Mathematical and computational foundations of infectious disease epidemiology

Bryan Lewis, Madhav V. Marathe and Anil Kumar Vullikanti
Network Dynamics and Simulation Science Laboratory
Biocomplexity Institute & Dept. of Computer Science Virginia Tech

August 7, 2017
Outline

1. Goals, History, Basic Concepts
2. Dynamics and Analysis
3. Surveillance and Forecasting
4. Control and optimization
5. Putting it all together: theory to practice
Goals

- **Overview and state of the art** – emphasis on formal models, computing, techniques in algorithms, artificial intelligence, machine Learning and data science

- **Describe open problems and future directions** – aim to attract researchers working in mathematical biology and network science 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.

- **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.

---

1Last JM, ed. Dictionary of Epidemiology.
Precursors to modern computational epidemiology

1796
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.

1854
CHOLERA // Bacteria
John Snow was the first to link the London cholera epidemic to a particular water source.
Disease status today: endemic; occurring in poverty-stricken countries.

1897
MALARIA // Parasite
Ronald Ross and George Macdonald developed a mathematical model of mosquito-borne pathogen transmission.
Anderson McKendrick studied with Ross on antimalarial operations, pioneering many discoveries in stochastic processes.
Disease status today: controlled in US; still prevalent in Africa, India.

1946
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.

1981
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.

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.
Role of computing, statistics and mathematics 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 & at-scale multi-agent networks
  - Graphical dynamical systems, e.g. deterministic and probabilistic Boolean networks
  - Markov chain methods
  - Intelligent steering of computational experiments
- Combinatorial techniques for optimization and control
  - Algorithmic game theory
  - Markov decision process
  - Mathematical programming
  - Heuristic search
- Compartmental models
- Networked Epidemiology
  - Branching process
  - Spectral radius characterization
- Extensions: threshold models, voter models
- Competing contagions
- Unifying framework: graph dynamical systems
Classifying formal models

Mathematical Models for Epidemiology

Differential Equation Based
  [Hethcote: SIAM Review]
  - ODE's (Bernoulli, Ross, McDonald, Kermack, McKendrick)
  - Stochastic ODE's (Bartlett, Bailey, Brauer, Castillo-Chavez)

Spatially explicit
  - Patch-based
  - Cellular automata

Network-Based Modeling
  [Keeling et al.]
  - Template-based
  - Realistic Social Net. (NDSSL, Salathe,)

ODE's (Bernoulli, Ross, McDonald, Kermack, McKendrick)
Stochastic ODE's (Bartlett, Bailey, Brauer, Castillo-Chavez)
Basic Terminology

- **Susceptibility** – level relative to others of an individual’s ability to become infected
- **Exposure** – process that allows infection to pass from infectious to susceptible
- **Infection** – process when pathogen is successfully transmitted to susceptible
- **Recovered** – state following infection
- **Incubation period** – duration between infection and symptoms/infectiousness
- **Infectious period** – duration an individual can transmit disease
- **Transmissibility** – degree to which an infected individual can transmit pathogen to susceptible individuals
- **Symptoms** – physical results of infection, often used in diagnosis of infection
Mass action compartmental Models

\[
\begin{align*}
\frac{ds}{dt} &= -\beta is \\
\frac{di}{dt} &= \beta is - \gamma i \\
\frac{dr}{dt} &= \gamma i
\end{align*}
\]

Assumption: complete mixing among population of size \( N \)

- **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.

Illustrating the problem using compartmental models: Impact of interventions

Many ways one can reduce transmission by 50%

Original group of people in contact

Split into 2 equal groups

Split into a few unequal groups

Reduce likelihood of transmission
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
Networks capture heterogeneity

Network for one small family

Challenging to visualize dense networks

Additional contacts for these individuals
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 $S$:
  - 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 $S$

- 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 $S$, phase space $\mathcal{P}(S)$, noisy observation $\mathcal{O}$

<table>
<thead>
<tr>
<th>Analysis Problems</th>
<th>Optimization Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does $\mathcal{P}(S)$ have a fixed point, GE configuration, transient of length $\geq k$?</td>
<td>Remove/Modify $\leq K$ nodes/edges in $G$ so as to infect minimum number of nodes.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference Problem</th>
<th>Forecasting and Situational Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find the most likely: (i) initial configuration, (ii) the transmission tree, (iii) underlying network or (iv) disease parameters</td>
<td>Assess total number of nodes in a particular state, Forecast total number of nodes (probabilistically) in a particular state after time $t$</td>
</tr>
</tbody>
</table>
## Mapping epidemic problems onto GDS problems

