Comparing Compartment and Agent-based Models

Shannon Gallagher

July 26, 2017

Committee:
William F. Eddy (Chair)
Joel Greenhouse
Howard Seltman
Cosma Shalizi
Samuel L. Ventura

Proposal: Combine two good models into a better one

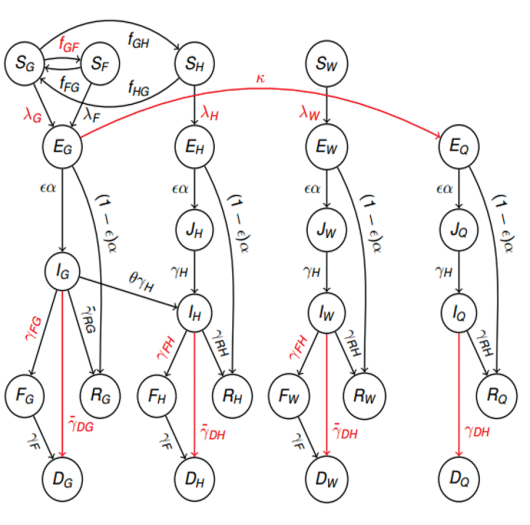
Studying infectious disease is important

Compartment and Agent-based Models are used to answer the same questions

- How bad is the flu going to be this year? (prediction)
- How do travel bans effect the spread of Ebola? (inference)
- What is the worst case scenario? (contingency planning)

Compartment vs. Agent-based Models

Compartment models (CMs) describe how individuals evolve over time



From (Pandey 2014)

S - susceptible
E - exposed
I, J - infectious
R - recovered
F - funeral
D - buried
G - general population
H - hospitalized
W - hospital worker
Q - quarantine

Examples of CMs

- Susceptible → Infectious → Removed/Recovered
  - (Kermack and McKendrick 1927)
  - plague
- S → Exposed → I → R
  - (Mills 2004, Lekone 2006, Althaus 2014, etc.)
  - influenza, Ebola
- IMmune → S → E → I → R
  - (Hethcote 2000)
  - measles, rubella, mumps
- Two-species SIR model
  - (Daley and Gani 2001)
  - malaria
- SEIR × # of Strains
  - (Blower and Chou 2004)
  - multi-strain Mycobacterium tuberculosis

Compartment models (CMs) describe how individuals evolve over time

Assumptions (Anderson and May 1992) :

1. Homogeneity of individuals

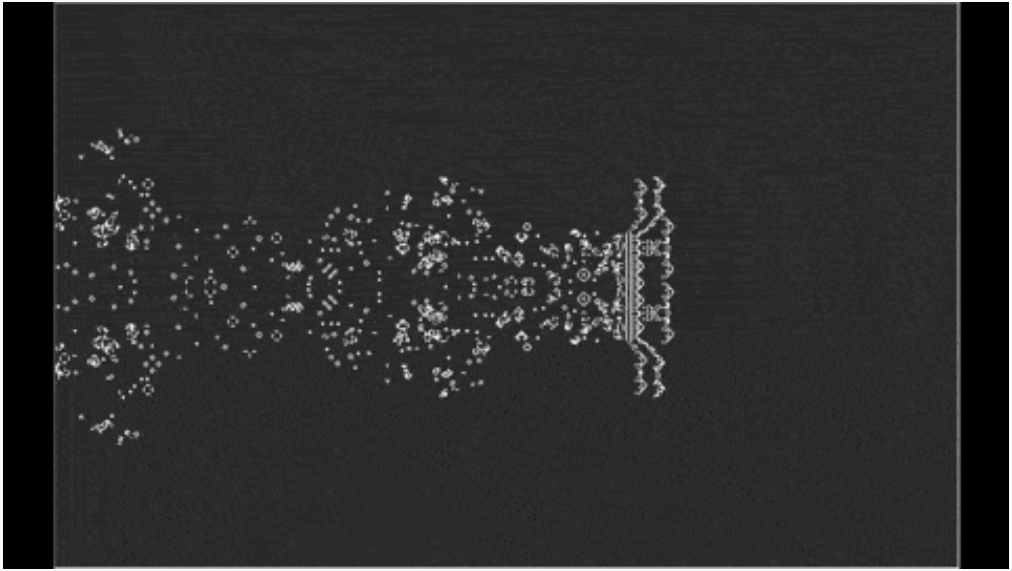
Compartment models (CMs) describe how individuals evolve over time

