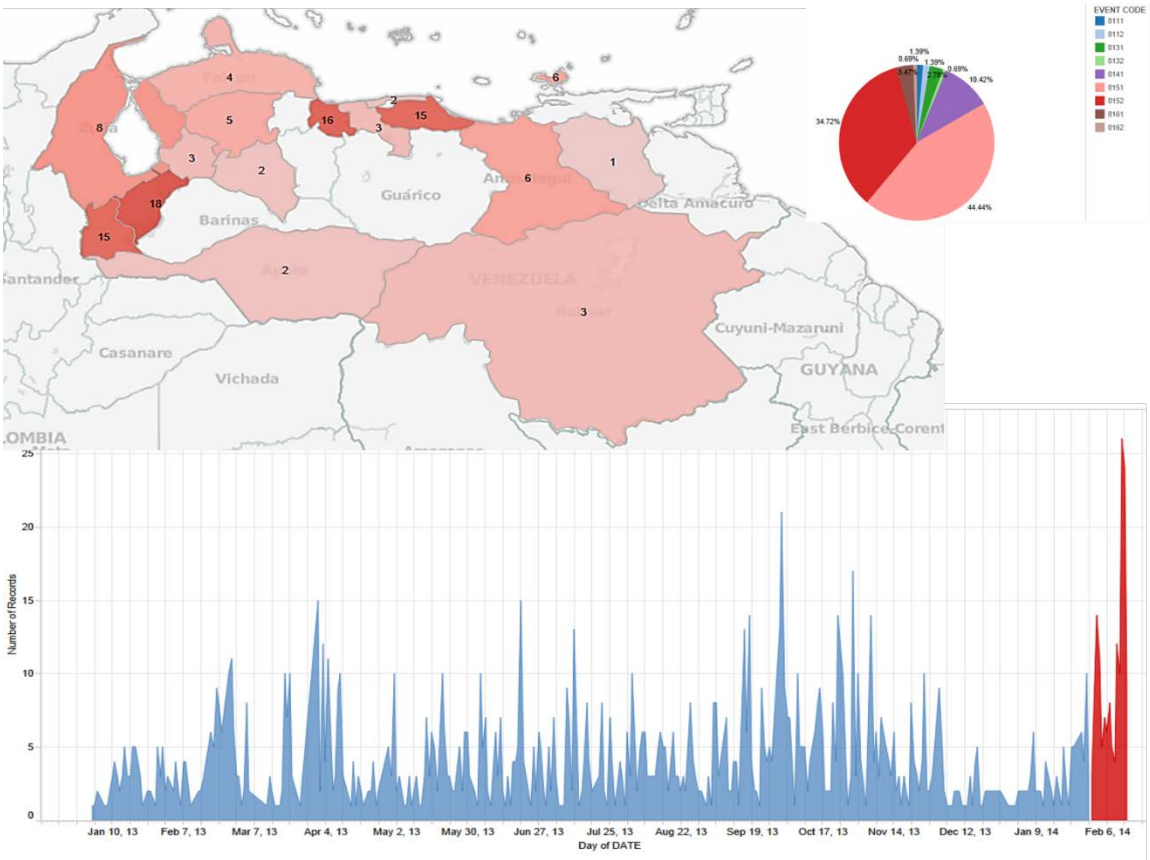


Figure 5. 1 GSR events in Venezuela from Jan.2013 to Feb.2014. The upper panel shows the accumulated number of events in every states of Venezuela. The darker of the color, the more events there are. The bottom panel is the trend curve of GSR events along with time axis, among which GSR activity in the month of February is marked in Red. The inset in the upper-right corner is the pie plot of the percentage of sub-codes of GSR: 011, 012, 013 …

Focusing on the GSR events in “Brazil Winter”, this research group have generated an illustration of events in Brazil during 2013 as shown in Figure 5.2. The bottom panel shows the time trend of GSR around “Brazil Winter”, with vertical horizon denoting the

number of events on that day. In the figure, it can be observed that there is no evident escalation of events on the real world ahead of the spike covering June 2013.

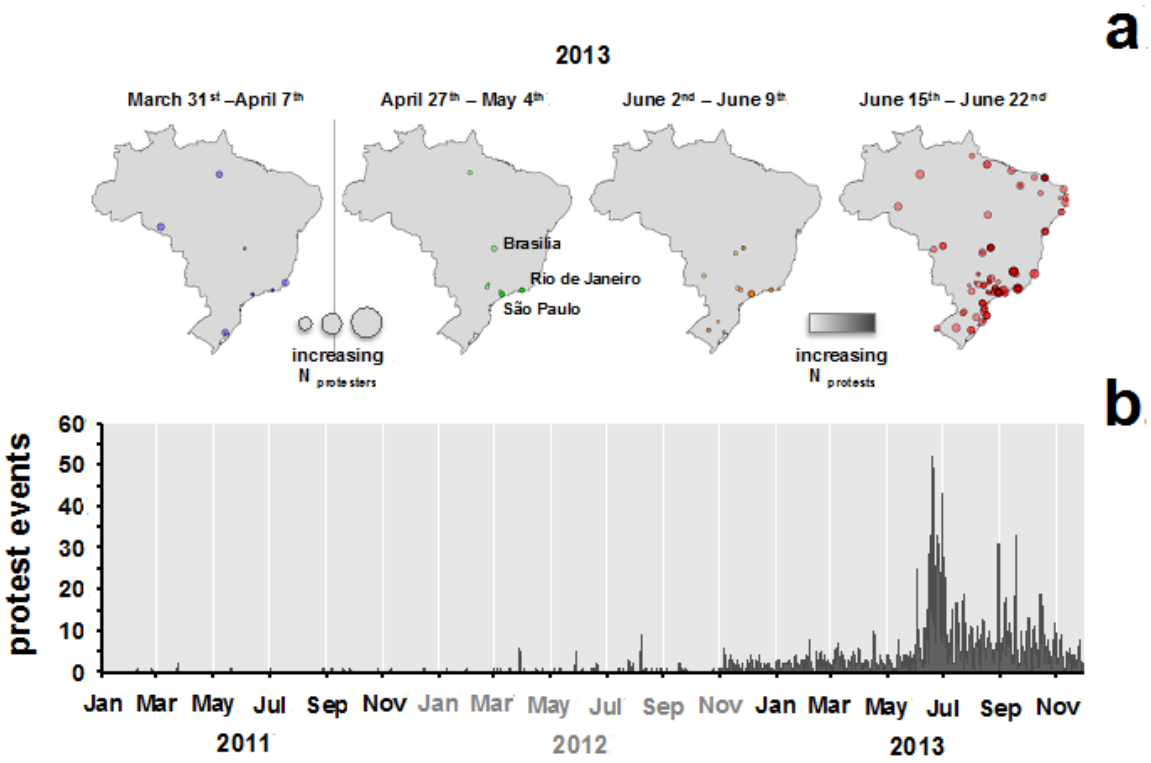


Figure 5. 2 Burst of social unrest on the streets of Brazil. (a) Snapshots of protest events in Brazil from March to June 2013 indicate limited localized activity in large cities, before the number of unrest events as well as their attendance drastically increase all over Brazil in mid-June of 2013. (b) The number of all events along with time line in Brazil, from 2011 to 2013. There is no noticeable precursor signal prior to the large burst of social unrest activity in mid-June of 2013, from the events on streets.

The absence of noticeable precursor in the GSR profile around the "Brazil Winter" implies that the traditional models of contagion among events may not be accountable to the process from build-up to transition and then to outburst. At the same time, however, analysis of social media sites required by that project (OSI) has indicated that spikes of online interactions can precede major events in a revolutionary process [121]. Therefore, it is encouraging to hinge the unexpected and volatile social uprisings on a convert build-up

in social media, specifically in a country like Brazil where there are highly developed digital infrastructure and devices.

Although most of the previous research assume macroscopic events as the consequence of individual action, just like the epidemic models, the potential microscopic mechanisms across social media is considered in this study. In particular, online interaction between individuals allowing them to switch opinions through pairwise social relationships should be taken into account. As mentioned in Ref. [122], there are three weak points of epidemic model when applied to social structure and information spreading online. First, the epidemic models implicitly fail to capture key features of this kind of highly non-stationary process, in spite of their ability to capture general patterns of social influence that govern dynamics of information spreading. Second, the structure of underlying social relationships among individuals online may vary significantly prior to an uprising, while the models of epidemic type adopt a relatively constant information space. Furthermore, with respect to such models, there is an absence of awareness of the existing collective dynamics between multiple individuals, whereas they have strong emphasis on the role in interactions between individual actors. The collective dynamics can support the emergence of mesoscopic groups during the uprising process, as indicated by Amir Ali [97].

5.2.2 Collection of Information of Organizations on Facebook

Accurate and detailed structural evolutions of social interactions online and the formations of virtual groups should be explored and investigated before analysis of potential mesoscopic dynamics. Hence, a long-term plan is set up to monitor the

emergence of organizations and communities. Self-developed script in Python implementing the Facebook API has provided the automatic engine to pull data and store it. Facilitated by some manual corrections, this research group have built up a dynamical software tool to collect the essential information about online groups.

As shown in Figure 5.3, many ad-hoc organizations have various pages open to public on Facebook, probably to attract followers. They are transient entities where people who may have never met are self-organizing around causes and issues. Although the term "organization" is usually the choice of groups on Facebook profile, they are actually self-organized clusters consisting of users/followers whose membership fluctuates over time [17]. Similar patterns are found in the dynamics of re-Tweets of organizations' messages, which is in the build-up to a transition [76]. In this research, Facebook profiles of online organizations with a political agenda are identified first, then those not active in coordinating protests for certain period are filtered. Later, this research group queried the Facebook platform with the API offered by Facebook to obtain the complete records of those profiles. Through this process, essential features of the group can be recognized, including the title, organizations name, and date of creation, mission and ideals, the number of likes, events created/co-created and geographical location in some cases. Further polishing work has to be done in order to get a more accurate and cleaner database of the activity of online organizations. More details and relevant discussions can be found in the subsection 5.5 in this chapter.

Figure 5. 3 Example of a typical Facebook organization, from SI of Ref.[17]**.** Pages of ad-hoc organizations are invariably publicly accessible, which is probably due to the fact that they desire to attract followers and participants in future planned events. In this way, we have managed to build up a dynamical picture of the creation and evolution of Facebook 'organizations' prior to visible appearance of new episodes of instability and unrest, as indicated in Figure 5.4. Since the structure of Facebook pages is the same for all users, our methodology is in principle applicable to any city, country, region or

language. Upper panel is the whole home page of an opposites of government. And in bottom panel, left and right parts show two examples of features recognizing and extracting.

5.3 Analysis of Underlying Dynamics and Result Presented

5.3.1 Non-Stationary Escalation of Organization Creation

As indicated in the above two sections, the emergence, development and evolution of organizations relevant to civil unrest activity is considered as a potential precursor signal source which can be tapped to predict the date of possible transition to outburst of events. Based on the data collection work done in section 5.2, the timeline of creation of such organizations on Facebook is observed and plotted in Figure 5.4 (A). Noticeable escalation of such kind of activity quantitatively plotted in Figure 5.4 (B) is found in the lead-up to the onset of civil unrest outburst in Brazil. The observed pattern of "seismic" activity online mimics Moore's Law of development, which is consistent with both theory of organizational development and contentious politics [123]. Particularly for quantitative analysis of escalation, we employ the progress curve equation 4.1 introduced in previous chapters: $\tau_n = \tau_1 n^{-b}$, capturing the basic fact that the time intervals between creation date of organizations are systematically decreasing. The definition and content of each variable is reiterated: τ_n is the time interval between two successive creation of organizations within a specific time window which are assigned an ordinal number n and n+1, respectively. τ_1 is the interval between first and second creations within the time window, and n is the ordinal label of the τ_n .

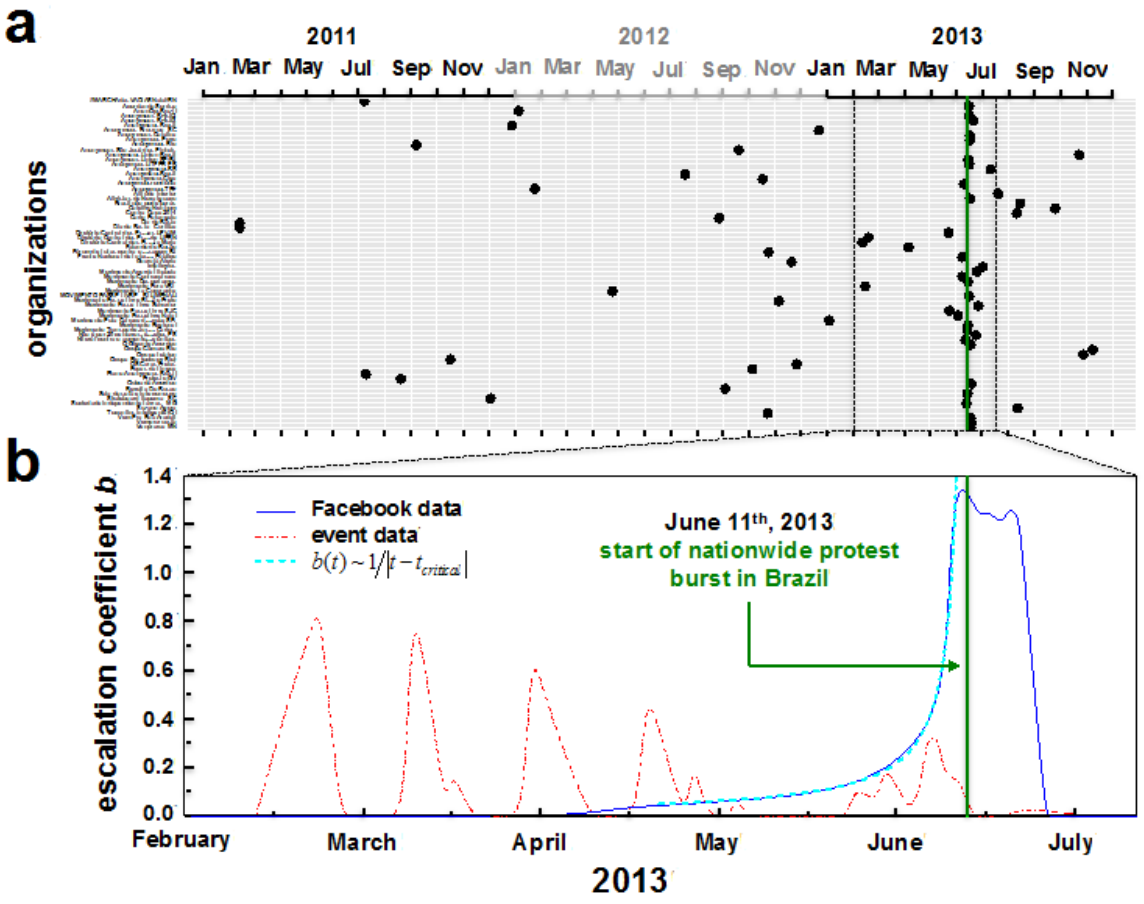


Figure 5. 4 A two-phase paradigm of dynamical process describing the life of the uprising. (a) The emergence of a list of non-stationary actors through large Facebook ad hoc "organizations". Each dot labels the date when a given organization initiated its online activity on Facebook, whose name is denoted on the right vertical axis. The vertical green line marks a transitional tipping point deduced from curve fitting, which coincided with a beginning of burst in Brazil in June 2013 (on June 11st). Furthermore, this line indicated a transition that separates a phase of online organizational dynamics (left side) from an outburst of protests on the streets. Meanwhile we can observe that the transition to protests on the streets are followed by decreasing of online activity (right side of the line). (b) Blue line shows the locally fitting result of the interval between the initiations of different organizations on Facebook from February to July 2013. Remarkably, the escalation parameter was best fit with a power equation whose exponent is 1.0 (dashed line), following the observed curve of b value closely. In contrast, same escalation analysis of time-series of GSR events during the same time period in Brazil (marked by red curve) did neither fit the observed b value curve well, nor generate any precursor signal.

The progress curve introduced before assumes a stationary process of development so that the escalation rate β or in other words escalation coefficient b is a constant

quantitative parameter during the time range of escalation. For the dynamics of online activity in social media, however, the value of escalation coefficient *b* can fluctuate due to the non-stationary nature of online dynamics. The hard point is that *b* could turn negative as a result of a momentary de-escalation of online discussion and activity of groups. It can be convincingly expected that the escalation coefficient *b* depends on the ordinal number *n*, and as a consequence *b* can be written in a function of time in the case of emerging uprisings. As illustrated in Figure 5.4 (b), escalation coefficient *b* is obtained as a function of time by locally optimized weighted estimation of scatterplot smoothing (LOWESS, details see section of Methods, subsection 5.5.2). In the bottom panel, negative escalation rate (temporary de-escalation part) are dropped out by setting those points at line *b=0.* Details of how to fit the escalation rate locally can be found in the Methods section 5.5.2. The segment of data of online activity of organizations on Facebook is the result out of the entire fitting result, which ignores those noise patterns but keeps only the long-run escalation part of non-stationary process.

