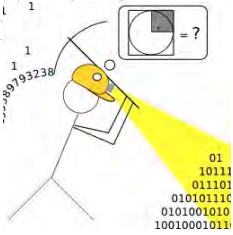
NDSSL NETWORK DYNAMICS & SIMULATION SCIENCE LABORATORY

VirginiaTech Biocomplexity Institute



Computational Epidemiology and Public Health Policy Planning

Prithwish Chakraborty, Madhav V. Marathe, Naren Ramakrishnan and Anil Kumar Vullikanti
Dept of Computer Science and Biocomplexity Institute of Virginia Tech

February 15, 2016

Slides for tutorial at AAAI, Phoenix, AZ, February 2016

We hope to update the slides and supplementary material continually over the next year.
The current version of the slides can be found at:
`https://ndssl.vbi.vt.edu/cms/files/9/f1e3606a-5b47-9464-851a-8dc19093c531/1787/aaai.pdf`

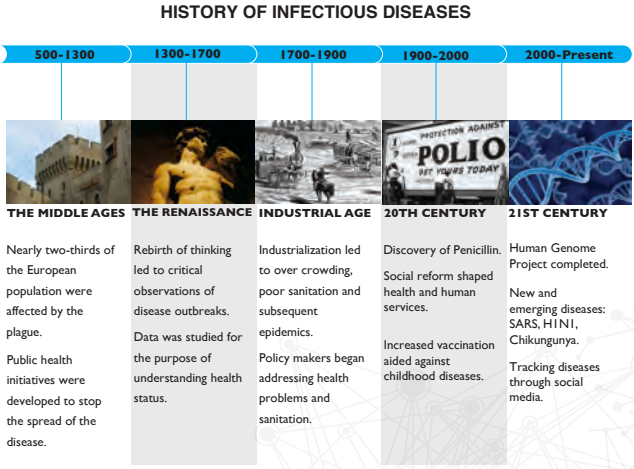
Outline

Goals

- **Overview and state of the art** – emphasis on computing, AI concepts and data science
- **Describe open problems and future directions** – aim to attract AI researchers to work in this exciting area
- **Unified framework** based on graphical dynamical systems and associated proof theoretic techniques; e.g. stochastic processes, spectral graph theory, randomized algorithms, mathematical programming, and Bayesian inference.
- **Computational epidemiology** as a multi-disciplinary science
- **Public health epidemiology** as an exemplar of data/computational science for social good
- *Does not aim to be extensive*; references provided for further exploration. *Important topics not* covered
  - Validation, verification and uncertainty quantification (UQ)
  - Different kinds of diseases

Epidemics and epidemiology in history

- *Good news:* Pandemic of 1918 lethality is currently unlikely Governments better prepared and coordinated : e.g. SARS epidemic But ..
- Planning & response to even a moderate outbreak is challenging: inadequate vaccines/anti-virals, unknown efficacy, hard logistics issues
- *Modern trends complicate planning:* increased travel, immuno-compromised populations, increased urbanization



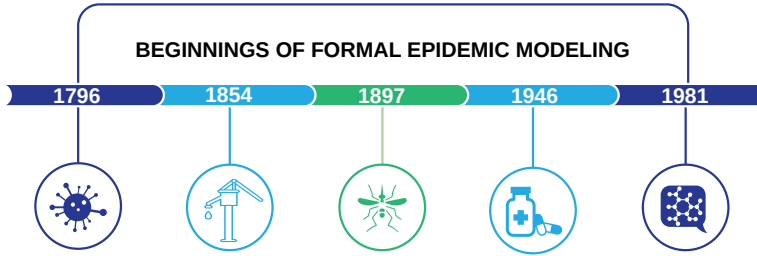
- 1918 Pandemic: 50 million deaths in 2 years (3-6% world pop) Every country and community was effected

What is epidemiology?

- Greek words **epi** = *on* or *upon*; **demos** = *people* & **logos** = *the study of.*[1]
- **Epidemiology:** study of the *distribution* and *determinants* of health-related states or events in specified populations, and the *application* of this study to the control of health problems.
  Now applies to non-communicable diseases as well as social and behavioral outcomes.
  - *Distribution*: concerned about population level effects
  - *Determinants*: causes and factors influencing health related events
  - *Application*: deals with public health action to reduce the incidence of disease.
- *Computational/mathematical epidemiology:* deals with the development of computational/mathematical methods, tools and techniques to support epidemiology.

[1] Last JM, ed. Dictionary of Epidemiology.

Precursors to modern computational epidemiology



SMALLPOX // Virus

Edward Jenner's research led to the development of vaccines.

Daniel Bernoulli mathematical models demonstrated the benefits of inoculation from a mathematical perspective.

Disease status today: eradicated.

CHOLERA // Bacteria

John Snow was the first to link the London cholera epidemic to a particular water source.

Disease status today: endemic; occuring in poverty-stricken countries.

MALARIA // Parasite

Ronald Ross and George Macdonald developed a mathematical model of mosquito-borne pathogen transmission.

Anderson McKendrick studied with Ross on anti-malarial operations, pioneering many discoveries in stochastic processes.

Disease status today: controlled in US; still prevalent in Africa, India.

TUBERCULOSIS // Bacteria

Albert Schatz discovered the antibiotic streptomycin under the direction of Selman Waksman.

Streptomycin was the first antibiotic that could be used to cure TB.

Disease status today: drug resistant TB strains persist since the 1980s.

HIV // Virus

Luc Montagnier discovered HIV and Robert Gallo determined HIV is the infectious agent responsible for AIDS.

The use of social network models have been initiated with the goal of controlling the virus.

Disease status today: no cure.

Epidemic science in real-time

Editorial, Fineberg and Harvey, Science, May 2009: Epidemics Science in Real-Time

Five areas: (i) Pandemic risk, (ii) vulnerable populations, (iii) available interventions, (iv) implementation possibilities & (v) pitfalls, and public understanding.

Epidemic science in real-time

Editorial, Fineberg and Harvey, Science, May 2009: Epidemics Science in Real-Time

Five areas: (i) Pandemic risk, (ii) vulnerable populations, (iii) available interventions, (iv) implementation possibilities & (v) pitfalls, and public understanding.

EDITORIAL

Harvey V. Fineberg is president of the Institute of Medicine.

Mary Elizabeth Wilson is associate professor of Global Health and Population at the Harvard School of Public Health and associate clinical professor at Harvard Medical School, Boston, MA.

Epidemic Science in Real Time

FEW SITUATIONS MORE DRAMATICALLY ILLUSTRATE THE SALIENCE OF SCIENCE TO POLICY THAN AN epidemic. The relevant science takes place rapidly and continually, in the laboratory, clinic, and community. In facing the current swine flu (H1N1 influenza) outbreak, the world has benefited from research investment over many years, as well as from preparedness exercises and planning in many countries. The global public health enterprise has been tempered by the outbreak of severe acute respiratory syndrome (SARS) in 2002–2003, the ongoing threat of highly pathogenic avian flu, and concerns over bioterrorism. Researchers and other experts are now able to make vital contributions in real time. By conducting the right science and communicating expert judgment, scientists can enable policies to be adjusted appropriately as an epidemic scenario unfolds.

In the past, scientists and policy-makers have often failed to take advantage of the opportunity to learn and adjust policy in real time. In 1976, for example, in response to a swine flu outbreak at Fort Dix, New Jersey, a decision was made to mount a nationwide immunization program against this virus because it was deemed similar to that responsible for the 1918–1919 flu pandemic. Immunizations were initiated months later despite the fact that not a single related case of infection had appeared by that time elsewhere in the United States or the world (www.iom.edu/swinefluaffair). Decision-makers failed to take seriously a key question: What additional information could lead to a different course of action? The answer is precisely what should drive a research agenda in real time today.

In the face of a threatened pandemic, policy-makers will want real-time answers in at least five areas where science can help: pandemic risk, vulnerable populations, available interventions, implementation possibilities and pitfalls, and public understanding. Pandemic risk, for example, entails both spread and severity. In the current H1N1 influenza outbreak, the causative virus and its genetic sequence were identified in a matter of days. Within a couple of weeks, an international consortium of investigators developed preliminary assessments of cases and mortality based on epidemic modeling.*

Specific genetic markers on flu viruses have been associated with more severe outbreaks. But virulence is an incompletely understood function of host-pathogen interaction, and the absence of a known marker in the current H1N1 virus does not mean it will remain relatively benign. It may mutate or acquire new genetic material. Thus, ongoing, refined estimates of its pandemic potential will benefit from tracking epidemiological patterns in the field and viral mutations in the laboratory. If epidemic models suggest that more precise estimates on specific elements such as attack rate, case fatality rate, or duration of viral shedding will be pivotal for projecting pandemic potential, then these measurements deserve special attention. Even when more is learned, a degree of uncertainty will persist, and scientists have the responsibility to accurately convey the extent of and change in scientific uncertainty as new information emerges.

A range of laboratory, epidemiologic, and social science research will similarly be required to provide answers about vulnerable populations; interventions to prevent, treat, and mitigate disease and other consequences of a pandemic; and ways of achieving public understanding that avoid both over- and underreaction. Also, we know from past experience that planning for the implementation of such projects has often been inadequate. For example, if the United States decides to immunize twice the number of people in half the usual time, are the existing channels of vaccine distribution and administration up to the task? On a global scale, making the rapid availability and administration of vaccine possible is an order of magnitude more daunting.

Scientists and other flu experts in the United States and around the world have much to occupy their attention. Time and resources are limited, however, and leaders in government agencies will need to ensure that the most consequential scientific questions are answered. In the meantime, scientists can discourage irrational policies, such as the banning of pork imports, and in the face of a threatened pandemic, energetically pursue science in real time.

– Harvey V. Fineberg and Mary Elizabeth Wilson

10.1126/science.1176297

*C. Fraser *et al.*, *Science* 11 May 2009 (10.1126/science.1176062).

www.sciencemag.org SCIENCE VOL 324 22 MAY 2009 987

Published by AAAS

Modeling before an epidemic

(*i*) Determine the (non)medical interventions required, (*ii*) feasibility of containment, (*iii*) optimal size of stockpile, (*iv*) best use of pharmaceuticals once a pandemic begins

Modeling during an epidemic

(*i*) Quantifying transmission parameters, (*ii*) Interpreting real-time epidemiological trends, (*iii*) measuring antigenic shift and (*iv*) assessing impact of interventions.

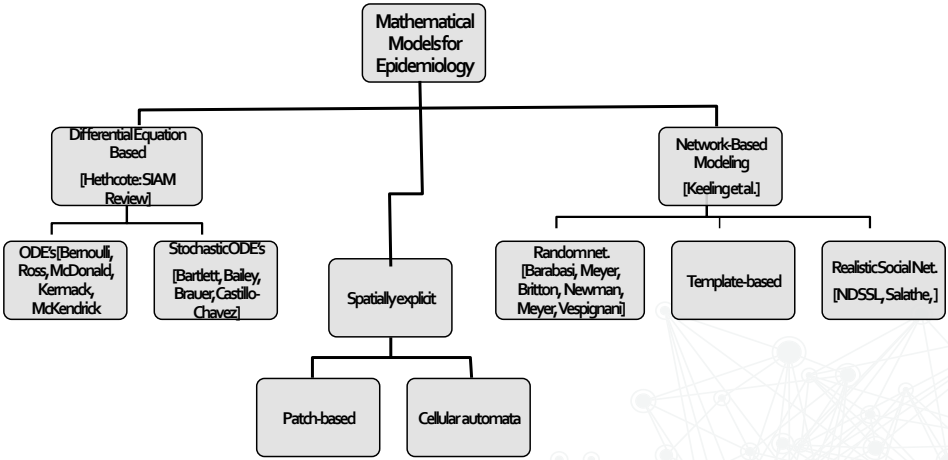
AI in epidemiology

- *Digital Disease Detection: using social media, news and other digital information for improved surveillance, forecasting and nowcasting*
  - Natural language processing, machine learning methods (e.g. matrix factorization)
- *Inference and analysis of Social contact networks and disease parameter*
  - Machine learning (CART trees, PCA)
  - Social and cognitive theories
  - Bayesian inference and graphical models
  - algorithmic graph theory
- *High performance computing and networked multi-agent modeling at scale*
  - Markov chain methods
  - Intelligent steering of computational experiments
- *Combinatorial techniques for optimization and control*
  - Algorithmic game theory
  - Markov decision process
  - Mathematical programming
  - Heuristic search

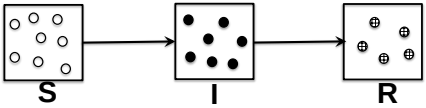
Outline

- Compartmental models
- Networked Epidemiology
  - Branching process
  - Spectral radius characterization
- Extensions: threshold models, voter models
- Competing contagions
- Unifying framework: graph dynamical systems

Classifying formal models



Mass action compartmental Models



Assumption: complete mixing among population of size N

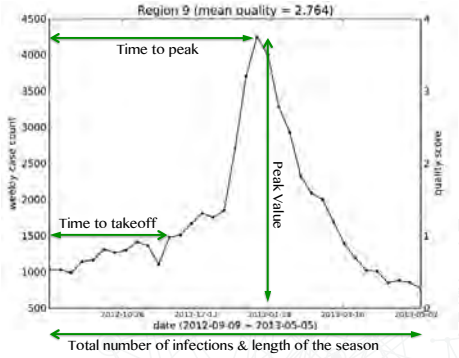
$$\begin{aligned}\frac{ds}{dt} &= -\beta is \\ \frac{di}{dt} &= \beta is - \gamma i \\ \frac{dr}{dt} &= \gamma i\end{aligned}$$

- *Susceptible* (S): An individual has never had the disease and is susceptible to being infected;
- *Infected* (I): An individual who currently has the disease and can infect other individuals, and
- *Resistant/Recovered* (R): An individual does not have the disease,cannot infect others, and cannot be infected (sometimes called removed)

Basic epidemic quantities

Typical epidemic quantities of interest

- *Epicurve:* time series of the number of infections
- *Peak* of the epidemic, *time to peak*, *total number of infections*
- *Basic Reproductive number* R_0: Average number of infections caused by a single infected individual in a completely susceptible population.
  - Condition for epidemic in terms of R_0
- *Take off time*: Time when epidemic takes off
- Time when number of daily infections falls below a threshold

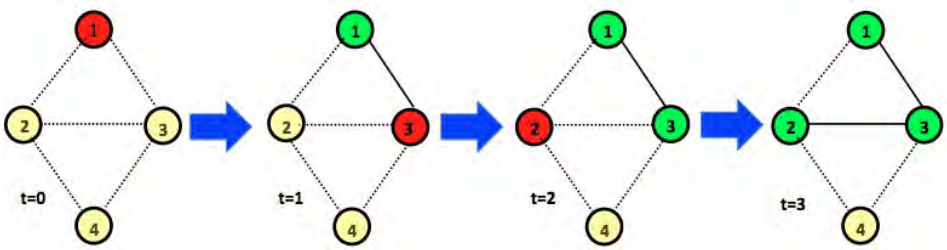


Pros and cons of compartmental models

- Compartmental models have been immensely successful over the last 100 years – (*i*) workhorse of mathematical epidemiology, (*ii*) easy to extend and quick to build; (*iii*) good solvers exist, simple ones can be solved analytically; (*iv*) mathematical theory of ODEs is well developed
- SARS was estimated to have $R_0 \in [2.2, 3.6]$[2]
  - Though it spread across many countries, small number of infections
  - Estimates were based on infections in crowded hospital wards, where complete mixing assumption was reasonable
- Compartmental models lack agency and heterogeneity of contact structure
  - True complexity stems from interactions among many discrete actors
  - Each kind of interaction must be explicitly modeled
  - Refinement is difficult
- Human behavioral issues – Inhomogeneous compliance; changes in the face of crisis
- Harder to design implementable interventions.

[2] Lipsitch et al., *Science*, 2003; Riley et al., *Science*, 2003

Networked epidemiology: Discrete time SIR model on a network



Fixed point: $R = \{1, 2, 3\}$ and $S = \{4\}$

$$p(1,3)(1-p(1,2))p(2,3)(1-p(2,4))(1-p(3,4))$$

- Each node is in states S (susceptible), I (infectious) or R (recovered)
- Time is discrete
- Each infected node u spreads the infection independently to each susceptible neighbor v with probability $p(u, v)$
- Infected node u recovers after 1 time step
- *Fixed point:* all nodes in states S or R

A general computational framework:
graphical models of dynamical (multi-agent) systems

Graphical Dynamical Systems (GDS)

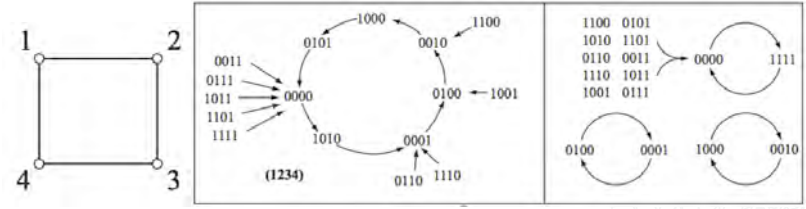
- Useful abstract model for networked interaction systems.
- Components of a GDS $\mathcal{S}$:
  - Undirected graph $G(V, E)$.
  - A state value from a finite domain $\mathbb{B}$ for each vertex v. (We use $\mathbb{B} = \{0, 1\}$.)
  - A local function f_v for each vertex v. (Inputs to f_v are the states of v and its neighbors; the output of f_v is from $\mathbb{B}$.)
  - The value of f_v gives the next state of v.
  - Vertices compute and update their states **synchronously**.

Contagions as graphical dynamical systems

- Contagion: (Cont = together with & Tangere = to touch): General term used to denote spread of "something" via interaction between agents
- Examples: financial contagion, product contagion, social contagion, malware contagion.
- Examples in social domain: rumors, fads, opinions, trust, emotions, ideologies, information, mass movements, riots, smoking, alcohol, drugs, contraceptive adoption, financial crises, repression, strikes, technology adoption

Example: Phase space of $\mathcal{S}$

- Directed graph with one vertex for each possible configuration.
- Directed edge (x, y) if the system transitions from the configuration corresponding to x to the one corresponding to y in one time step.
- Captures the **global behavior** of the system.
- Size of the phase space is **exponential** in the size of the SyDS.
- When the local functions are probabilistic, the phase space is best represented as a *Markov chain* (which is exponentially larger than the description).



Each node computes a Boolean NOR

Computational problems for GDS $\mathcal{S}$, phase space $\mathcal{P}(\mathcal{S})$, noisy observation $\mathcal{O}$

Analysis Problems

Does $\mathcal{P}(\mathcal{S})$ have a fixed point, GE configuration, transient of length $\geq k$?

Optimization Problems

Remove/Modify $\leq K$ nodes/edges in G so as to infect minimum number of nodes.

Inference Problem

Find the most likely: (i) initial configuration, (ii) the transmission tree, (iii) underlying network or (iv) disease parameters

Forecasting and Situational Assessment

Assess total number of nodes in a particular state, Forecast total number of nodes (probabilistically) in a particular state after time t

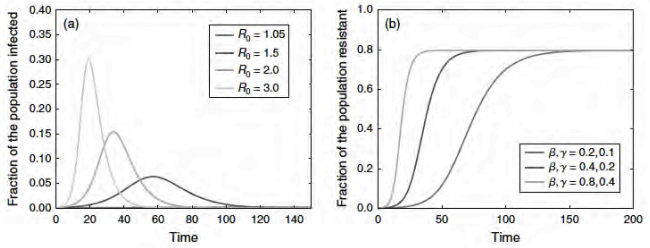
Mapping epidemic problems onto GDS problems

Quantity/Problem in epidemiology	GDS analogue
Epicurve	Analysis (e.g., #1's in configuration) of phase space trajectory
Computing epidemic characteristics	Analysis problem: Reachability problem in GDS
Inferring index case, given information about graph and observed infections	Predecessor inference problems in GDS
Inferring disease model, given the graph and observed infections	Local function inference problem

Dynamics in Compartmental Models

$$\begin{aligned}\frac{ds}{dt} &= -\beta is \\ \frac{di}{dt} &= \beta is - \gamma i \\ \frac{dr}{dt} &= \gamma i\end{aligned}$$

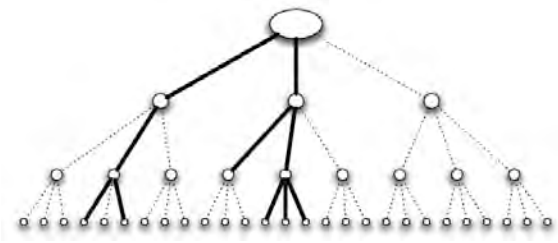
- $\frac{di}{dt} > 0$ (leads to a large epidemic) if $\frac{\beta s}{\gamma} > 1$
- At the start of epidemic: $s \approx 1$
- $R_0 = \beta/\gamma$: reproductive number
- Large epidemic if and only if $R_0 > 1$
- Modeling epidemic = estimating R_0
- Controlling epidemic: reducing R_0



Effect of R_0 on the dynamics[3]

[3] Dimitrov and Meyers, *INFORMS*, 2010

Dynamics over GDS: Trees



- Assume graph is an infinite d-ary tree, with transmission probability p on each edge. Using branching process as a proof technique.
- Assume the root is the only infected node, and everything else is susceptible
- Let q_n be the probability that the disease survives for atleast n waves (level of tree), in other words, that atleast one individual in the n^{th} level of the tree becomes infected.
- $q^* = \lim_{n \to \infty} q_n$

Image from: D. Easley and J. Kleinberg, 2010.

Analysis of the branching process on a tree

Theorem

Let $R_0 = pd$. If $R_0 < 1$ then $q^* = 0$. If $R_0 > 1$, then $q^* > 0$.

Case 1. $R_0 < 1$

- Let X_n denote the number of infected nodes in the nth level of the tree
- Pr[node i in nth level is infected] =

Analysis of the branching process on a tree

Theorem

Let $R_0 = pd$. If $R_0 < 1$ then $q^* = 0$. If $R_0 > 1$, then $q^* > 0$.

Case 1. $R_0 < 1$

- Let X_n denote the number of infected nodes in the nth level of the tree
- Pr[node i in nth level is infected] $= p^n$

Analysis of the branching process on a tree

Theorem

Let $R_0 = pd$. If $R_0 < 1$ then $q^* = 0$. If $R_0 > 1$, then $q^* > 0$.

Case 1. $R_0 < 1$

- Let X_n denote the number of infected nodes in the nth level of the tree
- Pr[node i in nth level is infected] $= p^n$
- $E[X_n] =$

Analysis of the branching process on a tree

Theorem

Let $R_0 = pd$. If $R_0 < 1$ then $q^* = 0$. If $R_0 > 1$, then $q^* > 0$.

