Algorithmic Considerations in Modeling and Simulation of Large Societal Infrastructures

Madhav Marathe
mmarathe@vbi.vt.edu

Network Dynamics and Simulation Science Lab
NDSSL Technical Report NDSSL-07-001
Acknowledgements

Virginia Tech: Members, Network Dynamics & Simulation Science Laboratory, VBI
Karla Atkins, Chris Barrett, Richard Beckman, Keith Bisset, Deepti Chafekar, Jiangzhou Chen, Stephen Eubank, Bryan Lewis, V.S. Anil Kumar, Achla Marathe, Henning Mortveit, Paula Stretz

External Collaborators
Riko Jacob (ETH), Goran Konjevod (ASU), Martin Holzer & Dorothea Wagner (U. Karlsruhe), Aravind Srinivasan (UMD), Sri Parthasarathy (IBM), Nan Wang (Morgan Stanley), Ravi Sundaram, (Northeastern), Mayur Thakur (UMR), Harry Hunt, S. Ravi, Dan Rosenkrantz and Richard Steams (U. Albany), Zoltan Toroczkai (Notre Dame), NISAC and TRANSIMS members at Los Alamos National Laboratory

Work funded in part by NIGMS, NIH MIDAS program, CDC, Center of Excellence in Medical Informatics, DTRA, NSF, NeTs program, DoT, VT Foundation.
Modeling Very Large Complex Systems (VLCS): An Algorithmic Challenge

- **Chicago Area and Population**: 400 sq Miles, 12 counties, 272 townships, 9 million people, every school, every business, every household in Cook county
- **Transportation System**: ~4 million edges/nodes, ~31 million trips, every road segment, bus, transit & rail.
- **Social Network for Public Health**: ~20 million nodes in the temporal network with 1 sec resolution
- **Telecommunication System**: Million IP addresses, ~125 million calls/day

Novel Algorithmic Techniques are essential to analyze such large systems
A collection of interoperable simulations of societal infrastructures

each mimics the time-dependent interactions of every individual in a regional area with the built environment (critical infrastructures)

based on:
  - where they are,
  - why they are there,
  - what they are doing,
  - how they got there, and
  - who they went with
Socio-technical system analysis using today’s computing technology motivate new & interesting algorithmic questions

Algorithmic viewpoint in their can make meaningful difference
## Today’s Talk: Two Algorithmic Problems

<table>
<thead>
<tr>
<th></th>
<th>Routing/Path Finding</th>
<th>Simulating Epidemics like diffusion processes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Familiarity in CS community</strong></td>
<td>Classical, old problem in CS</td>
<td>New problem in CS community</td>
</tr>
<tr>
<td><strong>Problem type</strong></td>
<td>Static problem over a graph</td>
<td>Dynamic problem over a network</td>
</tr>
<tr>
<td><strong>Primary Application Areas</strong></td>
<td>Transportation and Social Networks</td>
<td>Public Health, Viral Marketing, Sensor network protocols</td>
</tr>
<tr>
<td><strong>Solution Method</strong></td>
<td>Combinatorial</td>
<td>Discrete simulations and combinatorial</td>
</tr>
<tr>
<td><strong>Novelty</strong></td>
<td>Labeling Constraints</td>
<td>Network representation rather than mean field approximation</td>
</tr>
</tbody>
</table>
Computational Epidemiology and Public Health
**Good News**

- Pandemic of the 1918 lethality is **unlikely**
- 1918 virus was uniquely pathogenic
- More mutations required to approach the lethality and transmissibility of the 1918 virus

**But ..**

- Cannot handle even a moderate outbreak
- Not enough vaccine or production capacity to immunize the population
- Not enough anti-viral medications
- Logistical Issues
  - Who should get the limited supplies?
  - NOT enough hospital beds
  - How to get the supplies to everyone
Public Health officials can act to “ensure” people use their common sense

- Close schools
- Quarantine sick households
- Use limited treatments on Infected households
- Reduce contact at work and in public
- Close large public gatherings
- Promote Telecommuting

False alarms, economic losses

This is a lot to ask of Society, they need to know it will work!!!
Simple ODE Models

• Partition the population into compartments
  – Susceptible
  – Infected
  – Removed

• Assume mass action (uniform mixing)

Also see May, Hethcote, Bailey Heesterbeek, Murray,...
Weakness of ODE Models

- ODEs lack agency and heterogeneity of contact structure
  - Complexity stems from interactions among many discrete actors
  - Each kind of interaction must be explicitly modeled
  - Refinement is difficult

- ODE models say very little about transients

- Use constants such as $R_0$ as if they are universal constants
  - E.g. what is the $R_0$ for smallpox?