Assumptions **(Anderson and May 1992)** :

1. Homogeneity of individuals

2. Law of mass action
   $I(t+1) \propto I(t)$

Agent-based models (AMs) simulate phenomena



Examples of AMs

- Conway's Game of Life
  - **Adamatzky 2010**
  - cellular automata
- TRANSIMS – Transportation Analysis Simulation System
  - **(Beckman et al. 1996)**
  - data-driven traffic patterns
- EpiSims
  - **(Eubank et al. 2004)**
  - smallpox in Portland, Oregon
- FRED Framework – A Framework for Reconstructing Epiemiological Dynamics
  - **(Grefenstette et al. 2013)**
  - influenza, measles

Agent-based models (AMs) simulate the spread of disease

Assumptions (Helbing 2002):

1. Heterogeneity of agents

Agent-based models (AMs) simulate the spread of disease

Assumptions (Helbing 2002):

1. Heterogeneity of agents
2. Model adequately reflects reality

CMs and AMs: a side by side comparison

CMs

- Equation-based
- Computationally fast
- Homogeneous individuals
- No individual properties

AMs

- Simulation-based
- Computationally slow
- Heterogeneous individuals
- Individual properties

Combining the two together

(Bobashev 2007, Banos 2015, Wallentin 2017)

- ad hoc approaches
- perspective from non-statisticians

(Bobashev 2007, Banos 2015, Wallentin 2017)

- ad hoc approaches
- perspective from non-statisticians

Goal: Create a statistically justified hybrid model

Current Work

There are two main avenues of improvement

1. Quantifying how similar CMs and AMs are

2. Speeding up AM run-time

The SIR model: a detailed look

(Kermack and McKendrick 1927)

$$\begin{cases} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{cases}$$

- β – rate of infection
- γ – rate of recovery
- N – total population size

The SIR model: a detailed look

(Kermack and McKendrick 1927)

$$\begin{cases} \frac{\Delta S}{\Delta t} & = -\frac{\beta S I}{N} \\ \frac{\Delta I}{\Delta t} & = \frac{\beta S I}{N} - \gamma I \\ \frac{\Delta R}{\Delta t} & = \gamma I \end{cases}$$

- β – rate of infection
- γ – rate of recovery
- N – total population size

Our stochastic CM approach

$$\hat{S}(t+1) = \hat{S}(t) - s_t$$
$$\hat{R}(t+1) = \hat{R}(t) + r_t$$
$$\hat{I}(t+1) = N - \hat{S}(t+1) - \hat{R}(t+1),$$

with

$$s_{t+1} \sim \text{Binomial}\left(\hat{S}(t), \frac{\beta I(t)}{N}\right)$$
$$r_{t+1} \sim \text{Binomial}\left(\hat{I}(t), \gamma\right).$$

Our stochastic AM approach

For an agent $x_n(t)$, $n = 1, 2, \ldots, N$, the forward operator for $t > 0$ is