Case 1. $R_0 < 1$

- Let X_n denote the number of infected nodes in the nth level of the tree
- Pr[node i in nth level is infected] $= p^n$
- $E[X_n] = p^n d^n = R_0^n$

Analysis of the branching process on a tree

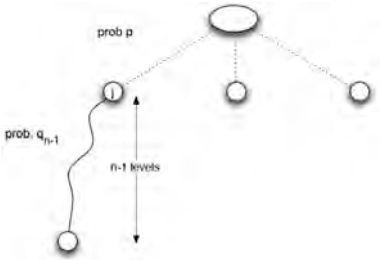
Theorem

Let $R_0 = pd$. If $R_0 < 1$ then $q^* = 0$. If $R_0 > 1$, then $q^* > 0$.

Case 1. $R_0 < 1$

- Let X_n denote the number of infected nodes in the nth level of the tree
- Pr[node i in nth level is infected] $= p^n$
- $E[X_n] = p^n d^n = R_0^n$
- Note that $E[X_n] = 1 \cdot Pr[X_n = 1] + 2 \cdot Pr[X_n = 2] + 3 \cdot Pr[X_n = 3] + \ldots$
- Equivalently: $E[X_n] = \Pr[X_n \geq 1] + \Pr[X_n \geq 2] + \ldots$; since $Pr[X_n = i]$
- $E[X_n] \geq \Pr[X_n \geq 1] = q_n$
- Therefore, $R_0 < 1 \Rightarrow \lim_{n \to \infty} q_n = 0$

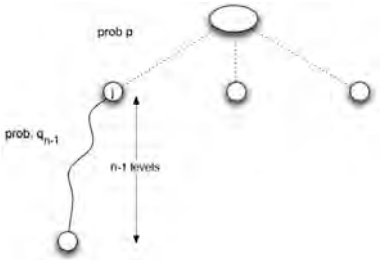
Analysis of the branching process (case 2): $R_0 > 1$ [4]



- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node

[4] Image from: D. Easley and J. Kleinberg, 2010.

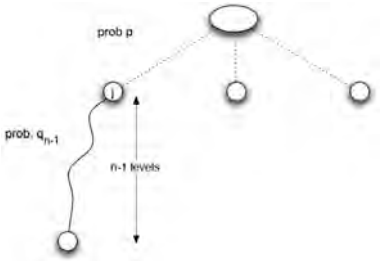
Analysis of the branching process (case 2): $R_0 > 1$ [4]



- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node
  - $\Pr[EP_j] =$

[4] Image from: D. Easley and J. Kleinberg, 2010.

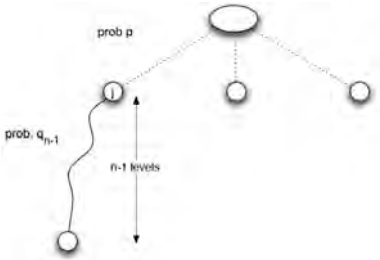
Analysis of the branching process (case 2): $R_0 > 1$ [4]



- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node
  - $\Pr[EP_j] = q_{n-1}$

[4] Image from: D. Easley and J. Kleinberg, 2010.

Analysis of the branching process (case 2): $R_0 > 1$ [4]

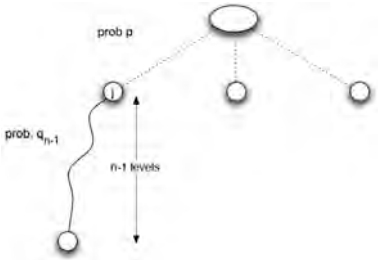


- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node
  - $\Pr[EP_j] = q_{n-1}$
  - Pr[epidemic persists, starting at the root and spreading via child j] =

.

[4] Image from: D. Easley and J. Kleinberg, 2010.

Analysis of the branching process (case 2): $R_0 > 1$ [4]

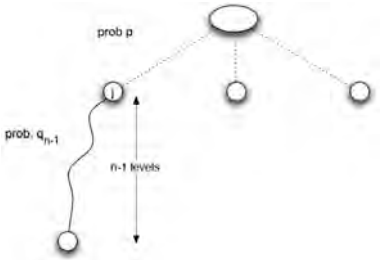


- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node
  - $\Pr[EP_j] = q_{n-1}$
  - Pr[epidemic persists, starting at the root and spreading via child j] = pq_{n-1}.

.

[4] Image from: D. Easley and J. Kleinberg, 2010.

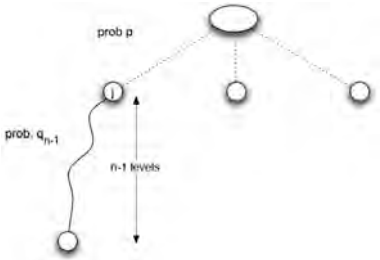
Analysis of the branching process (case 2): $R_0 > 1$ [4]



- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node
  - $\Pr[EP_j] = q_{n-1}$
  - Pr[epidemic persists, starting at the root and spreading via child j] = pq_{n-1}.
  - Pr[epidemic does not persist at level n] = .

[4] Image from: D. Easley and J. Kleinberg, 2010.

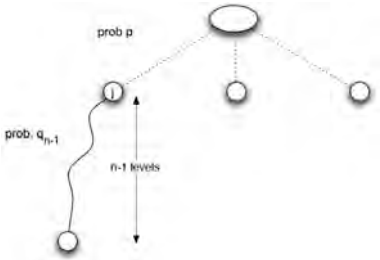
Analysis of the branching process (case 2): $R_0 > 1$ [4]



- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node
  - $\Pr[EP_j] = q_{n-1}$
  - Pr[epidemic persists, starting at the root and spreading via child j] = pq_{n-1}.
  - Pr[epidemic does not persist at level n] = $(1 - pq_{n-1})^d$.

[4] Image from: D. Easley and J. Kleinberg, 2010.

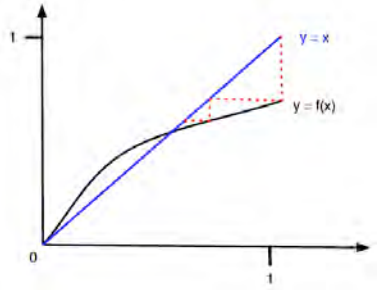
Analysis of the branching process (case 2): $R_0 > 1$ [4]



- Consider the subtree T_j rooted at child j of the root
- EP_j: event that epidemic persists until the nth level of T_j starting at child j of the root node
  - $\Pr[EP_j] = q_{n-1}$
  - Pr[epidemic persists, starting at the root and spreading via child j] = pq_{n-1}.
  - Pr[epidemic does not persist at level n] = $(1 - pq_{n-1})^d$.
- $q_n = 1 - (1 - pq_{n-1})^d$

[4] Image from: D. Easley and J. Kleinberg, 2010.

Analysis of the branching process: case 2



- $f(x) = 1 - (1 - px)^d$; $f(0) = 0$ and $f(1) < 1$
- $f'(x) = pd(1 - px)^{d-1}$
- $f'(x) > 0$ for $x \in [0, 1]$ and monotonically decreasing
- $f(\cdot)$ starts at origin, and ends up below the line $y = x$ at $x = 1$
- $f'(0) = R_0 > 1$, so $f(\cdot)$ starts above $y = x$ and then intersects it
- The sequence $1 = q_0, f(1) = q_1, f(f(1)) = q_2, \ldots$ converges to q^*

Dynamics and analysis on general graphs

Dynamics in the SIR model on other networks: impact of structure

- Phase transition for SIR model shown in many graph models: there exists a threshold p_t such that few infections if $p < p_t$ but large outbreak if $p > p_t$
- Technique: mainly extends branching process
- Clique on n nodes[5]: $p_t = 1/(n-1)$
- Lattice $\mathbb{Z}^d$: $p_t \to 1/(2d)$, as $d \to \infty$
- Random d-regular graphs: $p_t = 1/d$
- Not well understood in general graphs
  - Partial characterization in finite regular expander graphs with high girth[6]
  - Characterization in terms of the second moment[7]

[5] Erdős and Rényi, 1959
[6] Alon, Benjamini and Stacey, 2001
[7] Chung, Horn, Lu, 2009

Dynamics in the SIS model: preliminaries

- Nodes in Susceptible (S) or Infectious (I) states
- Each infected node spreads infection to each susceptible neighbor with rate β
- Each infected node becomes susceptible with rate δ
- $\rho(A)$: spectral radius of adjacency matrix A
- $T = \delta/\beta$
- Generalized isoperimetric constant: $\eta(G, m) = \inf_{S \subset V, |S| \leq m} \frac{E(S,\bar{S})}{|S|}$

Dynamics in the SIS model: preliminaries

- Nodes in Susceptible (S) or Infectious (I) states
- Each infected node spreads infection to each susceptible neighbor with rate β
- Each infected node becomes susceptible with rate δ
- $\rho(A)$: spectral radius of adjacency matrix A
- $T = \delta/\beta$
- Generalized isoperimetric constant: $\eta(G, m) = \inf_{S \subset V, |S| \leq m} \frac{E(S,\bar{S})}{|S|}$

- Spectral radius $\rho(A) = \max_x x^T Ax/x^x$
- Avg degree $\leq \rho(A) \leq \Delta(G)$, where $\Delta(G)$ is the maximum node degree

$\eta(G, 6) \leq 2/6$

Dynamics in the SIS model (informal) spectral characterization[10]

- $\rho(A)$: spectral radius of adjacency matrix A
- $T = \delta/\beta$
- Generalized isoperimetric constant: $\eta(G, m) = \inf_{S \subset V, |S| \leq m} \frac{E(S,\bar{S})}{|S|}$

- If $\rho(A) < T$: epidemic dies out "fast"
- If $\eta(m) > T$: epidemic lasts "long"

Similar implications but different assumptions, extended to SEIR models[8] [9]

[8] BA Prakash, D Chakrabarti, M Faloutsos, N Valler, C Faloutsos. *Knowledge and Information Systems*, 2012

[9] Y. Wang, D. Chakrabarti, C. Wang and C. Faloutsos, *ACM Transactions on Information and System Security*, 2008.

[10] A. Ganesh, L. Massoulie and D. Towsley, *IEEE INFOCOM*, 2005

Formally

Lemma (Sufficient condition for fast recovery)

Suppose $\rho(A) < T$. Then, the time to extinction τ satisfies

$$E[\tau] \leq \frac{\log n + 1}{1 - \rho(A)/T}$$

Lemma (Sufficient condition for lasting infection)

If $r = \frac{\delta}{\beta\eta(m)} < 1$, then the epidemic lasts for "long":

$$\Pr[\tau > r^{-m+1}/(2m)] \geq \frac{1-r}{e}(1 + O(r^m))$$

Implications for different network models

- Hypercube: $\rho(G) = \log_2 n$, and $\eta(m) = (1-a)\log_2 n$ for $m = n^a$
  - Fast die out if $\beta < \frac{1}{\log_2 n}$, slow die out if $\beta > \frac{1}{(1-a)\log_2 n}$
- Erdős-Rényi model: $\rho(G) = (1+o(1))np = (1+o(1))d$ and $\eta(m) = (1+o(1))(1-\alpha)d$ where $m/n \to \alpha$
  - Fast die out if $\beta < \frac{1}{(1+o(1))d}$, slow die out if $\beta > \frac{1}{(1+o(1))(1-\alpha)d}$
- Power law graphs (Chung-Lu model): assume degree distribution with power law exponent $\gamma > 2.5$
  - $E[\tau] = O(\log n)$ if $\beta < (1-u)/\sqrt{m}$ and $E[\tau]$ exponential if $\beta > m^\alpha/\sqrt{m}$ for some $u, \alpha \in (0,1)$ and $m = n^\lambda$, for $\lambda \in (0, \frac{1}{\gamma-1})$

Implications for different network models

- In general, gap between necessary and sufficient conditions for epidemic to last long
- Similar implications through different assumptions, extended to SEIR models[11] [12]

[11] Y. Wang, D. Chakrabarti, C. Wang and C. Faloutsos, *ACM Transactions on Information and System Security*, 2008.

[12] BA Prakash, D Chakrabarti, M Faloutsos, N Valler, C Faloutsos. *Knowledge and Information Systems*, 2012

Proof of sufficient condition for epidemic to die out fast (I)

Assume $\delta = 1$ for notational simplicity. Consider continuous version of the SIS model:

$$X_i : 0 \to 1 \qquad \text{at rate } \beta \sum_{(j,i)\in E} X_j$$

$$X_i : 1 \to 0 \qquad \text{at rate } 1$$

- Let τ denote the time to extinction
- $\Pr[\tau > t] \leq \Pr[X(t) = \sum_i X_i(t) \neq 0]$
- Goal: derive upper bound for $\Pr[X(t) = \sum_i X_i(t) \neq 0]$
- Challenging to derive this bound directly since X switches between 0 and 1. Instead, consider an alternative process which dominates $X(\cdot)$ and is easier to analyze

Main steps in proof

- Consider a random walk process $Y(\cdot)$ that upper bounds $X(\cdot)$

$$Y_i : k \to k+1 \qquad \text{at rate } \beta \sum_{(j,i)\in E} Y_j$$
$$Y_i : k \to k-1 \qquad \text{at rate } Y_i$$

- $X(t) \leq Y(t)$ for all $t \geq 0$
- $\frac{d}{dt}E[Y(t)] = (\beta A - I)E[Y(t)]$
- $E[Y(t)] = exp(t(\beta A - I))Y(0)$
- $\Pr[X(t) \neq 0] \leq \sum_i E[Y_i(t)] \leq ne^{(\beta\rho(A)-1)t}$

Proof of sufficient condition for epidemic to die out fast (II)

Consider an alternate random walk process $Y = \{Y_i\}_{i \in V}$:

$$Y_i : k \to k+1 \qquad \text{at rate } \beta \sum_{(j,i) \in E} Y_j$$

$$Y_i : k \to k-1 \qquad \text{at rate } Y_i$$

- Relaxation of $X(\cdot)$: $Y_i(t)$ is not upper bounded
- $X(t) \leq Y(t)$ for all $t \geq 0$ (formally: Y stochastically dominates X)
- $\Rightarrow \Pr[\sum_i X_i(t) \neq 0] \leq \Pr[\sum_i Y_i(t) \neq 0] = \Pr[\sum_i Y_i(t) > 0]$
- $\Pr[\sum_i Y_i(t) > 0] = \Pr[\sum_i Y_i(t) \geq 1] \leq \sum_i E[Y_i(t)]$ (Markov's inequality)
- Rest of the proof: derive upper bound on $\sum_i E[Y_i(t)]$

Proof of sufficient condition for epidemic to die out fast (III)

$$
\begin{aligned}
E[Y_i(t+dt) - Y_i(t)|Y(t)] &= \beta \sum_{(j,i)\in E} Y_j(t)dt - Y_i(t)dt + o(dt) \\
&= \beta \sum_{j} A_{ij} Y_j(t)dt - Y_i(t)dt + o(dt) \\
\Rightarrow \frac{d}{dt}E[Y(t)] &= (\beta A - I)E[Y(t)]
\end{aligned}
$$

Solution to this linear differential equation gives
$E[Y(t)] = exp(t(\beta A - I))Y(0)$

Proof of sufficient condition for epidemic to die out fast (IV)

- Recall: we need upper bound on $\sum_i E[Y_i(t)]$
- $\sum_i E[Y_i(t)] \leq \| E[Y(t)] \|_2 \| \mathbf{1} \|_2$ (by Cauchy-Schwartz: $\mathbf{a} \cdot \mathbf{b} \leq \| \mathbf{a} \|_2 \| \mathbf{b} \|_2$)
- Recall: $E[Y(t)] = exp(t(\beta A - I))Y(0)$

$$
\begin{aligned}
\| E[Y(t)] \|_2 &\leq \rho(exp(t(\beta A - I)) \| Y(0) \|_2 \\
&\leq e^{(\beta\rho(A)-1)t} \| Y(0) \|_2 \\
&\quad \left((\beta A - I)t \text{ symmetric} \Rightarrow \rho(exp(t(\beta A - I))) = e^{(\beta\rho(A)-1)t}\right) \\
\Rightarrow \sum_i E[Y_i(t)] &\leq \sqrt{n} e^{(\beta\rho(A)-1)t} \| Y(0) \|_2 \\
\Rightarrow \Pr[X(t) \neq 0] &\leq \sqrt{n} e^{(\beta\rho(A)-1)t} \| Y(0) \|_2
\end{aligned}
$$

Proof of sufficient condition for epidemic to die out fast (V)

Putting everything together:

$$
\begin{aligned}
E[\tau] &= \int_0^\infty \Pr[\tau > t]dt \\
&= \int_0^\infty \Pr[X(t) \neq 0]dt \\
&\leq \int_0^z \Pr[X(t) \neq 0]dt + \int_z^\infty \Pr[X(t) \neq 0]dt, \\
&\quad \text{where } z = \log n/(1-\beta\rho(A)) \\
&\leq z + \int_z^\infty n e^{(\beta\rho(A)-1)t}dt \\
&= \frac{\log n + 1}{1-\beta\rho(A)}
\end{aligned}
$$

Alternative approach[13]

- Let $X_{i,t}$ be the indicator random variable for the event that node i is infected at time t
- Let $p_{i,t} = \Pr[X_{i,t}]$
- Let $\zeta_{i,t} = \Pr[\text{node } i \text{ does not receive infection from neighbors at time } t]$
- Assuming independence between $X_{i,t}$'s

$$\begin{aligned}
\zeta_{i,t} &= \prod_{j \in N(i)} \Pr[\text{node } i \text{ does not get infected from } j] \\
&= \prod_{j \in N(i)} (p_{j,t-1}(1-\beta) + (1 - p_{j,t-1})) \\
&= \prod_{j \in N(i)} (1 - \beta p_{j,t-1})
\end{aligned}$$

[13] Y. Wang, D. Chakrabarti, C. Wang and C. Faloutsos, *ACM Transactions on Information and System Security*, 2008.

Non-linear dynamical system

$$\begin{aligned}\Pr[\text{node } i \text{ not infected at time } t] &= \Pr[\text{node } i \text{ not infected at time } t-1 \text{ and did not get infection from neighbors}] \\ &\quad + \Pr[\text{node } i \text{ infected at time } t-1, \text{ didn't get infection from nbrs and recovered}] \\ 1 - p_{i,t} &= (1 - p_{i,t-1})\zeta_{i,t} + \delta p_{i,t-1}\zeta_{i,t}\end{aligned}$$

Limiting state: epidemic need not die out

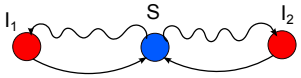
Theorem

Epidemic dies out if and only if $\rho(A) < \delta/\beta$

Extension to SIR and other models[14]

[14] BA Prakash, D Chakrabarti, M Faloutsos, N Valler, C Faloutsos. *Knowledge and Information Systems*, 2012

Competing contagions in the SIS model: the SI_1I_2S model



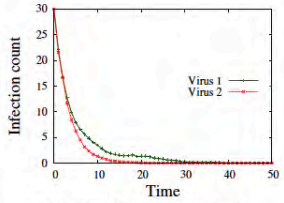
- $G = (V, E)$: undirected contact graph
- State transition for node u from S to I_j at rate β_j, $j = 1, 2$, depending on which infected neighbor of u is successful in infecting it
- Nodes switch back to susceptible state at rate δ_j from I_j to S
- What is the limiting distribution?

Steady state distribution in SI_1I_2S model: "winner takes all"

Theorem

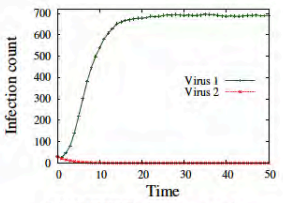
[a] *In the SI_1I_2S model in graph G with adjacency matrix A, and parameters $(\beta_1, \beta_2, \delta_1, \delta_2)$, virus 1 will dominate and virus 2 will completely die-out in the steady state if $\lambda_1 \frac{\beta_1}{\delta_1} > 1$ and $\frac{\beta_1}{\delta_1} > \frac{\beta_2}{\delta_2}$*

[a] B. Aditya Prakash, A. Beutel, R. Rosenfeld, C. Faloutsos, *WWW*, 2012



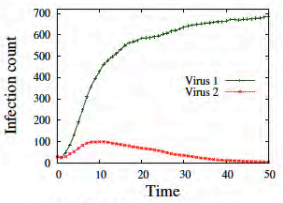
(a) BELOW (Time-plot)

Both viruses below threshold: $\lambda_1 \frac{\beta_1}{\delta_1} < 1$, $\lambda_1 \frac{\beta_2}{\delta_2} < 1$



(b) MIXED (Time-plot)

Virus 1 above threshold, virus 2 below: $\lambda_1 \frac{\beta_1}{\delta_1} > 1$, $\lambda_1 \frac{\beta_2}{\delta_2} < 1$



(c) ABOVE (Time-plot)

Both above threshold: $\lambda_1 \frac{\beta_1}{\delta_1} > 1$, $\lambda_1 \frac{\beta_2}{\delta_2} > 1$, $\frac{\beta_1}{\delta_1} > \frac{\beta_2}{\delta_2}$

Surveillance and Forecasting

What we will cover in this section

- Forecasting flu case counts using data-driven methods
- Forecasting flu epicurve characteristics
- Spatio-temporal models with applications to disease surveillance.

Breaking down the AI topics of interest

- Robust Information extraction
  - short text, long text, images
- Robust dynamic regressions
- Data Assimilation methods

Syndromic surveillance

Traditional vs syndromic surveillance

- Traditional: laboratory tests of respiratory specimens, mortality reports
- Syndromic: 'clinical features that are discernable before diagnosis is confirmed or activities prompted by the onset of symptoms as an alert of changes in disease activity' [15]

Issues in considering a syndromic surveillance system

- Sampling bias
- Veracity and reliability of syndromic data
- Granularity of space- and time-resolution
- Change point detection versus forecasting

Broad consensus is that syndromic surveillance provides some early detection and forecasting capabilities but nobody advocates them as a replacement for traditional disease surveillance.

[15] K Hope, DN Durrheim, ET d'Espaignet, C Dalton, *Journal of Epidemiology and Community Health*, 2006

Surrogate data sources: the good, bad, and ugly

Proposals for flu surveillance

- Search queries

  'Miley Cyrus cancels Charlotte Concert over Flu'

- OTC medication sales

  Discount sales, hoarding, lack of patient-specific data

- Wikipedia page views

  Lack of specificity about visitor locations

- Twitter

  Concerned awareness tweets versus infection reporting tweets

Surrogate data sources: the good, bad, and ugly

Proposals for flu surveillance

- Search queries

  'Miley Cyrus cancels Charlotte Concert over Flu'

- OTC medication sales

  Discount sales, hoarding, lack of patient-specific data

- Wikipedia page views

  Lack of specificity about visitor locations

- Twitter

  Concerned awareness tweets versus infection reporting tweets

Self-reinforcing and self-defeating prophecies abound!

