Generating Synthetic Populations for Social Modeling

Tutorial at IJCAI 2016, New York City

Network Simulation Science and Advanced Computing Division,
Biocomplexity Institute and Initiative,
University of Virginia.

{swarup, marathe}@virginia.edu

Technical report NDSSL-TR-2016-044

With contributions by: Abhijin Adiga, Richard Beckman, Jiangzhuo Chen, Youngyun Chungbaek, 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 IJCAI, New York City, July 10 2016

We hope to update the slides and supplementary material continually over the next year.

The current version of the slides can be found at:
http://people.virginia.edu/~ss7rs/synthetic_population_tutorial/

Suggested citation:
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
Outline

• Introduction to synthetic populations, networks, and information (a 20+ year effort by NDSSL)
  o The scale and scope of the problem

• Methodology
  o Population synthesis
  o Activity assignment
  o Location assignment
  o Network construction

• Applications

• Concluding remarks
INTRODUCTION
Section

SYNTHETIC POPULATION, NETWORK & INFORMATION
Synthetic information for multi-agent modeling at scale

• Emerging areas of AI application
  – social modeling, including computational sustainability and resilience, urban computing, disaster response, climate change, epidemiology, smart grids, personalized medicine and healthcare

• Complex social systems require detailed, data-driven models at scale to enable forecasting, planning, and intervention modeling.

• Provide a natural data structure to support AI and analytics at scale
What is a synthetic agent?

- A representation of elements’ and states that is not intended to precisely match any snapshot of the system, but to provide a statistically accurate overall picture:
  - people, places, things
  - cells, cytokines, organs

- A synthesis of incommensurate data

- E.g.: A synthetic human agent
  - Can have demographic, social, health, cognitive, cultural attributes
  - These attributes need to be statistically accurate to attributes of humans
What is a synthetic population?

- A synthetic population represents a set of synthetic agents (e.g., people) that share a common geographic, social, or biological characteristic (e.g., people in a rural or urban region, individuals from a given tribe, etc).
  
  - A synthetic population then is a set of synthetic humans that usually share a set of common distributions defining the three kinds described earlier.
  
  - Sharing can be done at a desired spatial, temporal, and social scale. So for e.g., we can have a synthetic population for a block group and they would share census measurements of that block group.

- By sharing these distributions (or alternatively said, being sampled from these distributions) automatically ensures certain that correlational structures will hold.
What is a synthetic network?

- A synthetic network is composed of synthetic agents combined with links that capture interactions among the agents
  - Links can be physical or a matter of convention
  - Links are often inferred based on the context
  - Multiple networks are possible over the same synthetic agent collection
Intended meaning and properties

• The term “Synthetic” is used in two ways:
  – data and attributes of the human are synthesized by integrating a diverse set of data sources and using models for interpolation and extrapolation of data,
  – “Statistically similar” to real individuals but is not identical to any individual in the population.

• Properties of Synthetic agents, populations and networks:
  – Individual attributes are based on real-world collected data,
  – Privacy of individuals is protected.
  – The correlations between the data sets agree with the measured correlations of data in the real world.

• E.g. For example, if I say there is a synthetic individual whose height is 20 feet then this is unlikely based on the existing data. In other words, a synthetic human is statistically similar to a group of individuals in the society but is not identical to any of them.
Why is synthetic information useful?

• Provides a spatially and individually resolved data structure
  – Enables fusion of diverse information in a statistically consistent manner, e.g. (demographic + energy + cell phone) usage data
  – Agents have nominal (age, income), declarative (activities they take) and procedural (e.g. how to drive) data
  – Enables information privacy and attribution

_A social coordinate system for incorporating new (intra- and inter) agentic information_
Why is synthetic information useful?

• Agentic representations and networks enable causal interaction-based models and data integration
  – Putting social and behavioral theories in motion e.g. contagion theories of Granovetter, Macy, Schelling, et al.
  – Explicitly represent interactions as networks
  – Model outputs can then be aggregated to any desired level

Supports next generation social science modeling and analytics at-scale
SYNTHETIC POPULATIONS: A BIG DATA PROBLEM
### Example: A Synthetic social proximity network

#### Facts in Delhi
- **Population**: 13.85M
- **Households**: 2.67M
- **Contacts**: >200M
- **Locations**: 2.64M

#### Volume

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Change every second</th>
<th>Node Status changes every second</th>
<th>They are modeled in minute scale</th>
</tr>
</thead>
</table>

#### Velocity
- Interactions: Change every second
- Node Status: changes every second
- They are modeled in minute scale

#### Variety
- Demographics
- Geographic
- Temporal Feature
- Virus Infectivity
- … …

#### Veracity
- **Data**
  - Do we collect enough raw data to render a clear picture?
- **Method**
  - Do we extract all useful information out of available raw data?
Scaling to global systems is even more challenging