The preliminary results in bottom panel of figure 5.4, indicates that the escalation coefficient *b* has a steady zero position at the beginning of the time range, implying that there is a short period of de-escalation from February to April. Then the escalation rate becomes positive and gradually increases before the sudden outburst in the 'Brazil Winter'. Specifically, the *b* value diverges as the uprising approaching in an inverse algebraic way $(T_c - t)^{-\alpha}$, where T_c is the critical time and the power exponent $\alpha = 1.0 \pm 0.10$ (Fig. 5.4 bottom panel, see Methods section 5.5.3.1). Moreover, the coefficient *b* actually serves as an order parameter in the lead-up to a transition, beginning from the phase of online activity whose key feature is organizational "many body". The kind of

transition strongly resembles transitions in dynamical networks and percolation [105,124], as well as a broad category of complex systems with "many body" phase and interaction between bodies prior to the transition [78].

On the contrary, when the same locally fitting method is applied to the time series of real events on the streets of Brazil in the same time range, the red curve in figure 5.4 (b) is obtained. The result from real events fails to provide a noticeable precursor signal prior to the transition to outburst. Consistent with the guess we have from figure 5.2, such a failure strongly implies that events on the street are insufficient to forecast any future unrest events. Instead, it highlights the essence and importance to investigate the convert build-up process occurring within social media, i.e. the "Dark Side" of our understanding of social movement.

5.3.2 Predicted Phase Transition Date

From the divergence equation of escalation rate b prior to transition, a prediction of the time to transition can be generated, i.e. critical time T_c in that inverse algebraic equation. Vice versa, the critical time T_c can serve as a testable prediction of our theory for the non-stationary dynamics and onset of unrest activity in a particular country or region. Assuming we are at the time before the outbursts of unrest and are monitoring the activity of all online organizations which are related to civil unrest on Facebook, the intervals of creation of organizations are thus fed in one by one. As a result, as more information is derived and accumulated, it should be possible to generate a prediction approximately self-consistent, meaning the predicted transition time approximately equal, and of more accurate.

In particular, at least three points is needed on the escalation curve fed in to begin the prediction of critical time. Also, with the exponent assumed as one, the fitting of predicted transition time can be simplified as the opposite number of ratio of intercept and slope from a linear regression. Under the condition of given data points, the joint distribution of numerator and dominator are a bivariate Student's *t*-distributions. The predicted transition time, therefore, is in a t-Ratio distribution [125]. Maximum Likelihood Estimation (MLE) method is adopted to estimate T_c. A 95% Confidence Interval (CI) is calculated subsequently by numerical approximation as to obtain the uncertainty of estimated T_c. In addition, the left boundary of CI, of course, have to be truncated at the position of "Today" when doing predictions, just as the shaded area in the four panels of figure 5.5 shows. For more details, please see section 5.5 - Methods.

Figure 5.5 illustrates the estimates of the critical time T_c as a function of time, narrower confidence interval (pink area) with more events information fed in. In this plot, the critical time is incredibly self-consistent: all on June 11th, 2013. The uncertainty of the prediction is shrinking remarkably as the time flows – more information of online activity is implemented into the transition equation. It is vital to point out here that all parameters and the prediction rely only on information from "history" and "today", so this model is not inapplicable when used for forecasting transition from online phase to the on-street phase.

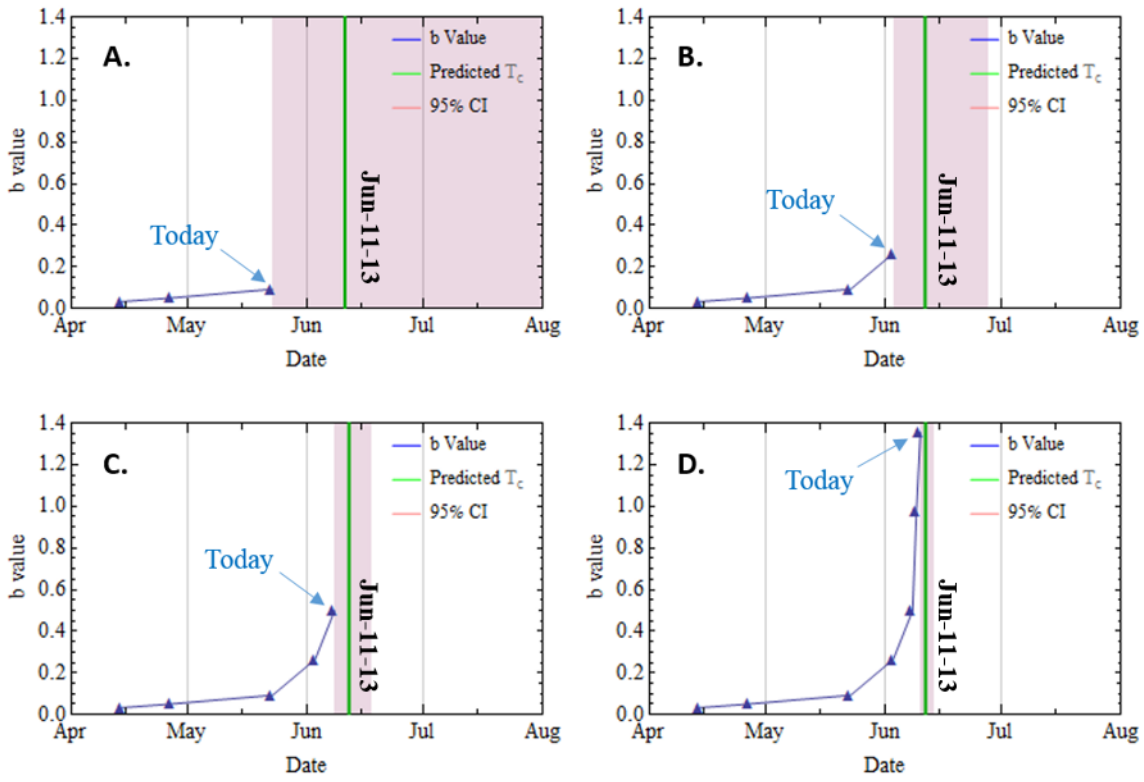


Figure 5. 5 Out-of-sample prediction of transition. Panel (A) shows a prediction made on 23 May of the date (vertical green line) and its uncertainty (pink shaded area) for the predicted upcoming transition; Panel (B) shows corresponding predictions made on 3 June, and Panel (C) shows them for 8 June. The last panel shows on 10 June the uncertainty nearly converges on the predicted transition time. These predictions are made by setting the order parameter exponent equal to 1 as for a wide class of percolation systems. The percolation is of organizational development in Facebook (Figure 1) not individuals in some contagion process. The unit of time measurement is one day.

Having provided a theory of the transition based on a mesoscopic order parameter *b*, a microscopic mechanistic model is now provided to complete the generative picture. It is acknowledged that other models may be possible just as in physical systems where the microscopic generative theory (Cooper pairs) underlying the mesoscopic order-parameter transition (Landau-Ginzburg) may not be unique, e.g. superconducting phase transition. However, the dynamical network model presented in this paper is innovative and based

on the acknowledged tendency of people to make and break links with organizations and hence with each other [126] as mimicked by the clustered network snapshot shown in Figure 5.4 corresponding to a famous subversive organization. The transition details can be described in the format as$(T_c - t)^{-1}$, which is common for a percolation type model. It is stressed that, however, here the percolation is among organizations with variable membership, not a simple contagion among people. A more complex generative model can be built in which a network of agents evolves with adaptive random process and peer attraction. It shows a similar divergence which arises as the links form and dissolve, making it a novel type of dynamical network transition.

5.3.3 Applying the Pattern to Other Countries

The paradigm introduced above describes the social uprisings as a two-phase process: successive phases consisting of covert build-up of organizations online and a burst of social unrests on the streets. We, furthermore, identify the phase transition stage of these two phase as the one from a broad class of models involving percolation, based on an inverse algebraic equation of transition:

$$b = A(T_c - t)^{-1}. \tag{5.1}$$

Because the Facebook over the world has a universal structure and our model relies only on internal interactions, not exogenous pressure from outside environment, we can reasonably expect our approach is language agnostic. Hence we also collect organizations with a political agenda in Venezuela from 2013 to 2014, totally 37 relevant Facebook groups as shown in Figure 5.6. The collection method is the same as what we use in the case of Brazil, which is introduced in the Methods section.

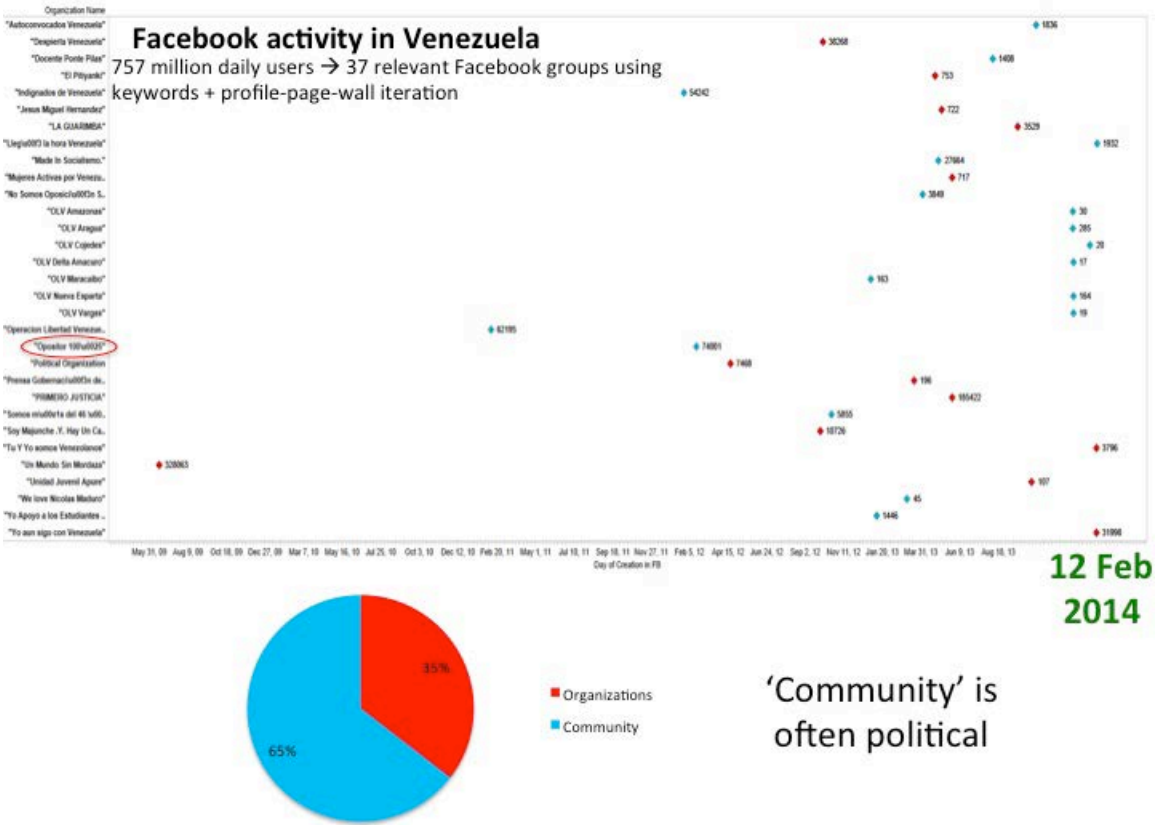


Figure 5. 6 Emergence of organizations and communities in Venezuela in 2014, from SI of Ref.[17]. Analogously to Figure 5.2 in the section 5.2.1, we collected organizations with a political agenda in Venezuela. In contrast to Brazil, the Venezuelan government suppressed those protests and thus disrupted the divergence toward a transition to outburst. Because of such retaliation from the government, the label 'community' was adopted in favor of 'organization'.

As the Venezuelan government usually strongly suppresses the protests thus pushes back any relevant activity online, the divergence of escalation rate towards transition is inevitably influenced by the government actions. Consequently, it is unable to observe a divergence curve of escalation coefficient (*b* value) as smooth as the one acquired in "Brazil Winter". Yet, a reasonable prediction of the onset of social outbursts in Venezuela is still accessible if resume the use of the previously observed mathematical equation governing the phase transition prior to an outburst. Figure 5.7 illustrates the data

of creation date of community (more political related) to our model, obtaining the non-stationary escalation rate/coefficient/parameter points (*t, b*).

In figure 5.7, an escalation parameter curve (blue points and curve) can be observed which is continuously pushed back compared to figure 5.4. Recall the nature of covert build-up, it is a self-evolved process with not only individual interaction but collective dynamics of cluster of individuals. Therefore, exogenous perturbation/suppression will definitely interrupt the normal build-up process, leading to a contorted dynamical build-up or even a failed attempt.

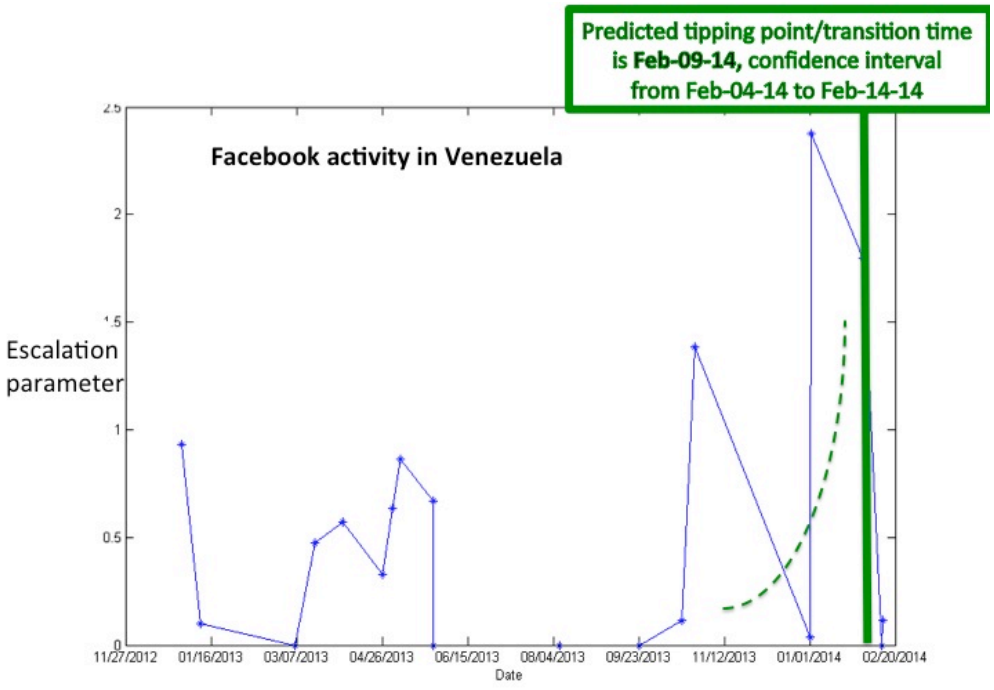


Figure 5. 7 Temporal development of escalation coefficient *b* based on the Venezuela data. The blue points mark the dynamical escalation parameter *b*, along with the time. The green dash line is the fitted curve from last five points, by the inverse algebraic equation. And the green bold line denotes the predicted critical date. Notably, we observed a sharp contrast compared to the smooth divergence observed in Figure 5.4 that we found in the Brazil data. Although data appear scattered, we predicted an onset of protests on the street in Venezuela, matching closely the observed start of nationwide protests on Feb. 12th.

On the left-hand side of figure 5.7, the blue points and curve shows an escalation parameter, fluctuating at the beginning and being oppressed at the end. That is a failed covert build-up online, possibly due to exogenous suppression from the government. Following the cease of the escalation there is a long peaceful time or de-escalation period. Meanwhile, on the right-hand side of figure 5.7, it is a diverging build-up of escalation parameter but with heavy noise. Although not smooth, this segment consisting of six points still can be utilized to make a prediction of transition date just as for “Brazil Winter". After fitting those points to the transition equation, a forecast could be made on Feb. 7th 2014 that an estimated nationwide protests would take place on Feb. 9th, 2014 with a 95% confidence interval of ±5 days. This is indeed close to the actual eruption of social unrests on the streets of Venezuela - Feb. 12th, 2014.

The inferior performance when analogously dealing with data from Venezuela indicates potential limitations of this model. It has been noticed that outbreaks in cities without significant volume of online media (e.g. Montevideo) seem to follow a distinct and slower process, with just one or two organizational events coordinated through existing union structure (e.g. teachers) or agricultural collectives (e.g. coffee and potato growers protests in Colombia in 2013). In these situations, the online social media that does exist plays the role of passively recording events as opposed to generating the two-phase process shown in Figure 5.4. Similar patterns has been found for the Arab Spring, suggesting that these countries were not sufficiently digitally enabled to exhibit the two-phase process. Given the rapidly growing ubiquity of social media, it is expected that most future uprisings will follow Figure 5.4 rather than the Arab Spring. As a final check of proposed generative mechanism, recognized socioeconomic percentages is utilized to

estimate the number of people in a given city involved in the Facebook organizational development phase, hence to estimate the number of people likely mobilized in future street protests. See section of Methods and Further Discussion for the example of Sao Paulo, which yields excellent agreement for the actual number hence confirms the value of the analytic sociology analysis proposed by Hedstrom [127].

