Computational Epidemiology

Madhav V. Marathe, Naren Ramakrishnan and Anil Kumar Vullikanti
Virginia Tech
Virginia Bioinformatics Institute, Technical Report 14-087, 2014.
Latest full version:
`ndssl.vbi.vt.edu/supplementary-info/vskumar/kdd-slides.pdf`

August 28, 2014

These slides were presented as a part of the tutorial titled *Computational Epidemiology* at KDD 2014, held in New York City, August 24-27, 2014; `http://www.kdd.org/kdd2014/tutorials.html`.

We hope to update the slides and supplementary material continually over the next year.
The current version of the slides can be found at:
`ndssl.vbi.vt.edu/supplementary-info/vskumar/kdd-slides.pdf`

Outline

1 Goals, History, Basic Concepts

2 Dynamics and Analysis

3 Inference problems

4 Control and optimization

5 Surveillance and Forecasting

6 Putting it all together: theory to practice

Goals

- **Overview and state of the art** – emphasis on computing and data science
- **Describe open problems and future directions** – aim to attract computing and data scientists to work in this exciting area
- **Unified framework** based on graphical dynamical systems and associated proof theoretic techniques; e.g. spectral graph theory, branching processes, 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
  - Game theoretic formulations, behavioral modeling, economic impact
  - Validation, verification and uncertainty quantification (UQ)

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

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

Epidemic Science in Real Time

Harvey V. Fineberg is president of the Institute of Medicine in Washington, DC, USA.

Mary Wilson Harvey is a professor of medicine and pediatrics at Stanford University School of Medicine, Stanford University, Stanford, CA, USA.

Epidemics are a recurrent feature of human life. Cholera, for example, has been with us since the early 19th century, when it first appeared in India and then spread widely to other parts of the world. Influenza, in contrast, is a recent arrival, appearing in its current form only in the 20th century. Yet, despite its relatively short tenure, influenza has caused more deaths than cholera. Influenza viruses are highly variable, and the virulence of influenza epidemics varies greatly from year to year. Influenza viruses also can cross species barriers, causing epidemics in animal populations and occasionally infecting humans with severe consequences. Influenza is not the only infectious disease that can spread rapidly around the world. In the current 2009 H1N1 influenza outbreak, for example, the World Health Organization (WHO) has reported more than 30,000 laboratory-confirmed cases in 74 countries, and the number of cases is increasing. Although the current H1N1 outbreak has not been associated with a high rate of severe illness or death, it has underscored the threat that infectious diseases can pose to the health of people around the world.

Epidemics are not the only health threat that can spread across borders. Tobacco products, for example, are a major cause of premature death worldwide, and the tobacco industry has been very successful in promoting its products around the world. Although the spread of tobacco products is not a biological process, it is a social process that has been compared to an epidemic. Other health threats that can spread across borders include the use of illicit drugs, the consumption of contaminated foods, and the emergence of antibiotic-resistant organisms in hospitals. Epidemics tend to suggest most powerfully the need for collaborative efforts across borders, but many other health threats also require collaborative efforts across borders. Epidemics provide a dramatic opportunity to demonstrate the value of such efforts, but they also have challenges. For example, if the United States decides to intervene to stem the number of people who fall ill, the consequences are the mirror image of the consequences if the United States decides not to intervene. In the case of H1N1 influenza, the availability and administration of vaccine (and/or of an other of antiviral medications) will be a major factor in the response to the current epidemic. Decisions about how to allocate vaccine doses will have consequences for many other countries. Time and resources are limited, however, and leaders in government agencies will need to ensure that the most consequential and cost-effective decisions are reached. Such decisions will be made more difficult if the United States is not able to share information with its partners in the face of a threatened pandemic, emergency measures cannot be carried out.

Harvey V. Fineberg and Mary Elizabeth Wilson

www.sciencemag.org SCIENCE VOL 324 22 MAY 2009 Published by AAAS

Downloaded from www.sciencemag.org on May 21, 2009

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.

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

$$\frac{ds}{dt} = -\beta is$$
$$\frac{di}{dt} = \beta is - \gamma i$$
$$\frac{dr}{dt} = \gamma i$$

- *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: $\mathrm{R} = \{1, 2, 3\}$ and $\mathrm{S} = \{4\}$

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

- 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 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

1 Goals, History, Basic Concepts

2 Dynamics and Analysis

3 Inference problems

4 Control and optimization

5 Surveillance and Forecasting

6 Putting it all together: theory to practice

Dynamics in Compartmental Models

$$\frac{ds}{dt} = -\beta i s$$
$$\frac{di}{dt} = \beta i s - \gamma i$$
$$\frac{dr}{dt} = \gamma i$$

- $\frac{di}{dt} > 0$ (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 nth 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
- $\Pr[\text{node } i \text{ in } n\text{th level is infected}] = p^n$
- $E[X_n] = p^n d^n = R_0^n$
- Note that $E[X_n] = 1 \cdot Pr[X_n = 1] + 2 \cdot Pr[X_n = 2] + 3 \cdot Pr[X_n = 3] + \ldots$
- This implies: $E[X_n] = \Pr[X_n \geq 1] + \Pr[X_n \geq 2] + \ldots$; since $Pr[X_n = i]$ contributes exactly i copies of itself to the sum
- $E[X_n] \geq \Pr[X_n \geq 1] = q_n$
- 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 descendant j of the root
- Let EP_j be the event that the epidemic persists until the nth level of T_j starting at one of the children of the root node $\Pr[EP_j] = q_{n-1}$
- The epidemic persists starting at the root and spreading via infecting j is pq_{n-1}. Since the roote has d neighbors, the probability that the epidemic does not persist is $(1 - pq_{n-1})^d$.
- $q_n = 1 - (1 - pq_{n-1})^d$

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

Analysis of the branching process: case 2

- $f(x) = 1 - (1 - px)^d; \quad f(0) = 0$ 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 $\rho(A) = \max_x x^T A x / x^x$
- 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[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

