Generating Synthetic Populations for Social Modeling

Tutorial at AAMAS 2017, São Paulo, Brazil

Samarth Swarup & Madhav V. Marathe,

Network Simulation Science and Advanced Computing Division,

Biocomplexity Institute and Initiative,

University of Virginia.

{swarup, marathe}@virginia.edu

Technical report NDSSL-TR-17-073

With contributions by: Abhijin Adiga, Richard Beckman, Jiangzhuo Chen, Youngyun Chungbaek, Stephen Eubank, Kristian Lum, Achla Marathe, Henning Mortveit, Yihui Ren, Srinivasan Venkatramanan

Special thanks to Marta C. González, MIT, for graciously sharing her work and recent slides

Acknowledgements

- Thanks to members of the Network Dynamics and Simulation Science Laboratory, BI and Discovery Analytics Center (DAC), both at Virginia Tech and our collaborators at SUNY Albany, UMD, Harvard and others.
- 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.

Slides for tutorial at AAMAS, São Paulo, Brazil, May 9, 2017

The current version of the slides are available at:
http://people.virginia.edu/~ss7rs/synthetic_population_tutorial_2

Suggested citation:
Samarth Swarup and Madhav V. Marathe. *Generating Synthetic Populations for Social Modeling: Second Tutorial* at the Sixteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS). São Paulo, Brazil, May 9 2017. Biocomplexity Institute, University of Virginia, Technical Report NDSSL-TR-2017-17-073.

Section

GOALS OF THE TUTORIAL

Objectives

- Overview the state of the art in synthetic populations
  - Discuss how to synthesize populations, networks and more broadly information
  - Discuss some of the challenges in the field
  - Illustrate how advances in computation and data are useful in advancing the field.
  - Provide audience an interesting topical area to work in.
- Describe open problems and challenges in the area
- Does not claim to be extensive. Due to time constraints some topics will not be covered
  - E.g. Detailed discrete choice theory, details of applications where these populations have been useful.
  - References are provided for additional reading

Section

POPULATION SYNTHESIS

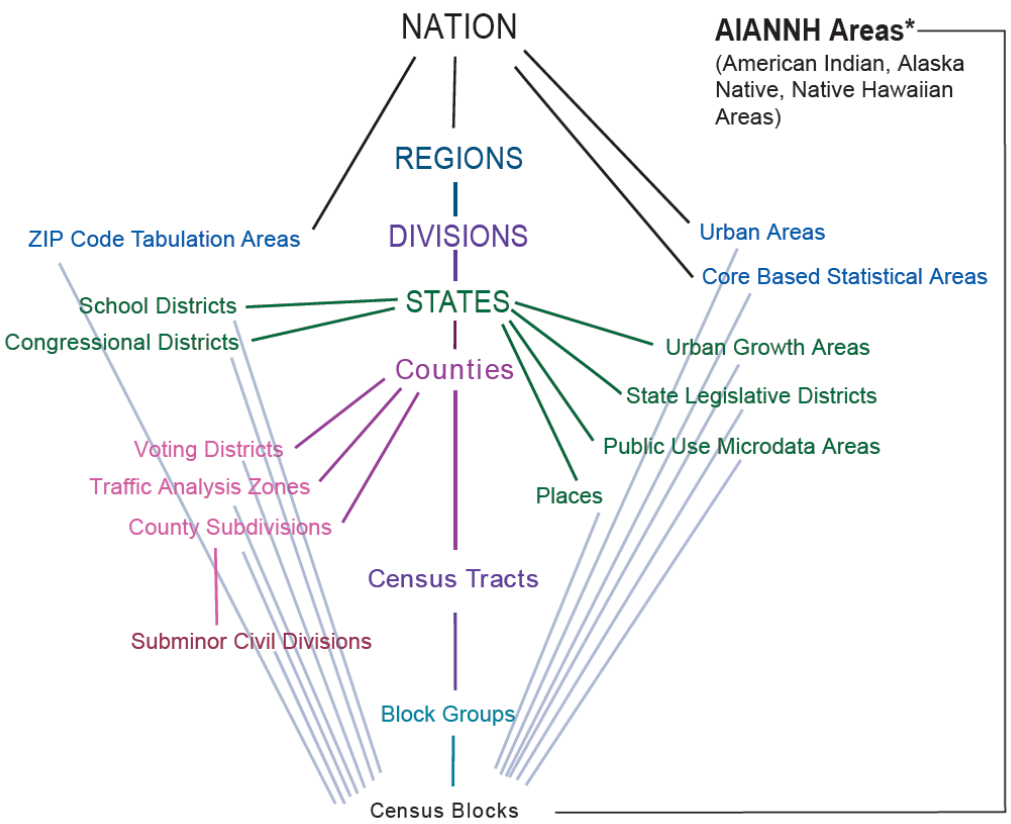
Goal: To generate a population of agents with realistic demographic attributes

Input:
- Distributions over demographics (*marginal distributions),*
- A sample of census records

Method: Iterative Proportional Fitting (IPF)

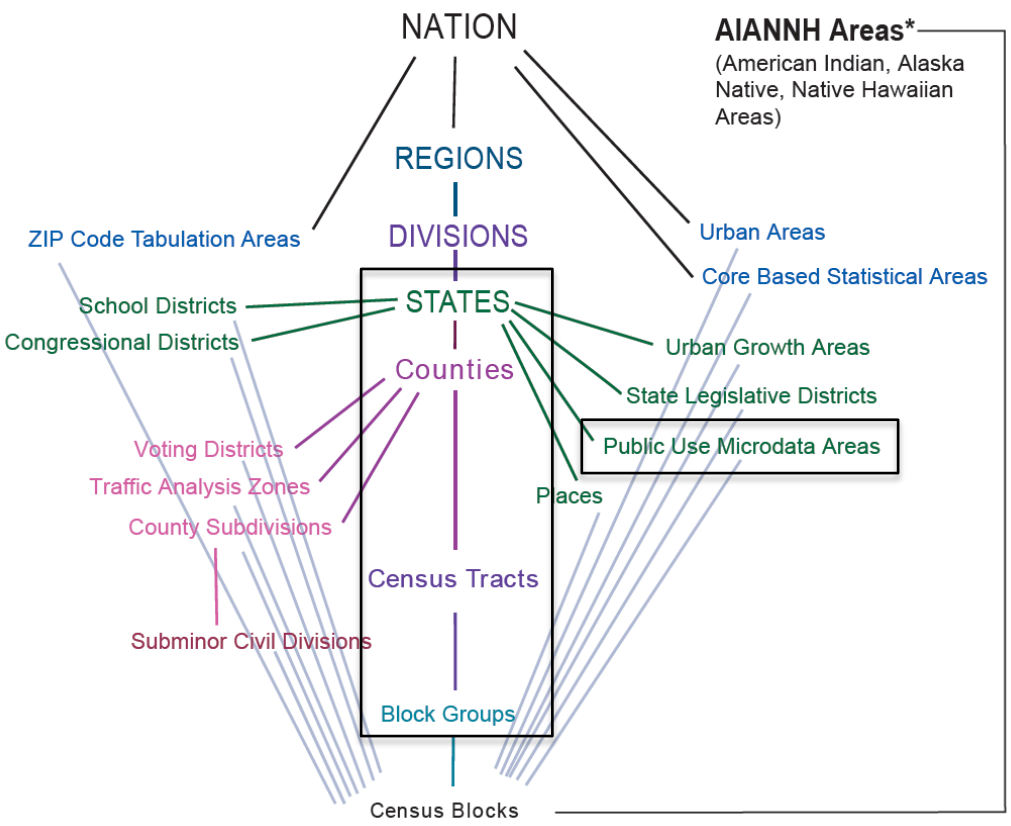
Geographical resolution

Standard hierarchy of US Census geographic entities



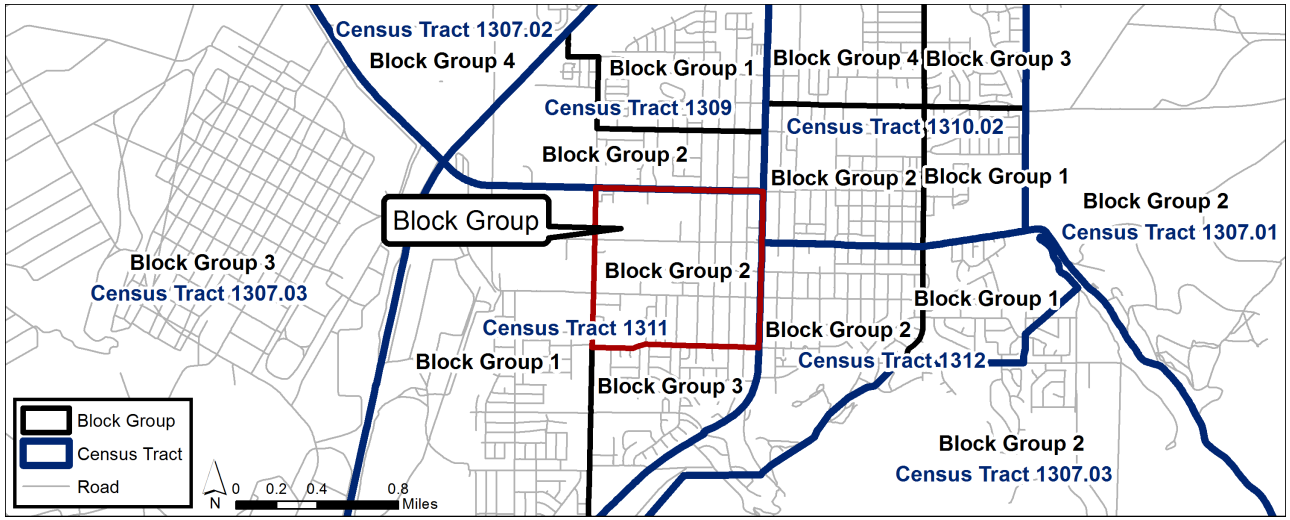
Geographical resolution

Standard hierarchy of US Census geographic entities



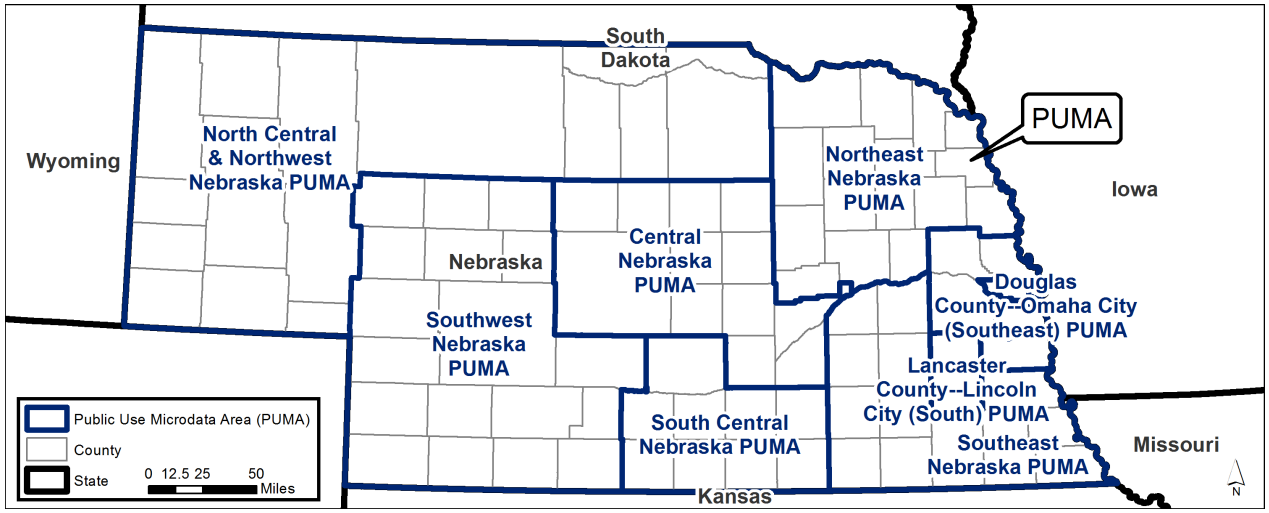
Geographical resolution

- **Block Groups (BGs)** are statistical divisions of census tracts, are generally defined to contain between 600 and 3,000 people.
  - The US Census provides various demographic distributions at the BG level.



Geographical resolution

- **Public Use Microdata Areas** (PUMAs) are statistical geographic areas defined for the dissemination of Public Use Microdata Sample (PUMS) data.
  - This is a 5% sample of the Census records.
  - A PUMA contains at least 100,000 people.
  - PUMAs are built on Census tracts and counties.



Generating a base population

- The Census gives marginal information about some variables at household level for each block group.
- Variables used (e.g.):
  - 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

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

- What we need:

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 detailed 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.

How Iterative Proportional Fitting works

There exists a universe of N (here 10,000) individuals that can be represented by a 2-way contingency table. For simplicity, let's say that there are 2 hair colors and 2 eye colors.

	Blue Eyes	Brown Eyes	Totals
Blond Hair	2000	1500	3500
Black Hair	500	6000	6500
Totals	2500	7500	10000

The Set-Up

There exists a universe of N = 10,000 individuals that can be represented by a 2-way contingency table. For simplicity, let's say that there are 2 hair colors and 2 eye colors.

	Blue Eyes	Brown Eyes	Totals
Blond Hair	?	?	3500
Black Hair	?	?	6500
Totals	2500	7500	10000

But, all we know are the marginal totals…

The Set-Up

The goal is to fill in as best we can the missing cells, while maintaining the interaction structure between hair and eye color

	Blue Eyes	Brown Eyes	Totals
Blond Hair	?	?	3500
Black Hair	?	?	6500
Totals	2500	7500	10000

But, all we know are the marginal totals…

A naïve approach

Assume hair and eye color are independent…

	Blue Eyes	Brown Eyes	Totals
Blond Hair	?	?	3500
Black Hair	?	?	6500
Totals	2500	7500	10000

We estimate the marginal probabilities as…

Pr(Blue) = .35 Pr(Blond) = .25

Pr(Brown) = .65 Pr(Black) = .75

Under the assumption of independence, the expected number of people in each of the cells should be…

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.35 x .25 x 10000= 875	.65 x .25 x 10000 = 1625	3500
Black Hair	.35 x .75 x 10000 = 2625	.65 x .75 x 10000 = 4875	6500
Totals	2500	7500	10000

A naïve approach

The marginal totals still match, but the dependence structure is totally changed.

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.35 x .25 x 10000= 875	.65 x .25 x 10000 = 1625	3500
Black Hair	.35 x .75 x 10000 = 2625	.65 x .75 x 10000 = 4875	6500
Totals	2500	7500	10000

Table estimated under independence assumption from marginal totals

	Blue Eyes	Brown Eyes	Totals
Blond Hair	2000	1500	3500
Black Hair	500	6000	6500
Totals	2500	7500	10000

Original table

The combination of blue eyes and blond hair is estimated to be much less common than it is in the real population because hair and eye color are dependent, i.e. Pr(Blue Eyes and Blond Hair) ≠Pr(Blue Eyes) Pr(Blond Hair)

The Set-Up

We obtain a sub-sample of size n=100 and obtain the following counts…

	Blue Eyes	Brown Eyes	Totals
Blond Hair	18	16	34
Black Hair	3	63	66
Totals	21	79	100

The Set-Up

The challenge is to maintain the dependence structure of the sub-sample while matching it to the whole population's marginal totals.

	Blue Eyes	Brown Eyes	Totals
Blond Hair	18	16	34
Black Hair	3	63	66
Totals	21	79	100

	Blue Eyes	Brown Eyes	Totals
Blond Hair	m11	m12	3500
Black Hair	m21	m22	6500
Totals	2500	7500	10000

What does this mean?

Intuition

Minimum discrimination information principle (Kullback)

Suppose we have a prior distribution over some variables, $p(x)$.
If we want to revise this distribution based on some new observations, we should choose a new distribution, $p'(x)$, that matches the new observations, while minimizing $D(p'||p)$.

$$I(p;\pi) = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} ln \frac{p_{ij}}{\pi_{ij}}, \quad where \quad \pi_{ij} = \frac{n_{ij}}{n}$$

In other words, we are trying to find a joint distribution that matches the marginals, while staying as close to the sample distribution as we can.

The IPF Algorithm

- Here we work in terms of cell probabilities instead of counts.
  - Let $m_{ij} = np_{ij}$
  - The marginal probabilities are then fixed.

$$\frac{N_{i\cdot}}{N} = p_{i\cdot} = \sum_j p_{ij} \qquad \frac{N_{\cdot j}}{N} = p_{\cdot j} = \sum_i p_{ij}$$

- Initialize $p_{ij}^{(0)} = n_{ij}/n$
- For $t \geq 1$
  - Set $p_{ij}^{(2t-1)} = p_{ij}^{(2t-2)} \dfrac{p_{i\cdot}}{p_{i\cdot}^{(2t-2)}}$ and $p_{ij}^{(2t)} = p_{ij}^{(2t-1)} \dfrac{p_{\cdot j}}{p_{\cdot j}^{(2t-1)}}$

IPF on our example data

	Blue Eyes	Brown Eyes	Totals
Blond Hair	18	16	34
Black Hair	3	63	66
Totals	21	89	100

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Initialize

	Blue Eyes	Brown Eyes	Totals
Blond Hair	18/100 = .18	17/100= .16	.34
Black Hair	3/100= .03	62/100= .63	.66
Totals	.21	.79	1.00

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Adjust rows to match given proportion… factor for this adjustment are .35/.34 and .65/.66…

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.18x.35/.34 = .185	.16x .35/.34 = .165	.35
Black Hair	.03x .65/.66 = .030	.63x .65/.66 = .620	.65
Totals	.215	.785	1.00

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Adjust columns to match given proportion… factor for this adjustment are .25/.215 and .75/.785…

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.185x .25/.215 = .215	.165x .75/.785 = .158	.373
Black Hair	.030x .25/.215 = .035	.620 x .75/.785 = .592	.627
Totals	.25	.75	1.00

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Adjust rows to match given proportion… factor for this adjustment are .35/.373 and .65/.627 …

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.215x .35/.373 = .202	.158x .35/.373 = .148	.35
Black Hair	.035x .65/.627 = .036	.592x .65/.627 = .614	.65
Totals	.238	.762	1.00

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Adjust columns to match given proportion… factor for this adjustment are .25/.238 and .75/.762 …

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.202x .25/.238 = .212	.148x .75/.762 = .146	.358
Black Hair	.036x .25/.238 = .038	.614 x .75/.762 = .604	.642
Totals	.25	.75	1.00

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Adjust rows to match given proportion… factor for this adjustment are .35/.358 and .65/.642 …

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.212x .35/.358 = .207	.146x .35/.358 = .143	.35
Black Hair	.038 x .65/.642 = .038	.604 * .65/.642 = .612	.65
Totals	.244	.755	1.00

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Adjust columns to match given proportion… factor for this adjustment are .25/.244 and .75/.755 …

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.210	.142	.352
Black Hair	.040	.608	.648
Totals	.25	.75	1.00

Subject to p1. = .35, p2. = .65, p.1=.25, p.2=.75

IPF on our example data

Adjust rows to match given proportion… factor for this adjustment are .35/.352 and .65/.648 …

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.209	.141	.350
Black Hair	.040	.609	.649
Totals	.249	.75	1.00

All changes were less than .001 from last iteration, so we've reached our stopping criterion. Done!
Notice that the marginal probabilities are very close to what we've constrained them to be. If we'd continued, we could get even closer.

Generating a base population

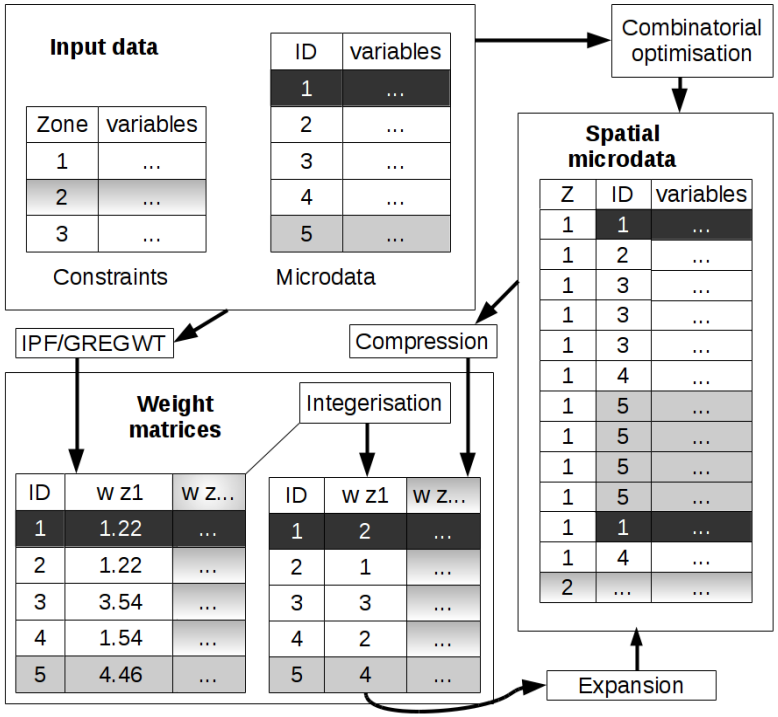
Householder's age								
Hsize	15-24	25-34	35-44	45-54	55-64	65-74	>74	Total
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.003	0.141	0.061	0.020	0.047	0.063	0.000	0.336
3	0.009	0.228	0.178	0.086	0.065	0.030	0.000	0.594
>3	0.000	0.003	0.022	0.022	0.016	0.007	0.000	0.069
Total	0.011	0.372	0.261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

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

- Once we have a joint distribution, we sample from it repeatedly and choose matching PUMS records to create the synthetic population.

Other approaches to population synthesis

- Combinatorial optimization
  - Estimating the micro-population by stochastically reweighting the given micro-data
  - Randomly allocate individuals from the micro-data to each zone, and iteratively replace to improve goodness-of-fit
- Generalized Regression Weighting (GREGWT), Simulated Annealing (GenSA)
- They can be considered to be broadly equivalent, as both involve some form of reweighting



Relationship between IPF and CO

Source: Lovelace, Robin, and Morgane Dumont. Spatial Microsimulation with R. CRC Press, 2016.

Examples

- TRANSIMS - https://www.fhwa.dot.gov/planning/tmip/resources/transims/
  - Used for accurate and sensitive travel forecasts for transportation planning and emission analysis
- URBANSIM - http://www.urbansim.com/urbansim/
  - To simulate metropolitan real estate markets and study the impact of land use policies
- EUROMOD - https://www.euromod.ac.uk/
  - EU based microsimulation to calculate effects of taxes and benefits on incomes and work incentives
- POHEM - http://www.statcan.gc.ca/eng/microsimulation/modgen/new/mods/pohem
  - Canada based Longitudinal population health microsimulation model to rationally compare competing health intervention alternatives
- TRIM - http://trim.urban.org/T3Welcome.php
  - to understand the potential outcomes of public policy changes such as welfare reform, tax reform, and national health care reform.

Some useful references

- Deming, W. Edwards; Stephan, Frederick F. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." *Ann. Math. Statist. 11* (1940), no. 4, 427—444.
- Ireland, C. T. and Kullback, S. "Contingency Tables with Given Marginals." *Biometrika 55* (1968), 179-188.
- Fienberg, Stephen E. "An Iterative Procedure for Estimation in Contingency Tables." *The Annals of Mathematical Statistics 41* (1970), 907-917.
- Orcutt, Guy H. "A New Type of Socio-economic System." *The Review of Economics and Statistics 39* (1957), 116-123.
- Beckman, Richard J., Keith A. Baggerly, and Michael D. McKay. "Creating synthetic baseline populations." *Transportation Research Part A: Policy and Practice 30.6* (1996): 415-429.
- Müller, Kirill, and Kay W. Axhausen. "Population Synthesis for Microsimulation: State of the Art." *Transportation Research Board 90th Annual Meeting*. No. 11-1789. 2011.
- Lovelace, Robin, and Morgane Dumont. "*Spatial Microsimulation with R.*" CRC Press, 2016.

Section

MULTI-LEVEL POPULATION SYNTHESIS

Goal: To generate a population of individuals *and households* with realistic demographic attributes

Input:
- Distributions over demographics (*marginal distributions),*
- A sample of census records

Method: Iterative Proportional Updating (IPU)

Motivation

- We may have marginals at both household and person level.
- The earlier IPF method is applied just at the household level.
  - Households are sampled from the joint distribution and copied over into the synthetic population
  - This may result in a discrepancy in the distributions at the person level

Multi-level constraints

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	

	Blue Eyes	Brown Eyes	Totals
Blond Hair	.209	.141	.350
Black Hair	.040	.609	.649
Totals	.249	.75	1.00

Simplified example

Household-level IPF results:

	Income < $50K	Income >= $50K	Totals
1 person	0.25	0.15	0.4
>= 2 people	0.1	0.5	0.6
Totals	0.35	0.65	1.00

Individual-level IPF results:

	Blue Eyes	Brown Eyes	Totals
Blond Hair	0.2	0.15	0.35
Black Hair	0.05	0.6	0.65
Totals	0.25	0.75	1.00

If we sample households using the household-level joint distribution, the synthetic population will not match the individual-level joint distribution.

Iterative Proportional Updating

- Do IPF separately for the household-level constraints and the individual-level constraints.
- The household records corresponding to each cell of the joint distribution will contain different numbers of people of each type.
- Adjust the sampling weights of the household records so that the distribution over individuals in the synthetic population more closely matches the results of the individual-level IPF.

Xin Ye, Karthik Konduri, Ram M. Pendyala, Bhargava Sana, Paul Waddell. "A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations." *Presented at the 88th Annual Meeting of the Transportation Research Board,* 2009.