5.4 Summary and Conclusion

In the previous sections, an innovative two-phase paradigm was introduced to analyze the outbursts of civil unrest in Brazil, 2013. The paradigm is proposed as the weakness of prevalent contagion algorithm has been found in forecasting and explaining modern uprisings. It is significantly inspiring to find a new model since the OSI project using the traditional approach of contagion failed to construct a valid prediction tool for civil unrest. Figure 5.2 demonstrates the GSR events profile in lower panel and shows that there is no evident precedent signals prior to the big outburst, known as “Brazil Winter”.

The key difference of this battery of unprecedented sudden uprising is that their activities of organizing, communicating and experimenting took place through tremendous virtual activities online rather than in real world. A huge volume of information and great amounts of users in social media boost and augment the preparation needed in initiating social unrests. Sharing ideas has never been this convenient in a cyber world, since one click can send messages to millions of people in the same region. The content of propaganda has never been as diverse thus its power and accessibility has achieved the historical maximum as video, music, words, speeches and many other forms can be used to persuade people.

Online activity is referred as covert phase of our two-phase model in this study and the phase transition time is of the most interests. Hence, it is necessary to analyze the dynamics of online activities and to find a theory to support the results of this study. To collect the data of online organizations relevant to the unrests, we keep monitoring the Facebook and recording the features, content and activities by key word searching together with manually filtering. Emphasis is placed on the non-stationary process of organization creation for the sake that the development of new organizations is the signal of information spread and people aggregation. Meanwhile, it can be analogized to the previous work in this study of insurgencies in Afghanistan and the general law of human confrontations. The theory of progress curve is augmented with local weighted regression (LOWESS) to fit the non-stationary dynamics of online creation.

A smooth curve of escalation rate versus time has been obtained for the case of "Brazil Winter". It well resembles the transition curve found in many complex systems in a power law equation. Notably, the exponent of the transition, obtained from best fitting, coincides the one found in high dimensional percolation ($D \geq 6$): 1.0. Thus the large-scale uprisings can be described as a systematically wide cluster in percolation. Also the development process can be viewed as the work of "opening edges" in lattices (see details in Methods section in this chapter). Even more excitingly, the order parameter value of 1.0 is homogenous to percolation in all space whose dimensions are no less than six. Hence, the heterogeneous structure of nodes can be neglected in online space for most of the users have more than 2*6 connections that definitely fall into the domain of high dimensional percolation.

Given the dynamics of transition, the escalation rate of any sequence of events can thus be monitored and the possible time of transition to the next phase can be predicted. This dynamical pattern is firstly applied to the original data, i.e. creation date of organizations online in Brazil, 2013.

Looking back at three months before the outburst (June.11 2013), this research group conduct simulation of forecasting by feeding the event one by one as time flows. On May 23, three points are discovered in the curve of order parameter which meets the minimum requirement for linear regression. Accordingly, the transition date has been predicted in advances of 18 days, whereas the uncertainty of this prediction is significant and unable to serve as a valid warning. Furthermore, the following events are recorded and achieve a narrower confidence interval while keeping the prediction date. In figure 5.3, it is shown that affirmative warning of large outburst is obtained two days before the actual uprising. This is authentically a remarkable achievement of our model that encourages us to devote to further investigations.

However, the limitation of this model in serving the society has be found before long. The theory basis of progress curve lays on the Red-Blue competition scheme of confrontations, while in some country, such as Venezuela, the Blue side (government) has very strong power to suppress the activities of Red side (protestors) in covert phase manifested by shutting down the network, monitoring the cellphone and messaging, and blocking the website etc. As a result, the escalation rate curve in these places has more noise and severer fluctuations, which undoubtedly damps the functionality of prediction of our model. Therefore, a curve as smooth as the case of Brazil is not able to be plotted but the predicted date of onset of unrests is still proportionately reliable: the date of

unrests from online to on streets is Feb. 09, 2014, while the real outburst is Feb. 12, 2014. In other countries where access to internet or cellphone is limited, this model is not the best fit for predicting social unrests.

This model of online development is generally the advanced version of the previous work of this research group, i.e. mathematical law of human confrontations. They both target the underlying dynamics of human confrontations in which a Red-Blue sides competing is present. Red side refers to someone who are using sporadic attacks to weaken the Blue and raising himself, while the red side is usually recognizably weaker than the blue side but good at making use of advantages in information, hiding in the dark or disguising in the public. This model implements new aspects of analysis of non-stationary progress, covert phase prior to real confrontations, and phenomenon of phase transition.

A simple mathematical law benchmarking confrontational behavior has been found by Dr. Johnson and his student several years ago. Now the work of this author in this project adds quantitative analysis of non-stationary dynamics of progress, especially those activities online. Digitally driven activity has a more aggressive development path, facilitated with modern information devices. Hence the development of online activity can serve in a covert phase as preparation to real confrontations and perform an increasing “escalation rate” in this phase. Additionally, it can be found in many other complex systems the divergence of order parameter in covert phase near the stage of transition to next phase, and a classical power law equation can fit it well. The transition equation of percolation in high dimensional space has a surprising coincidence to the divergence equation of escalation rate. Therefore, prediction of date of transition can be

administered with power exponent fixed and the underlying dynamics can be explained with percolations in lattices.

The future work on this model will be augmenting it for better accountability of underlying algorithm of development and phase transition in other fields, including finance, information spreading and behavioral economics. Next step will probably be applying this model in finance since some work has been done in previous chapters to analyze the ultra-fast trading in global financial market. Phase transition phenomenon is also pervasive in finance, for example, the forecast of shifts in global financial markets. The prediction in this field, however, differs greatly in other confrontations in that the discovery of predictability quickly annihilating the predictable pattern, as profit can be made from subsequent arbitrage [105]. As a result, although there is still some predictable quantities that can be used for forecasting by smart or lucky specialists [128,129], overall financial markets are notoriously unpredictable [130]. Nonetheless, evidences have been found to support the argument that dynamics in financial market may contain information presaging major events [131–133]. For example, some events are signaled by the growing spread between the value of put options and call options [131,132], and a shrinking or peaceful volatility surface can also be observed before the onset of abrupt change [131]. Moreover, research also addresses possible systematic relationships in variance and first-order autocorrelation [134].

All of above are of great interests to our research group; meanwhile, empirical challenges that will be encountered could enhance insights of two-sided human confrontations of this model. Hopefully, the new light shed on the confrontations in

financial market will improve the understanding and prediction of more transition phenomena.

5.5 Methods and Further Discussion

5.5.1 Facebook Organizations Monitoring

To identify Facebook profiles of organizations with a political agenda, key word search is firstly employed on the Facebook platform's search engine. After the recognition of profiles of Facebook organizations, private profiles belonging to individuals are discarded while profiles of public organizations are kept by "calling such a curated collection the Know Organizations Repository" (KOR, see figure below) [17]. The last step is to monitor the activities of organizations in the KOR. If a new organization "liked" an organizations that is already in our KOR or has interaction in coordinating further protests, it will be added to the KOR as illustrated by figure below. The automatic parts of the tool is done by Facebook Query Language (FQL), developed by Nicolas. The manual parts are finished by Nicolas and Ana.

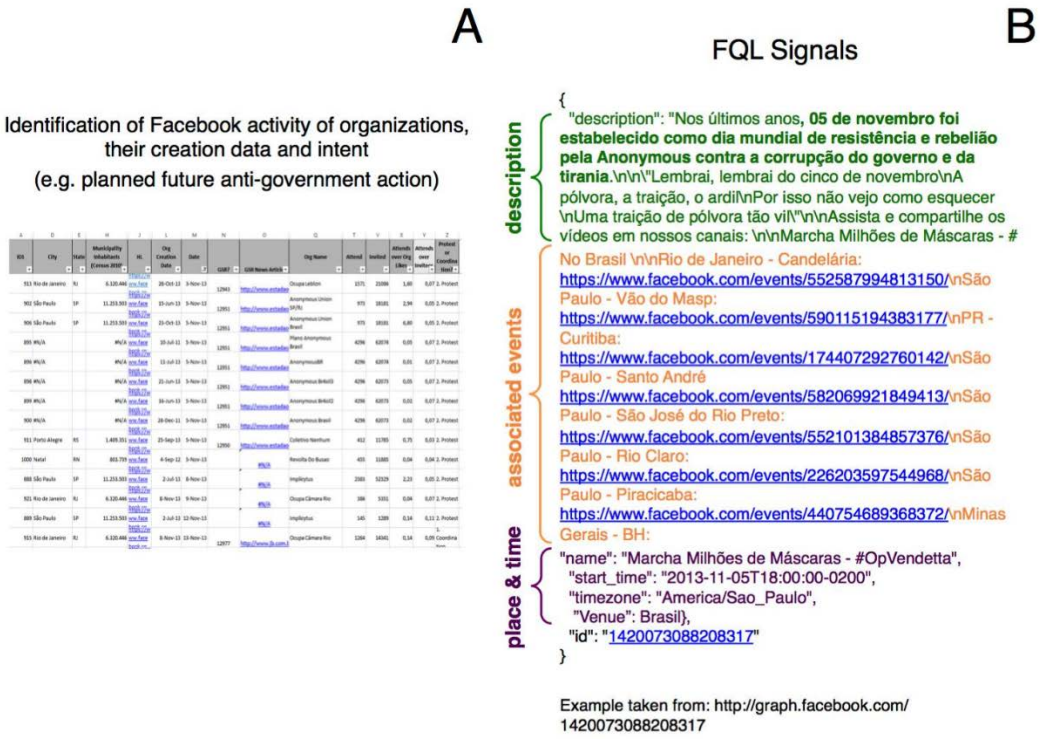


Figure 5. 8 Methodology to generate a list of active organizations. A) We call this list Knowable Organizations Repository (KOR) that can be generated for each city, country or region of interest. B) Information about the activities of organizations in our KOR's are retrieved and assembled using FQL (Facebook Query Language).

5.5.2 Determination of Order Parameter *b*

In reality, the observed time-interval values are scattered around the best-fit progress curve just as mentioned at the beginning. The escalation parameter/rate *b* is dynamical, fluctuating when events occurring. Now it's shown as an essential element for the locally weighted scatterplot smoothing (LOWESS) analysis.

The trend in time to formation of next organization is given by equation 4.1 $\tau_n = \tau_1 n^{-b}$ In other words, if the *n*th organization has just been formed, we will wait on average $\tau_1 n^{-b}$ until the next one is formed. However, in practice, a series of, say N "things", will need to fall into place happening before online Facebook users can form themselves into the next organization (*n*+1). Each of these steps may "fail" which makes

the creation of the next organization delayed or ahead of the trend. Following multiplicative degradation processes in engineering, it is assumed each of these steps multiplies the expected time interval by factor $(1 + \varepsilon_j)$ where the ε_j's are Gaussian distributed *i.i.d*. stochastic variables which mimic these exogenous factors. Therefore, the observed time interval is given by $\tau_n = \tau_1 n^{-b} \prod_{j=1}^{N}(1 + \varepsilon_j)$. Taking the logarithm of both sides, and assuming $\varepsilon_j \ll 1$ gives $\log \tau_n = \log \tau_1 - b \log n + \sum \varepsilon_n$ on a log-log plot since $\log(1 + \varepsilon_n)$ is approximately ε_n if it is small. This is a straight line fit (maximum likelihood) on a log-log plot as assumed by LOWESS, and with Gaussian-distributed *i.i.d*. residuals.

In this project, a robust version of LOWESS methods given by the function *lowess* in *Matlab 2013a* is used to estimate the trend, with data-factions of $f = ¼$ and number of steps = 2. Within the time window, LOWESS is applied to the scattered data points thus the corresponding smoothed value is obtained as shown in the following figure. The latest monotonically decreasing part of the smoothed curve is taken out and a simple linear regression is conducted on this partial sequence of intervals. Thereby the escalation value for specific time window is figured out.

When a new event happens, the time window is expanded to include it and a new finite size time window is initiated. The corresponding escalation value of that time window together with the date of the new event constitutes one point in Figure 5.9.

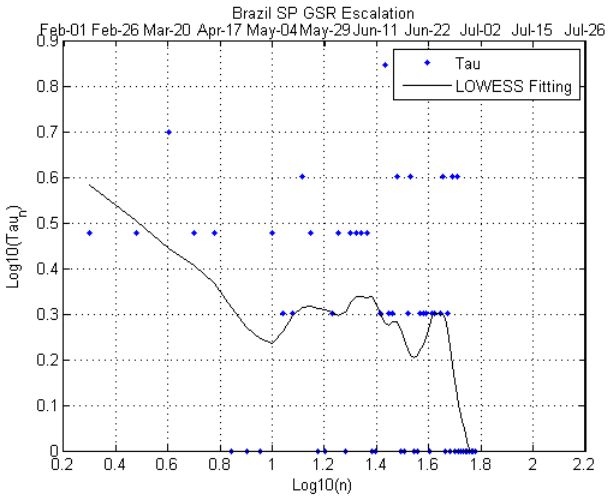


Figure 5. 9 LOWESS Fitting of Interval Points . Blue points are the interval $\boldsymbol{\tau_n}$ and ordinal label *n* in log-log scale, whose logarithm base is 10. The black curve is the fitted line as a result of LOWESS. We can see that the curve is in a systematic trend of escalation ($\boldsymbol{\tau_n}$ decreasing), but with fluctuation, somewhere even as de-escalation.

How to deal with the fluctuating *b* values? We care only about those latest parts that are in escalation stage. Hence when a new events (actually corresponding interval) occur, we put in a new point of ($\log\tau_n$,$\log n$), and check if the last point is in an decreasing segment. If not, it is the beginning of a de-escalation and the *b* value at that date is set as zero. If in escalation stage, the segment containing the last point is taken out and the *b* value from simple linear regression of that series of points is calculated, yielding a point in bottom panel in figure 5.4. Seven points are all obtained in the transition stage of the case of Brazil (from the first non-zero point to the peak point).

5.5.3 Dynamical Progress Curve: Fitting Curve of Escalation Rate

5.5.3.1 Fitting Exponent of Transition α

To fit the distribution of exponent α, the transition equation $b = A\,(T_c - t)^{-\alpha}$ is transformed to a linear equation as $ln\, b = -\alpha\, ln\,(T_c - t) + ln\, A$ (A is constant, natural logarithm as base). In the next step, a trial is adopted to and test method to search the critical time optimized for the best linear fitting. Defining a domain of T_c whose length is 30 days with a left boundary that is on the next day of the last moment of t series, T_c is taken from the beginning of the domain, and the exponent α and a subsequent confidence interval of 95% is estimated by a simple linear regression. Increasing the pre-assigned T_c by steps of 0.1 day until the end of the domain, it can be obtained a series of estimation of exponent α and corresponding R square value of the linear regression. The optimized estimation yields the highest R square value, leading to a result of 1.0 ± 0.10. The deviation 0.10 is from the standard error of linear regression of slope.

