

The Institute for Quantitative Social Science at Harvard University



# FORECASTING SOCIO-TECHNICAL SYSTEMS IN THE DATA SCIENCE AGE

Alessandro Vespignani @alexvespi Som MOBS LAB

LABORATORY FOR THE MODELING OF BIOLOGICAL AND SOCIO-TECHNICAL SYSTEMS

#### Northeastern University



*EPJ Data Science* covers a broad range of research areas and applications and particularly encourages contributions from techno-socio-economic systems, where it comprises those research lines that now regard the digital tracks of human beings as first-order objects for scientific investigation. Topics include, but are not limited to, human behavior, social interaction (including animal societies), economic and financial systems, management and business networks, socio-technical infrastructure, health and environmental systems, the science of science, as well as general risk and crisis scenario forecasting up to and including policy advice.

### Northeastern University

EPJ Data So a SpringerOpen Joint A SpringerOpen Joint	tience	Search EPJ Data Science 🗘 for	Go
			Advanced search
Home Articles	Authors Reviewers About this journal My EPJ	Data Science	
General information	Regular article Open Access Partisan asymmetries in online political activity		(A) Samia and
Editorial Board	Michael D Conover, Bruno Gonçalves, Alessandro Flar	mmini, Filippo Menczer	springer
FAQ	Abstract   Full text   PDF   ePUB		
Article processing charge FAQ	Regular article Open Access Social dynamics of Digg		EPJ.org
Contact us	EPJ Data Science 2012, 1:5 (18 June 2012)		••••
Privacy policy	Full text   PDF   ePUB		
EPJ Data Sc encourages lines that no investigation animal socie technical inf general risk	Regular article Open Access   Effects of time window size and placement on t   Gautier Krings, Márton Karsai, Sebastian Bernhardss   EPJ Data Science 2012, 1:4 (18 May 2012)   Abstract   Full text   PDF   ePUB   Regular article Open Access   Highly accessed   Positive words carry less information than negative   David Garcia, Antonios Garas, Frank Schweitzer   EPJ Data Science 2012, 1:3 (18 May 2012)   Abstract   Full text   PDF   ePUB	the structure of an aggregated com on, Vincent D Blondel, Jari Saramäki ative words	icularly those research scientific ction (including vorks, socio- as well as
	Regular article Open Access Long trend dynamics in social media Chunyan Wang, Bernardo A Huberman EPJ Data Science 2012, 1:2 (18 May 2012) Abstract   Full text   PDF   ePUB		

### Northeastern University

#### NETWORK SCIENCE

#### Welcome to Cambridge Journals Online

To access subscriptions and personalised features please log in or register Home > NETWORK SCIENCE

#### NETWORK SCIENCE

Register for an Account

Ð.

TOOLS

Announcing a new journal, coming in Spring 2013 Published by Cambridge University Press

#### EDITORS

Lada Adamic, *University of Michigan, USA* - Editor for Information Science Ulrik Brandes, *University of Konstanz, Germany* - Editor for Computer Science and Mathematics Noshir Contractor, *Northwestern University, USA* - Editor for Communication, Engineering, and Management Sanjeev Goyal, *University of Cambridge, UK* - Editor for Economics Garry Robins, *University of Melbourne, Australia* - Editor for Psychology and Political Science Thomas Valente, *University of Southern California, USA* - Editor for Public Health and Medicine Alessandro Vespignani, *Northeastern University, USA* - Editor for Physics Stanley Wasserman, *Indiana University USA* - Editor for Statistics and Sociology and Coordinating Editor

D. Balcan F. Ciulla **B.** Goncalves M.Gomes H. Hu A. Pastore N. Perra D. Mocanu L. Rossi Qian Zhang K.Borner J.Sherman L.Weng

V. Colizza L.Cappa • C. Cattuto • P. Bajardi • C.Gioannini D.Paolotti C. Poletto • M.Quagiotto • Suyu Liu • J. Ramasco • M. Tizzoni • V. Van den Broeck M.Roncaglione

Andrea Baronchelli S. Merler

- R. Pastor Satorras
- A.J.Valleron
- S.Meloni

• M. Ajelli

- A. Barrat
- M. Barthelemy
- Y.Moreno
- M. Karsai
- N.Samay
- F. Menczer
  - A. Flammini

NIH, DTRA, NAKFI, NSF, ARL, IARPA, Lilly foundation, Abbott, EC-FET, ERC, **CRT**-foundation, ISI foundation





**Prediction**= probabilistic on the future based on what is known today (generally mostly influenced by initial conditions).