<table>
<thead>
<tr>
<th>Quantity/Problem in epidemiology</th>
<th>GDS analogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epicurve</td>
<td>Analysis (e.g., #1’s in configuration) of phase space trajectory</td>
</tr>
<tr>
<td>Computing epidemic characteristics</td>
<td>Analysis problem: Reachability problem in GDS</td>
</tr>
<tr>
<td>Inferring index case, given information about graph and observed infections</td>
<td>Predecessor inference problems in GDS</td>
</tr>
<tr>
<td>Inferring disease model, given the graph and observed infections</td>
<td>Local function inference problem</td>
</tr>
</tbody>
</table>
Networks capture heterogeneity
Unmitigated disease spread

Initial infection at Daycare

- Uninfected
- Infected
- Recovered
Intervention: Social Distancing

Compliant individuals don’t engage in other activities (non-home, work, school)
Intervention: Close Work

Compliant individuals don’t go to work
Intervention: Close Schools

Compliant individuals don’t go to school
Intervention: Pharmaceuticals

- **Efficacy** – Reduction in transmission from pharmaceutical
- **Regimen duration** – Duration a pharmaceutical is active in an individual
- **Supply** – Pharmaceuticals are finite and logistical constraints can limit availability
Intervention: Vaccination

Vaccine administered: Delay till full efficacy, provides effect throughout simulation following administration
Intervention: Treatment

Administering medicine upon diagnosis Shortens symptomatic periods, limits transmissibility
Intervention: Prophylaxis

Providing medicine to those at risk of infection regardless of diagnosis – Only effective will medicine is taken.
Intervention: Treatment

Each interaction has two probabilities: Likelihood of infection (susceptibility) & Likelihood of transmission (transmissibility). Different pharmaceuticals have different effects on these parameters.
Intervention: Sequestration

Members of sub-population remove all links except with each other
1 Goals, History, Basic Concepts
2 Dynamics and Analysis
3 Surveillance and Forecasting
4 Control and optimization
5 Putting it all together: theory to practice
Dynamics in Compartmental Models

\[
\begin{align*}
\frac{ds}{dt} &= -\beta is \\
\frac{di}{dt} &= \beta is - \gamma i \\
\frac{dr}{dt} &= \gamma i
\end{align*}
\]

- \( \frac{di}{dt} > 0 \) (leads to a large epidemic) if \( \frac{\beta s}{\gamma} > 1 \)
- At the start of epidemic: \( s \approx 1 \)
- \( R_0 = \frac{\beta}{\gamma} \): 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
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 at least $n$ waves (level of tree), in other words, that at least one individual in the $n^{th}$ level of the tree becomes infected.

$q^* = \lim_{n \to \infty} q_n$

**Theorem**

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

---

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$^4$: $p_t = 1/(n - 1)$
- Lattice $\mathbb{Z}^d$: $p_t \rightarrow 1/(2d)$, as $d \rightarrow \infty$
- Random $d$-regular graphs: $p_t = 1/d$
- Not well understood in general graphs
  - Partial characterization in finite regular expander graphs with high girth$^5$
  - Characterization in terms of the second moment$^6$

---

$^4$Erdős and Rényi, 1959
$^5$Alon, Benjamini and Stacey, 2001
$^6$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 $\beta$
- Each infected node becomes susceptible with rate $\delta$
- $\rho(A)$: 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 $\beta$
- Each infected node becomes susceptible with rate $\delta$
- $\rho(A)$: 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
  $\rho(A) = \max_x x^T A x / x^x$
- 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

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

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

Similar implications but different assumptions, extended to SEIR models

---

7 BA Prakash, D Chakrabarti, M Faloutsos, N Valler, C Faloutsos. *Knowledge and Information Systems*, 2012
9 A. Ganesh, L. Massoulié and D. Towsley, *IEEE INFOCOM*, 2005
Formally

**Lemma (Sufficient condition for fast recovery)**

Suppose $\rho(A) < T$. Then, the time to extinction $\tau$ 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\textsuperscript{10} \textsuperscript{11}

