On the use of ACS data to construct synthetic populations

Samarth Swarup, Henning S. Mortveit, Ana Aizcorbe, Stephen G. Eubank, Madhav V. Marathe

Network Dynamics and Simulation Science Laboratory
Virginia Bioinformatics Institute
Virginia Tech
April 2015
Motivation: Detailed individual-based modeling of epidemics and other network phenomena

Need good representation of where people are at what times
Methodology overview for detailed network-based modeling of epidemics

- Create a synthetic base population
- Assign activity sequences (using e.g. CART trees) to each individual
- Assign a location to each activity of every person
- Derive a social contact network $G$
- Create a model of disease transmission
  - Design probabilistic timed finite state automata based on data
  - Simulate disease spreads over $G$
- Compute effects of interventions: co-evolution of $G$, behavior, policy and disease progression
Mapping from activities to social contacts

SYNTHETIC POPULATION
Demographic information, population densities, activity surveys and other data sources are fused by modeling and computation to construct a representation of the actual population and the people interactions.
Example: activity sequence induced contact network for Liberia (w/ long distance travel)
Data sources: American Community Survey

- Gives marginal information about some variables at household level.
- Variables used:
  - Householder’s age
  - Household income
  - Household size

- What we need:

For census tract 1, block group 2 of Los Alamos county, NM

<table>
<thead>
<tr>
<th>Householder’s age</th>
<th>Hsize</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;74</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>121</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>214</td>
</tr>
<tr>
<td>&gt;3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>134</td>
<td>94</td>
<td>46</td>
<td>46</td>
<td>36</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Generating a base population

- Use PUMS data (5% sample data)
  - A PUMA can contain multiple census block groups.
  - Gives detail information about household and person demographics.

<table>
<thead>
<tr>
<th>Hsize</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;74</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>11</td>
<td>9</td>
<td>3</td>
<td>26</td>
<td>64</td>
<td>42</td>
<td>157</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>108</td>
<td>122</td>
<td>48</td>
<td>80</td>
<td>61</td>
<td>18</td>
<td>448</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>135</td>
<td>274</td>
<td>156</td>
<td>85</td>
<td>22</td>
<td>6</td>
<td>706</td>
</tr>
<tr>
<td>&gt;3</td>
<td>0</td>
<td>3</td>
<td>65</td>
<td>76</td>
<td>40</td>
<td>10</td>
<td>3</td>
<td>197</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>257</td>
<td>470</td>
<td>283</td>
<td>231</td>
<td>157</td>
<td>69</td>
<td>1508</td>
</tr>
</tbody>
</table>

For PUMA containing census tract 1, block group 2 of Los Alamos county, NM
Generating a base population

- Use Iterative Proportional Fitting (IPF) Algorithm.
- Uses block group marginal information and PUMA data.
- Generates joint distribution for each block group in given PUMA.

<table>
<thead>
<tr>
<th>Hsize</th>
<th>15-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>&gt;74</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.003</td>
<td>0.141</td>
<td>0.061</td>
<td>0.020</td>
<td>0.047</td>
<td>0.063</td>
<td>0.000</td>
<td>0.336</td>
</tr>
<tr>
<td>3</td>
<td>0.009</td>
<td>0.228</td>
<td>0.178</td>
<td>0.086</td>
<td>0.065</td>
<td>0.030</td>
<td>0.000</td>
<td>0.594</td>
</tr>
<tr>
<td>&gt;3</td>
<td>0.000</td>
<td>0.003</td>
<td>0.022</td>
<td>0.022</td>
<td>0.016</td>
<td>0.007</td>
<td>0.000</td>
<td>0.069</td>
</tr>
<tr>
<td>Total</td>
<td>0.011</td>
<td>0.372</td>
<td>0.261</td>
<td>0.128</td>
<td>0.128</td>
<td>0.100</td>
<td>0.040</td>
<td></td>
</tr>
</tbody>
</table>

For census tract 1, block group 2 of Los Alamos county, NM

- Sample the required number households from PUMS data from the same category.
Assign Activities to the Base Population

- Data Used:
  - National Household Travel Survey
  - Activities are matched at household level
  - Matching synthetic households with survey households
  - Matching adults within household and assigning activities
  - Kids are assigned activities separately.
Matching Synthetic households with Survey households

- Select household demographic variables and create a binary tree.
- All survey households are assigned to one of the terminal nodes.
- Each synthetic household is mapped to a terminal node.
- A survey household is chosen from that terminal node to match to the synthetic household.
Assigning activities to the individuals