- Necessary and sufficient conditions tight for some graphs
  - Hypercube: $\rho(G) = \log_2 n$, and $\eta(m) = (1-a)\log_2 n$ for $m = n^a$
  - Erdős-Rényi model: $\rho(G) = (1+o(1))np = (1+o(1))d$ and $\eta(m) = (1+o(1))(1-\alpha)d$ where $m/n \to \alpha$
  - 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 $\lambda \in (0, \frac{1}{\gamma-1})$
- 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:

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

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

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

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

1 Goals, History, Basic Concepts

2 Dynamics and Analysis

3 Inference problems

4 Control and optimization

5 Surveillance and Forecasting

6 Putting it all together: theory to practice

Inference problems: limited information

- Partial information available about network and disease model parameters
- Source of infections and time progression not fully known

Inference problem	Inputs
Estimate source	Network and observed infections
Infer network parameters and structure	Observed infection cascades
Estimate disease model parameters	Network and observed infections
Infer behavior model	Network, observed infections, surveys
Forecasting epidemic characteristics	Partial infection counts, network structure

Inference problems: GDS perspective

- Recall: GDS computation specified by graph, local functions and initial configuration
- Inference problems: one of more of these components are not known or partially known
  - Source inference = find initial configuration (predecessor existence problem in dynamical systems)
  - Disease model inference = find local functions
  - Network inference = find graph

Source inference problems

- Patient zero: the first/index case of the disease
- Finding patient zero: key public health concern during every epidemic

Source inference problems

- Patient zero: the first/index case of the disease
- Finding patient zero: key public health concern during every epidemic

Gaëtan Dugas: presumed to be index case for AIDS epidemic in the US

Source inference problems

- Patient zero: the first/index case of the disease
- Finding patient zero: key public health concern during every epidemic

Gaëtan Dugas: presumed to be index case for AIDS epidemic in the US

Mary Mallon: presumed to be responsible for typhoid outbreak in New York in early 1900's

Image source:
http://en.wikipedia.org/wiki/File:AIDS_index_case_graph.svg
http://upload.wikimedia.org/wikipedia/commons/thumb/f/fd/Mallon-Mary_01.jpg/330px-Mallon-Mary_01.jpg

Source inference problems: different formulations

General framework

Assume graph $G = (V, E)$ and set of infected nodes I are known. Find the source(s) which would result in outbreak "close" to I.

- Likelihood maximization, assuming single source in the SI model[15]
- Minimize difference between resulting infections and I in SIR model[16]
- Formulation in SI model based on Minimum Description Length, that extends to multiple sources[17]

[15] Shah and Zaman, *ACM SIGMETRICS*, 2009
[16] T. Lappas, E. Terzi, D. Gunopulos and H. Mannila, *KDD*, 2010.
[17] B. Aditya Prakash, J. Vrekeen and C. Faloutsos, *ICDM*, 2012

Source inference as likelihood maximization in SI model assuming single source[18]

- Single source $\Rightarrow$ the subgraph G_I induced by infected set I is connected
- SI model with time τ_{ij} for infection spread on edge (i, j) exponentially distributed and independent on edges.
  - Reduces problem to partial orderings which result in set I

Source inference problem

Given graph $G = (V, E)$ and a set I of infected nodes under the SI model of diffusion, find node

$$\hat{v} \in \text{argmax}_{v \in G_I} \Pr[G_I | v],$$

where G_I denotes the graph induced by set I.

[18] Shah and Zaman, *ACM SIGMETRICS*, 2009

Results

- Notion of rumor centrality $R(v, G_I)$ as ML estimator in trees
  - Efficient algorithm for estimating $R(v, G_I)$
- Heuristic to extend to general graphs: assume transmission on BFS tree.

Theorem

1. *Assume d-regular tree.*
   1. *If $d = 2$, the source detection probability after time t is $O(\frac{1}{\sqrt{t}})$.*
   2. *If $d > 2$, the source detection probability is at least α_d as $t \to \infty$.*
2. *In non-regular trees which satisfy "polynomial growth" and "regularity" properties, the source detection probability of the estimator $\to 1$ as $t \to \infty$.*

Estimating $\Pr[G_I|v]$: preliminaries

Observation

Only some permutations of node infections can result in G_I

- Consider permutations of the order in which nodes got infected
- With node 1 as source, permutations $(1, 2, 3, 4)$ and $(1, 2, 4, 3)$ are feasible, but $(1, 3, 2, 4)$ is not

Estimating $\Pr[G_I|v]$ in trees (I)

- $\Omega(v, G_I)$: set of all permitted permutations starting with v and resulting in G_I
- Example: consider permutation $\sigma = (v_1 = v, \ldots, v_{|I|})$
- Let $G_k(\sigma)$ denote the graph induced by nodes $\{v_1, \ldots, v_k\}$

Estimating $\Pr[G_I|v]$ in trees (I)

- $\Omega(v, G_I)$: set of all permitted permutations starting with v and resulting in G_I
- Example: consider permutation $\sigma = (v_1 = v, \ldots, v_{|I|})$
- Let $G_k(\sigma)$ denote the graph induced by nodes $\{v_1, \ldots, v_k\}$

Consider permutation $\sigma = (1, 2, 3, 4)$

$\#$ uninfected nbrs $= 4$

$\Pr[3 \text{ infected} | G_2(\sigma), \text{node } 1] = \frac{1}{4}$

Estimating $\Pr[G_I|v]$ in trees (I)

- $\Omega(v, G_I)$: set of all permitted permutations starting with v and resulting in G_I
- Example: consider permutation $\sigma = (v_1 = v, \ldots, v_{|I|})$
- Let $G_k(\sigma)$ denote the graph induced by nodes $\{v_1, \ldots, v_k\}$

Consider permutation $\sigma = (1, 2, 3, 4)$

uninfected nbrs $= 4$

$\Pr[3 \text{ infected}|G_2(\sigma), \text{node } 1] = \frac{1}{4}$

uninfected nbrs $= 5$

$\Pr[4 \text{ infected}|G_3(\sigma), \text{node } 1] = \frac{1}{5}$

Estimating $\Pr[G_l|v]$ in trees (II)

- Let $n_k(\sigma)$ denote number of uninfected neighbors in $G_k(\sigma)$
- $n_k(\sigma) = n_{k-1}(\sigma) + d_k(\sigma) - 2$, where $d_k(\sigma)$ is the degree of node $v_k(\sigma)$