Simplified example

HH ID	HH Size	HH Income	Blond, blue	Blond, brown	Black, blue	Black, brown	Wt
1	1	40,000	1	0	0	0	1.1
2	1	45,000	0	1	0	0	1.1
3	2	35,000	1	0	0	1	0.75
4	2	42,000	0	1	1	0	0.75
5	4	80,000	2	0	1	1	1.4
6	3	60,000	0	1	0	2	1.3

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	1
2	1	0	0	0	0	1	0	0	1
3	0	1	0	0	1	0	0	1	1
4	0	1	0	0	0	1	1	0	1
5	0	0	0	1	2	0	1	1	1
6	0	0	0	1	0	1	0	2	1

Total number of households = 2000
Total number of people = 4000

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	1
2	1	0	0	0	0	1	0	0	1
3	0	1	0	0	1	0	0	1	1
4	0	1	0	0	0	1	1	0	1
5	0	0	0	1	2	0	1	1	1
6	0	0	0	1	0	1	0	2	1

$$\rho = \frac{c_k}{\sum_q d_{s_{qk},k} \times w_{s_{qk}}} = \frac{0.25 \times 2000}{1 \times 1 + 1 \times 1} = \frac{500}{2} = 250$$

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	250
2	1	0	0	0	0	1	0	0	250
3	0	1	0	0	1	0	0	1	1
4	0	1	0	0	0	1	1	0	1
5	0	0	0	1	2	0	1	1	1
6	0	0	0	1	0	1	0	2	1

$$\rho = \frac{c_k}{\sum_q d_{s_{qk},k} \times w_{s_{qk}}} = \frac{0.1 \times 2000}{1 \times 1 + 1 \times 1} = \frac{200}{2} = 100$$

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	250
2	1	0	0	0	0	1	0	0	250
3	0	1	0	0	1	0	0	1	100
4	0	1	0	0	0	1	1	0	100
5	0	0	0	1	2	0	1	1	500
6	0	0	0	1	0	1	0	2	500

$$\rho = \frac{c_k}{\sum_q d_{s_{qk},k} \times w_{s_{qk}}} = \frac{0.2 \times 4000}{1 \times 250 + 1 \times 100 + 2 \times 500} = \frac{800}{1350} = 0.593$$

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	148.25
2	1	0	0	0	0	1	0	0	250
3	0	1	0	0	1	0	0	1	59.3
4	0	1	0	0	0	1	1	0	100
5	0	0	0	1	2	0	1	1	296.5
6	0	0	0	1	0	1	0	2	500

$$\rho = \frac{c_k}{\sum_q d_{s_{qk},k} \times w_{s_{qk}}} = \frac{0.15 \times 4000}{1 \times 250 + 1 \times 100 + 1 \times 500} = \frac{600}{850} = 0.706$$

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	148.25
2	1	0	0	0	0	1	0	0	176.5
3	0	1	0	0	1	0	0	1	59.3
4	0	1	0	0	0	1	1	0	70.6
5	0	0	0	1	2	0	1	1	296.5
6	0	0	0	1	0	1	0	2	303

$$\rho = \frac{c_k}{\sum_q d_{s_{qk},k} \times w_{s_{qk}}} = \frac{0.05 \times 4000}{1 \times 70.6 + 1 \times 296.5} = \frac{200}{367.1} = 0.545$$

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	148.25
2	1	0	0	0	0	1	0	0	176.5
3	0	1	0	0	1	0	0	1	59.3
4	0	1	0	0	0	1	1	0	38.48
5	0	0	0	1	2	0	1	1	161.6
6	0	0	0	1	0	1	0	2	303

$$\rho = \frac{c_k}{\sum_q d_{s_{qk},k} \times w_{s_{qk}}} = \frac{0.6 \times 4000}{1 \times 59.3 + 1 \times 161.6 + 2 \times 303} = \frac{2400}{826.9} = 2.9$$

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	148.25
2	1	0	0	0	0	1	0	0	176.5
3	0	1	0	0	1	0	0	1	171.97
4	0	1	0	0	0	1	1	0	38.48
5	0	0	0	1	2	0	1	1	468.64
6	0	0	0	1	0	1	0	2	878.7

$$\delta = \frac{\sum_j \frac{\sum_i |d_{i,j} w_i - c_j|}{c_j}}{m}$$

Simplified example

HH ID	HH Type				Person type				Wt
	<50K, 1	<50K, >=2	>=50K, 1	>=50K, >=2	Blond, blue	Blond, brown	Black, blue	Black, brown	
1	1	0	0	0	1	0	0	0	148.25
2	1	0	0	0	0	1	0	0	176.5
3	0	1	0	0	1	0	0	1	171.97
4	0	1	0	0	0	1	1	0	38.48
5	0	0	0	1	2	0	1	1	468.64
6	0	0	0	1	0	1	0	2	878.7
	0.35	0.05	1	0.347	0.014	0.823	1.52	.0008	

$$\delta = 0.5131$$

Example

Table 1. An Example of the Iterative Proportional Updating (IPU) Algorithm

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Weights 1	Weights 2	Weights 3	Weights 4	Weights 5	Final Weights
1	1	1	0	1	1	1	11.67	11.67	9.51	8.05	12.37	1.36
2	1	1	0	1	0	1	11.67	11.67	9.51	9.51	14.61	25.66
3	1	1	0	2	1	0	11.67	11.67	9.51	8.05	8.05	7.98
4	1	0	1	1	0	2	1.00	13.00	10.59	10.59	16.28	27.79
5	1	0	1	0	2	1	1.00	13.00	13.00	11.00	16.91	18.45
6	1	0	1	1	1	0	1.00	13.00	10.59	8.97	8.97	8.64
7	1	0	1	2	1	2	1.00	13.00	10.59	8.97	13.78	1.47
8	1	0	1	1	1	0	1.00	13.00	10.59	8.97	8.97	8.64
Weighted Sum		3.00	5.00	9.00	7.00	7.00						
Constraints		35.00	65.00	91.00	65.00	104.00						
δ_b		0.9143	0.9231	0.9011	0.8923	0.9327						
Weighted Sum 1		***35.00***	5.00	51.67	28.33	28.33						
Weighted Sum 2		35.00	***65.00***	111.67	88.33	88.33						
Weighted Sum 3		28.52	55.38	***91.00***	76.80	74.39						
Weighted Sum 4		25.60	48.50	80.11	***65.00***	67.68						
Weighted Sum 5		35.02	64.90	104.84	85.94	***104.00***						
δ_a		0.0006	0.0015	0.1521	0.3222	0.0000						
Final Weighted Sum		35.00	65.00	91.00	65.00	104.00						

Xin Ye, Karthik Konduri, Ram M. Pendyala, Bhargava Sana, Paul Waddell. "A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations." *Presented at the 88th Annual Meeting of the Transportation Research Board,* 2009.

Some useful references

- Theo Arentze, Harry Timmermans, Frank Hofman. "Creating Synthetic Household Populations: Problem and Approach." *Transportation Research Record: Journal of the Transportation Research Board 2014*, 85-91, 2007.
- Jessica Y. Guo and Chandra R. Bhat, "Population Synthesis for Microsimulation Travel Behavior," *Transportation Research Record: Journal of the Transportation Research Board 2014*, 92–101, 2007.
- Xin Ye, Karthik Konduri, Ram M. Pendyala, Bhargava Sana, Paul Waddell. "A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations." *Presented at the 88th Annual Meeting of the Transportation Research Board,* 2009.
- Joshua Auld and Abolfazl Mohammadian, "Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels," *Transportation Research Record: Journal of the Transportation Research Board 2175*, 138–147, 2010.
- Kirill Müller, Kay W. Axhausen. "Population Synthesis for Microsimulation: State of the Art." *Transportation Research Board 90th Annual Meeting*. No. 11-1789. 2011.

Section

POPULATION EVOLUTION

Goal: To evolve a population of individuals and households with realistic demographic attributes over time

Input:
- An existing synthetic population,
- Various data on probabilities of life events

Method: Logistic regression; direct lookup in probability tables

Modeling with life events: FPOP

- FPOP models six kinds of life events:
  - aging,
  - mortality,
  - birth,
  - marriage/union formation
  - marriage/union dissolution,
  - migration

Susan M. Rogers et al. (2014), A Geospatial Dynamic Microsimulation Model for Household Population Projections, *International Journal of Microsimulation 7*, 119-146.

Modeling with life events: FPOP

Table 1 Initial Demographic Parameters applied to FPOP for a U.S. Population

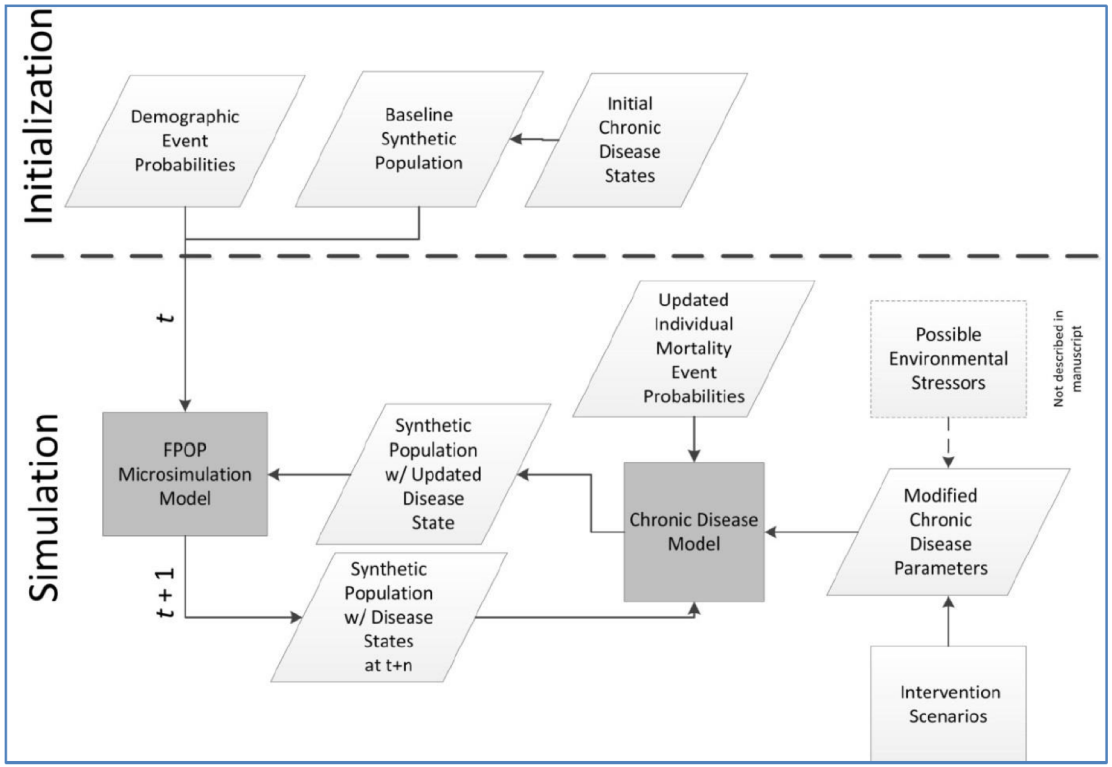
Parameter	Source of estimate	Geospatial level of estimate	Statistical model	Explanatory variables
Aging	2007-2011 ACS		Time *t* -> *t*+1	
Mortality	NCHS/NVSS	National	Lifetable	Age, sex, race, Hispanic origin
Fertility	NCHS/NVSS	State	Age-race-parity specific probability	Age, race, parity Hispanic origin
Union				
Formation	ACS 2009	State	Logistic regression	Age, sex, race Hispanic origin, education, employment,
Dissolution	ACS 2009	State	Logistic regression	Age, sex, race, Hispanic origin, education, employment
Migration	ACS 2009; 2008-10	State	Logistic regression	Age, sex, race, Hispanic origin

ACS is the American Community Survey; NCHS NVSS is the National Center for Health Statistics / National Vital Statistics System.

Susan M. Rogers et al. (2014), A Geospatial Dynamic Microsimulation Model for Household Population Projections, *International Journal of Microsimulation 7*, 119-146.

Modeling with life events: FPOP

Figure 1 A Model Framework Showing FPOP as the Demographic Core for a Chronic Disease Scenario



Susan M. Rogers et al. (2014), A Geospatial Dynamic Microsimulation Model for Household Population Projections, *International Journal of Microsimulation 7*, 119-146.

Modeling with life events: UrbanSim

- Economic and demographic transition models
- Household and employment mobility models
- Household and employment location models
- Real estate development model
- Land price model

Paul Waddell (2002), UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning, *Journal of the American Planning Association 68*, 297-314.

Modeling with life events

Namazi-Rad et al. (2014)

- Generalized Malthusian model of population growth:

$$P^{(t+1)} = P^{(t)} + \sum_{i=1}^{w}\left[F_i^{(t)} \times \varphi\left(F_i^{(t)}\right)\right]$$
$$+ \sum_{j=1}^{w}\left[M_j^{(t)} \times \varphi\left(M_j^{(t)}\right)\right] + \sum_{i=1}^{w}\sum_{j=1}^{w}\left[C_{(i,j)}^{(t)} \times \varphi\left(C_{(i,j)}^{(t)}\right)\right]$$

- Growth rates for males, females, and couples are derived from data.

M.-R. Namazi-Rad, P. Mokhtarian, and P. Perez (2014), Generating a Dynamic Synthetic Population - Using an Age-Structured Two-Sex Model for Household Dynamics, *PLoS ONE 9(4)*: e94761.

Modeling with life events

Namazi-Rad et al. (2014)

- Household transitions:
  - Death rate: $\boldsymbol{D}'_{ijkl} = \left[D_i^{(M1)}, D_j^{(F1)}, D_k^{(F2)}, D_l^{(M2)}\right]$,
  - Marriage rate: $\boldsymbol{\Gamma}'_{ij} = \left[\boldsymbol{\Gamma}_{ij}^{(1)}, \boldsymbol{\Gamma}_{ij}^{(2)}\right]$,
  - Birth rate: $\boldsymbol{B}'_{ij} = \left[\boldsymbol{B}_{ij}^{C}\right]$,
  - Divorce rate: $\boldsymbol{\Lambda}'_{ij} = \left[\boldsymbol{\Lambda}_{ij}^{C}\right]$, and
  - Leaving the parental home: $\boldsymbol{\Delta}'_{kl} = \left[\boldsymbol{\Delta}_k^{(F)}, \boldsymbol{\Delta}_l^{(M)}\right]$.

M.-R. Namazi-Rad, P. Mokhtarian, and P. Perez (2014), Generating a Dynamic Synthetic Population - Using an Age-Structured Two-Sex Model for Household Dynamics, *PLoS ONE 9(4)*: e94761.

Modeling with life events

Namazi-Rad et al. (2014)

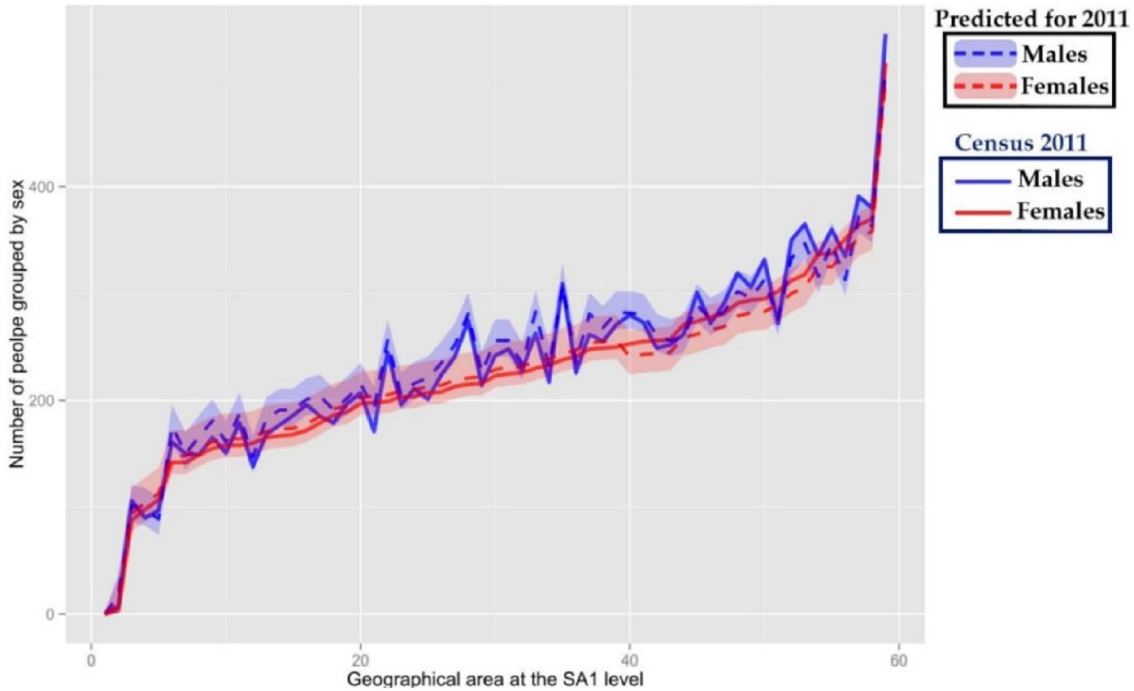


Figure 7. Age-specific population predictions vs the 2011 Australian Census at SA1 level.
doi:10.1371/journal.pone.0094761.g007

M.-R. Namazi-Rad, P. Mokhtarian, and P. Perez (2014), Generating a Dynamic Synthetic Population - Using an Age-Structured Two-Sex Model for Household Dynamics, *PLoS ONE 9(4)*: e94761.

Some useful references

- Paul Waddell (2002), UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning, *Journal of the American Planning Association 68*, 297-314.
  - https://udst.github.io/urbansim/
- Susan M. Rogers, James Rineer, Matthew D. Scruggs, William D. Wheaton, Phillip C. Cooley, Douglas J. Roberts & Diane K. Wagener (2014), A Geospatial Dynamic Microsimulation Model for Household Population Projections, *International Journal of Microsimulation 7*, 119-146.
- M.-R. Namazi-Rad, P. Mokhtarian, and P. Perez (2014), Generating a Dynamic Synthetic Population - Using an Age-Structured Two-Sex Model for Household Dynamics, *PLoS ONE 9(4)*: e94761.
- Kristian Lum, Samarth Swarup, Stephen G. Eubank, James Hawdon (2014). The Contagious Nature of Imprisonment: An Agent-based Model to Explain Racial Disparities in Incarceration Rates, *J. R. Soc. Interface 11(98):20140409*. Code available: http://ndssl.vbi.vt.edu/synthetic-data

Section

ACTIVITY ASSIGNMENT

Goal: To assign a realistic daily activity sequence to each agent

Input:
- A synthetic population of agents with demographics,
- A household activity survey

Methods:
- Classification and Regression Trees
- Fitted Values Method

Problem

- Given a population of agents, we want to assign them realistic activity sequences that
  - Obey the within-household dependence structure of the given data
  - Are consistent with the individual demographic characteristics of each agent.
- Why is this important?
  - In transportation modeling, this will help determine which locations people go to.
  - In epidemiology, this will determine contact patterns.

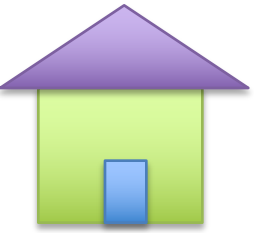
The National Household Travel Survey

- The NHTS/NPTS serves as the nation's inventory of daily travel. Data is collected on daily trips taken in a 24-hour period, and includes:
  - purpose of the trip (work, shopping, etc.);
  - means of transportation used (car, bus, subway, walk, etc.);
  - how long the trip took, i.e., travel time;
  - time of day when the trip took place;
  - day of week when the trip took place; and
  - if a private vehicle trip:
    - number of people in the vehicle , i.e., vehicle occupancy;
    - driver characteristics (age, sex, worker status, education level, etc.); and
    - vehicle attributes (make, model, model year, amount of miles driven in a year).
- These data are collected for:
  - all trips,
  - all modes,
  - all purposes,
  - all trip lengths, and
  - all areas of the country, urban and rural.

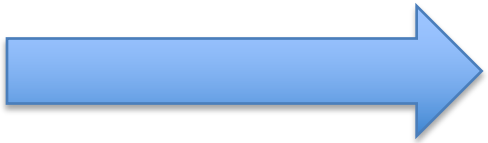
NHTS

http://nhts.ornl.gov

Activity Selection



What do I want to do?



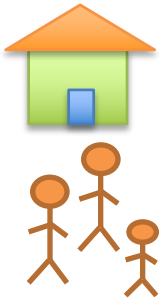
Where do I want to go?

Notation

Variable	Description
P_{ij}	Synthetic person i in synthetic household j.
H_j	Synthetic household j.
$P^*_{i^*j^*}$	Survey person i^* in survey household j^*.
$H^*_{j^*}$	Survey household j^*.
X_{ij}	Personal and household demographics of the ith synthetic person in the jth synthetic household. Example: X = [age, sex, highest income in household]
$X^*_{i^*j^*}$	Personal and household demographics of the i^{*th} person in the j^{*th} survey household.
$A^*_{i^*j^*}$	Activity sequence of the i^{*th} person in the j^{*th} survey household.
$Y^*_{i^*j^*k}$	Number of minutes the i^{*th} person in the j^{*th} survey household spends performing activity-type k (home, work, school, shopping, …)

CART-based re-sampling

Vaughn K., Speckman P. & Pas, E. (1997). Generating Household Activity-Travel Patterns (HATPs) for Synthetic Populations. *Transportation Research Board Annual Meeting.*



Survey Household

Consists of...

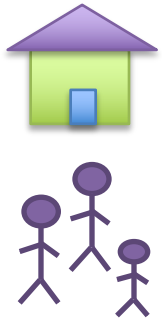
Survey Individuals

Split 1 (ex: age of oldest member)

Synthetic Household

Consists of...

Synthetic Individuals

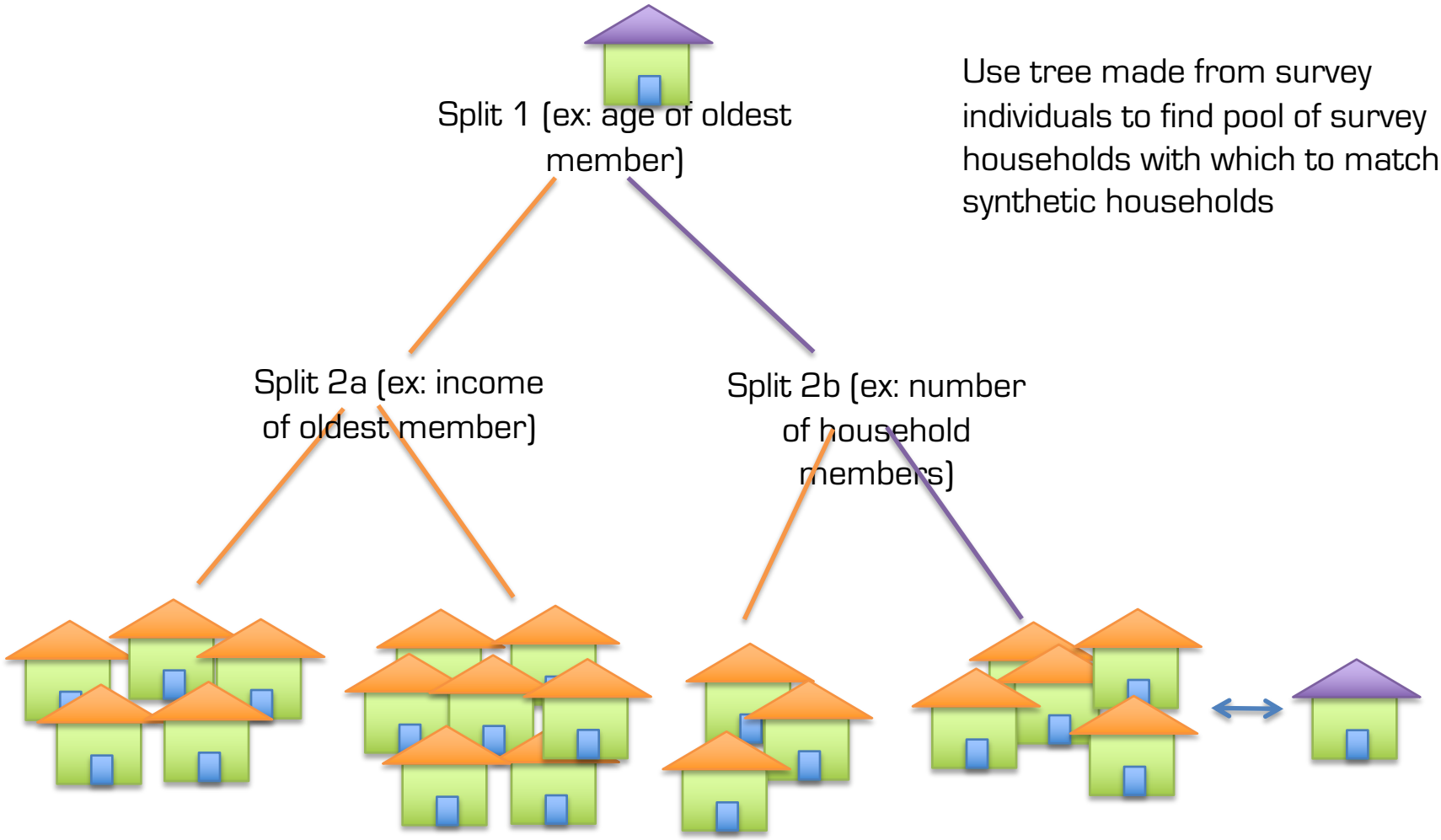


Split 2a (ex: income of oldest member)

Split 2b (ex: number of household members)

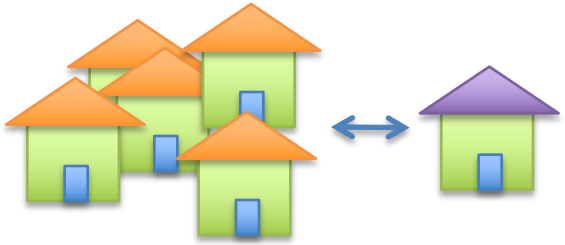
Use survey households in CART to break into leaf nodes based on total time spent doing each of several activities

CART-based re-sampling

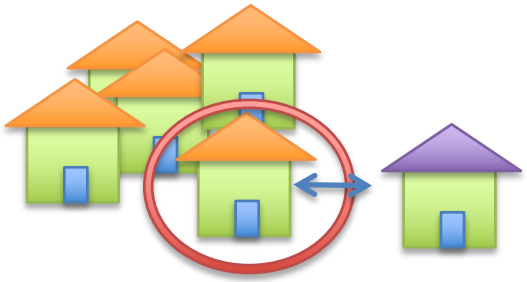


Use tree made from survey individuals to find pool of survey households with which to match synthetic households

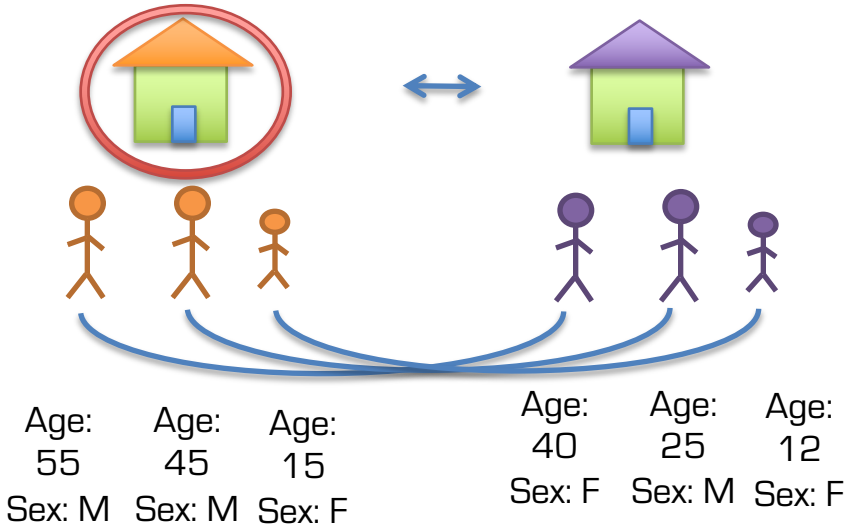
CART-based re-sampling



How to assign individual activity schedules?



1) Select at random from the households in the leaf node



2) Assign the schedules according to age (i.e. oldest member of synthetic household is assigned the schedule of survey household's oldest member).

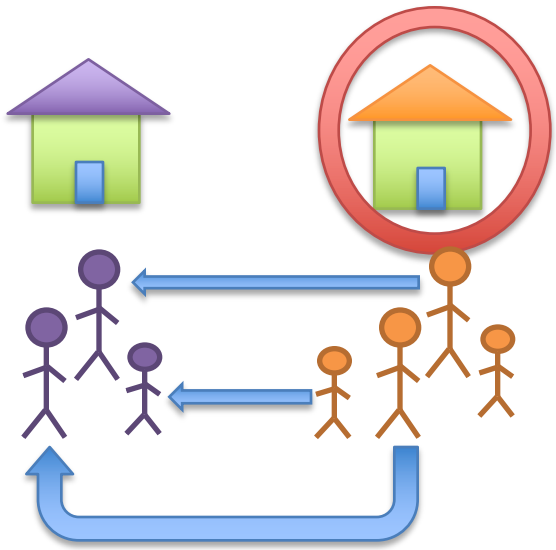
Fitted Values Method: general idea

We want to match survey households to synthetic households based on some measure of similarity.

Step 1: Select a survey household based on the similarity between the it and the synthetic household (probabilistic, minimum distance, etc.)