**Forecast** = best prediction given the present knowledge on the system.

**Projection** = attempt to describe what would happen under certain assumptions and hypotheses (what if)

#### Standard MOBS LAB





# THE SOCIAL COMPONENT

Human science requires a deep understanding of mental processes and reasoning (Psychology and cognitive).

Social science is hard because it involves the infinite psychological and cognitive reactions of individuals. We're too complex and unpredictable.

From "social atom" or "social molecules" (i.e. small social groups) to the quantitative analysis of "social aggregate states" (Lundberg, Moreno).

"social aggregate states" = large-scale social systems consisting of millions of individuals that can be characterized statistically in space (geographic and social) and time.



# THE SOCIAL COMPONENT

Human science requires a deep understanding of mental processes and reasoning (Psychology and cognitive).

Social science is hard because it involves the infinite psychological and cognitive reactions of individuals. We're too complex and unpredictable.

From "social atom" or "social molecules" (i.e. small social groups) to the quantitative analysis of "social aggregate states" (Lundberg, Moreno).

"social aggregate states" = large-scale social systems consisting of millions of individuals that can be characterized statistically in space (geographic and social) and time.







« »

# THE SOCIAL COMPONENT

Human science requires a deep understanding of mental processes and reasoning (Psychology and cognitive).

Social science is hard because it involves the infinite psychological and cognitive reactions of individuals. We're too complex and unpredictable.

From "social atom" or "social molecules" (i.e. small social groups) to the quantitative analysis of "social aggregate states" (Lundberg, Moreno).

"social aggregate states" = large-scale social systems consisting of millions of individuals that can be characterized statistically in space (geographic and social) and time.



States MOBS LAB

# **BIG DATA**

- Big data has fueled spectacular advances in the natural sciences over the last 100 years.
- So what is different now?
- Every 1.2 years, more human-driven socioeconomic data is produced than during all previous history
- Embedded within the data are the raw ingredients for understanding socio-technical and socio-economic systems
- The focus is on understanding these data sets in a statistical sense and more deeply the real world processes which produced the data



### MICRO-SCALE: RFID, BLUETHOOTH, SOCIOBADGE



Sociopatterns experiment in school





« »

Paris Fête de la Musique / mouvements des mobiles

12:43 21/06/2008





## **AIRLINE DATA**

### viz by Aron Koblin



Wednesday, November 6, 13

« »

## **AIRLINE DATA**







Standard MOBS LAB

**« »** 

# A COMPLEX WORLD

Large numbers of heterogeneous individuals

Multiple time and length scales

Non-linearity, threshold effects, discreteness, cooperation

Systemic approach/complex systems



## **NETWORK THINKING**





Black death in1347: a continuous diffusion process

SARS epidemics: a discrete network driven process





### GLEAM

### GLOBAL EPIDEMIC AND MOBILITY MODEL



SMOBS LAB





«

**>>** 





#### Wednesday, November 6, 13

Standard MOBS LAB

# WHAT IS UNDER THE HOOD

### Stochastic Intra population disease evolution: chain binomial process

– REPEAT

- $\circ$  CALL  $\texttt{RANBin}(S,\beta I/N)$  and  $\texttt{RANBin}(I,\mu)$
- S=S-RANBin $(S, \beta I/N)$
- $\circ \quad \texttt{I=I+RANBin}(S,\beta I/N)\texttt{-RANBin}(I,\mu)$
- $\circ \quad \texttt{R=R+RANBin}(I,\mu)$
- $\circ \quad t=t+\Delta t$
- $\circ$  PRINT S, I, R, t

```
- UNTIL I = 0
```



### Stochastic Inter population dynamics:

- data driven Explicit stochastic simulation of slow mode traveling patterns.
- Time-Scale separation through effective force of infection for fast commuting patterns
  - > 3,500 single populations coupled models
  - > 40,000-100,000 stochastic discrete equations

Standard MOBS LAB



#### States MOBS LAB

### **REAL TIME FORECAST FOR THE H1N1PDM (2009)**

Seasonal transmission potential and activity peaks of the new influenza A(HINI): a Monte Carlo likelihood analysis based on human mobility