- $\sigma = (1, 2, 3, 4)$
- $n_2(\sigma) = 4$
- $n_3(\sigma) = 5 = n_2(\sigma) + d_2(\sigma) - 2$

Estimating $\Pr[G_I|v]$ in trees (III)

- $n_k(\sigma) = n_{k-1}(\sigma) + d_k(\sigma) - 2 = d_1(\sigma) + \sum_{i=2}^{k}(d_i(\sigma) - 2)$, where $d_k(\sigma)$ is the degree of node $v_k(\sigma)$ (by induction)

$$\begin{aligned}
\Pr[\sigma|v] &= \prod_{k=2}^{N} \Pr[k\text{th infected node is } v_k | G_{k-1}(\sigma), v] \\
&= \prod_{k=2}^{N} \frac{1}{n_{k-1}(\sigma)} \\
&= \prod_{k=2}^{N} \frac{1}{d_1(\sigma) + \sum_{i=2}^{k}(d_i(\sigma) - 2)}
\end{aligned}$$

Estimating $\Pr[G_I|v]$ in trees (IV)

- If G is d-regular, $n_k(\sigma) = d_1(\sigma) + \sum_{i=2}^{k}(d_i(\sigma) - 2) = dk - 2(k-1)$
- For d-regular trees, $\Pr[\sigma|v] = \prod_{k=1}^{|I|-1} \frac{1}{dk-2(k-1)} \equiv p(d, I)$
  - $\Rightarrow \Pr[G_I|v] \propto p(d, I)|\Omega(v, G_I)|$
  - Let $R(v, G_I)$ denote $|\Omega(v, G_I)|$
- For d-regular trees

$$\begin{aligned}
\hat{v} &\in \operatorname{argmax}_v \Pr[G_I|v] \\
&= \operatorname{argmax}_v \sum_{\sigma \in \Omega(v, G_I)} \Pr[\sigma|v] \\
&= \operatorname{argmax}_v R(v, G_I) p(d, I) \\
&= \operatorname{argmax}_v R(v, G_I)
\end{aligned}$$

Regular trees

Estimator reduces to $R(v, G_I) \equiv$ **Rumor centrality**

Estimator for general trees

- Permutations are not all equally likely. So $\Pr[G_I|v] \not\propto |\Omega(v, G_I)|$
- Heuristic: assume the infections spread along a BFS tree. Let σ^* be a permutation consistent with a BFS tree (ties broken arbitrarily).

Estimator for general trees

Let σ_v^* be a permutation consistent with the BFS tree at v. Then, the estimator is

$$\hat{v} \in \operatorname{argmax}_{v \in G_I} \Pr[\sigma_v^*|v] R(v, G_I)$$

where G_I denotes the graph induced by set I.

Similarly extend to general graphs: assume infection spread on BTS tree $T_{bfs}(v)$

Computing $R(v, G)$ in trees

- Let T_u^v denote the subtree rooted at u, with v as source
- Let $n_u^v = |T_u^v|$

Recursion:

$$\begin{aligned} R(v, T_v^v) &= (n-1)! \prod_{u \in \text{child}(v)} \frac{1}{n_u^v!} \prod_{u \in \text{child}(v)} R(u, T_u^v) \\ &= (n-1)! \prod_{u \in \text{child}(v)} \frac{(n_u^v - 1)!}{n_u^v!} \prod_{w \in \text{child}(u)} \frac{R(w, T_w^u)}{n_w^u!} \\ &= n! \prod_{u \in G_N} \frac{1}{n_u^v} \text{ (Continuing the recursion)} \end{aligned}$$

$R(v, G)$ can be computed in a tree in linear time.

Alternative approach in the SIR model[19]

- Given: graph G, an infected set I.
- Define activation vector $\mathbf{a}$ as: $\mathbf{a}(v) = 1$ if $v \in I$
- Define Pr[node v gets infected|set X initially infected] $= \alpha(v, X)$

k-Effectors problem

Given graph G, infected set I and a parameter k, find set X (the *effectors*) such that

$$C(X) = \sum_{v} |\mathbf{a}(v) - \alpha(v, X)|$$

is minimized and $|X| \leq k$.

- Generalizes influence maximization problem
- When G is a tree, can be solved optimally by dynamic programming

[19] T. Lappas, E. Terzi, D. Gunopulos and H. Mannila, *KDD*, 2010.

General graphs and extensions

- Maximum likelihood tree in general graphs:
  - For tree $T = (V_T, E_T)$, likelihood $L(T) = \prod_{(u,v) \in E_T} p(u, v)$
    - Objective: find maximum likelihood tree that contains set I
    - Equivalently, consider log-likelihood: $LL(T) = -\sum_{(u,v) \in E_T} \log p(u, v)$
    - Corresponds to directed steiner tree problem
- Several extensions and variations
  - Find set X so that number of infections within I is maximized[20]
  - Analysis of computational complexity[21] [22]
  - Other estimators based on distances[23]

[20] D. Nguyen, N. Nguyen and M. Thai, *MilCom*, 2012
[21] G. Askalidis, R. Berry and V. Subramanian, 2014
[22] M. Marathe, S. S. Ravi, D. Rosenkrantz, 2014
[23] K. Zhu and L. Ying, 2013

Formulation based on Minimum Description Length[24]

- Given: graph G and infected subgraph G_I in the SI model
- Ripple R: list of node ids by time, representing the order in which they get infected in G_I.
- For seed set S and ripple R, description length $\mathcal{L}(G_I, S, R) = \mathcal{L}(S) + \mathcal{L}(R|S)$

Minimum Infection Description Problem

Given graph G and infected subgraph G_I in the SI model, find seed set S and associated ripple R that minimize $\mathcal{L}(G_I, S, R)$.