Step 2: Find the survey individual in the selected survey household that is most similar to each synthetic individual

Step 3: Assign schedules accordingly



Schedule:

8:00am	wake up
9:00am	travel to work
9:15am	arrive at work
1:00pm	go to lunch
2:00pm	return to work
6:00pm	travel to store
6:25pm	shopping
7:00pm	travel home
7:30pm	home

Similarity measures between survey and synthetic households

Define:

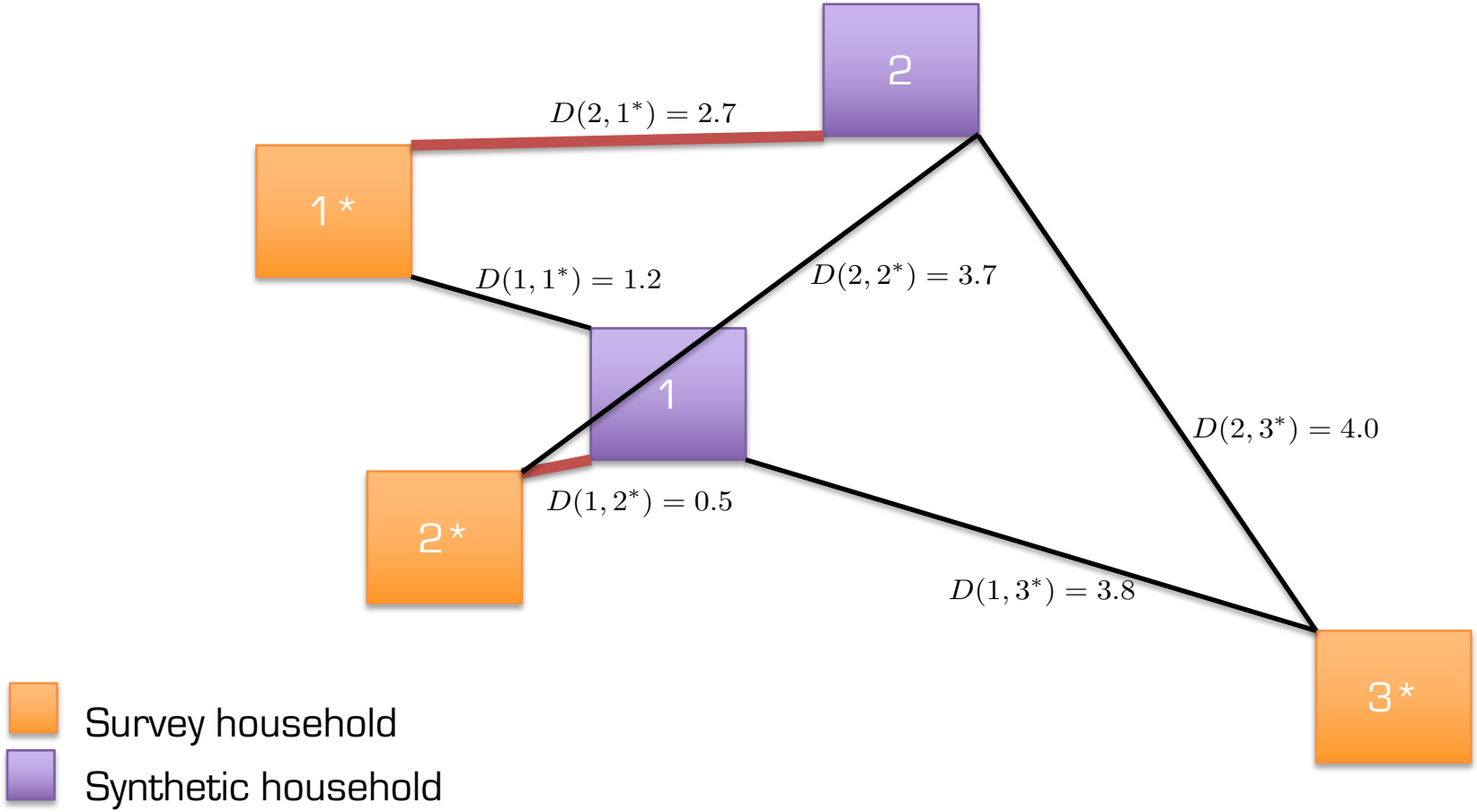
(1) $$D_{HH}(H_j, H^*_{j*}) = max_i\{D_{PH}(P_{ij}, H^*_{j*})\}$$

(2) $$D_{PH}(P_{ij}, H^*_{j*}) = min_{i*}\{D(P_{ij}, P^*_{i*j*})\}$$

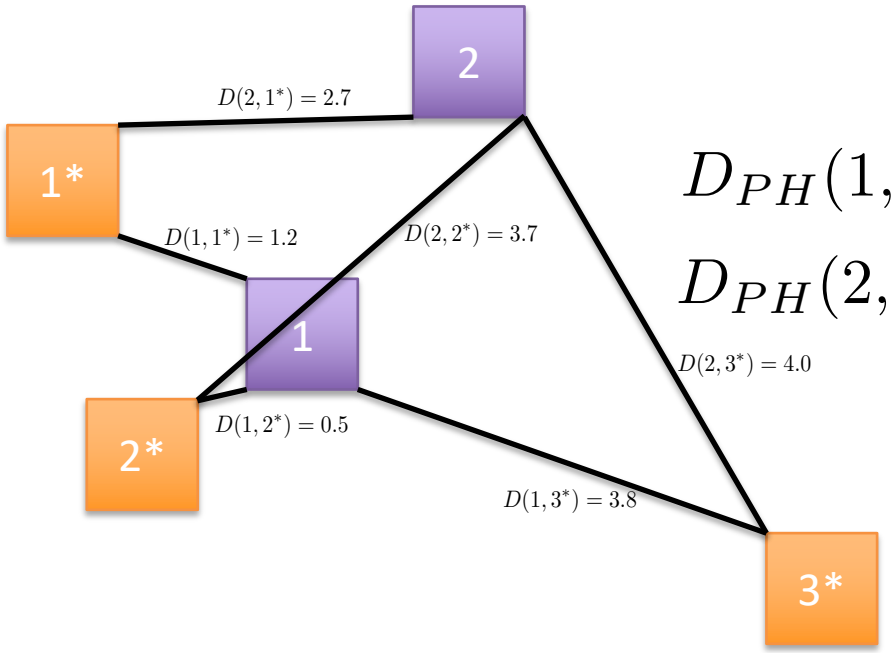
This is the (asymmetric) **Hausdorff distance** between synthetic household j and survey household j*.

Algorithmically, for each person in synthetic household j, find the person in survey household j* that most closely resembles them— this is that person's distance from the survey household (represented by Eq. (2)). The distance between the households j and j* is the farthest any of the members of household j is from j*.

Similarity measures between survey and synthetic households



Similarity measures between survey and synthetic households



$$D_{PH}(1, H^*_{j*}) = min(1.2, 0.5, 3.8) = 0.5$$

$$D_{PH}(2, H^*_{j*}) = min(2.7, 3.7, 4.0) = 2.7$$

$$D_{HH}(H_j, H^*_{j*}) = max(D_{PH}(1, H^*_{j*}), D_{PH}(2, H^*_{j*})) = 2.7$$

What about D?

Euclidean distance between covariate vectors?

$$||X_{ij} - X^*_{i^*j^*}||$$

Mahalanobis distance between covariate vectors?

$$(X_{ij} - X^*_{i^*j^*})S^{-1}(X_{ij} - X^*_{i^*j^*})$$

What about D?

We use...

$$D(P_{ij}, P^*_{i^*j^*}) = ||\hat{\mathbf{Y}}_{ij} - \hat{\mathbf{Y}}^*_{i^*j^*}||$$

where

$$\hat{\mathbf{Y}}_{ij} = \{\hat{f}_k(X_{ij}) : k = 1 \ldots K\}$$

i.e. the set of fitted values from some model, f_k, for each of the activity times. For example, if f_k is just a regression function*, for each activity type, k, we'd have

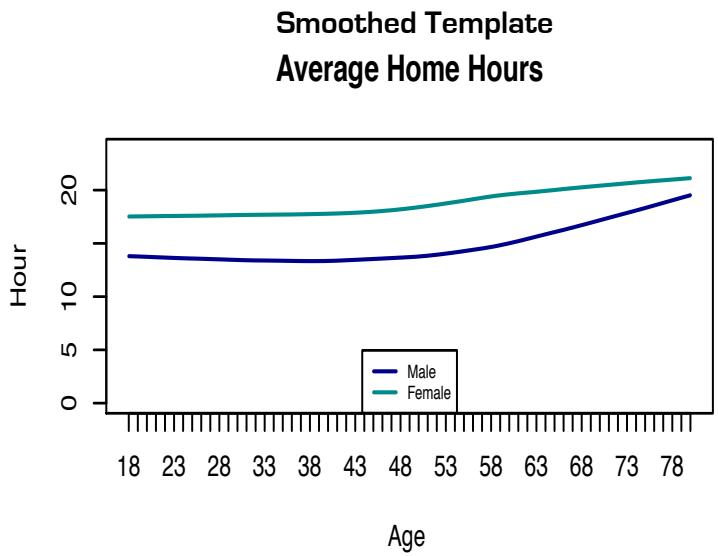
$$\hat{Y}_{ijk} = X_{ij}\hat{\beta}_k$$

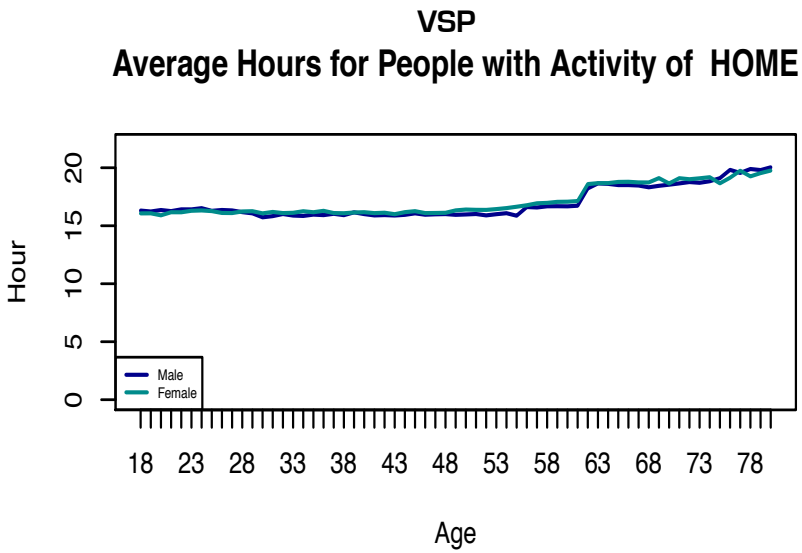
Where $\hat{\beta}_k$ is vector of estimated coefficients from the model

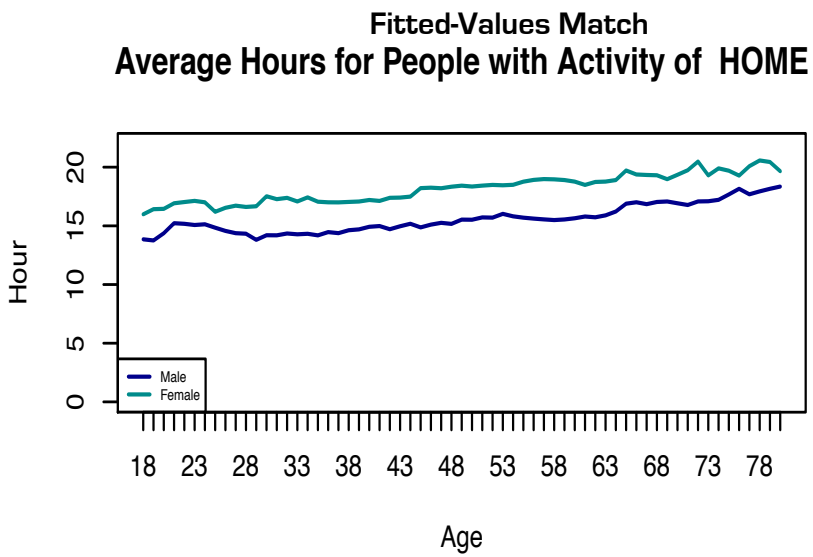
$$\hat{Y^*}_{i^*j^*k} = X_{i^*j^*}\beta_k + \epsilon_{i^*j^*k}$$

* We don't have to use linear regression. Something like CART, random forest, etc. could also be used here for f_k.

Results: Home



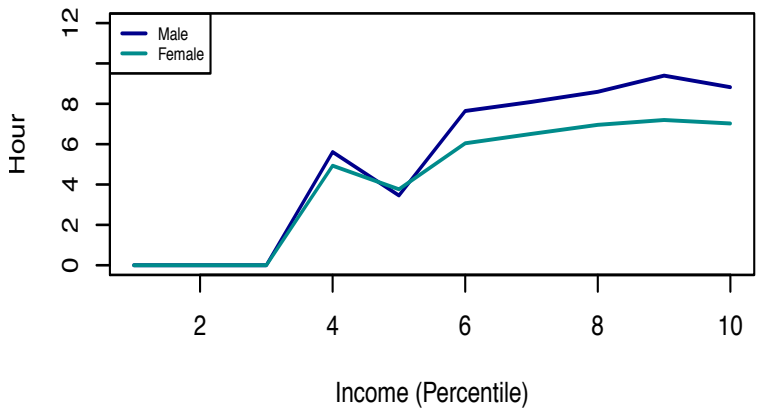




Results: Work

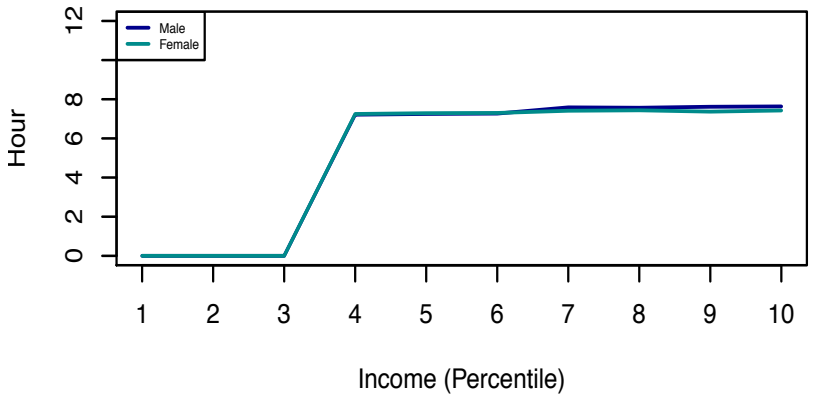
Smoothed Template

Average Work Hours



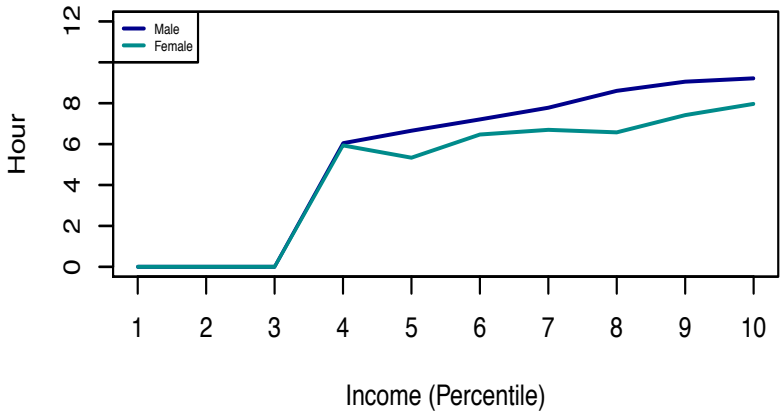
VSP

Average Hours for People with Activity of WORK

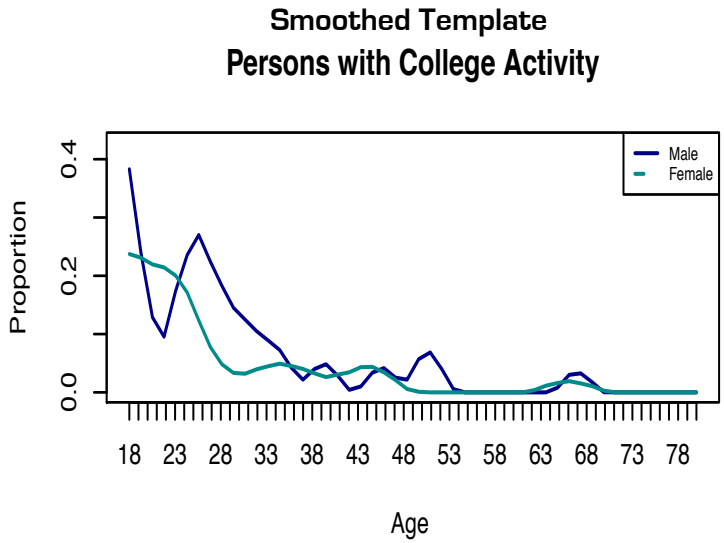


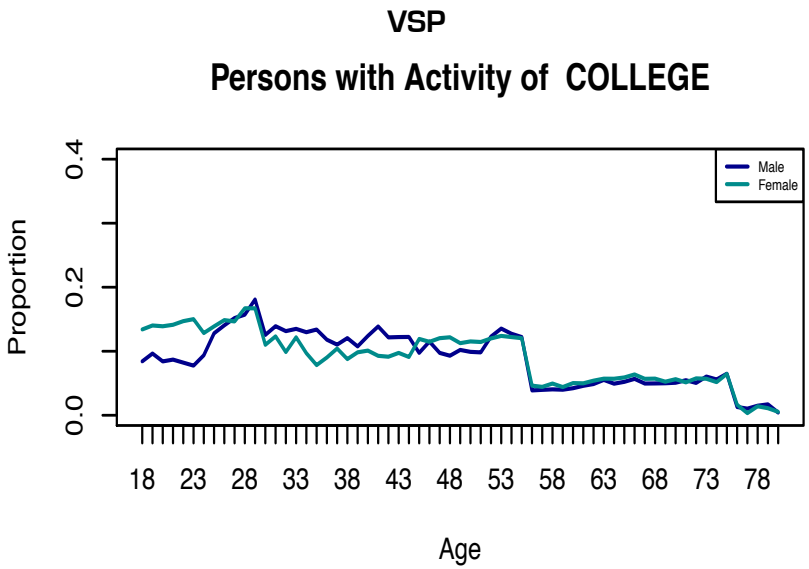
Fitted-Values Match

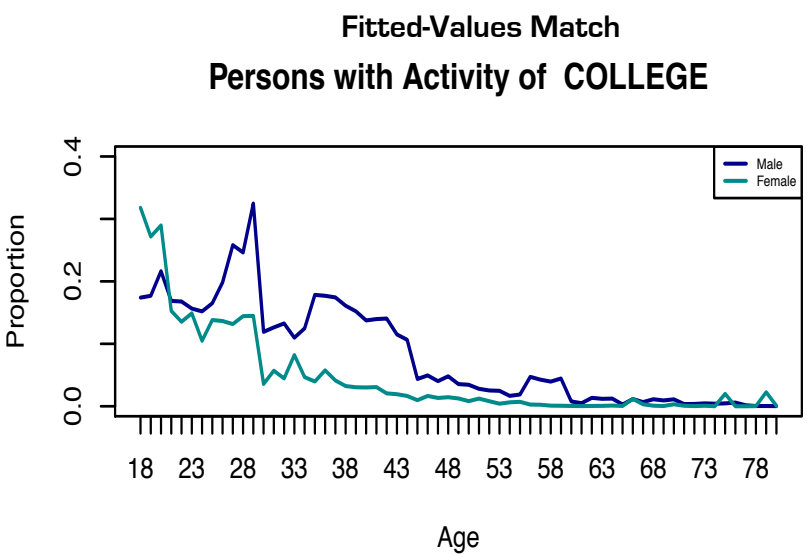
Average Hours for People with Activity of WORK



Results: College







Some useful references

- Kitamura, R. (1983). Sequential, history-dependent approach to trip-chaining behavior. *Transportation Research Record*.
- Kitamura, R., & Kermanshah, M. (1983). Identifying time and history dependencies of activity choice. *Transportation Research Record*.
- Vaughn K., Speckman P. & Pas, E. (1997). Generating Household Activity-Travel Patterns (HATPs) for Synthetic Populations. *Transportation Research Board Annual Meeting.*
- Bowman, J. L., & Ben-Akiva, M. E. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice, 35*, 1-28.
- Gliebe, J. P., & Koppelman, F. S. (2002). A model of joint activity participation between household members. *Transportation, 29*, 49-72.
- Goulias, K. G. (2002). Multilevel analysis of daily time use and time allocation to activity types accounting for complex covariance structures using correlated random effects. *Transportation, 29*, 31-48.
- Borgers, A., Hofman, F., & Timmermans, H. (2002). Conditional choice modelling of time allocation among spouses in transport settings. *European Journal of Transport and Infrastructure Research, 2*, 5-17.
- Zhang, J., & Fujiwara, A. (2004). Representing heterogeneous intra-household interaction in the context of time allocation. In *83rd Annual Meeting of the Transportation Research Board, Washington, DC*.
- Zhang, J., Timmermans, H., & Borgers, A. (2002). Utility-maximizing model of household time use for independent, shared, and allocated activities incorporating group decision mechanisms. *Transportation Research Record: Journal of the Transportation Research Board, 1807*, 1-8.
- Bradley, M., & Vovsha, P. (2005). A model for joint choice of daily activity pattern types of household members. *Transportation, 32*, 545-571.
- Lum, K., Chungbaek, Y., Eubank, S. G., & Marathe, M. V. (2016). A two-stage fitted-values approach to activity matching. *International Journal of Transportation 4(1),* 41-56.

Section

ASSIGNING NEW ATTRIBUTES/ACTIVITIES TO SYNTHETIC AGENTS

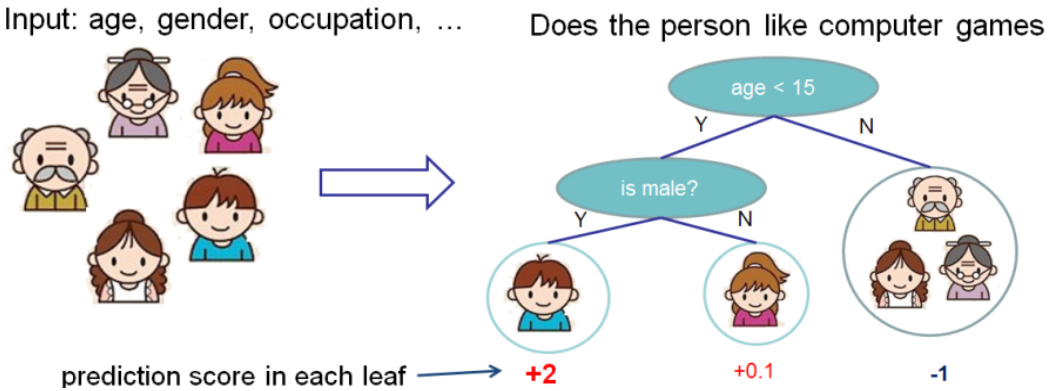
Augment the list of attributes in synthetic population

- The synthetic population has a set of available features but additional features may be desirable to expand its application to other areas of interest.
- For example to study obesity, tobacco addiction, activity on social media etc.

Available Features	Example Desirable Features
Age	Personal income
Gender	Have a computer/cell phone
Marital status	Level of physical activity
Employment status	Number of virtual friends
Household size	Time spent on social media
Household income	Have a credit card
Education level	Tobacco use
Number of vehicles	Gym membership
Religion	Body mass index
Race	Have a bank account

Classification and regression tree (CART)

- Predictive modeling approach
- CART recursively partitions data into disjoint sets by fitting a simple prediction model within each set.
- The dependent variable is the desired attribute (owns a cell phone) and the independent variables are the available attributes (such as age, gender, income).
- The resulting tree has a flow-chart structure that is easy to interpret.

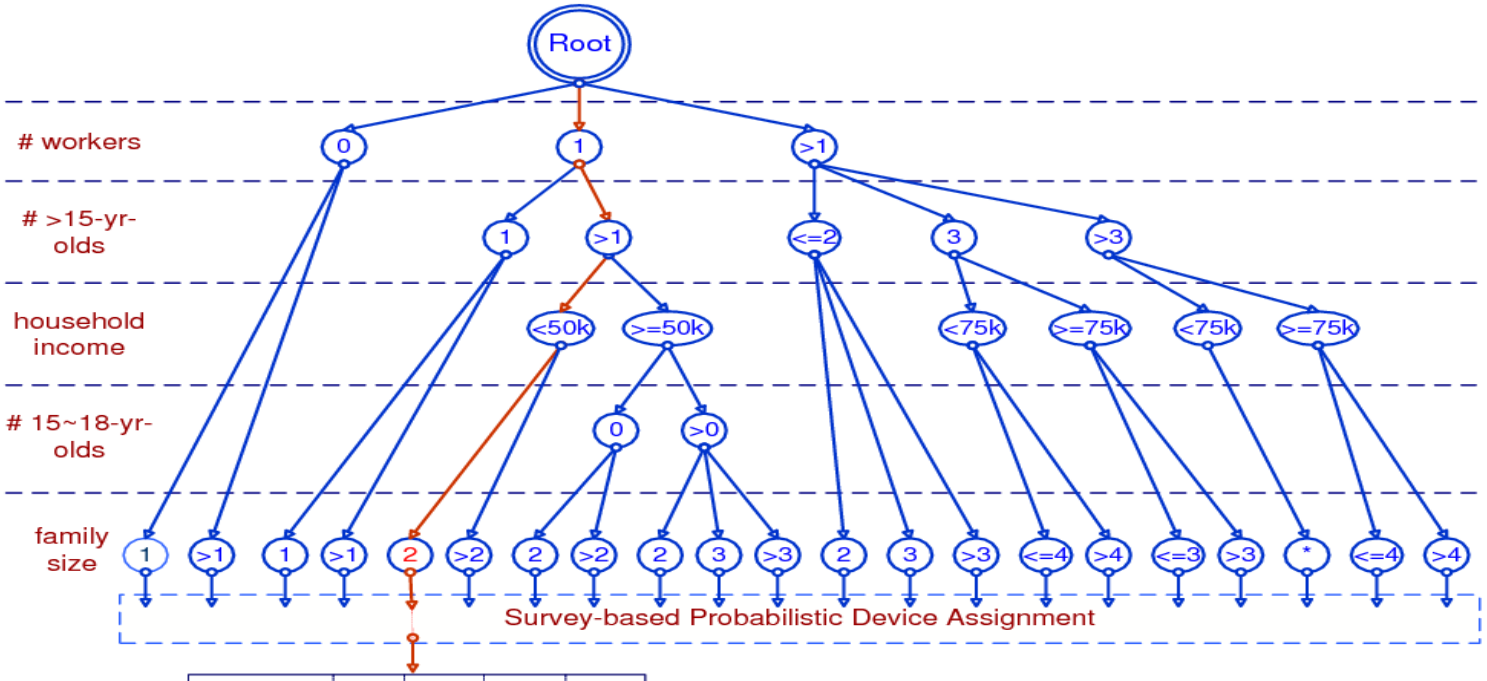


Example application of CART

- Cell phone assignment to synthetic population using CART
- Use National Health Interview Survey (NHIS) data which contains demographic information and cell phone ownership data for households.
- CART model is built using NHIS data. Dependent variable is average number of cell phones among householders and independent variables are demographics e.g. household income, # of workers in the family, family size etc.

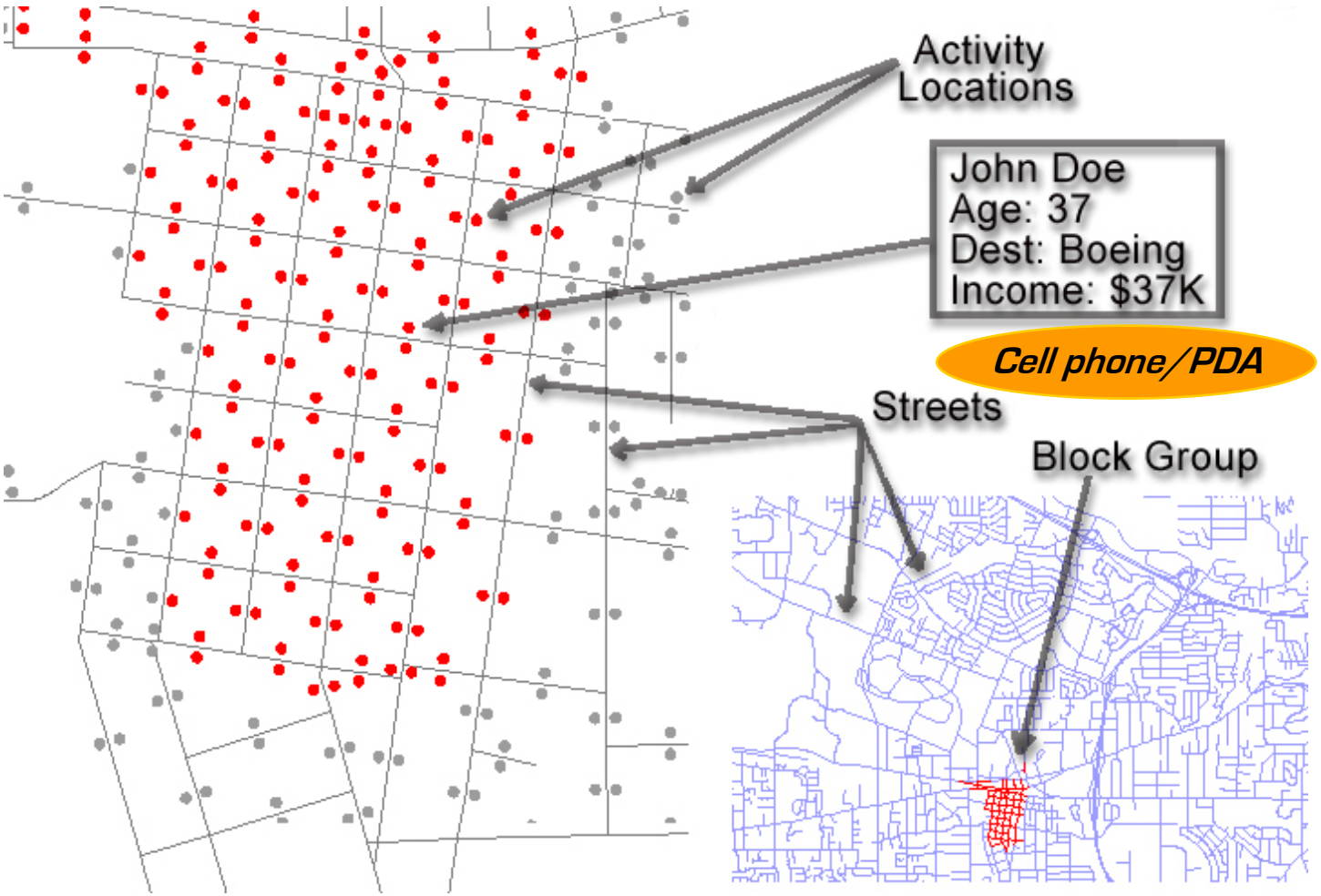
Example Application of CART

- The fitted model is applied to the synthetic population to determine the number of cell phones among householders.
- Once the number of cell phones for each family is determined, individuals are assigned cell phone randomly with a higher weight to workers in the family who are between age group 18-50 years.