5.5.3.2 Prediction of Transition Time T_c

Figure 5.4 suggested that the best fit for the underlying data is

$$b = A(T_c - t)^{-1}. \tag{5.1}$$

We can easily rewrite this equation in a linear form

$$b^{-1} = \frac{T_c}{A} - \frac{1}{A}t. \tag{5.2}$$

Assuming simple linear regression, the coefficients of equation (2) are both *t*-distributed with identical degrees of freedom. The format of the underlying equation indicates that the ratio of the intercept T_cA^{-1} and the slope A^{-1} is simply T_c, allowing us to consider T_c as a *t*-ratio distribution [3] with a degree of freedom denoted as *n.* In other words, we need to estimate the probability distribution of the ratio of the underlying two distributions to estimate T_c. The *t*-ratio distribution describes a random variable that is the result of the ratio of two other random variables that follow a *t*-distribution with equal degrees of freedom. Denoting these random variables by *u* and *v,* their joint distribution can be written as [3]:

$$f_n(u,v) = \frac{n^{\frac{n}{2}}\left(\frac{n}{2}\right)|\boldsymbol{H}|^{\frac{1}{2}}}{\pi[n + h_{11}(u-\theta_u)^2 + 2h_{12}(u-\theta_u)(v-\theta_v) + h_{12}(v-\theta_v)^2]^{\frac{n+2}{2}}} \tag{5.3}$$

Where $\boldsymbol{H}$ is a 2 × 2 symmetric matrix, usually called the scaled matrix defined as

$$\boldsymbol{H} \equiv \boldsymbol{T}'\boldsymbol{T} = \begin{pmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{pmatrix}, \qquad \boldsymbol{T} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_{n+2} \end{pmatrix} \tag{5.4}$$

Where $\boldsymbol{T}'$ is the transpose of the matrix $\boldsymbol{T}$, which denotes the (n+2) × 2 matrix of regressors**.** In addition, h_{11} and h_{12} are elements of $\boldsymbol{H}$, and θ_u and θ_v are constants defined as:

$$h_{11} = n + 2, \;\; h_{12} = \sum_{i=1}^{n+2} t_i \tag{5.5}$$

$$\binom{\theta_u}{\theta_v} \equiv \frac{n}{\sqrt{(\boldsymbol{B}-\boldsymbol{T}\hat{\beta})'(\boldsymbol{B}-\boldsymbol{T}\hat{\beta})}}\hat{\beta} \tag{5.6}$$

$$\hat{\beta} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{B} \tag{5.7}$$

Where

$$\boldsymbol{B} = \begin{pmatrix} {b_1}^{-1} \\ \vdots \\ {b_{n+2}}^{-1} \end{pmatrix} \tag{5.8}$$

A new random variable *w*, is constructed as the ratio of random variables *u* and *v*. For suitability, we define a linear variable *x* of *w* as

$$x = \frac{h_{11}}{|\boldsymbol{H}|^{1/2}}w + \frac{h_{12}}{|\boldsymbol{H}|^{1/2}}. \tag{5.9}$$

Similarly, we define four constants, m_x, m_y, k_1, k_2, and consider two quantities, q and Q:

$$m_x = h_{11}^{1/2}\theta_u + \frac{h_{12}}{h_{11}^{1/2}}\theta_v \qquad m_y = \frac{|\boldsymbol{H}|^{1/2}}{h_{11}^{1/2}}\theta_v \tag{5.10}$$

$$k_1 = \frac{1}{\pi(1+\frac{m_x^2+m_y^2}{n})^{n/2}} \qquad k_2 = \frac{\sqrt{\pi}n^{\frac{n+2}{2}}\Gamma(\frac{n+1}{2})}{2\Gamma(\frac{n+1}{2})(1+\frac{m_x^2+m_y^2}{n})^{-n/2}} \tag{5.11}$$

$$q = -\frac{m_x x + m_y}{(1+x^2)^{\frac{1}{2}}} \qquad Q = (m_x^2 + m_y^2 + n - q^2)^{1/2} \tag{5.12}$$

The probability density function of *x* can therefore be written as[3]

$$f_n(x; \boldsymbol{H}, \theta_u, \theta_v, n) = \frac{k_1}{1+x^2}\left\{1 + \frac{k_2 q}{Q^{n+1}}\left[2F_{n+1}\left(\frac{q\sqrt{n+1}}{Q}\right) - 1\right]\right\} \tag{5.13}$$

while the density function of *w* is

$$g_n(w; \boldsymbol{H}, \theta_u, \theta_v, n) = \frac{h_{11}}{|\boldsymbol{H}|^{1/2}} f_n\left(\frac{h_{11}}{|\boldsymbol{H}|^{1/2}} w + \frac{h_{12}}{|\boldsymbol{H}|^{1/2}}\right). \tag{5.14}$$

The function F_{n+1} in equation (13) represents the cumulative density function of standard Student's *t*-distribution with *n+1* degrees of freedom. The domain of variable *w* is (-∞, +∞), and we assume that h_{11} is non-negative in the result. Since *w* is the density function of the ratio of two random variables, we easily see that

$$p(T_c) = g_n(w; \boldsymbol{H}, \theta_u, \theta_v, n). \tag{5.15}$$

Recalling equation (2), we can consider it in matrix notation as a linear regression problem. After obtaining the density function, we can estimate the critical time and the 95% confidence interval with a Maximum Likelihood Estimator. Specifically, the maximum of the distribution refers to T_c. To numerically approximate the 95% confidence interval, we utilize the following algorithm:

1. Calculate the maximum of $p(T_c)$, which is T_c. Then select one value t_1 on the x-axis on the left side of the peak.
2. With the aid of the density function, find another point on the x-axis on the right side of the peak, t_2, satisfying $p(t_2) = p(t_1)$.
3. The integral value α of density function p(t) from $t = t_1$ to $t = t_2$ indicates the probability of a random variable T falling into this range. Testing a different value of t_1 in the range of (-100, T_c) we considered the range (t_1, t_2) that yields an α nearest to 0.95 as 95% confidence interval of the predicted transition time T_c.

(In particular, we used the *FindRoot* function in the software *Mathematica 9* in the 3rd

step to find the appropriate t_1.)

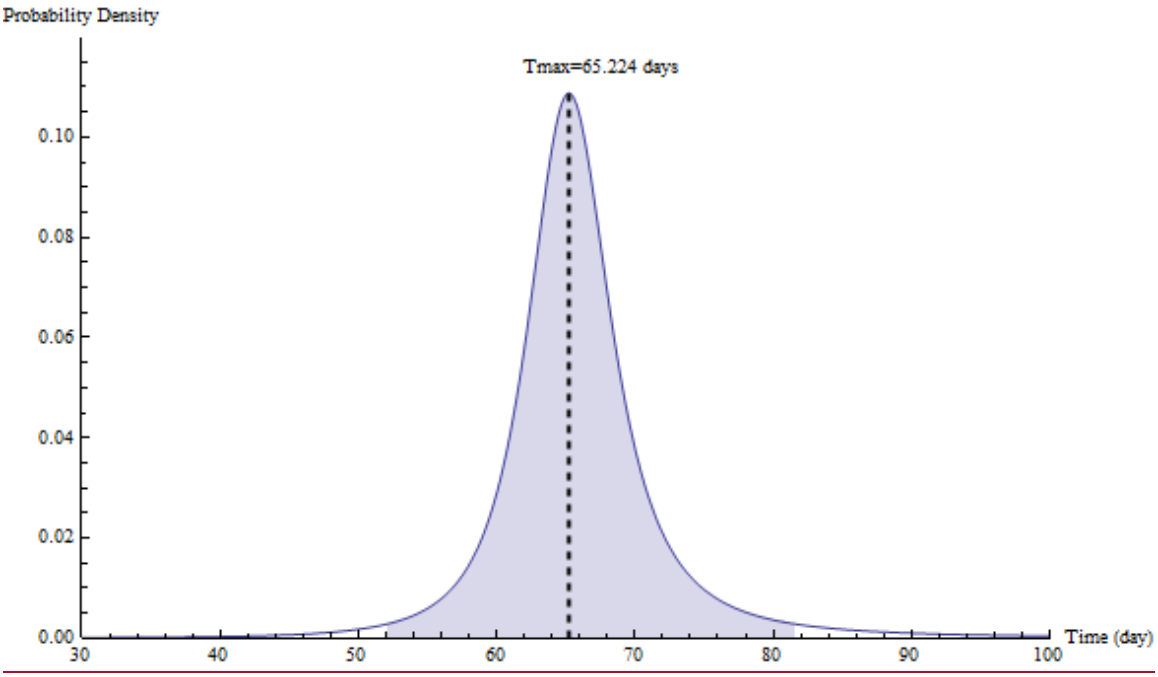


Figure 5. 10 Example of t-ration Distribution of Predicted T_c . The curve is the density function of the predicted critical time. The blue shadow represents the 95% confidence interval and the bold dash line marks the maximum likelihood of T_c

5.5.4 Sociological Survey Supporting Our Data

The following figure shows that, taking socioeconomic data from surveys etc. and following the elements of our organizational development, the predicted number of participants is actually the correct order of magnitude observed for protestors:

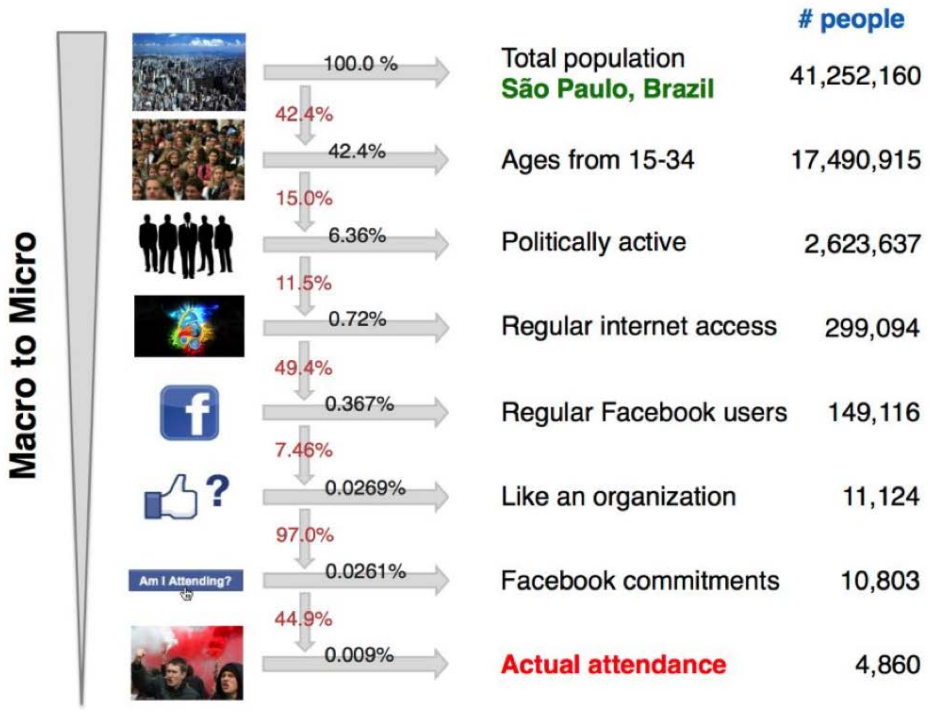


Figure 5. 11 The estimated actual attendance to protest in Sao Paulo, Brazil, from the census of city and Facebook.

It supports our scheme of two-phase development that people possibly influenced by Facebook is in the same order of observed protestors. At least it confirms that the number of people influenced by the Facebook are able to onset a protests as strong as observed on the streets in Sao Paulo, Brazil.

5.5.5 Simple Introduction of Percolation Theory

In previous sections, we mention that the exponent α in transition equation coincides the exponent in percolation in high-dimension space (dimensions larger than six). Now here we present an introductory explanation of the reason.

Percolation theory, commonly used in statistical physics and mathematics, describes the behavior of self-connection of nodes in a random graph. And it has many applications

to materials science and other domains. Let us start with the easiest example of square lattice in 2D space, which is a special case of bond percolation. The following figure is the work of Dr. Michael Gastner of Imperial College London, available at website: http://wwwf.imperial.ac.uk/~mgastner/percolation/percolation.html

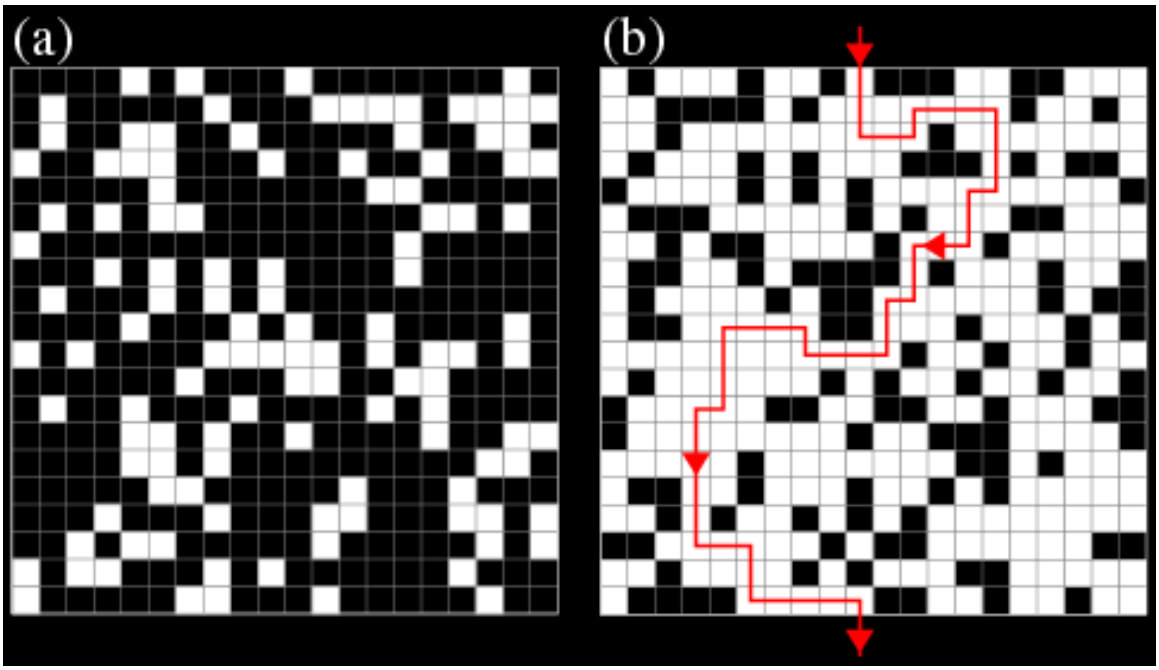


Figure 5. 12 (a) The square lattices are randomly popped up, i.e. one Black square replaced by White Square at each step. At the beginning, there are only a few white squares, clusters of white are too small to connect two opposite sides of the square lattice. (b) When there are a lot of white sites, it is possible to connect the top and the bottom of the lattice with white sites, for example the highlighted path (denoted by red arrow).

Just as figure 5.12, the site percolation is physically described as follows

1. Begin with a $n \times n$ square lattice, where the squares are usually called “sites”.
2. The sites may be destroyed (turning to white) with probability p, or restored (turning to black) with probability 1-p. A white site will allow liquids to flow through, while the black site does not. The event of opening/closing a site is in independently identical distributions.
3. We care about the probability P_{path} , which denotes the probability of the event: there is a path from the top to bottom allowing the liquid to flow through. For example, the red curve in Figure 5.12

The P_{path} is a function of size n and opening probability p. Under the limitation of infinite size, P_{path} depends only on p. And it has a critical phenomenon, as illustrated below (from http://wwwf.imperial.ac.uk/~mgastner/percolation/percolation.html)

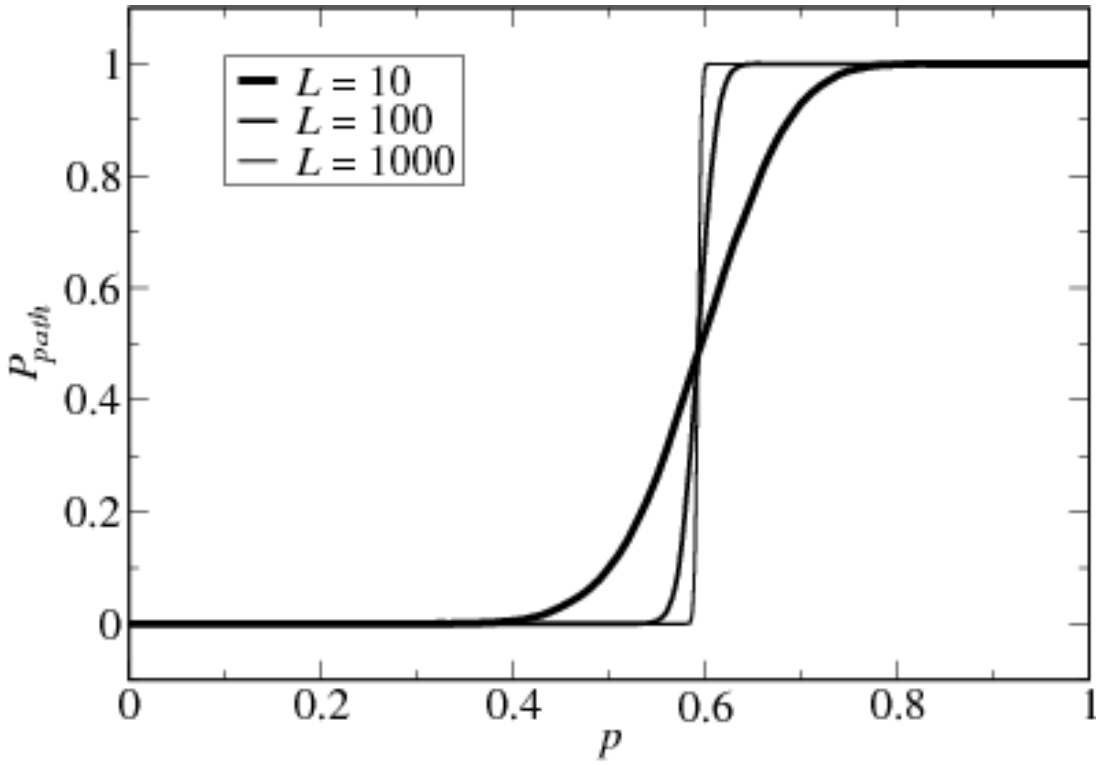


Figure 5. 13 The Phase Transition in Percolation. The probability of a path emerging from top to bottom P_{path} , has a transition phenomenon, depending on "opening probability" p. There is a critical threshold of p, beyond which P_{path} comes close to 1. Near the critical point, the P_{path} has a divergence relationship to p.

The behavior of systematic characteristics of percolation are diverged around the critical probability p_c , usually in a power law equation with a critical exponent.

$$P_{path} \propto (p_c - p)^{-\beta} \tag{5.16}$$

.

And the expected size of largest cluster during percolation C also has a divergence performance near the transition point $C \propto (p_c - p)^{-\alpha}$

The probability p or equivalently the existing open edges in lattices is proportional to the time in our case of online organizations, assuming the successful rate of opening an edge is constant. So we can get the transition equation of b

$$b = A(T_c - t)^{-\alpha}. \tag{5.17}$$

CHAPTER 6 Common Space Contagion Model

6.1 New Contagion Dynamics in the Cyberworld

In chapter 5, the weakness of traditional contagion models for real unrests events are indicated, as its application to "Brazil Winter" is insufficient to explain the underlying algorithm. The covert phase online is pointed out as the key difference of modern unrests to traditional ones. Although a simple model capturing the non-stationary dynamical pattern called hyper-escalation of online activities is found, there is still no clear model to simulate the whole scheme.

