**Meeting: 2/1/18**

Before Sunday:

1. Sign up for Overleaf account: <https://www.overleaf.com/dash> with purdue.edu account
2. Do LaTeX tutorial, at least [part 1](https://www.overleaf.com/latex/learn/free-online-introduction-to-latex-part-1) but [part 2](https://www.overleaf.com/latex/learn/free-online-introduction-to-latex-part-2) is actually really helpful for us

If you have time:

1. Find .csv files on demographics: <https://datausa.io/profile/geo/west-lafayette-in/>
2. Upload them to Google Drive folder “.csv files”
3. Work on extracting them in R “Population\_Generation.R” in an organized way

**Meeting: 2/4/18**

What we did:

* We looked at the number of websites that have data and which format they have it in.
* Then we downloaded the csv files and imported them into R
* We figured out how to clean up the files so that we just have the column names and the number of people.

Thangs we need to do:

* Inform the rest of the group which file you have worked on (Age, Race etc)

Karuna- Commute Transport,

Michael - Race

Julia - Age

* Have all our cleaned up data by wed

Example R code:

transdata\_raw<- read.csv("CommuteTransport.csv",header = T)

transdata\_raw<- transdata\_raw[6,seq(4,length(transdata\_raw))] //to get the 2015 info

transdata<-c(transdata\_raw[1,c(1,3,4)],transdata\_raw[1,seq(6,24,2)],transdata\_raw[1,38])

//to get only data we need, not MOEs)

**Meeting: 2/11/18**

* Agenda
  + Complete plan for Monte Carlo simulation
    - Determine how we are breaking down the logic for creating people
    - Determine additional information we need to collect
    - Set deadlines for when portions of the code need to be completed
  + Collect data that we need to complete simulation and document it
  + Create plan moving forward to write the report and get Stage 1 completed
* School information
  + We need to determine the exact schools each region will go to for elementary, middle, and high school
  + Will create documentation with the name of each school, the type of school, and what regions apply to it in the same Sheets book on a new sheet as the Population Stats
* Things we will need to add later
  + Bus routes
  + Work places
* Monte Carlo Simulation
  + Steps of the simulation:
    - First step: read in csv file with all the relevant data
    - Second step: determine what region a person lives in
      * How this will work: We will take the population of each region and divide it by the total population. This will work as a multinomial distribution where each region has a population
    - Third step: create households
      * Have probabilities of family households versus non-family households
      * Use info about people living alone/with roommates/esc for non-family households and family sizes in regions to determine each household size
        + Each household will be a row and the columns will be attributes of the household

How many people, race, income, number of kids, esc.

* + - Fourth step: determine the characteristics that apply to the household
      * These will be things like race, household income, below poverty level, esc
      * Will need to simulate the number of people in the household and the factors that affect each
* VERY basic breakdown of the logic of the program in terms of what it handles and assigns to each household

Region

Household: # of People

# of Adults and Children

# >60 and <18

Where kids go to school

Income

Race

Public Transportation

Jobs

* Plans moving forward
  + Most important
    - Finish Parker’s part of data collection
    - Brainstorm ways to account for where people work
      * Maybe determine largest employers for WL + Lafayette
      * Maybe incorporate travel times to work
    - Complete prototype of code by Wednesday
      * Expectation is for everyone to contribute and keep everyone else informed about any changes/additions that they made
    - Have Stage 1 model completed by Friday
    - Work on report due next Monday

**Meeting: 2/15/18**

* Agenda
  + Work on Code
  + Website Design
  + Break down responsibilities for report
* Web Design Features
  + User Selection to Map:
    - Strength of Disease
    - Number of initially infected
    - Age
    - Region
    - Grouping
  + Recommended Policies
    - Schools
    - Quarantine
    - Bus routes
    - Specific workplaces
* Home Page
  + Has information about us, about page, pretty basic
* Contagion Page
  + (See picture in GroupMe)
* Action Items:
  + Michael and Karuna: Finishing code
  + Mrun: Implementing disease features and selection menu into website/finding some way

**Meeting: 2/25/18**

ER diagram: Population, Businesses, Hospitals, Schools.

