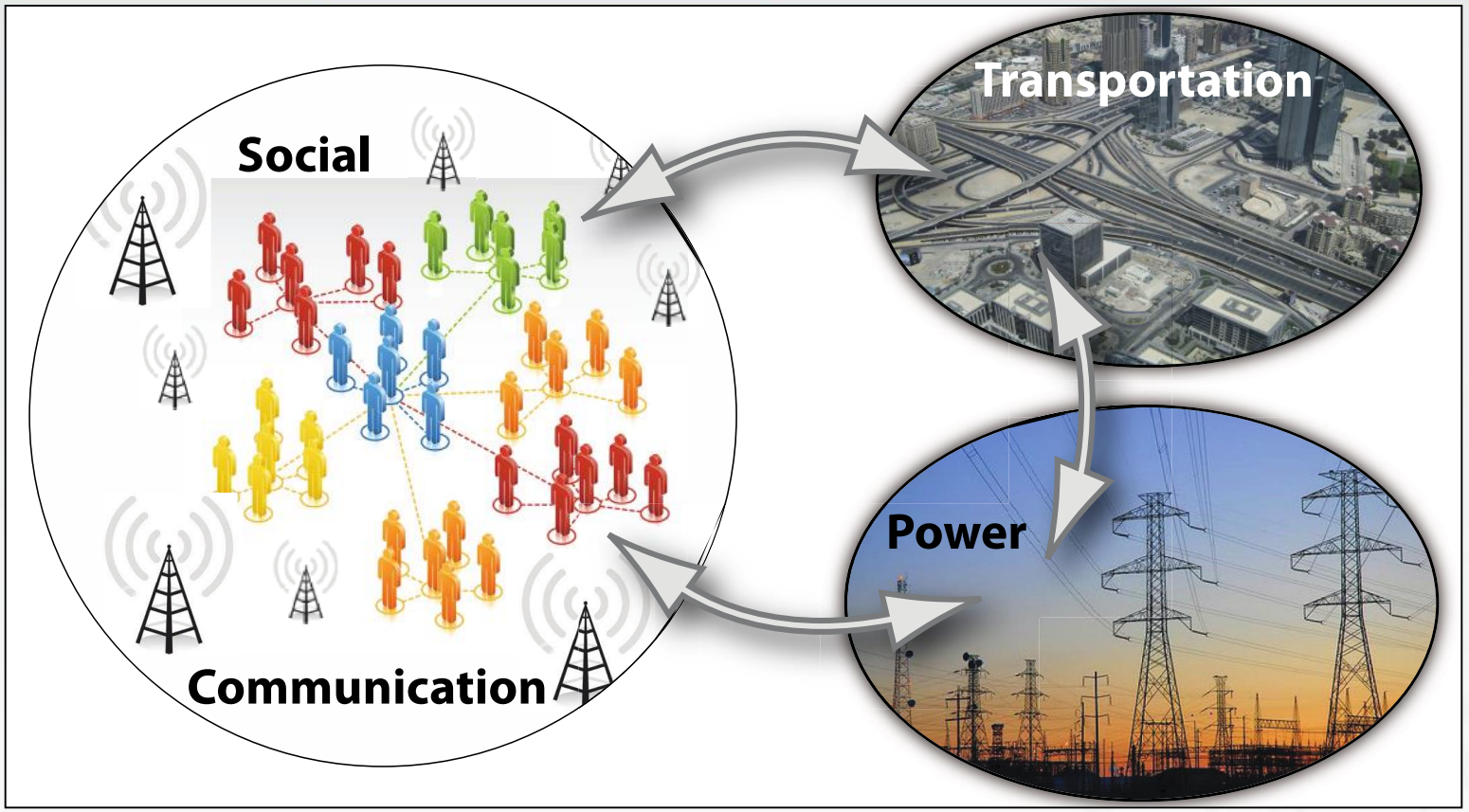
On the use of ACS data to construct synthetic populations