- Human behavioral issues
  - Non-homogeneous compliance
  - Changes in the face of crisis

- Are not naturally suited for formulating and testing policies that can be implemented
  - E.g. reduce contact rates by 30% --- is an ambiguous statement
Alternative Approach: Simdemics
High Resolution Network Based Modeling

How do you start to assess how an infectious disease will be transmitted through an urban population and how different interventions will affect this process?

1. Create a synthetic population
   - Sampling Contingency Tables, Assignment Problems,
2. Derive a social network
   - Construction and analysis of large networks
3. Create a model of disease transmission
   - Design probabilistic timed finite state automata based on data
4. Study how the disease spreads over this network
   - Simulation of a diffusion process
5. Formulate and compare outcomes for the different policies
   - Optimal Selection of critical individuals, whom do we isolate, vaccinate, Implementable policies

Step 1: Synthetic Populations

- **Who**: People
  - Individuals
  - Household structure
  - Statistically identical to U.S. Census

- **What**: Activities
  - Activity sequences based on data from national travel survey
  - Includes travel between cities/countries

- **Where**: Locations
  - Disaggregate locations and infrastructure information
  - Locations assigned to activity sequence

Beckman et al. Tran. Science, NISS technical reports, Barrett et al. TRANSIMS technical reports
Step 2: Urban Social Contact Network

People Vertex:
- age
- household size
- gender
- income...

Location Vertex:
- (x,y,z)
- land use.
- Business type

Edge labels:
- activity type: shop, work, school
- (start time 1, end time 1)
- (start time 2, end time 2)
On each time step, each person’s state of health can change

- (S) susceptible -> infectious in t time steps with probability depending on health and duration of contacts in a social network
- (I) infectious -> removed or recovered and thus susceptible after t time step (dependency is stochastic and on demographics)
- (R) removed -> terminal state (could become Susceptible again)
Putting it all Together gives ..

Spatial and Temporal details on spread of disease at this scale and fidelity
Too Much Too Early
Over enthusiastic response, too expensive

Too Little Too Late
Lack of planning and not enough vaccination

Just Right
Proper planning, targeted vaccination and response

Allows us to Compare Different Policies
Yields Data that is not easy to Collect.

What individuals were doing when they were infected

Allows for more informative analysis

for example:
First person infected in a household
Initially node 1 is infected. Each node runs SIR model.

Node remains infected for 2 steps and then recovers.

At each step an infected node infects its uninfected neighbors with probability $p$ and this is done independently.

A node is infected if one or more of its neighbor successfully infects it.
Output of one run of the simulation

- $V_i =$ set of nodes that got infected at step $i$
- $V_\infty =$ set of uninfected nodes
- $E_{inf} =$ set of edges on which infection spreads (solid)
- Dashed edges: no infection

SIR process is captured by the tuple

$$\mathcal{I} = ((V_0, V_1, \ldots, V_T, V_\infty), E_{inf})$$
## Comparison of the Three Methods

<table>
<thead>
<tr>
<th></th>
<th>EpiSims</th>
<th>EpiSimdemics</th>
<th>EpiFast/FastDiffuse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Solution Method</strong></td>
<td>Discrete Event Simulation</td>
<td>Hybrid Simulation</td>
<td>Combinatorial</td>
</tr>
<tr>
<td><strong>Generality</strong></td>
<td>Most General</td>
<td>Parameterized</td>
<td>Less General</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>50 hours</td>
<td>4 hours</td>
<td>1 sec- 4 min</td>
</tr>
<tr>
<td>180 days</td>
<td>9M hosts &amp; 40 proc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dynamic Social Network</strong></td>
<td>Can work</td>
<td>Can work</td>
<td>Can’t work</td>
</tr>
<tr>
<td><strong>Disease transmission model</strong></td>
<td>Edge as well as vertex based</td>
<td>Edge as well as vertex based</td>
<td>Edge based only</td>
</tr>
<tr>
<td><strong>Disease Progression Latency</strong></td>
<td>No restriction</td>
<td>Reasonable time between infection and infectious</td>
<td>No restriction</td>
</tr>
</tbody>
</table>
Algorithm Idea for FastDiffuse: Directed Percolation & Shortest Path