Businesses: Number of employees, locations. Classification of businesses? - Michael+ Julia

Hospitals: Locations, capacities - Karuna

Parks: Locations, size - Richard

Bus routes: Routes - Michael

Homeless shelters: Locations, capacities - Mrunmayi

Libraries/Grocery stores: Locations - Parker

Health insurance assumptions? - Karuna

Ask Mario:

Include Purdue in West Lafayette? (Official city boundaries do not include Purdue)

What kind of “existing data” on public entities are we looking for? API

Links:

Data:

Tutorials: <https://www.youtube.com/watch?v=yPu6qV5byu4> (mySQL)

<https://www.youtube.com/watch?v=QpdhBUYk7Kk> (ER Diagram part1)

<https://www.youtube.com/watch?v=-CuY5ADwn24> (ER Diagram part2)

**Meeting: 2/28/18**

**ER Diagram:** <https://www.lucidchart.com/documents/edit/d150454a-2bf0-4d00-b49e-5795e4c2ea14/0?shared=true&>

**Important Not:**: Every action word must have a real reason

**Links:** <https://www.gocitybus.com/wp-content/uploads/2018/01/System-Map-2018-v3-detailed-1.compressed.pdf>

**Assignments:**

**Michael:** Bus routes for regions

**Julia:** Research businesses

**Karuna:** Research hospitals

**Richard:** Shelters & Parks

**Mrunmayi:** Grocery stores

**Ask Mario:**

Improvements to be made in **Phase I** (K-nodes & whatever)

How specific does data need to be for the model? (businesses and hospitals are very cryptic with their information)

How do we go about finding this data?

If data unavailable, what assumptions can be made? (axioms and false axioms?)

Is our ER diagram logic sound? Red flags? Suggestions?

Nodes? Entities, what is igraph for?

Purdue is a entity?

**Meeting 3/4/18**

Purdue University Regions: 68, 70,12,14,15,17, 18.

<http://www.city-data.com/city/West-Lafayette-Indiana.html>

* Watch the igraph package tutorial -( link in the assignment 2 pdf?)
* Add any capacity data found, into the data base (Database Entity sheet).
* Add questions for Mario.

**Meeting 3/21/18**

Tasks:

* ~~Divide each household into population - Michael~~
  + Potential idea is to generate people from household and then assign the attributes belonging to individuals
  + Individual is linked to separate table of households
* Website Design - Julia
  + Need to make from scratch
  + Google html codes and reference
* Optimization Algorithm (igraph?) - Michael, Karuna
* ~~ER Diagram Finalization & Database CREATE statements - Julia~~
  + ~~Possibly remove stores~~
  + ~~Have connection to homeless shelters if income is less than $10k (?)~~
* ~~Monte Carlo Simulation into Database - Michael~~
* Sources + Assumptions: Richard, Mrun, everyone else to a lesser degree
  + Research public policies for node research
  + Probability of getting sick in school, in office, on bus
  + Probability of getting sick vs contact time
  + Figure out how to do businesses for adults…
  + Factors that affect likelihood of spread
  + Likelihood to transmit to family members
  + ~~Vaccination costs (will use to remove all edges from vaccinated node)~~
  + Does CDC recommend closing
* Machine Learning Ideas - Karuna
* Idea for making the network
  + Create database
  + Populate with household.csv
  + Ask database for information about different parameters
    - Nodes will be individuals
    - Edges will be accumulation of contact between each pairing
  + Need to also include cost of each action
    - Machine learning in R to balance degree of the nodes with cost of action
* Things we’ll need to add in later
  + Susceptibility to disease
  + Initial health status
  + Spreading disease through network

**Changes to Database Required:**

* Households: Race needs to be added
* Small renaming of the database entities for Homeless shelters, contacted several administrators and who told me that they are actually day facilities. These locations will remain included and the table will be changed to Homeless\_Shelters/Day\_Facilities. The foot traffic will be taken as the yearly average visitors converted to weekly average as traffic depends on season and situation - Richard