<table>
<thead>
<tr>
<th>Coarse populations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Countries:</td>
<td>220</td>
</tr>
<tr>
<td>Workplaces:</td>
<td>219.5 Million</td>
</tr>
<tr>
<td>Individuals:</td>
<td>7.2 Billion</td>
</tr>
<tr>
<td>Interactions:</td>
<td>45 Billion</td>
</tr>
<tr>
<td>Households:</td>
<td>2.6 Billion</td>
</tr>
<tr>
<td>Schools:</td>
<td>51.6 Million</td>
</tr>
</tbody>
</table>

- Initial storage size of data: 7.7 Terabytes
- Storage: rule of thumb 1-2 GB/million individuals
- Current computation time estimate (including testing, data transfers, etc.): 45 days

Computations and performance data assume *extensive and powerful hardware*
Needs multitude of data sources

- **Activity locations:**
  - LandScan
  - D&B
  - InfoGrid
  - NAVTEQ/HERE POIs
  - OSM POIs
  - Wikipedia

- **Residence locations:**
  - LandScan
  - NAVTEQ/HERE
  - OSM

- **Activity template data**
  - NHTS
  - MTUS
  - ATUS
  - Custom surveys
  - Country similarity measure (matching algorithm)

- **Administrative boundaries**
  - GADM
  - NAVTEQ/HERE
  - OSM
  - US Census
  - ADC Worldmap

- **General demographic information [distributions and micro-samples]**
  - UN
  - 5 UNICEF
  - CIA
  - Eurostat
  - NationMaster
  - European Social Survey
  - IMF
  - WHO
  - Worldbank
  - World Education News and Reviews
  - OECD

- **Detailed, country specific demographic information**
  - US Census Bureau;
  - Country’s statistical agencies
  - Wikipedia

- **Examples of data sources for West African countries:**
  - Liberia National Data Archive Center
    - Liberia household data
    - Liberia Census data
  - Liberia Labor Force Survey 2008
  - Statistics Sierra Leone
  - OpenDataForAfrica
  - Trading Economics
  - Social Security Website
Example: Synthesizing networks from Synthetic populations

Demographic information, population densities, activity surveys and other data sources are fused by modeling and computation to construct a representation of the actual population and the people interactions.

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

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

METHODOLOGY
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

- NATION
  - REGIONS
    - DIVISIONS
      - STATES
        - Counties
          - Census Tracts
            - Block Groups
              - Census Blocks
              - Public Use Microdata Areas
              - State Legislative Districts
              - Urban Growth Areas
              - Core Based Statistical Areas
              - Urban Areas
              - AIANNH Areas* (American Indian, Alaska Native, Native Hawaiian Areas)
              - Places
              - County Subdivisions
              - Traffic Analysis Zones
              - Voting Districts
              - Subminor Civil Divisions
              - Congressional Districts
              - School Districts
              - ZIP Code Tabulation Areas
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
- What we need:

<table>
<thead>
<tr>
<th>The number of Households with household size in given ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsize</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>n</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The number of Households with age in given ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>n</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Householder’s age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsize</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
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.

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

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

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>2000</td>
<td>1500</td>
<td>3500</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>500</td>
<td>6000</td>
<td>6500</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>?</td>
<td>?</td>
<td>3500</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>?</td>
<td>?</td>
<td>6500</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>?</td>
<td>?</td>
<td>3500</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>?</td>
<td>?</td>
<td>6500</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

But, all we know are the marginal totals...
A naïve approach

Assume hair and eye color are independent...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>?</td>
<td>?</td>
<td>3500</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>?</td>
<td>?</td>
<td>6500</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

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...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>.35 x .25 x 10000 = 875</td>
<td>.65 x .25 x 10000 = 1625</td>
<td>3500</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>.35 x .75 x 10000 = 2625</td>
<td>.65 x .75 x 10000 = 4875</td>
<td>6500</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>
A naïve approach

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.35 x .25 x 10000 = 875</td>
<td>.65 x .25 x 10000 = 1625</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>.35 x .75 x 10000 = 2625</td>
<td>.65 x .75 x 10000 = 4875</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>2000</td>
<td>1500</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>500</td>
<td>6000</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

Table estimated under independence assumption from marginal totals

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:

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>21</td>
<td>79</td>
<td>100</td>
</tr>
</tbody>
</table>
### The Set-Up

**UNIVERSE**

\[
\begin{array}{cccc}\hline
i=1 & N_{11} & N_{12} & \cdots & N_{1s} & N_1. \\
& \vdots & & \ddots & \vdots & \vdots \\
i=2 & N_{21} & N_{22} & \cdots & N_{2s} & N_2. \\
& \vdots & & \ddots & \vdots & \vdots \\
r & N_{r1} & N_{r2} & \cdots & N_{rs} & N_r. \\
\hline
N_1 & N_2 & \cdots & N_j & \cdots & N_s & N
\end{array}
\]