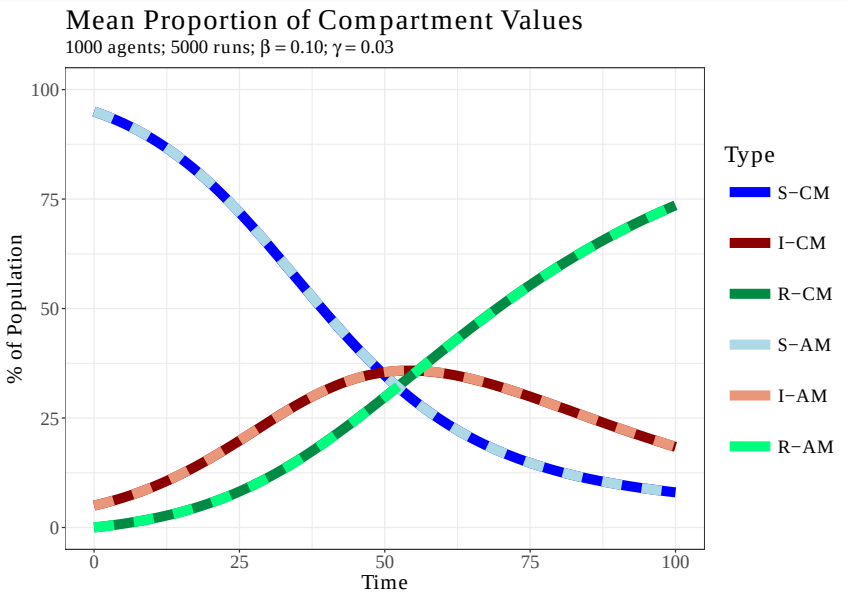
$$x_n(t+1) = \begin{cases} x_n(t) + \text{Bernoulli}\left(\frac{\beta I(t)}{N}\right) & \text{if } x_n(t) = 1 \\ x_n(t) + \text{Bernoulli}(\gamma) & \text{if } x_n(t) = 2 \\ x_n(t) & \text{otherwise} \end{cases} .$$

where $x_n(t) = k$, $k \in \{1, 2, 3\}$ corresponds to state S, I, and R, respectively

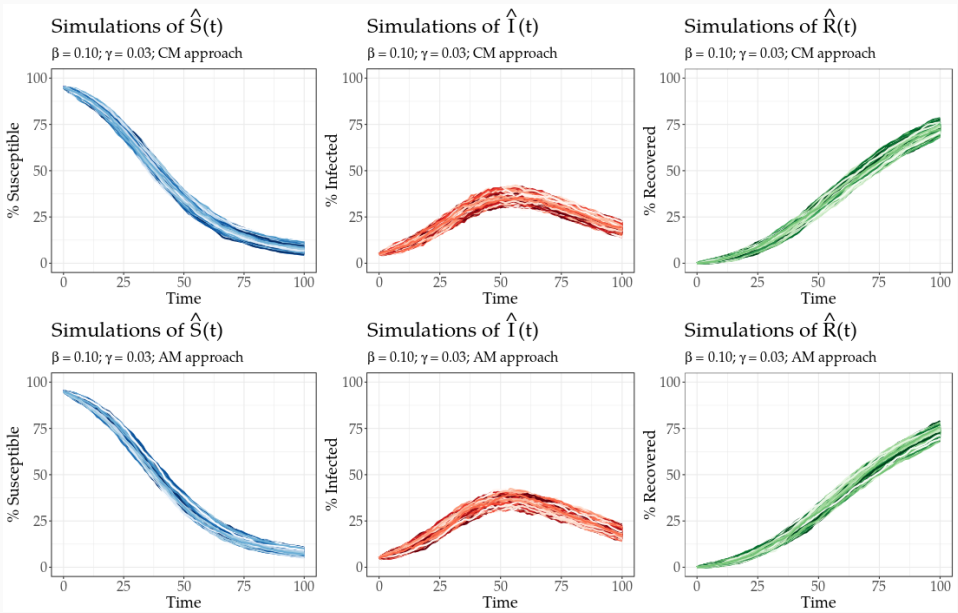
Let the aggregate total in each compartment be

$$\hat{X}_k(t) = \sum_{n=1}^{N} \mathcal{I}\{x_n(t) = k\}$$

The means overlap



The distributions look the same



These approaches are equivalent

Theorem

Let the CM and AM be as previously described. Then for all $t \in \{1, 2, \ldots, T\}$,

$$\begin{aligned} \hat{S}(t) &\stackrel{d}{=} \hat{X}_S(t) \\ \hat{I}(t) &\stackrel{d}{=} \hat{X}_I(t) \\ \hat{R}(t) &\stackrel{d}{=} \hat{X}_R(t). \end{aligned} \tag{1}$$