**Meeting 3/28**

Removed Tables from ER diagram:

* Stores, Restaurants

Changes to ER diagram

* Add Belongs To Table to relate bus routes to regions

Additional things to check/add

* For parks:
  + Each person tends to visit 29 recreational sites/parks per year
  + Uniform distribution of 29\*number of people in each region/number of parks in region/52 weeks per year
* For now, nursing home is only referencing id number of 999, which will need to be changed later once we have information about how many nursing homes there are and which region(s) they service
* Workplaces: Work\_ID and Work\_Name are TBD, Region is 1, Traffic is 100
  + This will need to be changed later to actually be useful
* Schools need to be checked to see that all the school names match between the schools table and the individuals table

**Office Hours 4/2/18**

-Database stores data on population only, no temporary variables or count variables - no updating anything (health status).

-Machine learning done on K-nodes to determine best policy

-User inputs variables such as disease strength, number of K nodes, etc (many variables not necessary, only if want extra credit).. Precomputed results under these conditions will be output directly to website. (Can make another database for results, but not necessary)

-Network in iGraph can be multi-edge or aggregate edge network (former bonus points, hard to implement). After removal of nodes based on policy, we see what the effect on the network is via summary statistics such as largest component size, etc (see homework 2). Instead of trying to delete a node (weights of edges etc might have to change for that) we should re-run the simulation without the nodes.

- Research on economic loss due to shutting one office or school down data needs to be collected. Policy must be based off that.

-Disease can be assumed to be one type, with different strengths. These strengths will serve as parameters for our disease network model.

- We assume that our database only has information that the government actually has- so things like who contacts who need to be estimated for simulation, but can’t be put in database.

- We need to model hospitals like we do stores etc- have some people (not necessarily the ones sick) be there and have a probability of getting sick.

-The project report should be concise, words are a last resort, let your facts and figures display themselves

- Basic business model type descriptions- rest in Appendix. (We should definitely look at the project we found online like ours for this)

-Look into simulation of a network in igraph. More questions, may have to ask Mario next week.

- Still not sure if we are actually meant to be simulating interaction and disease spread. ?

**Meeting 4/4/18**

Things to do:

|  |  |
| --- | --- |
| Finalize Database | 10 page report |
| Research Economic Costs of Policies | Presentation (business, overall etc). |
| Ask Mario about simulation of our population on Friday. Using edgelist to identify K nodes. |  |
| Purdue Students add to Individuals Table under one household (enrollment at Purdue) |  |
| Get small network working ~25-50 nodes |  |
| Apply machine to learning to the K nodes to determine what policy. |  |
| Querying database to find number of people connected to each location, then determine the individuals who are connected to the locations with most connections. (Identifying K-nodes) |  |
| Use edgelists to build a network and use greedy algorithm to find the most central K nodes, ignoring all locations. |  |
| Once we get the K-nodes, we need to classify policies for the data set using machine learning. |  |
| Implement policy by removing the K-nodes (vaccination) OR remove nodes that are covered by policy. See before/after in disease spread. |  |

K-node threshold decided on severity of disease

4/06: Fixing Workplace Data

* How It Works
  + Will use business directory for greater Lafayette area (<http://www.greaterlafayettecommerce.com/list>) to determine an estimate of the number of businesses in our project area
  + Will assign people to work at one of the random businesses in the following way:
    - Teachers will work at one of the schools/Purdue
    - Healthcare officials will work at a healthcare facility
    - Farm-related people will all work for farming
    - Everyone else will randomly be assigned to work at one of the other firms
    - Students under age of 19 will be assigned to “work” at their schools
    - People under the age of 25 and in region of Purdue students will be assigned to “work” at Purdue
    - Based off of this, we don’t even need the schools table in the database since it will be included in Workplaces

