Structural Aspects of Massive Social Networks

Stephen Eubank
Basic and Applied Simulation Sciences
Los Alamos National Laboratory
Outline

• Simulation

• Social network estimation

• Structural analysis and generation of large social networks
Epidemiological Simulation System: EpiSims

• One of a set of interoperable simulations of socio-technical systems,

• each mimics the time-dependent interactions of every individual in a regional area with the built environment

• based on:
  • where they are,
  • when they arrived,
  • what they are doing,
  • how they got there, and
  • who they went with
Urban Infrastructure Suite

Energy Sector
- Project: Interdependent Energy Infrastructure Simulation System (IEISS)
  - Load and generation information
  - Infrastructure capacity
  - Outage areas
  - People, locations, activities

Interacting Urban Infrastructure & Users
- Project: Urban Population Mobility Simulation Technologies (UPMoST)
  - User mobility
  - Constraints
  - Disease population
  - Telecommunication activities

Public Health Sector
- Projects: Epidemiological Simulations (EpiSims), Biological Surveillance (BSAFER)
  - People, locations, activities, infrastructure information

Financial Sector
- Project: Marketecture
- People, locations, activities

Transportation Sector
- Project: Transportation Analysis Simulation System (TRANSIMS)
  - People, vehicles, locations, device ownership

Telecommunications Sector
- Project: AdHopNet
  - Telecommunication activities
Why Agent Based?

- Mitigation strategies operate on individuals
- $\times$ complex group interactions $\checkmark$ simple individual behaviors $\checkmark$
- Conditions change $\rightarrow$ agent behaviors change producing demand-driven interdependencies

Direct representation not feasible (combinatorial explosion)
Demand-driven Interdependence

- Transport accident causes cell phone outage
- Evacuation causes traffic jams in unexpected places
- SARS in Toronto affects NY Chinese restaurants
Components of EpiSims

• Disease model
  – Within host (multiple manifestations)
  – Transmission

• Contact patterns
  – Interaction based computing
  – Inherently discrete
  – Irregular, anisotropic, inhomogeneous, …
Synthetic Population: Typical Household

7
$0
student
Typical Household’s Day

Carpool → Work → Lunch → Work → Carpool

Home → Shopping → Home

Car → Daycare → Car

Bus → School → Bus
Others Use the Same Locations
Time Slice of a Typical Household’s Day
People Adapt to Events
Structural Analysis Goals

• Prevent or efficiently stop outbreaks of disease
  – Understand structural properties relevant for epidemiology
  – Quickly generate realistic instances of social network
    • experimental testbed,
    • sensitivity testing,
    • anonymization,
    • generalization
Structural Analysis Problems

• What properties are relevant to epidemiology?
  – Degree distribution? Clustering?
  – Non-local measures of centrality

• What are the observed values?
  – NP hard problem, with very large N

• How to generate constrained random graphs?

• Need
  – efficient algorithms
  – for approximate solutions
  – with bounded error
What is “the” social network?

• Depends on definition of contact
  – Telecommunications (routing, collaboration/citation networks)
  – Sexual partnership (STDs)
  – Physical co-location (aerosol transmitted diseases)

• Dynamic: changes as people react to events

• Bipartite, labeled graph
Social Network: bipartite labeled graph

People (1.6 million)
Social Network: bipartite labeled graph

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

[Diagram showing bipartite graph with vertex attributes]
Social Network: bipartite labeled graph

Locations (1/4 million)
Social Network: bipartite labeled graph

Locations (1/4 million)

Vertex attributes:
- (x,y,z)
- land use
- ...

Locations (1/4 million)
Social Network: bipartite labeled graph
Social Network: bipartite labeled graph

Edge attributes:
- activity type: shop, work, school
- (start time 1, end time 1)
- (start time 2, end time 2)
- ...