\textsuperscript{11}BA Prakash, D Chakrabarti, M Faloutsos, N Valler, C Faloutsos. \textit{Knowledge and Information Systems}, 2012
Competing contagions in the SIS model: the $Sl_1I_2S$ model

- $G = (V, E)$: undirected contact graph
- State transition for node $u$ from $S$ to $I_j$ at rate $\beta_j$, $j = 1, 2$, depending on which infected neighbor of $u$ is successful in infecting it
- Nodes switch back to susceptible state at rate $\delta_j$ from $I_j$ to $S$
- What is the limiting distribution?
Steady state distribution in $SI_1I_2S$ model: “winner takes all”

**Theorem**

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

\[ a \text{B. Aditya Prakash, A. Beutel, R. Rosenfeld, C. Faloutsos, } WWW, \text{ 2012} \]

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

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

Both above threshold: $\lambda_1 \frac{\beta_1}{\delta_1} > 1$, $\lambda_1 \frac{\beta_2}{\delta_2} > 1$, $\frac{\beta_1}{\delta_1} > \frac{\beta_2}{\delta_2}$
1 Goals, History, Basic Concepts
2 Dynamics and Analysis
3 Surveillance and Forecasting
4 Control and optimization
5 Putting it all together: theory to practice
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’

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.

---

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.  

- 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.

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

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

---

Was Google Flu Trends a pioneer?

Polgreen et al. \(^{14}\) 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’

- Explored searches with one- to ten-week lead times as explanatory variables; reports 1-3 week lead time over CDC reporting

\(^{14}\)PM Polgreen, Y Chen, DM Pennock, FD Nelson, *Clinical Infectious Diseases*, 2008
Google Flu Trends vs. traditional surveillance

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

- 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)

---

How does GFT fare when used in conjunction with other indicators?\textsuperscript{16}


- 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

More murmurs of discontent

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

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

---

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

---


Recent improvements to Google Flu Trends\textsuperscript{20}

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

\textsuperscript{20}http://patrickcopeland.org/papers/isntd.pdf
Designing your own vocabulary

- 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.’

---

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 \(^{22}\) and Lampos et al. \(^{23}\) 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

\(^{22}\)A Culotta, *Proceedings of the First Workshop on Social Media Analytics*, 2010

\(^{23}\)V Lampos, N Cristianini, *ACM TIST*, 2012
Using Twitter during the H1N1 pandemic

Signorini et al. 24 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

---

Getting into more detailed content analysis

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

---

Multi-level classification of tweets

Even finer distinctions

- 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

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Words in Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infection</td>
<td>getting, got, recovered, have, having, had, has, catching, catch, cured, infected</td>
</tr>
<tr>
<td>Possession</td>
<td>bird, the flu, flu, sick, epidemic</td>
</tr>
<tr>
<td>Concern</td>
<td>afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified</td>
</tr>
<tr>
<td>Vaccination</td>
<td>vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray</td>
</tr>
<tr>
<td>Past Tense</td>
<td>was, did, had, got, were, or verb with the suffix “ed”</td>
</tr>
<tr>
<td>Present Tense</td>
<td>is, am, are, have, has, or verb with the suffix “ing”</td>
</tr>
<tr>
<td>Self</td>
<td>I, I’ve, I’d, I’m, im, my</td>
</tr>
<tr>
<td>Others</td>
<td>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</td>
</tr>
</tbody>
</table>

---

The SIRS equations are given by:

\[
\begin{align*}
\frac{dS}{dt} &= N - S - I - L - \beta(t)SI - \alpha \\
\frac{dI}{dt} &= \beta(t)SI - \frac{I}{D} + \alpha
\end{align*}
\]

where the AH modulated reproductive number is given by

\[
R_0(t) = \exp(a \times q(t) + b) + R_{0\text{min}}
\]

where, \(a = -180\) and \(b = \log(R_{0\text{max}} - R_{0\text{min}})\). \(q(t)\) 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

---

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

---

32 LC Brooks, DC Farrow, S Hyun, RJ Tibshirani, R Rosenfiled, *PLOS Comp. Biol.*, 2015
OpenTable reservation monitoring

- 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

---

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

---

Global disease monitoring with Wikipedia

- 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

---

Parking lot imagery

Estimating “fill rate” in parking lots

Cars: 151
Spaces: 400
Fill Rate: 37.8%
# Hospital parking lot study in Latin America