[24] B. Aditya Prakash, J. Vrekeen and C. Faloutsos, *ICDM*, 2012

Network Inference problems

- Nodes/edges are unknown or partially known
- Graph models with unknown parameters
- Surveillance information, e.g., cascades or time series of infections
- Formulations and results include:
  - Maximum likelihood formulation for network inference[25]
  - Upper and lower bounds on the number of cascades needed for inferring graph and its properties[26] [27]

[25] M. Gomez-Rodriguez, J. Leskovec and A. Krause, *ACM Trans. Knowledge Discovery from Data, 2012*

[26] B. Abrahao, F. Chierichetti, R. Kleinberg and A. Panconesi, *KDD*, 2013

[27] P. Netrapalli and S. Sanghavi, *ACM SIGMETRICS*, 2012

Inferring networks from cascades[28]

- Hidden underlying network G^*
- Given: a set $\mathcal{C}$ of observed cascades created by a contagion
  - Cascade c is specified by triples $(u, v, t_v)_c$: where contagion reaches node v from u, at time t_v
  - Only the time $\mathbf{t}_c = [t_1, \ldots, t_n]$ is observed, where t_u is the time when node u gets infected by cascade $c \in \mathcal{C}$ ($t_u = \infty$ if u is not infected)
- Let $\mathcal{T}_c(G)$ denote the set of all directed spanning trees (arborescences) in G induced by the nodes infected in cascade c. Then

$$\Pr[c|G] = \sum_{T \in \mathcal{T}_c(G)} \Pr[c|T] \Pr[T|G]$$

$$\Pr[\mathcal{C}|G] = \prod_{c \in \mathcal{C}} \Pr[c|G]$$

[28] M. Gomez-Rodriguez, J. Leskovec and A. Krause, *ACM Trans. Knowledge Discovery from Data, 2012*

Formulation

Diffusion network inference problem

Given the vector of node infection times $\mathbf{t}_c$ for a set of cascades $c \in \mathcal{C}$, SIR model with transmission probability β and incubation time distribution $P_c(u, v)$, find the network $\hat{G}$ such that:

$$\hat{G} = \text{argmax}_{|G| \leq k} \Pr[\mathcal{C}|G]$$

Approximations and main ideas

- Assume all trees $T \in \mathcal{T}_c(G)$ are equally likely: $\Pr[T|G] = 1/|\mathcal{T}_c(G)|$
  - $\Rightarrow \Pr[c|G] \propto \sum_{T \in \mathcal{T}_c(G)} \prod_{(u,v) \in E_T} P_c(u,v)$
- External influence: add all edges (u,v) with transmission probability ϵ
  - Captures external influence
  - E_ϵ: ϵ-edge set forms a clique
- q: #network edges in T, q': #ϵ-edges that transmit, s: #network edges that did not transmit, s': #ϵ-edges that did not transmit

$$\begin{aligned}\Pr[c|T] &= \beta^q \epsilon^{q'} (1-\beta)^s (1-\epsilon)^{s'} \prod_{(u,v) \in E_T} P_c(u,v) \\ &\approx \beta^q \epsilon^{q'} (1-\epsilon)^{s+s'} \prod_{(u,v) \in E_T} P_c(u,v)\end{aligned}$$

- Consider only the most likely propagation tree
  - $\Pr[\mathcal{C}|G] \approx \prod_{c \in \mathcal{C}} \max_{T \in \mathcal{T}_c} \Pr[c|T]$

NetInf approximation algorithm[29]

- Log-likelihood reformulation:
  $F_c(G) = \max_{T \in \mathcal{T}_c(G)} \log \Pr[c|T] - \max_{T \in \mathcal{T}_c(\bar{K})} \log \Pr[c|T]$, where $\bar{K}$ is the clique induced by E_ϵ
- Redefine $F_{\mathcal{C}}(G) = \sum_{c \in \mathcal{C}} F_c(G)$
- Modified problem: find
  $\hat{G} = \text{argmax}_G F_{\mathcal{C}}(G) = \sum_{c \in \mathcal{C}} \max_{T \in \mathcal{T}_c(G)} \sum_{(i,j) \in E_T} w_c(i,j)$

Theorem

$F_{\mathcal{C}}(G[W])$ is a submodular function of the set W of edges.

- Greedy algorithm for finding graph
- Heuristics to speed up

[29] M. Gomez-Rodriguez, J. Leskovec and A. Krause, *ACM Trans. Knowledge Discovery from Data, 2012*

Trace complexity of network inference

Trace/cascade: subgraph of infections

Trace complexity

Determine the smallest number of traces needed to infer the edge set of a graph.
Similarly, the smallest number of traces needed to infer properties of the graph.

- $\Omega(\frac{n\Delta}{\log^2 \Delta})$ traces are necessary and $O(n\Delta \log n)$ traces are sufficient to infer the edge set of a graph in the SIR model[30]
- Exact inference of trees using $O(\log n)$ traces
- Infer the degree distribution using $O(n)$ traces
- Similar results under other assumptions[31]

[30] B. Abrahao, F. Chierichetti, R. Kleinberg and A. Panconesi, *KDD*, 2013
[31] P. Netrapalli and S. Sanghavi, *ACM SIGMETRICS*, 2012

Inference problems: disease model parameter estimation

- Uncertainty about disease model parameters
  - Every flu season: uncertainty about transmissibility of flu strain

Disease parameter estimation problem

Given a network G and set I of infections, estimate transmission probability $\hat{\beta}$ such that

$$\hat{\beta} = \text{argmax}_{\beta} \Pr[\text{Set } I \text{ is infected in } G | \text{probability } \beta]$$

- Other parameters—incubation and infectious period– which affect dynamics
- Network only partially known
- Set I not exactly known
- Confounding with other diseases with similar symptoms

Model based reasoning approach[32]

- Nelder-Mead technique used for finding disease model from the current set of models that is closest
- Discussion of results in the section on surveillance

[32] E. Nsoesie, R. Beckman, S. Sashaani, K. Nagaraj and M. Marathe, *PLoS One, 2013*

Sensor sets for epidemics

Can one get information about an epidemic in the full population by monitoring carefully selected individuals (sensors)?