Google Flu Trends

Google Flu Trends (http://www.google.org/flutrends/) is a nowcasting system for monitoring health-seeking behavior through Google queries. [16]

- 50 million candidate queries were narrowed down to a set of 45 (proprietary) queries that most accurately fit CDC ILI data in the US
- Queries merely correlated with flu season (e.g., 'high school basketball') were hand pruned
- Relative query volumes (w.r.t. weekly search volume per location) were used as independent variables
- Simple linear model from query fraction $\mapsto$ ILI physician visits.

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon \qquad (1)$$

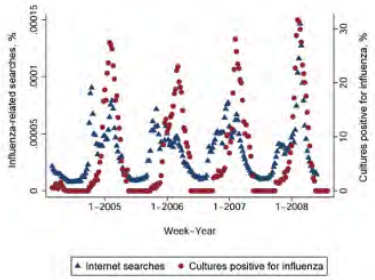
where P is the percentage of ILI related physician visits and Q is the ILI-related search query fraction.

[16] J Ginsberg, MH Mohebbi, RS Patel, L Brammer, MS Smolinski, L Brilliant, *Nature*, 2008

Was Google Flu Trends a pioneer?

Polgreen et al. [17] was the original paper that proposed the use of search queries for influenza surveillance

- Yahoo search queries from March 2004–May 2008
  1. Fraction of US search queries that contain the term 'influenza' or 'flu' but NOT 'bird', 'avian', or 'pandemic'
  2. Fraction of US search queries that contain 'influenza' or 'flu' but NOT 'bird', 'avian', 'pandemic', 'vaccination', or 'shot'
- Explored searches with one- to ten-week lead times as explanatory variables; reports 1-3 week lead time over CDC reporting

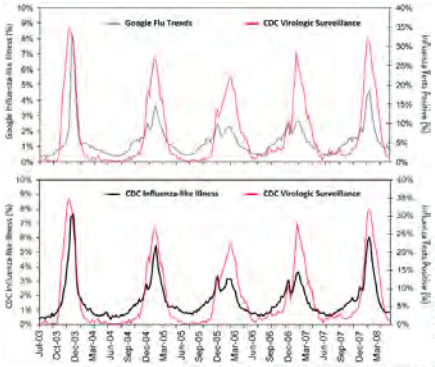


[17] PM Polgreen, Y Chen, DM Pennock, FD Nelson, *Clinical Infectious Diseases*, 2008

Google Flu Trends vs. traditional surveilance

Comparisons of GFT as well as CDC ILI surveillance data against US Influenza Virologic Surveillance data [18]

- First study evaluating Google Flu Trends against laboratory confirmed infections
- Pearson correlation coefficients: GFT-Virological (0.72), CDC/ILI-Virological (0.85), GFT-CDC/ILI (0.94)



[18] JR Ortiz, H Zhou, DK Shay, KM Neuzil, AL Fowkes, CH Goss, *PloS ONE*, 2011

Google Flu Trends w/ other data sources

How does GFT fare when used in conjunction with other indicators? [19]

- 5 typical seasons (2004–2008, 2010–2011) and 2 atypical seasons (2008–2009 and 2009-2010) studied in an urban tertiary care provider in Baltimore, MD
- Response variable: influenza-related ED visits; independent variables: GFT, local temperature, local relative humidity, Julian weeks; connected using a GARMA model

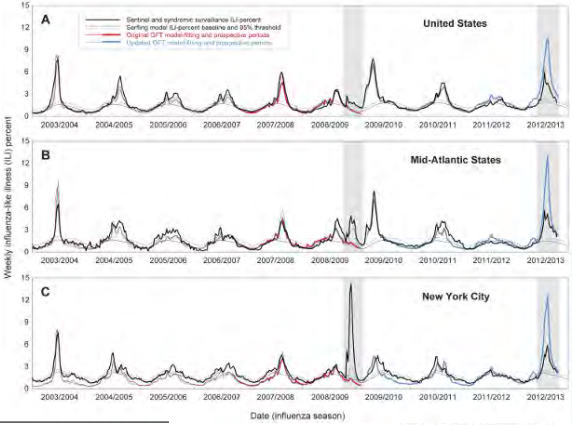
- Autoregressive component had the strongest influence

[19] AF Dugas, YH Hseih, SR Levin, JM Pines, DP Mareiniss, A Mohareb, CA Gaydos, TM Perl, RE Rothman, *Clinical Infectious Diseases*, 2012

More murmurs of discontent

GFT evaluated at three geographic scales: national (US), regional (mid-Atlantic), and local (NY city)levels [20]

- Correlations can be misleading
  1. GFT completely missed the first wave of the 2009 H1N1 pandemic flu
  2. GFT overstimated the intensity of the H3N2 epidemic during 2012–2013

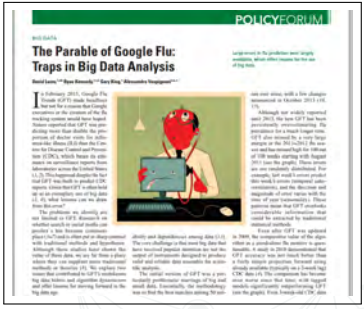


[20] DR Olson, KJ Konty, M Paladini, C Viboud, L Simonsen, *PloS computational biology*, 2013

The Final Straw [21] [22]

- Search algorithm continually being modified
- Additional search term suggestions
- Lack of transparency
- Big data 'hubris'

For the two years ending Sep 2013, Google's estimates were high in 100 out of 108 weeks. After Oct 2013 update, Google's estimates are over by 30% for 2013–2014 season



POLICYFORUM

BIG DATA

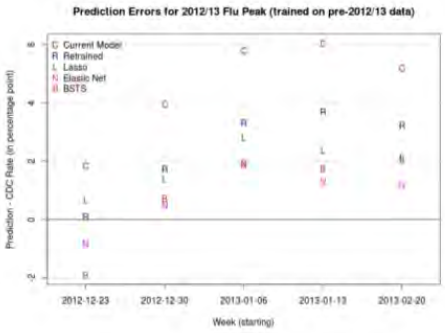
The Parable of Google Flu: Traps in Big Data Analysis

[21] DM Lazer, R Kennedy, G King, A Vespignani, *Science*, 2014

[22] DM Lazer, R Kennedy, G King, A Vespignani, http://gking.harvard.edu/files/gking/files/ssrn-id2408560_2.pdf, 2014

Recent improvements to Google Flu Trends[23]

- Handling 'inorganic queries' resulting from heightened media coverage - spike detectors (long term and short term).
- Handling drift
  - Retraining after each season
  - Use of regularizers



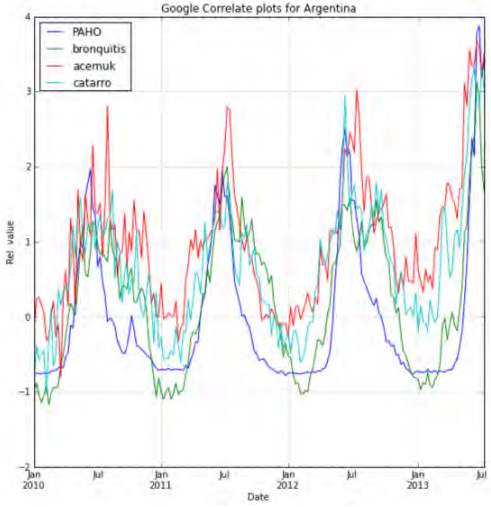
[23] http://patrickcopeland.org/papers/isntd.pdf

Designing your own vocabulary[24]

- Pseudo-query expansion methods
  - Health ministry website.
  - News articles.
- Google Correlate
  - Correlate search query volumes with disease case count time series.
  - Compare against different time shifted case counts.
- Example keywords
  - From search query words such as 'flu',
  - through correlation analysis words we can discover such as 'ginger.'

[24] P Chakraborty, P Khadivi, B Lewis, A Mahendiran, J Chen, P Butler, EO Nsoesie, SR Mekaru, JS Brownstein, M Marathe, N Ramakrishnan, *SDM*, 2014

Designing your own vocabulary (contd..)



Symptomatic words: **“**bronquitis”, “catarro”, “tos seca” (whooping cough)

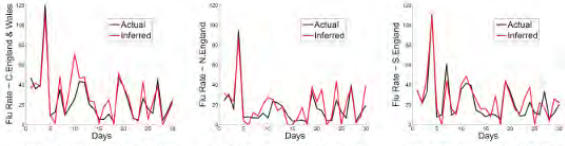
Medicinal words: **“**acemuk”, “claritromicina” (clarithromycin)

Interesting words: ginger (“jengibre”), leave letter (“letra de deja”)

Nowcasting with Twitter

Culotta [25] and Lampos et al. [26] adapted GFT-like ideas to forecasting ILI case counts using Twitter

- Geolocation to narrow down to regions of interest
- Document filtering to first identify ILI-related tweets
- Prediction models:
  1. Regression with multiple keyword independent variables performs better than simple linear regression (as used in GFT)
  2. LASSO with n-grams as features



(a) *C. England & Wales* – RMSE: 11.781 (b) *N. England* – RMSE: 9.757 (c) *S. England* – RMSE: 9.599

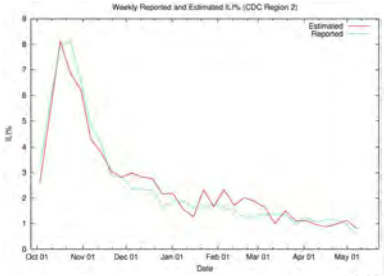
[25] A Culotta, *Proceedings of the First Workshop on Social Media Analytics*, 2010

[26] V Lampos, N Cristianini, *ACM TIST*, 2012

Using Twitter during the H1N1 pandemic

Signorini et al. [27] study the use of Twitter to nowcast the 2009 season.

- Geolocated tweets (US home locations) containing specific flu-related keywords were filtered and used to create a dictionary (after stemming, stopword removal]
- Support vector regression from dictionary to CDC ILI rates
- Model trained on 9 of the 10 CDC US regions and evaluated on the 10th



[27] A Signorini, AM Segre, PH Polgreen, *PLoS One*, 2011

Getting into more detailed content analysis [28]

- Coding rules to help categorize tweets
- 52.6% of tweets were about news and information; 4.5% were misinformation

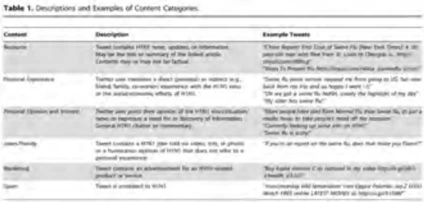
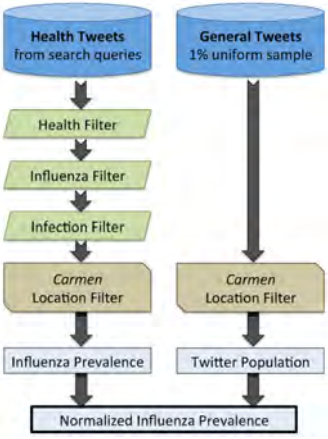


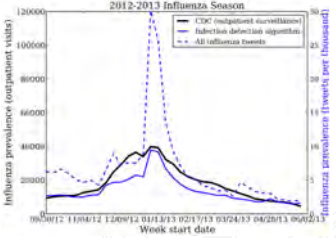
Table 1. Descriptions and Examples of Content Categories.

Content	Description	Example Tweets
Resource	Tweet contains H1N1 news, updates, or information. May be the title or summary of the linked article. Contents may or may not be factual.	[illegible]
Personal Experience	[illegible]	[illegible]
Personal Opinion and Interest	[illegible]	[illegible]
Jokes/Parody	[illegible]	[illegible]
Marketing	[illegible]	[illegible]
Spam	Tweet is unrelated to H1N1	[illegible]

[28] C Chew, G Eysenbach, *PloS one*, 2010

Multi-level classification of tweets [29]





[29] DA Broniatowski, MJ Paul, M Dredze, *PloS one*, 2013

Even finer distinctions [30]

- Infection vs concerned awareness.

  going over to a friends house to check on her son. he has the flu and i am worried about him

  starting to get worried about swine flu...

- Self vs other
- Part of speech templates constructed from word class features

Class Name	Words in Class
Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
Possession	bird, the flu, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
Vaccination	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
Past Tense	was, did, had, got, were, or verb with the suffix "ed"
Present Tense	is, am, are, have, has, or verb with the suffix "ing"
Self	I, I've, I'd, I'm, im, my
Others	your, everyone, you, it, its, u, her, he, she, he's, she's, she, they, you're, she'll, he'll, husband, wife, brother, sister, your, people, kid, kids, children, son, daughter

[30] A Lamb, MJ Paul, M Dredze, *HLT-NAACL*, 2013

Data Assimilation Methods [31]

The SIRS equations are given by:

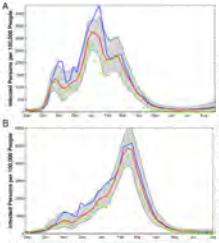
$$\begin{aligned} \frac{dS}{dt} &= \frac{N-S-I}{L} - \frac{\beta(t)SI}{N} - \alpha \\ \frac{dI}{dt} &= \frac{\beta(t)SI}{N} - \frac{I}{D} + \alpha \end{aligned} \tag{2}$$

where the AH modulated reproductive number is given by

$$R_0(t) = \exp(a \times q(t) + b) + R_{0min} \tag{3}$$

where, $a = -180$ and $b = \log(R_{0max} - R_{0min})$. $q(t)$ is the time varying specific humidity.

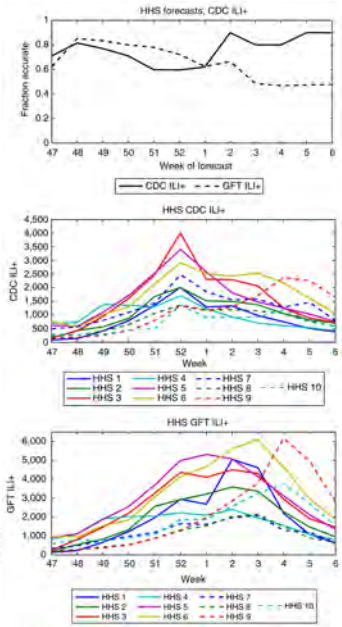
- GFT ILI estimates are assimilated to generate a posterior estimate of infection rates
- Captures long rise and single peak of infection during 2007–2008 as well as multiple modes during 2004–2005



[31] J Shaman, A Karspeck, *PNAS*, 2012

Data Assimilation Methods (contd) [32]

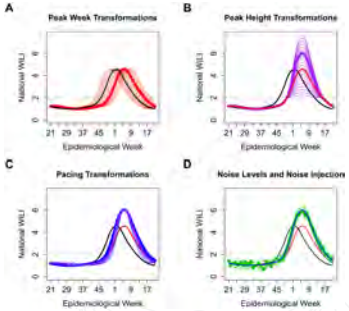
- First example of real-time forecasting
- Evaluated peak timing and peak value prediction
- By week 52, prior to peak for majority of cities, 63% of forecasts were accurate



[32] J Shaman, A Karspeck, W Yang, J Tamerius, M Lipsitch, *Nat. Comm.*, 2013

Data Assimilation Methods (contd) [33] [34] [35]

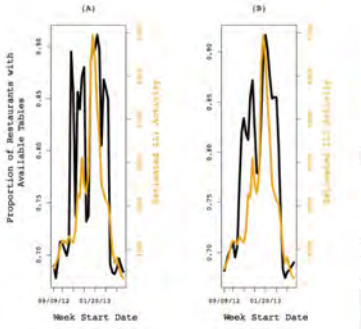
- Peformance comparison of different Kalman and Particle Filters.
- Geographical considerations (Hong Kong)
- Empirical Bayes Framework: improves flexibility of modeling.



[33] W Yang, A Karspeck, J Shaman, *PLOS Comp. Biol.*, 2015
[34] W Yang, BJ Cowling, EHY Lau, J Shaman, *PLOS Comp. Biol.*, 2015
[35] LC Brooks, DC Farrow, S Hyun, RJ Tibshirani, R Rosenfiled, *PLOS Comp. Biol.*, 2015

OpenTable reservation monitoring [36]

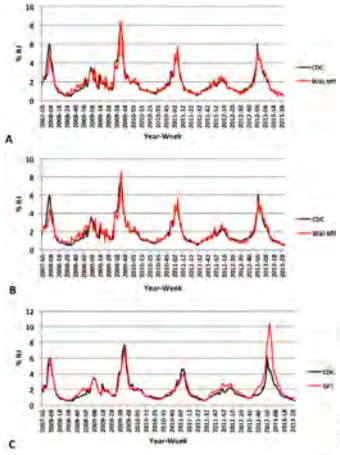
- Daily search performed for restaurants with available tables for 2 at the hour and half past the hour for 22 distinct times: between 11am–3:30pm and 6pm–11:30pm
- Multiple cities in US and Mexico



[36] EO Nsoesie, DL Buckeridge, JS Brownstein, *Online Journal of Public Health Informatics*, 2013

Monitoring Wikipedia usage [37]

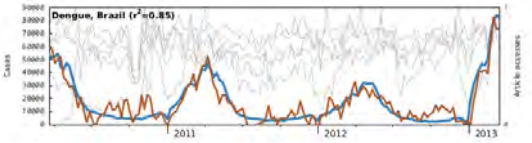
- Handful of pages were identified and tracked for daily article view data
- LASSO model gives comparable performance to a full model



[37] DJ McIver, JS Brownstein, *PLoS Computational Biology*, 2014

Global disease monitoring with Wikipedia [38]

- Cholera, Dengue, Ebola, HIV/AIDS, Influenza, Plague, Tuberculosis
- Haiti, Brazil, Thailand, Uganda, China, Japan, Poland, Norway, US

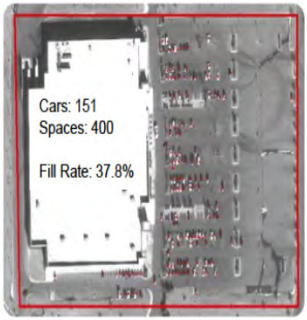


- Reasons it doesn't work: noise, too slow, or too fast disease incidence

[38] N Generous, G Fairchild, A Deshpande, SY Del Vallem, R Preidhorsky, *arXiv preprint*, 2014

Parking lot imagery

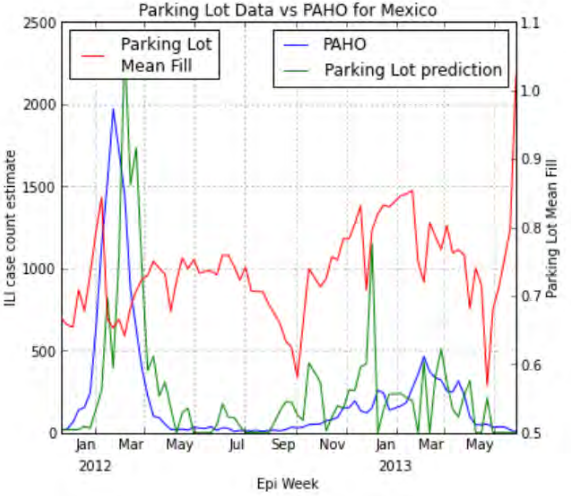
Estimating
"fill rate"
in parking lots



Hospital parking lot study in Latin America

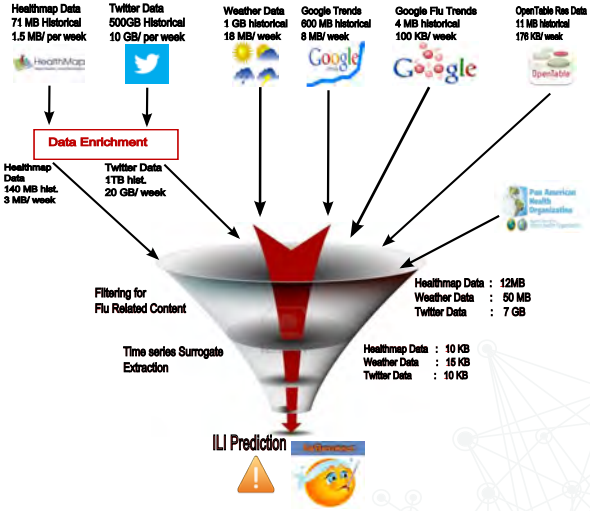
Country	Images Ordered	Images Used	Time Period
Argentina	327	231	11/1/2011–5/26/2013
Mexico	1,566	781	11/1/2011–5/26/2013
Chile	682	292	11/1/2011–5/26/2013

Parking lot imagery results [39]



[39] P. Butler, N. Ramakrishnan, E. Nsoesie, J. Brownstein, *IEEE Computer*, 2014.