[Diagram of a bipartite labeled graph showing nodes and edges with different attributes]
Social Network: explicitly time dependent form

Edge attributes:
- activity type: shop, work, school

[t₁, t₂]

[t₂, t₃]

[t₃, t₈₆₄₀₀]
Distance 2 induced graphs

- Edge in new graph iff vertices were at distance 2 in original graph
- Bipartite -> falls into 2 disconnected pieces
  - People-people (if simultaneously present)
  - Location-location (if person moved from one to other -> directed)
A Social Network: projection onto people

[t_1, t_2]

becomes

[t_1, t_2]
A Social Network: people-people graph
A Social Network: people-people graph
Social Network: location-location graph

\[ [t_1, t_2] \quad [t_2, t_3] \quad t_2 \]
Social Network: location-location graph
Social Network: location-location graph

\[ [t_1, t_2] \quad [t_2, t_3] \quad [t_3, t_{86400}] \]
Static graphs

- Still too big, take union of edge sets over time
  
  (identify vertices connected by dotted lines)

  - Operations don’t commute:
    - union first, then distance 2 gives more connectivity
  
  - Time scale & mode of transmission determine relevance
    - Incubation, infectivity duration
    - Aerosol, fomite, …
A Social Network: static people-people graph
Social Network: static location-location graph
Measurements on static graphs

- Measures of centrality:
  - Degree distributions,
    - Vaccinate hi degree? 2 reasons: vulnerability & mitigability – not always!
  - Clustering
    - Special construction: 1 person connects cliques
    - $k$ cliques gives $1/k$ clustering $\gg$ Erdos -Renyi
    - Tells something about 2nd cohort
  - Betweenness
    - # shortest paths
    - Need # paths of length $k$ to find out about $k$th cohort

- Expansion
  - Expands like tree, but has high clustering

- Dominating set
  - Greedy algorithm provably good: choose high degree locations first
Degree distributions for static graphs

People-people

Locations

slope ~ -2.8
Measurements on static graphs

- **Measures of centrality:**
  - Degree distribution,
    - Vaccinate hi degree? 2 reasons: vulnerability & mitigability – not always!
  - Clustering
    - Special construction: 1 person connects cliques
    - $k$ cliques gives $1/k$ clustering $>>$ Erdos -Renyi
    - Tells something about 2nd cohort
  - Betweenness
    - # shortest paths
    - Need # paths of length $k$ to find out about $k$th cohort

- **Expansion**
  - Expands like tree, but has high clustering

- **Dominating set**
  - Greedy algorithm provably good: choose high degree locations first
Clustering by degree for static graphs

People-people

Locations

![Graphs showing clustering coefficient for people-people and locations.](image)
Measurements on static graphs

- **Measures of centrality:**
  - Degree distributions,
    - Vaccinate hi degree? 2 reasons: vulnerability & mitigability – not always!
  - Clustering
    - Special construction: 1 person connects cliques
    - $k$ cliques gives $1/k$ clustering >> Erdos -Renyi
    - Tells something about 2nd cohort
  - Betweenness
    - # shortest paths
    - Need # paths of length $k$ to find out about $k$th cohort

- **Expansion**
  - Expands like tree, but has high clustering

- **Dominating set**
  - Greedy algorithm provably good: choose high degree locations first
Measures of Centrality

Most vulnerable vertices
Measures of Centrality, cont’d

Most *vulnerable* vertices

Best vertex to remove
Measures of Centrality, cont’d

High degree

High between-ness
Measurements on static graphs

• Measures of centrality:
  – Degree distributions,
    • Vaccinate hi degree? 2 reasons: vulnerability & mitigability – not always!
  – Clustering
    • Special construction: 1 person connects cliques
    • $k$ cliques gives $1/k$ clustering $>>$ Erdos-Renyi
    • Tells something about 2nd cohort
  – Betweenness
    • # shortest paths
    • Need # paths of length $k$ to find out about $k$th cohort