- \(N_{ij}\) unknown
- Marginal totals \(N_j\) and \(N_i\) known
- \(N\) known

**SAMPLE**

\[
\begin{array}{cccc}\hline
j=1 & n_{11} & n_{12} & \cdots & n_{1s} & n_1. \\
& \vdots & & \ddots & \vdots & \vdots \\
j=2 & n_{21} & n_{22} & \cdots & n_{2s} & n_2. \\
& \vdots & & \ddots & \vdots & \vdots \\
& \vdots & & \cdots & \vdots & \vdots \\
r & n_{r1} & n_{r2} & \cdots & n_{rs} & n_r. \\
\hline
n_1 & n_2 & \cdots & n_j & \cdots & n_s & n
\end{array}
\]

- \(n_{ij}\) known
- Marginal totals \(n_j\) and \(n_i\) known
- \(n\) known

* This is taken from the seminal paper of Deming and Stephan (1940)
The challenge is to maintain the dependence structure of the sub-sample while matching it to the whole population’s marginal totals.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td>Black Hair</td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>Totals</td>
<td>21</td>
<td>79</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>m11</td>
<td>m12</td>
<td>3500</td>
</tr>
<tr>
<td>Black Hair</td>
<td>m21</td>
<td>m22</td>
<td>6500</td>
</tr>
<tr>
<td>Totals</td>
<td>2500</td>
<td>7500</td>
<td>10000</td>
</tr>
</tbody>
</table>

What does this mean?
Possible approaches

• Maximum likelihood [Smith, 1947]

\[
prob\{n_{ij}\} = \frac{n!}{\prod_{i=1}^{r} n_{ij}! \prod_{j=1}^{c} p_{ij}^{n_{ij}}}
\]

• Minimize the chi-squared statistic [Deming and Stephan, 1940]

\[
\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - np_{ij})^2}{n_{ij}}
\]

• Minimize discrimination information [Ireland and Kullback, 1968]

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

• It turns out the latter two are equivalent.

Notation from Ireland and Kullback (1968)
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 \( D(p' || p) \).

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

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.
    \[
    \frac{N_i}{N} = p_i = \sum_j p_{ij} \quad \frac{N_j}{N} = p_j = \sum_i p_{ij}
    \]

- Initialize $p_{ij}^{(0)} = \frac{n_{ij}}{n}$
- For $t \geq 1$
  - Set $p_{ij}^{(2t-1)} = p_{ij}^{(2t-2)} \frac{p_i}{p_i^{(2t-2)}}$ and $p_{ij}^{(2t)} = p_{ij}^{(2t-1)} \frac{p_j}{p_j^{(2t-1)}}$
# IPF on our example data

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>21</td>
<td>89</td>
<td>100</td>
</tr>
</tbody>
</table>

Subject to \( p_1 = .35 \), \( p_2 = .65 \), \( p_1 = .25 \), \( p_2 = .75 \)
IPF on our example data

Initialize

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>18/100 = 0.18</td>
<td>17/100 = 0.16</td>
<td>0.34</td>
</tr>
<tr>
<td>Black Hair</td>
<td>3/100 = 0.03</td>
<td>62/100 = 0.63</td>
<td>0.66</td>
</tr>
<tr>
<td>Totals</td>
<td>0.21</td>
<td>0.79</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to $p_1 = 0.35$, $p_2 = 0.65$, $p_1 = 0.25$, $p_2 = 0.75$
IPF on our example data

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>.18x.35/.34 = 4 = .185</td>
<td>.16x.35/.34 = .165</td>
<td>.35</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>.03x.65/.66 = .030</td>
<td>.63x.65/.66 = .620</td>
<td>.65</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.215</td>
<td>.785</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to \( p_1 = .35 \), \( p_2 = .65 \), \( p_1 = .25 \), \( p_2 = .75 \)
IPF on our example data

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>0.185x</td>
<td>0.165x</td>
<td>0.373</td>
</tr>
<tr>
<td></td>
<td>(\frac{25}{215} = 0.215)</td>
<td>(\frac{75}{785} = 0.158)</td>
<td></td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>0.030x</td>
<td>0.620x</td>
<td>0.627</td>
</tr>
<tr>
<td></td>
<td>(\frac{25}{215} = 0.035)</td>
<td>(\frac{75}{785} = 0.592)</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>0.25</td>
<td>0.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to \(p_1. = 0.35, p_2. = 0.65, p_1 = 0.25, p_2 = 0.75\)
IPF on our example data

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td></td>
<td></td>
<td>.35</td>
</tr>
<tr>
<td>Blue Eyes</td>
<td>.215x</td>
<td>.158x</td>
<td>.35/.373 = .202</td>
</tr>
<tr>
<td>Brown Eyes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black Hair</td>
<td>.035x</td>
<td>.592x</td>
<td>.65/.627 = .036</td>
</tr>
<tr>
<td>Blue Eyes</td>
<td>.238</td>
<td>.762</td>
<td>.238/.762 = 1.00</td>
</tr>
<tr>
<td>Brown Eyes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 ...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.202x .25/.238 =</td>
<td>.148x .75/.762 =</td>
<td>.358</td>
</tr>
<tr>
<td></td>
<td>.212</td>
<td>.146</td>
<td></td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.036x .25/.238 =</td>
<td>.614x .75/.762 =</td>
<td>.642</td>
</tr>
<tr>
<td></td>
<td>.038</td>
<td>.604</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.25</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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 ...