<table>
<thead>
<tr>
<th>Country</th>
<th>Images Ordered</th>
<th>Images Used</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>1,566</td>
<td>781</td>
<td>11/1/2011–5/26/2013</td>
</tr>
<tr>
<td>Chile</td>
<td>682</td>
<td>292</td>
<td>11/1/2011–5/26/2013</td>
</tr>
</tbody>
</table>

**Diagram:**
- Parking Lot Border in Red
- Street Parking Border in Green
- Hospital Perimeter Border in Blue
Parking lot imagery results

---

Putting it all together

37

Issues to consider

- How to combine different data sources: 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 ($X$) represented as a multivariate weekly time-series with flu counts ($P$) as target variable - heavily skewed matrix and hence regression problem ill-defined
Putting it all together - contd (2)

Problem framework and solution sketches

- Solution \((MFN)\): Matrix Factorization (dim reduction) + nearest neighbor search (non-linear) - similar to recommender systems

- Data-Target Matrix \(M = [XP]\)

- Model

\[
\hat{M}_{i,j} = \underbrace{b_{i,j}}_{\text{avg. value}} + \underbrace{U_i^T F_j}_{\text{dense subspace}} + F_j|\mathcal{N}(i)|^{-\frac{1}{2}} \sum_{k \in \mathcal{N}(i)} (M_{i,k} - b_{i,k})x_k
\]  

\[\text{(4)}\]

- Fitting

\[
b_*, F, U, x_* = \arg\min \left( \sum_{i=1}^{m-1} \left( M_{i,n} - \hat{M}_{i,n} \right)^2 \right) + \lambda_2 \left( \sum_{j=1}^{n} b_j^2 + \sum_{i=1}^{m-1} \|U_i\|^2 + \sum_{j=1}^{n} \|F_j\|^2 + \sum_{k} \|x_k\|^2 \right)
\]

\[\text{(5)}\]

\[^{38}\text{Y. Koren, KDD 2008}\]
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.
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.
Advantages of *MFN*

- Theoretical properties: Handles non-linearity and sparsity in a well defined framework. \(^{39}\)
- 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

\(^{39}\)Y. Koren, KDD 2008
Extending Forecasting Boundary

- Dynamic Model

\[ y_i = w_i z_i^T + \epsilon_i \]

- Similarity Constraints on time points

---

\[ \sum_{i=1}^{N} (y_i - w_i z_i^T)^2 + \eta \sum_{i,j} s_{ij} \| w_i - w_j \|_2^2 \]

---

\(^{40}\)Z Wang, P Chakraborty, SR Mekaru, JS Brownstein, J Ye, N Ramakrishnan, *KDD*, 2015
Dynamic GLMs to account for local variations.

DARX: Identity Link

\[ \sum_{i=1}^{N} (y_i - w_i z_i^T)^2 + \eta \sum_{i,j} S_{ij} \| w_i - w_j \|_2^2 \]

DPARX: Poisson Link

\[ \min_{w} \sum_{i} (w_i z_i^T - y_i \log(w_i z_i^T)) + \eta \sum_{i,j} S_{ij} \| w_i - w_j \|_2^2 \]

st. \( w z_i^T \geq 0, \forall i. \)

DPARX performs better overall.
Recommendations for future forecasting programs

- 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

---

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

Dynamical models, differential equations, network based models, etc.

Use models to simulate epidemics

Syndromic data
- Data cleaning, integration, manipulation, querying, etc.

Incidence data

Epidemic Library
- Time series curves
- Disease model parameters (Database)

Historical data from past epidemics and pandemics

Compare epidemic to those in the library

Optimization methods for finding new model parameters

Event forecasts with appropriate confidence
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

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

---

Forecasting Global Epidemic Spread

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

---

Mapping interactions using Twitter

- 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%

---

A sobering study

Smallpox simulation under human mobility assumptions

- 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 suffices over mass vaccination

---

Some Notable Mentions

- Exploring novel data streams to enhance disease surveillance \(^{47}\).
- Spatio-temporal modeling of 2014 Ebola outbreak \(^{48}\).
- Individualizing prediction of Disease Trajectories \(^{49}\).