Meanwhile, new digitally driven unrests take places in many regions throughout the world. For example, ISIS (Islamic State) has recently got an unexpected rise with online support [135], which highlights the urgency to quantify modern-day human contagion phenomena [71,136]. However whether they are online [137], offline [138], social [11] or disease-based [16,139–150], it is unfortunate that all of them lie well beyond the predictions of standard infection models [11,140,151,152].

Here it is shown in this chapter that adding a simple mobility module of individual's alternation of online and offline spaces, gives rise to contagion dynamics that differs a lot to Susceptible-Infected-Recovered (SIR) model. Contagion in our model involves not only epidemics, but a list of more general items. For instance, the propagation of a rumor, idea, message or virus through a population, online or offline [11,39,136,140,144].

Traditional SIR model is used extensively for such outbreaks with mechanisms described by the three generic differential equations for susceptible $S(t)$, infected $I(t)$ and

recovered $R(t)$, including specific social network topologies and rewirings, geographical features, birth-death processes and super-spreaders [11,71,136,137,142]. As illustrated in figure 1 (e)-(g), however, modern style contagion profiles can be notably different from the standard SIR model. Those profiles are usually characterized by a fast time-to-peak (T_m), slow non-exponential decays, subsequent long duration T as well as small ratio of T_m/T and occasional revivals (Fig.6.1).

Aiming at solving those un-matches of SIR, this research group develop a novel contagion model called common space contagion model (CSC), which can be viewed as combination of mobility from online/public space to offline/private space and the SIR process. The new feature of mobility from public space to private space features the fact that people online are more open to "virus" and thus easier to be "infected", whereas people offline is relatively hard to be susceptible to other "epidemics". This mobility structure is adopted in order to replicate the characteristics of modern style contagions.

Another creative job done by this research group is using technology of data mining and science to collect automatically the logs of ISIS aggregates, the data of their activity and their opponents, which has not yet been done by other researchers. This data is the latest example of contagions in cyber world – modern style contagion that are digitally driven. As a consequence, the novel contagion model developed by our group is able to capture the modern-style contagion dynamics online, from online pro-ISIS activity to the spread of financial rumors [7] and from online organizations to massive street protests [9,10]. Surprisingly, our result of the model is counter-intuitive such as that increasing the throughput of a common/public space, the severity can on the contrary be decreased.

In the following sections, the details of our innovative model is demonstrated, as well as how well it fit the data from ISIS activity, mainly on aggregates serving as ad-hoc to recruiters and new members. The work is a result of multidisciplinary cooperation from computer science, sociology and physics. As it is still undergoing, the results shown here is preliminary.

6.2 Structure of Common Space and Its Contagion Profile

Each example (Fig.6.1 (e)-(g)) recurs frequently in the corresponding datasets (Details in methods section 6.6) and hence is unlikely to be dictated by one-off exogenous factors. Rather than issues such as potential power-law decays [151,152], the focus here is on the overall profile shape, i.e. the peak value H, time-to-peak T_m and duration T since these play a critical role in planning. The three panels of Fig. 6.1 (e)-(g) are from bursts in real world and correspond to the three representative infection profiles: red, green and blue. It supports that the infection profiles generated by CSC model are not just fiction but are able to replicate some outbreaks in real world. Red type of profile has a sharp rise and decays slowly, while green type has a relatively slower rise but decay faster. With respect to the blue type, it has a recursive peak which is usually smaller than the initial one.

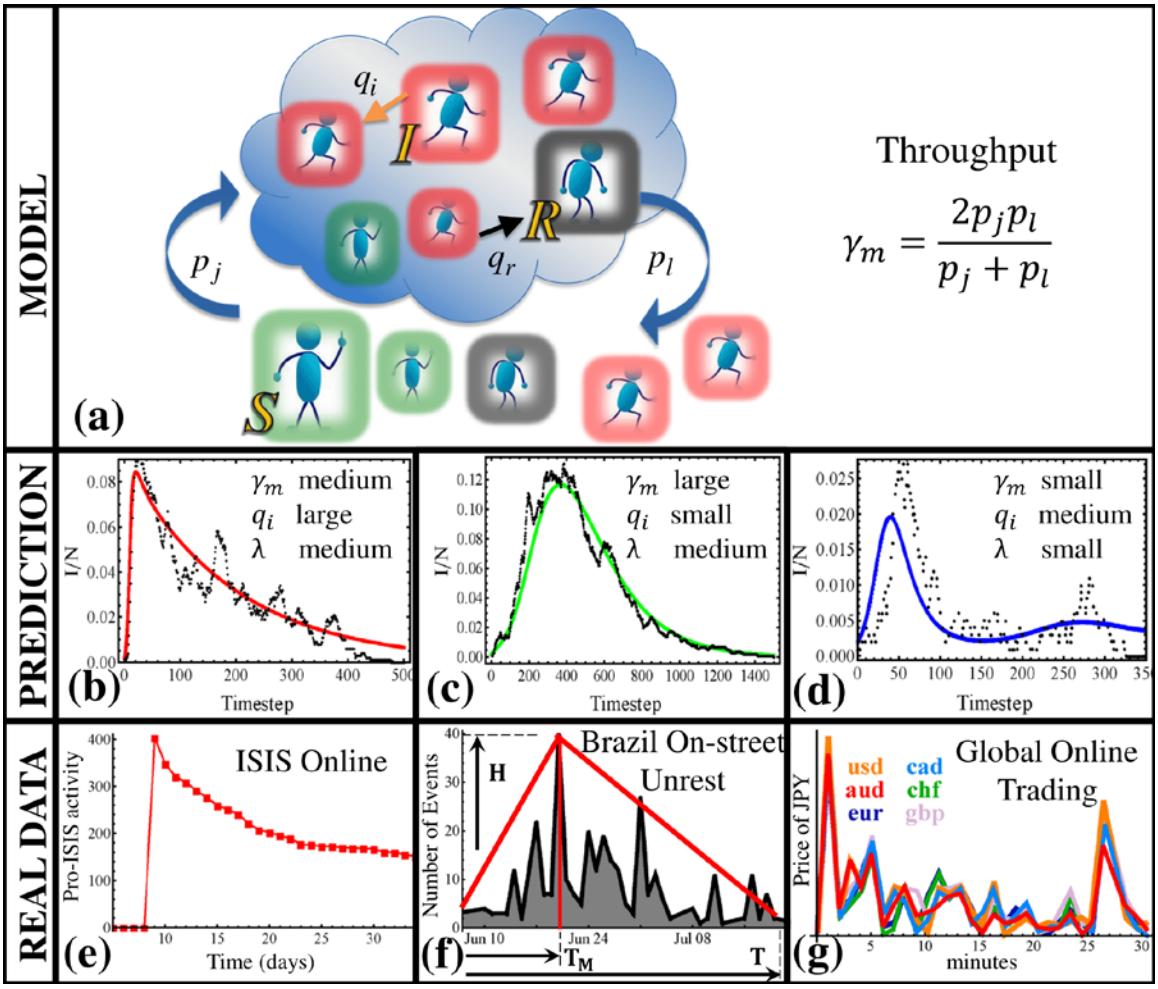


Figure 6. 1 Outbreak profiles in real-world systems and CSC model. (a) Common space theoretical contagion model. Common space (blue cloud) can be an online site, a network group, a physical place etc. and the transmission can range from information or rhetoric to a real virus. An individual outside (or inside) the common space has a probability p_j (or p_l to enter (or leave) at each time step. Infected individuals (i.e. activated: red) inside the common space have a probability q_i to infect other susceptibles (green) inside the common space. Infecteds both inside and outside recover (black) with probability q_r, $\lambda=q_i/q_r$ (b)-(d) illustrates the qualitatively distinct outbreak profiles predicted by our model, with the corresponding parameter regime. Black line is simulation, colored line is from integrating the coupled differential equations (see Methods section 6.6). (e)-(g) shows how these theoretical profiles capture various modern-day outbreak profiles: (e) pro-ISIS activity online 2014 matches (b); (f) protests on-street in Brazil 2013 matches (c); (g) global online currency trading during transmission of rumor of re-evaluation of Yuan, matches panel (d). Profile shows the variation of all major currency exchange rates [7]. For (e)-(g), similar profiles appear repeatedly in our datasets

(see Fig. 6.2) confirming that the $I(t)$ variation is a reproducible signal, e.g. for (g) an almost identical profile occurred several months later when the same rumor circulated again [7,11].

6.3 Feeding Islamic State Data in CSC Model

Figure 6.2 confirms this wide variability in modern-day infection profiles, using detailed datasets (see Methods section) for contagion activity surrounding online pro-ISIS support (color triangles) and on-street civil unrest (circles). The pro-ISIS data were obtained from the Russian-language equivalent of Facebook called VKontakte (www.vk.com) during 2014, and double-checked by subject-matter experts. VKontakte provides an ideal platform for study since many of its users are of Chechen origin focused in the Caucasus near ISIS' main area of influence. It regularly featured infection-like outbreaks of pro-ISIS rhetoric prior to the shutdown campaigns that started in late 2014, with the copying and spreading of material (e.g. for recruitment) becoming an essential part of ISIS' online power and global threat. The civil unrest data were obtained during a unique multi-year, national research project in which exhaustive event analysis was carried out across an entire continent, with resolution at the level of individual cities [10,26].

Bursts of activity were recognized using the methods of Ref. [153] and cross-checked manually by subject experts. For each burst of events, its best-fit profile $I(t)$ is obtained and hence its features (H, T_m, T) value (e.g. Fig. 6.1(f)). Even if its parameters are allowed to vary freely, the standard SIR model (black line, Fig.6.2) cannot capture this range of profiles. Figure 6.2 inset illustrates the underlying pro-ISIS infection profiles

(*I(t)*) with the corresponding color triangles. Again, the best-fit SIR profile provides a poor description.

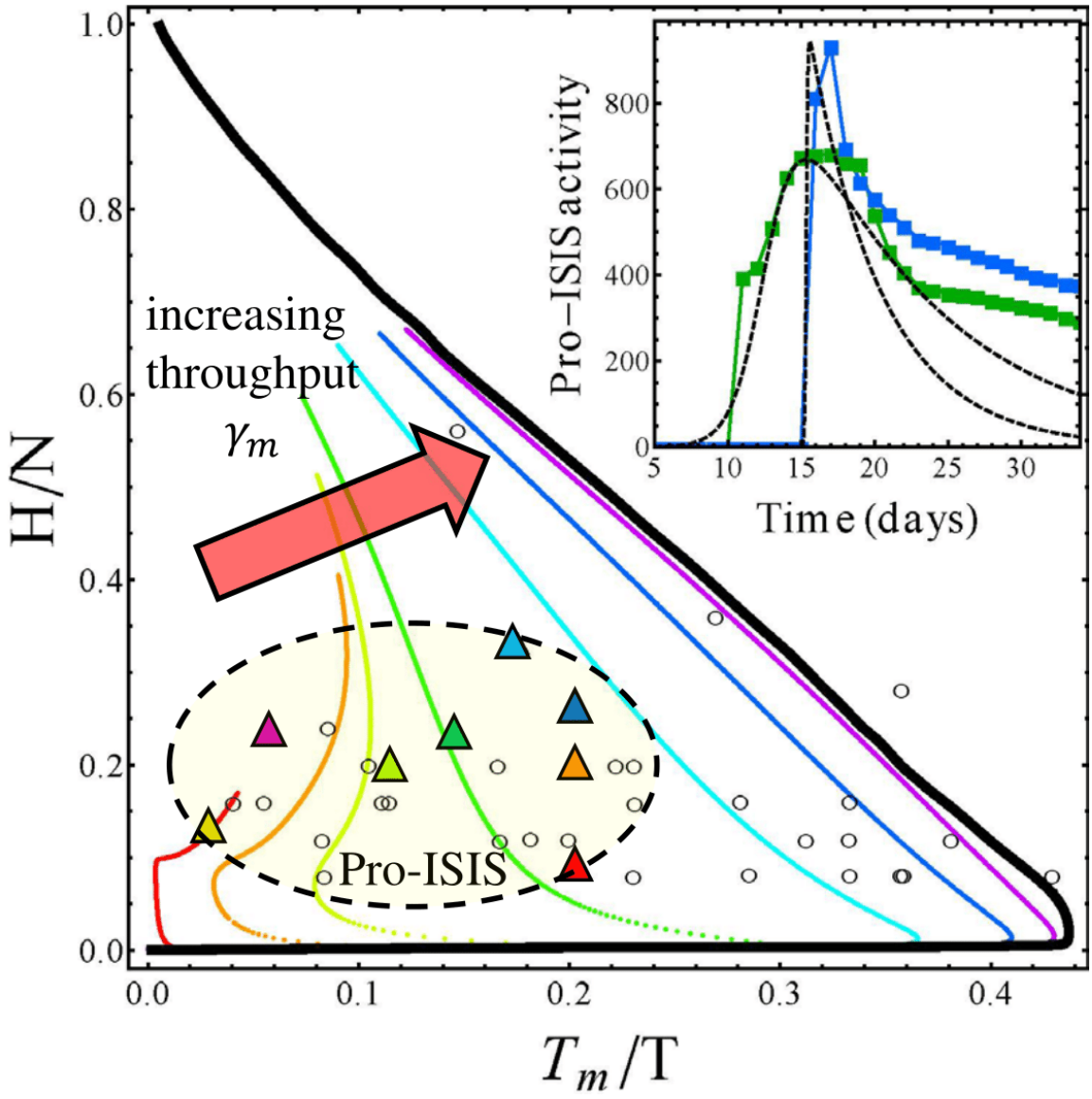


Figure 6. 2 Empirical versus theoretical results as a function of infection features: peak H, time-to-peak T_m and duration T. Empirical data are outbreaks of online pro-ISIS activity (colored triangles); and on-street civil unrest (circles) across Latin American during 2012-2014 national project [26]. See text for details. Theoretical results (colored lines) obtained from integrating the six coupled differential equations in our model, for different values of throughput γ_m (red line, γ_m =0.001; light green line, γ_m =0.01; blue line, γ_m =0.1). The same quantities are also calculated for the standard SIR model (black line). In all simulations the population size is identical (N=1000), and the same infection

probability q_i = 0.002). Each trajectory starts near to the origin for $\boldsymbol{\lambda \equiv q_i/q_r}$ and grows until $\boldsymbol{\lambda = 1}$ in steps of $\boldsymbol{\delta\lambda = 10^{-3}}$. Inset: Two typical empirical profiles for online outbreaks of pro-ISIS activity, named as club81567093 (blue) and interes.publics (green). Colors match corresponding triangle. Black line shows that even a best-fit standard SIR curve with the benefit of freely varying parameters, fails to capture the overall *I*(*t*) profile -- as is also the case in Figure 6. 1(e)-(g).

This theoretical model (scheme as Fig. 6.1(a)) allows the number and membership in a common space to change on a timescale which may be fast, slow or comparable to the intrinsic timescales of the transmission and recovery process. The *Common Space* can be an online social media site such as a chatroom, an online community or group, an offline place such as popular bar, airport or workplace, or even a well-defined cluster within a dynamic network. To reflect the increased chance of infection within the confines of the common space (i.e. transmission of a call to protest, an idea or information etc.), viral transmission can occur between people inside, but not outside, the common space. Figure 1(a) mimics how individuals may sporadically connect online (i.e. enter the cloud) and hence interact with others (e.g. through a spiritual site sympathetic to ISIS as in Fig.6. 3(a)); then become infected by a particular rhetoric that starts circulating and hence join an activist group online (e.g. pro-ISIS in Figs. 6.2 and 1(e)) or on-street (Fig.6.1(f)). The real-time analysis of VKontakte shows that this is indeed the mechanism through which ISIS manages to spread recruitment material and files, audio lectures, new prayers and official ISIS videos. Even if not all infected participate, the actual numbers will be proportional to the profile of infected individuals *I*(*t*), assuming each individual has a similar probability to participate. As an example of other applicable contexts, foreign exchange traders in trader chatrooms [7] can become infected by a given rumor that is circulating and hence become active in trading particular currencies, thereby moving prices as in Fig. 6.1(g).

Though more elaborate variants of this theoretical model are possible, they would all necessarily feature the same fundamental competition between the timescales of throughput, transmission and recovery. Hence, the simplest possible variant is presented. At each time-step *t*, each individual leaves or joins the common space with probabilities p_l and p_j. Any correlated leave-join behaviors is ignored for simplicity. In the steady state, the total number of people leaving and joining the common space, e.g. a group or community, per time-step is $N\gamma_m$ on average, where $\gamma_m = 2p_j p_l/(p_j + p_l)$ (see Methods section). γ_m is the throughput parameter. The mean number of people in the common space is:

$$\langle N_g \rangle = \frac{Np_j}{p_j + p_l} = N\gamma_s \text{ where } \gamma_s = p_j/(p_j + p_l) \tag{6.1}$$

Varying $\gamma_m$at fixed γ_s amounts to changing the throughput while keeping $\langle N_g \rangle$ fixed. The system evolution is described by six coupled equations involving ($S(t)$, $I(t)$, $R(t)$) and ($S_g(t)$, $I_g(t)$, $R_g(t)$) for the susceptible, infected, and recovered sub-populations in the entire system and common space respectively (see Methods section 6.6).