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td></td>
<td></td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>.212x</td>
<td>.146x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.35/.358</td>
<td>.35/.358</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>= .207</td>
<td>= .143</td>
<td></td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>.038 x</td>
<td>.604 *</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>.65/.642</td>
<td>.65/.642</td>
<td></td>
</tr>
<tr>
<td></td>
<td>= .038</td>
<td>= .612</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>.244</td>
<td>.755</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to \( p_1. = .35, p_2. = .65, p_1 = .25, p_2 = .75 \)
IPF on our example data

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>.210</td>
<td>.142</td>
<td>.352</td>
</tr>
<tr>
<td>Black Hair</td>
<td>.040</td>
<td>.608</td>
<td>.648</td>
</tr>
<tr>
<td>Totals</td>
<td>.25</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Subject to $p_1. = .35$, $p_2. = .65$, $p_1 = .25$, $p_2 = .75$
IPF on our example data

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

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>.209</td>
<td>.141</td>
<td>.350</td>
</tr>
<tr>
<td>Black Hair</td>
<td>.040</td>
<td>.609</td>
<td>.649</td>
</tr>
<tr>
<td>Totals</td>
<td>.249</td>
<td>.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

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

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

- 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

Examples

- **TRANSIMS** - [https://www.fhwa.dot.gov/planning/tmip/resources/transims/](https://www.fhwa.dot.gov/planning/tmip/resources/transims/)
  - Used for accurate and sensitive travel forecasts for transportation planning and emission analysis

  - To simulate metropolitan real estate markets and study the impact of land use policies

- **EUROMOD** - [https://www.euromod.ac.uk/](https://www.euromod.ac.uk/)
  - EU based microsimulation to calculate effects of taxes and benefits on incomes and work incentives

  - Canada based Longitudinal population health microsimulation model to rationally compare competing health intervention alternatives

  - to understand the potential outcomes of public policy changes such as welfare reform, tax reform, and national health care reform.
Some useful references

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.

http://nhts.ornl.gov
Activity Selection

Where do I want to go?

What do I want to do?
## Notation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{ij}$</td>
<td>Synthetic person $i$ in synthetic household $j$.</td>
</tr>
<tr>
<td>$H_j$</td>
<td>Synthetic household $j$.</td>
</tr>
<tr>
<td>$P^*_{i^<em>j^</em>}$</td>
<td>Survey person $i^<em>$ in survey household $j^</em>$.</td>
</tr>
<tr>
<td>$H^<em>_{j^</em>}$</td>
<td>Survey household $j^*$.</td>
</tr>
</tbody>
</table>
| $X_{ij}$   | Personal and household demographics of the $i^{th}$ synthetic person in the $j^{th}$ synthetic household.  
Example: $X = \{\text{age, sex, highest income in household}\}$ |
| $X^*_{i^*j^*}$ | 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

Survey Household
Consists of...
Survey Individuals

Split 1 (ex: age of oldest member)

Split 2a (ex: income of oldest member)

Split 2b (ex: number of household members)

Synthetic Household
Consists of...
Synthetic Individuals

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.

Split 1 (ex: age of oldest member)

Split 2a (ex: income of oldest member)

Split 2b (ex: number of household members)
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:

\[ D_{HH}(H_j, H_{j^*}) = \max_i \{ D_{PH}(P_{ij}, H_{j^*}) \} \]  

\[ D_{PH}(P_{ij}, H_{j^*}) = \min_i \{ D(P_{ij}, P_{i^*j^*}) \} \]

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

Survey household

Synthetic household

\[ D(1, 1^*) = 1.2 \]

\[ D(1, 2^*) = 0.5 \]

\[ D(2, 1^*) = 2.7 \]

\[ D(1, 3^*) = 3.8 \]

\[ D(2, 2^*) = 3.7 \]

\[ D(2, 3^*) = 4.0 \]
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?

$$\left( X_{ij} - X^*_{i* j*} \right) S^{-1} \left( X_{ij} - X^*_{i* j*} \right)$$
What about D?