\(^{47}\) BA Althouse et al., *EPJ Data Science*, 2015

\(^{48}\) S Merler et al., *The Lancet Infectious Diseases*, 2015

\(^{49}\) P Schulam, S Saria, *NIPS*, 2015
Rounding Up: The Big Picture

Data Sources
- Signal strength w.r.t target
- Sparsity: Ideally complimentary sources
- Robustness

Statistical Models
1. Linear vs. Non-Linear
2. Dynamic vs. Static
3. Point process vs. Continuous Process

Fusion/Ensembles
1. Data vs. Model
2. Majority voting vs. Metric based
3. Random Forest, Bayesian Fusion, etc.

Forecast Attribute
1. Short-term vs. Long term
2. Statistical vs. Causal Methods
3. Dis. Characteristics: seasonal vs non-seasonal

Disease Forecasting Pipeline
Weekly predictions sent out for 15 Latin American countries

EMBERS, http://dac.cs.vt.edu/research-project/embers/
Disease Forecasting Competitions

- IARPA Open Source Indicator Program. 51, 2010-2014
- CDC Influenza Prediction Challenge, 2013-current
- DARPA Chikungunya Prediction Challenge, 2014
- RAPIDD Ebola Challenge, 2015

51 EMBERS, http://dac.cs.vt.edu/research-project/embers/
1. Goals, History, Basic Concepts
2. Dynamics and Analysis
3. Surveillance and Forecasting
4. Control and optimization
5. Putting it all together: theory to practice
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 $\mathcal{P}(S)$ denote the phase space of a given SyDS $S(G)$. Modify $S$ optimally so that the set of reachable states in $\mathcal{P}(S)$ 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 $\mathcal{P}(S)$ satisfy $\mathcal{P}_1$
  - Similarly, reducing epidemic duration $\equiv$ reducing transient length in $\mathcal{P}(S)$

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\(^{52}\)
  - Decrease \(\beta\), 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 \(\delta\): culling animals

Interventions can be modeled in networks as node deletions (vaccination), edge deletion (quarantining) and reducing \(\beta\) on edges

\(^{52}\)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).
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 $\alpha$ if for any instance $I$ of the problem, it finds a solution $S$ such that $\text{cost}(S) \leq \alpha \cdot \text{cost}(\text{OPT}(I))$
- 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 $B/\epsilon$ vaccines, so that $\#\text{infections}$ is at most $1/(1 - \epsilon)$ times optimal\textsuperscript{53} \textsuperscript{54}
  - If initial infection is random: $O(\log n)$ approximation\textsuperscript{55}

\textsuperscript{53}A. Hayrapetyan, D. Kempe, M. Pal and Z. Svitkina, \textit{ESA}, 2005
\textsuperscript{54}S. Eubank, V. S. Anil Kumar, M. Marathe, A. Srinivasan and N. Wang, \textit{AMS DIMACS}, 2005
\textsuperscript{55}V. S. Anil Kumar, R. Rajaraman, Z. Sun and R. Sundaram, \textit{IEEE ICDCS}, 2010
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 $\beta > 2$. 
Compartmental differential equation based approach for OVAP\textsuperscript{59}

- Age-structured differential equation model for H1N1
  - Mixing between age groups based on survey data\textsuperscript{56}
  - $R_0 = 1.4$ for swine flu\textsuperscript{57}
  - 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\textsuperscript{58}

- CDC guidelines for swine-flu: prioritize vaccination for children 6 months to 5 years, and 5-18 years

\textsuperscript{56} J. Mossong et al., \textit{PLoS Med.}, 2008
\textsuperscript{57} C. Fraser et al., \textit{Science}, 2009
\textsuperscript{58} Such as: A. C. Haddix et al., \textit{Oxford University Press}, 1996; M. Meltzer et al., \textit{Emerg. Infec. Dis.}, 1999
\textsuperscript{59} Medlock and Galvani, \textit{Science}, 2009
Coupled differential equation model

\[
\begin{align*}
\frac{dU_{Sa}}{dt} &= -\lambda_a U_{Sa} \\
\frac{dU_{Ea}}{dt} &= \lambda_a U_{Sa} - \tau_a U_{Ea} \\
\frac{dU_{la}}{dt} &= \tau_a U_{Ea} - (\gamma_{Ua} + \nu_{Ua}) U_{la} \\
\frac{dU_{Ra}}{dt} &= \gamma_{Ua} U_{la} \\
\frac{dV_{Sa}}{dt} &= -(1 - \epsilon_a) \lambda_a V_{Sa} \\
\frac{dV_{Ea}}{dt} &= (1 - \epsilon_a) \lambda_a V_{Sa} - \tau_a V_{Ea} \\
\frac{dV_{la}}{dt} &= \tau_a V_{Ea} - (\gamma_{Va} + \nu_{Va}) V_{la} \\
\frac{dV_{Ra}}{dt} &= \gamma_{Va} V_{la}
\end{align*}
\]