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

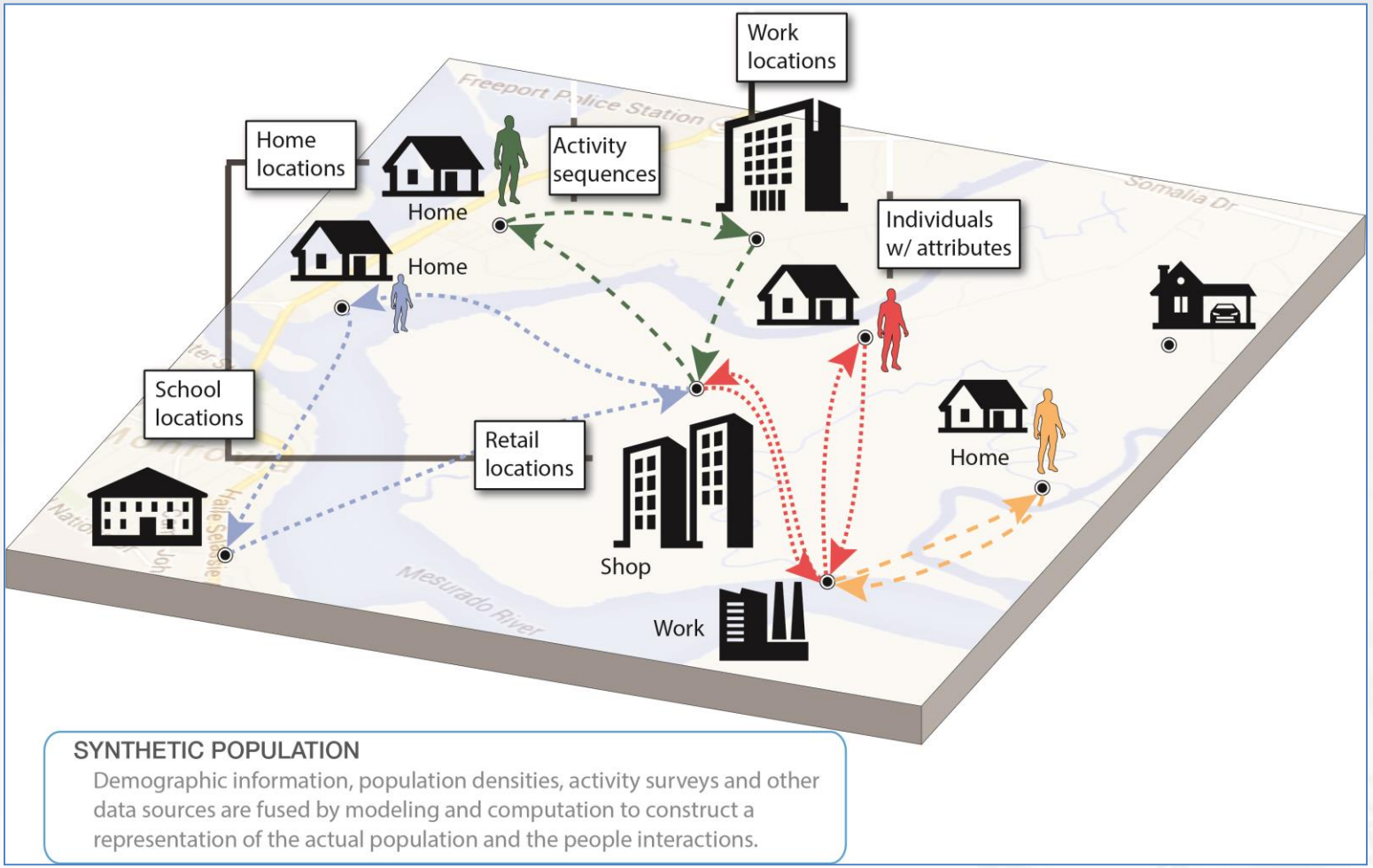


Need good representation of where people are at what times

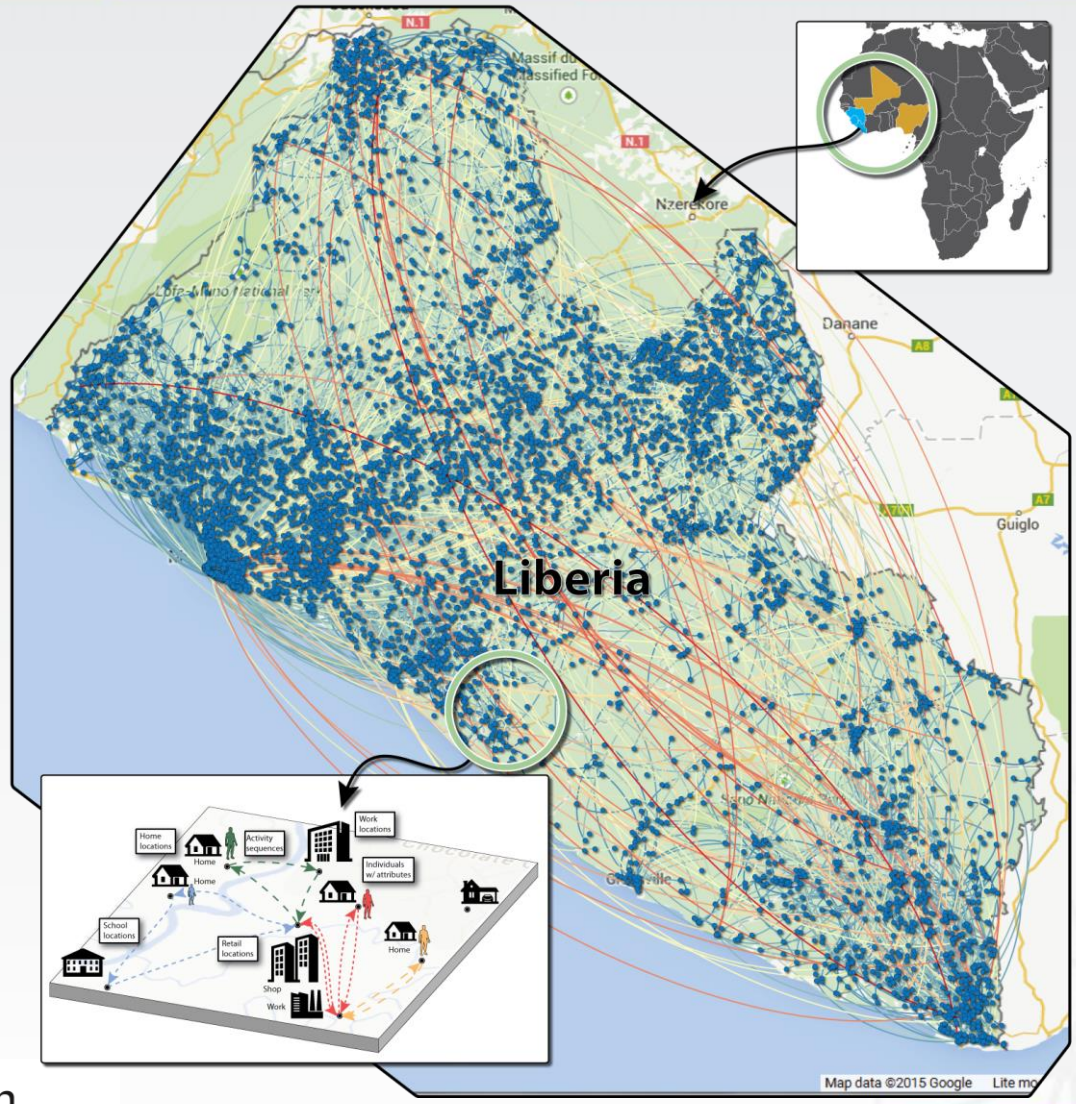
Methodology overview for detailed network-based modeling of epidemics

- 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



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**

The number of Households with household size in given ranges				
Hsize	1	2	3	>3
n	0	121	214	25

The number of Households with age in given ranges							
Age	15-24	25-34	35-44	45-54	55-64	65-74	>74
n	44	134	94	46	46	36	0

- What we need:

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

Householder's age								
Hsize	15-24	25-34	35-44	45-54	55-64	65-74	>74	Total
1	?	?	?	?	?	?	?	0
2	?	?	?	?	?	?	?	121
3	?	?	?	?	?	?	?	214
>3	?	?	?	?	?	?	?	25
Total	44	134	94	46	46	36	0	