#Devices	0	1	2	3
Prob.	0.35	0.33	0.30	0.02

Device assignment: example



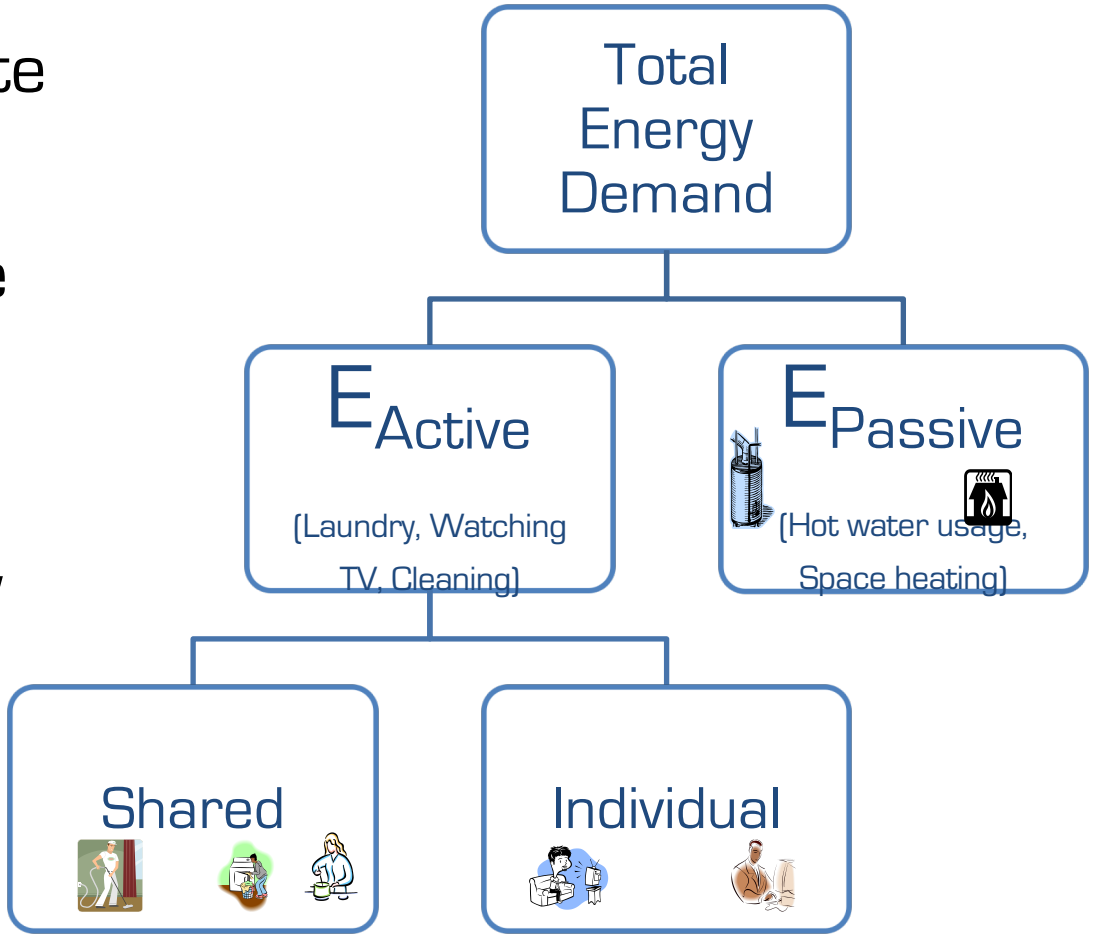
Limitations of CART

- Locally optimal decisions are made at each branch. Does not guarantee globally-optimal decision tree.

- Over-fitting can occur and result in a very complex tree model that does not generalize beyond training dataset.

Assigning energy related activities to individuals

- Goal is to calculate energy demand profiles of households in the synthetic population
- Currently will only focus on active energy demand calculation



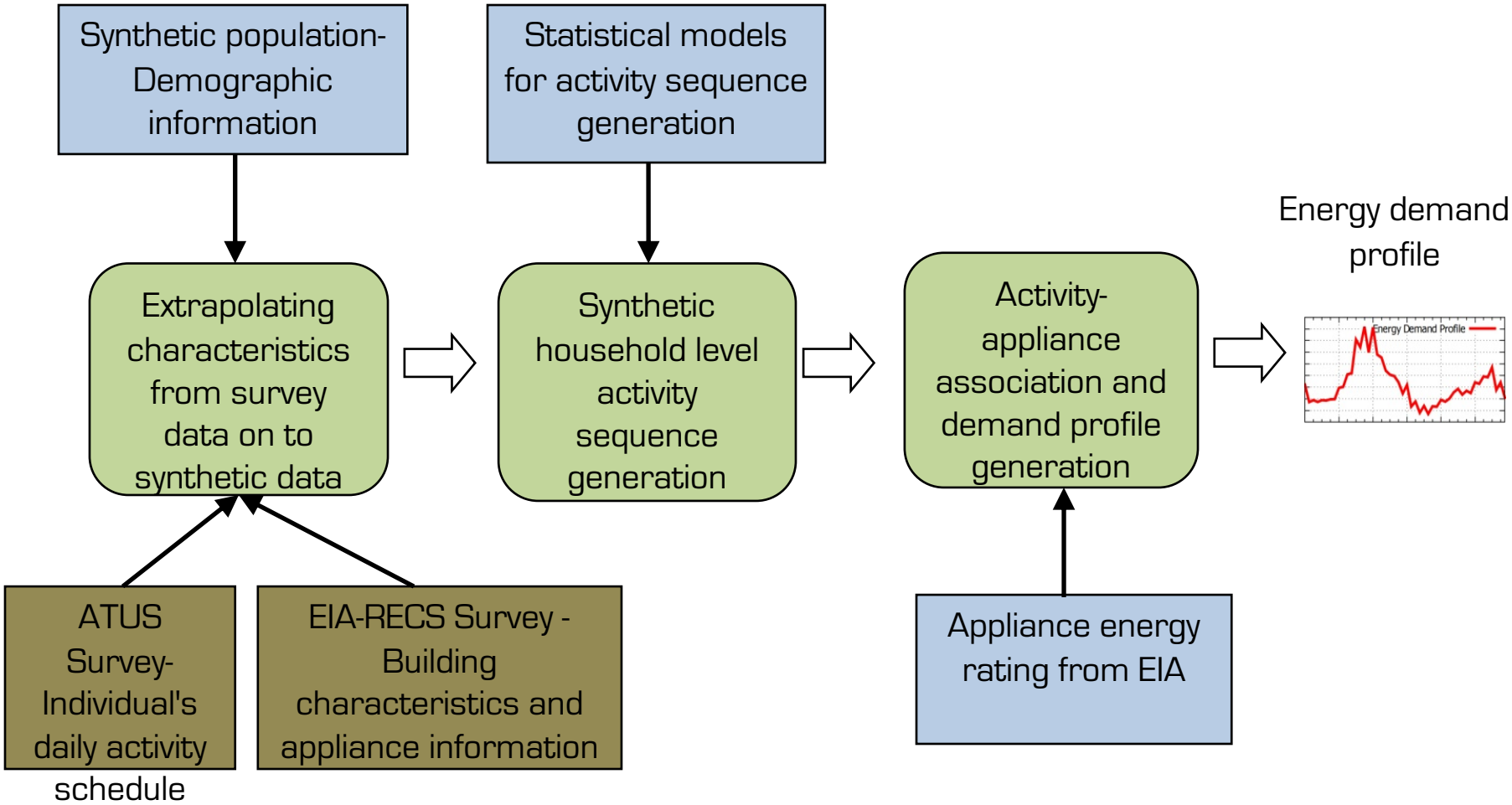
Active demand

- The root node represents the complete set of ATUS (American time use survey) respondents
- Contains 24 hour period activity dairies for 13, 260 respondents
- The demographic variables are used as the splitting variables namely
  - Age, Employment Status, Household Size, School Enrollment Status, Number of Children and Martial Status
- We use CART to construct the binary decision tree
- At each stage, the algorithm tries to split the node into two groups based on the best possible splitting variable

Active energy demand

- Each household member in the synthetic population is assigned a leaf node based on his/her demographic variables
- From the matching demographic variables, we select a ATUS survey respondent at random and assign the activity pattern of the ATUS survey respondent to the synthetic population household member.
- For each energy activity, a list of appliances are associated to calculated the energy used.

Overall view



Some useful references

- R. Subbiah, K. Lum, A. Marathe, and M. Marathe, "Activity Based Energy Demand Modeling for Residential Buildings," in *Proceedings of the IEEE PES Conference on Innovative Smart Grid Technologies*, Washington DC, 2013.
- G. Bustos-Turu, K. H. van Dam, S. Acha, C. N. Markides, and N. Shah, "Simulating residential electricity and heat demand in urban areas using an agent-based modelling approach," in *2016 IEEE International Energy Conference (ENERGYCON)*, 2016, pp. 1–6.

Section

LOCATION ASSIGNMENT

Goal: To assign a geographical location for each activity for each agent

Input:

- A synthetic population of agents with demographics and daily activity sequences
- Geographical data on roads, residence types, business locations, school locations, and other points of interest

Methods:

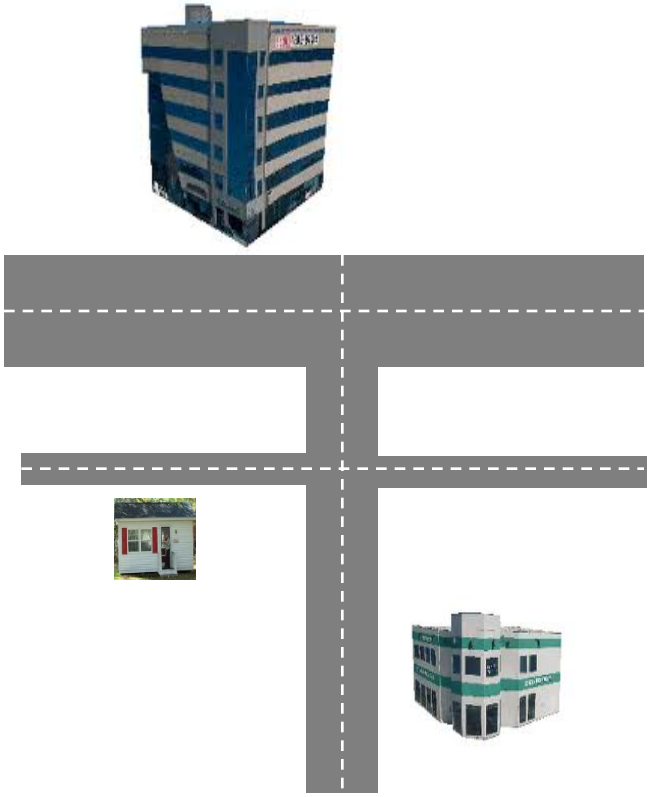
- Gravity model
- Trip chaining model
- Radiation model

Problem

- Given a population of agents with demographics and daily activity sequences, we want to assign a location for each activity
  - Activity types should be consistent with location types
  - It should result in sensible travel patterns

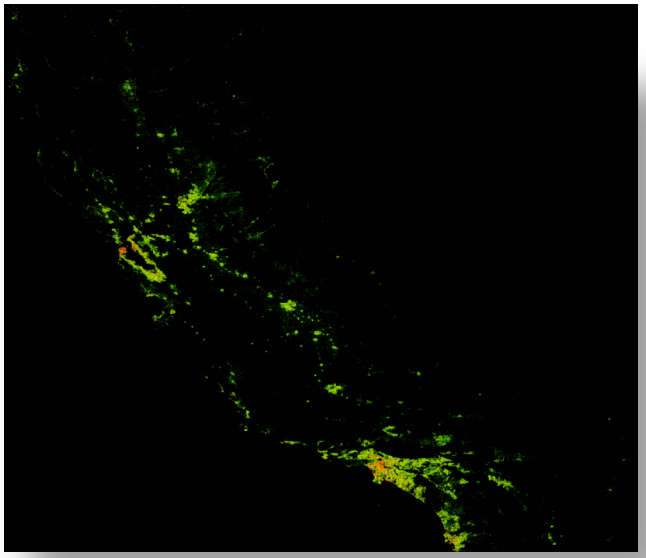
Home location assignment

- 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.

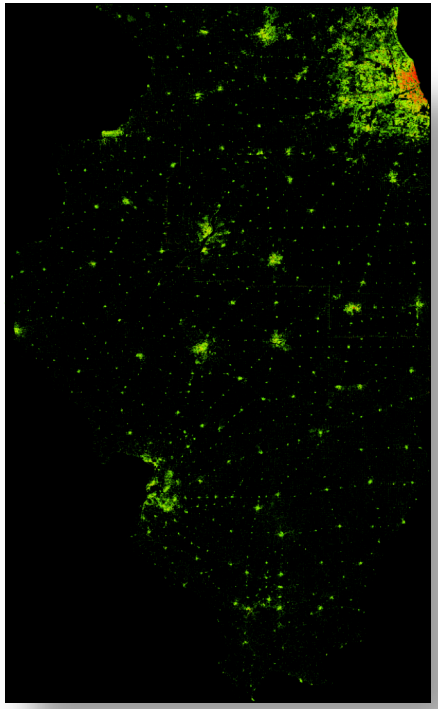


Home location assignment

- Synthetic households are assigned home locations with probability proportional to home location weights (home location type – building capacity)
- Output: located base population



California



Illinois

Assigning locations for all other activities

- Step 1: Use a discrete choice model to generate all work or school locations.
- Step 2: Use trip-chaining discrete choice models to generate locations for other
- activities.

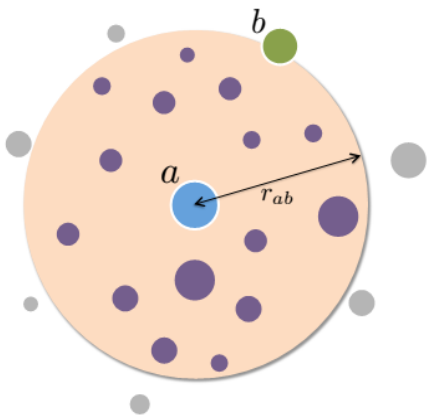
Step 1: work & school locations

- A work/school location, *L*, is chosen with probability,

$$p(L) = \frac{e^{a(L)+b_m t(H,L)}}{\sum e^{a(L')+b_m t(H,L')}}$$

a(L) : Attractor weight
b_m : Travel mode coefficient
t(H,L) : Travel time

For computational reasons, only locations within a radius r_{ab} of the home location are considered.



Step 2: other activity locations

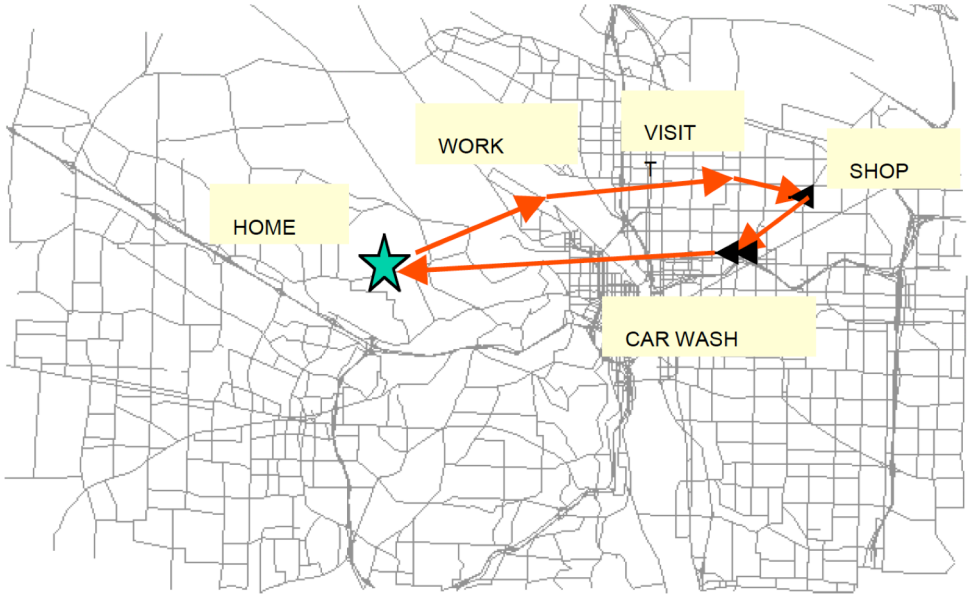
- To generate locations for other activities, we use a logistic multinomial choice chaining model.

$$p(L_1) = \frac{e^{b_{m1}t(L,L_1)+a(L_1)+b_{m2}t(L_1,H)}}{\sum e^{b_{m1}t(L,L_1')+a(L_1')+b_{m2}t(L_1',H)}}$$

- This means that the location choice for a non-anchor activity takes into account the locations of the previous and next anchor activities.

Step 2: other activity locations

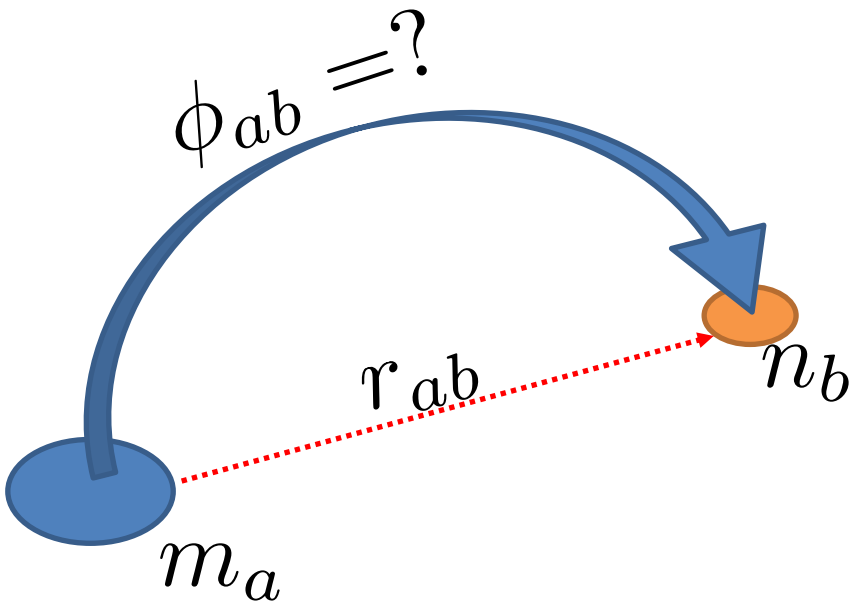
- To generate locations for other activities, we use a logistic multinomial choice chaining model.



Question:
How many people travel from *a* to *b*?

1. Gravity Model

2. Radiation Model



Gravity Models:
In Analogy with Newton's Law of Gravity

The Gravity Model is a fitting based model with a form of:

$$\phi_{ab} = C \frac{m_a^{\alpha} n_b^{\beta}}{f(r_{ab})}$$

m_a - source population size n_b - destination population size

r_{ab} - distance between and (Euclidian distance).

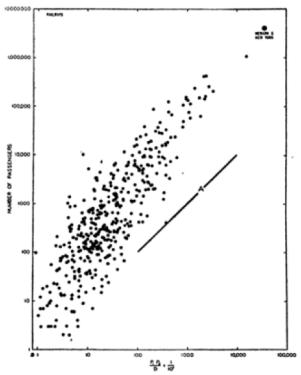
$f(r)$ - deterrence function, typically either $f(r) = r^{\gamma}$ or $f(r) = e^{\delta r}$

$\alpha, \beta, \gamma, \delta$ are the <u>fitting parameters</u>

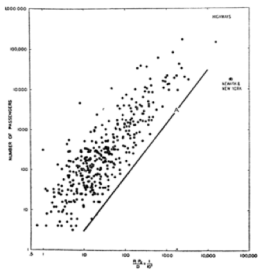
Gravity Models:
In Analogy with Newton's Law of Gravity

Originally, introduced by Zipf, G. K. (1946) as a hypothesis with the form:

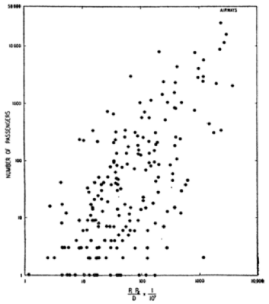
$$\phi_{ab} = \frac{m_a n_b}{r_{ab}^{\gamma}}$$



Railway, γ=1



Highway, γ=1.25



Airway, γ=?

Zipf, G. K. The P_1P_2/D hypothesis: on the intercity movement of persons. Am. Sociol. Rev. 11, 677–686 (1946)

Gravity Models:
In Analogy with Newton's Law of Gravity

Original 1-parameter fitting form is not satisfying.

Ex.1. Global cargo ship movements:

$$\phi_{ab} = \frac{\alpha_a \beta_b m_a n_b}{f(r_{ab})}$$

$$f(r_{ab}) = r_{ab}^{\gamma} e^{\delta r_{ab}}$$

Kaluza, P, et al.. The complex network of global cargo ship movements. J. R. Soc. Interface 7, 48, (2010)

Ex.2. Airway Traffic:

$$\phi_{ab} = C \frac{m_a^{\alpha} n_b^{\beta}}{f(r_{ab})}$$

$$f(r_a b) = e^{\delta r_{ab}}$$

Balcana, D, et al.. Multiscale mobility networks and the spatial spreading of infectious diseases. PNAS. 106, 51 21484 (2009)

*Requires two sets of parameters, for $r_{ab} > 300\ km$ and $r_{ab} < 300\ km$

Gravity Models:
In Analogy with Newton's Law of Gravity

Some Limitations of Gravity Models:

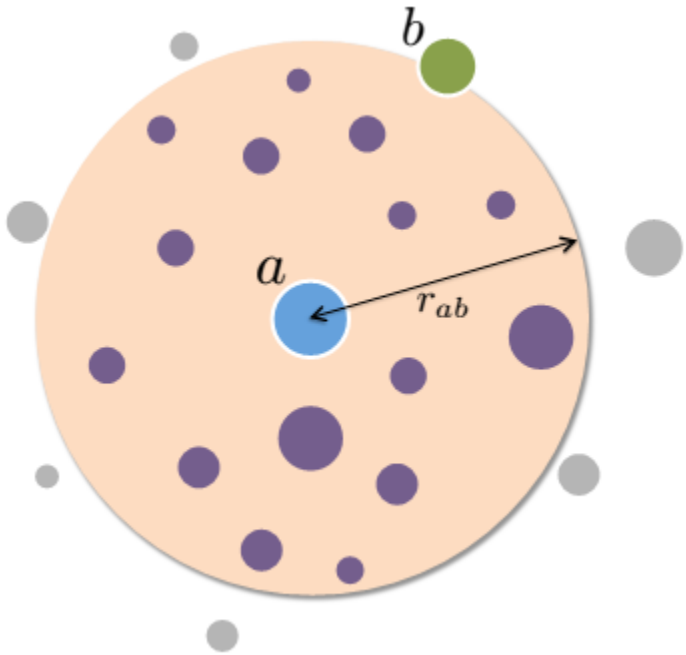
- Lacking theoretical guidance or a rigorous derivation
- Parameters having no physical meaning
- Requiring data to fit. Once the system is changed, previous parameters are no longer valid.

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." — John von Neumann

Nature **427**, 297 (22 January 2004) | doi:10.1038/427297a

Radiation Model

Simini F, et. al. (2012) Nature 484: 96-100
"A universal model for mobility and migration patterns"



$$\phi_{ab} = \zeta m_a p_{ab}$$

ζ – fraction of traveling people. An overall scaling factor for all locations

p_{ab} – probability for a person from location *a* (population m_a) to travel to location *b* (population n_b)

$$p_{ab} = \frac{m_a n_b}{(s_{ab} + m_a)(s_{ab} + m_a + n_b)}$$

s_{ab} – total population of nodes within the disk of radius r_{ab}, except m_a

Where does this equation come from?

A person living at location , is looking for a job with salary or benefits expectation of at least .

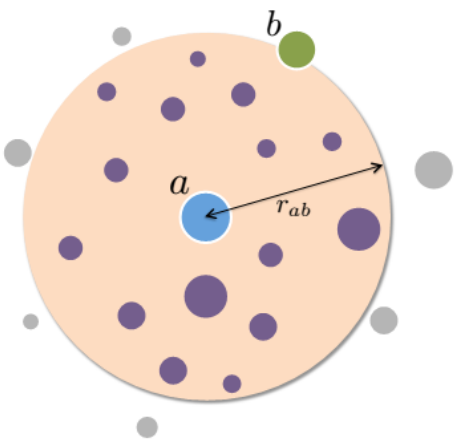
The number of job offers at a location is proportional to the population of that location.

Values of offers are denoted by .

This person only takes the job that satisfies his expectation ($z' \geq z$ the closest to .

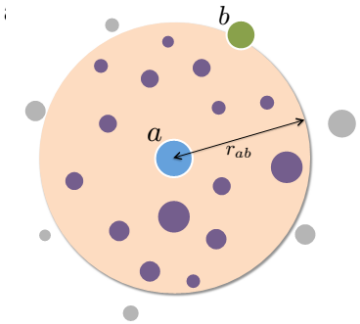
Let's assume that

and are drawn from the same distribution .



What is the probability that a person with salary expectation , living at location will take a job at location and not anywhere closer?

What is the probability that a person with salary expectation , living at location will take a job at location and not anywhere closer?



Simini et al. use an analogy with radiation emission and absorption processes to formalize this problem.

- The source, *a*, is assumed to be emitting particles with absorption thresholds, z_x^a
- The locations around *a* are assumed to absorb particles with probabilities, z_x^b
- A particle is absorbed by the closest location whose absorbance is higher that its absorption threshold.

From this, they work out the probability of one emission/absorption event between any two locations, and thereby obtain an analytical expression for the flux between them.

$$p_{ab} = \frac{m_a n_b}{(s_{ab} + m_a)(s_{ab} + m_a + n_b)}$$

Note that this turns out to be independent of the distribution, *p(z)*, of the emission and absorption thresholds.

The Generalized Radiation Model

Ren Y., et. al. (2014) Nature Comm. 5: 5347 "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges." DOI:10.1038/ncomms6347

The problem: s_{ab} is the population within r_{ab}. Here r_{ab} acts as a cost measure.

Notice: there is no coupling between the mobility flux and the network.