We use...

\[
D(P_{ij}, P_{i^*j^*}) = ||\hat{Y}_{ij} - \hat{Y}_{i^*j^*}||
\]

where

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

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

\[
\hat{Y}_{i^*j^*k} = X_{i^*j^*} \hat{\beta}_k + \epsilon_{i^*j^*k}
\]

*We don’t have to use linear regression. Something like CART, random forest, etc. could also be used here for \(f_k\).
Results: Home

Smoothed Template
Average Home Hours

VSP
Average Hours for People with Activity of HOME

Fitted-Values Match
Average Hours for People with Activity of 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

Smoothed Template
Persons with College Activity

VSP
Persons with Activity of COLLEGE

Fitted-Values Match
Persons with Activity of COLLEGE
Some useful references

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 fitting parameters
Gravity Models: In Analogy with Newton’s Law of Gravity

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

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

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

Original 1-parameter fitting form is not satisfying.

Ex.1. Global cargo ship movements:

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

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

Ex.2. Airway Traffic:

\[
\phi_{ab} = \frac{m^\alpha_n^\beta}{C f(r_{ab})}
\]

\[
f(r_{ab}) = e^{\delta r_{ab}}
\]

*KRequires two sets of parameters, for \( r_{ab} > 300 \text{ km} \)

and \( r_{ab} < 300 \text{ km} \)


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

Radiation Model

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

\[
\phi_{ab} = \zeta m_a p_{ab}
\]

- \(\zeta\) – 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 $a$, is looking for a job with salary or benefits expectation of at least $z$.

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

Values of offers are denoted by $W_z$.

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

Let's assume that

and are drawn from the same distribution.

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

Here, $c_{ab}$ is a general travel cost on the network.
Some useful references

- Yingxiang Yang, Shan Jiang, Daniele Veneziano, Shounak Athavale, Marta C. Gonzalez, “TimeGeo: a spatiotemporal framework for modeling urban mobility without surveys.” PNAS.
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

**Structural properties**
- Network measures
  - Degree distribution
  - Number of components
  - Diameter
  - Spectral radius
- Sensitivity to structural uncertainty.

**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 to locations with durations of visit and through this determine their contacts and interactions.
Relational, dynamic social contact networks

**People (8 million)**

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

**Locations (1 million)**

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

- Edge attributes:
  - activity type: shop, work, school
  - \([\text{start time 1, end time 1}]\)
  - \([\text{start time 2, end time 2}]\)
  - ...

Locations (1 million)

People (8 million)
Various projections of person-location temporal social visitation network

Person Projections

Static Projections

Location Projections

Temporal Projections

G_P

G_{PL}

G_L

G_{SP}

G_{SL}
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

1-1 mapping
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:
  - 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

Hybrid approach that combines digital traces with administrative and survey information

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

Table 1: Statistic of real class schedules (data from registration information)
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.
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.
## Comparing subnetwork refinement

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

<table>
<thead>
<tr>
<th>type</th>
<th>subpopulation identification</th>
<th>subnetwork construction</th>
<th>subnetwork embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>high school</td>
<td>individuals in synthetic population who have school type activities in the three schools</td>
<td>usual method is used to combine activity sequences with sublocation model to derive co-location contacts; we have studies different sublocation models</td>
<td>replace existing in-school contacts with new contacts</td>
</tr>
<tr>
<td>hospital</td>
<td>patients can be identified in an epidemic simulation; healthcare workers are those having work type activities in this hospital</td>
<td>activities are created for each healthcare worker, patient, and visitor, from which a people-location network is generated</td>
<td>N/A [not done yet]</td>
</tr>
<tr>
<td>military base</td>
<td>synthetic households whose home locations are within [not enough] or near the bases</td>
<td>on-base activity sequences are assigned to individuals in subpopulation by cohort; usual method is used to combine activity sequences with sublocation model to derive co-location contacts</td>
<td>replace existing on-base contacts with new contacts</td>
</tr>
<tr>
<td>slum</td>
<td>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</td>
<td>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</td>
<td>replace existing edges of individuals in slum subpopulation with new edges</td>
</tr>
</tbody>
</table>
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


APPLICATIONS & EXTENSIONS
Three significant programs: 1992 - present

• TRANSIMS: Urban transport planning
  – Long term policies, first use of HPC for social sciences

• CNIMS: National Incident management system
  – Interdependent infrastructure modeling and simulation, short-term planning and response for large scale disasters

• Simdemics: Real-time epidemic planning and response
  – Planning and response; short time scale (1-6 months)
Examples of related efforts

- Urbansims and Synthicity.com (UC Berkeley)
- Mistra Urban Futures (Initiative at Chalmers U.)
- VT: Social and Decision Informatics Laboratory (SDAL) and Global Forum on Urban and Regional Resilience (GFURR)
- IIT Mumbai: C-USE: center for Urban Science and Engineering
- CUSP: NYU center for Urban Science
- IBM: Smart Cities initiatives
- Warwick Institute for Science of Cities
- ETH: Future Cities Laboratory
Example 1: Urban transport planning
TRANSIMS: Activity based urban transport planning environment