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.

Householder's age								
Hsize	15-24	25-34	35-44	45-54	55-64	65-74	>74	Total
1	2	11	9	3	26	64	42	157
2	11	108	122	48	80	61	18	448
3	28	135	274	156	85	22	6	706
>3	0	3	65	76	40	10	3	197
Total	41	257	470	283	231	157	69	1508

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.

	Householder's age							
Hsize	15-24	25-34	35-44	45-54	55-64	65-74	>74	Total
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.003	0.141	0.061	0.020	0.047	0.063	0.000	0.336
3	0.009	0.228	0.178	0.086	0.065	0.030	0.000	0.594
>3	0.000	0.003	0.022	0.022	0.016	0.007	0.000	0.069
Total	0.011	0.372	0.261	0.128	0.128	0.100	0.040	

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

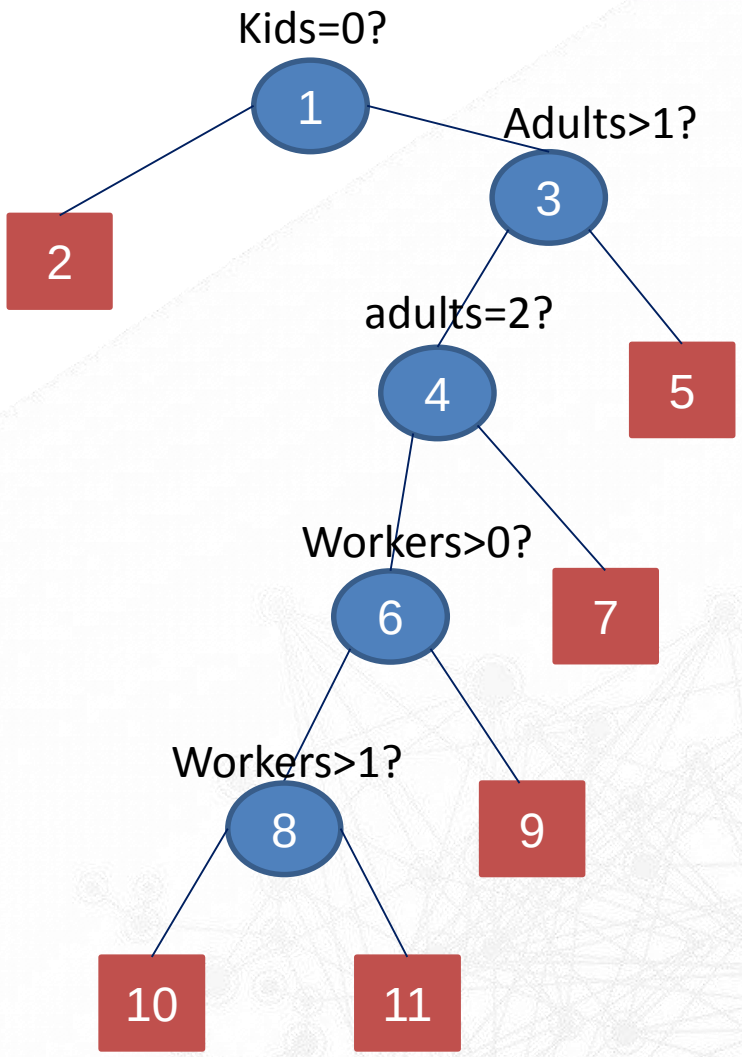
- Sample the required number households from PUMS data from the same category.