- Step I: Construct a random weighted subgraph $H$ of the original Social Network $G$
- Step II: Shortest paths on random subgraph $H$
- Step III: Compute probability of each run

**Theorem:** Probability $P$ of producing a configuration $C$ using FastDiffuse is identical to simulating the SIR process. Time $\sim O(\text{Shortest Path computation})$

---

Case Studies and Experimental Analysis

• Studies in Vulnerability and Criticality using EpiFast
  - Chicago Network with 9 Million individuals
  - [Experiment 1] Phase transition: threshold for epidemic outbreak.
  - ✓[Experiment 2] Criticality: people who when isolated/vaccinated will help stop the spread
  - [Experiment 3] Sensitivity: impact on epidemic from various parameters.

• Case Studies to Support Policy Planning
  - Smallpox Study for OHS:
  - Disaster planning NISAC DHS Study
  - Pandemic Preparedness for civilians: DHHS NIH MIDAS project
  - Pandemic Preparedness for military: DoD Study (Ongoing)
FastDiffuse Performance

- **Setup**
  - 1GHz CPU + 1.8 GB Memory
  - Chicago Social Network: 9 Million individuals
  - One run using 40 CPUs takes 7.3 minutes.
  - Amortized time per run can be reduced greatly if Fast-Diffuse is run continuously.

- **Such large number of runs on a network consisting of millions of nodes were previously infeasible.**

<table>
<thead>
<tr>
<th></th>
<th>Phase transition</th>
<th>Criticality</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td># experimental cells</td>
<td>704</td>
<td>90</td>
<td>135</td>
</tr>
<tr>
<td>(configurations)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># iterations per</td>
<td>1000</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>configuration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total running time (hour)</td>
<td>198</td>
<td>7.2</td>
<td>18.6</td>
</tr>
<tr>
<td>Time per run (second)</td>
<td>1</td>
<td>5.7</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Vulnerability versus Criticality

- **Vulnerability**
  - Probability of becoming infected
  - Depends on where disease starts and spreading dynamics
- **Criticality**
  - Expected # infections stopped by removing vertex/edges
- Involve time, initial conditions, topology & dynamics
- General problem is computationally hard

Node 1 is most vulnerable and 7 most critical

Vulnerability distribution (gen1)
Few Open Questions

- Extending FastDiffuse to dynamic networks
- Structural Properties of Large Scale Social Networks
  - Expansion, Betweenness
- Stochastic Vaccination Policy problem:
  - **Given** graph G(V,E), edge transmission probabilities p(e), and some distribution x on the initial infected nodes
  - **Objective**: Choose a set of nodes S, |S| ≤ k, so that the expected number of infections in the epidemic on G[V-S] is minimized - [complement of the problem studied by Kempe, Kleinberg, Tardos KDD]
  - **Stochastic optimization version**: Choose sets S_i on step i of the process so that ∑_i |S_i| ≤ k, and the expected number of infections is minimized.
Routing in Inter-modal Urban Transportation Systems

How do we address questions related to congestion planning, ITS, Clean air act and economic analysis of adding roads, rail, etc.

1. Create synthetic population & assign activities

2. Route Planning
   • Constrained Shortest path problems

3. Traffic Micro-simulation
   • Simulating vehicular traffic: granular flows
   • Game theoretic problems (feedback loop with routing)

4. Case studies to support policy planning
Simulated traffic using TRANSIMS in downtown Portland

Vehicular adhoc network formed by using simulated traffic in downtown Washington DC
Mathematical Model

- Labels of edges $\in \Sigma$, finite alphabet (red, green, orange, yellow, blue)
- Label of path $P = \text{concatenation of edges on } P$
- Label of Path denotes a unique word $l(P) \in \Sigma^*$
- Modal preference specified as a formal language $R$
- Goal: Find shortest path $P$ from $s$ to $t$ s.t. the label $l(P) \in L(R)$
- Example: Reagan National to College park: $(b^* + y^*) g^*$

Mendelson and Wood SICOMP 91, Yannakakis PODS 01, Romuef IPL 87
1. Labeled transportation network interpreted as a NFA G.
   - Source $s_G$ (start state) destination $t_G$ (final state)
   - Edge labels describe the allowable moves.