Duygu Balcan<sup>†1,2</sup>, Hao Hu<sup>†1,2,3</sup>, Bruno Goncalves<sup>†1,2</sup>, Paolo Bajardi<sup>†4,5</sup>, Chiara Poletto<sup>†4</sup>, Jose J Ramasco<sup>4</sup>, Daniela Paolotti<sup>4</sup>, Nicola Perra<sup>1,6,7</sup>, Michele Tizzoni<sup>4,8</sup>, Wouter Van den Broeck<sup>4</sup>, Vittoria Colizza<sup>4</sup> and Alessandro Vespignani<sup>\*1,2,4</sup>

Published: 10 September 2009 BMC Medicine 2009, 7:45 doi:10.1186/1741-7015-7-45 Received: 31 July 2009 Accepted: 10 September 2009



Transmissibility Generation time Seasonality scaling

Monte-Carlo Likelihood estimate



#### Plausible parametrization

Parameter	Description	Value	Sensitivity Analysis Range
r <sub>β</sub>	Relative infectiousness of asymptomatic individuals	0.5	0.2 - 0.8
p <sub>a</sub>	Probability of becoming an asymptomatic individual	0.33	0.33 - 0.5
p,	Probability of becoming a traveling symptomatic individual	0.5	0.4 - 0.6
β	Transmission rate	$\mu^{-1}R_0/(1 - p_a - r_{\beta}p_0)$	
01 <sub>max</sub>	Maximal seasonality rescaling	1.1	1.0 - 1.1

#### Standard MOBS LAB

« »

### Monte Carlo likelihood parameters' estimate

Backtrack of the number of infections in Mexico from case importation and transportation data. (Fraser et al. Science 2009,324,1557)

Numerical generation of thousands of infection trees (Importation/generation of the first symptomatic infectious in a given subpopulation).



Statistical distribution of the seeding time after >10<sup>3</sup>-10<sup>4</sup> numerical stochastic realizations for each set of the parameters.

### Monte Carlo likelihood parameters' estimate



Backtrack of the number of infections in Mexico from case importation and transportation data. (Fraser et al. Science 2009,324,1557)

Numerical generation of thousands of infection trees (Importation/generation of the first symptomatic infectious in a given subpopulation).



Statistical distribution of the seeding time after >10<sup>3</sup>-10<sup>4</sup> numerical stochastic realizations for each set of the parameters.



### Monte Carlo likelihood parameters' estimate

Stop Record

EE2.pptx

Backtrack of the number of infections in Mexico from case importation and transportation data. (Fraser et al. Science 2009,324,1557)

Numerical generation of thousands of infection trees (Importation/generation of the first symptomatic infectious in a given subpopulation).



Sec. 10

Statistical distribution of the seeding time after >10<sup>3</sup>-10<sup>4</sup> numerical stochastic realizations for each set of the parameters.



History

The Poisson Experiment

Bookmarks

www.math.bme.hu/~nandori/Virtual\_lab/stat/applets/F

# **STOCHASTIC FORECAST OUTPUT SETS**

Ab-initio estimates of the epidemic timeline in each country or urban area without assumptions on case importation.

Model calibration based on case importation and epidemic arrival time in first infected countries

Specific for H1N1 Inclusion of inter-country commuting data Traffic reduction to and from Mexico

Monte-Carlo likelihood analysis on >6x 106 synthetic epidemic (on supercomputer). Each realization produces 300-500 MB.

Data from countries reporting the first cases (93 countries by June 18.

Transmissibility determined with the first 12 countries seeded from Mexico.

Seasonality signal by using 60 countries (determine the time window worth of data).



	Baseline	Reference	Pre-exposure immunity
Monte-Carlo likelihood calibration	~	~	~
Worldwide air traffic reduction after April 25 <sup>th</sup> , 2009			~
Real vaccinations campaigns in the northern hemisphere			~
Pre-exposure immunity of the elderly			~
Full mobility database	~	~	~

States MOBS LAB

Feb 18 2009

La Gloria Sao Paulo Mexico City Rio De Janeiro San Juan Bogota	

#### Infection tree

#### Standard MOBS LAB

« »



Wednesday, November 6, 13



Wednesday, November 6, 13



#### Wednesday, November 6, 13

Standard MOBS LAB

« »



Wednesday, November 6, 13

Standard MOBS LAB

« »

# LINEAR-THINKING DOES NOT WORK

### Non-linear Behavior

- an infinitesimal perturbation or change in pattern, sets up macroscopic changes and generate new patterns (and viceversa)
- Surprising phase transitions: one "state" of the system may give way to another, with no precursors or warning
- Counter intuitive change in collective behavior.
- Systemic and interdependent risk
- Global perspective