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

Onset of epidemic: sequester guard to base
- Complete mixing within rooms
- Some people are already exposed to disease and can get infected and then infectious
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\textsuperscript{60}
- Significantly outperforms random allocation

\textsuperscript{60}C. Barrett et al., ACM SIGHIT International Health Informatics Symposium, 2012
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
1. **Goals, History, Basic Concepts**

2. **Dynamics and Analysis**

3. **Surveillance and Forecasting**

4. **Control and optimization**

5. **Putting it all together: theory to practice**
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

Timeline for H1N1

28 Mar: First Case of H1N1
4 year Boy in Mexico

13 April: First death
A woman in Mexico

17 April: First Cases of H1N1 start being detected in US

24 April: 8 Cases of H1N1 in US

30 April: 30 schools across US closed

20th May: 10,000 cases worldwide

29th April: WHO raises pandemic level 5
Mexico suspends non-essential services at government offices

29th April: WHO declares a pandemic. First pandemic in 41 years

4th Sept: 625 deaths in last week. 2500 deaths so far.

10th Sept: Just one shot of vaccine appears to be protective

625 deaths in last week. 2500 deaths so far.
### 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**: EpiSimms 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 predictive 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 respiratory illness will create complex demands on the medical infrastructure, which will require high-level coordination to maximize the delivery of care.
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:

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?
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

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

---

61 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/featured-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 ⇒ 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 $w = (w(v_1, V), w(v_2, V), ..., w(v_n, V))$ for nodes $v_i \in V$, a random graph $G(w)$ 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 $\text{deg}(u)$.

- 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

### Methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey &amp; direct</td>
<td>Multi purposed use; fewer biases; can capture multiple correlations</td>
<td>can be expensive to collect data observations;</td>
</tr>
<tr>
<td>Wi-Fi localization</td>
<td>Accuracy; Energy usage 50% GPS</td>
<td>Providing access point is expensive</td>
</tr>
<tr>
<td>GPS localization</td>
<td>High spatial precision: 5m; Can distinguish between transportation modes</td>
<td>High battery (energy) usage; expensive; sampling biases; No (low quality) signal in indoor environment</td>
</tr>
<tr>
<td>Cellular network localization (passive) (Call Data Records);</td>
<td>Automatically generated;</td>
<td>Sparse in time; Lower spatial resolution ( 175m); Needs more filtering; sampling biases; Proprietary</td>
</tr>
<tr>
<td>Cellular network localization (active)</td>
<td>More accuracy than passive localization; Less expensive than previous methods</td>
<td>More costly than passive form; sampling biases; Proprietary and thus not publicly available</td>
</tr>
</tbody>
</table>
Characterizing human travel patterns using CDR\textsuperscript{65, 66}

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

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

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

- Study by Lu et al.\textsuperscript{64} highlights that algorithms are capable of approaching the theoretical limits of predictability

\textsuperscript{65}M Gonzalez, CA Hidalgo, A Barabasi, \textit{Nature}, 2008
\textsuperscript{66}Becker et al, CACM 2013
Synthesizing social proximity networks using RFID tags and wifi localization \(^{67, 68}\)

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

\(^{67}\) N Eagle, A Pentland, D Lazer, *PNAS*, 2009

\(^{68}\) M. Salathe et al., *PNAS*, 2010
Step 2: Within host disease progression models.
Parameter estimation is done using derivative free optimization method.

Simulation based optimization approach
Step 3: Simulations to unravel the disease dynamics over a network
Recall: Theoretical results on computing dynamics for special classes of networks

Worst case complexity: We are given an SIR GDS $S$, an initial configuration $I$ and a final configuration $B$. The goal is to decide whether $S$ starting from $I$ reaches $B$ with a non-zero probability or $B$ 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 \textbf{NP}-hard. It is \#P-hard if we want to assure that $B$ reaches with a probability $\geq \pi$. Moreover, this result holds even when the initial configuration has one infected node.