Sensor sets for epidemics

Let $t_{pk}(S)$ denote the time when the epicurve restricted to a set S "peaks"

Peak lead time maximization problem

Given parameters ϵ and k, network $G = (V, E)$, and the epidemic model, find a set S of nodes such that

$$S = \text{argmax}_S E[t_{pk}(V) - t_{pk}(S)]$$
$$\text{s.t.} \quad f(S) \geq \epsilon,$$
$$|S| = k,$$

where $f(S)$ = probability that the infection hits S

- Other objectives for sensor set
  - Early detection: first infection in S should happen close to the first infection in S
  - Identify peak: the peak of the epidemic restricted to S should happen before that in V
  - Parameter estimation: determine parameters, e.g., transmissibility of disease from characteristics in S
- More complex cost objectives for sensor set

Heuristic based on friends of random nodes[33]

- Pick a random subset R
- For each $v \in R$, pick a "friend" as part of sensor set S

[33] N. Christakis and J. Fowler, *PLoS One*, 2010

Heuristic based on dominator trees[34]

- Node x dominates a node y in a directed graph iff all paths to node y must pass through node x.
- Node x is the *unique immediate dominator* of y iff (i) x dominates y and (i) there exists no node z s.t. x dominates z and z dominates y.
- A node can have at most one immediate dominator, but may be the immediate dominator of any number of nodes.
- Dominator tree $D = (V^D = V, E^D)$ for $G = (V, E)$ is a tree s.t. $(u \to v) \in E^D$ iff u is the immediate dominator of v in G.

[34] H. Shao, K. S. M. T. Hossain, H. Wu, M. Khan, A. Vullikanti, B. A. Prakash, M. Marathe and N. Ramakrishnan, 2014

Improved lead time for peak

Dominator based heuristic performs better than Christakis-Fowler heuristic

Inference problems: so far

Missing information	Inference problem
Unknown initial conditions (source)	Source inference problem
Unknown network	Network inference problem
Unknown disease model parameters	Disease model inference problem

In reality: source, network, disease model are all unknown or noisy

Other research challenges

- More realistic formulations with partial surveillance information
- Sensors for detection and inference of epidemic characteristics
- Coevolution of behavior with network

1 Goals, History, Basic Concepts

2 Dynamics and Analysis

3 Inference problems

4 Control and optimization

5 Surveillance and Forecasting

6 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 $\mathcal{P}(\mathcal{S})$ 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

Strategies for controlling epidemics and objectives

- Different kinds of strategies[35]
  - 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
- Different objectives
  - Expected outbreak size
    - In the whole population and in different subpopulations
    - Other economic costs
  - Duration of epidemic
  - Size and time of peak
- Interventions can be modeled in networks as node deletions (vaccination), edge deletion (quarantining) and reducing β on edges

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

Illustrative formulations

- Vaccine allocation problems in the SIR model
- Optimal vaccination policies using ODE approach
- Vaccination strategies based on the spectral characterization
- Sequestration of critical populations

Effective vaccination allocation problem

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?

- 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[36] [37]
  - If initial infection is random: $O(\log n)$ approximation[38]

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

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

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

Bicriteria approximation algorithm for OVAP

Assume: transmission probability 1

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

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

- 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 k/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 $Exp[|E'|] \leq k/\epsilon$.

Compartmental differential equation based approach for OVAP[42]

- Age-structured differential equation model for H1N1
  - Mixing between age groups based on survey data[39]
  - $R_0 = 1.4$ for swine flu[40]
  - 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[41]
- CDC guidelines for swine-flu: prioritize vaccination for children 6 months to 5 years, and 5-18 years

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

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

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

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

Coupled differential equation model

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

- 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

Vaccination strategies based on the spectral characterization

- Recall: Epidemic dies out fast if $\rho(G) < T = \delta/\beta$ in the SIS model, where $\beta =$ transmission rate and $\delta =$ recovery rate
- Motivates: interventions that lower the spectral radius
- 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
- *Objective*: choose cheapest set $E' \subseteq E$ so that $\lambda_1(G[E - E']) \leq T$.

- Similarly: node and labeled versions
- NP-hard to approximate within a constant factor

Heuristics for SRM

- Eigenscore heuristic[43] [44]
  - Let $\mathbf{x}$ be the first eigenvector of $\mathbf{A}$, the adjacency matrix of G
  - For edge $e = (u, v)$, define score$(e) = x_u \cdot x_v$
  - Pick the top k edges with the highest score
  - Works well in real world graphs
- Product degree heuristic[44]
  - For edge $e = (u, v)$, define score$(e) = deg(u) \cdot deg(v)$
  - Pick the top k edges with the highest score
  - Works well in random graphs with assortativity

[43] H. Tong, B. A. Prakash, E. Eliassi-Rad, M. Faloutsos and C. Faloutsos, *CIKM*, 2012

[44] P. V. Mieghem, D. Stevanovic, F. F. Kuipers, C. Li, R. van de Bovenkamp, D. Liu, and H. Wang. *IEEE Transactions on Networking*, 2011

Greedy algorithm for SRM problem[45]

Some notation

- 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

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 $\frac{n(e, G[E \setminus E'])}{c(e)}$
  - $E' \leftarrow E' \cup \{e\}$

[45] S. Saha, A. Adiga, B. Aditya Prakash, A. Vullikanti, 2014

Analysis of greedy algorithm

Lemma

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

Similar bound for node version

Other results

- Analysis of vaccination strategies based on degrees[46]
- Vaccination schemes based on PageRank[47]
- Vaccination strategies in terms of the cut width of the graph[48]

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

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[49]
- Significantly outperforms random allocation

[49] 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

1 Goals, History, Basic Concepts

2 Dynamics and Analysis

3 Inference problems

4 Control and optimization

5 Surveillance and Forecasting

6 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
- Human mobility modeling with applications to disease surveillance

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' [50]

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.

[50] 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. [51]

- 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.

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

Was Google Flu Trends a pioneer?

Polgreen et al. [52] 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

[52] 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 [53]

- 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)

[53] 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? [54]

- 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

[54] 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 [55]

- 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

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

The Final Straw [56] [57]

- 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

David Lazer,[1,2*] Ryan Kennedy,[1,3*] Gary King,[4] Alessandro Vespignani[5,6,7]

Large errors in flu predictions were largely avoidable, which offers lessons for the use of big data.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplar of the "big data" (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict x has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses-driven research. To ensure that we know the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

ability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 mil-

run ever since, with a few changes announced in October 2013 (*10*, *11*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT has been higher than CDC estimates in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of errors depend on the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a predictor of the monitor is questionable. A study in 2011 reported that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*9*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data

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

[57] 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[58]

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

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

Designing your own vocabulary[59]

- 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.'