Putting it all together [40]



[40] P Chakraborty, P Khadivi, B Lewis, A Mahendiran, J Chen, P Butler, EO Nsoesie, SR Mekaru, JS Brownstein, M Marathe, N Ramakrishnan, *SDM*, 2014

Putting it all together - contd (1)

Issues to consider

- How to combine different data soures: Model level fusion vs. data level fusion?
- How to accounting for initial, unreliable, estimates of official flu case counts
- Relationship between final flu counts and associated data sources is non-linear: How to get robust non-linear methods?
- Each data source ($\mathcal{X}$) represented as a mutlivariate weekly time-series with flu counts (P) as target variable - heavily skewed matrix and hence regression problem ill-defned

Putting it all together - contd (2)

Problem framework and solution sketches

- Solution (*MFN*): Matrix Factorization (dim reduction) + neeares neighbor search (non-linear) - similar to recommender systems [41]
  - Data-Target Matrix $M = [XP]$
  - Model

$$\widehat{\mathcal{M}}_{i,j} = \underbrace{b_{i,j}}_{\text{avg. value}} + \underbrace{U_i^T F_j}_{\text{dense subspace}} + \underbrace{F_j |\mathcal{N}(i)|^{-\frac{1}{2}} \sum_{k \in N(i)} (\mathcal{M}_{i,k} - b_{i,k}) x_k}_{\text{weighted nearest neighbor}} \quad (4)$$

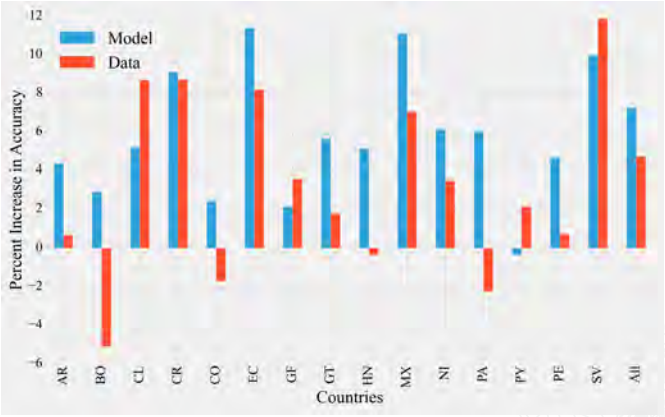
  - Fitting

$$b_*, F, U, x_* = \operatorname{argmin} \Big(\underbrace{\sum_{i=1}^{m-1} \left(\mathcal{M}_{i,n} - \widehat{\mathcal{M}}_{i,n} \right)^2}_{\text{loss on target variable}} + \underbrace{\lambda_2 \Big(\sum_{j=1}^{n} b_j^2 + \sum_{i=1}^{m-1} ||U_i||^2 + \sum_{j=1}^{n} ||F_j||^2 + \sum_{k} ||x_k||^2 \Big)}_{\text{L2 regularization on parameters}} \Big) \quad (5)$$

[41] Y. Koren, KDD 2008

Putting it all together - contd (3)

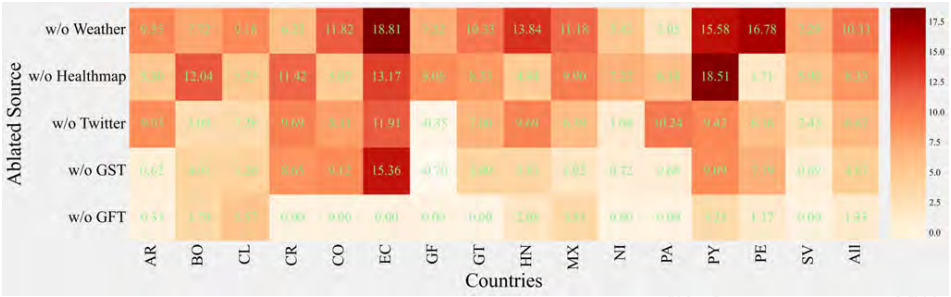
Model level (*MFN* with all data sets as $\mathcal{X}$ vs. Source level (Two level *MFN* on each data source + output of data source)



- Model level involves lower computational complexity.
- Model level fusion on avg. produces more accurate forecasts.

Putting it all together - contd (4)

Which sources are most important?



- Weather sources appear to contribute most to performance gains.
- Importance of sources such as Twitter can also be seen - able to capture changes from baseline.

Putting it aell together - contd (5)

Advantages of *MFN*

- Theoretical properties: Handles non-linearity and sparsity in a well defined framework. [42]
- Targeted: Custom non-linear and sparse methods allows for better reconstruction of target variables, sacrificing accuracy for non-target variables.
- Flexible: Easy to add different data sources and evaluate importance of each source via ablation.

Disadvantage

- Order of factorization needs to be found empirically
- Model is static - doesn't update over time
- reliable forecasting horizon is short

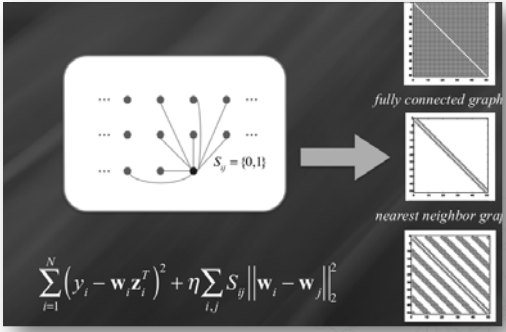
[42] Y. Koren, KDD 2008

Extending Forecasting Boundary [43]

- Dynamic Model

$$y_t = \mathbf{w}_t \mathbf{z}_t^T + \varepsilon_t$$

- Similarity Constraints on time points



[43] Z Wang, P Chakraborty, SR Mekaru, JS Brownstein, J Ye, N Ramakrishnan, *KDD*, 2015

Extending Forecasting Boundary - contd

- Dynamic GLMs to account for local variations.
- DARX: Identity Link

$$\sum_{i=1}^{N}\left(y_i-\mathbf{w}_i\mathbf{z}_i^T\right)^2+\eta\sum_{i,j}S_{ij}\left\|\mathbf{w}_i-\mathbf{w}_j\right\|_2^2$$

- DPARX: Poisson Link

$$\min_{\mathbf{w}}\quad \sum_i\left(\mathbf{w}_i\mathbf{z}_i^T-y_i\log(\mathbf{w}_i\mathbf{z}_i^T)\right)+\eta\sum_{i,j}S_{ij}\left\|\mathbf{w}_i-\mathbf{w}_j\right\|_2^2$$
$$s.t.\quad \mathbf{w}_i\mathbf{z}_i^T\geq 0,\ \forall i.$$

Step	Method	AR	BO	CL	MX	PE	PY	US
1	ARX	2.85	2.63	3.18	2.61	2.51	2.82	3.71
	MFN	2.33	2.41	2.34	2.69	2.48	2.54	3.73
	SARX	3.02	2.42	3.11	2.90	2.81	2.69	3.67
	DARX	3.05	2.74	3.12	2.78	2.50	2.65	3.71
	DPARX	3.13	2.82	3.18	2.97	2.64	2.81	3.72
2	ARX	2.38	2.22	2.83	1.88	1.90	2.57	3.47
	MFN	2.12	2.00	2.13	2.33	2.21	2.19	3.63
	SARX	2.75	2.03	2.76	2.61	2.43	2.43	3.61
	DARX	2.94	2.68	3.02	2.58	2.38	2.58	3.60
	DPARX	2.86	2.70	2.89	2.64	2.52	2.65	3.61
3	ARX	2.11	1.86	2.61	1.28	1.44	2.31	3.19
	MFN	1.99	1.87	2.11	2.14	2.10	2.09	3.33
	SARX	2.33	1.61	2.46	2.42	2.16	2.23	3.40
	DARX	2.66	2.36	2.77	2.37	2.26	2.46	3.41
	DPARX	2.58	2.53	2.56	2.45	2.37	2.52	3.42
4	ARX	1.84	1.61	2.39	0.88	1.12	2.22	2.92
	MFN	1.85	1.83	2.00	2.05	2.01	1.94	3.15
	SARX	2.12	1.41	2.30	2.22	2.02	2.09	3.30
	DARX	2.34	2.21	2.52	1.98	2.19	2.22	3.18
	DPARX	2.29	2.35	2.32	2.26	2.20	2.40	3.20

DPARX performs better overall.

Recommendations for future forecasting programs [44]

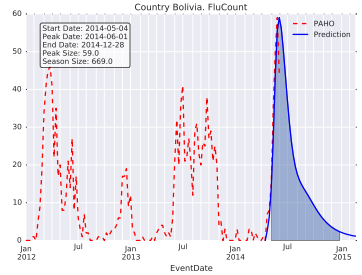
- Development of best practices for forecasting studies
- Head-to-head comparison of forecasting methods
- Assessment of model calibration
- Methods to incorporate subjective input into forecasting models
- Pilot studies to assess usefulness in real-world settings
- Improved mutual understanding between modelers and public health officials

[44] J Chretien, D George, E McKenzie *Online Journal of Public Health Informatics*, 2014

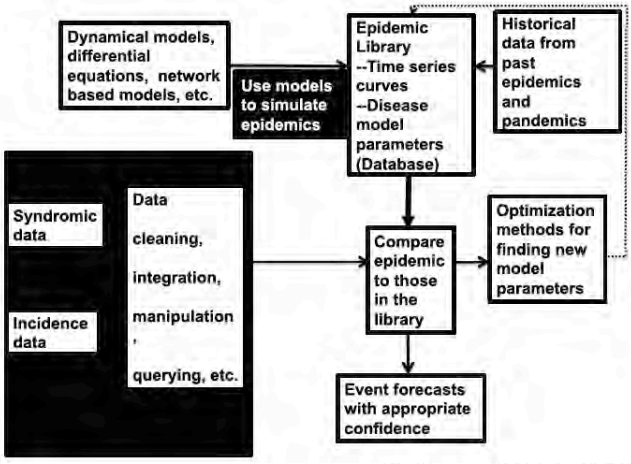
Epicurve Forecasting

Provide more actionable information for public health surveillance

- Start of season
- End of season
- Peak time
- Peak number of infections
- Total number of infections



Simulation Optimization Approach

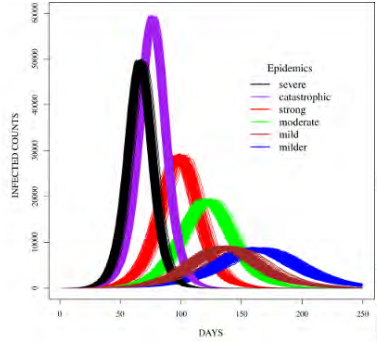


Simulation Optimization Approach - more details

- Parameters
  1. Transmissibility: The rate at which disease propagates through propagation
  2. Incubation period: Duration between infection and onset of symptoms
  3. Infectious period: Period during which infected persons shed the virus
- Typical strategy
  1. Seed a simulation (e.g., with simulated ILI count or with GFT data)
  2. Use a direct search parameter optimization algorithm (Nelder-Mead, Robbins-Monro) to find parameter sets
  3. Use the discovered parameter sets to forecast for next time frame (e.g., week)
  4. Repeat for the whole season

Classifying epidemic curves [45]

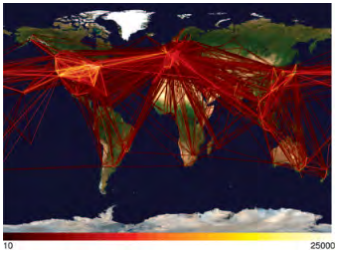
- Dirichlet process model to classify epidemic curves
- CRP representation of Dirichlet process model enabled classification into (Normal, Poisson, Negative Binomial)

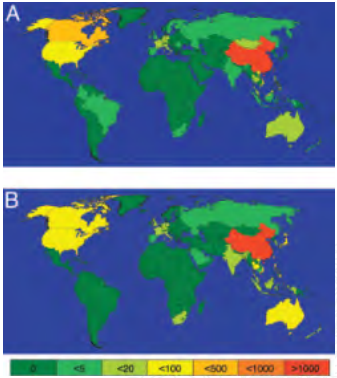


[45] EO Nsoesie, SC Leman, MV Marathe, *BMC infectious Dis.*, 2014

Forecasting Global Epidemic Spread [46]

- Uses aviation data to define a weighted network between airports
- Aims to replicate the global spread of SARS
- Stochastic SIR model to capture fluctuations





[46] L. Hufnagel, D. Brockmann, T. Geisel, *PNAS*, 2004

Mapping interactions using Twitter [47]

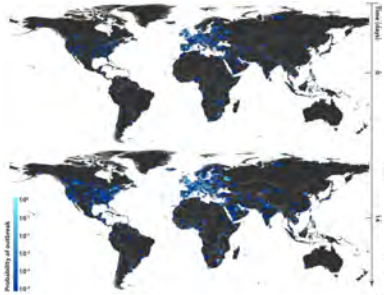
- Latent variable modeling to capture interactions between people solely through their Twitter status updates
- 51,000 individuals traveling between 100 airports in 75 cities
- 73,460 flights inferred and 445,812 meetings inferred from Twitter updates
- Goal was to explain variation in flu incidence across cities
  1. Raw airline traffic volume: 56%
  2. Health of individual passengers: 17%
  3. Physical encounters between healthy and sick individuals: 5%

[47] S Brennan, A Sadilek, H Kautz, *IJCAI*, 2013

A sobering study

Smallpox simulation under human mobility assumptions [48]

- Intentional release can have global effects
- Outbreaks can spread to different continents even before detection
- Outbreaks can happen in countries without necessary health infrastructure



Early detection + targeted vaccination sufficers over mass vaccination [49]

[48] B Goncalvez, D Balcan, A Vespignani, *Nature Scientific Reports*, 2013

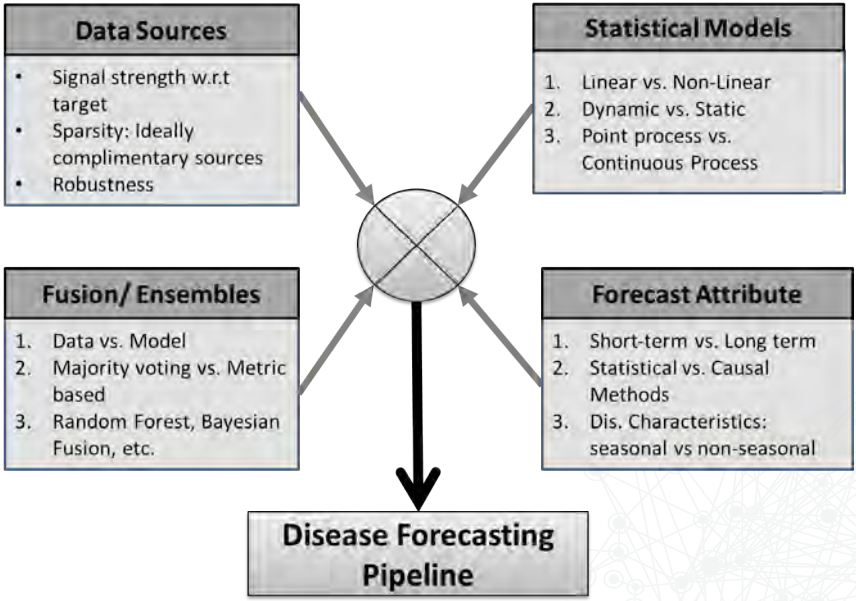
[49] S. Eubank et al., *Nature*, 2004.

Some Notable Mentions

- Exploring novel data streams to enhance disease surveillance [50].
- Spatio-temporal modeling of 2014 Ebola outbreak [51]
- Individualizing prediction of Disease Trajectories [52].

[50] BA Althouse et al., *EPJ Data Science*, 2015
[51] S Merler et al., *The Lancet Infectious Diseases*, 2015
[52] P Schulam, S Saria, *NIPS*, 2015

Rounding Up: The Big Picture



DEMO: EMBERS Disease Prediction Pipeline [53]

- Weekly predictions sent out for 15 Latin American countries

[53] EMBERS, http://dac.cs.vt.edu/research-project/embers/

Disease Forecasting Competitions

- IARPA Open Source Indicator Program. [54], 2010-2014
- CDC Influenza Prediction Challenge, 2013-current
- DARPA Chikungunya Prediction Challenge, 2014
- RAPIDD Ebola Challenge, 2015

[54] EMBERS, http://dac.cs.vt.edu/research-project/embers/

Controlling the spread of epidemics

General problem

Given a partially known network, initial conditions and disease model:

- Design interventions for controlling the spread of an epidemic
- Different objectives, such as: number of infections, peak (maximum number of infections at any time) and time of peak, logistics

Complement of influence maximization: much more challenging

Correspondence with GDS

Optimization Problems in GDS

Let $\mathcal{P}(\mathcal{S})$ denote the phase space of a given SyDS $\mathcal{S}(G)$. Modify $\mathcal{S}$ optimally so that the set of reachable states in $\mathcal{P}(\mathcal{S})$ satisfies a given property.

- **Property** $\mathcal{P}_1$: #1's in configuration is "small" $\equiv$ interventions that try to minimize the outbreak size
  - Modify graph by removing nodes (vaccination) or edges (quarantining) so that fixed points in $\mathcal{P}(\mathcal{S})$ satisfy $\mathcal{P}_1$
- Similarly, reducing epidemic duration $\equiv$ reducing transient length in $\mathcal{P}(\mathcal{S})$

GDS view: enables many algorithmic and complexity results to be translated across systems

Objectives and strategies for controlling epidemics

- Different objectives
  - Expected outbreak size
    - In the whole population and in different subpopulations
    - Other economic costs
  - Duration of epidemic
  - Size and time of peak
- Different kinds of strategies[55]
  - Decrease β, the transmissibility
    - Quarantining and social distancing of infected individuals
    - Hand washing and other hygienic precautions
    - Treating infected individuals with antimicrobials
  - Reduce number of susceptibles: vaccination
  - Reduce infectious duration: treatment with antimicrobials
  - Increase δ: culling animals
- Interventions can be modeled in networks as node deletions (vaccination), edge deletion (quarantining) and reducing β on edges

[55] Dimitrov and Meyers, *INFORMS*, 2010

Different kinds of issues in studying interventions as optimization problems

- Resource constraints, e.g., budget for vaccines to use
- Complex and multiple objective functions, e.g., expected outbreak size, peak size and duration
  - Need multi-criteria optimization
- Implementability and compliance
  - Interventions should be described succinctly
- Individual utility: game-theoretical issues
- Computationally very hard problems
  - Computing basic properties related to epidemics (e.g., probability that a node gets infected) is #P-hard in network models
  - Optimization problems NP-hard even for very simplistic settings (e.g., SI model or simple contagion)
  - Metaheuristics do not give any insights into how well they perform (relative to the best possible).

Outline for this section

- Centralized interventions: minimizing social cost
  - Vaccine allocation problems in the SI model on arbitrary networks: bicriteria approximation algorithm
  - Vaccination strategies based on the spectral characterization in the SIS model: approximation algorithms
  - Optimal vaccination policies using ODE approach
- Vaccination games
  - Network based vaccination game in the SI model
  - Game based on spectral characterization
- Sequestration of critical populations
- Combining individual and social objectives: anti-viral distribution problem (discussed later)

Vaccination allocation problems

Optimal vaccine allocation problem (OVAP)

Given a graph G and limited supply of vaccine (B doses), how should it be allocated to different sub-populations so that different epidemic outcomes are optimized?

- Algorithm has approximation ratio α if for any instance $\mathcal{I}$ of the problem, it finds a solution S such that $\text{cost}(S) \leq \alpha \cdot \text{cost}(OPT(\mathcal{I}))$
- Simplest setting: SIR model with transmission probability 1 ("highly contagious disease")
  - NP-hard to approximate within factor of $O(n^{\delta})$ for any $\delta < 1$
  - If initial infected set is given: bicriteria-approximation, which uses B/ϵ vaccines, so that #infections is at most $1/(1-\epsilon)$ times optimal[56] [57]
  - If initial infection is random: $O(\log n)$ approximation[58]

[56] A. Hayrapetyan, D. Kempe, M. Pal and Z. Svitkina, *ESA*, 2005

[57] S. Eubank, V. S. Anil Kumar, M. Marathe, A. Srinivasan and N. Wang, *AMS DIMACS*, 2005

[58] V. S. Anil Kumar, R. Rajaraman, Z. Sun and R. Sundaram, *IEEE ICDCS*, 2010

Integer program for OVAP (edge version)

I: set of initial infections, B: budget on #edges to remove

$$\begin{aligned} \min \sum_v x(v) & \quad \text{subject to} \\ \forall e = (u,v) : y(e) &\geq x(u) - x(v) \\ \forall u \in I : x(u) &= 1 \\ \sum_e y(e) &\leq B \\ x(u), y(e) &\in \{0,1\} \end{aligned}$$

Lemma

The above IP is equivalent to the edge version of OVAP.

- Let $S = \{u : x(u) = 1\}$, $E' = \{e : y(e) = 1\}$
- cut $(S, V - S) \subseteq E'$
  - For edge $e = (u, v)$ with $u \in S, v \in V - S$, $x(u) - x(v) = 1$
- All nodes reachable from I are contained in S
- How do we get a polynomial time algorithm?

Bicriteria approximation algorithm for OVAP (edge version)

- Let (x^*, y^*) be the optimal solution to the following LP:

$$\begin{aligned} \min \sum_v x(v) \qquad & \text{subject to} \\ \forall e = (u, v) : y(e) &\geq x(u) - x(v) \\ \forall u \in I : x(u) &= 1 \\ \sum_e y(e) &\leq B \\ x(u), y(e) &\in [0, 1] \end{aligned}$$

- Choose $r \in [1 - \epsilon, 1]$ uniformly at random.
- Let $S = \{v : x^*(v) \geq r\}$. Choose critical set $E' = \{e = (u, v) : u \in S, v \in \bar{S}\}$.

Analysis

Lemma (Hayrapetyan et al., 2005, Eubank et al., 2005)

The above algorithm chooses at most B/ϵ edges, and ensures that the number of infected nodes is at most $1/(1-\epsilon)$ times the optimal.

1. $\sum_{v \in V} x^*(v) \geq \sum_{v \in S} x^*(v) \geq (1-\epsilon)|S|$, which implies $|S| \leq \frac{1}{1-\epsilon} \sum_{v \in V} x^*(v)$
2. Edge $e = (u, v) \in E'$ if r is between $x^*(u)$ and $x^*(v)$.
3. $\Pr[e \in E'] \leq \frac{|x^*(u) - x^*(v)|}{\epsilon} \leq y(e)/\epsilon$, so that $Exp[|E'|] \leq B/\epsilon$.

Controlling epidemics in the SIS model

- Reduce spectral radius below T to ensure the epidemic dies out fast.
- Spectral radius can be reduced by deleting nodes (vaccination) or edges (social distancing)

Spectral Radius Minimization (SRM) problem

- *Given*: graph G=(V, E), threshold T and cost $c(e)$ for edges
- *Objective*: choose cheapest set $E' \subseteq E$ of edges to delete, so that $\lambda_1(G[E - E']) \leq T$.

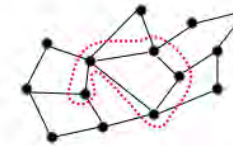
Similarly: node version

Reducing the spectral radius to control epidemic spread

- Interventions (node/edge deletion) to reduce spectral radius below given threshold
- NP-hard to approximate within a constant factor
- Heuristics based on components of the first eigenvector and degree: [Tong et al., 2012], [Van Mieghem et al., 2011]
- Node version: if G has a power law degree sequence with exponent $\beta > 2$ and $T^2 \leq cd_{max}$, then a high degree strategy gives an $O(T^{\beta-1})$ approximation ([S. Saha, A. Adiga and A. Vullikanti, AAAI 2014])
- Node version: $\Theta(1)$ approximation by a high degree strategy in Chung-Lu random graphs with power law weights with exponent $\beta > 2$.