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.

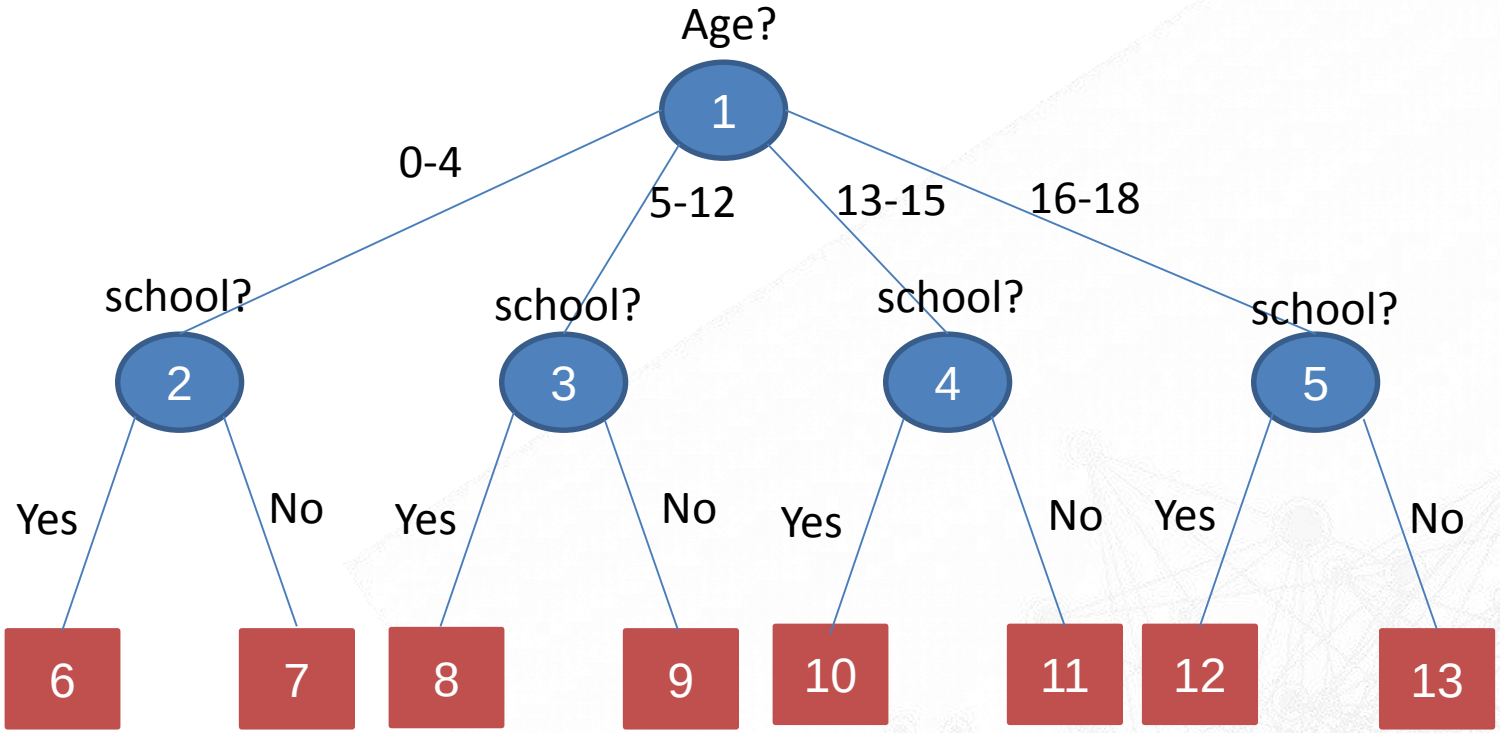


Assigning activities to the individuals

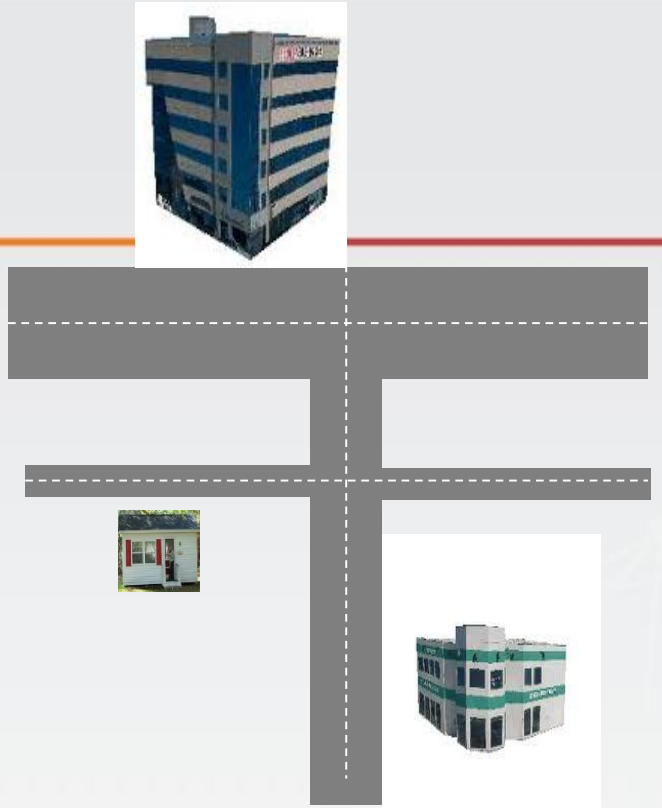
- For adults
  - One-to-one match is done between adults and activities are copied from survey members to synthetic members.
  - If synthetic household has more adults than survey household, the activities of the last adult survey member are copied as many time as required.
  - If survey household has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