Implications: need to develop \textit{fast simulations} to compute the epidemic dynamics in general.
## Fast high performance simulations: From 40 hours to 40 seconds

<table>
<thead>
<tr>
<th>Distinguishing Features</th>
<th>EpiSims (Nature’04)</th>
<th>EpiSimdemics (SC’09, WSC’10)</th>
<th>EpiFast (ICS’09)</th>
<th>Indemics (ICS’10, TOMACS’11)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Solution Method</strong></td>
<td>Discrete Event Simulation</td>
<td>Interaction-Based Simulation</td>
<td>Combinatorial + discrete time</td>
<td>Interaction-based, Interactive Simulations</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>~40 hours</td>
<td>1 hour for 300 Million nodes</td>
<td>~40 seconds</td>
<td>15 min-1 hour</td>
</tr>
<tr>
<td><strong>Co-evolving Social Network</strong></td>
<td>Can work</td>
<td>Works Well</td>
<td>Works only with restricted form</td>
<td>Very general</td>
</tr>
<tr>
<td><strong>Disease transmission model</strong></td>
<td>Edge as well as vertex based</td>
<td>Edge as well as vertex based (e.g. threshold functions)</td>
<td>Edge based, independence of infecting events</td>
<td>Edge based</td>
</tr>
<tr>
<td><strong>Query and Interventions</strong></td>
<td>Restrictive</td>
<td>Scripted, groups allowed but not dynamic</td>
<td>Scripted and specific groups allowed</td>
<td>Very general: no restriction on groups</td>
</tr>
</tbody>
</table>
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
- 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 → data-analytics → 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

- \( X(1, T) = \langle X(1), \ldots, X(T) \rangle \) 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 \( \theta \) such that the outcomes \( Y_\theta \) are “close” to \( X \), 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 \( r \)th replicate.

- The forecasting problem: Given a quantity of interest \( Q(\cdot) \) (e.g., number infected), compute the time-series \( Q(1, T + r) = \langle Q(1), \ldots, Q(T + r) \rangle \), 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 help

- **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
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
  - \( \beta \) 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-infrastructure-html

Gleam

www.gleamviz.org
CIEPI: Cyber-infrastructure for computational epidemiology, NDSSL, BI, Virginia Tech

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 (DBV)**
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.

**TransSim**

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

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

- **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.

---

69Alexander et al., Vectore Borne & zoonotic diseases, 2012
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**

---

70Kuhlman et al. WSC 2012, AAMAS 2014
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

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

vaccination study

Details
- Replicates: 25
- Total Cells: 1
- Simulated Days: 200
- Model: EpiFast

Region

Disease Models
- Select disease model

Show Cells

vaccination study

Region: No Region Attached

Search region: Montgomery County VA

Cancel

Save
Select disease model
Specify initial conditions and interventions
Specify intervention
Specify intervention

### Subpopulation
- **Selection:** Seniors
- **Type:** Age
- **Category:**
  - Preschool: 6.05% (4507), 100% (4507)
  - School-age: 14.9% (11097), 100% (11097)
  - Adults: 69.9% (52047), 100% (52047)
  - Seniors: 9.04% (6725), 100% (6725)

### Compliance
- **Type:** Sweep
- **Linear/Customized:**
  - **Initial Value:** 20
  - **Final Value:** 60
  - **Increment Value:** 10

### Efficacy
- **Type:** % Value
- **Value:** 46%

### Rate of Administration
- **Rate Per Day:** Unlimited

### Trigger
- **On Day/Every Day/Infectious:**
  - **Name:** day15
  - **Value:** 15

### Duration
- **Default Exp Simulation Days**
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, NSF DIBBS Grant ACI-1443054, NSF BIG DATA Grant IIS-1633028,
- Defense Threat Reduction Agency grant HDTRA1-11-1-0016, DTRA CNIMS contract HDTRA1-11-D-0016-0001, DTRA BSVE HDTRA1-11-D-0016-0001,
- National Institute of Health, Midas grant 2U01GM070694-09, and 1R01GM109718.
- 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:

Bryan Lewis (blewis@bi.vt.edu)
Madhav V. Marathe (mmarathe@bi.vt.edu)
Anil Vullikanti (akumar@bi.vt.edu)


Additional references: Forecasting and Surveillance


Additional references: Forecasting and Surveillance


Additional references: Forecasting and Surveillance

Additional references: Forecasting and Surveillance


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.