4/07: Queries from DB to build Network

|  |  |
| --- | --- |
| **Entity Tables Used:** | Individuals I, Households H, Belongs\_To B, Region |
| **Attributes Used:** | I.ID, I.Workplaces, I.Age I.House\_ID,  H.House\_ID, H.Region\_Num, H.Income, H.Race, H.PubTrans  B.Region\_Num, B.Bus\_ID |
| Unnecessary Entities | Parks, Stores, Homeless Shelters, Restaurants, Schools, Hospitals |
| Unnecessary Attributes | I.Child, I.School, I.Health\_Status, I.OHL\_ID, I.Homeless\_Shelter, I.Susceptibility, H.nChildren, H.nPeople, H.Males, H.Females, H.nOld |

**4/08/18 Meeting**

Uploading to practice database:

library(RMySQL)

library(fields)

mydb <- dbConnect(MySQL(), user= 'jmonti', password= 'rockymountainhigh', dbname= 'jmonti', host= 'mydb.ics.purdue.edu')

on.exit(dbDisconnect(mydb))

P <- read.csv("Household.csv", header = TRUE)

for ( i in 1:nrow(P) ) {

dbSendQuery(mydb,sprintf("insert into Household(Household\_ID, Region\_Num, Race, Income, PubTrans) values('%d','%d','%s', '%s', '%s')",

P[i,1],P[i,2], P[i,4], P[i,10], P[i,11]))

}

Questions to Ask Mario

* Is database meant to include all that government would have regardless of if we are actually using it or not?
* Hospitals: are we supposed to model people going to them?
* ER Diagram: how basic can we go?
* Network: discuss whether it would be better to do everything in R using SIR model or do simulation idea with network connections determining who gets sick per day with SIR equations to determine how many people get infected
* Machine learning...in general...

**Meeting with Mario 4/9/18 (By Richard)**

-Some of our attributes are not actually attributes

-A simulation iterates day by day will be more accurate/complex (bonus points), a network analysis takes the aggregate over several months. As a note, the length of the simulation should be over several months

-As an example pick 20 as a parameter for a Poisson distribution, say it gives 24 back, randomly select 24 people at that location and those connections for the edgelists (this is related to the birthday question on Exam 1). People only come in contact with a set amount of individuals (provide justification) and therefore those who have no chance of coming contact with sick individuals can be ignored

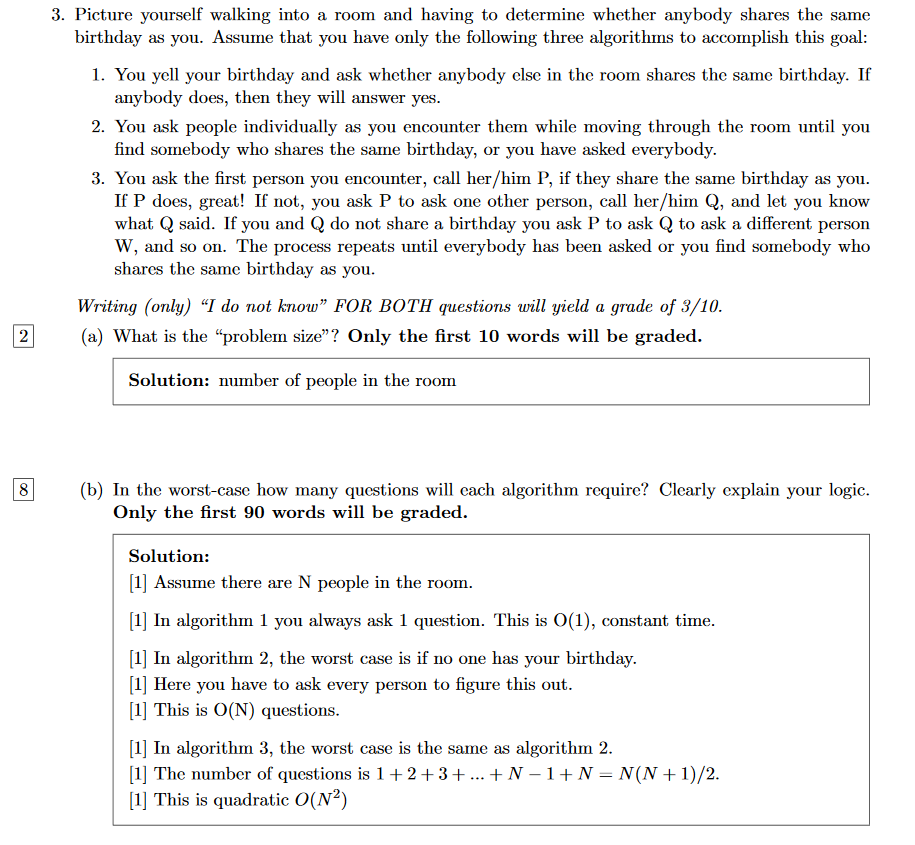
-Note, that typically the seed is unknown to the user (number of people who are sick, who is sick is unknown depending on the approach we use), policy makers would look at it as people seem to be getting sick