Assigning activities to the individuals

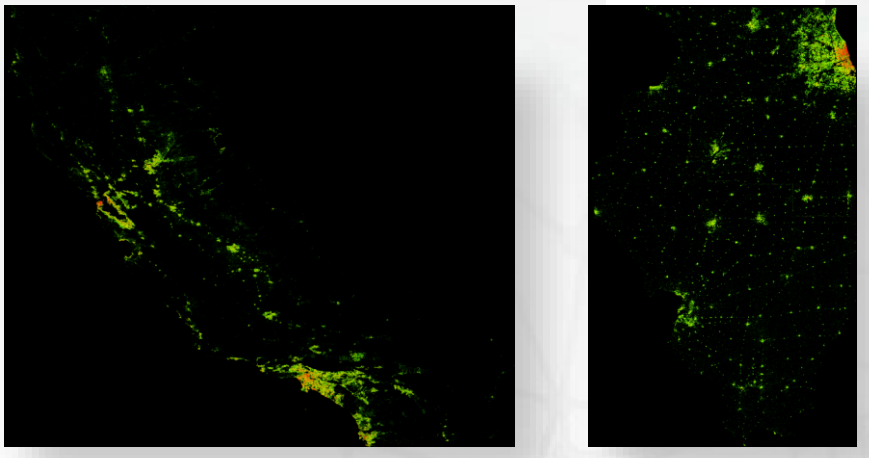
- For kids



Activity locations



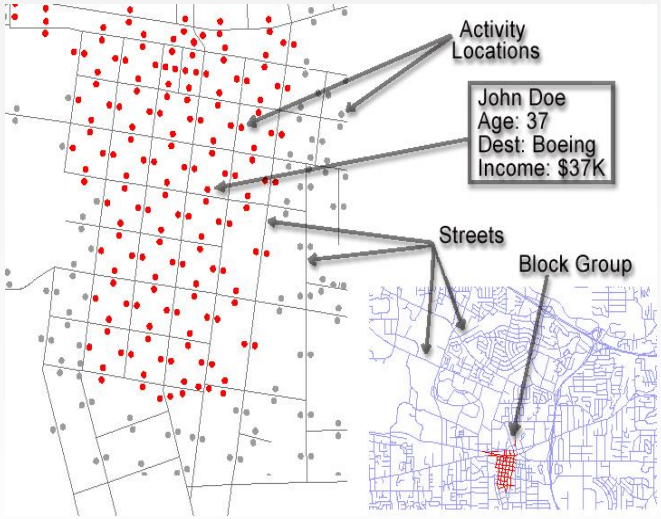
- Data used:
  - Household structure (type of the building, capacity) i.e. single family household, duplex, apartment etc.
  - Street data from NAVTEQ/HERE, i.e. name, type of the road/street, length and other geometry info
- Housing unit (home location) is assigned to a link of given category with probability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