- For adults
  - One-to-one match is done between adults and activities are copied from survey members to synthetic members.
  - If synthetic household has more adults than survey household, the activities of the last adult survey member are copied as many time as required.
  - If survey household has more adults than synthetic household, the extra adults in survey household are ignored.
Assigning activities to the individuals

- For kids

1. Age?
   - 0-4
   - School?
     - Yes
       - 6
     - No
       - 7
   - 5-12
   - School?
     - Yes
       - 8
     - No
       - 9
   - 13-15
   - School?
     - Yes
       - 10
     - No
       - 11
   - 16-18
   - School?
     - Yes
       - 12
     - No
       - 13
Activity locations

- Data used:
  - Household structure (type of the building, capacity) i.e. single family household, duplex, apartment etc.
  - Street data from NAVTEQ/HERE, i.e. name, type of the road/street, length and other geometry info
  - Housing unit (home location) is assigned to a link of given category with probability proportional to its length.

California and Illinois
 Locate Activities

- Home activity already located at home location
- All activities of an individual is assigned a location within 60 miles of radius.
- Two types of activities
  - Anchor Activities - work and school
  - Non-anchor activities - all other activity types
Generate the Social Contact Network

- **Input:**
  - Person
  - Location
  - Activities

- **Output:**
  - 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)
Sub-location modeling

- Counts the number of households at each location and each household is assigned a different sub-location.
- For each location, count the number of activities(visits) for each activity type

<table>
<thead>
<tr>
<th>Activity type</th>
<th>Work</th>
<th>School</th>
<th>Shop</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>55</td>
<td>100</td>
<td>35</td>
<td>5</td>
</tr>
</tbody>
</table>

- A sub-location is bounded by a capacity based on the activity type

<table>
<thead>
<tr>
<th>Activity type</th>
<th>Work</th>
<th>School</th>
<th>Shop</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-location capacity</td>
<td>25</td>
<td>50</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

- Estimate the number of sub-locations required for each activity type at given location
- Update activities with sub-location information.
Methodology – Summary

- Demographic samples
- Demographic distributions

Base population w/ demographic variables

- Activity templates w/ decision trees
- Activity surveys
- Activity templates

Activity templates mapping to individuals

Location assignment to all activities

Location mixing model assignment

- Geographic region augmentation
- Zone neighbor map, zone attractor weight and global travel coefficients

Contact network Construction

- Activity locations w/ capacities and geo-coordinates
- Population density estimates
- Population/household counts

Residence locations w/ geo-coordinates

- Administrative region boundaries
- Residential road data
- Building type distributions
- Population/household counts

Residence mapping to individuals

- Population density estimates
- Population/household counts

Individuals mapped to households

- Demographic samples

Input data
- Work-flow
- Tested output data
- Tested and curated data

Modeling, analysis and visualization

Network Dynamics & Simulation Science Laboratory

Virginia Tech
Virginia Bioinformatics Institute
Verification and Validation
# Validation: Network Measures