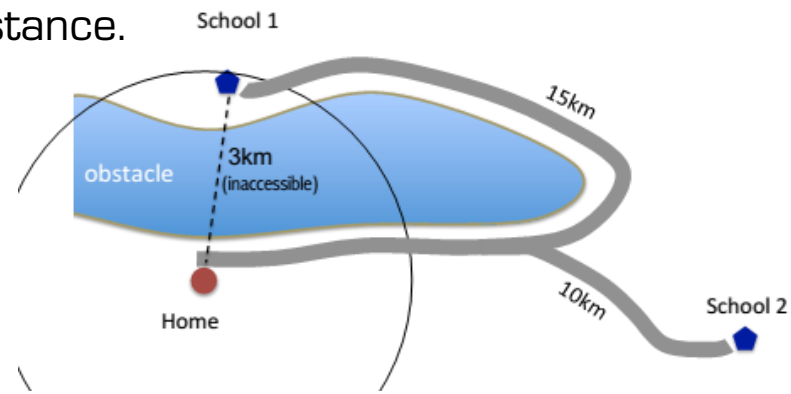
However:

People don't estimate costs based on Euclidean distance.

They have to travel on paths of the network.

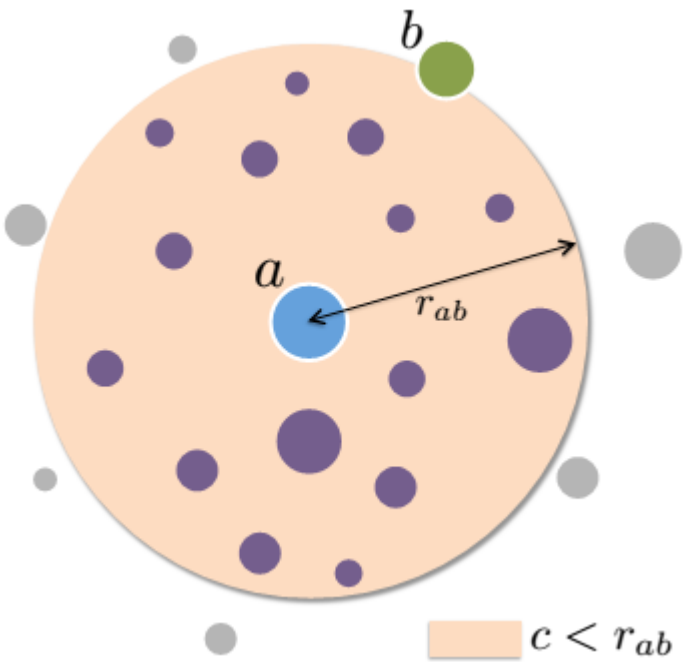
Hence, they will estimate costs on the network.

This couples the radiation model with the network.

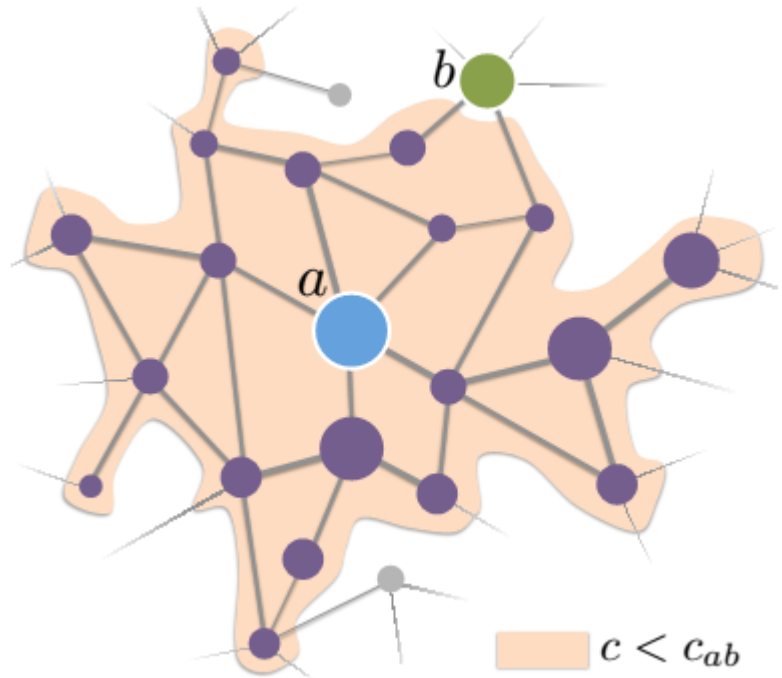


Therefore, we need to redefine the area population s_{ab}.

The Generalized Radiation Model



Original Radiation Model



Network Based Radiation Model

Here, c_{ab} is a general travel cost on the network.

Some useful references

- Zipf, G. K. The P_1P_2/D hypothesis: on the intercity movement of persons. Am. Sociol. Rev. 11, 677–686 (1946)
- Erlander, S. & N. F. Stewart. *The Gravity Model in Transportation Analysis: Theory and Extensions,* VSP, (1990)
- Kaluza, P, et al. The complex network of global cargo ship movements. J. R. Soc. Interface 7, 48, (2010)
- Balcana, D, et al.. Multiscale mobility networks and the spatial spreading of infectious diseases. PNAS. 106, 51 21484 (2009)
- Marta C. González; César A. Hidalgo & Albert-László Barabási, Understanding Human Mobility Patterns, *Nature 453*, 779-782 (2008)
- Simini F, et. al. "A universal model for mobility and migration patterns" Nature 484: 96-100 (2012)
- Ren Y., et. al. "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges." Nature Comm. 5: 5347 (2014)
- Yingxiang Yang, Shan Jiang, Daniele Veneziano, Shounak Athavale, Marta C. Gonzalez, "TimeGeo: a spatiotemporal framework for modeling urban mobility without surveys." PNAS.

Section

NETWORK CONSTRUCTION

Constructing synthetic information and networks: an overview

Synthetic Agents

Agents

- People
- Computer, mobile devices, etc.
- Hosts (plant models)

Places of interaction

- work
- residence
- school
- trading facility

Data

- Demographics
- Activity surveys
- Infrastructure
- GIS
- Trade
- Timeline, surveillance data

Interactions

Data driven

- Activities of people
- Social & professional relationships
- Transportation
- Commodity flow
- Long distance travel

Procedure driven

- Behavior
- Network protocols

- Social networks
- Friend networks
- Disease networks
  - HIV
  - Flu
  - Ebola
  - ⋮
- Addiction networks
- Radio networks
- ⋮

Structural properties

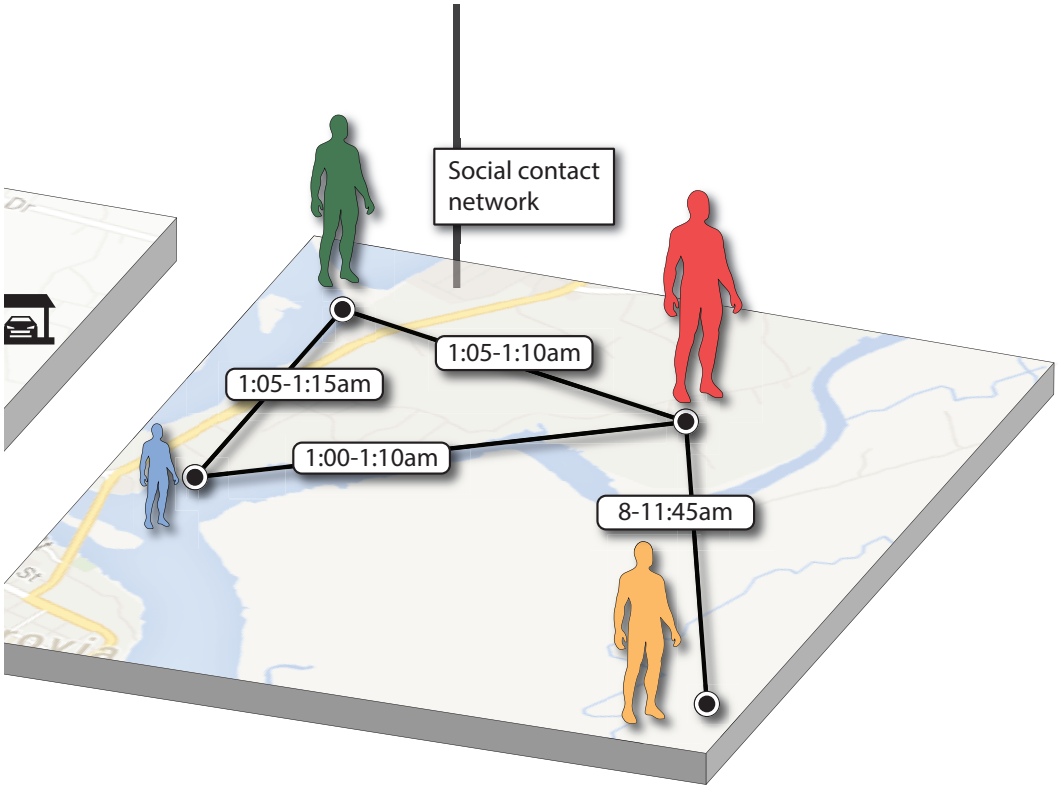
- Network measures
  - degree distribution
  - number of components
  - diameter
  - spectral radius
- Sensitivity to structural uncertainity.

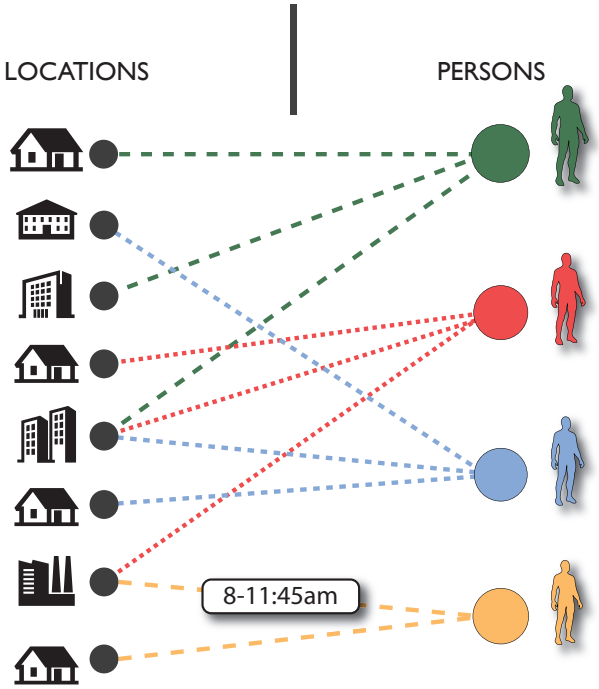
Dynamical properties

- Simulations
- Sensitivity to model choice and parameters

Analyses
V&V
UQ

Example: An induced contact network





SOCIAL CONTACT NETWORK

Using co-occupancy at residence and activity locations, we can infer contacts and their durations. From this we infer the social contact network.

PEOPLE-LOCATION NETWORK

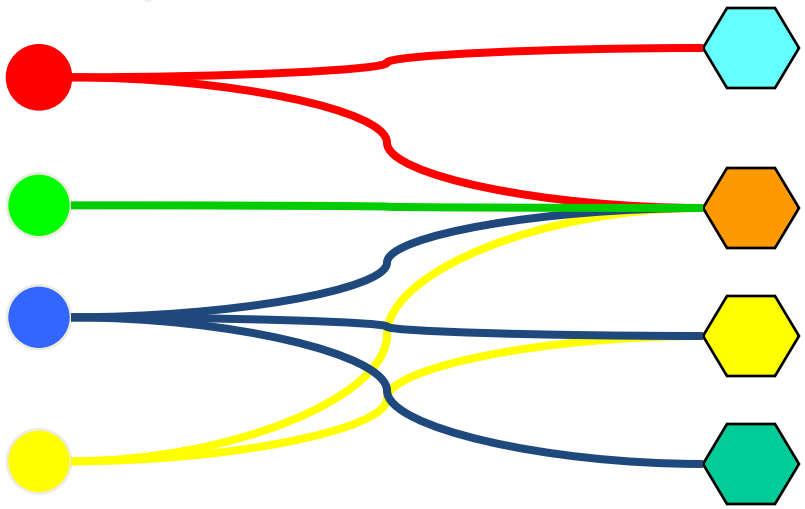
The computational model allows us to assign people t locations with durations of visit and through this determine their contacts and interactions.

Relational, dynamic social contact networks

People (8 million)

Locations (1 million)

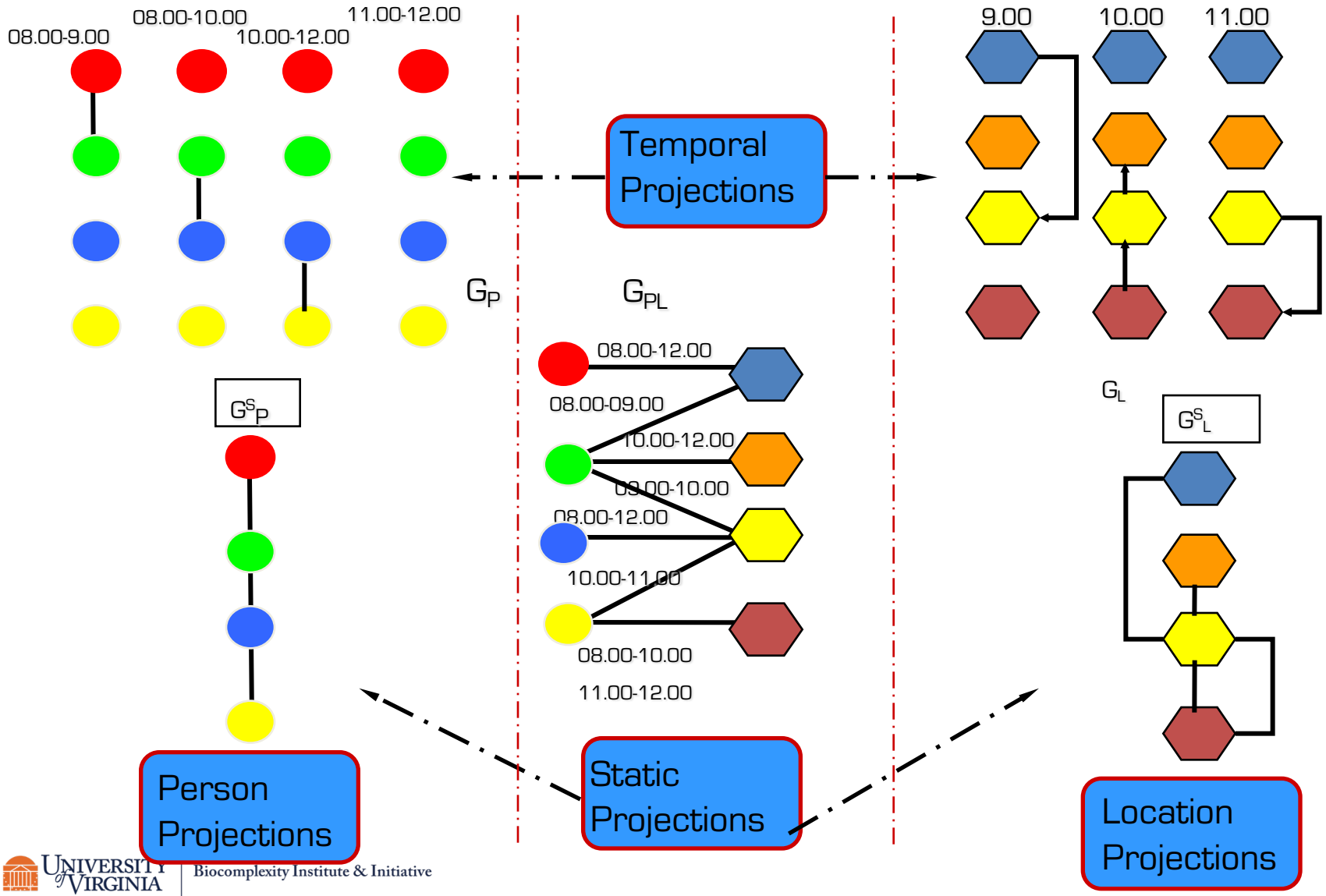
Vertex attributes:
- age
- household size
- gender
- income
- ...



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

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

Various projections of person-location temporal social visitation network



Refining subnetworks

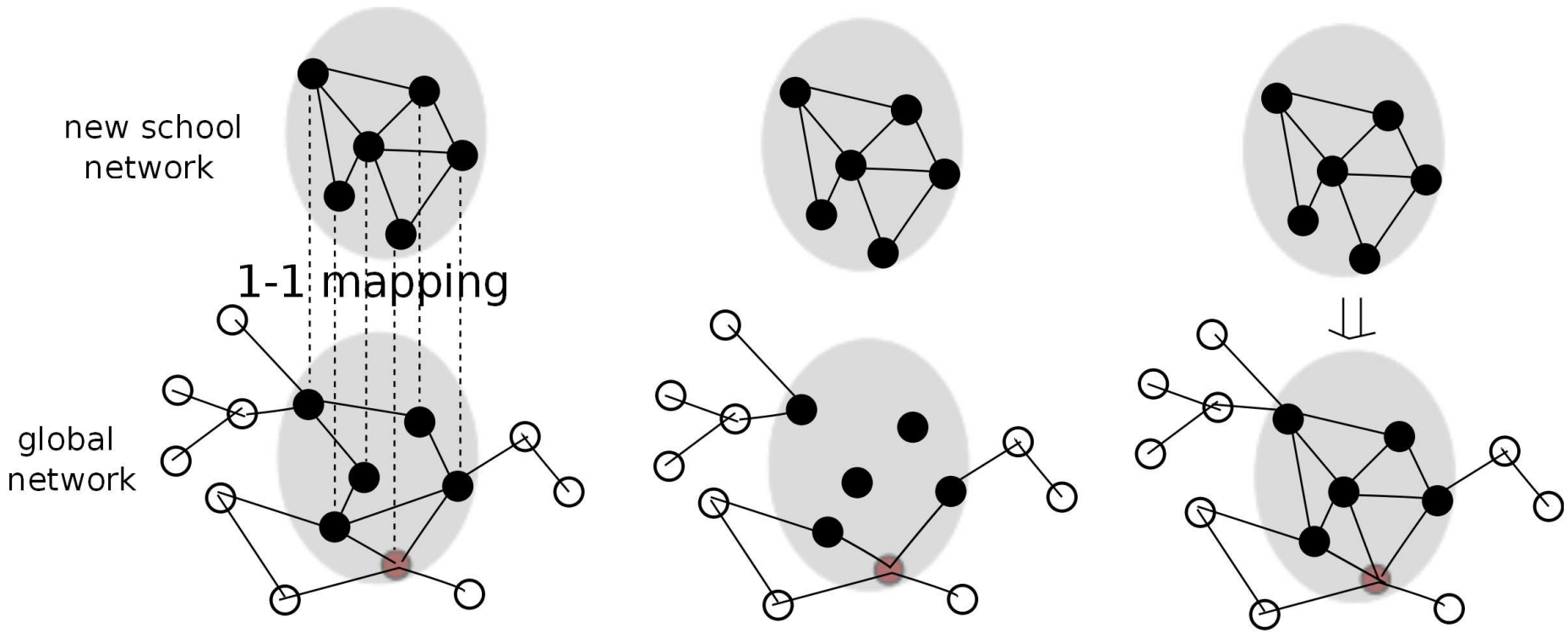
- Refine a subpopulation and the associated subnetwork in an existing synthetic population / network when more detailed data on the subpopulation becomes available.
  - Special locations: schools, office work/college campuses, hospitals, military bases, hotels, etc.
  - Additional data: mobility data from surveys, sensors, phones, fitbits, etc.

Subnetwork construction

- Identify subpopulation in the whole synthetic population
- Construct subnetwork as follows:
  - Activity data: can replace existing activities occurring at the location in each individual's daily activity sequence
    - Class registration and schedule data in a high school describes how students move between rooms (sublocations) during the day.
    - Activity vignettes for different cohorts describe activities of military personnel and civilians on post.
    - Patient admission data and hospital work log data describe how patients and health care workers move between rooms (sublocations).
    - Activity survey data for slum subpopulations in Delhi describes the daily activities of people living in slums.

    Note that these individuals' activities outside of the location can be kept.
  - Connection data: is sometimes directly available.
    - Collected by close proximity sensors carried by participants: e.g. Salathe et al. 2010 describes such a data set from a high school.

Embedding a subnetwork



(a) identifying and mapping (b) removing school edges (c) embedding

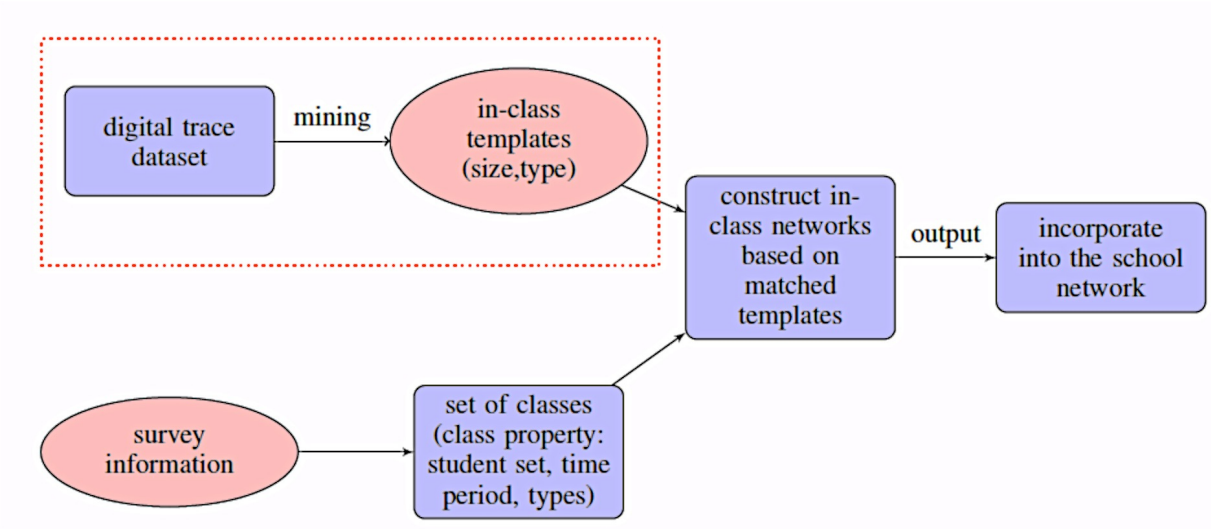
Subpopulation and network inference: Survey and administrative information versus digital traces

- Subpopulation and network attributes can be obtained via various sources:
  - Subjective surveys, administrative data, digital traces from sensors, phones, call records.
- Each method has its pros and cons:
  - (*i*) Accuracy of the data collection method, (*ii*) human error, (*iii*) Cost, (*iv*) Coverage,(*v*) Resolution (*vi*) Access to data and (*vii*) generalizability
  - E.g. Call data records are not the best for synthesizing networks for specific locations; data although collected by phone companies is not available below the resolution of cell towers at best

Example: Synthesizing high school networks and embedding them

	school 1	school 2	school 3
no. of students	1116	958	425
no. of students in grade [9,10,11,12] respectively	[293,266,298,259]	[311,231,220,196]	[107,115,105,98]
no. of classrooms	78	66	45
no. of classes per day	9	5	7
duration of classes (minutes)	45	90	50-56

Table 1: Statistic of real class schedules (data from registration information).



Hybrid approach that combines digital traces with administrative and survey information

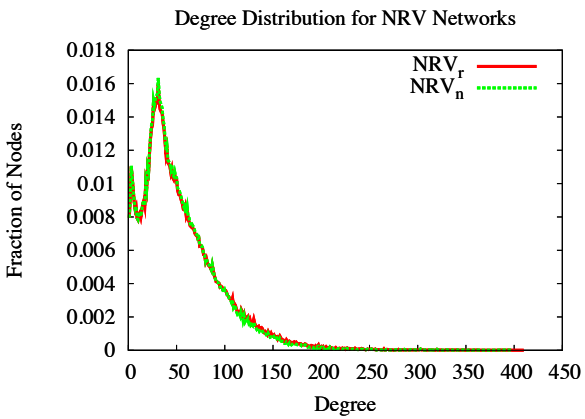
- Administrative information obtained from schools and anonyimized
- Digital traces collected by RFID sensors and provide in-class networks

Challenges

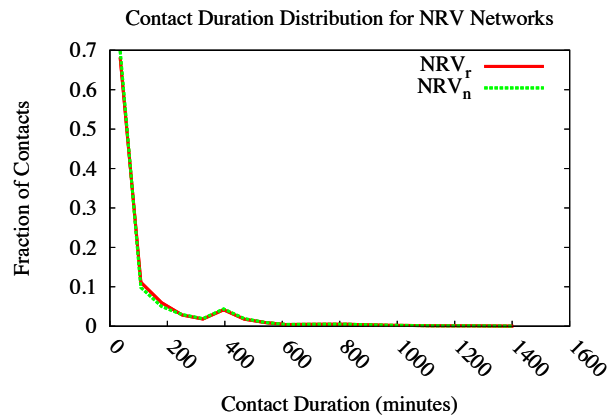
- Uncertainty and misalignment in subpopulation identification
  - Uncertainty: Real (anonymized) data of high schools (in NRV): Multiple ways to match real and synthetic high school students
  - Misalignment: In the whole synthetic population, the number of people identified through activity locations may not equal the actual size of subpopulation.
    - Identified > actual: need random selection
    - Identified < actual: need expansion (relaxation of identification criteria)
- Adequacy of the refinement: not all details matter.

Do details matter?

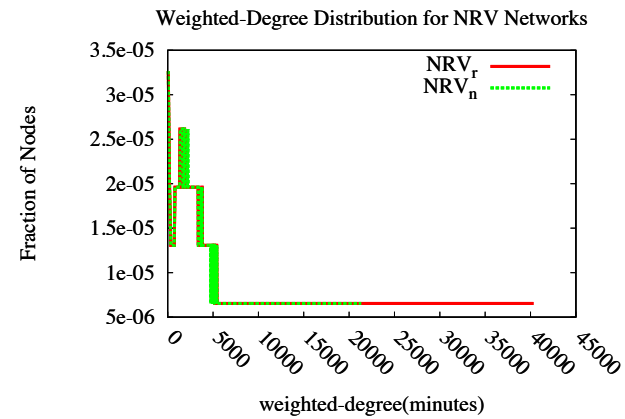
Embedding refined NRV high school subnetworks in NRV synthetic network did not significantly affect the network structure of the whole network.