2. Construct cross product of $M$ (mode choice NFA) and $G = (M \times G)$
   - Edge weights inherited from $G$

3. Find shortest path from start state $(s_G, s_M)$ to final state $(t_G, t_M)$ in $(M \times G)$
Applications of the General Framework

• **Using Transit:** Shortest path from x to using transit.
• **$k$-similar paths:** Given a path $P$, find a path $Q$ that does not overlap with $P$ on more than $k$ edges (ITS application)
• **Turn Complexity:** Find Path from $s$ to $t$ that has no more than 3 left turns (variant apparently used by UPS)
• **Trip Chaining & Business Closures:** Find optimal path to visit a set of locations (home, gas station, shop, post office, etc) taking into account closures

Algorithmic Improvements

- General formalism and computational results for routing with labeling constraints: \([R: \text{Regular expression and } G: \text{graph}]\)
  
<table>
<thead>
<tr>
<th></th>
<th>Regular</th>
<th>Context Free</th>
<th>Context Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Shortest path</em></td>
<td>(</td>
<td>R</td>
<td>\cdot</td>
</tr>
</tbody>
</table>

- Implicit construction of the product graph: crucial for scaling
  - Allows us to work with a single copy of the transport network
- Parameterized A* algorithm (with overdo factor \(\mu\)): works well
  - label of a vertex = distance to the current tree + \(\mu\) (Euclidean distance to destination)
  - 10-100 times speed up
- Number of other improvements
  - Sorting the edge lists so that edges for a given mode are contiguous
  - Thread based parallel implementation
Case Studies and Experimental Analysis

- **Linear Expressions**
  - Dallas Ft Worth Case Study 1997 & Portland Case Study 2002
  - Approx 450,000 Nodes and 3,000,000 edges
  - Dijkstra, A* and A* with overdo.
  - Expressions: $w^*$, $w_c w^*$, $w_t w^*$

- **Arbitrary regular-language**
  - Dijkstra, bidirectional Dijkstra, A* and A* with overdo.
  - Phoenix Metro Area
  - 140,000 nodes and 300,000 edges
  - Highways, primary roads, secondary roads, local/rural roads

Barrett et al SWAT, SICOMP, ESA02, JEA, DIMACS 06, ATMOS 02
Comparing A* and A*+Overdo

Trade-off: running time & quality of paths as a function of the overdo parameter. X axis: overdo factor from 0 to 100. Y axis represents three quantities on a log scale: (i) running time, (ii) maximum relative error, and (iii) fraction of plans with relative error greater than a threshold value (0%, 1%, 2%, 5%, 10%).
Further Work

- Engineering the Shortest Path Algorithm (DIMACS Shortest Path challenge)
  - Scaling to national scale networks
  - Improvements using heuristic methods and data structures such as containers, highway hierarchies, landmarks, reach etc. need to be investigated
  - Parallel algorithms?

---

Goldberg et al Alenex 05, Madduri et al Alenex 07 and DIMACS 06, Gutman, WAE, Sanders et al ESA 02, Wagner et al. ATMOS,
Where are we headed:
Simfrastructure & National Scale Modeling

- Grid Based, Service Oriented Distributed Modeling Framework
- Decision Support Tools are being integrated
Problem Domains I did not Cover

• Data Management
  – Methods to build efficient distributed data warehouses (both for input data as well as data generated by simulations)
  – Algorithms to support digital libraries

• Data Mining
  – Analyzing large social networks
  – Mining output of simulations to actively guide them

• Parallel & Distributed Algorithms
  – Resource management on Grids
  – Parallel algorithm development
Conclusions and Summary

- Algorithmic Viewpoint *matters* when analyzing socio-technical systems
- Generality/Extensibility of algorithms is *as useful* as its efficiency
- Policies and Human Behaviors can be cast as algorithms!
  - Closing schools: When? Who else is affected?
  - Evacuating an area: How will it be staged? Means of transportation? Where will congestion appear?

Questions?