[59] 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 [60] and Lampos et al. [61] 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

[60] A Culotta, *Proceedings of the First Workshop on Social Media Analytics*, 2010
[61] V Lampos, N Cristianini, *ACM TIST*, 2012

Using Twitter during the H1N1 pandemic

Signorini et al. [62] 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

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

Getting into more detailed content analysis [63]

- 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.	*"China Reports First Case of Swine Flu (New York Times): A 36-year-old man who flew from St. Louis to Chengdu is.. http://tinyurl.com/nbfkrg"* *"Ways To Prevent Flu http://tinyurl.com/r4l4cx #swineflu #h1n1"*
Personal Experience	Twitter user mentions a direct (personal) or indirect (e.g., friend, family, co-worker) experience with the H1N1 virus or the social/economic effects of H1N1.	*"Swine flu panic almost stopped me from going to US, but now back from my trip and so happy I went :)!"* *"Oh we got a swine flu leaflet. clearly the highlight of my day"* *"My sister has swine flu!"*
Personal Opinion and Interest	Twitter user posts their opinion of the H1N1 virus/situation/news or expresses a need for or discovery of information. General H1N1 chatter or commentary.	*"More people have died from Normal Flu than Swine flu, its just a media hoax, to take people's mind off the recession"* *"Currently looking up some info on H1N1"* *"Swine flu is scary!"*
Jokes/Parody	Tweet contains a H1N1 joke told via video, text, or photo; or a humourous opinion of H1N1 that does not refer to a personal experience.	*"If you're an expert on the swine flu, does that make you Fluent?"*
Marketing	Tweet contains an advertisement for an H1N1-related product or service.	*"Buy liquid vitamin C as featured in my video http://fb.gd/y87r @health @h1n1"*
Spam	Tweet is unrelated to H1N1	*"musicmonday Mid lamondom Yom Kippur Polanski Jay-Z H1N1 Watch FREE online LATEST MOVIES at http://o.gd/b1S8d"*

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

Multi-level classification of tweets [64]

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

Even finer distinctions [65]

- 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

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

Atmospheric modeling [66]

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

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

Atmospheric modeling (contd) [67]

- 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

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

OpenTable reservation monitoring [68]

- 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

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

Monitoring Wikipedia usage [69]

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

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

Global disease monitoring with Wikipedia [70]

- 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

[70] 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 [71]

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

Putting it all together [72]

[72] 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

- Model level fusion versus data level fusion
- Accounting for initial, unreliable, estimates of official flu case counts
- Matrix factorization methods similar to those used in recommender systems research
  - Model

$$\begin{aligned}\widehat{\mathcal{M}}_{i,j} &= b_{i,j} + U_i^T F_j \\ &+ F_j |\mathcal{N}(i)|^{-\frac{1}{2}} \textstyle\sum_{k \in N(i)} (\mathcal{M}_{i,k} - b_{i,k}) x_k\end{aligned} \tag{4}$$

  - Fitting

$$\begin{aligned}b_*, F, U, x_* &= \operatorname{argmin}\Big(\sum_{i=1}^{m-1} \Big(\mathcal{M}_{i,n} - \widehat{\mathcal{M}}_{i,n}\Big)^2 \\ &+ \lambda_2 \Big(\sum_{j=1}^{n} b_j^2 + \sum_{i=1}^{m-1} ||U_i||^2 + \sum_{j=1}^{n} ||F_j||^2 + \sum_{k} ||x_k||^2\Big)\Big)\end{aligned} \tag{5}$$

Putting it all together - contd (2)

Which sources are most important?

Sources	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
All	**3.21**	**3.24**	**3.05**	**2.89**	**2.96**	**3.19**	2.87	**3.00**	**2.89**	**3.13**	**2.77**	**2.93**	**3.08**	**2.92**	**2.88**	**3.00**
w/o $\mathcal{W}$	2.91	2.99	2.77	2.71	2.61	2.59	2.66	2.69	2.49	2.78	2.62	2.87	2.60	2.43	2.67	2.69
w/o $\mathcal{H}$	3.04	2.85	2.89	2.56	2.81	2.77	2.61	2.75	2.75	2.82	2.57	2.75	2.51	2.87	2.71	2.75
w/o $\mathcal{T}$	2.92	3.14	2.95	2.61	2.72	2.81	2.88	2.79	2.61	2.93	2.74	2.63	2.79	2.74	2.81	2.80
w/o $\mathcal{S}$	3.19	3.11	2.92	2.64	2.69	2.70	**2.89**	2.88	2.78	3.07	2.75	2.91	2.80	2.71	2.86	2.86
w/o $\mathcal{F}$	3.20	3.12	2.88	**2.89**	**2.96**	**3.19**	2.87	**3.00**	2.83	3.02	**2.77**	**2.93**	2.98	2.88	**2.88**	2.96

- 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.

Recommendations for future forecasting programs [73]

- 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

[73] 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 [74]

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

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

Forecasting Global Epidemic Spread [75]

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

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

Human mobility modeling [76]

- (Anonymized) call data records (CDRs) provide a ready source of location information that sheds insight into human mobility patterns
- Data format: (time call/text was placed/received, cellular antenna location, cellular antenna direction)
- Key limitations
  1. Gathered only during active periods
  2. Coarseness of geographic resolution
- Has been used for studying work/travel patterns, carbon footprints

[76] R Becker, C Ramon, K Hanson, S Isaacman, JM Loh, M Martonosi, J Rowland, S Urbanek, A Varshavsky, C Volinsky, *CACM*, 2013

Characterizing human travel patterns [78]

100,000 anonoymized 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. [77] highlights that algorithms are capable of approaching the theoretical limits of predictability