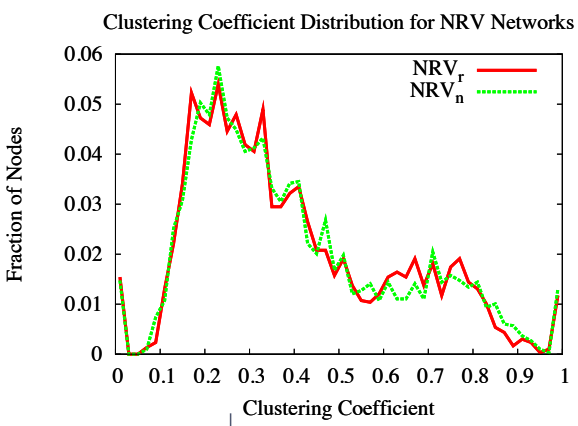
(a)



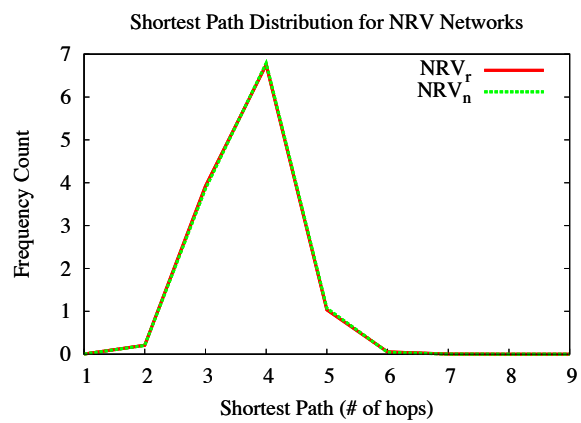
(b)



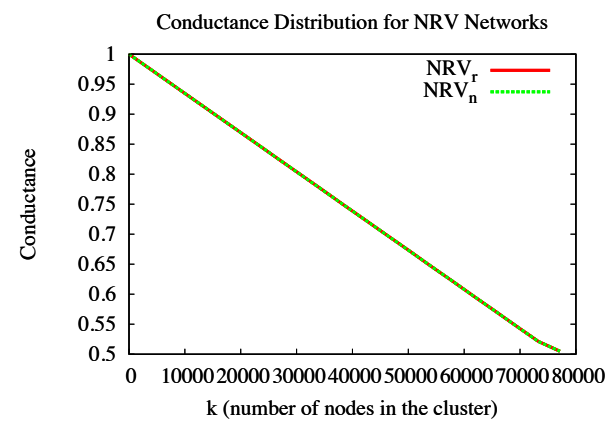
(c)



(d)



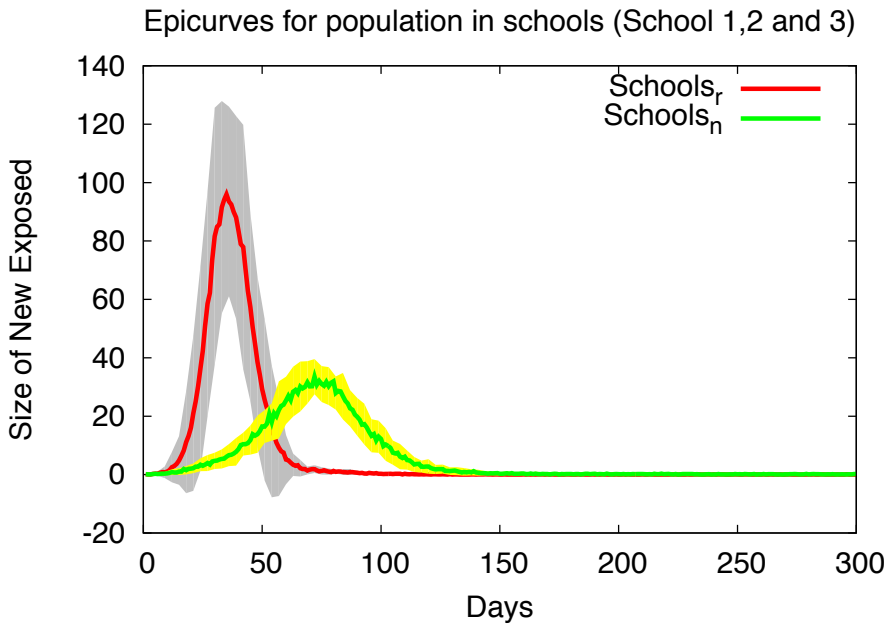
(e)



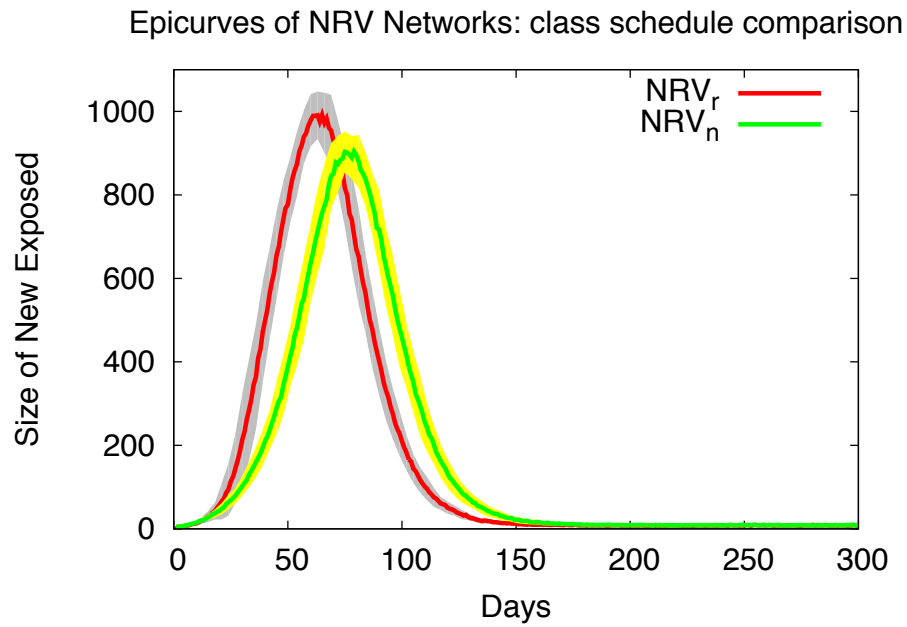
(f)

Do details matter?

But the refinement significantly affects the infectious disease dynamics on the whole network, even though the subpopulation is only a small fraction (*1.7%*) of the whole population.



(a)



(b)

Comparing subnetwork refinement

type	whole population	subpopulation	extra detailed data
high school	New River Valley, VA	students attending three high schools	class registration data of each student
hospital	Virginia	patients and health care workers in a hospital	electronic medical records and direct observations of healthcare workers
military base	several metropolitan regions in US	people residing and working at military bases	demographic distributions and activity vignettes (templates) of different cohorts
slum	Delhi, India	people residing in slum areas	demographic and activity survey on households in slum areas, Delhi

Comparing subnetwork refinement

type	subpopulation identification	subnetwork construction	subnetwork embedding
high school	individuals in synthetic population who have school type activities in the three schools	usual method is used to combine activity sequences with sublocation model to derive co-location contacts; we have studies different sublocation models	replace existing in-school contacts with new contacts
hospital	patients can be identified in an epidemic simulation; healthcare workers are those having work type activities in this hospital	activities are created for each healthcare worker, patient, and visitor, from which a people-location network is generated	N/A (not done yet)
military base	synthetic households whose home locations are within (not enough) or near the bases	on-base activity sequences are assign to individals in subpopulation by cohort; usual method is used to combine activity sequences with sublocation model to derive co-location contacts	replace existing on-base contacts with new contacts
slum	synthetic households whose home locations are within the slum areas (defined by polygons); some living in the same wards are also selected and moved into slum areas	activity templates are derived from survey data; activity sequence is assigned to each individuals in subpopulation according demographics; usual method is used to combine activity sequences with sublocation model to derive co-location contacts	replace existing edges of individuals in slum subpopulation with new edges

Extensions

- Other networks we have constructed
  - Large Geographical areas:
    - Military base: Administrative data available; daily activities are different; mixture of civilians and military personnel; work and residential quarters
    - Slums: high density living spaces, data is not trivial to obtain; within-house network is not easy; can span large geographical area
  - Small set of Buildings
    - Hospitals: detailed data obtained by shadowing the personnel; patients in rooms for long periods; staff moves in and out at short time scales
    - Office Building: Detailed maps available, people reside in their space for long periods; relatively low levels of mobility

Some useful references

1. C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. PLOS ONE, 5(7), 2010.
2. H. Xia, J. Chen, M. V. Marathe, and H. S. Mortveit. Synthesis and refinement of detailed subnetworks in a social contact network for epidemic simulations. In Social Computing, Behavioral-Cultural Modeling and Prediction, volume 6589 of Lecture Notes in Computer Science, pages 366–373. 2011.
3. M. Salathe, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. A high-resolution human contact network for infectious disease transmission. Proc. National Academy of Science,, Dec. 2010.
4. G. E. Potter, M. S. Handcock, I. M. Longini, & M. E. Halloran. Estimating Within-School Contact Networks To Understand Influenza Transmission. The annals of applied statistics, 6(1):1–26, 2012.
5. C. Hlady, D. Curtis, M. Severson, J. Fries, S. Pemmaraju, A. Segre, T. Herman, and P. Polgreen. A near-real-time method for discovering healthcare worker social networks via wireless devices. In 47th Annual Meeting of the Infectious Disease Society of America, Oct. 2009.
6. Huadong Xia, Jiangzhuo Chen, Madhav V. Marathe, Henning S. Mortveit, Marcel Salathé: Synthesizing Social Proximity Networks by Combining Subjective Surveys with Digital Traces. IEEE CSE 2013: 188-195
7. Leecaster M, Toth DJA, Pettey WBP, Rainey JJ, Gao H, et al. (2016) Estimates of Social Contact in a Middle School Based on Self-Report and Wireless Sensor Data. PLoS ONE 11(4).
8. Beutels P, Shkedy Z, Aerts M, Van Damme P. Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. Epidemiology and infection. 2006;134:1158–66.
9. A. Conlan, K. Eames KTD, et al. Measuring social networks in British primary schools through scientific engagement. Proc. Royal Society B: Biological Sciences. 2010.

Section

POPULATION SYNTHESIS FOR DATA-POOR REGIONS

New data sources and/or classes

- WorldPop.org — Population density
- ADCW 7.2-7.3 — NDSSL-customized database from ADCi
- US Census Bureau — Auto-harvested age/gender distributions
- European Social Survey — Household composition
- ILO — Labor force statistics
- The World Bank — School enrollment rates
- OECD — Daycare and preschool rates
- ORNL — LandScan Population Counts
- ESRI — Geographic Boundary Shapefiles
- OSM — Road networks
- Additional data for specific target countries

Methodology Overview

- For regions with limited available data
- We model people w/attributes, household compositions, home- and activity locations, human activities and where they occur, and all contacts between people on a typical day.
- Activity types: Home; Work; School; Other

Model Input Data Sets

- Joint age/gender distributions (U.S. Census)
- Household size distribution
- School distributions by type
- Workplace size distributions
- Main daytime activity type by age and location type

Model Inputs and Parameters

- Population density estimates (LandScan; WorldPop)
- Activity templates
- Adult age limit (Eric – what is the list of features governed)
- Maximal commute distance (default is 80km)
- Target commute distance (currently 6km for mean value; exponent in gravity model for distance term)

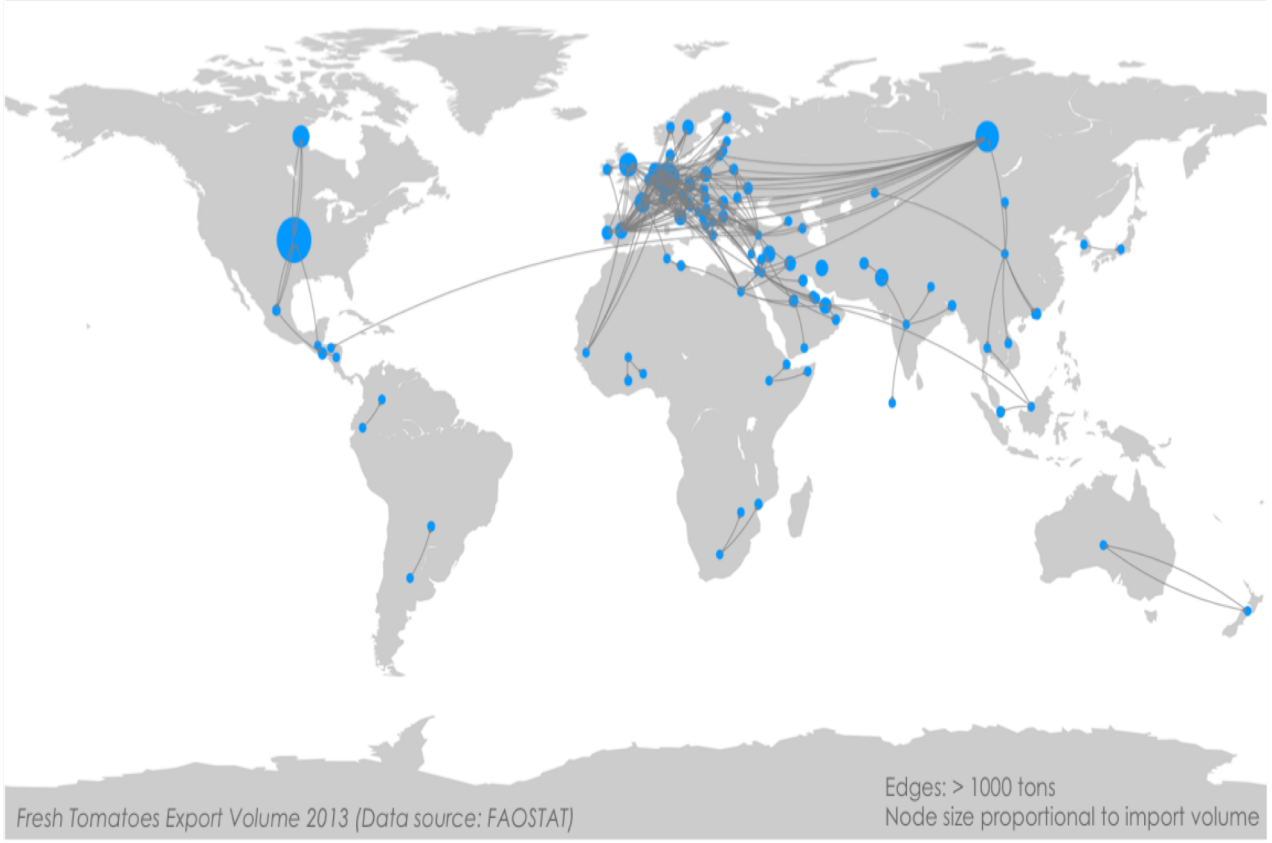
Population augmentation: activities & locations

- The constructed populations have basic demographic information, properties and relations
- For specific scenarios and analyses, one generally needs to introduce additional, scenario-specific properties. Examples:
  - Funeral attendance (West Africa; Ebola)
  - Funeral locations (West Africa; Ebola)
  - Military sub-population with activities (National guard force readiness)
  - Integration of mosquito prevalence map and risk maps (vector-borne diseases: Malaria; Zika; Dengue)

Population augmentation: long distance travel and transportation of commodities

- For pandemics and many other phenomena, it is essential to capture the impact of travel (people) and transportation (commodities)
- Populations and models must consider long distance domestic and international T&T
- Challenge: a myriad of existing data source; new ones come online: how to ingest and integrate?
- From a systems perspective, we need a framework consisting of a ***harmonizing database***, flow modeling, and translations/augmentation
- Purpose: support quick, unified handling of broad ranges of data sources when augmenting for T&T

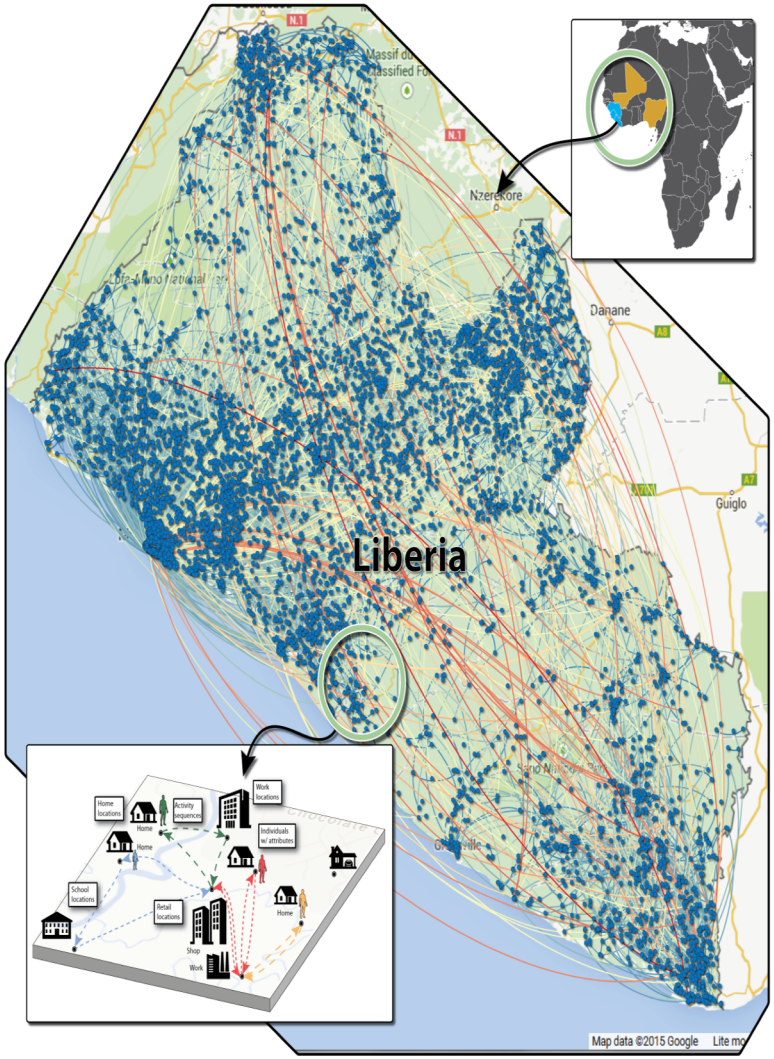
Population augmentation: long distance travel and transportation of commodities



Commodity flow

Population augmentation: long distance travel and transportation of commodities

- Framework naturally supports most patch models
- A radiation-based model can be used for capturing commuter flow on road networks



Section

VALIDATION, VERIFICATION AND SENSITIVITY ANALYSIS

V&V and UQ

- Validation based on
  - Independently collected data: surveys, surveillance information, etc.
  - Expert opinion
- Uncertainty quantification
  - How does uncertainty in the network structure affect its properties?
  - How good is a network based on sampled data?
  - E.g. robustness of rankings based on k-core [PKDD'13]

Criteria for measuring the utility of network modification

- Along the three versions of methodologies, we
  - Add New information (residential networks)
  - Replace raw data with higher resolution data (MMI)
  - Correct wrong data (activity survey)
- We expect to see a higher "quality" network through the modifications
  - What does "quality" mean?
  - How do we compare two networks?
- Metrics of data quality
  - Accuracy: How close the measured value is to the actual value in the real world
  - Precision: how consistent the simulation outcomes are across replicates.
  - Fidelity: The number of featural components included in the model
  - Resolution: The granularity of the features represented
  - Model Purpose: Affected Epidemics

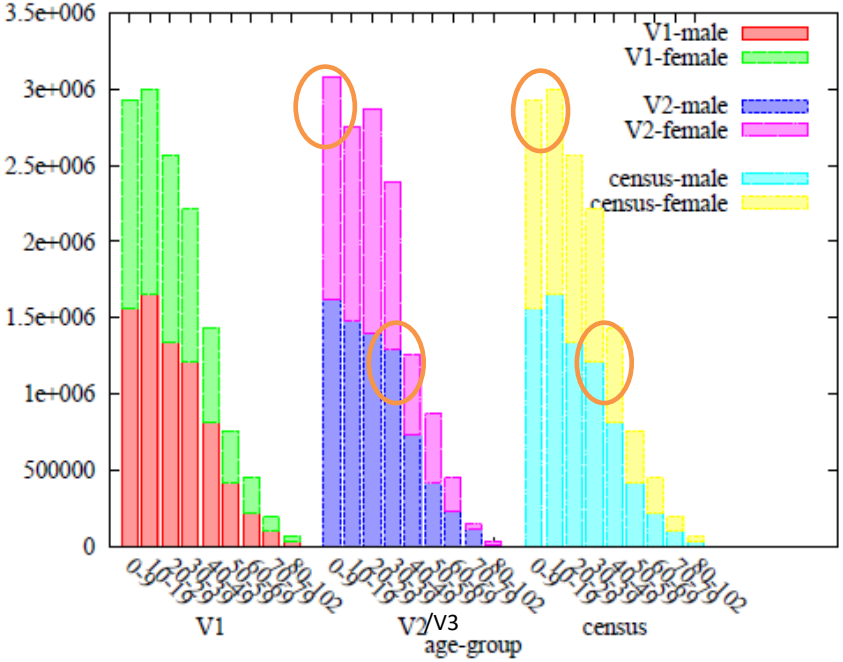
How can we compare two synthetic networks?

- No single metric which serves to "validate" a system dynamics model (Jay Forrester)
  - A number of metrics, organized as *multiple levels/parts of the synthetic network*
    - Entity level: the population, built infrastructure and their layout
    - Collective level: validate against aggregate statistics.
    - Network level: structural properties
    - Epidemic dynamics level: policy effects

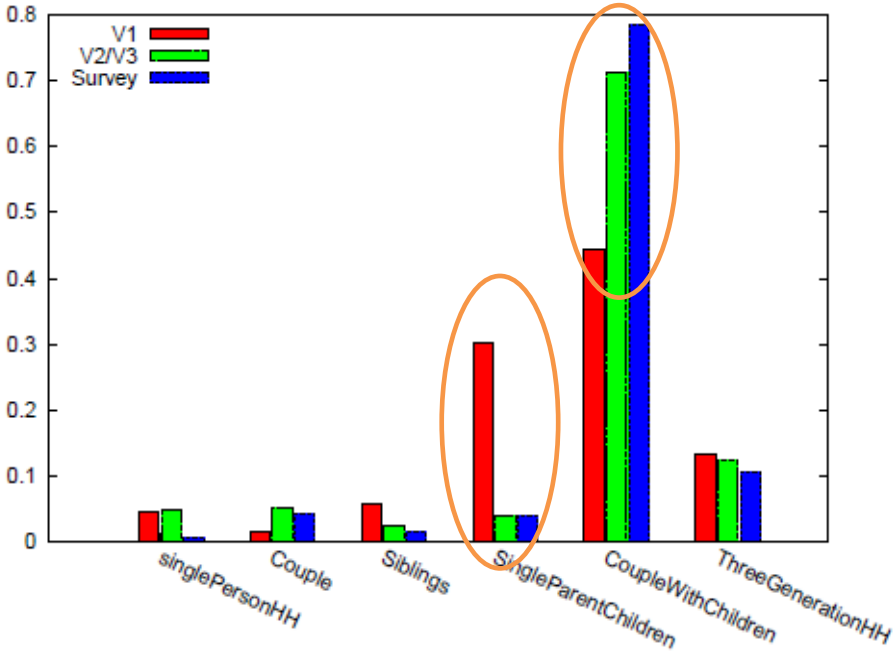
Analysis and comparison of three synthetic networks for Delhi NCR

- Different population demographic structure
  - Individual level: V1 matches census data better
  - Household level: V2/V3 matches survey data

Individual level age-gender structure

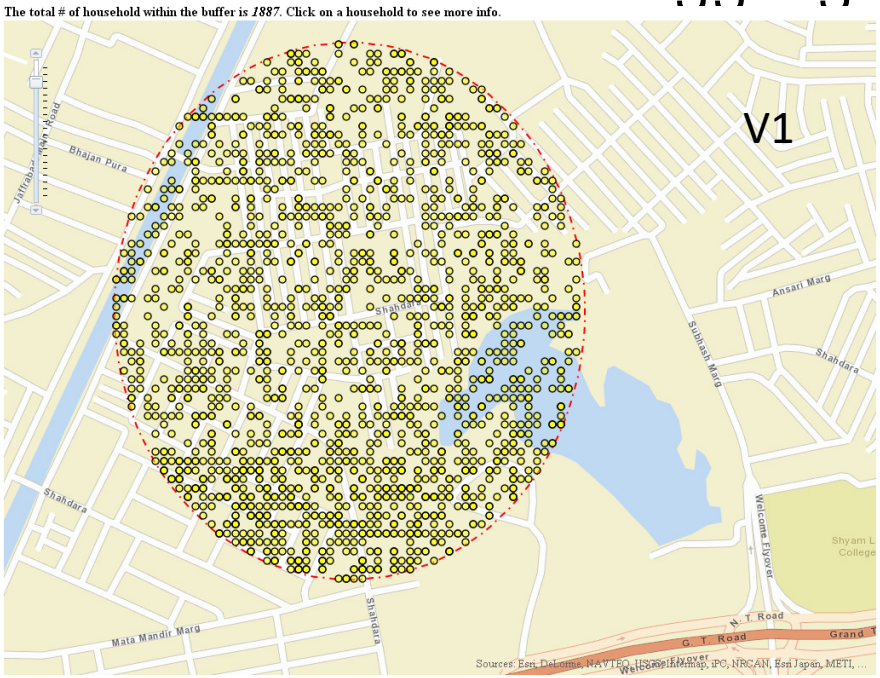


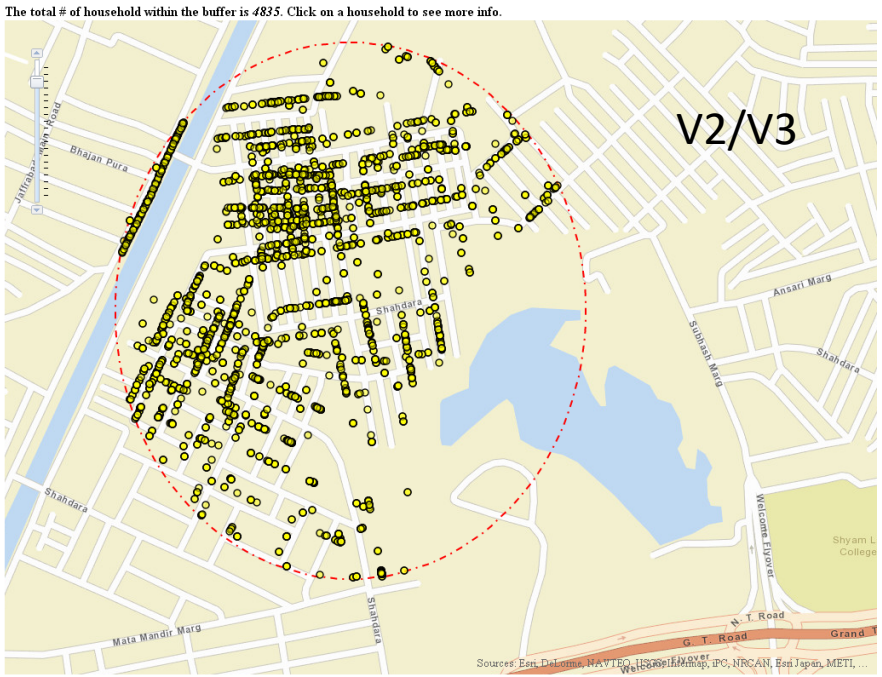
Household structure distribution



Analysis and comparison of three synthetic networks for Delhi NCR

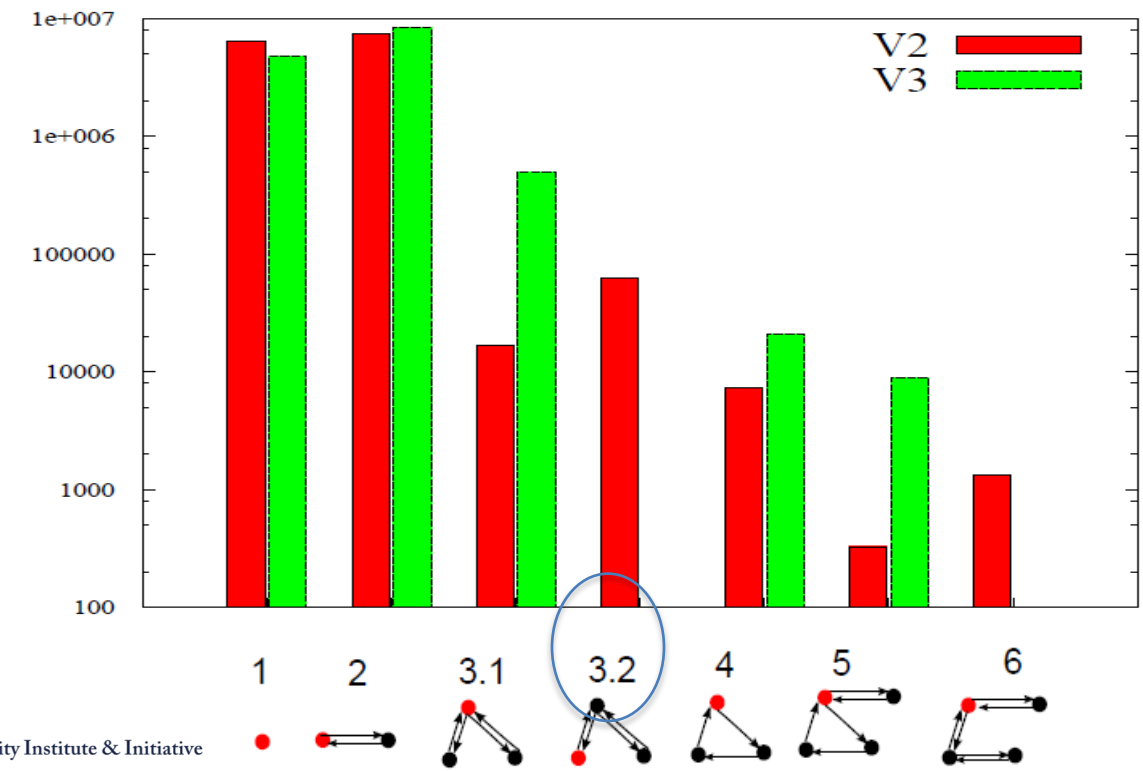
- Location layout
  - The three versions are progressively of higher "realism" in disaggregated structure of location