California and Illinois

Locate Activities

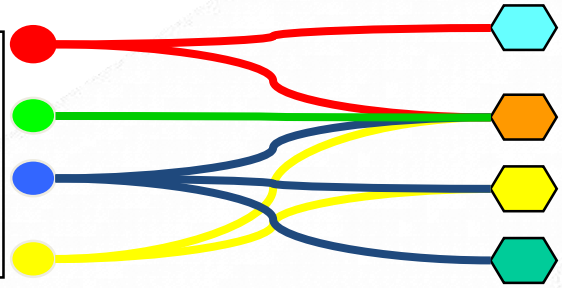
- Home activity already located at home location
- All activities of an individual is assigned a location within 60 miles of radius.
- Two types of activities
  - Anchor Activities - work and school
  - Non-anchor activities - all other activity types



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

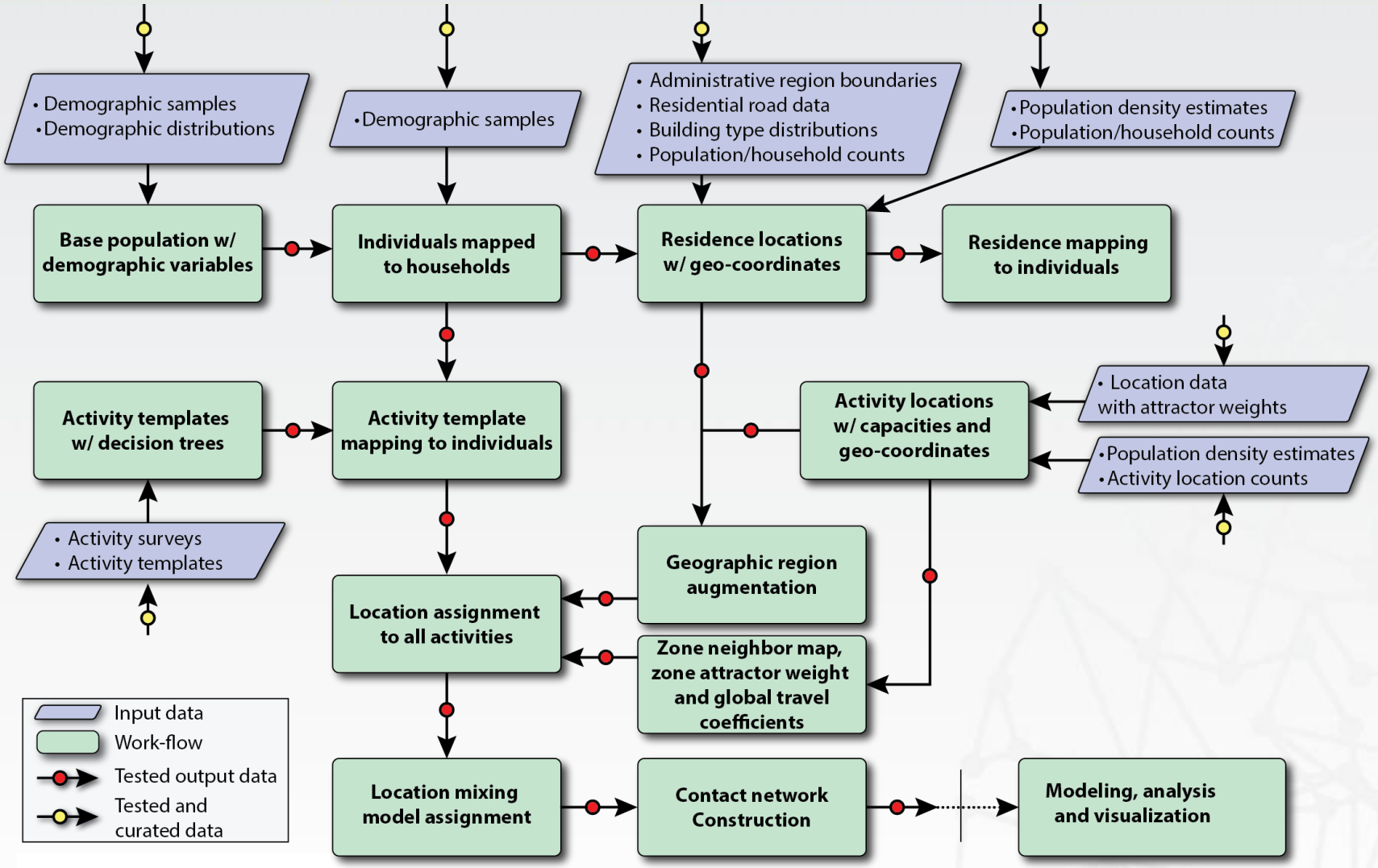
Activity type	Work	School	Shop	Other
location	55	100	35	5