### **Computational thinking**

- Computational thinking as the "macroscope" for the mind in exploring collective surprises.
- Simulation as an analytical tool for the quantitative understanding





## **NON-LINEAR THINKING AT WORK**

#### OPEN BACCESS Freely available online

PLos one

Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic

Paolo Bajardi<sup>1,1,8</sup>, Chiara Poletto<sup>1,9</sup>, Jose J. Ramasco<sup>9</sup>, Michele Tizzoni<sup>1,4</sup>, Vittoria Colizza<sup>3,6,7</sup>, Alessandro Vespignani<sup>8,9,10</sup>,



Non-linear features of traffic restriction

Large traffic reduction leads to small delays of the epidemic peak.

[Hollingsworth et al 2006; Brownstein et al. 2006; Cooper et al. 2006; Scalia Tomba et al. 2008; Gautreau et al. 2008; Colizza et al.2007; Bajardi et al. 2011]

 $\Delta t \sim - \ln (1-\alpha)$ 

States MOBS LAB

# IS A SUCCESSFUL CONTAGION PROCESS IN A SINGLE SUBPOPULATION ABLE TO SPREAD IN A COLLECTION OF SUBPOPULATIONS?



Infection/information is carried by Particles/agents diffusing interacting with rate p from sub-population to sub-population:

p=0: no spreading

p=1: equivalent to a fully mixed situation



## **MULTI-POPULATION GLOBAL THRESHOLD**



Basic equations describing the dynamics



Basic generation equation for the invasion of new populations

$$\partial_t N_{kk} = -\sigma_k N_{kk} + \tau_k k \sum_{k'} N_{kk'} P(k'|k)$$
  
$$\partial_t N_{kk'} = \sigma_{kk'} N_{kk} - \tau_k N_{kk'} \quad ,$$

$$D_k^n = \frac{2\overline{N}(R_0 - 1)^2}{R_0^2 \langle k^{1+\theta} \rangle \langle k \rangle} kP(k) \sum_{k'} D_{k'}^{n-1}(k' - 1) \times \left[ \frac{k'^{1+\theta}\sigma_{k'k}}{(1+\rho_{k'})\tau_{k'}} + \frac{k^{1+\theta}\sigma_{kk'}}{(1+\rho_k)\tau_k} \right]$$

#### 🗱 MOBS LAB

« »

# INVASION THRESHOLD AS A FUNCTION OF INTERACTION RATE AND DURATION OF THE INTERACTION.





Analytical results

#### Numerical results

#### Standard MOBS LAB

### **TIPPING POINT AS A FUNCTION OF HUMAN MOBILITY**



### Tenfold traffic reduction is needed to achieve containment effects



## FROM GEOGRAPHY TO SOCIAL SPACE





#### Geographical areas/census Mobility



# FROM GEOGRAPHY TO SOCIAL SPACE





Structured communities in the abstract social space define by knowledge and information



« »





## MICROBLOGGING (I.E.TWITTER)



**Twitter Signal** 

#### Standard MOBS LAB



Wednesday, November 6, 13





Wednesday, November 6, 13

OPEN OR ACCESS Freely available online

#### The Twitter of Babel: Mapping World Languages through Microblogging Platforms



PLOS ONE

### Universal user activity





### **TURKEY PROTESTS**

Mapping Events Through Microblogging Platform

On May 28<sup>th</sup> 2013, a group of people began protesting against the disruption of Gezi Park in Istanbul, Turkey. The protest spread to the main Turkish cities, generating a massive confrontation with the current Turkish Government. Since the beginning, Twitter has been used as an effective media to coordinate and spread news about the protest. Following Twitter's most popular hashtags related to the protest, it is possible to have a spatiotemporal mapping of the Turkish events.