For $I(0)$ - initial infected individuals - in the system, not necessarily inside the common space, the criterion for an initial increase in I(t) is

$$\lambda\, \gamma_s^2\, S(0) > 1 \tag{6.2}$$

If instead the infection starts with one infected agent inside the common space, the criterion is

$$\lambda\, (N\, \gamma_s - 1) > 1 \tag{6.3}$$

As shown in Figs.6. 1(b)-(d) and Fig.6. 2, the resulting dynamics reproduce the wide range of empirical values as well as detailed shapes (e.g. Figs. 6.1 (e) vs. 6.1(b), and the revivals in Figs. 6.1(g) and 6.1(d)).

6.4 Empirical Value of Mobility Parameters

Figure 6.3 shows that the key throughput parameters p_l and p_j have remarkably similar dependencies for two of the most important current examples of online common spaces. Figure 6.3(a) is derived from our dataset of online groups that are sympathetic to ISIS. Since these groups are largely spiritual as opposed to being reactionary to exogenous events, each has a fairly constant size over time and hence represents a steady-state common space of size $\sim N\gamma_s$. From the average number joining and leaving per day, as well as the total number of followers, we deduce p_j and p_l values. To check the validity of our common space picture, we then use these values to determine the average group size and find the predicted values remarkably close to the observed ones. Following the injection of an individual(s) with a contagious item (e.g. rhetoric or story), the resulting infecteds ($I(t)$) then vent their pro-ISIS stance in particular chatroom groups (e.g. Fig. 6.2 inset and 1(e)). Figure 6.3(b) shows that the massively multiplayer online role-playing game World of Warcraft (WoW) features similarly self-organized groups (guilds) which also act as a common space, producing a similar dependence to ISIS in Fig. 6.3(a). WoW is the world's largest, most-subscribed and most popular online game with more than 100 million accounts created [35]. Despite their different online origins, the similarity of Figs. 3(a) and (b) suggests a robust pattern. This means that for a group of a given average

size, the likely profile characteristics of a future outbreak can be estimated by reading p_j and p_l off from Fig. 6.3, calculating γ_m and γ_s, and then reading off the results in Fig.6. 2. Using $p_j \sim 10^{-3}$ and $p_l \sim 10^{-2}$ for ISIS as in (a), yields $\gamma_m \sim 2 \times 10^{-3}$ which predicts (Fig. 6.2) outbreaks of ISIS activity with small T_m/T profiles as indeed observed (Fig.6. 2 inset and Fig. 6.1(e)).

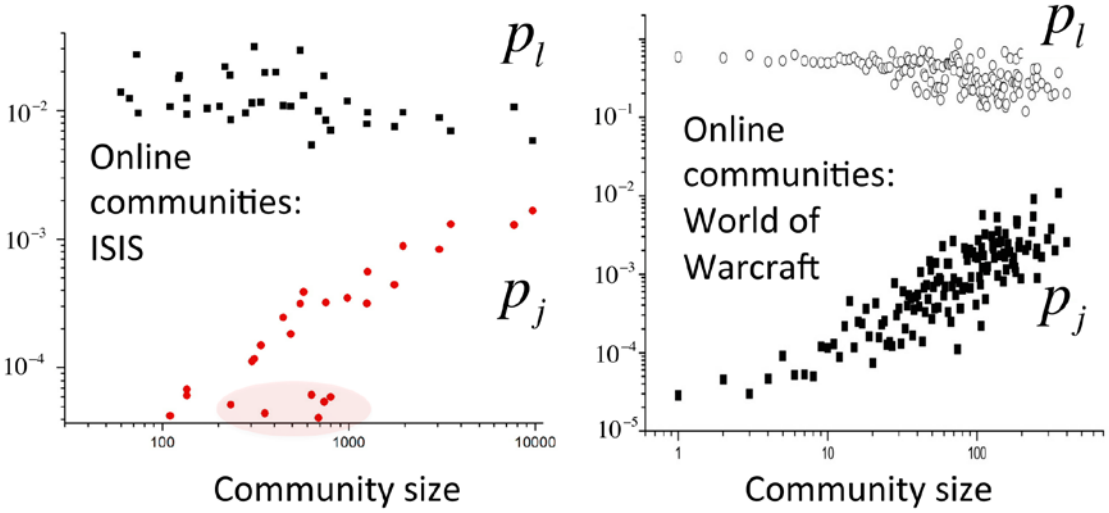


Figure 6. 3 Empirical p_j and p_l values across different online systems. The left and right panels show a common monotonic dependence as a function of community (i.e. group) size. Examples shown are (a) online pro-ISIS groups and (b) groups in the massively multiplayer online role-playing game World of Warcraft, in which groups are known as guilds [35]. In (a), shaded ellipse includes 6 groups who are news-based (e.g. *imandela*) and hence attract people infected by new information or material that they wish to spread quickly. They are not spiritual groups. Hence their p_j values for joining fluctuate significantly in time but are largely independent of group size, meaning that their time-averages will not follow the same monotonic variation as the other groups, all of which approximate a steady-state common space.

6.5 Diverse Behavior of Infection Profiles

Figure 6.4 shows the highly nonlinear dependence that emerges on the throughput parameter γ_m, which is missing from standard SIR models. Figure 6.4(a) shows there is a

critical value of the throughput γ_m^{crit} for which the peak intensity H is a maximum. For $\gamma_m < \gamma_m^{\text{crit}}$, infections are confined mainly within an initial group in the common space, while for $\gamma_m > \gamma_m^{\text{crit}}$people can be swept in and out of the common space before they have had a chance to be infected.

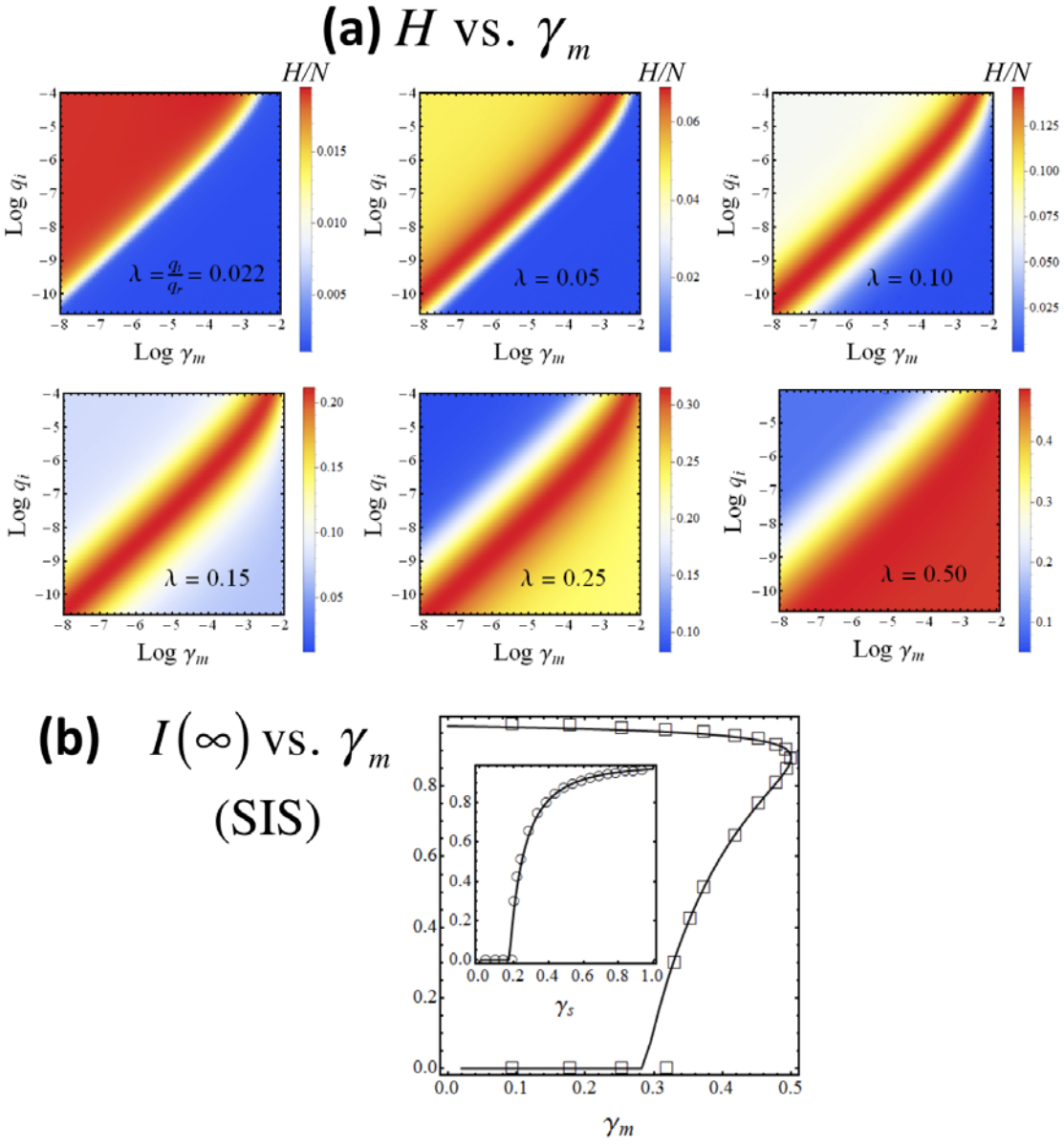


Figure 6. 4 Nonlinear effect of infection profile parameters on throughput γ_m. (a) The panel shows infection peak H normalized by N as a function of γ_m and q_i, for different values of $\lambda = q_i/q_r$. (b) shows SIS version of model. $I(\infty)/N$ is the

normalized fraction of infecteds in the long-time limit, shown as a function of $\boldsymbol{\gamma_m}$. N = 1000, q_i = 0.0005, q_r = 0.015. The condition of $\boldsymbol{p_j + p_l = 1}$ is satisfied, i.e. $\boldsymbol{\gamma_m = 2\gamma_s(1-\gamma_s)}$. Inset shows as function of $\boldsymbol{\gamma_s}$. Lines are from integrating the coupled differential equations, symbols are simulation results.

Also, the peak intensity H the time-to-peak T_m and the total number of infected individuals $R(\infty)$ can all decrease with an increase in throughput γ_m. This contradicts conventional wisdom that in order to reduce the severity of an outbreak, one should always increase the isolation of the common space (e.g. school, pro-terrorism website) and hence decrease γ_m.

From Fig. 4(a), $\gamma_m^{\text{crit}} \propto q_i$ (specifically $\gamma_m^{\text{crit}} \sim \mathrm{e}^3\, q_i$) for a range of λ values. The above findings are robust to different N due to intrinsic scaling. The lack of throughput in the standard SIR model can be partially corrected by re-normalizing its parameters. This could prove useful given the wealth of existing machinery and intuition surrounding the standard SIR model. We start with the discrete time version of $\mathrm{d}S_g/\mathrm{d}t$ given by:

$$S_g(t+1) = S_g(t) - q_i\, S_g(t)I_g(t) - p_l\big(\mathrm{S_g}(t) - q_i S_g(t)I_g(t)\big) + p_j(S(t) - S_g(t)) \tag{6.4}$$

From $dS/dt = -q_i S_g(t)I_g(t)$, we have $S(t+1) = S(t) - q_i S_g(t)I_g(t)$. The approximation $S_g(t+1) = \gamma_s S(t+1)$ yields

$$\big(1-p_j-p_l\big)S_g + \big(p_j-\gamma_s\big)S - (1-p_l-\gamma_s)q_i S_g I_g = 0 \tag{6.5}$$

This equality holds at all times only if $p_j + p_l = 1$ which says that the probability for the individuals outside the common space not to enter (i.e. $(1-p_j)$) should be equal to the probability of those inside the common space to leave. Under this condition, the full

differential equations (see Methods section) collapse to the three standard SIR equations for a well-mixed population but with an effective infection probability of $\gamma_s^2 q_i$ and an effective recovery probability of q_r. The reason that high $\gamma_m$happens to approach the standard SIR model in Fig. 6.2 is simply because large γ_{m} mimics a well-mixed limit. Similarly strong nonlinear dependences on throughput γ_m emerge from other disease models (e.g. SIS in Fig. 6.4 (b)) and also from multiple, interacting common spaces. Person-to-person transmission can also be replaced by a simple broadcast mechanism, to mimic situations where everyone is watching or reading the same news bulletin.

6.6 Methods and Supplement Information

6.6.1 SIR and CSC Model

CSC model (Common Space Contagion, Fig. 6.1(a)) is a generalization of a standard SIR process to take into account individuals passing through a dynamical group (common space) where they can become infected. The SIR model describes the spreading of an infection on a population over time. SIR stands for Susceptible (*S*), Infective (*I*) and Recovered (*R*), which are the three possible states that an individual within a population can have. The mechanism of infection is through direct contact between a susceptible agent and an infected agent with a probability q_i per time-step. Once an agent is infected, there is a probability q_r to recover per time step and hence become immune i.e., it can no longer get infected. It is convenient to define an infection's contact rate as $\lambda = q_i/q_r$. An agent outside the common space has a probability p_j per unit time to join the space, while

an agent inside has a probability p_l per unit time to leave the space. The equation of motion that describes the population dynamics within the common space, where N_g is the number of individuals within the common space at a given time *t*, can be written as

$$dN_g/dt = -p_l N_g + p_j(N - N_g) \tag{6.6}$$

In the steady state (i.e. $dN_g/dt = 0$) the mean number of agents within the common space $<N_g>$ is given by:

$$\langle N_g \rangle = N\frac{p_j}{(p_j + p_l)} = N\gamma_s \tag{6.7}$$

To quantify the turnover in the common space, the sum of the mean number of agents joining and leaving the space in a unit time is given by:

$$\begin{aligned} \mu &= \big(N - \langle N_g \rangle\big)p_j + \langle N_g \rangle p_l = N\big[(1-\gamma_s)p_j + \gamma_s p_l\big] \\ &= N\frac{2p_l p_j}{p_l + p_j} = N\gamma_m \end{aligned} \tag{6.8}$$

where $\gamma_m = \frac{2p_l p_j}{p_l + p_j}$ is the throughput parameter referred in section 6.3.

6.6.2 Simulation and Analytic Equations

It is found that remarkably close agreement between the integration of our six coupled differential equations is present, as well as a discrete time simulation of our model. For the simulation, initially all the agents are susceptibles and the system run the common space process until the average size reaches its equilibrium value. Then an individual in the common space (group) is randomly picked and made infected. In the next time step,

this infected individual will carry out the SIR process and the co-evolution of the throughput and SIR begins.

The six coupled differential equations that describe the process are:

$$\frac{dS_g}{dt} = -q_i S_g I_g - p_l\left(S_g - q_i S_g I_g\right) + p_j(S - S_g) \tag{6.9}$$

$$\frac{dI_g}{dt} = q_i S_g I_g - q_r I_g - p_l\left(I_g + q_i S_g I_g - q_r I_g\right) + (1 - q_r)p_j(I - I_g) \tag{6.10}$$

$$\frac{dR_g}{dt} = q_r I_g - p_l\left(R_g + q_r I_g\right) + p_j\left(\left(R - R_g\right) + q_r\left(I - I_g\right)\right) \tag{6.11}$$

$$\frac{dS}{dt} = -q_i S_g I_g \tag{6.12}$$

$$\frac{dI}{dt} = q_i S_g I_g - q_r I \tag{6.13}$$

$$\frac{dR}{dt} = q_r I \tag{6.14}$$

From the above six equations, $\frac{dI}{dt} = q_i S_g I_g - q_r I$. Hence an initial $dI/dt > 0$ will yield an epidemic. For $I(0)$ initial infected people in the system, no matter in the common space or private space, the initial $S(0)$ can be assumed to be randomly distributed and thus at the well-mixed limit. As a consequence, $S_g(0) = \gamma_s S(0)$ and $I_g(0) = \gamma_s I(0)$. The restriction that both variables in the right hand side implies the criterion $\lambda\gamma_s^2 S(0) > 1$. For our simulation, we set the initial seed of infection as one infected individual in the common space, which requires $\lambda(N\gamma_s - 1) > 1$ for initial increase of the infected.