• TRANSIMS: (1991-2001)
  – Set up to address important national problems
  – Assessing traffic flows in periods of high congestion
  – Predicting household behavior when the infrastructure changes
  – Assessing the social justice issues in proposed infrastructure changes
  – Predicting the environmental impact of the transportation system

• Requirements set by:
  – Government (DOT), Regional planning organizations
  – University researchers, Transportation consultants

• Technology developed, and demonstrated in close collaboration with Metropolitan planning offices
  – Case studies with Albuquerque, Dallas, Portland MPOs
  – Open source system managed by Argonne National Laboratory
Simulated Traffic in Portland
Exposure and air quality
White House Area Transportation Study
(National Capitol Planning Commission)

• Purpose: Examine traffic problems around White House due to street closures
  – Objective: Mitigate the impacts of closures for travelers in the downtown area
  – Study by USDOT: 2005-2010
  – AECOM

• Metrics examined include:
  – Travelers experiences
  – Travelers interactions with each other
  – Amount of time to travel
  – Cost to travel

  ▪ Compare solution alternatives:
  ▪ Times & cost between scenarios
    ▪ New tunnels construction – Infrastructure
    ▪ Open or close streets
    ▪ Transit improvements – bus
    ▪ Traffic management and traffic operations

http://www.ncpc.gov/ncpc/Main%28T2%29/Planning%28Tr2%29/PlanningStudies%28Tr3%29/Transportation.html
High Level Findings: White house area transportation study (WHATS)

- **TRANSIMS** was used to collect over 20M regional trip/day and create models to examine congestion relieving options.

- **Effect of closures**
  - Reduced the number of people traveling downtown. Increased the amount of delays to 14,000 hours

- **Effect of improvements**
  - Combination of traffic operations and transit operation improvements have the potential to provide measurable benefits and travel improvement to all downtown travelers
Example 2: Societal response in the event of large unanticipated disaster
CNIMS: Comprehensive National Incident Management System (2005-present)

- **CNIMS**
  - Broad interdisciplinary effort to incorporate details of interdependent social, behavioral, economic and societal infrastructure, mainly for counter-WMD/CBNT decision informatics.
  - Scalable agent-based simulation technology
  - Uses high performance computing grids and web-based services
  - Methods scale to 100M individual agents interacting in very small time intervals.
  - Epidemiological models, Interdependent infrastructure models, transportation models, economic models

- Requirements set by
  - DoD

- Developed and deployed in close collaboration with DoD
  - Initial version tested in spiral development during actual studies performed for the DoD, DHHS, BARDA.
Hypothetical Scenario:
- Unannounced 10 kt detonation of an Improvised Nuclear Device (IND)
- 11:15am May 15th, 2006
- 16th and K Street, Washington DC

Traditional Focus
- Prompt impact on overall health
- Responders and their safety
- Victims as essentially passive
- Cold war scenarios or low level exposure to radiation
SI based causal approach to investigate role of individual and collective dynamic behavior

- Will Fred shelter in place? How long?
- How far can Fred walk?
- Which direction will Fred walk?
- What if Fred could contact his family? Where will he be when he can?
- Where will Fred encounter decontam/triage facilities? What supplies are needed to care for him?
- How do we align response with Fred’s goals?
# The Multi-Network Synthesis

## Communication Network
- TowerMaps
- Sprint Sites
- AT&T Towers
- CDC Tobacco Use Survey (TUS)

## Electrical networks
- Substations
- DC Infrastructure details
- Electricity generation
- Electricity Consumption
- Distribution network
- Transmission network

## Transportation Network
- NAVTEQ
- Washington Metropolitan Area Transit Authority (wmata.com)
- Washington Metropolitan Area Transit Authority (wmata.com)

## Populations and Built infrastructure
- American Community Survey
- Dun & Bradstreet
- Navteq
- Tiger/LINE
- NCES
- NHTS
- Destination DC
- Smithsonian Institution
- DC.gov data catalog
- DC MPD

<table>
<thead>
<tr>
<th>Population Segment</th>
<th>Population Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSA occupancy at t=0:00</td>
<td>730,833</td>
</tr>
<tr>
<td>Base population (living in DC)</td>
<td>498,801</td>
</tr>
<tr>
<td>Tourists</td>
<td>26,810</td>
</tr>
<tr>
<td>Business travelers</td>
<td>17,499</td>
</tr>
<tr>
<td>College students in dorms</td>
<td>19,321</td>
</tr>
<tr>
<td>Fire stations, fire fighters</td>
<td>35,1050</td>
</tr>
<tr>
<td>Police stations, policemen</td>
<td>19,3803</td>
</tr>
<tr>
<td>Base households (in DC)</td>
<td>250,723</td>
</tr>
<tr>
<td>Single person households (in DC)</td>
<td>117,760</td>
</tr>
<tr>
<td>School locations (in DC)</td>
<td>594</td>
</tr>
</tbody>
</table>
Interactions between behavior and infrastructures
Relevance of social details even in physically dominated event:
  e.g., cumulative individual exposures to radiation & hazards as behaviors change
Relevance of social details even in physically dominated event:

e.g., cumulative individual exposures to radiation & hazards as behaviors change
Illustrative conclusions