<table>
<thead>
<tr>
<th>Country</th>
<th>ISO</th>
<th>Table Prefix</th>
<th>Population</th>
<th>Household</th>
<th>Home-loes</th>
<th>Work-loes</th>
<th>College-loe</th>
<th>Total Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guinea</td>
<td>324</td>
<td>GIN 324</td>
<td>11,521,656</td>
<td>2,375,532</td>
<td>2,375,532</td>
<td>250,096</td>
<td>6</td>
<td>55,069,905</td>
</tr>
<tr>
<td>Korea, Republic of</td>
<td>410</td>
<td>KOR 410</td>
<td>49,039,986</td>
<td>18,263,532</td>
<td>18,263,532</td>
<td>1,555,909</td>
<td>371</td>
<td>152,109,511</td>
</tr>
<tr>
<td>Liberia</td>
<td>430</td>
<td>LBR 430</td>
<td>4,092,310</td>
<td>844,066</td>
<td>844,066</td>
<td>88,016</td>
<td>20</td>
<td>22,202,262</td>
</tr>
<tr>
<td>Liberia</td>
<td>430</td>
<td>LBR 430 2GROUPS</td>
<td>4,092,310</td>
<td>844,066</td>
<td>844,066</td>
<td>85,395</td>
<td>25</td>
<td>22,502,300</td>
</tr>
<tr>
<td>Liberia</td>
<td>430</td>
<td>LBR 430 9GROUPS</td>
<td>4,092,310</td>
<td>844,066</td>
<td>844,066</td>
<td>85,395</td>
<td>25</td>
<td>20,144,766</td>
</tr>
<tr>
<td>Nigeria</td>
<td>566</td>
<td>NGA 566</td>
<td>175,288,000</td>
<td>36,087,439</td>
<td>36,087,439</td>
<td>2,525,551</td>
<td>88</td>
<td>938,891,578</td>
</tr>
<tr>
<td>Poland</td>
<td>616</td>
<td>POL 616</td>
<td>38,535,872</td>
<td>12,479,530</td>
<td>12,479,530</td>
<td>703,468</td>
<td>3,793</td>
<td>112,414,786</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>694</td>
<td>SLE 694</td>
<td>5,743,725</td>
<td>989,917</td>
<td>989,917</td>
<td>134,847</td>
<td>5</td>
<td>23,663,983</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>324</td>
<td>GIN 324</td>
<td>11,521,656</td>
<td>11,425,439</td>
<td>183,208,217</td>
<td>1</td>
<td>255</td>
<td>32.60</td>
<td>124.84</td>
<td>124.70</td>
<td>115,224</td>
<td>11,113,955</td>
<td>0.973</td>
<td>21</td>
<td>1,457,330,472</td>
<td>0.67</td>
</tr>
<tr>
<td>410</td>
<td>KOR 410</td>
<td>49,039,986</td>
<td>47,484,264</td>
<td>537,131,527</td>
<td>1</td>
<td>133</td>
<td>22.62</td>
<td>554,972</td>
<td>42,605,860</td>
<td>0.897</td>
<td>30</td>
<td>2,210,000,000</td>
<td>0.62</td>
<td>1,457,330,472</td>
<td>0.67</td>
</tr>
<tr>
<td>430</td>
<td>LBR 430</td>
<td>4,092,310</td>
<td>4,084,569</td>
<td>84,789,847</td>
<td>1</td>
<td>249</td>
<td>41.52</td>
<td>125.48</td>
<td>125.35</td>
<td>14,073</td>
<td>4,031,059</td>
<td>0.992</td>
<td>18</td>
<td>720,629,723</td>
<td>0.59</td>
</tr>
<tr>
<td>430</td>
<td>LBR 430 2GROUPS</td>
<td>4,092,310</td>
<td>4,077,272</td>
<td>87,255,111</td>
<td>1</td>
<td>254</td>
<td>42.80</td>
<td>126.83</td>
<td>126.41</td>
<td>10,106</td>
<td>4,053,906</td>
<td>0.994</td>
<td>16</td>
<td>824,000,000</td>
<td>0.59</td>
</tr>
<tr>
<td>430</td>
<td>LBR 430 9GROUPS</td>
<td>4,092,310</td>
<td>4,077,426</td>
<td>78,830,117</td>
<td>1</td>
<td>250</td>
<td>38.67</td>
<td>111.97</td>
<td>111.93</td>
<td>10,014</td>
<td>4,034,303</td>
<td>0.994</td>
<td>17</td>
<td>679,000,000</td>
<td>0.59</td>
</tr>
<tr>
<td>566</td>
<td>NGA 566</td>
<td>175,288,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>616</td>
<td>POL 616</td>
<td>38,535,872</td>
<td>27,523,619</td>
<td>245,114,775</td>
<td>1</td>
<td>104</td>
<td>17.81</td>
<td>65.53</td>
<td>65.46</td>
<td>667,720</td>
<td>23,698,455</td>
<td>0.861</td>
<td>43</td>
<td>1,212,865,142</td>
<td>0.62</td>
</tr>
<tr>
<td>694</td>
<td>SLE 694</td>
<td>5,743,725</td>
<td>5,733,911</td>
<td>93,734,286</td>
<td>1</td>
<td>164</td>
<td>32.69</td>
<td>97.84</td>
<td>97.81</td>
<td>10,549</td>
<td>5,704,544</td>
<td>0.995</td>
<td>16</td>
<td>683,523,753</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Validation: Networks and Measures

Guinea

Liberia

Sierra Leone
Synthetic populations form a foundation for detailed, interaction-based simulation models for processes over coupled networks with humans in the loop.

This is a technology that we use in the lab for many different projects. Examples: epidemics in urban populations; evacuation scenarios; transportation analyses;

Synthetic populations offer great flexibility and can handled a broad range of policy- and what-if analyses often without changes to the model.
Data provided by the ACS is a cornerstone in our construction process for the U.S.

ACS data permits us to connect many other data sources (e.g. NHTS) with a demographic component.

Through our approach, we obtain a natural coordinate system for this type of information.

Our modeling approach naturally integrates anonymized data – we do not require access to original data. Obtaining aggregated data (distributions) and anonymized samples is sufficient. Sensitive data never has to be given to us in any form.
Contact information

- Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech
- Web: [http://www.vbi.vt.edu/ndssl](http://www.vbi.vt.edu/ndssl)

Thank you!