For a special case of $p_j + p_l = 1$, the differential equation of three sizes variables are simplified to the following effective SIR equations:

$$\frac{dS}{dt} = -q_i S_g I_g = -\frac{q_i(N^2 \gamma_s^2 SI)}{N^2} = -\gamma_s^2 q_i SI \tag{6.15}$$

$$\frac{dI}{dt} = q_i S_g I_g - q_r I = \gamma_s^2 q_i SI - q_r I \tag{6.16}$$

$$\frac{dR}{dt} = q_r I \tag{6.17}$$

6.6.3 Data Collection and Details

The data for online ISIS support are obtained as follows. Both manual and computer-based techniques are used by this research group. First, as a pilot study groups of followers that support ISIS are manually selected by searching ISIS-related hashtags through the VK search option, e.g. popular hashtags are: ISN, isn, ISIS, khilafah, islamicstate, IS, shamtoday, fisyria, IslamicCaliphate, IslamicState, caliphate, DaulahIslamiyah. Relevant groups of followers were then selected based on the following criteria: a) They must explicitly express support for ISIS by publishing ISIS-related news and/or other propaganda. b) Their messages must contain calls for jihad in Syria (Sham). Once the groups of followers supporting ISIS were identified, an additional search is performed using followers and groups to whom they are linked, and to whom the links are linked etc. The search stopped when any new groups of followers that has been come across was already in our database. Thus even though its completion cannot be proved, our data is at least a highly representative sample. Moreover, once automated, it went through a rigorous process of cross-checking and occasionally augmentation manually using random sampling by subject matter experts. This research group logged into VK.com on a daily basis at the same time of the day and manually searched for new batches of followers by:

a) Analyzing posts and reposts in the known groups of followers,

b) Following selected individual profiles that actively publish ISIS news as well as analyzing the groups they follow, if any.

Since the goal of these followers and ISIS is to spread the message across the social network, they constantly publish links to related material on their pages. The manual content analysis assisted the automated process by helping identify new groups of followers every day. The automated version of this process is as follows: The data collection script was written in Python and was hosted on our servers. Followers were found both automatically (and cross-checked manually) online on vk.com and documented in a CSV file that served as input to the data collection script.

Among the statistics collected were (1) the date that posts were created, (2) who posted the post, including gender, (3) the receiver of the post, including gender. Our time resolved ISIS data is obtained by setting up automated API's that run continually, together with manual cross-checked lists. Individuals are linked when they form part of the same online aggregate (group) supporting ISIS. For ISIS, there are typically >40,000 actors at any one time, with more than 100 million actual links in less than two months.

6.6.4 Bursts Features of Civil Unrest

The bursts of civil unrest we collected from Latin America have information on: category of events, location and date (as shown in Figure 6.5). The key to extract burst features (H, T, T_m) from events sequence is to construct the infection profile(s) from the events sequence. First, a long sequence of unrest events is segmented by a pre-specified threshold d (usually 3 days). That is, if interval between two consecutive events is not larger than d, the latter event is in the same segment with the prior. Second, the curve of

infection is built based on one segment of events, by making the reciprocal of intervals between events as vertical value and time step as horizontal value. Then the features of that infection curve can be extracted, just as figures 6.6 show.

1	Argentina	Thu 10 May 2012 00:00:00	Wed 16 May 2012 00:00:00
2	Argentina	Fri 3 Aug 2012 00:00:00	Thu 9 Aug 2012 00:00:00
3	Argentina	Mon 27 Aug 2012 00:00:00	Fri 7 Sep 2012 00:00:00
4	Argentina	Mon 1 Oct 2012 00:00:00	Thu 4 Oct 2012 00:00:00
5	Argentina	Mon 8 Oct 2012 00:00:00	Thu 18 Oct 2012 00:00:00
6	Argentina	Mon 22 Oct 2012 00:00:00	Fri 23 Nov 2012 00:00:00
7	Argentina	Tue 27 Nov 2012 00:00:00	Thu 6 Dec 2012 00:00:00
8	Brazil	Tue 11 Dec 2012 00:00:00	Fri 18 Jan 2013 00:00:00
9	Brazil	Mon 18 Feb 2013 00:00:00	Fri 29 Mar 2013 00:00:00
10	Brazil	Thu 14 Feb 2013 00:00:00	Wed 27 Feb 2013 00:00:00
11	Brazil	Mon 4 Mar 2013 00:00:00	Sat 16 Mar 2013 00:00:00
12	Brazil	Mon 25 Mar 2013 00:00:00	Wed 10 Apr 2013 00:00:00
13	Brazil	Mon 15 Apr 2013 00:00:00	Thu 25 Apr 2013 00:00:00
14	Chile	Mon 10 Jun 2013 00:00:00	Wed 31 Jul 2013 00:00:00
15	Chile	Wed 16 May 2012 00:00:00	Sat 26 May 2012 00:00:00
16	Chile	Sun 10 Jun 2012 00:00:00	Thu 14 Jun 2012 00:00:00
17	Chile	Tue 25 Sep 2012 00:00:00	Sat 29 Sep 2012 00:00:00
18	Chile	Thu 17 Jan 2013 00:00:00	Sat 19 Jan 2013 00:00:00
19	Colombia	Wed 27 Feb 2013 00:00:00	Fri 8 Mar 2013 00:00:00
20	Colombia	Wed 13 Mar 2013 00:00:00	Fri 29 Mar 2013 00:00:00
21	Colombia	Mon 17 Sep 2012 00:00:00	Tue 25 Sep 2012 00:00:00
22	Colombia	Tue 9 Oct 2012 00:00:00	Thu 18 Oct 2012 00:00:00
23	Colombia	Tue 4 Dec 2012 00:00:00	Mon 10 Dec 2012 00:00:00
24	Colombia	Tue 18 Dec 2012 00:00:00	Sat 22 Dec 2012 00:00:00
25	Ecuador	Mon 25 Feb 2013 00:00:00	Thu 21 Mar 2013 00:00:00
26	Ecuador	Mon 15 Oct 2012 00:00:00	Tue 23 Oct 2012 00:00:00
27	Paraguay	Wed 30 May 2012 00:00:00	Fri 15 Jun 2012 00:00:00
28	Paraguay	Mon 17 Sep 2012 00:00:00	Sun 7 Oct 2012 00:00:00
29	Argentina	Thu 10 May 2012 00:00:00	Wed 23 May 2012 00:00:00
30	Argentina	Wed 25 Apr 2012 00:00:00	Wed 2 May 2012 00:00:00
31	Argentina	Wed 16 May 2012 00:00:00	Sat 15 Sep 2012 00:00:00
32	Argentina	Tue 25 Sep 2012 00:00:00	Tue 16 Oct 2012 00:00:00
33	Argentina	Fri 2 Mar 2012 00:00:00	Fri 9 Mar 2012 00:00:00
34	Argentina	Wed 1 Aug 2012 00:00:00	Mon 20 Aug 2012 00:00:00
35	Brazil	Tue 18 Dec 2012 00:00:00	Sat 22 Dec 2012 00:00:00
36	Brazil	Thu 27 Dec 2012 00:00:00	Thu 28 Mar 2013 00:00:00
37	Chile	Sat 30 Jun 2012 00:00:00	Sun 30 Dec 2012 00:00:00
38	Chile	Mon 18 Jun 2012 00:00:00	Tue 26 Jun 2012 00:00:00
39	Chile	Thu 19 Jul 2012 00:00:00	Wed 22 Aug 2012 00:00:00
40	Chile	Sun 28 Oct 2012 00:00:00	Tue 13 Nov 2012 00:00:00

Figure 6. 5 Bursts of civil unrest in Fig. 6.2. Left column indicates the country where the burst occurred. Middle column shows start time of the burst and right column shows end time.

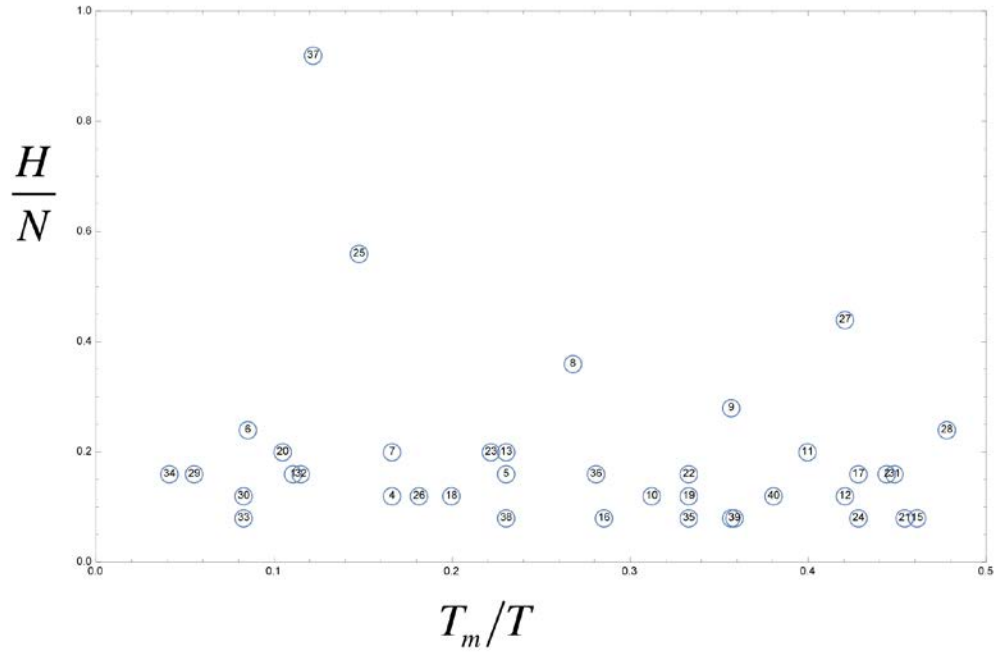


Figure 6. 6 Corresponding positions in Fig. 6.2. Result of the bursts features of civil unrest extracted from sequence in Fig. 6.5. The number inside the circle is the index of bursts, the same with row number in Figure 6.5. The circles mark the ratio of T_m/T and H/N.

In the two figures above, it is shown that bursts of events from different countries can have diverse distribution of their feature points, which is hard to explain by the standard SIR model and other agent-based models with only contagion process.

References

1. Ladyman, J., Lambert, J. & Wiesner, K. What Is a Complex System ? *Eur. J. Philos. Sci.* **3,** 33–67 (2011).

2. Bagnold, R. A. & Taylor, G. Motion of Waves in Shallow Water. Interaction between Waves and Sand Bottoms. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **187,** 1–18 (1946).

3. Kocurek, G., Townsley, M., Yeh, E., Havholm, K. G. & Sweet, M. L. Dune and Dune-Field Development on Padre Island, Texas, with Implications for Interdune Deposition and Water-Table-Controlled Accumulation. *SEPM J. Sediment. Res.* **Vol. 62,** 622–635 (1992).

4. Werner, B. T. & Gillespie, D. T. Fundamentally Discrete Stochastic-model for Wind Ripple Dynamics. *Phys. Rev. Lett.* **71,** 3230–3233 (1993).

5. Werner, B. T. Complexity in Natural Landform Patterns. *Sci.* **284,** 102–104 (1999).

6. Stolum, H.-H. River Meandering as a Self-Organization Process. *Sci.* **271,** 1710–1713 (1996).

7. McDonald, M., Suleman, O., Williams, S., Howison, S. & Johnson, N. F. Impact of Unexpected Events, Shocking News, and Rumors on Foreign Exchange Market Dynamics. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **77,** 046110 (2008).

8. Johnson, N. F. Pattern in Escalations in Insurgent and Terrorist Activity. *Sci.* **333,** 81–84 (2011).

9. Morgenstern, A. P. *et al.* Resource Letter MPCVW-1: Modeling Political Conflict, Violence, and Wars: A Survey. *Am. J. Phys.* **81,** 805 (2013).

10. Manrique, P. *et al.* Context matters: Improving the Uses of Big Data for Forecasting Civil Unrest: Emerging Phenomena and Big Data. *IEEE International Conference on Intelligence and Security Informatics* 169–172 (IEEE, 2013). doi:10.1109/ISI.2013.6578812

11. Zhao, Z. *et al.* Effect of Social Group Dynamics on Contagion. *Phys. Rev. E* **81,** 056107 (2010).

12. Whitesides, G. M. & Grzybowski, B. Self-assembly at All Scales. *Sci.* **295,** 2418–21 (2002).

13. Johnson, N. F. *Two's Company, Three is Complexity: A Simple Guide to the Science of All Sciences*. (Oneworld, 2007).

14. Mahan, G. D. *Many-Particle Physics*. (Springer Science & Business Media, 2000).

15. Medvinsky, A. B., Petrovskii, S. V., Tikhonova, I. A., Malchow, H. & Li, B.-L. Spatiotemporal Complexity of Plankton and Fish Dynamics. *SIAM Rev.* **44,** 311–370 (2002).

16. Zuzek, L. G. A., Stanley, H. E. & Braunstein, L. A. Epidemic Model with Isolation in Multilayer Networks. *Sci. Rep.* **5,** 12151 (2015).

17. Qi, H. *et al.* New Online Ecology of Adversarial Aggregates: ISIS and Beyond. *Sci.* **Reviewing,** (2015).

18. Johnson, N. *et al.* Financial Black Wwans Driven by Ultrafast Machine Ecology. *Physics (College. Park. Md).* 1–18 (2012). doi:10.2139/ssrn.2003874

19. Nicolis, G. & Wiley, J. *Self-Organization in Nonequilibrium Systems* (A Wiley-Interscience, 1982)

20. Mogul, J. Emergent (Mis) Behavior vs . Complex Software Systems Introduction. *Proc. First Eur. Conf. Comput. Syst.* 293–304 (2006). doi:10.1145/1217935.1217964

21. Sutula, T. P. & Dudek, F. E. Unmasking Recurrent Excitation Generated by Mossy Fiber Sprouting in the Epileptic Dentate Gyrus: An Emergent Property of a Complex System. *Prog. Brain Res.* **163,** 541–63 (2007).

22. Crutchfield, J. P. & Young, K. Inferring Statistical Complexity. *Phys. Rev. Lett.* **63,** 105–108 (1989).

23. Coco, G. & Murray, a. B. Patterns in the Sand: From Forcing Templates to Self-organization. *Geomorphology* **91,** 271–290 (2007).

24. Murray, A. B. Reducing Model Complexity for Explanation and Prediction. *Geomorphology* **90,** 178–191 (2007).

25. Murray, A. B. *et al.* Geomorphology, Complexity, and the Emerging Science of the Earth's Surface. *Geomorphology* **103,** 496–505 (2009).

26. Matheny, J. *Test and Evaluation in Ace and Osi Iarpa*. (Office of the Director of National Intelligence, 2013).

27. Choi, H. & Varian, H. Predicting the Present with Google Trends. *Econ. Rec.* **88,** 2–9 (2012).

28. Baker, S. R. & Fradkin, A. What Drives Job Search? Evidence from Google Search Data. *Tech. Report of Stanford Univ.* **94305,** (2011).

29. Askitas, N. & Zimmermann, K. F. Google Econometrics and Unemployment Forecasting. *Appl. Econ. Q.* **55,** 107–120 (2009).

30. Keegan, B., Ahmed, M. A., Williams, D., Srivastava, J. & Contractor, N. Dark Gold: Statistical Properties of Clandestine Networks in Massively Multiplayer Online Games. *2010 IEEE Second Int. Conf. Soc. Comput.* 201–208 (2010). doi:10.1109/SocialCom.2010.36

31. Krzakala, F., Ricci-Tersenghi, F., Sherrington, D. & Zdeborová, L. No Spin Glass Phase in the Ferromagnetic Random-field Random-temperature Scalar Ginzburg–Landau model. *J. Phys. A Math. Theor.* **44,** 042003 (2011).

32. Ashkenazi, J., Barnes, S., Zuo, F., Vezzoli, G. *High-Temperature Superconductivity: Physical Properties, Microscopic Theory, and Mechanisms*. (Springer Science & Business Media, 2012).

33. Rechenberg, H. *The Historical Development of Quantum Theory, Volumes 1-2*. (Springer Science & Business Media, 2001).

34. Nakamura, K. Review of Particle Physics. *J. Phys. G Nucl. Part. Phys.* **37,** 075021 (2010).

35. Johnson, N. F. *et al.* Human Group Formation in Online Guilds and Offline Gangs Driven by a Common Team Dynamic. *Phys. Rev. E* **79,** 066117 (2009).

36. Stumpf, M. P. H. & Porter, M. A. Mathematics. Critical Truths About Power Laws. *Sci.* **335,** 665–6 (2012).

37. Helbing, D. *et al.* Saving Human Lives: What Complexity Science and Information Systems Can Contribute. *J. Stat. Phys.* **158,** 735–781

38. Song, C., Koren, T., Wang, P. & Barabási, A. L. Modelling the Scaling Properties of Human Mobility. *Nat. Phys.* **6,** 818–823 (2010).

39. Vespignani, A. Predicting the Behavior of Tech-social Systems. *Sci.* **325,** 425–428 (2009).

40. Makse, H. A., Havlin, S. & Stanley, H. E. Modelling Urban Growth Patterns. *Nature* **377,** 608–612 (1995).

41. Krings, G., Calabrese, F., Ratti, C. & Blondel, V. D. Urban Gravity: A Model for Inter-city Telecommunication Flows. *J. Stat. Mech.-Theor. Exp.* L07003 (2009).

42. Gabaix, X., Gopikrishnan, P., Plerou, V. & Stanley, H. E. A Theory of Power-law Distributions in Financial Market Fluctuations. *Nature* **423,** 267–70 (2003).