Some notation and properties

- A: adjacency matrix of G with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$
- Let $\mathcal{W}_k(G)$ denote the set of closed walks of length k
- Let $W_k(G) = |\mathcal{W}_k(G)|$
- Edge e "hits" walk w if $e \in w$.
- $n(e, G)$: #walks in $\mathcal{W}_k(G)$ containing edge e
- Let $E_{opt}(T)$ denote the optimum set of edges, whose deletion reduces the spectral radius below T
- $\sum_i \lambda_i^k = \sum_i A_{ii}^k = \sum_{w \in \mathcal{W}_k(G)} d(w)$, where $d(w)$ is the number of distinct nodes in walk w



An $O(\log^2 n)$-approximation algorithm

Algorithm GreedyWalk: Pick the smallest set of edges E' which hit at least $W_k(G) - nT^k$ walks, for even $k = c \log n$

- Initialize $E' \leftarrow \phi$
- Repeat while $W_k(G[E \setminus E']) \geq nT^k$:
  - Pick the $e \in E \setminus E'$ that maximizes $\frac{n(e, G[E \setminus E'])}{c(e)}$
  - $E' \leftarrow E' \cup \{e\}$

Lemma

We have $\lambda_1(G[E \setminus E']) \leq (1+\epsilon)T$, *and*
$c(E') = O(c(E_{\mathrm{OPT}}(T)) \log n \log \Delta/\epsilon)$ *for any* $\epsilon \in (0,1)$.

Similar bound for node version

Proof: bounding spectral radius of residual graph

By construction: $W_k(G') \leqslant nT^k$, where $G' = G[E - E']$. Therefore,

$$\sum_{i=1}^{n} \lambda_i(G')^k = \sum_i A_{ii}^k = \sum_{w \in \mathcal{W}(G')} d(w) \leq kW_k(G')$$

$$\Rightarrow \sum_{i=1}^{n} \lambda_i(G')^k \leqslant nkT^k$$

and therefore, $\lambda_1(G') \leqslant 2^{(\log n + \log k)/k} T$

$$\leqslant (1+\epsilon)T, \text{ for } k \geqslant \frac{2}{\epsilon} \log n.$$

Proof: bounding $c(E')$

- Let E_{HITOPT} be optimal solution for the partial covering instance: cheapest subset of edges that hits at least $W_k(G) - nT^k$ walks.
- Standard greedy analysis $\Rightarrow c(E') = O(c(E_{\text{HITOPT}}) \log H)$, where $H =$ #elements in covering instance.
- Elements= walks $\Rightarrow H = |W_k(G)| \leqslant n\Delta^k$
- By definition, $\lambda_1(G[E - E_{\text{OPT}}(T)]) \leq T$. Therefore, $W_k(G[E - E_{\text{OPT}}(T)]) \leq \sum_{i=1}^{n} \lambda_i(G[E - E_{\text{OPT}}(T)])^k < nT^k$.
- $\Rightarrow c(E_{\text{HITOPT}}) \leq c(E_{\text{OPT}}(T))$
- $c(E') = O(E_{\text{OPT}}(T) \log n \log \Delta)$.

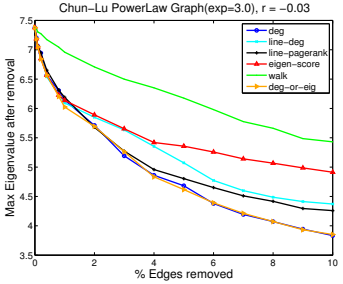
Improvement to $O(\log n)$ factor

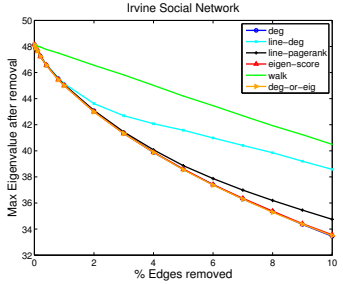
- Partial coverage problem: primal-dual algorithm of [Gandhi et al., 2004] for selecting a minimum cost collection of sets that covers at least k elements, with $O(f)$-approximation, where f is the maximum number of sets containing any element
- Our set system:
  - Sets $\equiv$ edges, elements $\equiv$ walks in $\mathcal{W}_k$
  - $f = O(\log n)$, since walks have length $k = O(\log n)$
- Set system of size $n^{O(\log n)}$, so cannot apply primal-dual algorithm of [Gandhi et al., 2004] directly
  - Can do updates implicitly and get polynomial time $O(\log n)$-approximation
  - Results in $c(E') = O(c(E_{\mathrm{OPT}}(T)) \log n)$, $\lambda_1(G[E - E']) \leq (1 + \epsilon)T$
- Constant factor approximation by semidefinite programming based rounding.

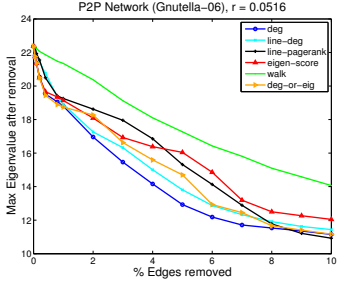
Heuristics that work well

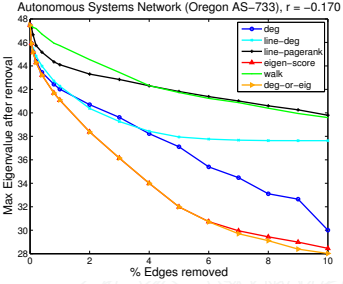
- Pick edges $e = (i, j)$ in decreasing order of $\mathit{eigenscore}(i, j) = x^1(i) \cdot x^1(j)$ [Tong et al., 2012], [Van Mieghem et al., 2011]
- Pick edges $e = (i, j)$ in decreasing order of $\mathit{degscore}(i, j) = d(i)d(j)$ [Van Mieghem et al., 2011]
- Hybrid rule: pick edge from either order whose removal causes the largest reduction in λ_1

Empirical analysis of different heuristics









Analysis of degree heuristic

Lemma

Let G be a power law graph with exponent $\beta > 2$, where β is a constant and let threshold T satisfy $T^2 \leqslant c\Delta$ for a constant $c < 1$. Then, the number of edges removed by the degree heuristic is $O(T^{\beta-2}|E_{\mathrm{OPT}}(T)|)$.

Lemma

Let $G(\mathbf{w})$ be a Chung-Lu random power law graph on n nodes with exponent $\beta > 2$ and $w(V)$ a constant. Let T be the threshold satisfying $\max_{i \in V} w_i > T^2$ and $T = \Omega(\log n)$. Then, the number of edges removed by the degree heuristic is $O\big((\log n)^{\beta-1}\big)|E_{\mathrm{OPT}}|$.

Other results for network models

- Analysis of vaccination strategies based on degrees[59]
- Vaccination schemes based on PageRank[60]
- Vaccination strategies in terms of the cut width of the graph[61]

[59] C. Borgs, J. Chayes, A. Ganesh and A. Saberi, *Random Structures and Algorithms*, 2009
[60] F. Chung, P. Horn and A. Tsiatas, *Internet Mathematics*, 2009
[61] K. Drakopoulos, A. Ozdaglar and J. Tsitsiklis, 2014

Compartmental differential equation based approach for OVAP[65]

- Age-structured differential equation model for H1N1
  - Mixing between age groups based on survey data[62]
  - $R_0 = 1.4$ for swine flu[63]
  - Different outcomes: deaths, infections, years of life lost, contingent valuation, and economic costs
    - Mortality considerations based on 1957 and 1918 pandemics
    - Valuations and economic costs of sickness and death from health economics literature[64]
- CDC guidelines for swine-flu: prioritize vaccination for children 6 months to 5 years, and 5-18 years

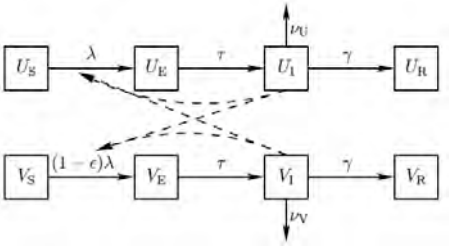
[62] J. Mossong et al., *PLoS Med.*, 2008

[63] C. Fraser et al., *Science*, 2009

[64] Such as: A. C. Haddix et al., *Oxford University Press*, 1996; M. Meltzer et al., *Emerg. Infec. Dis.*, 1999

[65] Medlock and Galvani, *Science*, 2009

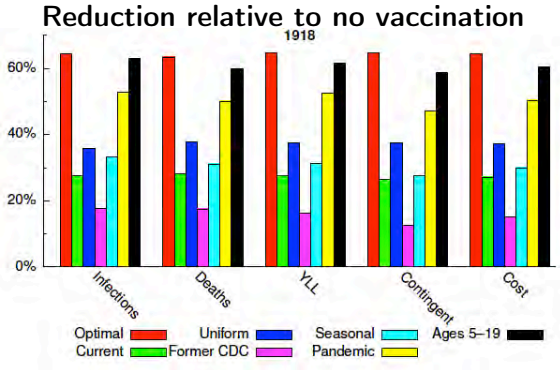
Coupled differential equation model



$$
\begin{aligned}
\frac{dU_{Sa}}{dt} &= -\lambda_a U_{Sa} \\
\frac{dU_{Ea}}{dt} &= \lambda_a U_{Sa} - \tau_a U_{Ea} \\
\frac{dU_{Ia}}{dt} &= \tau_a U_{Ea} - (\gamma_{U_a} + \nu_{U_a}) U_{Ia} \\
\frac{dU_{Ra}}{dt} &= \gamma_{U_a} U_{Ia} \\
\frac{dV_{Sa}}{dt} &= -(1-\epsilon_a)\lambda_a V_{Sa} \\
\frac{dV_{Ea}}{dt} &= (1-\epsilon_a)\lambda_a V_{Sa} - \tau_a V_{Ea} \\
\frac{dV_{Ia}}{dt} &= \tau_a V_{Ea} - (\gamma_{Va} + \nu_{Va}) V_{Ia} \\
\frac{dV_{Ra}}{dt} &= \gamma_{Va} V_{Ia}
\end{aligned}
$$

- 17 different age classes, indexed by a
- $U_{Sa}(t), U_{Ea}(t), U_{Ia}(t), U_{Ra}(t)$: number of unvaccinated susceptible, latent, infectious and recovered
- $V_{Sa}(t), V_{Ea}(t), V_{Ia}(t), V_{Ra}(t)$: number of vaccinated susceptible, latent, infectious and recovered
- Vaccine allocation: $\sum_a V_{Sa}(t) + V_{Ea}(t) + V_{Ia}(t) + V_{Ra}(t)$

Results



- Optimal allocation to different age groups depends on the objective and the number of available doses
- Significantly better than CDC guidelines at that time, allocation to age group 30-39
- High sensitivity to disease model and other parameters

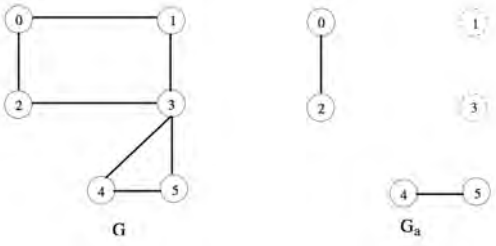
Decentralized vaccination decisions

- Vaccination or other interventions leads to cost (say C) for an individual
- Infection cost L (typically higher than C)
- *Herd immunity*: if enough neighbors get vaccinated, then low probability of infection
- Individual utility for vaccination: natural game-theoretical setting

Network vaccination game [Aspnes, Chang and Yampolskiy, 2006]

- Players: each node $i \in V$
- Strategies: $a_i = 1$ if i is vaccinated (else $a_i = 0$)
  - Strategy profile denoted by $\mathbf{a}$
  - $P(\mathbf{a}) = \{i : a_i = 1\}$ is the set of vaccinated nodes
  - $G_{\mathbf{a}} = G - P(\mathbf{a})$ (attack graph: remove all vaccinated nodes)
  - Assume vaccine has 100% efficacy and cost C_i for node i
- Epidemic model: SI with random initial infection
  - If $a_i = 0$, and (random) source is in the component containing i in $G_{\mathbf{a}}$ then node i incurs infection cost L_i
  - $p_i(\mathbf{a}) = \Pr[i \text{ gets infected} | a_i = 0]$
    - $p_i(\mathbf{a}) = k_i/n$, where k_i is the size of the component in $G_{\mathbf{a}}$ containing i
- $\text{cost}_i(\mathbf{a}) = a_i C_i + (1 - a_i) L_i p_i(\mathbf{a})$

Network vaccination game



$$\begin{aligned}
\text{cost}(\mathbf{a}) &= \sum_j \text{cost}_j(\mathbf{a}) = \sum_j a_j C + (1-a_j) L p_j(\mathbf{a}) \\
&= C|P(\mathbf{a})| + \sum_{j \in \text{comp } i} \frac{L k_i}{n} \\
&= C|P(\mathbf{a})| + \frac{L}{n} \sum_i k_i^2,
\end{aligned}$$

where $G_{\mathbf{a}}$ has components of sizes $k_1, k_2, \ldots$

Nash equilibrium and social optimum

- Strategy $\mathbf{a}$ is a Nash equilibrium (NE) if no node i has incentive to deviate unilaterally
- Pure NE: $a_i \in \{0, 1\}$ for all i
- Social optimum: strategy $\mathbf{a}$ that minimizes cost($\mathbf{a}$)
- General questions:
  - Structure of NE, complexity of computing NE
  - Price of Anarchy: maximum ratio of cost of NE to social optimum
  - Which nodes are likely to be vaccinated in NE? How to increase that fraction?

Structure of NE

- Given strategy $\mathbf{a}$, let $S(i)$ denote the component containing node i, when all vaccinated nodes (other than i, in case $a_i = 1$) are removed
- Threshold $t = nC/L$

Theorem

Strategy $\mathbf{a}$ *is a NE if and only if*

1. *For all* i *such that* $a_i = 1$: $S(i) \geq t$
2. *For all* i *such that* $a_i = 0$: $S(i) < t$
3. *For all* i *such that* $0 < a_i < 1$: $S(i) = t$

Corollary

If $\mathbf{a}$ *is a pure NE:*

1. *Every component in attack graph* $G_{\mathbf{a}}$ *has size at most* t
2. *If* $a_i = 1$, *then adding* i *to* $G_{\mathbf{a}}$ *leads to a component of size* $> t$

Computing NE

- $S_{big}(\mathbf{a})$: set of components of size $> t$ in $G_{\mathbf{a}}$
- $S_{small}(\mathbf{a})$: set of components of size $\leq t$
- For $\mathbf{a}$, define potential function

$$\Phi(\mathbf{a}) = \sum_{A \in S_{big}(\mathbf{a})} |A| - \sum_{A \in S_{small}(\mathbf{a})} |A|$$

- Best response method: starting from any $\mathbf{a}$, in each iteration, a node i switches if its cost reduces
  - Converges to pure NE in $O(n)$ iterations

Computing NE

$$\Phi(\mathbf{a}) = \sum_{A \in S_{big}(\mathbf{a})} |A| - \sum_{A \in S_{small}(\mathbf{a})} |A|$$

- $-n \leq \Phi(\mathbf{a}) \leq n$
- Suppose node i switches from $a_i = 0$ to $a_i = 1$
  - $\Rightarrow i$ in a big component in $G_{\mathbf{a}}$
  - If a_i becomes 1, one big component reduces in size by at least one
  - $\Rightarrow \Phi$ decreases by at least one (more if smaller components are formed when i is removed from $G_{\mathbf{a}}$
- Suppose node i switches from $a_i = 1$ to $a_i = 0$
  - $\Rightarrow$ node i becomes part of small component
  - Sum of sizes of small components increases by $1 \Rightarrow \Phi$ decreases
- Convergence in at most $2n$ steps

Computing the social optimum

- $\text{cost}(\mathbf{a}) = C|P(\mathbf{a})| + \frac{L}{n}\sum_i k_i^2$
- If we "guess" #vaccinated nodes $M = |P(\mathbf{a})|$, this reduces to the *minimum sum of squares partitioning* problem:
  - remove at most M nodes from G, partitioning it into components $V_1, \ldots, V_k$ such that $\sum_i |V_i|^2$ is minimized
- $O(\log^2 n)$-approximation[66]
- Improved to $O(\log n)$ approximation
  If initial infection is random: $O(\log n)$ approximation[67]

[66] [Aspnes et al., 2006]

[67] V. S. Anil Kumar, R. Rajaraman, Z. Sun and R. Sundaram, *IEEE ICDCS*, 2010

Simplifying assumptions

- Simplistic disease model, arbitrary contact network
  - Aspnes et al. assume SI model (highly contagious disease)
- Complex disease models, simplistic networks
  - Differential equation approachs, starting from [Bauch and Earns, 2004]
  - Solution can be characterized fully in completely mixed models in terms of R_0 and vaccine risk/cost
- Challenge: how can we take both into account?

Epidemic containment game in the SIS model

- Let $\mathbf{a} = (a_1, a_2, ..., a_n)$ denotes the strategy profile with $a_x = 1$ denoting that node x is vaccinated
- Let $P = P(\mathbf{a}) = \{x \in V : a_x = 1\}$ denote the set of vaccinated nodes
- Cost for node v, given strategy vector $\mathbf{a}$:

$$\text{cost}(v, \mathbf{a}) = \begin{cases} C, & \text{if } a_v = 1, \\ L, & \text{if } a_v = 0 \text{ and } \lambda_1(G[V - P(\mathbf{a})]) < T, \\ L_e, & \text{if } a_v = 0 \text{ and } \lambda_1(G[V - P(\mathbf{a})]) \geq T. \end{cases}$$

Nash equilibrium $\mathbf{a}$: if no node v has incentive to switch unilaterally, given that other players' strategies are fixed
Social cost $\text{cost}(\mathbf{a}) = \sum_v \text{cost}(v, \mathbf{a})$

Structure of Nash equilibria

- Assume $C = 1, L = 0, L_e > 1$
- The strategy corresponding to any *minimal* set S such that $\lambda_1(G[V - S]) < T$ is a NE.
- Finding the social optimum of an EC game is NP complete. Moreover, the cost of social optimum cannot be approximated within a factor of 1.3606 unless P=NP.

Results: price of anarchy

- Let G be a power law graph with exponent $\beta > 2$, where β is a constant and let $T^2 \leq c\Delta$ for a constant $c < 1$, where Δ is the maximum node degree. Then, the price of anarchy is $O\left(T^{2(\beta-1)}\right)$.
- Erdős-Rényi random graph model: if $G = G(n, p)$, for $p \geq \frac{c}{n}$, where c is a suitably large constant and $np \geq (1+\delta)T^2$ for any positive constant δ, the price of anarchy is almost surely $O\left(\frac{\log n}{np}\right)$.

Results: price of anarchy

- Chung-Lu random graph model: given a weight sequence $\mathbf{w} = (w(v_1, V), w(v_2, V), ..., w(v_n, V))$ for nodes $v_i \in V$, the random graph $G(\mathbf{w})$ is constructed in the following manner:
  - add edge (v_j, v_k) with probability $\frac{w(v_j,V)w(v_k,V)}{\sum_{v_i \in V} w(v_i,V)}$

Theorem

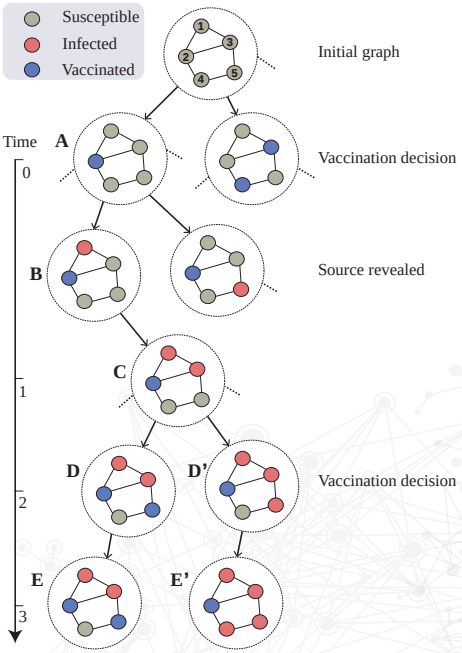
Consider a Chung-Lu random power law graph $G(\mathbf{w})$ of n nodes and power law exponent $\beta > 2$. Suppose $w(V) = \sum_v w(v)/|V| = O(1)$ and $w_{\max} = \max_v\{w_v\} \geq (1+\delta)T^2 w(V)$ for some constant δ and $T = \Omega(log^2 n)$. The price of anarchy in $G(\mathbf{w})$ is $\theta(T^{2(\beta-1)})$ almost surely.

Temporal vaccination games

- Vaccination games so far: decisions only made at time 0 (before epidemic starts)
- Typically: people delay vaccination decisions
  - Vaccination rates follow disease incidence rates
- How do we model delay in vaccination?
- What factors affect delay?

Temporal vaccination game in the SI model

- $\mathcal{T} = \{t_0, t_1, \ldots\}$: times when vaccination decisions can be made
- C_v^t: cost of vaccination at time t for node v
- L_v: cost of infection for node v
- SI model, with random source of infection
- $Y(v, s, t) \in \{0, 1\}$: decision by node v at time t, when source is s
- Structure of game
  - Vaccination game played at time $t_0 = 0$
  - Random source revealed (full information)
  - At each time $t \in \mathcal{T}$, $t > t_0$, vaccine game played



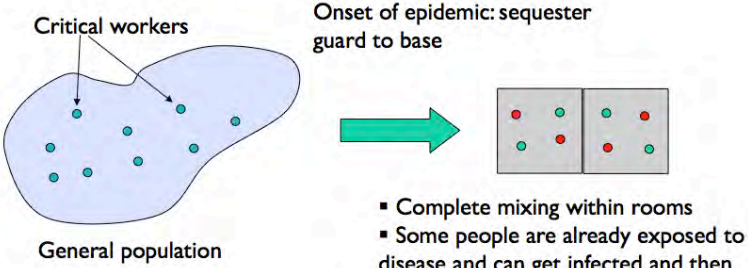
Our results[69]

- All vaccination decisions made either at time 0 or the first time $T \in \mathcal{T}$, $T > 0$.
- Significant variation in number of nodes that defer vaccination decision
- Pure NE need not exist, determining when is NP-complete
- Social optimum can be approximated within a factor of $2T$
- Resource constraints lead to big changes in structure of equilibria[68]

[68] A. Adiga and A. Vullikanti, AAAI 2016

[69] A. Adiga, S. Venkat and A. Vullikanti, *IEEE INFOCOM*, 2016

Sequestration for protecting critical sub-populations

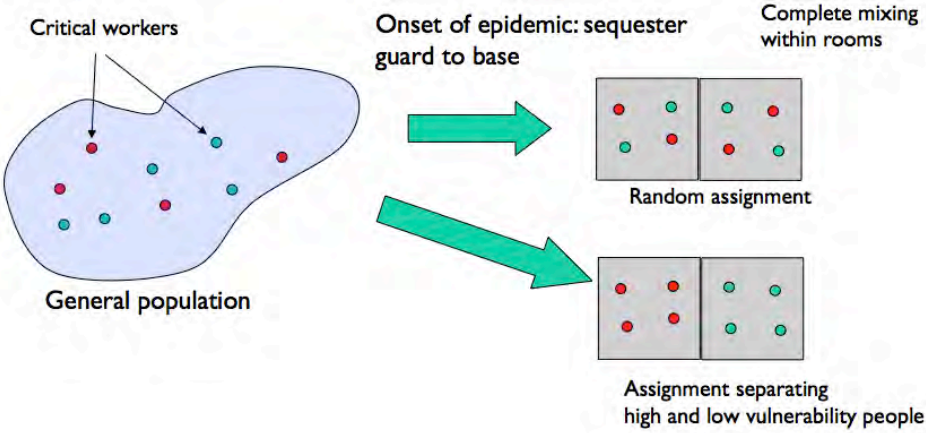


Goal

Partition people into groups so that overall outbreak is minimized

- 1918 epidemic: thought to have spread primarily through military camps in Europe and USA
- Large outbreaks in naval ships

Sequestration problem



- Cannot do much without any information about the individuals
- Assume estimates $f(i)$ of vulnerability: probability the node i gets infected (for some initial conditions)

Sequestration problem

Sequestration Problem

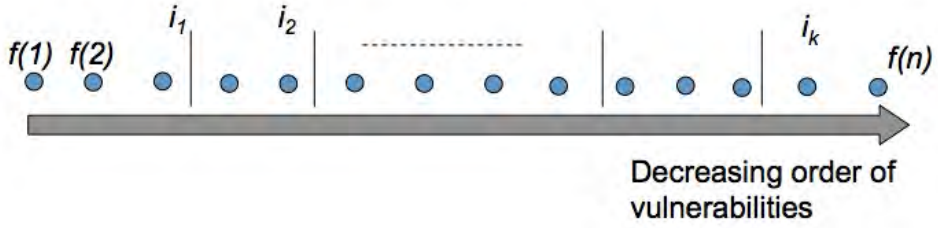
Given: a set V of people to be sequestered in a base, group size m, number of groups k and vulnerability $f(i)$ for each $i \in V$.
Objective: partition V into groups $V_1, V_2, \ldots, V_k$ so that the expected number of infections is minimized.