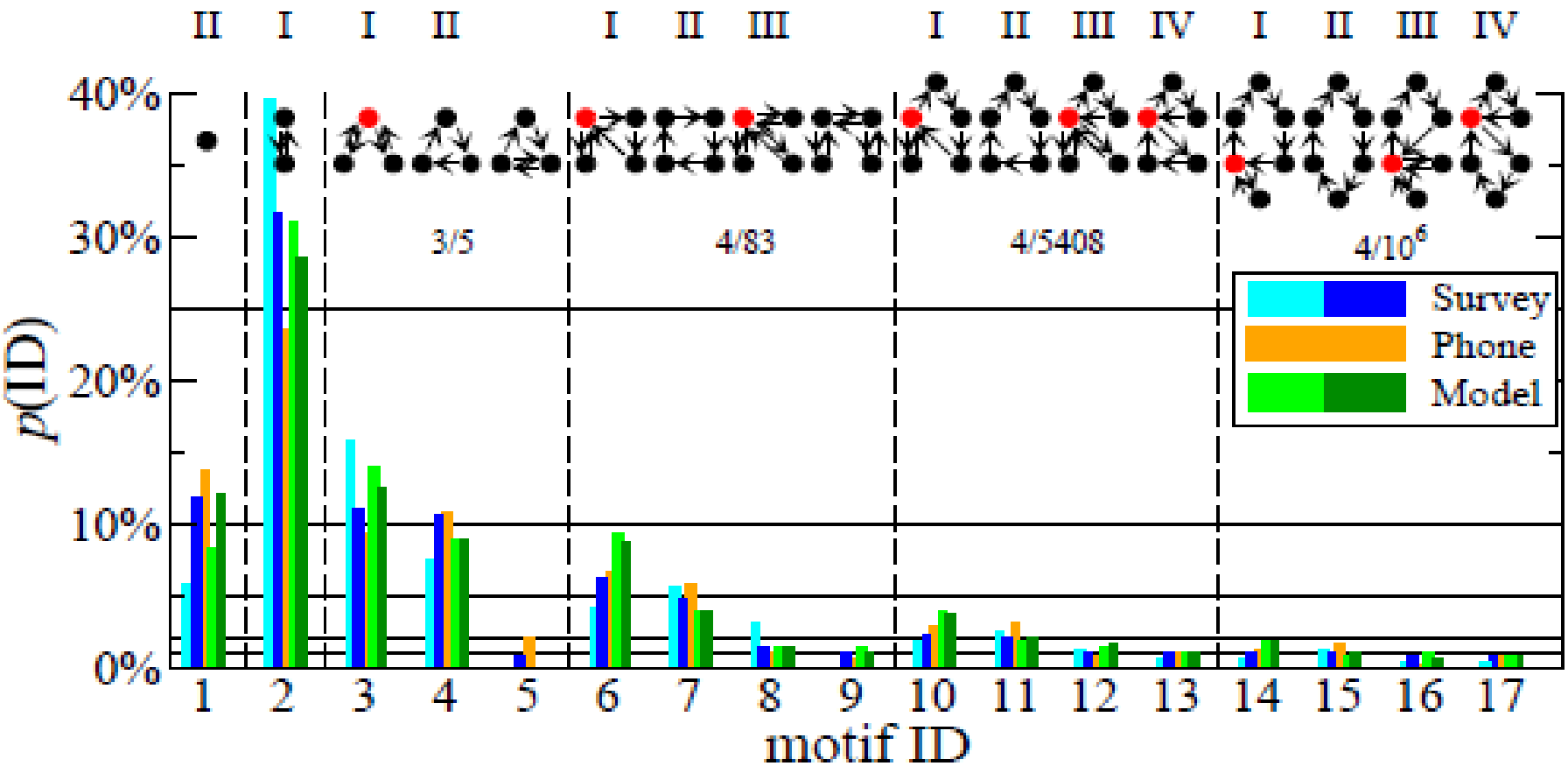


Analysis and comparison of three synthetic networks for Delhi NCR

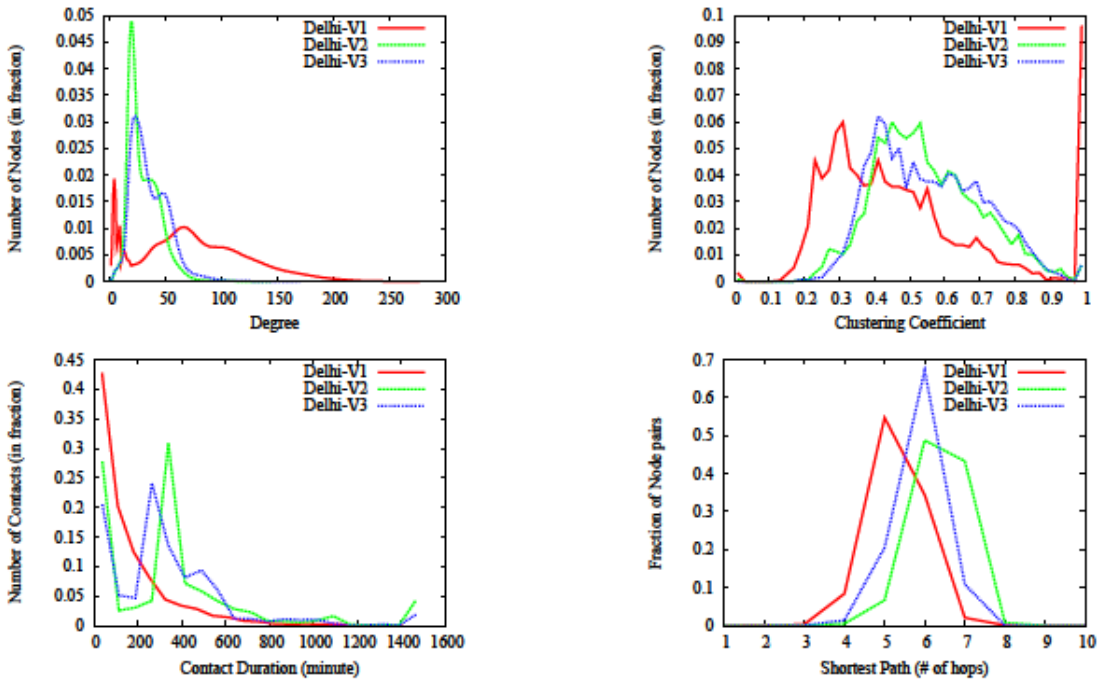
- Different activity patterns
  - In the figure are distributions of 7 most frequent human mobility motifs in V2 and V3 (red dots represent home).



Analysis and comparison of three synthetic networks for Delhi NCR



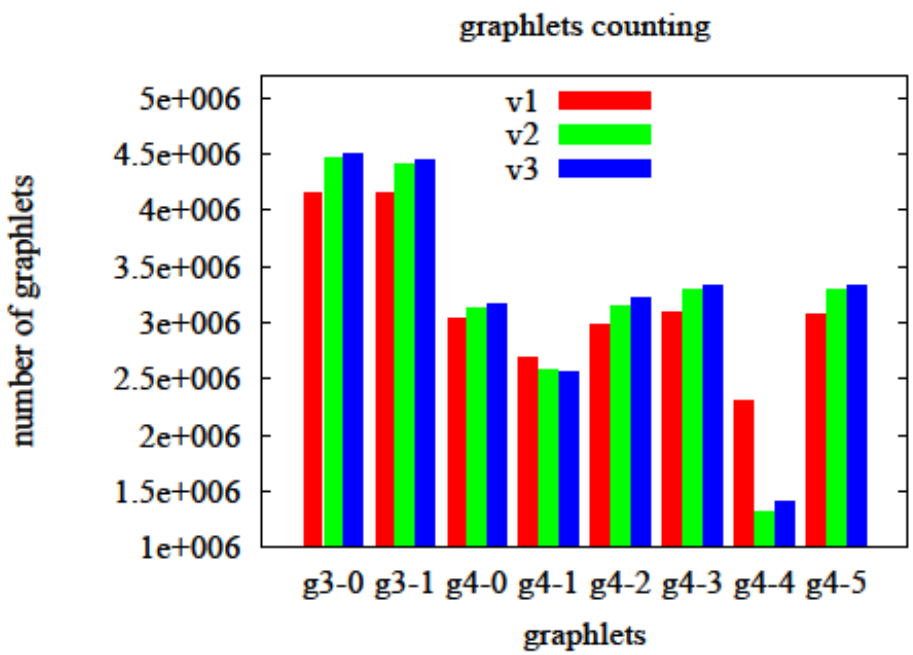
Analysis and comparison of three synthetic networks for Delhi NCR

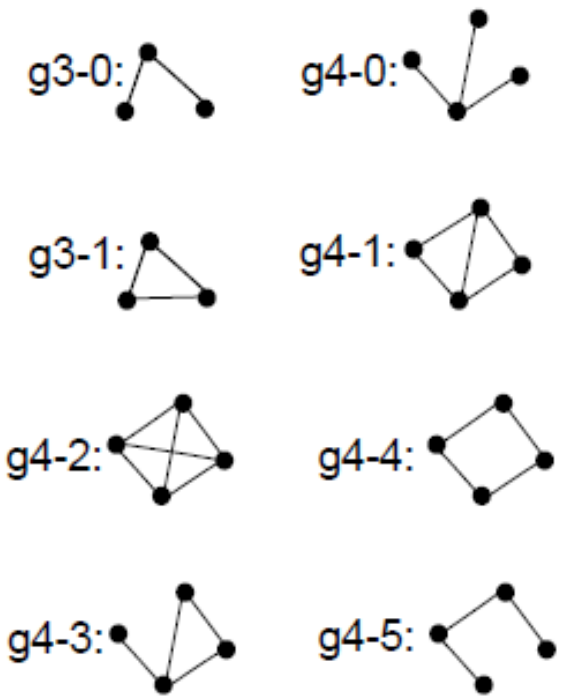


- Delhi-V1 has a group of people with very low degrees and high clustering coefficients.
- The abnormal peaks disappear in Delhi-V2/Delhi-V3 due to the modeling of residential contacts
- Residential contacts modeling increases the model fidelity

Analysis and comparison of three synthetic networks for Delhi NCR

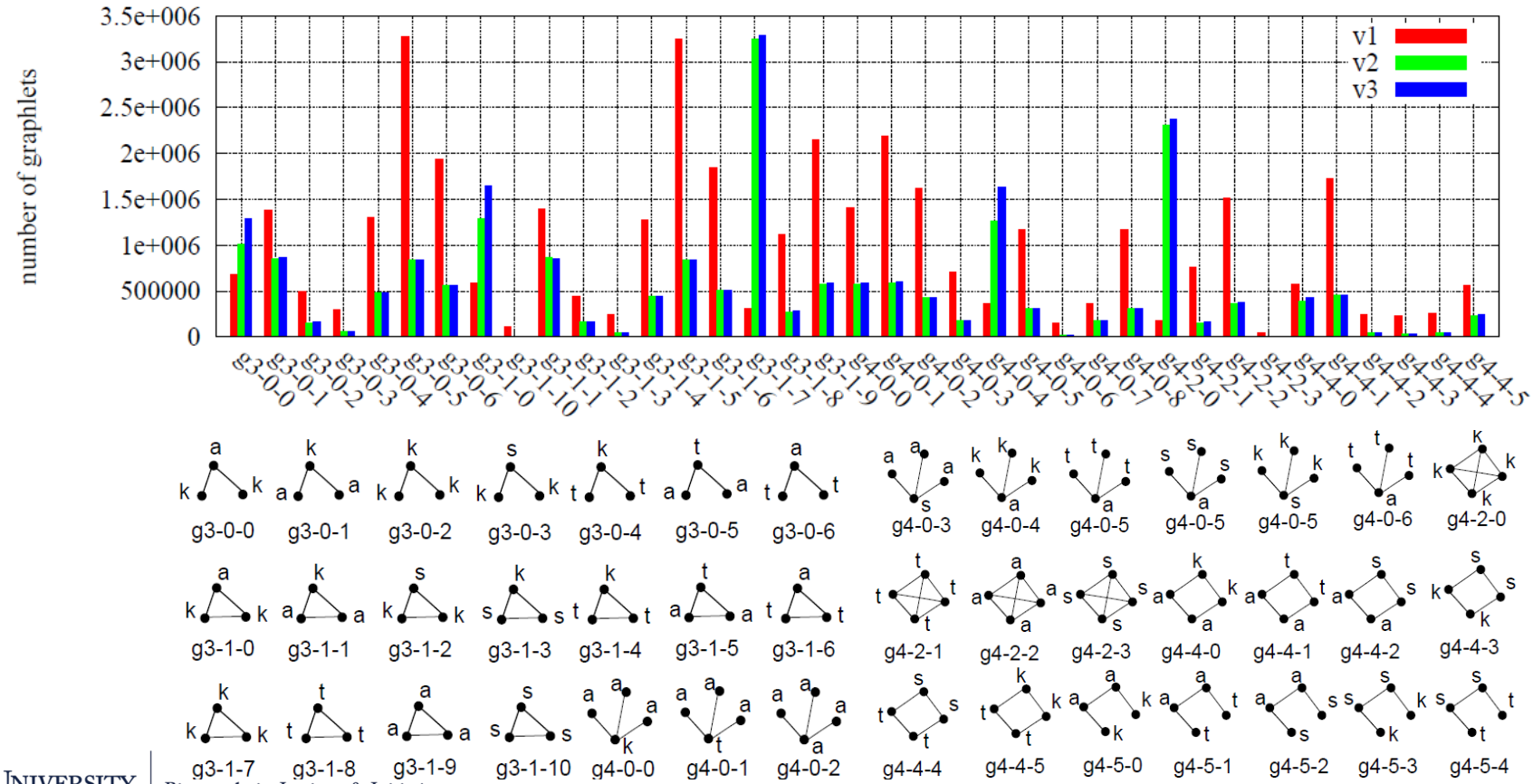
- The distribution of unlabeled graphlets in three networks are very similar.





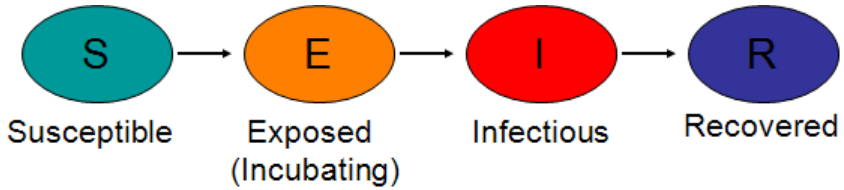
Analysis and comparison of three synthetic networks for Delhi NCR

- The distribution of labeled graphlets in three networks differs significantly.



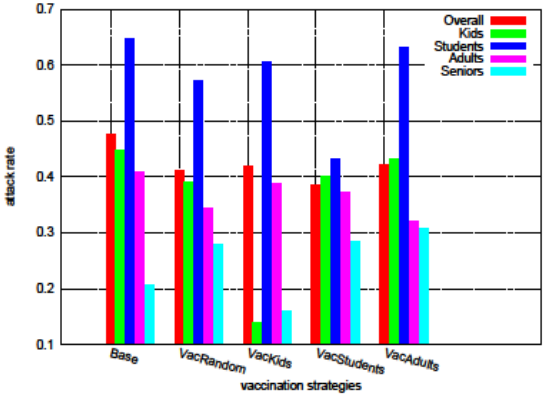
Disease Spread in a Social Network

- *Within-host disease model*: SEIR

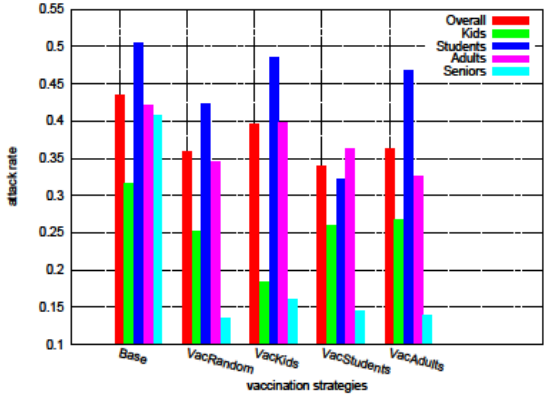


- *Between-host disease model*:
  - probabilistic transmissions along edges of social contact network
  - from infectious people to susceptible people
- *Vaccination Strategies:*
  - implemented at the beginning of epidemic;
  - vaccinate only a selected subpoplation;
  - vaccine stockpile is 10% of the whole population;
  - compliance rate 67%

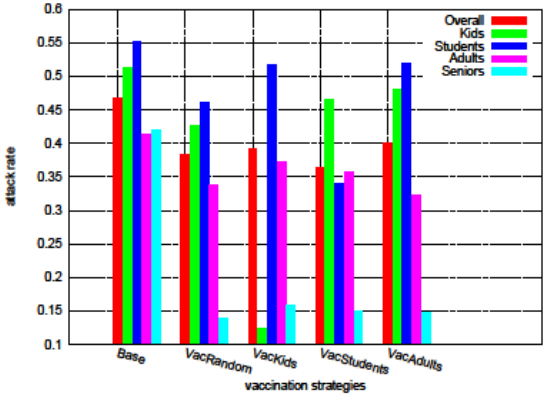
Analysis and comparison of three synthetic networks for Delhi NCR: Intervention Comparison: the Impact of Age-Mixing Patterns to Epidemics



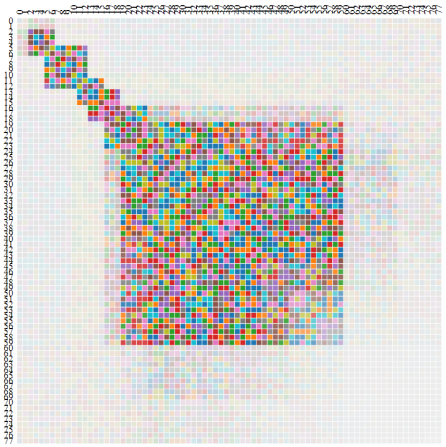
(a) *DelhiNetV1*

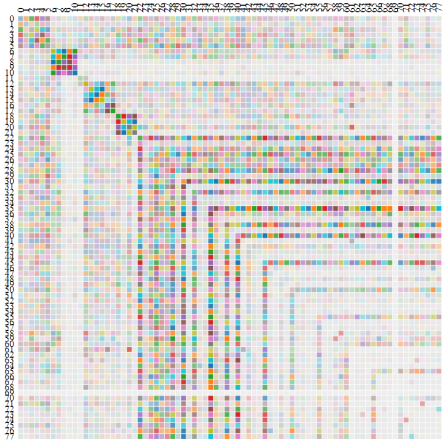


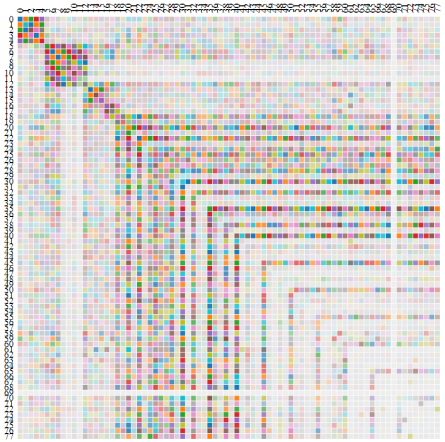
(b) *DelhiNetV2*



(c) *DelhiNetV3*







SubSection

SENSITIVITY ANALYSIS VIA LOCAL PERTURBATION

Network Sensitivity and importance of details

- A common assumption in the current literature: local properties of a network (such as degree distribution) are *often adequate* to determine dynamical properties of the system abstracted by the network

- We show that local properties such as the degree distribution or assortativity might not be sufficient to characterize the global dynamics of reaction-diffusion processes in realistic social networks.

- Define: Network Sensitivity (NS) as the change in the dynamical outcome as a function of network structure and propose a systematic way to study NS.

- Method: By a simple edge switch operations, we generate a Markov chain of graphs with fixed structural constraints (e.g. degrees) and study disease dynamics on these families of graphs

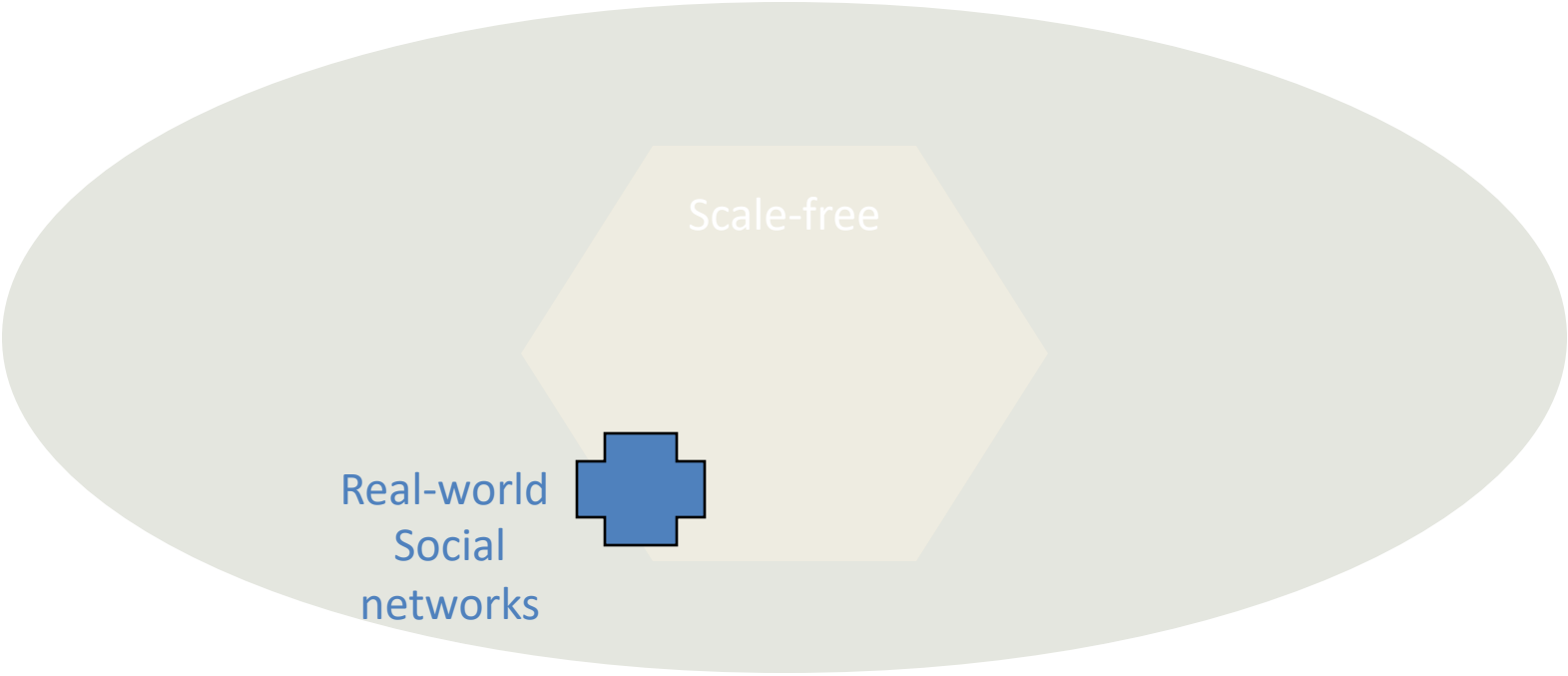
Consider the space of all graphs

There are $2^{O(N^2)}$ graphs on N vertices. When the vertices are distinctly labeled, many of these graphs are topologically distinct.

Some of these graphs have a particular degree distribution

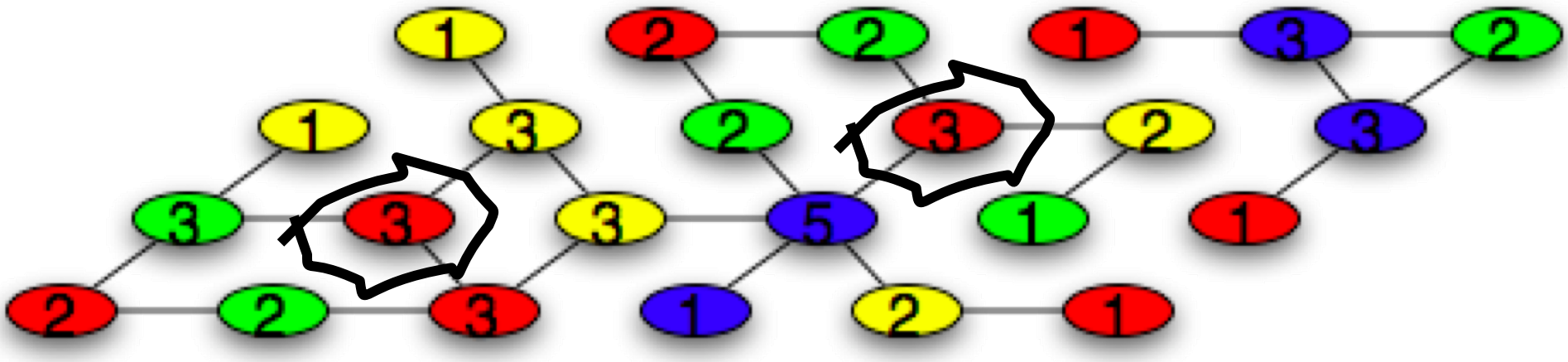
Proofs about properties of scale-free graphs (e.g. “no epidemic transition in a scale-free graph”) describe mean outcomes across the hexagonal region, sometimes variances.

Some of these might represent synthetic networks

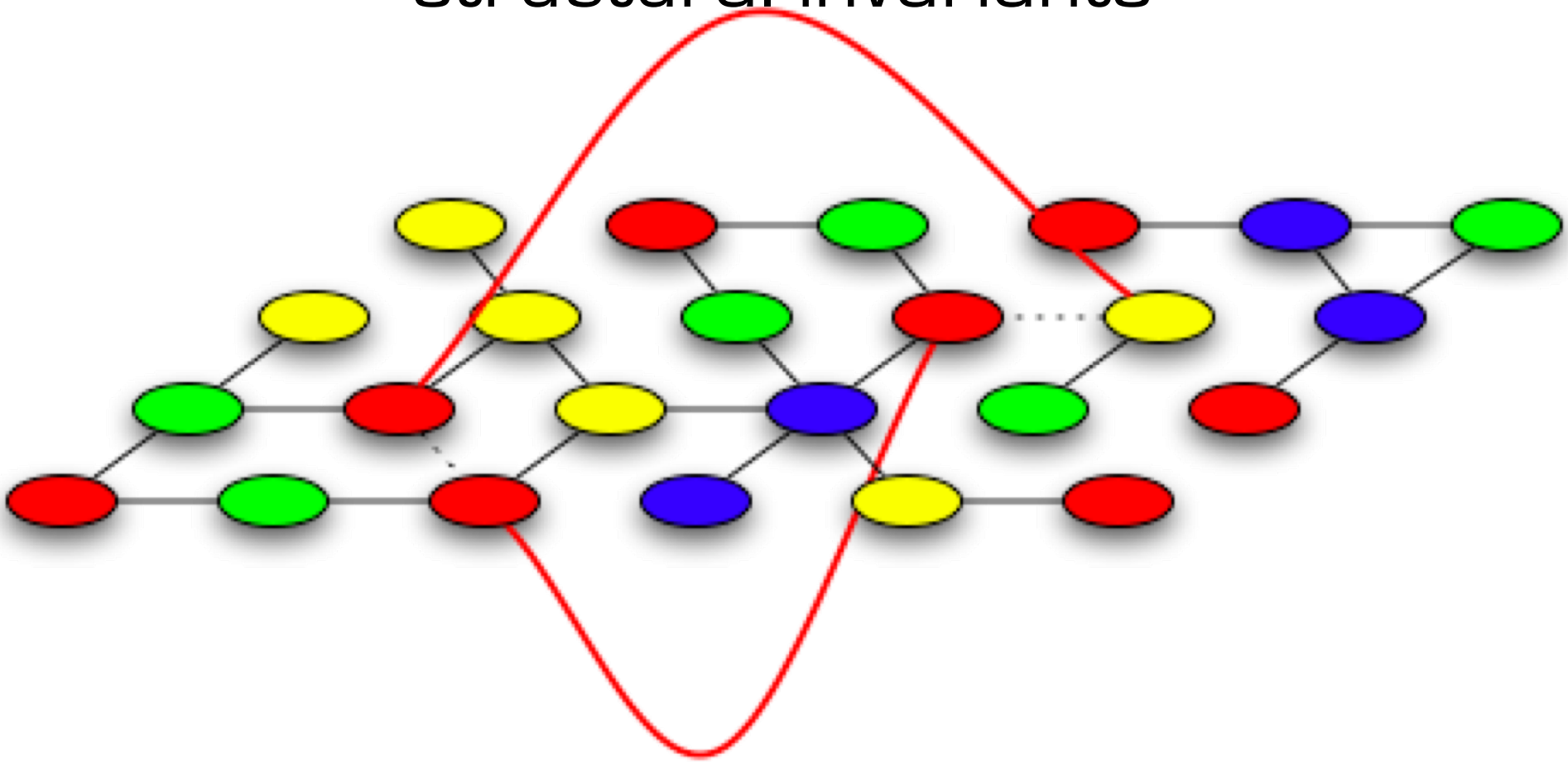


The set of synthetic networks may overlap the set of scale free graphs, but may occupy a "small region". Analytical proofs that weigh each network in this region "equally" therefore are not applicable since most of the weight in an average is not from realistic social networks.