• Even a *partially restored* communication system has disproportionately positive impact on the overall behavior, leading to fewer deaths, better health outcomes & reduction in anxiety.
  – Policy question: how do we build a self forming comm. network using components that belong to diverse stakeholders with few strategic nodes

• The power network suffers a huge loss and large portions of the network will unlikely be operational for at least year or two.
  – Nevertheless if the protection devices work as planned, NO significant cascading failures beyond a small area
  – Important implications on how the city and its surroundings will be reconstituted.

• Individual emergent behaviors: personal choices + contextualized dynamic decisions + institutionalized processes and policies
  – Computational representations need to place individuals within such a well defined spatial socio-technical system
Example 3: Epidemiology and Public Health
Application: analysis of epidemics

- 1918 Pandemic
  - 50 million deaths in 2 years (3-6% world pop)
  - Every country and community was effected
- Good news
  - Pandemic of 1918 lethality is currently unlikely
  - Governments seem to be better prepared and coordinated: e.g. SARS epidemic
- But ... WHAT DETAILS MATTER TO DECISION MAKING?
  - Planning and responding to even a moderate outbreak is challenging:
    - inadequate vaccines/anti-virals, unknown efficacy, hard logistics issues
  - Modern trends further complicate planning:
    - increased travel, immuno-compromised populations, increased urbanization
Simdemics: Modeling environment for networked epidemiology

- Simdemics: 2001-present
  - a modeling environment for pandemic planning and response.
  - designed specifically to scale to networks with 300 million agents – the underlying algorithms and methods are all HPC-oriented methods.
  - a better understanding of the characteristics of the underlying network and individual behavioral adaptation can give better insights into contagion dynamics and response strategies.

- Requirements set by
  - DoD, NIH, CDC and other health agencies

- Technology demonstrated and used in 15+ case studies involving pandemic planning and response
Supporting public policies

• White House Homeland Security Council for smallpox mass vaccination
  – Do we need mass vaccination? How do we protect critical workers? \textit{[Nature’04]}
• Federal Influenza Plan: OHS & DHHS – NIH MIDAS project
  – TLC: Targeted Layered Containment, Importance of Social Distancing \textit{[PNAS’08]}
• Pandemic Planning for Military Preparedness: DoD
  – Impact of layered interventions for force projection: Public versus military health epidemiology \textit{[WSC’09,IHI’12]}
• DHHS Planning and Response to H1N1 studies
  – Markets + public distribution of A/V \textit{[SBP’10]}
  – Optimal vaccination strategies
Examples of analytic support for Ebola response

- *Estimating* basic epidemiological parameters
- *Forecasting* the ongoing epidemic with & without control
- *Assessing* the threat of imported cases in the US causing secondary infections
- *Efficiently allocating* of potential pharmaceutical and protective equipment
- *Location of Emergency treatment centers* and assessing the impact these centers will have on the outbreak
- Analyzing Twitter data to assess public mood & sentiments
Some useful references

CONCLUDING REMARKS
Things we did not cover

• Adding new attributes to an individual:
  – e.g. Creating synthetic electronic medical records
    • Time course of a person’s disease history needs to be tagged to the basic set of attributes
    • Privacy and anonymity concerns are critical

• Modeling behavior
  – Endogenous decision-making

• Creating new class of synthetic populations
  – E.g. population of animals, cells within a body, IOT devices

• Synthesizing the relationship between synthetic agents
  – Inferring interactions (edges, relational structures and hyperedges)
  – The relationships are usually captured as multi-theory multi-networks

• Synthetic information environment
  – New machine learning and big data methods for synthesis, storage, reasoning about synthetic data.
  – Methods for making the data accessible as well allowing individuals to interact with synthetic data.
Things we did not cover

• Infrastructure modeling
  – Power grids, energy markets, cell communication

• Verification & validation, uncertainty quantification, sensitivity analysis
  – Graph Dynamical Systems

• Coupling with data-gathering
  – Active data gathering
  – Social media surveillance

• Synthesizing information at different levels of quality
  – What to do in data-poor regions?
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
Extra Slides
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.
    \[
    \frac{N_i}{N} = p_i = \sum_j p_{ij} \quad \frac{N_j}{N} = p_j = \sum_i p_{ij}
    \]

• Initialize \( p_{ij}^{(0)} = n_{ij}/n \)
• For \( t \geq 1 \)
  • Set \( p_{ij}^{(2t-1)} = p_{ij}^{(2t-2)} \frac{p_i}{p_i^{(2t-2)}} \) and \( p_{ij}^{(2t)} = p_{ij}^{(2t-1)} \frac{p_j}{p_j^{(2t-1)}} \)

Important to notice:
• At the 2t-1 iteration, everything in the same row [same \( i \)] is multiplied by the same factor, hence “proportional” fitting.
• At the 2t-1 iteration, everything is adjusted so that the row proportions agree with the constraint– this may jitter the column proportions
• At the 2t iteration, the column proportions are then adjusted to agree with their constraints.