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

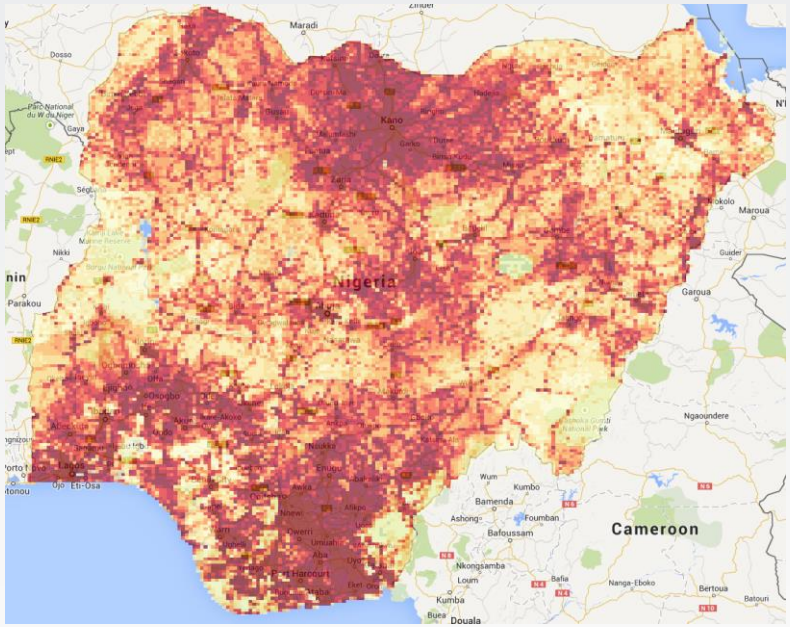
Activity type	Work	School	Shop	Other
Sub-location capacity	25	50	10	10

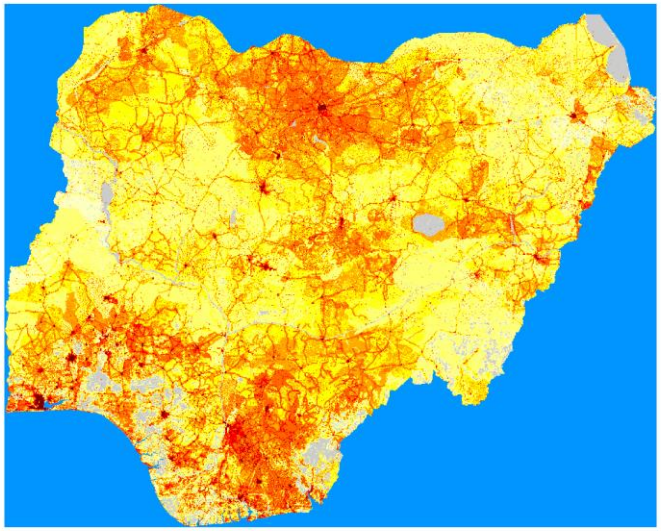
- Estimate the number of sub-locations required for each activity type at given location
- Update activities with sub-location information.