• Expansion
  – Expands like tree, but has high clustering

• Dominating set
  – Greedy algorithm provably good: choose high degree locations first
Local Network Topology

Household (clique) of 4
Local Network Topology

Contacts of people in the household
(distance 0 -> 1)
Local Network Topology

Contacts among the household’s contacts
(distance 1 - 1)
Local Network Topology

Contacts’ contacts
(distance 1 --> 2)
Local Network Topology

Distance 2 --> 2
Distance 2 -> 3
Distance 3 -> 3
Local topology variability

5 strangers’ contacts

5 infecteds’ contacts
Size of giant component vs degree of people vaccinated

Size of the largest component after each iteration

Degrees in decreasing order
Fraction of people in giant component vs fraction of locations closed
Measurements on static graphs

• Measures of centrality:
  – Degree distributions,
    • Vaccinate hi degree? 2 reasons: vulnerability & mitigability – not always!
  – Clustering
    • Special construction: 1 person connects cliques
    • $k$ cliques gives $1/k$ clustering $\gg$ Erdos -Renyi
    • Tells something about 2nd cohort
  – Betweenness
    • # shortest paths
    • Need # paths of length $k$ to find out about $k$th cohort

• Expansion
  – Expands like tree, but has high clustering

• Dominating set
  – Greedy algorithm provably good: choose high degree locations first
Dominating Set

Performance of the two fast greedy algorithms on the EpiSim graph

- FastGreedy-1
- FastGreedy-2
Generating Random Social Networks:
The Chung-Lu Model for the Full Bipartite Graph

• Given: expected degree \( d(v) \) for each \( v \)
• Edge between person \( p \) and location \( l \) added with prob. \( \alpha d(p) \cdot d(l) \)
• No additional constraints on path lengths or clustering
<table>
<thead>
<tr>
<th></th>
<th>Episims Network</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giant ppl comp size</td>
<td>1615813</td>
<td>1507054</td>
</tr>
<tr>
<td># edges</td>
<td>6060679</td>
<td>6065637</td>
</tr>
<tr>
<td>Avg ppl deg</td>
<td>3.7507</td>
<td>4.0227</td>
</tr>
<tr>
<td>Mean c.c.</td>
<td>0.6376</td>
<td>0.6161</td>
</tr>
<tr>
<td>Time for generation</td>
<td>&gt; 10 hrs</td>
<td></td>
</tr>
</tbody>
</table>
# Chung-Lu vs Episims data

<table>
<thead>
<tr>
<th></th>
<th>Episims Network</th>
<th>CL</th>
<th>Fastgen-1</th>
<th>Fastgen-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giant ppl comp size</td>
<td>1615813</td>
<td>1507054</td>
<td>1507148</td>
<td>1615803</td>
</tr>
<tr>
<td># edges</td>
<td>6060679</td>
<td>6065637</td>
<td>6060679</td>
<td>6060679</td>
</tr>
<tr>
<td>Avg ppl deg</td>
<td>3.7507</td>
<td>4.0227</td>
<td>4.0209</td>
<td>3.7507</td>
</tr>
<tr>
<td>Mean c.c.</td>
<td>0.6376</td>
<td>0.6161</td>
<td>0.6235</td>
<td>0.7021</td>
</tr>
<tr>
<td>Time for generation</td>
<td>10 hrs</td>
<td>&lt; 40 sec</td>
<td>&lt; 30 sec</td>
<td>O(&lt;E&gt;)</td>
</tr>
<tr>
<td></td>
<td>O(</td>
<td>P</td>
<td>*</td>
<td>L</td>
</tr>
</tbody>
</table>
Degree Distribution of locations in Bipartite Graph
Degree Distribution in people-people graph

[Graphs showing degree distribution (log-log) of the giant component of the people-people graph for different simulations and models.]
Clustering Coefficient
Summary

• Different graphs for different needs

• Metrics, “meter sticks”, rules of thumb needed

• Constrained, fast random graph generation needed