[77] X Lu, E Wetter, N Bharti, AJ Tatem, L Bengtsson, *Nature Scientific Reports*, 2013
[78] M Gonzalez, CA Hidalgo, A Barabasi, *Nature*, 2008

Reconstructing high-resolution human contact networks [79][80]

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

[79] N Eagle, A Pentland, D Lazer, *PNAS*, 2009
[80] M. Salathe et al., *PNAS*, 2010

Mapping interactions using Twitter [81]

- 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%

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

A sobering study

Smallpox simulation under human mobility assumptions [82]

- 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 [83]

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

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

1 Goals, History, Basic Concepts

2 Dynamics and Analysis

3 Inference problems

4 Control and optimization

5 Surveillance and Forecasting

6 Putting it all together: theory to practice

Putting it all together: outline

- A real-world example: H1N1 Pandemic
- Data, Synthetic realistic social networks
- Detailed agent-based simulations
- Case studies
- Computational Ecosystems
- Extensions

Unfolding of a pandemic

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

Current example: Ebola outbreak in Africa

- Largest Ebola outbreak yet: 3 countries; 2000 cases; 1000 deaths.

Current example: Ebola outbreak in Africa

- Largest Ebola outbreak yet: 3 countries; 2000 cases; 1000 deaths.
- 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

Challenges[84]

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

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

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 perfromed (Where)
  - The reasons for doing the activties (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[85]

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

Step 2: Within host disease progression models.

Within host disease progression model

Disease Outcomes (depends on susceptibility)			
Name	Typical	Mild	Asymp
Fully_susceptible	0.73	0.25	0.02
Waned_natural1	0.12	0.35	0.53
Waned_natural2	0.09	0.27	0.64
Waned_natural3	0.00	0.25	0.75
Waning_vaccine1	0.28	0.28	0.44
Waning_vaccine2	0.27	0.27	0.46
Waning_vaccine3	0.25	0.25	0.50
Full_vaccination	0.5	0.25	0.25
Partial_vaccination	0.52	0.24	0.24

- 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

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

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.

Day 275
Median Infections: 20.7M47
Median of Simulations: Low Income Rural Area
Median of Simulations: Low Income Urban Area
Median of Simulations: Middle Income Rural Area
Median of Simulations: Middle Income Urban Area
Median of Simulations: High Income Rural Area
Median of Simulations: High Income Urban Area
Median of Simulations: Nationwide
Infections
Deaths
Hospitalizations
Ventilators
Antivirals
Vaccines
Medical Supplies
Hospital Beds
Ventilators
Medical Personnel

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/

CIEPI Cyber-infrastructure for computational epidemiology by NDSSL

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

ISIS User Interface by NDSSL

NDSSL epidemiology app ecosystem
ndssl.vbi.vt.edu/apps

VirginiaTech
Virginia Bioinformatics Institute

Modeling | Case Studies | Press/Videos | Publications | Applications | About us

Welcome to the Network Dynamics & Simulation Science Laboratory applications page. In our pursuit of advanced research in synthetic information systems, we have developed a number of integrative applications that help to answer questions posed by scientists, policy makers, and planners involved with very large complex systems. Our tools and applications support real world decisions in areas that include public health, transportation, urban planning, communication, community resilience, and much more. From tabletop exercises to classroom lessons, the breadth of our applications reaches a diverse audience.

ISIS | FluCaster | my4Sight | Granite

DBV | SIV United States | SIV International | GDSCalc

EpiViz | DSI:DC | Virus Tracker | TranSims

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[86]

- *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.

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

Extension 2: Non-communicable Diseases and social contagions[87]

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

[87] 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

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:

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.

Threshold model

- Large literature in statistical physics and social science: also referred to as *bootstrap percolation* and *complex contagion*
- More complex dynamical properties

Theorem (Characterization for Erdős-Rényi model[a])

[a] S. Janson, T. Luczak, T. Turova and T. Vallier, 2010

Let $G \in G(n,p)$. Let $a = \#$initial infections and $A^* = \#$final infections. Assume $r \geq 2$.

1. *(Varying a) Suppose $\frac{1}{n} < p < \frac{1}{n^{1/r}}$ There exists constant a_c, such that:*
   1. *a "smaller" than a_c implies $A^* = o(n)$*
   2. *a "larger" than a_c implies $A^* \geq n - o(n)$*
2. *(Varying p, fixed a) There exists a threshold p_c such that:*
   1. *$p \leq (1-\delta)p_c$ implies $A^* = o(n)$*
   2. *$p \geq (1+\delta)p_c$ implies $A^* \geq n - o(n)$*

SIR Model: network of communicating FSMs

Transition probabilities for a node depend on neighbors' states

Goal

Fast algorithm for exact stochastic simulation of the SIR model on a graph.

Example: SIR Process on a graph

$t = 0$

A simple instance:

1. Each infected node v is infectious for 2 steps
2. Infection spreads along edge (u, v) with probability $p(u, v)$
3. Discrete time process
4. Node 1 infected initially

Example

$t = 0$

$t = 1$

Probability $= p(1,3)(1 - p(1,2))$

Example

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

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

Example: $t = 4$

$E_{inf} = \{(1,3),(3,2)\}$

Output of one run of the SIR process

1. V_i: set of nodes that got infected in step i
2. V_∞: set of uninfected nodes
3. Solid edge: infection spreads
4. Dashed edge: no infection
5. E_{inf}: set of solid edges

SIR process captured by the tuple

$\mathcal{I} = ((V_0, V_1, \ldots, V_T, V_\infty), E_{inf})$

Example: Random run of SIR Process

$E_{inf} = \{(1,3),(3,2)\}$

$\mathcal{I} = ((V_0, V_1, V_2, V_\infty), E_{inf})$
Probability of this run

$$p_{SIR}(\mathcal{I}) = p(1,3)p(3,2)(1-p(1,2))^2 \cdot (1-p(2,4))^2(1-p(3,4))^2$$

Algorithm FastDiffuse

Given: Contact graph $G(V, E)$ and disease model

Step 0: Make every edge bi-directed

Step I: Generating a random weighted subgraph

Step II: Computing shortest paths in the random subgraph

Step III: Computing a dendogram (transmission tree) and the associted probability of this fixed point