43. Manrique, P. D. *et al.* Anomalous Contagion and Renormalization in Dynamical Networks. **Writing,** 1–4 (2015).

44. Richardson, Lewis Fry. *Statistics of Deadly Quarrels*, 1809-1949. (Inter-university Consortium for Political and Social Research, 1960). http://doi.org/10.3886/ICPSR05407.v1

45. Watts, D. J. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. (Princeton University Press, 1999).

46. Granovetter, M. S. Threshold Models of Collective Behavior. *Am. J. Sociol.* **83,** 1420–1443 (1978).

47. Lichbach, M. I. Nobody Cites Nobody else: Mathematical Models of Domestic Political Conflict. *Def. Econ.* **3,** 341–357 (1992).

48. Marwell, G., Oliver, P. E. & Prahl, R. Social Networks and Collective Action: A Theory of the Critical Mass. III. *Am. J. Sociol.* **94,** 502–534 (1988).

49. Tilly, C. & Tarrow, S. *Contentious Politics*. (Oxford University Press, 2006).

50. McCarthy, J. D. & Zald, M. N. Resource Mobilization and Social Movements: A Partial Theory. *Am. J. Sociol.* **82,** 1212–1241 (1977).

51. Della Porta, D. & Diani, M. *Social Movements: An Introduction*. (Blackwell Publishing, 2006).

52. Della Porta, D. & Tarrow, S. *Transnational Processes and Social Activism: An Introduction*. (Rowman & Littlefield, 2005).

53. Palla, G., Barabasi, A. L. & Vicsek, T. Quantifying Social Group Evolution. *Nature* **446,** 664–667 (2007).

54. Macy, M. W. Chains of Cooperation: Threshold Effects in Collective Action. *Am. Sociol. Rev.* **56,** 730 (1991).

55. Biggs, M. Strikes as Forest Fires: Chicago and Paris in the Late Nineteenth Century. *Am. J. Sociol.* **110,** 1684–1714 (2005).

56. Boot, M. *Invisible Armies: An Epic History of Guerrilla Warfare from Ancient Times to the Present*. (Liveright, 2013).

57. Mungiu-Pippidi, A. & Munteanu, I. Moldova's 'Twitter Revolution'. *J. Democr.* **20,** 136–142 (2009).

58. Morozov, E. Iran Downside to the 'Twitter Revolution'. *Dissent* **56,** 10–14 (2009).

59. Guzik, K. Discrimination by Design: Predictive Data Mining as Security Practice in the United States 'War on Terrorism'. *Surveillance & Society* **7,** 3–20 (2009).

60. Valenzuela, S., Arriagada, A. & Scherman, A. The Social Media Basis of Youth Protest Behavior: The Case of Chile. *J. Commun.* **62,** 299–314 (2012).

61. Lazer, D. Computational Social Science. *Sci.* **323,** 721–723 (2009).

62. Cohn, J. F. & Tronick, E. Z. Mother–infant Face-to-face Interaction: Influence is Bidirectional and Unrelated to Periodic Cycles in Either Partner's Behavior. *Dev. Psychol.* **24,** 386–392 (1988).

63. Francisco, R. A. *Collective Action Theory and Empirical Evidence* (Springer, New York. *Dyn. Confl.*, 2010).

64. Spagat, M., Mack, A., Cooper, T. & Kreutz, J. Estimating War Deaths: An Area of Contestation. *J. Confl. Resolut.* **53,** 934–950 (2009).

65. Beddington, J. The Future of Computer Trading in Financial Markets, *Foresight Final Project Report*. (2013).

66. Johnson, N. F. *et al. Financial Market Complexity* (Publ. Oxford University Press, 2013).

67. Ekas, N., Haltigan, J. D. & Messinger, D. S. The Dynamic Still-face Effect: Do Infants Decrease Bidding Over Time When Parents are Not Responsive? *Dev. Psychol.* **49,** 1027–1035 (2013).

68. Kappler, K. E. & Kaltenbrunner, A. The Power Laws of Violence against Women: Rescaling Research and Policies. *PLoS One* **7,** e40289 (2012).

69. Song, C., Qu, Z., Blumm, N. & Barabasi, A. L. Limits of Predictability in Human Mobility. *Sci.* **327,** 1018–1021 (2010).

70. Wuchty, S., Jones, B. F. & Uzzi, B. The Increasing Dominance of Teams in Production of Knowledge. *Sci.* **18,** 1036–1039 (2007).

71. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A. L. Understanding Individual Human Mobility Patterns. *Nature* **453,** 779–782 (2008).

72. Preis, T., Schneider, J. J. & Stanley, H. E. Switching Processes in Financial Markets. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 7674–7678 (2011).

73. Horgan, J. Arlington, Virginia: Office of Naval Research Code 30. *From Bomb To Bomb Maker. A Soc. Netw. Anal. Socio Psychol. Cult. Dyn. Ied Process Final Report.*

74. Kenney, M. Organisational Adaptation in an Activist Network: Social networks, Leadership, and Change in Al-Muhajiroun. *Appl. Ergon.* **44,** 739e747 (2013).

75. Johnson, N. F. Bias in Epidemiological Studies of Conflict Mortality. *J. Peace Res.* **45,** 653–659 (2008).

76. Bohorquez, J. C. Common Ecology Quantifies Human Insurgency. *Nature* **462,** 911–914 (2009).

77. Clauset, A., Young, M. & Gleditsch, K. S. On the Frequency of Severe Terrorist Events. *J. Confl. Resolut.* **51,** 58–87 (2007).

78. Clauset, A. & Gleditsch, K. The Developmental Dynamics of Terrorist Organizations. *PLoS One* **7,** e48633 (2012).

79. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Martino Fine Books, 2012).

80. Pareto, V. *Cours d'Économie Politique*. (Librairie Droz, 1964).

81. Barabási, A. Emergence of Scaling in Random Networks. *Sci.* **286,** 509–512 (1999).

82. Barabási, a. L., Albert, R. & Jeong, H. Mean-field Theory for Scale-free Random Networks. *Phys. A Stat. Mech. its Appl.* **272,** 173–187 (1999).

83. Grabowicz, P. a. & Eguíluz, V. M. Heterogeneity Shapes Groups Growth in Social Online Communities. *EPL (Europhysics Lett.* **97,** 28002 (2012).

84. Gibrat, R. *Les Inégalités Economiques.* (Recueil Sirey, 1931).

85. Rozenfeld, H. D. *et al.* Laws of Population Growth. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 18702–7 (2008).

86. Rybski, D., Buldyrev, S. V, Havlin, S., Liljeros, F. & Makse, H. A. Scaling Laws of Human Interaction Activity. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12640–5 (2009).

87. Barabási, a. L. *et al.* Evolution of the Social Network of Scientific Collaborations. *Phys. A Stat. Mech. its Appl.* **311,** 590–614 (2002).

88. Johnson, N. F. *et al.* Simple Mathematical Law Benchmarks Human Confrontations. *Sci. Rep.* **3,** 3463 (2013).

89. Cohn, J. F. & Tronick, E. Z. Mother-infant Face-to-face Interaction: Influence is Bidirectional and Unrelated to Periodic Cycles in Either Partner's Behavior. *Developmental Psychology* **24,** 386–392 (1988).

90. Clauset, A., Shalizi, C. & Newman, M. Power-Law Distributions in Empirical Data. *Siam Rev.* **51,** 661–703 (2009).

91. Dutton, J. M. & Thomas, A. Treating Progress Functions as a Managerial Opportunity. *Acad. Manag. Rev.* **9,** 235–247 (1984).

92. Argote, L. & Epple, D. Learning Curves in Manufacturing. *Sci.* **247,** 920–924 (1990).

93. Saraswat, S. P. & Gorgone, J. T. Organizational Learning Curve in Software Installation: An Empirical Investigation. *Inf. Manag.* **19,** 53–59 (1990).

94. Mottrie, A. *et al.* Impact of the Learning Curve on Perioperative Outcomes in Patients Who Underwent Robotic Partial Nephrectomy for Parenchymal Renal Tumours. *Eur. Urol.* **58,** 127–32 (2010).

95. Crossman, E. R. F. W. A Theory of the Acquisition of Speed-skill. *Ergonomics* **2,** 153–166 (1959).

96. Johnson, E. J., Bellman, S. & Lohse, G. L. Cognitive Lock-In and the Power Law of Practice. *J. Mark.* **67,** 62–75 (2003).

97. Ali, A. H. The Power of Social Media in Developing Nations : New Tools for Closing the Global Digital Divide and Beyond. *Harv. Hum. Rights J.* **24,** 185–219 (2011).

98. Horgan J. & J., H. *From Bomb to Bomb-maker: A Social Network Analysis of the Socio-Psychological and Cultural Dynamics of the IED Process - Final Report*. (2013).

99. Segerberg, A. & Bennett, W. L. Social Media and the Organization of Collective Action: Using Twitter to Explore the Ecologies of Two Climate Change Protests. *Commun. Rev.* **14,** 197–215 (2011).

100. Mourtada, R. & Salem, F. Civil movements: The Impact of Facebook and Twitter. *Arab Soc. Media Rep.* **1,** 1–30 (2011).

101. Radinsky, K. & Horvitz, E. Mining the Web to Predict Future Events. *Proc. sixth ACM Int. Conf. Web search data Min. - WSDM '13* 255 (2013). doi:10.1145/2433396.2433431

102. Asur, S. & Huberman, B. A. Predicting the Future with Social Media. (2010). doi:10.1016/j.apenergy.2013.03.027

103. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. & Watts, D. J. Predicting Consumer Behavior with Web Search. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 17486–90 (2010).

104. Kallus, N. Predicting Crowd Behavior with Big Public Data. (2014). doi:10.1145/2567948.2579233

105. Scheffer, M. *et al.* Early-warning Signals for Critical Transitions. *Nature* **461,** 53–9 (2009).

106. Lenton, T. M. Tipping Elements in the Earth's Climate System. *Proc. Natl Acad. Sci. USA* **105,** 1786–1793 (2008).

107. Scheffer, M. Catastrophic Shifts in Ecosystems. *Nature* **413,** 591–596 (2001).

108. Duralappah, A. K. *Ecosystems and Human Well-being: Synthesis Report*. (World Resources Institute, 2005).

109. Kambhu, J., Weidman, S. & Krishnan, N. *New Directions for Understanding Systemic Risk: A Report on a Conference Cosponsored by the Federal Reserve Bank of New York and the National Academy of Sciences*. (The National Academies Press, 2007).

110. May, R. M., Levin, S. A. & Sugihara, G. Ecology for Bankers. *Nature* **451,** 893–895 (2008).

111. Venegas, J. G. Self-organized Patchiness in Asthma as a Prelude to Catastrophic shifts. *Nature* **434,** 777–782 (2005).

112. Litt, B. Epileptic Seizures May Begin Hours in Advance of Clinical Inset: A report of five patients. *Neuron* **30,** 51–64 (2001).

113. McSharry, P. E., Smith, L. A. & Tarassenko, L. Prediction of Epileptic Seizures: Are Nonlinear Methods Relevant? *Nat. Med.* **9,** 241–242 (2003).

114. Kuznetsov, Y. A. *Elements of Applied Bifurcation Theory*. (Springer-Verlag, 2004).

115. Schroeder, M. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. (Dover Publications, 2009).

116. Wissel, C. A Universal Law of the Characteristic Return Time Near Thresholds. *Oecologia* **65,** 101–107 (1984).

117. Guttal, V. & Jayaprakash, C. Changing Skewness: An Early Warning Signal of Regime Shifts in Ecosystems. *Ecol. Lett.* **11,** 450–460 (2008).

118. Scheffer, M., Westley, F. & Brock, W. Slow Response of Societies to New Problems: Causes and Costs. *Ecosystems* **6,** 493–502 (2003).

119. Holyst, J. A., Kacperski, K. & Schweitzer, F. Social Impact Models of Opinion Dynamics. *Annu. Rev. Comput. Phys.* **9,** 253–273 (2002).

120. Ramakrishnan, N. *et al.* 'Beating the news' with EMBERS: Forecasting Civil Unrest using Open Source Indicators. 1–17 (2014).

121. Howard, P. N. & Hussain, M. M. The Role of Digital Media. *J. Democr.* **22,** 35–48 (2011).

122. Qi, H., Velasquez, N., Manrique, P., Morgenstern, A. & Johnson, D. Prediction of Digitally Enabled Uprisings. *Sci.* Reviewing (2015)

123. Bennett, W. L. & Segerberg, A. the Logic of Connective Action. *Information, Commun. Soc.* **15,** 739–768 (2012).

124. Epstein, J. M. *Nonlinear Dynamics, Mathematical Biology, and Social Science*. (Westview Press, 1997).

125. Press, S. The *t*-ratio Distribution. *J. Am. Stat. Assoc.* **64,** 242–252 (1969).

126. Aureli, F. *et al.* Fission Fusion Dynamics: New Research Frameworks. *Curr. Anthropol.* **49,** 627–654 (2008).

127. Hedström, P. & Ylikoski, P. Causal Mechanisms in the Social Sciences. *Annu. Rev. Sociol.* **36,** 49–67 (2010).

128. Brock, W., Lakonishok, J. & Lebaron, B. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *J. Financ.* **47,** 1731–1764 (1992).

129. Lo, A. W., Mamaysky, H. & Wang, J. Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *J. Financ.* **55,** 1705–1765 (2000).

130. Lebaron, B. The Stability of Moving Average Technical Trading Rules on the Dow Jones Index. *Deriv. Use Trad. Regul.* **5,** 324–338 (2000).

131. Bates, D. S. The Crash of 87 - Was It Expected? The Evidence From Options Markets. *J. Financ.* **46,** 1009–1044 (1991).

132. Bates, D. S. Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options. *Rev. Financ. Stud.* **9,** 69–107 (1996).

133. Hens, T. & Schenk-Hoppe, K. R. *Handbook of Financial Markets: Dynamics and Evolution.* (North-Holland, 2009).

134. Lebaron, B. Some Relations Between Volatility and Serial Correlations in Stock-Market Returns. *J. Bus.* **65,** 199–219 (1992).

135. Panin, A. & Smith, L. Russian Students Targeted as Recruits by Islamic State - BBC News. *BBC News* (2015).

136. Estrada, E. *The Structure of Complex Networks: Theory and Applications*. (OUP Oxford, 2011).

137. Feng, L. *et al.* Competing for Attention in Social Media under Information Overload Conditions. *PLoS One* **10,** e0126090 (2015).

138. Tomasello, M. V, Perra, N., Tessone, C. J., Karsai, M. & Schweitzer, F. The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks. *Sci. Rep.* **4,** 5679 (2014).

139. Gracia, J. F., Onnela, J.-P., Barnett, M. L., Eguíluz, V. M. & Christakis, N. A. Spread of Pathogens in the Patient Transfer Network of US Hospitals. *arXiv*:1504.08343 (2015).

140. Watts, D. J., Muhamad, R., Medina, D. C. & Dodds, P. S. Multiscale, Resurgent Epidemics in a Hierarchical Metapopulation Model. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 11157–62 (2005).

141. Karsai, M., Perra, N. & Vespignani, A. Time Varying Networks and the Weakness of Strong ties. *Sci. Rep.* **4,** 4001 (2014).

142. Liu, S.-Y., Baronchelli, A. & Perra, N. Contagion Dynamics in Time-varying Metapopulation Networks. *Phys. Rev. E* **87,** 032805 (2013).

143. Karsai, M., Iñiguez, G., Kaski, K. & Kertész, J. Complex Contagion Process in Spreading of Online Innovation. *J. R. Soc. Interface* **11,** 20140694 (2014).

144. Onnela, J.-P. & Reed-Tsochas, F. Spontaneous Emergence of Social Influence in Online Systems. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 18375–80 (2010).

145. Lin, Y.-R., Bagrow, J. P. & Lazer, D. More Voices Than Ever? Quantifying Media Bias in Networks. *arXiv:1111.1227* (2011).

146. Palchykov, V., Kaski, K. & Kertész, J. Transmission of Cultural Traits in Layered Ego-centric Networks. *Condens. Matter Phys.* **17,** 33802 (2014).

147. Perra, N., Balcan, D., Gonçalves, B. & Vespignani, A. Towards a Characterization of Behavior-Disease Models. *PLoS One* **6,** e23084 (2011).

148. Jo, H.-H., Perotti, J. I., Kaski, K. & Kertész, J. Analytically Solvable Model of Spreading Dynamics with Non-Poissonian Processes. *Phys. Rev. X* **4,** 011041 (2014).

149. Asztalos, A., Sreenivasan, S., Szymanski, B. K. & Korniss, G. Cascading Failures in Spatially-embedded Random Networks. *PLoS One* **9,** e84563 (2014).

150. Murase, Y., Török, J., Jo, H.-H., Kaski, K. & Kertész, J. Multilayer Weighted Social Network Model. *Phys. Rev. E* **90,** 052810 (2014).

151. Sornette, D., Deschâtres, F., Gilbert, T. & Ageon, Y. Endogenous Versus Exogenous Shocks in Complex Networks: An Empirical Test Using Book Sale Rankings. *Phys. Rev. Lett.* **93,** 228701 (2004).

152. Deschâtres, F. & Sornette, D. Dynamics of Book Sales: Endogenous versus Exogenous Shocks in Complex Networks. *Phys. Rev. E* **72,** 016112 (2005).

153. Karsai, M., Kaski, K., Barabási, A.-L. & Kertész, J. Universal Features of Correlated Bursty Behaviour. *Sci. Rep.* **2,** 397 (2012).