This notation and explanation borrowed from Fienberg [1970]
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.

\[
\frac{N_i}{N} = p_i = \sum_j p_{ij} \quad \frac{N_j}{N} = p_j = \sum_i p_{ij}
\]

- Initialize
  - $p_{ij}^{(0)} = \frac{n_{ij}}{n}$
- For $t \geq 1$
  - Set
    - $p_{ij}^{(2t-1)} = p_{ij}^{(2t-2)} \frac{p_i}{p_i^{(2t-2)}}$
    - and $p_{ij}^{(2t)} = p_{ij}^{(2t-1)} \frac{p_j}{p_j^{(2t-1)}}$

Important to notice:
- At the 2t-1 iteration, everything in the same row (same $i$) is multiplied by the same factor, hence “proportional” fitting.
- At the 2t-1 iteration, everything is adjusted so that the row proportions agree with the constraint—this may jitter the column proportions.
- At the 2t iteration, the column proportions are then adjusted to agree with their constraints.

This notation and explanation borrowed from Fienberg [1970]
How does this preserve dependence?

Notice that for any T,
\[ p^{(T)}_{ij} = a^{(T)}_i b^{(T)}_j \frac{n_{ij}}{n} \]

A non-rigorous explanation is that at iteration 2t-1, everything in the same row is multiplied by some factor \([a]\). At iteration 2t, everything in the same column is multiplied by some other factor \([b]\).

Look at \(t=1\):
\[
\begin{align*}
p^{(1)}_{ij} &= a_i p^{(0)}_{ij} = a_i n_{ij} / n \\
p^{(2)}_{ij} &= b_j p^{(1)}_{ij} = b_j a_i n_{ij} / n 
\end{align*}
\]

As this process continues, this relationship holds.

This notation and explanation borrowed from Fienberg [1970]
How does this preserve dependence?

Notice that for any $T$,  \[ p_{ij}^{(T)} = a_i^{(T)} b_j^{(T)} n_{ij} / n \]

This implies that the estimated cross-product ratios,

\[ \frac{p_{ij}^{(T)} p_{hk}^{(T)}}{p_{ik}^{(T)} p_{hj}^{(T)}} = \frac{n_{ij}^{(T)} n_{hk}^{(T)}}{n_{ik}^{(T)} n_{hj}^{(T)}} \]

remain constant through all iterations of the procedure.
How does this preserve dependence?

Notice that for any $T$, 

$$ p_{ij}^{(T)} = a_i^{(T)} b_j^{(T)} n_{ij}/n $$

This implies that the estimated cross-product ratios,

$$ \frac{p_{ij}^{(T)} p_{hk}^{(T)}}{p_{ik}^{(T)} p_{hj}^{(T)}} = \frac{n_{ij}^{(T)} n_{hk}^{(T)}}{n_{ik}^{(T)} n_{hj}^{(T)}} $$

remain constant through all iterations of the procedure.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blond Hair</strong></td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td><strong>Black Hair</strong></td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>21</td>
<td>79</td>
<td>100</td>
</tr>
</tbody>
</table>

$$ r_1 = p_{11}/p_{12} $$

$$ r_2 = p_{21}/p_{22} $$

$$ r_1/r_2 = \text{sample ratio} $$
How does this preserve dependence?

\[
\begin{align*}
    r_1 &= \frac{\text{Blue and Blond}}{\text{Brown and Blond}} \\
    r_2 &= \frac{\text{Blue and Black}}{\text{Brown and Black}} \\
    r_1/r_2 &= \text{sample ratio}
\end{align*}
\]

Ratio of Blue:Brown for Blond divided by Ratio of Blue:Brown for Black remains constant.

<table>
<thead>
<tr>
<th></th>
<th>Blue Eyes</th>
<th>Brown Eyes</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond Hair</td>
<td>18</td>
<td>16</td>
<td>34</td>
</tr>
<tr>
<td>Black Hair</td>
<td>3</td>
<td>63</td>
<td>66</td>
</tr>
<tr>
<td>Totals</td>
<td>21</td>
<td>79</td>
<td>100</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
    r_1 &= p_{11}/p_{12} \\
    r_2 &= p_{21}/p_{22} \\
    r_1/r_2 &= \text{sample ratio}
\end{align*}
\]

This notation and explanation borrowed from Fienberg [1970].