Methodology – Summary



Verification and Validation



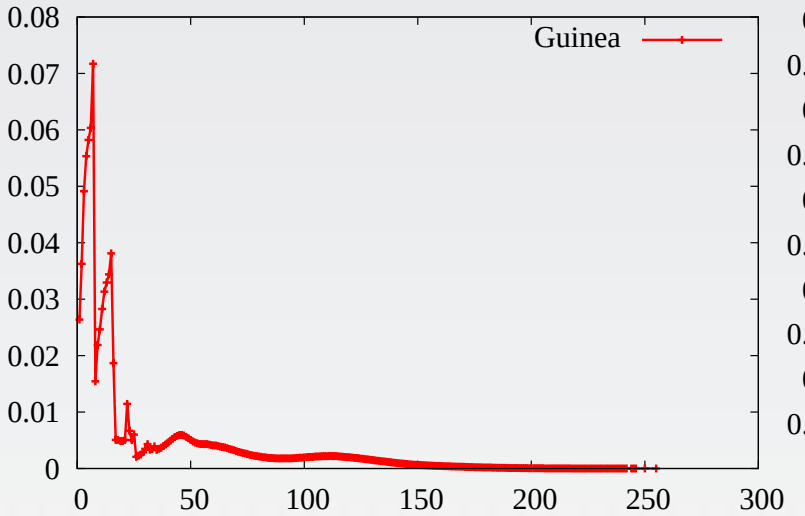


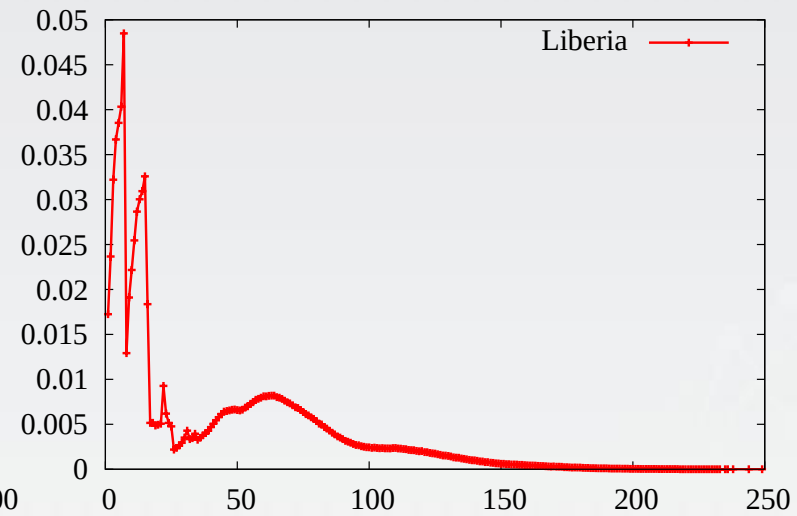
Validation: Network Measures

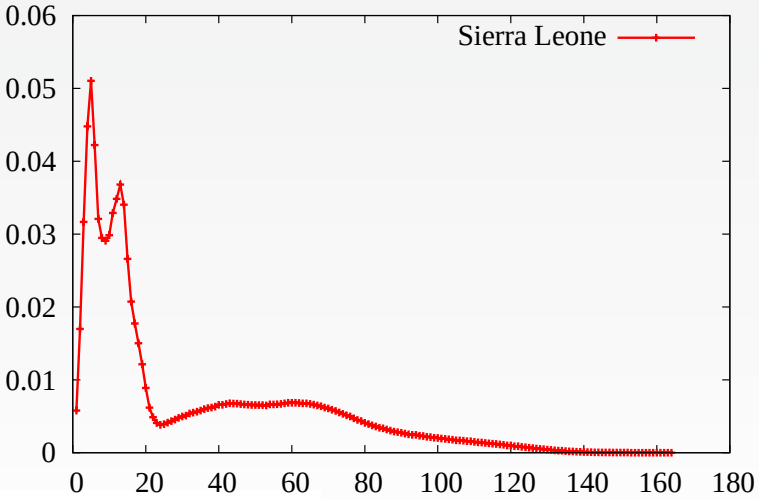
Country	ISO	Table Prefix	Population	Household	Home-locs	Work-locs	College-loc*	Total Activities
Guinea	324	GIN 324	11,521,656	2,375,532	2,375,532	250,096	6	55,069,905
Korea, Republic of	410	KOR 410	49,039,986	18,263,532	18,263,532	1,555,909	371	152,109,511
Liberia	430	LBR 430	4,092,310	844,066	844,066	88,016	20	22,202,262
Liberia	430	LBR 430 2GROUPS	4,092,310	844,066	844,066	85,395	25	22,502,300
Liberia	430	LBR 430 9GROUPS	4,092,310	844,066	844,066	85,395	25	20,144,766
Nigeria	566	NGA 566	175,288,000	36,087,439	36,087,439	2,525,651	88	938,801,578
Poland	616	POL 616	38,535,872	12,479,530	12,479,530	703,468	3,793	112,414,786
Sierra Leone	694	SLE 694	5,743,725	989,917	989,917	134,847	5	23,663,983

ISO	Table Prefix	Population	Vertices	Edges	Min deg.	Max deg.	Avg deg.	Lambda_1	Lambda_2	Comps.	Largest Comps.	Largest Comps. Ratio	Diameter	Triangles	CC
324	GIN 324	11,521,656	11,425,439	183,268,217	1	255	32.69	124.84	124.70	115,224	11,113,955	0.973	21	1,457,330,472	0.67
410	KOR 410	49,039,986	47,484,264	537,131,827	1	133	22.62			554,972	42,605,860	0.897	30	3,210,000,000	0.62
430	LBR 430	4,092,310	4,084,569	84,789,847	1	249	41.52	125.48	125.35	14,073	4,051,099	0.992	18	720,629,723	0.59
430	LBR 430 2GROUPS	4,092,310	4,077,272	87,255,911	1	254	42.80	126.83	126.41	10,106	4,053,906	0.994	16	824,000,000	0.59
430	LBR 430 9GROUPS	4,092,310	4,077,426	78,830,017	1	250	38.67	111.97	111.93	10,014	4,054,303	0.994	17	679,000,000	0.59
566	NGA 566	175,288,000													
616	POL 616	38,535,872	27,523,619	245,114,775	1	104	17.81	65.53	65.46	667,730	23,698,455	0.861	43	1,212,865,142	0.62
694	SLE 694	5,743,725	5,733,911	93,734,286	1	164	32.69	97.84	97.81	10,543	5,704,544	0.995	16	683,523,753	0.61

Validation: Networks and Measures







Conclusion (1/2)

- 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.

Conclusion (2/2)

- 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

Thank you!