- Assume complete mixing within each group with transmission probability p among any pair of nodes
- Individual i is (externally) infected with probability $f(i)$. Additionally, the disease can spread within each group, following an SIR process.

- Efficient exact algorithm for group sequestration[70]
- Significantly outperforms random allocation

[70] C. Barrett et al., *ACM SIGHIT International Health Informatics Symposium*, 2012

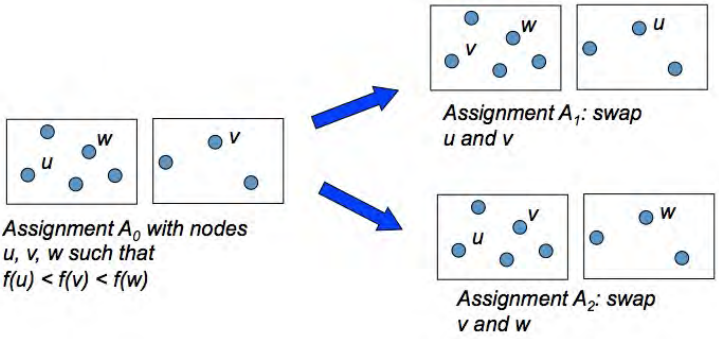
Structural Property of Optimum Solution



Theorem

There exist integers $i_1, \ldots, i_k$ and an optimal solution such that the jth group contains all the nodes between $i_1 + i_2 + \ldots + i_{j-1} + 1$ and $i_1 + i_2 + \ldots + i_{j-1} + i_j$.

Main idea of proof: swapping lemma



$Cost(A)$: expected outbreak size for a specific assignment A of people to groups

Lemma

$\min\{Cost(A_1), Cost(A_2)\} < Cost(A_0)$

Research challenges

- Need to find implementable strategies
  - Identifiable attributes such as: demographics, geographical locations
  - Temporal strategies: Markov Decision Processes
- Complex objectives and constraints
  - Logistics of production and delivery of medicines
  - Economies of scale
  - Resource constraints, e.g., public health staff
- Uncertainty in network and disease parameters
  - Network, state and model parameters not known
  - Multiple and evolving disease strains
- Compliance and behavioral changes
  - Network co-evolves with epidemic

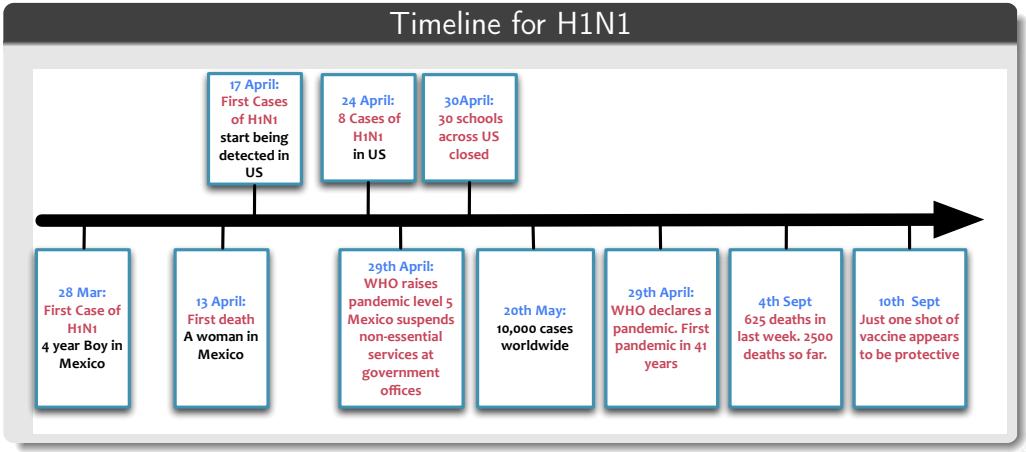
Putting it all together: outline

- Recent real-world examples: 2009 H1N1 and 2014-15 Ebola Outbreak
- Data, Synthetic realistic social networks
- Detailed agent-based simulations
- Case studies
- Computational Ecosystems
- Extensions

Supporting outbreak response and preparedness exercises

Unfolding of a pandemic

Timeline: http://www.nbcnews.com/id/30624302/ns/health-cold_and_flu/t/timeline-swine-flu-outbreak/#.U_LBJUgdXxs



Pandemic Influenza Planning

Problem

How can we prepare for a likely influenza pandemic?

Study Design

Population: Chicago Metropolitan area, 8.8 million individuals

Disease: Pandemic Influenza, R0 1.9, 2.4, and 3.0,varying proportion symptomatic

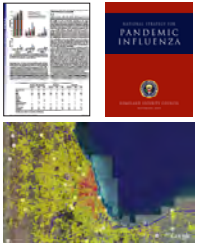
Interventions: Social distancing, School closure, and prophylactic anti-virals triggered when 0.01%, 0.1%, and 1% of population is infected

Modeling tool

EpiSims manually configured to 6 different scenarios specified by decision-maker

Policy recommendations

Non-pharmaceutical interventions can be very effective at moderate levels of compliance if implemented early enough



Antiviral Distribution Planning

Problem

What is the impact of encouraging the private stockpiling of antiviral medications on an Influenza pandemic?

Study Design

Population: Chicago Metropolitan area, 8.8 million individuals

Disease: Pandemic Influenza, calibrated to 33% total attack rate

Parameter of interest: Different methods of antiviral distribution (private insurance-based, private income-based, public, random

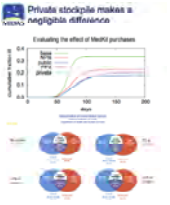
Other parameters: Percent taking antivirals, Positive predicitive value of influenza diagnosis, School closures, Isolation

Modeling tool

EpiFast, specifically modified for this study and manually configured to explore multiple parameter interactions and sensitivities

Policy recommendations

Private stockpiling of antiviral medications has a negligible impact on the spread of the epidemic and merely reduces demands on the public stockpile.



Emergence of H1N1 Influenza

Problem

What are the characteristics of this novel H1N1 influenza strain and their likely impact on US populations?

Study Design

Population: Various metropolitan areas throughout the US

Disease: Novel H1N1 Influenza

Parameters Studied:

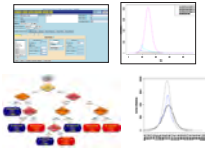
- Levels and timing of Social Distancing, School Closure, and Work Closure
- Viral mutation causing diminished immunity, seasonal increase in transmissibility, size of 2nd wave, timing of changes, reduced vaccine uptake

Modeling tool

Initial configurations with DIDACTIC, then manual configurations were made to web-enabled epidemic modeling and analysis environment based on EpiFast simulation engine

Policy recommendations

- The novel strain of H1N1 influenza presents a risk to becoming a pandemic, limited data make predicting exact disease characteristics difficult
- Several conditions would have to align to allow a sizeable 3rd wave to occur



Adenovirus Pandemic Simulation and Analysis

Problem

How can decision makers become familiar with the challenges and decisions they are likely to encounter during a national pandemic, mainly centered on the allocation of scarce resources?

Study Design

Population: US (contiguous 48 states)

Disease: Adenovirus 12v

Interventions: None (request for unmitigated disease)

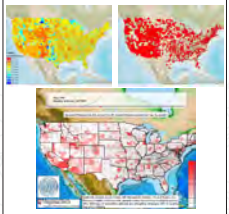
Other Details: Novel fusion of both coarse scale national level model and high resolution state-wide transmission to generate estimates of demand for scarce medical resources

Modeling Tool

National Model and EpiFast

Policy Recommendations

Nationwide epidemics of a severe respitory illness will create complex demands on the medical infrastructure, which will require high-level coordination to maximize the delivery of care.



Fall 2006 — Summer 2007 — Spring - Fall 2009 — Summer 2013

2006 — 2009 — 2012

Recent example: Ebola outbreak in Africa

- Largest Ebola outbreak till date: 5 countries; 28000 cases; 10500 deaths (WHO, Sept 2015)

Recent example: Ebola outbreak in Africa

- Largest Ebola outbreak till date: 5 countries; 28000 cases; 10500 deaths (WHO, Sept 2015)
- Beautifully done NY Times webpage: `http://www.nytimes.com/interactive/2014/07/31/world/africa/ebola-virus-outbreak-qa.html`

Important Questions

1. How many people have been infected?
2. Where is the outbreak?
3. How did it start; tracing the first few cases.
4. Chances of getting Ebola in the US?
5. How does this compare to past outbreaks?
6. How contagious is the virus? Why is Ebola so difficult to contain?
7. How does the disease progress? How is the disease treated?
8. Where does the disease come from?

NY Times Graphics

Details of the disease

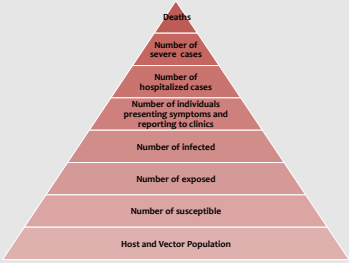
- Started in a village near Gueckedou, Guinea,
- Natural reservoirs (fruit bats)
- Cultural practices, lack of strong public health infrastructure and lack of trust due in public officials due to long civil war all played a part.
- Human to human transmission: body fluids
- No effective vaccines, or anti-vitals developed during the outbreak
- **Disease parameters:** (*i*) Average incubation period: approximately 12 days; (*ii*) Symptom onset to recovery/death: approximately 9 days; (*iii*) Overall case-fatality ratio: approximately 54%
- **What worked?:** Social interventions and boots on the ground public health response: Placement of Ebola treatment units (ETU); Personal protective equipment (PPE) for Healthcare workers; Improved surveillance and contact tracing; Community engagement - Safe burial practices
- Long term socio-economic and health impacts still being ascertained

Challenges[71]

Considerations

- Data is noisy, time lagged and incomplete
  - E.g. How many individuals are currently infected by Ebola?
- Policy is influenced not just by optimality of solutions but real-world considerations
- Good models are used as a part of evidence based decision making process

Epidemiology and Surveillance Pyramid



[71] Lipsitch et al., 2011, Van Kerkhove & Ferguson 2013, National Pandemic Influenza Plan

Questions our group worked on

1. Estimating basic epidemiological parameters
2. Forecasting the ongoing epidemic with & without control
3. Assessing the threat of imported cases in the US and Latin America causing secondary infections
4. Efficiently allocating potential pharmaceutical treatments
5. Location of Emergency treatment centers and assessing their impact
6. Estimating the need for supplies such as personal protective equipment
7. Analyzing social media for public mood & sentiments

https://www.vbi.vt.edu/ndssl/fea
projects/ebola

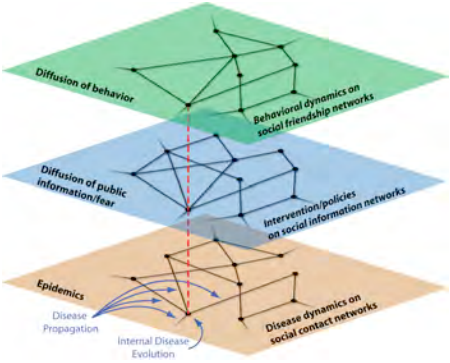
AI in the real world

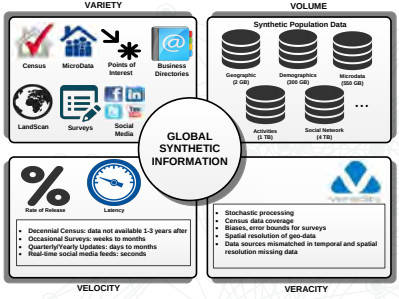
Elements of real-time computational epidemiology

- ***Step 1.*** Construct a synthetic realistic social contact network by integrating a variety of commercial and public sources.
- ***Step 2.*** Develop models of within-host disease progression using detailed case-based data and serological samples to establish disease parameters.
- ***Step 3.*** Develop high-performance computer simulations to study epidemic dynamics (exploring the Markov chain M).
- ***Step 4.*** Develop multitheory behavioral models and policies formulating and evaluating the efficacy of various intervention strategies and methods for situation assessment and epidemic forecasting. Use Markov decision processes to formulate and evaluate these policies.
- ***Step 5.*** Develop Cyber-ecosystems to support epidemiologists and policy makers for effective decision making.

Big data problem

- Synthesis of realistic networks
  - Data is noisy and time-lagged
  - Need new methods for information fusion and ML: Currently using 34 databases
- Large complex networks
  - $>$ 100GB input data: 300M people , 22B edges, 100M locations, 1.5B daily activities
  - Irregular network: Dimension reduction techniques (e.g. renormalization group techniques) do not apply
  - Coevolving behaviors and networks
- Large experimental design $\Rightarrow$ multiple configurations





Step 1: Synthesizing Social Contact networks

Modeling social networks: random graph models

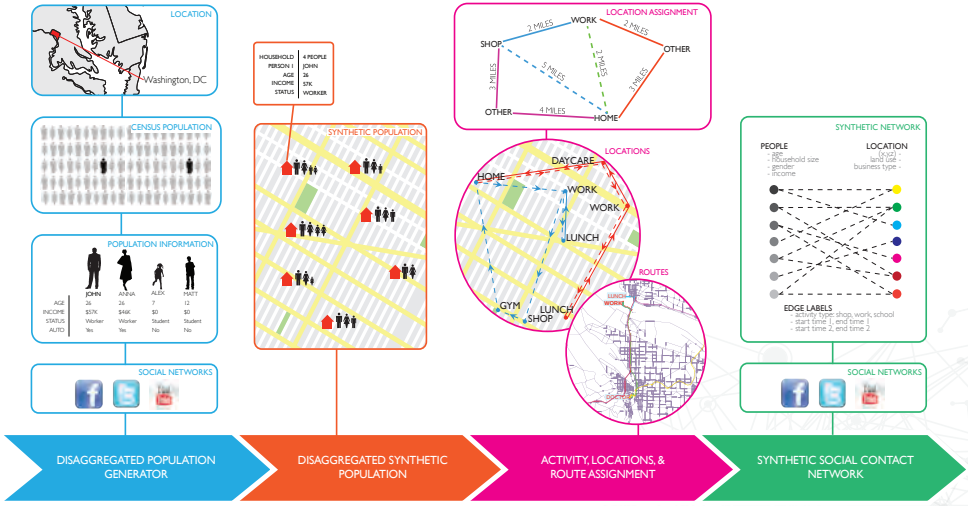
- *Erdős-Rényi* model, $G(n, p)$: Each edge $e = (u, v)$ is selected independently with probability p
- *Chung-Lu model*: given a weight sequence $\mathbf{w} = (w(v_1, V), w(v_2, V), ..., w(v_n, V))$ for nodes $v_i \in V$, a random graph $G(\mathbf{w})$ is constructed as follows:
  - add each edge (v_j, v_k) independently with probability $\frac{w(v_j,V)w(v_k,V)}{\sum_{v_i \in V} w(v_i,V)}$
- Evolutionary models (e.g., preferential attachment): new node v connects to earlier nodes u with probability proportional to $deg(u)$
- Network models capture simple local properties, e.g., degree sequence, clustering coefficient
- Primary goal was to obtain analytical bounds
- Cannot model higher order properties, heterogeneities

Step 1: Synthesizing Social Contact networks

First principles based network synthesis

- For individuals in a population (representation of individuals):
  - Their demographics (Who)
  - The sequences of their activities (What)
  - The times of the activities (When)
  - The places where the activities are performed (Where)
  - The reasons for doing the activities (Why)
- No explicit data sets available for such networks
- Synthesis of a number of public and commercial data sets and expert knowledge
- Can explicitly model the impact of behavioral changes

A methodology for synthesizing social contact networks[72]



[72] Beckman et. al. 1995, Barrett et al. WSC, 2009, Eubank et al. Nature 2005, TRANSIMS project, 1997, 1999

Comparing various data collection techniques to infer mobility patterns[73]

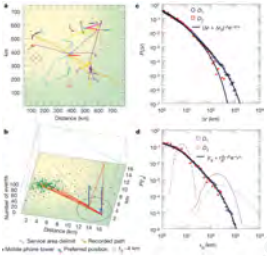
Methods	Advantages	Disadvantages
Survey & direct	Multi purposed use; fewer biases; can capture multiple correlations	can be expensive to collect data observations;
Wi-Fi localization	Accuracy; Energy usage 50% GPS	Providing access point is expensive
GPS localization	High spatial precision: 5m; Can distinguish between transportation modes	High battery (energy) usage; expensive; sampling biases; No (low quality) signal in indoor environment
Cellular network localization (passive) (Call Data Records);	Automatically generated;	Sparse in time; Lower spatial resolution (175m); Needs more filtering; sampling biases; Proprietary
Cellular network localization (active)	More accuracy than passive localization; Less expensive than previous methods	More costly than passive form; sampling biases; Proprietary and thus not publicly available

Characterizing human travel patterns using CDR [75], [76]

100,000 Anonymized mobile phone users tracked for a 6-month period

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-\Delta r/\mathcal{K})$$

- Radius of gyration distribution rules out a traditional Levy flight distribution of step lengths



- Study by Lu et al. [74] highlights that algorithms are capable of approaching the theoretical limits of predictability

[74] X Lu, E Wetter, N Bharti, AJ Tatem, L Bengtsson, *Nature Scientific Reports*, 2013

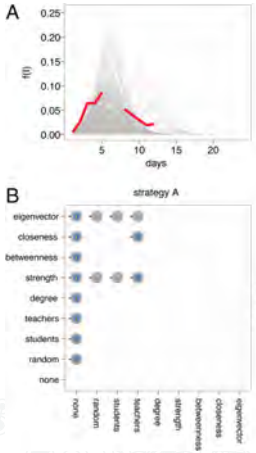
[75] M Gonzalez, CA Hidalgo, A Barabasi, *Nature*, 2008

[76] Becker et al, CACM 2013

Synthesizing social proximity networks using RFID tags and wifi localization [77], [78]

Wireless sensor network motes distributed to students, teachers, staff at an American high school

- Social network reconstructed using 762,868 CPIs (close proximity interactions) at a maximal distance of 3 meters across 788 individuals
- Network exhibits typical small-world properties with high modularity
- SEIR model imposed over the network with 100 runs for each individual (78800 simulations)
- Secondary infections and R_0 in agreement with school absenteeism data during this period

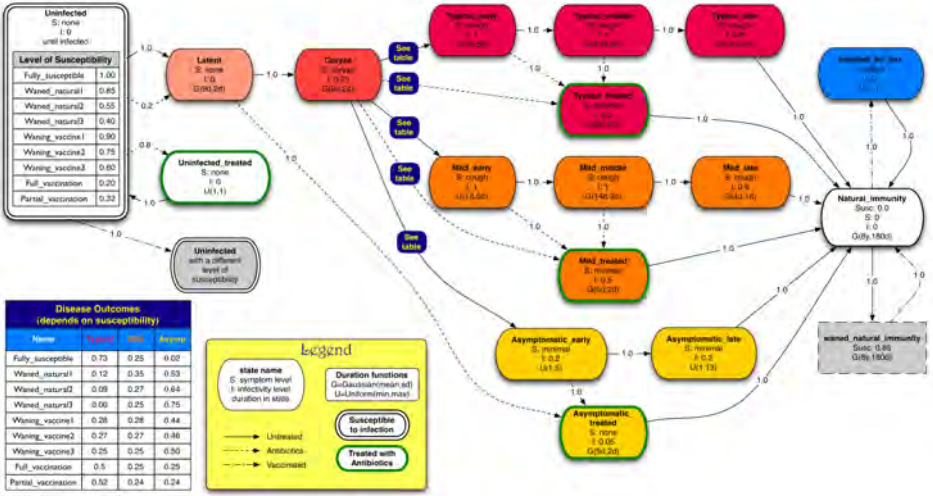


[77] N Eagle, A Pentland, D Lazer, *PNAS*, 2009

[78] M. Salathe et al., *PNAS*, 2010

Step 2: Within host disease progression models.