Percolation and epidemics: Related Work

- Original motivation for percolation theory [Broadbent-Hammersley, 1957]
- Percolation and epidemic size: [Grassberger, 1983] [Sander, Warren, Sokolov, Simon, Koopman, 2002], ...
- Directed Percolation for analytically computing epidemic quantities in random contact graphs: [Meyers, Newman, Pourbohloul, 2006]

. . . this mapping gives us results only for the ultimate outcome of the disease in the limit of long times, in which all individuals are in either the S or R states, and no new cases of the disease are occurring. [Newman, SIAM Review, 2003]

Simplifying assumptions

Assumptions (for this talk alone):

1. Initially: only one node s infected, all other nodes susceptible
2. All nodes become recovered after getting infected (simple SIR model)
3. No incubation period
4. Symptomatic duration $S(v)$ for each node v
5. Uniform probability p of infection on each edge, invariant with time

More general disease models can be handled

Step I: Random Weighted Subgraph

Compute a random length $\ell(e)$ for each $e = (u, v) \in E$:

$$\ell(e) = \begin{cases} i \in \{1, \ldots, S(u)\}, & \text{with probability } (1 - p(e))^{i-1} p(e); \\ \infty, & \text{with probability } (1 - p(e))^{S(u)}. \end{cases}$$

Sample random length function. For each edge $e = (u, v)$ compute $\ell(e)$:

$$\ell(e) = \begin{cases} 1, & \text{with probability } p; \\ 2, & \text{with probability } (1 - p)p; \\ \infty, & \text{with probability } (1 - p)^2. \end{cases}$$

Step II: Shortest Paths

Compute the shortest path distance $dist_\ell(s, v)$ of each node $v \in V$ from s

- Let $V_t = \{v : dist_\ell(s, v) = t\}$, $V_\infty = \{v : dist_\ell(s, v) = \infty\}$.
- Let $E_{inf} = \{e = (u, v) : \ell(e) = dist_\ell(s, v) - dist_\ell(s, u)\}$ and $E_1 = \{e = (u, v) : dist_\ell(s, v) > dist_\ell(s, u)$ and either $dist_\ell(s, v) - dist_\ell(s, u) < \ell(e)$ or $dist_\ell(s, v) = \infty\}$.
- Output $\mathcal{I} = \mathcal{I}(\ell) = (\mathcal{P}, E_{inf})$ where $\mathcal{P} = (V_0, \ldots, V_T, V_\infty)$, and $T = \max_v\{dist_\ell(s, v) : dist_\ell(s, v) < \infty\}$.

Example: Step II

Different $\ell()$'s might result in the same $\mathcal{I}$

Several $\ell()$ functions might result in the same output $((V_0, V_1, V_2, V_\infty), E_{inf})$.

Step III: Probability

Compute the probability $p(\mathcal{I})$ of instance $\mathcal{I} = ((V_0, \ldots, V_T, V_\infty), E_{inf})$ in the following manner:

$$\begin{aligned} p_{\mathrm{FD}}(\mathcal{I}) &= \Pi_{e \in E_{inf}} (1 - p(e))^{\ell(e)-1} p(e) \cdot \\ &\quad \Pi_{e=(u,v) \in E_1} (1 - p(e))^{\min\{S(u), dist_\ell(s,v) - dist_\ell(s,u)\}} \end{aligned}$$

FastDiffuse on the example contact graph

$E_{inf} = \{(1,2),(3,2)\}$
$p_{SIR} = p^2(1-p)^6$

Lemma

FastDiffuse produces $\mathcal{I} = ((V_0, V_1, V_2, V_\infty), E_{inf})$ *with probability* $p^2(1-p)^6$

Algorithm FastDiffuse

Lemma

Let $\mathcal{I} = ((V_0, V_1, \ldots, V_T, V_\infty), E_{inf})$. The probability that FastDiffuse generates $\mathcal{I}$ is exactly equal to $p_{SIR}(\mathcal{I})$.

Lemma

FastDiffuse takes $O(|E| + |V| \log |V|)$ time to produce one run.

Proof sketch

Several $\ell()$ functions might result in the same output $((V_0, V_1, V_2, V_\infty), E_{inf})$.

Structural constraints on $\ell()$ producing same $((V_0, V_1, V_2, V_\infty), E_{inf})$

1. $\ell(1,3) = 1$
2. $\ell(3,2) = 1$
3. $\ell(1,2) = \infty$
4. $\ell(2,4), \ell(3,4) = \infty$
5. Any length for edges $(3,2)$, $(2,1)$, $(3,1)$, $(4,2)$ and $(4,3)$.

Let $\mathcal{L}$ denote the set of length functions $\ell()$ that result in $\mathcal{I} = ((V_0, V_1, V_2, V_\infty), E_{inf})$.

Goal: Prove that

$$\Pr[\ell \in \mathcal{L}] = p_{SIR}(\mathcal{I})$$

Proof

Consider $\ell() \in \mathcal{L}(((L_0, L_1, L_2, L_\infty), E'))$.

$$\begin{aligned}
\Pr[\ell()] &= p(1,3)p(3,2)(1-p(1,2))^2(1-p(2,4))^2(1-p(3,4))^2 \cdot \\
&\quad f(2,1)f(3,1)f(2,3)f(4,2)f(4,3) \\
&= \alpha f(2,1)f(3,1)f(2,3)f(4,2)f(4,3), \text{ where} \\
&\quad \alpha = p(1,3)p(3,2)(1-p(1,2))^2(1-p(2,4))^2(1-p(3,4))^2
\end{aligned}$$

Summing up the probabilities

$$
\begin{aligned}
\sum_{\ell()\in\mathcal{L}} \Pr[\ell] &= \alpha \sum_{(2,1)} f(2,1) \sum_{(3,1)} f(3,1) \sum_{(2,3)} f(2,3) \sum_{(4,2)} f(4,2) \sum_{(4,3)} f(4,3) \\
&= \alpha \cdot (p(2,1) + (1-p(2,1))p(2,1) + (1-p(2,1))^2) \cdot \ldots \\
&= \alpha
\end{aligned}
$$