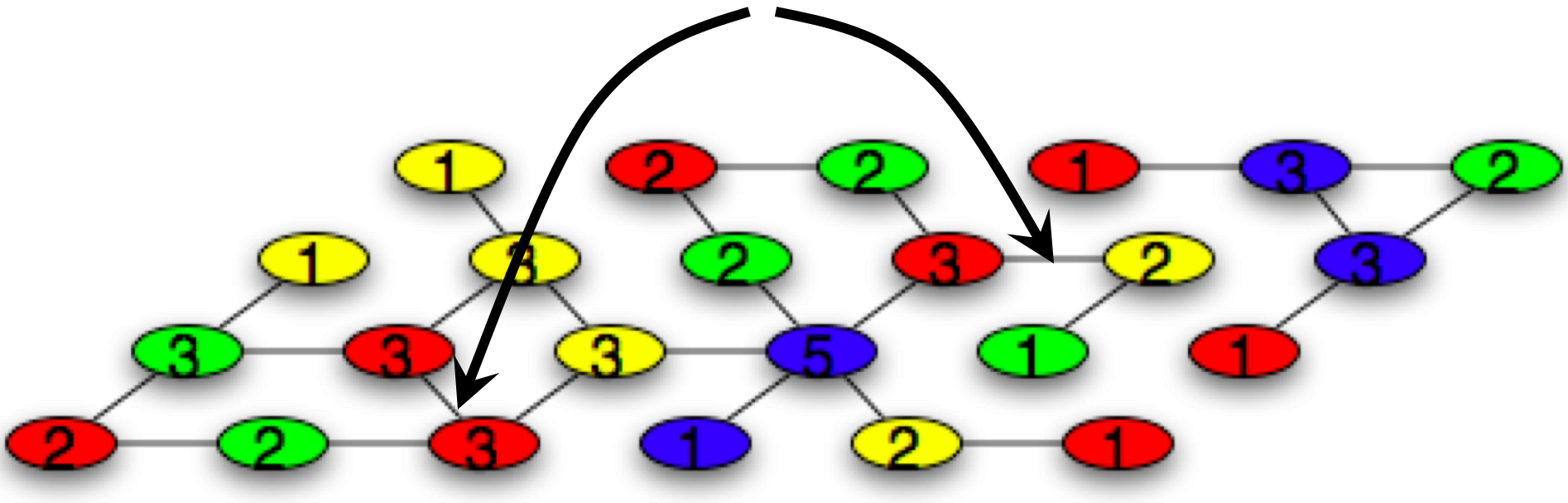
Create “more random” graphs by switching edge endpoints



Constrained edge switch keep local structural invariants

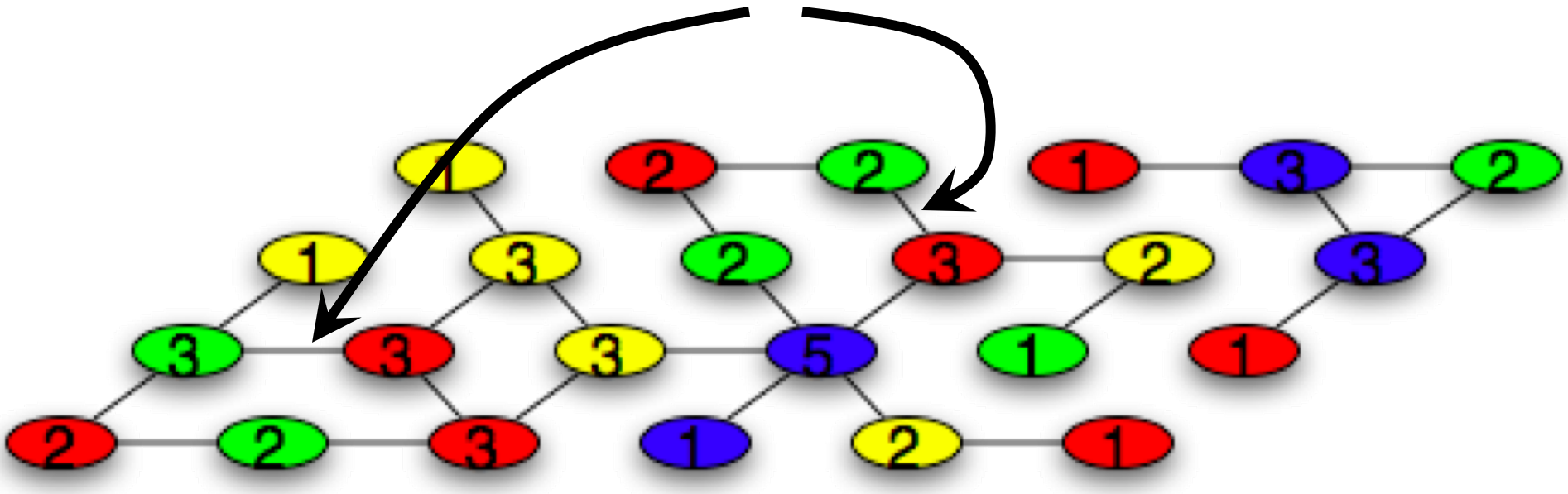


Different switching rules keep different structural constraints



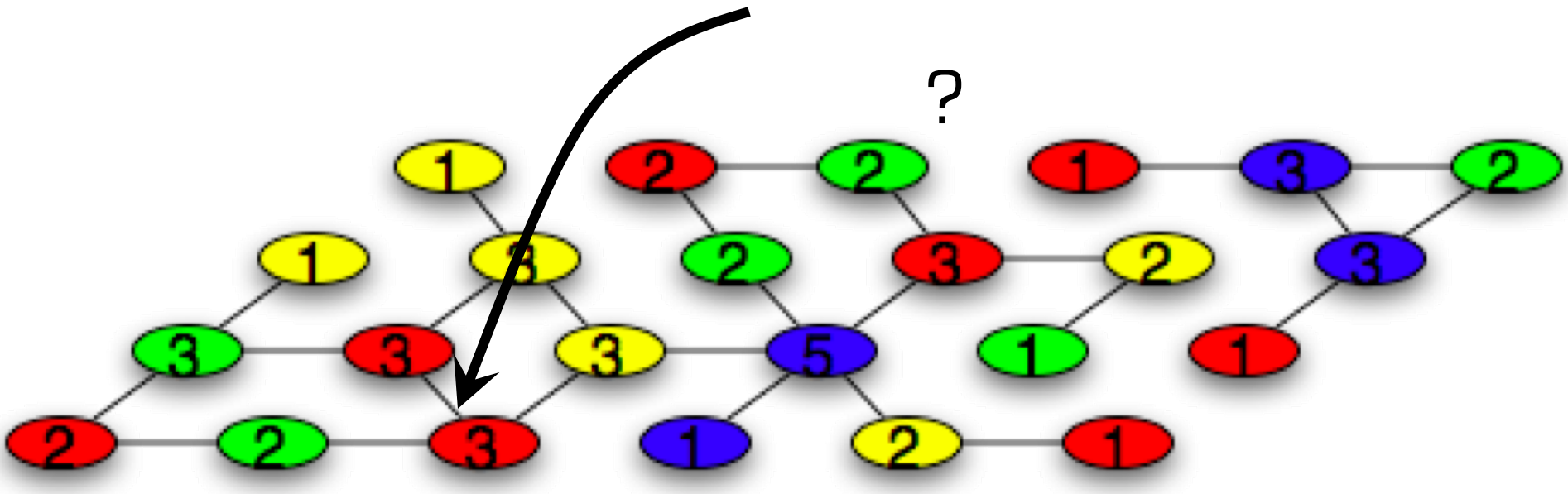
This choice maintains degree distribution (actually it keeps the degree of all the nodes unchanged)

Different switching rules keep different structural constraints



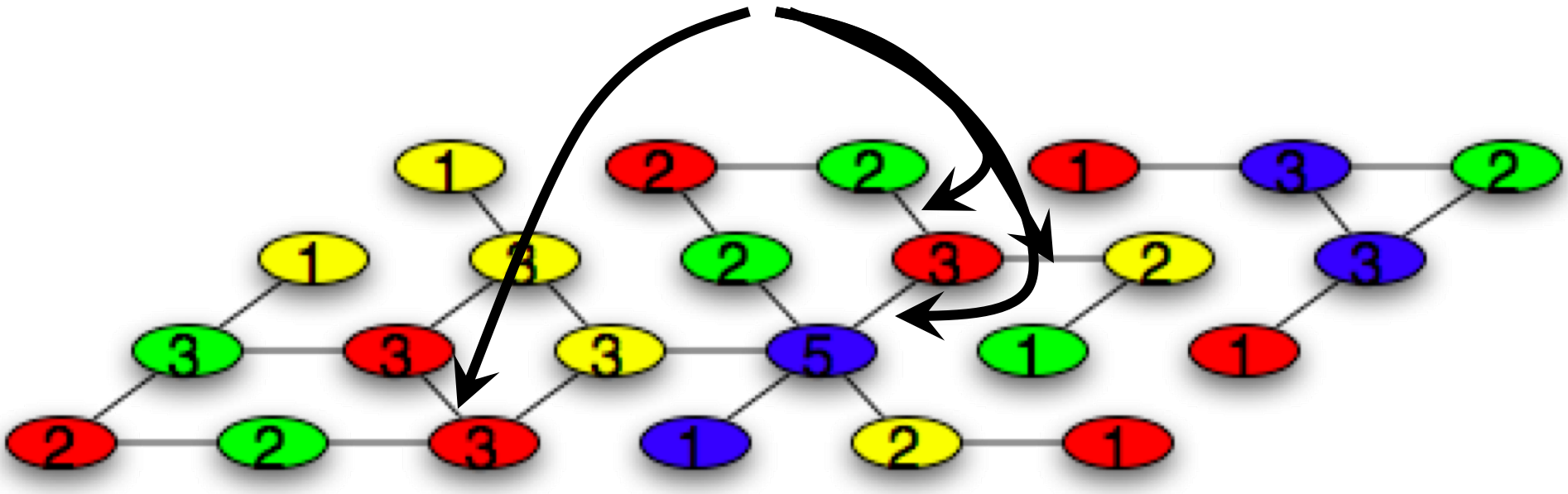
This choice maintains assortativity by color

Different switching rules keep different structural constraints



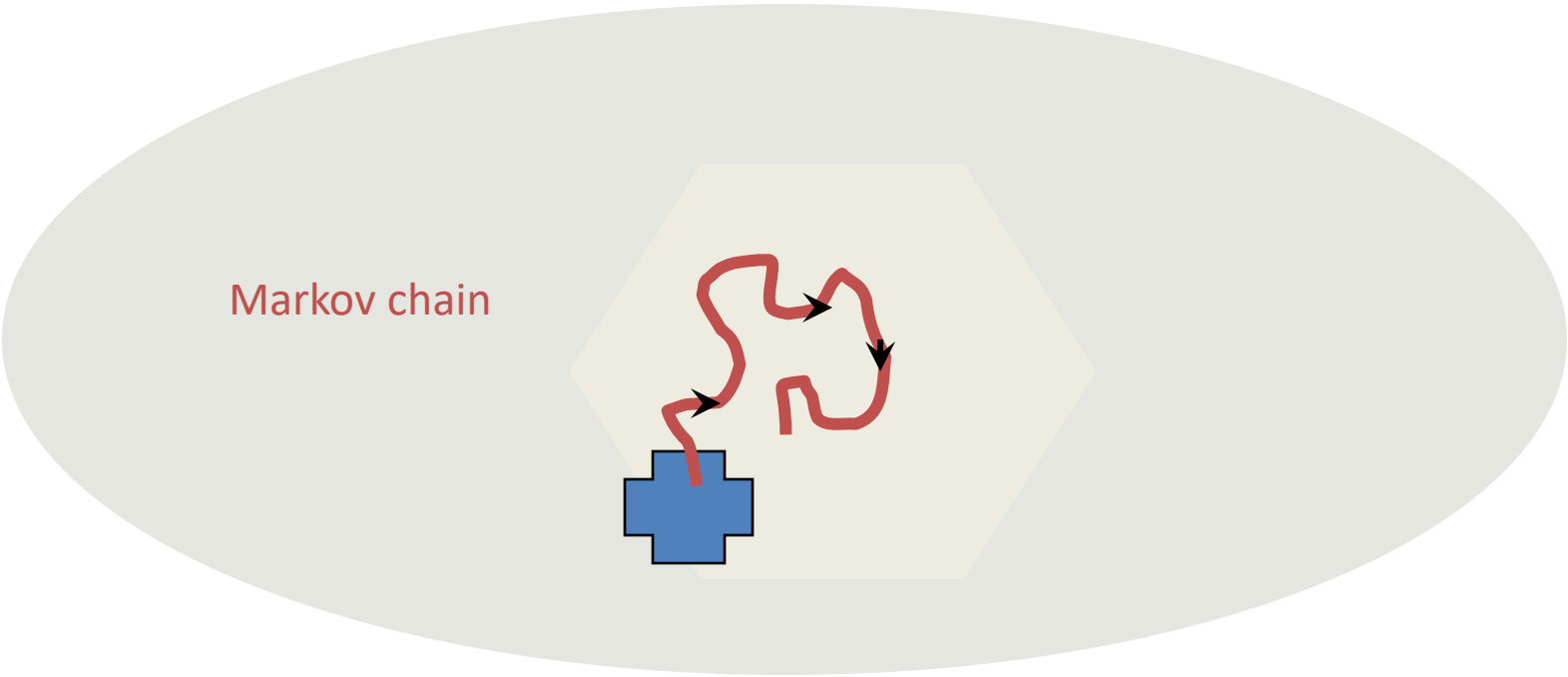
No choice maintains assortativity by degree and color

Different switching rules keep different structural constraints



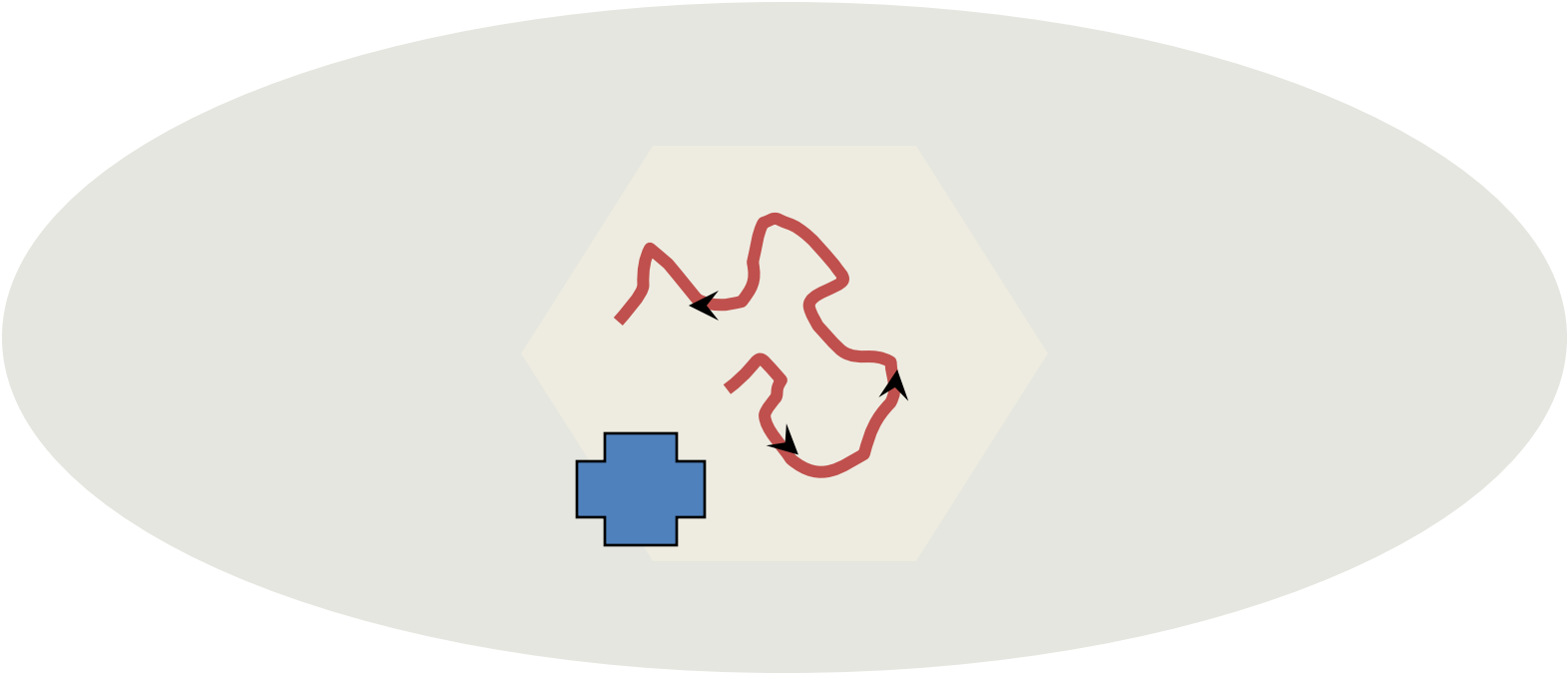
Any choice maintains clustering distribution invariant

Edge switch is one step in a Markov chain



Our Markov chain mixes quickly (empirically), sampling from all graphs with the constrained property.

The chain's starting point is important



Similar studies have gone the other direction, introducing structure into random graphs. Harder since one does not know the minimal structural features that need to be maintained.

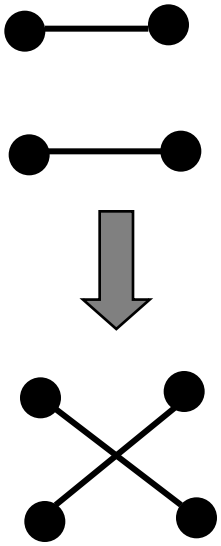
Simple approach for deciding which details matter

- Given
  - a representational detail in the system interaction generator
  - A labeled graph abstraction *G* of the interaction structure
  - A dynamical process *P* over the graph

- Let *S(G)* be a edge-switch operation that holds certain structural constraints invariant and let $S^i(G)$ denote the iterated application (walk on the Markov chain). If *P* changes "sufficiently" as a result, then the features held invariant are not sufficient descriptors of the network model.

Switching operations

Switch operations

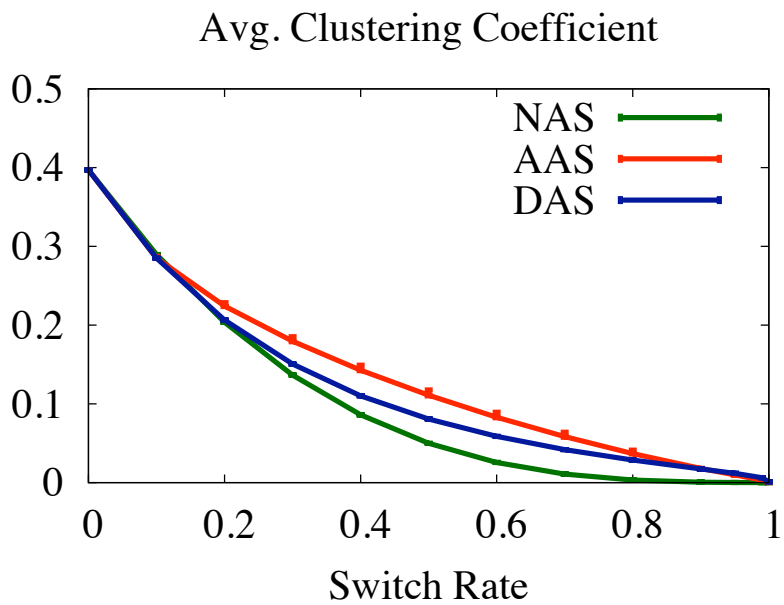


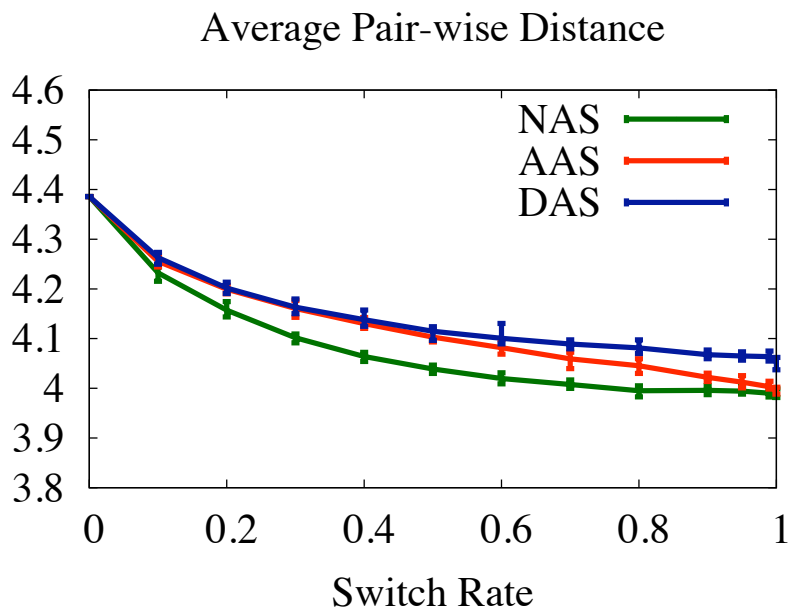
Preserves degree distribution

Assortativity preserving Switching

- Age assortativity(AAS): switch edges if end points have similar age
- Degree assortativity (DAS): switch edges if end points have similar degrees

Edge switch changes structural properties

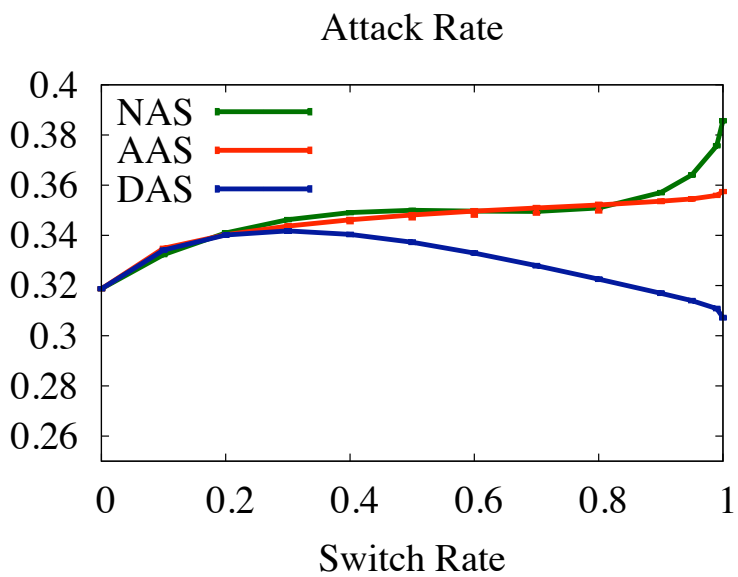


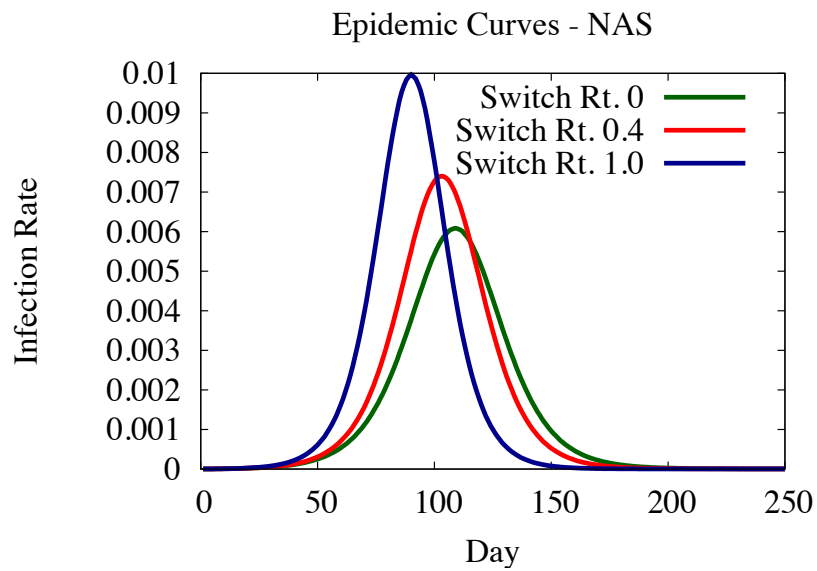


Switch rate: fraction of the edges switched

- Structure of the networks changes significantly with switch rate.

Effect of switching on disease dynamics

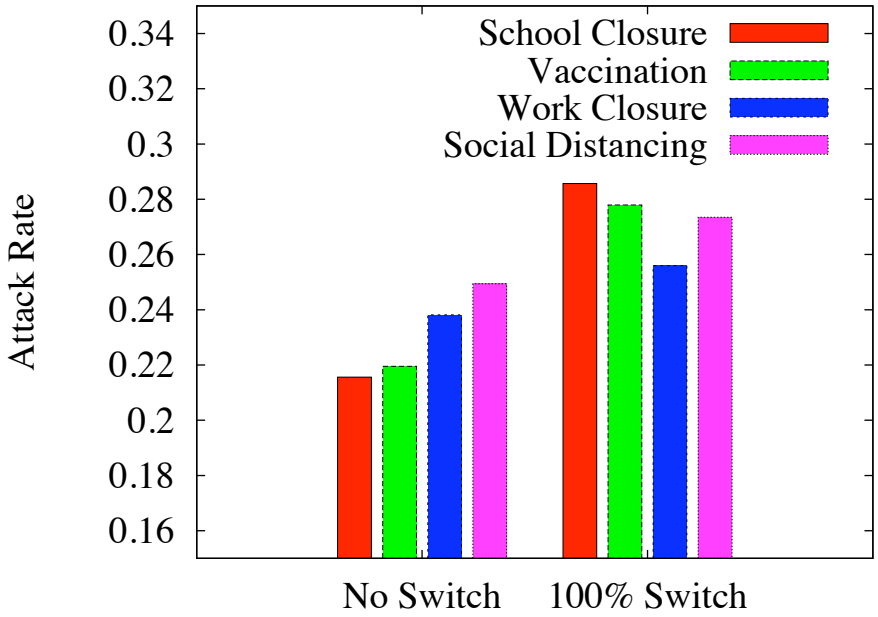




Attack rate: fraction of population that gets infected

- Disease spreads quickly in the switched networks

Effect of switching on intervention policies



In original network:
SC > VC > WC > SD

In the shuffled network:
WC > SD > VC > SC

"A > B" means intervention A is more effective than B.

Efficacy of various interventions: School Closure (SC), Vaccination (VC), Work Closure (WC), and Social Distancing (SD)

- Order of efficacy changes in the switched networks

Section

CONCLUSION

Paths forward

- Synthetic populations provide a data-driven framework for microsimulation and agent-based modeling.
- Applications are numerous and growing:
  - Infectious and chronic disease epidemiology
  - Disaster response
  - Infrastructure modeling
  - Land use
  - Regional economic modeling
  - Environmental epidemiology
  - Refugees, …
- Each application brings new methodological challenges
- There is also a need for integrated software platforms for data management, provenance, version tracking, quality evaluation, and more.

THANK YOU!

Samarth Swarup

swarup@virginia.edu

http://people.virginia.edu/~ss7rs

Madhav V. Marathe

marathe@virginia.edu

https://biocomplexity.virginia.edu/madhav-marathe

Useful websites:

NSSAC: https://biocomplexity.virginia.edu/nssac

Biocomplexity Institute: http://biocomplexity.virginia.edu