Within host disease progression model



- Parameter estimation is done using derivative free optimization method.
- Simulation based optimization approach

Step 3: Simulations to unravel the disease dynamics over a network

Computing epidemic dynamics over networks

- Recall: Theoretical results on computing dynamics for special classes of networks
- Worst case complexity: We are given an SIR GDS $\mathcal{S}$, an initial configuration $\mathcal{I}$ and a final configuration $\mathcal{B}$. The goal is to decide whether $\mathcal{S}$ starting from $\mathcal{I}$ reaches $\mathcal{B}$ with a non-zero probability or $\mathcal{B}$ reaches with a probability $\geq \pi$ in $\leq t$ steps
- **Theorem :** For simple SIR GDS systems and for each $t \geq 3$, reachability in t time steps is **NP**-hard. It is #P-hard if we want to assure that $\mathcal{B}$ reaches with a probability $\geq \pi$ Moreover, this result holds even when the initial configuration has one infected node

Implications: need to develop *fast simulations* to compute the epidemic dynamics in general

Fast high performance simulations: From 40 hours to 40 seconds

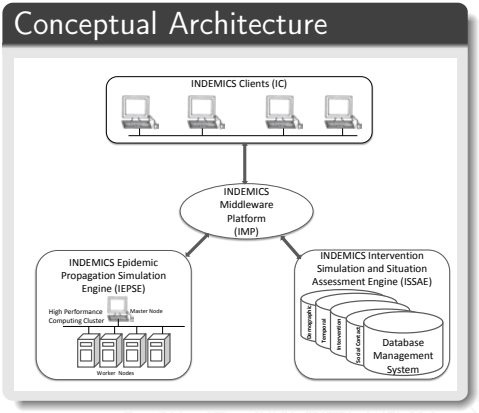
Distinguishing Features	EpiSims (Nature'04)	EpiSimdemics (SC'09,WSC'10)	EpiFast (ICS'09)	Indemics (ICS'10,TOMACS'11)
Solution Method	Discrete Event Simulation	Interaction-Based Simulation	Combinatorial +discrete time	Interaction-based, Interactive Simulations
Performance 180 days 9M hosts & 40 proc.	**~40 hours**	**1 hour for 300Million nodes**	**~40 seconds**	**15min-1hour**
Co-evolving Social Network	Can work	Works Well	Works ***only*** with restricted form	Very general
Disease transmission model	Edge as well as vertex based	Edge as well as vertex based (e.g. threshold functions)	Edge based, independence of infecting events	Edge based
Query and Interventions	restrictive	Scripted, groups allowed but not dynamic	Scripted and specific groups allowed	Very general: no restriction on groups

Simulating coevolving epidemics and interventions

- Computer experiments pertaining to control, resource allocation, planning in epidemiology are best viewed as a Markov decision process (MDP).
- Most simulations discussed in the literature have focused on disease progression; framing of interventions and their execution within a simulation is not well studied
- Epidemic analysis = disease spread + situation assessment + interventions
- Leads to separation of concerns: disease spread is best computed by interaction based methods, interventions are best specified as a (query, action) tuple.
- What is the best way to structure such modeling environments?
  - Need a natural language for representing (query, action) tuple
  - Need the ability for the simulation to start and stop
  - Typically, a public health analyst should be able to formulate a new intervention: they need not have to work with complex parallel code

Indemics

- Simulation can start and stop at any desired point
- Detailed state assessment (e.g. is Tom infected, or how many folks between ages 15-25 are infected)
- Supports (simulation $\rightarrow$ data-analytics $\rightarrow$ simulation) loop
- Interventions and statement assessment questions specified as SQL queries
- New data-centric architecture for interactive epidemic simulation environments
- Decouples data, disease diffusion, intervention and user interaction

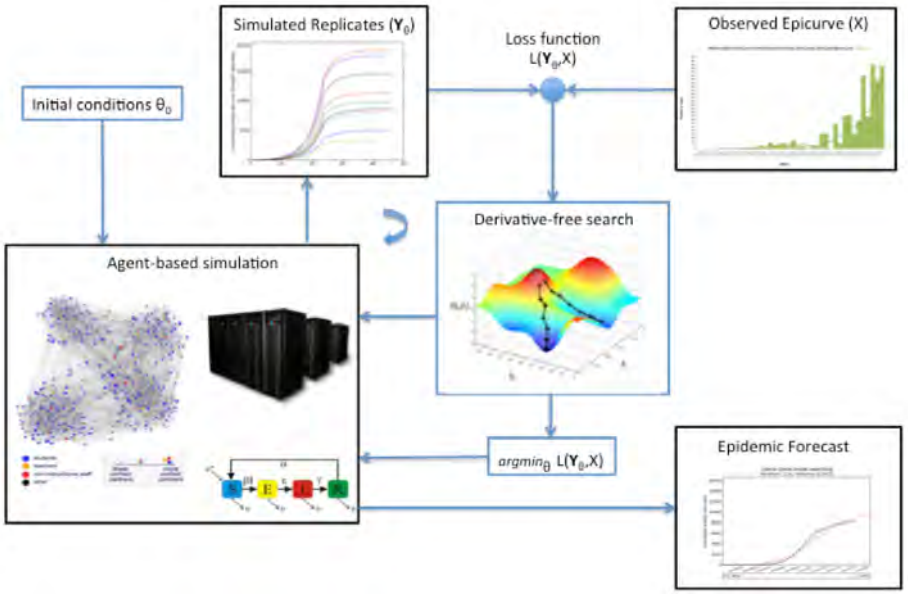


Step 4: Evaluating Policies, Forecasting and Situation Assessment

Calibration and Forecasting

- $\mathcal{X}(1, T) = \langle X(1), \ldots, X(T) \rangle$ denote observations till time T. (e.g. incidence rates for the entire population, or attack rates for different regions or subpopulations.)
- *The calibration problem* find a model θ such that the outcomes Y_θ are "close" to $\mathcal{X}$, using weighted L_1 distance:
  $\min_\theta \frac{1}{R} \sum_{r=1}^{R} \left(\sum_{i=1}^{T} \alpha^{T-i} | Y_{\theta,r}(i) - X(i) | \right)$,
  where $Y_{\theta,r}$ denotes the epicurve for the rth replicate.
- *The forecasting problem:* Given a quantity of interest $Q(\cdot)$ (e.g., number infected), compute the time-series $\mathcal{Q}(1, T + r) = \langle Q(1), \ldots, Q(T + r) \rangle$, where $Q(i)$ denotes the value of the quantity at time i and associated confidence probability $p(i)$ for the value to occur.
- *Metrics:* (i) how far ahead in time (r) it projects (ii) how high the confidence ($p(\cdot)$) is (iii) the quantities $Q(\cdot)$ it can project (e.g., peak, time to peak, total infections, etc).

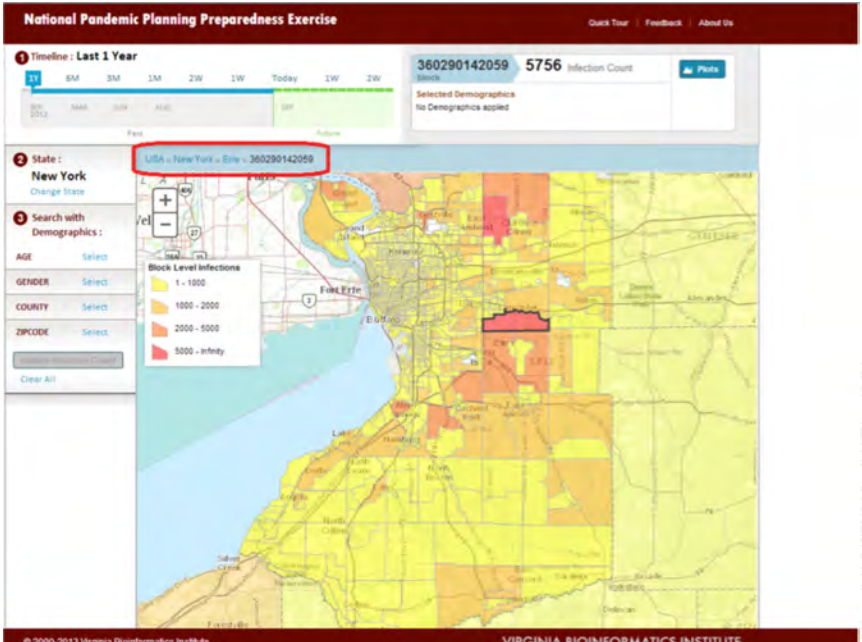
Using derivative free optimization approach



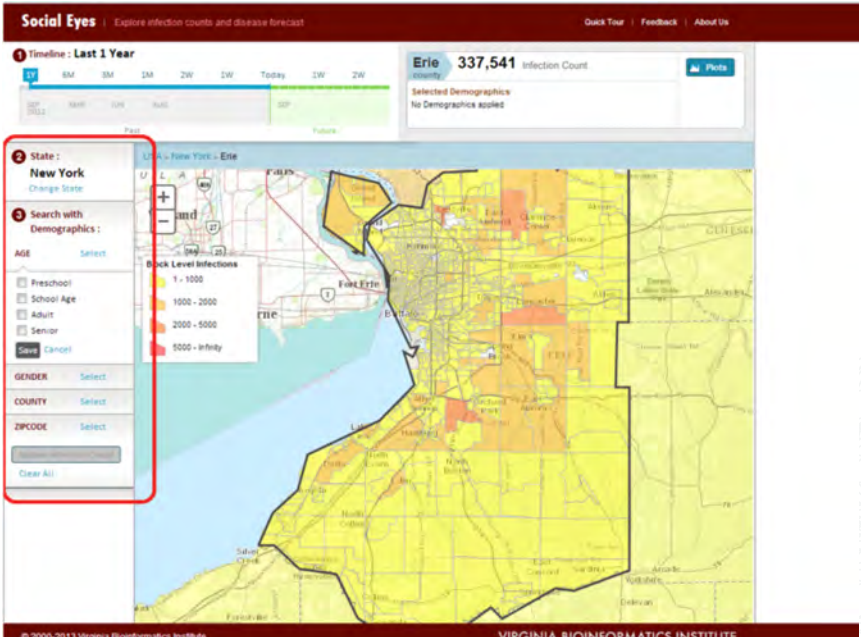
How do network-based causal models helps

- Statistical Models
  - Specific model/predictor for each data source
  - Weights for fusing predictions are learned using cross-validation
  - Low computational complexity, produces good short term forecasts.
  - Easy to extend when new data sources are found (e.g. weather, flu-near you, Athena)
- Causal Models
  - Both ODE (low compute resources) and network-based models (high compute resource) can serve as causal models
  - Network models: Simulation + Blackbox optimization+ Machine learning detailed yields *dendograms*
  - *Universality:* Dendograms + Spatially explicit social network yields forecast for many reasonable quantities of interest

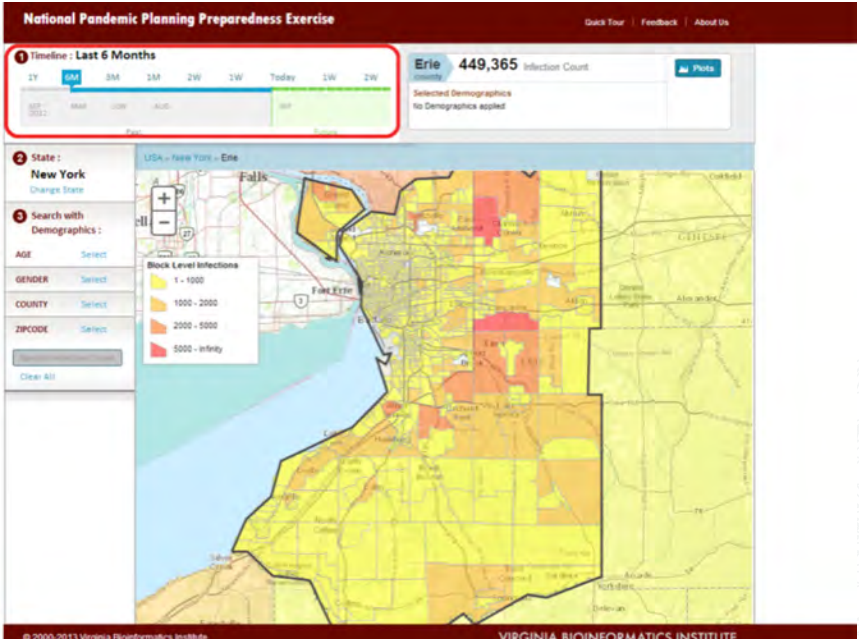
Forecasting using causal models: improved spatial resolution



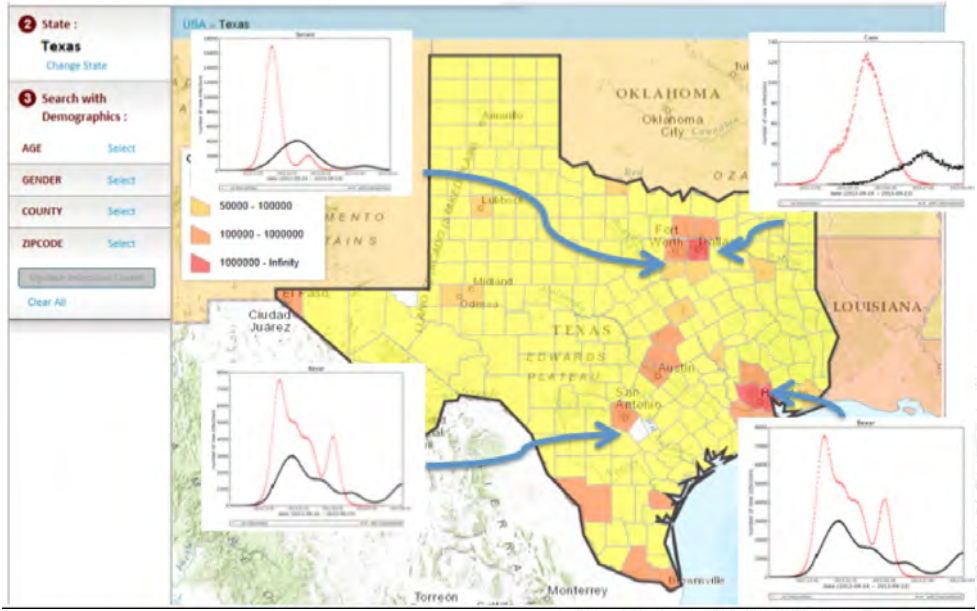
Forecasting using causal models: improved demographic resolution



Forecasting using causal models: improved temporal resolution



Forecasting using causal models: capturing interventions



Outline

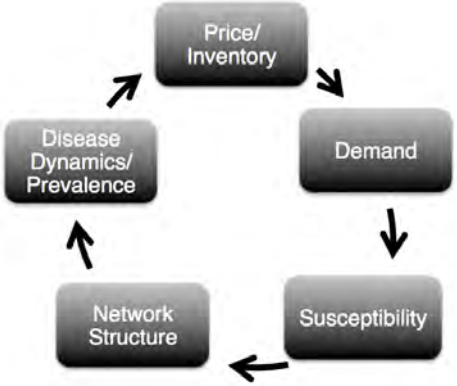
- Mathematical models for epidemic spread
- Intervention design as optimization problems
  - Social objective: designing interventions to minimize outbreak (centralized)
  - Social objective with limited compliance: group level interventions (partially centralized)
  - Individual level objective: game-theoretical interventions (decentralized)
  - Combining individual and social objectives: anti-viral distribution problem

Epidemic Analysis problem as a Markov Decision Process

- Computer experiments pertaining to control, resource allocation, planning in epidemiology are best viewed as a Markov decision process (MDP).

Combining social and individual incentives: anti-viral distribution problem

- Policy Problem: Is there an optimum strategy to partition the scarce AV doses between public stockpile administered through hospitals and private stockpile distributed using a market-mechanism
- Measures of Effectiveness: Number of infected, peak infections, cost of recovery, equitable allocation
- Additional issues: How do disease prevalence, individual behavior, network structure, disease dynamics and AV demand co-evolve?



Models of individual behaviors and adaptation

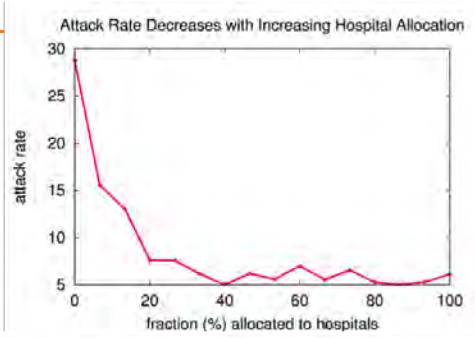
- Isolation based on Prevalence (fear contagion)
  - Entire household isolated when perceived prevalence $>$ threshold
  - Compliance rate: 40%
- Economic Behavior: Demand elasticity based on Prevalence
  - Household demand: $D_{t,h} = \frac{B_{t,h}}{P_t}(1 - e^{-\beta x_t})$
  - Increases with disease prevalence x_t
  - Increases with household budget $B_{t,h}$, decreases with price P_t, and price is linear in remaining supply
  - β reflects risk aversion or prevalence elastic demand to AV.
- Disease Reporting and treatment
  - Anti-virals are administered to individuals who are symptomatic, report clinic and are correctly diagnosed.

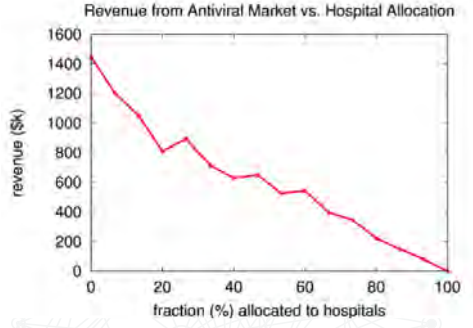
Organizational Behavioral models

- Hospitals
  - Total AV supply is 15K: allocated between hospitals and market
  - Hospitals: give to diagnosed as infected
- Markets
  - Market: sells to households according to demand and price
  - Markets provide A/Vs on a first come first serve basis (are not spatially sensitive in this version)
  - Assume a centralized market. Linear price curves (as a function of remaining A/V stock)

Results (I): both Private and Public Distribution are important

- Suggests optimal allocation strategy of AVs between public and private stockpile
  - Hospitals (public sector) should be given priority
  - If $>$ threshold, the remaining stockpile be distributed via market.
  - Private stockpile useful for individuals who are infectious but not symptomatic
- Optimal split (40% to hospitals, 60% to the market) recovers the cost of antiviral manufacturing if the unit cost is below a bound.

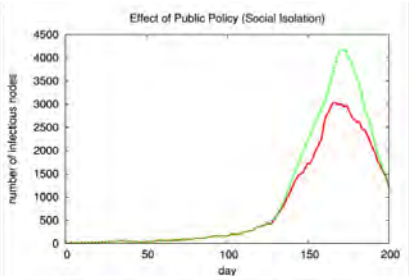


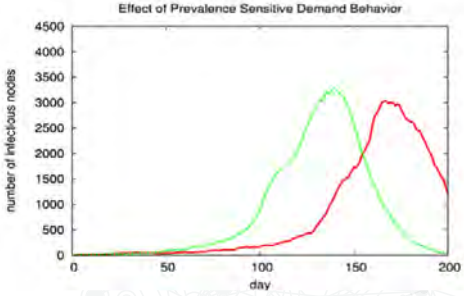


Results (II): Role of Behavioral Adaptations

- Both behavioral adaptation were critical in controlling the epidemic
  - Household isolation reduces the peak infection rate by 30%.
  - Prevalence based demand delays the peak infection rate by 30 days.

Natural behavior adaptation to an epidemic in conjunction with well established logistics (markets + public distribution) reduce and delay the peak infection rate





Step 5: Developing Cyber-ecosystems to support decision making

Cyber-ecosystems: Examples

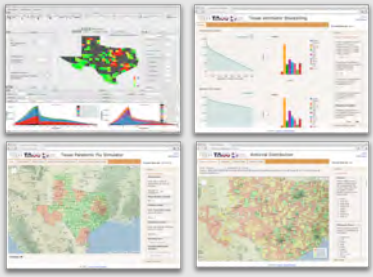
BSVE by DTRA CB

The Biosurveillance Ecosystem (BSVE) at DTRA: a cloud-based, social, self-sustaining web environment to enable real-time biosurveillance

BARD by LANL

Tools to (*i*) validate/confirm disease surveillance information (*ii*) rapidly select appropriate epidemiological models for infectious disease prediction, forecasting and monitoring; (*iii*) provide context and a frame of reference for disease surveillance information

Texas Pandemic tool kit

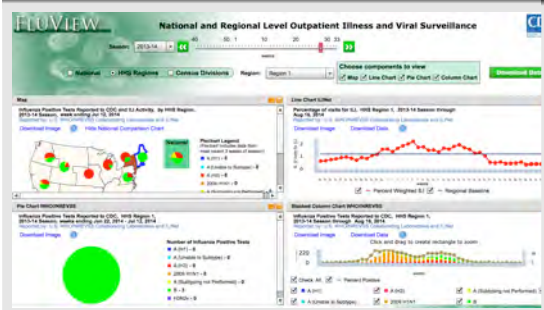


Tools for (*i*) anti-viral scheduling & distribution; (*ii*) ventilator stockpiling; (*iii*) vaccine allocation; (*iv*) pandemic exercise tool; (*v*) flu simulator; (*vii*) sample size calculators for public health labs.

`http://flu.tacc.utexas.edu/`

Cyber-ecosystems: Examples

CDC FluView



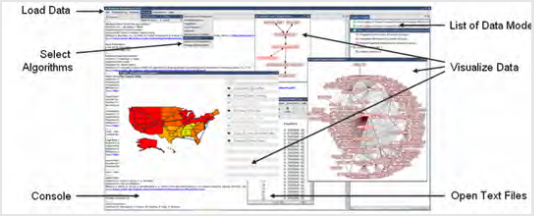
http://www.cdc.gov/flu/weekly/

HealthMap

http://healthmap.org/en/

EpiC: A Computational Infrastructure for Epidemics (MOBS Lab, Northeastern U.)

EpiC



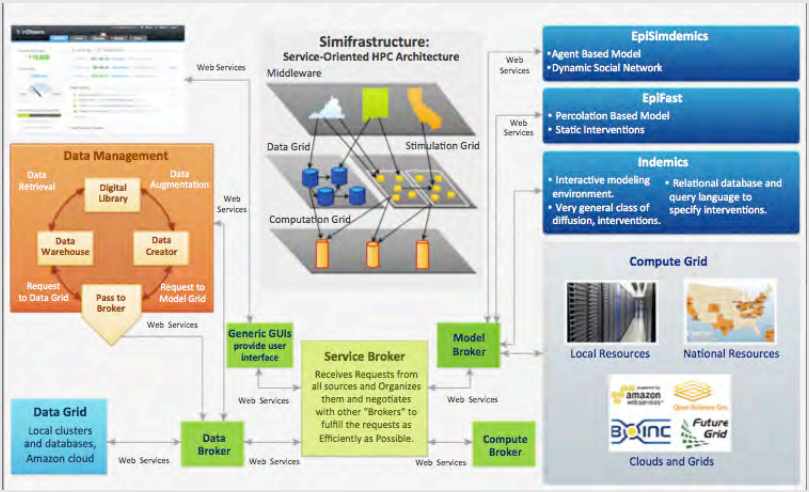
http://www.mobs-lab.org/-epic-a-computational-infrastructur html

Gleam

www.gleamviz.org

CIEPI: Cyber-infrastructure for computational epidemiology, NDSSL, BI, Virginia Tech

CIEPI Cyber-infrastructure for computational epidemiology by NDSSL



Provides seamless access to high performance computing models, libraries and data

CINET: A cyberinfrastructure for Network Science

- An open access cyberinfrastructure.
  - A web portal that hides the details of computation and data management, thereby minimizing the learning effort required.
- A flexible framework
  - Allows easy extension by integrating off-the-shelf network analysis suites for analysis and visualization; this means new algorithms can be added easily over time.
- A common repository
  - Managing data, models, and results through a digital library that maintains metadata.
- Fostering research, teaching and collaboration
  - Building a broad user base, from multiple disciplines, including incorporation into courses on network science at many different universities.