-Models in order of least to greatest: Differential Model, Network Analysis, Simulation

-For the first iteration, assume no policy intervention at all. Record information such as number of contacts and who they contact. Take this as the base case, integrate that into a K score, pick that top K score(the people) and see if the policy affects those individuals (if school was closed, financial cost of days of work missed, etc). Data mining context with a decision tree, interpret the tree, some branches may have no effect, some branches might be infeasible

-The decision tree represents the rules of machine learning on what it thinks the best outcome is,

-Assign the homeless shelters to a large home and a capacity, determine the percentage of it that is full, provide justification



**Meeting with Mario 4/16/18 (By Richard)**

-The way the results shown can be a database, file structure, plots, a directory with results somewhere. Have the results be retrievable with SQL queries

-Results are more than just the SIR model, the summary statistics, financial cost, days of work missed, death toll. Anything that policy makers would be interested in knowing. Bonus points: Connect R and be able to plot on the fly about the population not the disease spread

-Poverty is something that should be calculated. Instead of a poverty column there should be a salary column (continuous income), calculate the poverty in R and then add an extra column or replace a column

-Replace age with date of birth and use the exact same process as above, every analysis is run on a fixed age however. Choose a method to determine this

-What matters in terms of race/income is the process that we go about determining why we decided to use said factors

-Bonus: Doing two analyses for decision tree, one to see what policy might be effective and then the second to show that policy in effect

-Bonus: Use algorithms to prune compactable branches for complexity. Prune by hand using the algorithm, explain process for which we did this for.

**TO DO**

-Poverty is something that should be calculated. Instead of a poverty column there should be a salary column (continuous income), calculate the poverty in R and then add an extra column or replace a column **DONE**

-Replace age with date of birth and use the exact same process as above, every analysis is run on a fixed age however. Choose a method to determine this **DONE**

-NOT NULL/Other constraints for SQL Database **DONE**

-Initial infected add a category for 1 person infected **DONE**

-Population Query plots (Karuna)

-Rerun with enacted policy - (Query for affected individuals by policy, and then remove anyone who matches ID of removed ID from inf\_IDs)

-Trim Decision tree (Richard)

-Display policy information on website, infected days, cost, etc

-Update ER Diagram with relationship labels **DONE**

-Update Database Integrity - See lecture 24 slide 3 **DONE**

- Report - IN PROGRESS

- Presentation Powerpoint - -Mrunmayi

**Bonus Points Considerations**

* Phase 1, Phase 2 and Final Report all finished in LaTeX (IN PROCESS)
* Workplace names are generated with string matching algorithm (DONE)
* Ran a simulation with 100 time steps/days
* Animated gif for SIR model of disease spread (DONE)
* 36 different fully-functional results of different disease parameter combinations (DONE)
* Visual aids on population statistics (Karuna)
* Pruned decision trees displayed for user to understand the policy logic (TO DO?Richard)
* Information regarding each type of disease explained to user (TO DO? - Julia)
* Description of pros/cons of possible policies (TO DO? - Julia)
* Re-running cases with policies enacted (TO DO - Vaccination only?)

**Last Changes Made (Richard Amayo)**

**-**Static and Interactive trees created, decision tree node information, lift plots, gain curve

-