### TAP THE "GLOBAL CONVERSATION" TO PREDICT THE FUTURE

- Politics (emergence of consensus, election)
- Social collective phenomena (riots, political protests, etc.)
- Economics (stock market)

Many authors have pointed out, there are several challenges: • intrinsic biases,

- uneven sampling across location of interest
- causality assumption



# THIS IS NOWCAST !!!

We miss the microscopic generative foundations and models



**« »** 

Wednesday, November 6, 13

### **TWITTER NETWORK ON MAY 15TH PROTEST MOVEMENT IN SPAIN**

#### Moreno et al. (BIFI, Universidad de Zaragoza















1974-1984







1974





## **NETWORK MODELING**

### Connectivity driven models:

- Connectivity patterns as basic ingredient of models generation (Erdös-Rényi model to Logit models, p\*-models, Markov random graphs)
- Preferential attachment dynamical models
- [N.B. Topological properties merely represent a time-integrated perspective of the system]

### Relational event-based network analyses.

### Feedback models

#### Models generally considered in a time-scale separation regime

- Process dynamics decoupled by network evolution
- Frozen network (process time scale << network evolution time scale)</li>
- Random homogenous mixing (network evolution time scale << process time scale)</li>

#### Standard MOBS LAB

### **DEFINITION OF THE SIMPLEST NON-TRIVIAL NETWORK GENERATIVE MODEL**

- Encodes the heterogeneous activity of the nodes in one observable
- Generates fluctuating connectivity patterns
- Amenable of refinement
- Allowing the analytical treatment of dynamical processes

Perra, Goncalves, Pastor-Satorras, Vespignani, Activity driven modeling of time-varying networks, Scientific Reports, 2, 469, (2012). Perra et al. Random walk and search in time-varying networks, Phys. Rev. Letters, 109, 238701, (2012)



## IDENTIFICATION OF A FUNCTION THAT ENCODES THE INSTANTANEOUS DYNAMICS OF THE NETWORK (EVOLUTION RATE)

Activity potential  $x_i$  = relative probability of activity of agent i

 $x_i = \frac{\# \operatorname{int}_i}{\sum_i \# \operatorname{int}_i}$ 

The distribution Fc(x) is virtually independent of the time scale over which the activity potential is measured.





### **A SIMPLE MODEL**

•We consider N nodes (agents) .

Each node i an activity/firing rate ai =  $\eta$  xi, defined as the probability per unit time to create new contacts or interactions with other agents.

 $\eta$  is a rescaling factor defining the average number of interactions per unit time in the system).

At each discrete time step t the network Gt starts with N disconnected vertices;

•With probability ai each vertex i becomes active and generates m links that are connected to m other randomly selected vertices. Non-active nodes can still receive connections from other active vertices;

• At the next time step t + 1, all the edges in the network Gt are deleted. From this definition it follows that all interactions have a constant duration  $\tau i = 1$ .



#### Standard MOBS LAB

« »























### **RANDOM WALK PROPERTIES IN TIME-VARYING NETWORKS**

Master equation for the number of walkers in nodes of activity a

$$\frac{\partial W_a(t)}{\partial t} = -aW_a(t) + amw$$
$$-m\langle a \rangle W_a(t) + \int a' W_{a'}(t) F(a') da'$$

$$W_a = \frac{amw + \phi}{a + m\langle a \rangle} \text{ where } \qquad \phi = \int aF(a) \frac{amw + \phi}{a + m\langle a \rangle} \mathrm{d}a$$



### **RANDOM WALK PROPERTIES IN TIME-VARYING NETWORKS**



• Walkers diffuse each time a node is active.

•Nodes are active with rate a

#### Trapping of walkers

### • "Slave mode" with respect to the network dynamic



### **RANDOM WALK PROPERTIES IN TIME-VARYING NETWORKS**

Static/time-aggregated networks

Time-Varying networks

$$W_k \sim k \sim a$$

$$W_a = \frac{amw + \phi}{a + m\langle a \rangle}$$



# **FUTURE QUESTIONS/CHALLENGES**

Definition of general co-evolving network models (no-time scale separation)

Explorations of different dynamical processes in time-varying networks (consensus, game-theoretical, non-linear contagion...)

Non-trivial correlations in time and connectivity (assortativity, link persistence, bursts activity...etc.)

Response function in co-evolving networks



### TRANSFORMING THE WAY WE APPROACH SOCIO-TECHNICAL SYSTEMS

- Social computational science
- Computational epidemiology
- Science of Science
- Information systems and data science
  - General classes of tools methods and models generally applicable to complex techno-social systems:
  - New techniques for the generation of models, social analytics, real time empirical data generation.
  - Scenario analysis: systemic risk, economic impact etc.



## More Info

#### http://mobs-lab.org

@alexvespi



Alain Barrat, Harc Rarthelemy, Alexiandro Verpignarii



LABORATORY FOR THE MODELING OF BIOLOGICAL AND SOCIO-TECHNICAL SYSTEMS