CINET components

- *Granite:* Structural analysis tool
  - 110+ networks; 18+ network generators.
  - 70+ network algorithms (measures).
    - Adapted from 3 graph libraries: GaLib, SNAP, NetworkX.
  - Visualization of networks (uses Gephi toolkit).
  - Services: Adding new networks, measures
  - *NetScript:* Python-based domain specific language for supporting complex workflows.
- *Edison:* Simulators for broad class of contagion dynamics
  - 20+ networks (graphs).
  - 4 model families.
  - Query system to identify subsets of vertices and edges.
  - Generalized contagion simulator.
- *GDS Calc:* Graph Dynamical System calculator
  - Analyzing the phasic structure of a graph dynamical system (can only deal with smaller networks but complete phase structure).

NDSSL's epidemiological application

ecology:https://www.bi.vt.edu/ndssl/tools

DSI:DC

DSI:DC is an educational adventure game where players take on the role of a public health agent asked to solve a growing crisis resulting from a disease outbreak. The game focuses on the science and mathematics of infectious disease, the application of computational modeling for decision-making, and the civic and health implications of these actions.

Dynamic Behavior Visualizer

Dynamic Behavior Visualizer (DBV) is an interactive visualization of people, a group of people, or a family used to help understand behaviors and movements over time during a natural or man-made disaster. DBV is used to study the resilience of critical infrastructures such as transportation, communication, and public health.

EDISON

Dynamics on networked populations are useful in understanding social processes on networked populations. Contagion dynamics can take various forms such as epidemics, social protest, smoking and drug use, use of social media, etc. EDISON utilizes big data and data mining to perform social dynamics on networks.

NDSSL's epidemiological application

ecology: https://www.bi.vt.edu/ndssl/tools

EpiCaster

Users can view Ebola (or Flu) activity for the past four weeks and view forecast predictions for the next two weeks. They can also view forecast trends and compare them to surveillance data. EpiCaster allows users to see what impact various strategies, such as vaccines and social distancing, have on disease spread.

EpiViewer

EpiViewer is a data repository for epidemiologists. Users can upload and compare Ebola forecasts and surveillance data from a variety of sources and see how forecasts change over time. Users can also load and share their own forecasting predictions.

EpiViz

EpiViz is a highly dynamic system that provides a platform to track and study subjects' decision making and information search strategies, under controlled and repeatable conditions using simulated disease outbreaks. Data visualization supports a multiple views environment and parallel simulation runs.

NDSSL's epidemiological application

ecology: https://www.bi.vt.edu/ndssl/tools

Eyes on the Ground

Road conditions can be variable in some of the rural areas in Western Africa. Eyes on the Ground allows people in affected areas to report their road conditions. Other travelers can then view these reports and plan their trips accordingly. This is especially useful when planning the delivery of patients and supplies between cities.

FluCaster

FluCaster is a disease surveillance and situation assessment tool that uses social media crowd sourcing and complex mathematical models to track and predict the spread of communicable diseases. It enables users to see how many people in their area are infected with the flu, displaying area influenza activity similar to a weather forecast.

GDS Calculator

Agent-based simulations are used to understand disease transmission, the spread of social unrest, and the propagation a host of other contagions such as fads, rumors, and influence. Contagions may be spread, for example, by face-to-face interaction and/or electronic means (e.g., social media). Simulation is an effective way to study these dynamics of contagion spread.

NDSSL's epidemiological application

ecology:https://www.bi.vt.edu/ndssl/tools

Granite

Networks are an effective abstraction for representing real systems. Consequently, network science is increasingly used in academia and industry to solve problems in many fields. Computations that determine structure properties and dynamical behaviors of networks are useful because they give insights into the characteristics of real systems.

my4Sight

my4Sight uses human computation to enhance disease forecasting. Similar to games like Foldit, this web application allows users to assist computational models by performing tasks that humans are uniquely good at, in this case pattern matching.

PATRIC

PATRIC is the Bacterial Bioinformatics Resource Center, an information system designed to support the biomedical research community's work on bacterial infectious diseases via integration of vital pathogen information with rich data and analysis tools.

NDSSL's epidemiological application

ecology:https://www.bi.vt.edu/ndssl/tools

SIBEL

SIBEL allows bioinformatics researchers to design experiments and create analysis for epidemiological disease studies based on realistic social network simulations. It enables improved readiness, planning, and decision making in the domains of public safety and national security by delivering sophisticated modeling and simulation capabilities directly into the hands of the analyst.

SIV

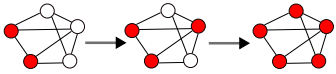
Synthetic Information Viewer (SIV) is a synthetic population visualization tool that allows users to explore demographic information at several levels of resolution - from national to individual. It supports our 2009 version of U.S. data and all international countries NDSSL has constructed, total population of 800+ million.

TranSims

The TRansportation ANalysis SIMulation System is a transportation planning and decision support tool capable of simulating second-by-second movements of every person and every vehicle through the transportation network of a large urban environment. TRANSIMS was the first successful example of high performance computational social sciences and policy informatics.

Three Extensions

Generalized contagions as models of influence and (mis)information


Example: threshold-2 model

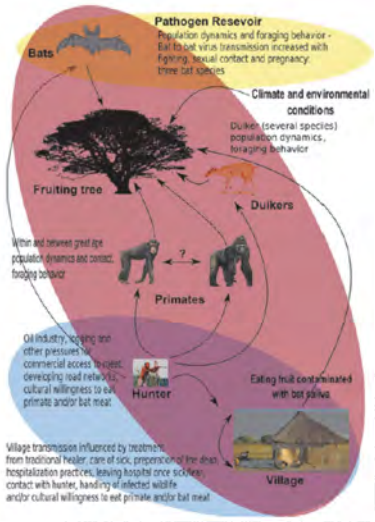
Model	Description	Example Applications
Percolation & extensions: SI/SIS/SIR/Independent cascades	Each red node infects each neighbor independently with some probability	Malware, failures, infections
Complex contagion: threshold and variants	Each node switches to red if at least k neighbors are red	Spread of innovations, peer pressure
Non-monotone multi-threshold models	Thresholds for switching to red and from red to uncolored	More complex social behavior
Voter models	Each node picks the state of a random neighbor	Spread of ideologies

Generalized contagions

- GDS and its generalizations are well suited to capture generalized contagion processes.
- *Example:* Social Contagions
  - *Local Mechanism:* Thresholds, Voter, Linear threshold, Independent cascade, Purely stochastic, Generalized contagion, Cooperative action, Learning, Multi-contagion
  - *Mechanism for social interaction:* Individual (local) interactions (e.g., face-to-face, phone, skype), Joint (group) behaviors (e.g., cadre, team, school, club), Global interactions (e.g., use of social media), & Regional interactions (e.g., TV, news, newspapers, crowds)

Extension 1: Zoonoses and emerging diseases[79]

- *Zoonosis:* Disease that is naturally transmissible from vertebrate animals to humans and vice-versa. Includes all types of pathogenic agents, including bacteria, parasites, fungi, and viruses as causative agents (e.g., Ebola)
- *Spillover:* Process by which a zoonotic pathogen moves from an animal host to a human host.
- Two different R_0 values: capturing human-human and human-animal transmission. Intervention strategies are quite different in the two cases.
- Pharmaceutical interventions are unavailable.



[79] Alexander et al., Vectore Borne & zoonotic diseases, 2012

Extension 2: Non-communicable Diseases and social contagions[80]

Obesity is a growing epidemic

- Obesity, Smoking, Memes, product adoption, social unrest: e.g. of epidemic "like" processes.
- Social media important in the recent uprisings, e.g., Arab Spring, Occupy Wall Street. One Egyptian said, *"facebook used to set the date, twitter used to share logistics, youtube to show the world, all to connect people."* (Gonzalez-Bailon et al. 2011)
- **Intersim:** High performance computing modeling environment to simulate general contagion processes.

Social contagions

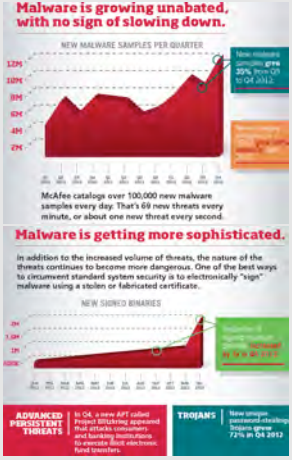
[80] Kuhlman et al. WSC 2012, AAMAS 2014

Extension 3: Malware propagation and Internet epidemiology

Malware as generalized contagion

- Amplified as Internet of Things takes hold; the malware ecosystem is becoming rich and diverse.
- **EpiCure:** High performance scalable and expressive modeling environment to study mobile malware in large dynamic networks (Channakesava, et al. IPDPS, 2012)
  - 3.5 million mobile devices in a city as large as Miami can be simulated in 1.5 hours.
  - Model approximations used for bluetooth protocol

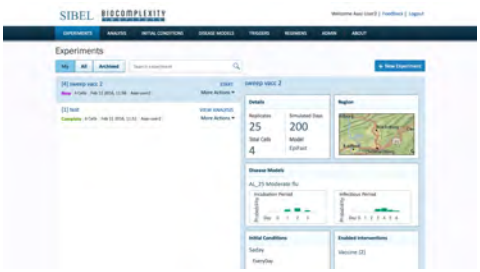
Growth of Malware



http://www.mcafee.com/us/resources/misc/infographic-state-of-malware.pdf

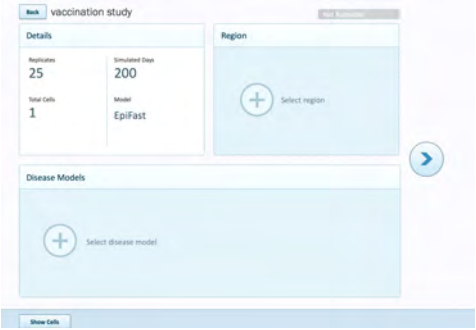
Exercises on SIBEL

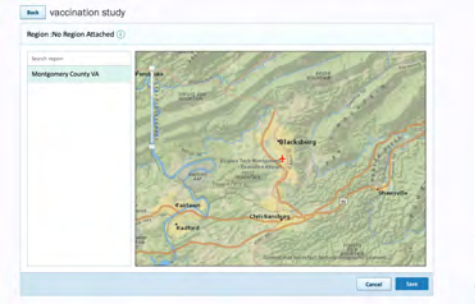
- `link`: http://sibel.vbi.vt.edu:8080/demo/sibel.html#login
- `userid`: aaai-user1, ..., aaai-user12
- `Password`: P@ssw0rd



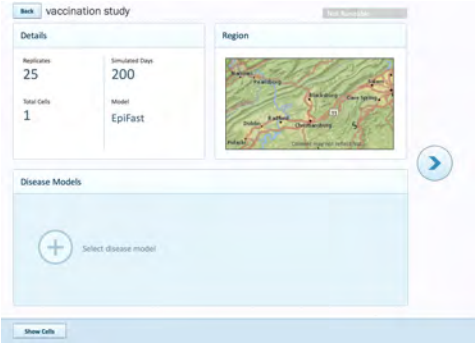
Set up experiment

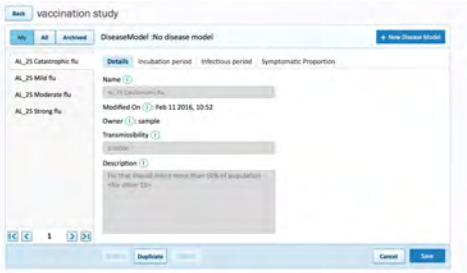
Select region



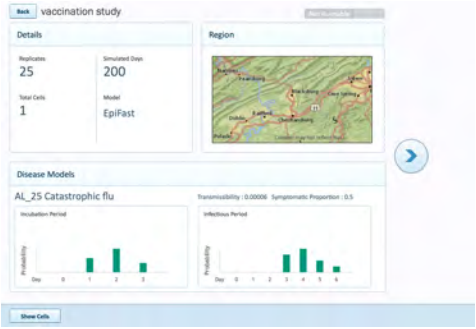


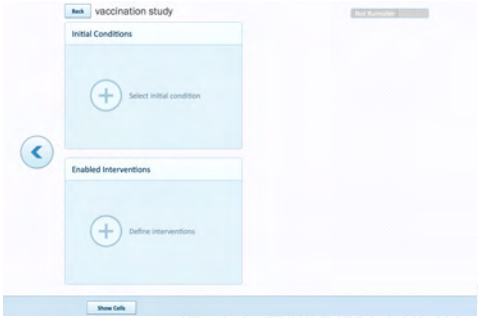
Select disease model



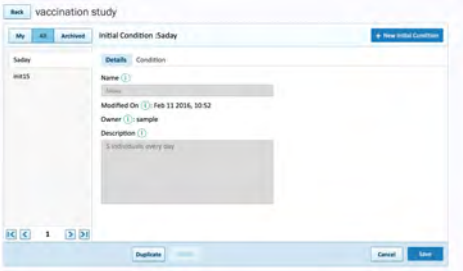


Specify initial conditions and interventions

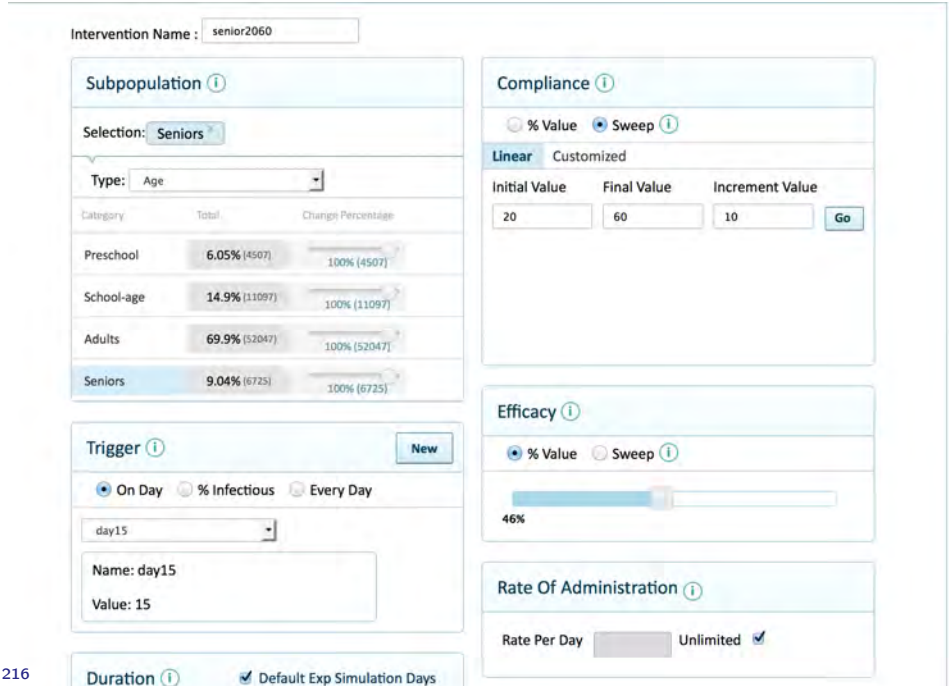




Specify intervention



Specify intervention



Summary and concluding remarks

Summary and conclusions

- *Controlling and responding to future pandemics is a hard problem*; emerging global trends make it challenging
  - (i) increased and denser urbanization, (ii) increased local & global travel, (iii) older and immuno-compromised population.
- *Public health epidemiology is a complex system problem*. Epidemics, social-contact networks, individual and collective behavior, and public policies ***coevolve*** during a pandemic — a system-level understanding must represent these components and their coevolution.
- Computational Epidemiology: fascinating field at the intersection of many disciplines. Excellent example of computing for social good.
  - GDS as a unifying framework
  - *Mathematical and computational models and methods are critical in public health epidemiology*.
  - *Advances in computing, big data, and computational thinking* have created entirely new *opportunities to support real-time epidemiology* – a move towards ***pervasive computational epidemiology***

Important topics not covered in the tutorial

- Important topics we did not cover
  - Game theory, economics of pandemics, behavioral modeling
  - Uncertainty quantification, Validation and Verification
  - Prediction Markets
- Directions for future research
  - *Ecological Epidemiology:* One-health: understanding epidemiology in a broader context of health and well being across human, animals and plant species: combining ecology and epidemiology
  - *Immunology+Epidemiology:* Current models of disease manifestation are statistical in nature. Use immunological modeling to understand disease progression. Will help understand the role of therapeutics and novel interventions
  - *Phylogeography:* Combining phylogenetics and epidemiology to understand the drift and shift of viruses and their relationship to geography (e.g. Flu, HIV).

Acknowledgments: Thanks to members of the Network Dynamics and Simulation Science Laboratory, VBI and Discovery Analytics Center (DAC), both at Virginia Tech and our collaborators.

Support

- National Science Foundation: HSD grant SES-0729441, NSF PetaApps grant OCI-0904844, NSF NetSE grant CNS-1011769, NSF SDCI grant OCI-1032677,
- Defense Threat Reduction Agency grant HDTRA1-11-1-0016, DTRA CNIMS contract HDTRA1-11-D-0016-0001,
- National Institute of Health Midas grant 2U01GM070694-09,
- Intelligence Advanced Research Projects Activity (IARPA) via the US Department of Interior (DoI) National Business Center (NBC): D12PC000337.

The US government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US government.

Course notes, data and some of the tools are available on the web:
`ndssl.vbi.vt.edu/apps`, `http://ndssl.vbi.vt.edu/synthetic-data/`

Comments and questions are welcome

Contacts:

Prithwish Chakraborty (prithwi@cs.vt.edu)

Madhav V. Marathe (mmarathe@vbi.vt.edu)

Naren Ramakrishnan (naren@cs.vt.edu)

Anil Vullikanti (akumar@vbi.vt.edu)

Additional References: overview

- R.M. Anderson and R.M. May. *Infectious Diseases of Humans*. Oxford University Press, Oxford, 1991.
- N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Hafner Press, New York, 1975.
- F. Brauer, P. van den Driessche, and J. Wu, editors. *Mathematical Epidemiology*. Springer Verlag, Lecture Notes in Mathematics 1945.
- M. Gersovitz and J. S. Hammer. Infectious diseases, public policy, and the marriage of economics and epidemiology. *The World Bank Research Observer*, 18(2):129–157, 2003.
- M. Newman. The structure and function of complex networks. *SIAM Review*, 45, 2003.
- Marcel Salathé et al. Digital epidemiology. *PLoS Computational Biology*, 8(7), 2012.

Additional references: Forecasting and Surveillance

- Sean Brennan, Adam Sadilek, and Henry Kautz. Towards understanding global spread of disease from everyday interpersonal interactions. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2783–2789. AAAI Press, 2013.
- David A Broniatowski, Michael J Paul, and Mark Dredze. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672, 2013.
- Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O Nsoesie, Sumiko R Mekaru, John S Brownstein, Madhav Marathe, et al. Forecasting a moving target: Ensemble models for ili case count predictions. In *2014 SIAM International Conferenc on Data Mining*, SDM'14, page 9. SIAM, 2014.
- Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118, 2010.

Additional references: Forecasting and Surveillance

- Jean-Paul Chretien, Dylan George, and Ellis McKenzie. A systematic review of influenza forecasting studies. *Online Journal of Public Health Informatics*, 6(1), 2014.
- Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.
- Andrea Freyer Dugas, Yu-Hsiang Hsieh, Scott R Levin, Jesse M Pines, Darren P Mareiniss, Amir Mohareb, Charlotte A Gaydos, Trish M Perl, and Richard E Rothman. Google flu trends: correlation with emergency department influenza rates and crowding metrics. *Clinical infectious diseases*, 54(4):463–469, 2012.
- Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

Additional references: Forecasting and Surveillance

- Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.
- Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y Del Valle, and Reid Priedhorsky. Detecting epidemics using wikipedia article views: A demonstration of feasibility with language as location proxy. *arXiv preprint arXiv:1405.3612*, 2014.
- Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
- Bruno Gonçalves, Duygu Balcan, and Alessandro Vespignani. Human mobility and the worldwide impact of intentional localized highly pathogenic virus release. *Scientific reports*, 3, 2013.

Additional references: Forecasting and Surveillance

- Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- Kirsty Hope, David N Durrheim, Edouard Tursan d'Espaignet, and Craig Dalton. Syndromic surveillance: is it a useful tool for local outbreak detection? *Journal of Epidemiology and Community Health*, 60(5):374–374, 2006.
- Lars Hufnagel, Dirk Brockmann, and Theo Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42):15124–15129, 2004.
- Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 239–252. ACM, 2012.

Additional references: Forecasting and Surveillance

- Alex Lamb, Michael J Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795, 2013.
- Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72, 2012.
- D Lazer, R Kennedy, G King, and A Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203, 2014.
- David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. Google flu trends still appears sick: An evaluation of the 2013-2014 flu season. 2014.
- Elaine O Nsoesie, David L Buckeridge, and John S Brownstein. Who's not coming to dinner? evaluating trends in online restaurant reservations for outbreak surveillance. *Online Journal of Public Health Informatics*, 5(1), 2013.
- Elaine O Nsoesie, Scotland C Leman, and Madhav V Marathe. A dirichlet process model for classifying and forecasting epidemic curves. *BMC infectious diseases*, 14(1):12, 2014.

Additional references: Forecasting and Surveillance

- Donald R Olson, Kevin J Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS computational biology*, 9(10):e1003256, 2013.
- Justin R Ortiz, Hong Zhou, David K Shay, Kathleen M Neuzil, Ashley L Fowlkes, and Christopher H Goss. Monitoring influenza activity in the united states: a comparison of traditional surveillance systems with google flu trends. *PloS one*, 6(4):e18687, 2011.
- Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.
- Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 2010.

Additional references: Forecasting and Surveillance

- Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.
- Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature communications*, 4, 2013.
- Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.