These approaches are equivalent

Theorem

Let the CM and AM be as previously described. Then for all $t \in \{1, 2, \ldots, T\}$,

$$\begin{aligned} \hat{S}(t) &\stackrel{d}{=} \hat{X}_S(t) \\ \hat{I}(t) &\stackrel{d}{=} \hat{X}_I(t) \\ \hat{R}(t) &\stackrel{d}{=} \hat{X}_R(t). \end{aligned} \tag{1}$$

We can construct CM-AM pairs for a given deterministic CM

Deterministic CM with K compartments X_k and difference equations

$$\frac{\Delta X_k}{\Delta t} = \sum_{i=1}^{K} D_{ik}(t) - \sum_{j=1}^{K} D_{kj}(t)$$

and $p_{ij}(t) = \frac{D_{ij}(t)}{X_i(t)}$ when $i \neq j$ and $p_{ii}(t) = 1 - \sum_{k \neq i} p_{ik}$

We can construct CM-AM pairs for a given deterministic CM

Deterministic CM with K compartments X_k and difference equations

$$\frac{\Delta X_k}{\Delta t} = \sum_{i=1}^{K} D_{ik}(t) - \sum_{j=1}^{K} D_{kj}(t)$$

and $p_{ij}(t) = \frac{D_{ij}(t)}{X_i(t)}$ when $i \neq j$ and $p_{ii}(t) = 1 - \sum_{k \neq i} p_{ik}$

CM

$$Z_i \sim \text{Multinomial}\left(\hat{X}_i, (p_{i1}(t), \ldots, p_{iK}(t))\right)$$

$$\hat{X}_k^{CM}(t+1) = \hat{X}_k^{CM}(t) + \sum_{i=1}^{K} Z_{i,k} - \sum_{j=1}^{K} Z_{k,i}$$

We can construct CM-AM pairs for a given deterministic CM

Deterministic CM with K compartments X_k and difference equations

$$\frac{\Delta X_k}{\Delta t} = \sum_{i=1}^{K} D_{ik}(t) - \sum_{j=1}^{K} D_{kj}(t)$$

and $p_{ij}(t) = \frac{D_{ij}(t)}{X_i(t)}$ when $i \neq j$ and $p_{ii}(t) = 1 - \sum_{k \neq i} p_{ik}$

CM

$$Z_i \sim \text{Multinomial}\left(\hat{X}_i, (p_{i1}(t), \ldots, p_{iK}(t))\right)$$

$$\hat{X}_k^{CM}(t+1) = \hat{X}_k^{CM}(t) + \sum_{i=1}^{K} Z_{i,k} - \sum_{j=1}^{K} Z_{k,i}$$

AM

$x_n(t+1) = j$ with probability $p_{kj}(t)$

$$\hat{X}_k^{AM}(t) = \sum_{n=1}^{N} \mathcal{I}\{x_n(t) = k\}.$$

The previous theorem generalizes

Theorem

Let the stochastic CM and AM be as in the previous slide. Then for all $t \in \{1, 2, \ldots, T, \}$,

$$\left(\hat{X}_1(t), \hat{X}_2(t), \ldots, \hat{X}_K\right)^{CM} \stackrel{d}{=} \left(\hat{X}_1(t), \hat{X}_2(t), \ldots, \hat{X}_K(t)\right)^{AM}$$

The previous theorem generalizes

Theorem

Let the stochastic CM and AM be as in the previous slide. Then for all $t \in \{1, 2, \ldots, T, \}$,

$$\left(\hat{X}_1(t), \hat{X}_2(t), \ldots, \hat{X}_K\right)^{CM} \stackrel{d}{=} \left(\hat{X}_1(t), \hat{X}_2(t), \ldots, \hat{X}_K(t)\right)^{AM}$$

Proof.

The initial conditions are designed to be the same in each model. Noting that the Multinomial draws in the CM model can be thought of a sum of independent Multinomial draws of size 1, the claim quickly follows. □

The previous theorem generalizes

Theorem

Let the stochastic CM and AM be as in the previous slide. Then for all $t \in \{1, 2, \ldots, T, \}$,

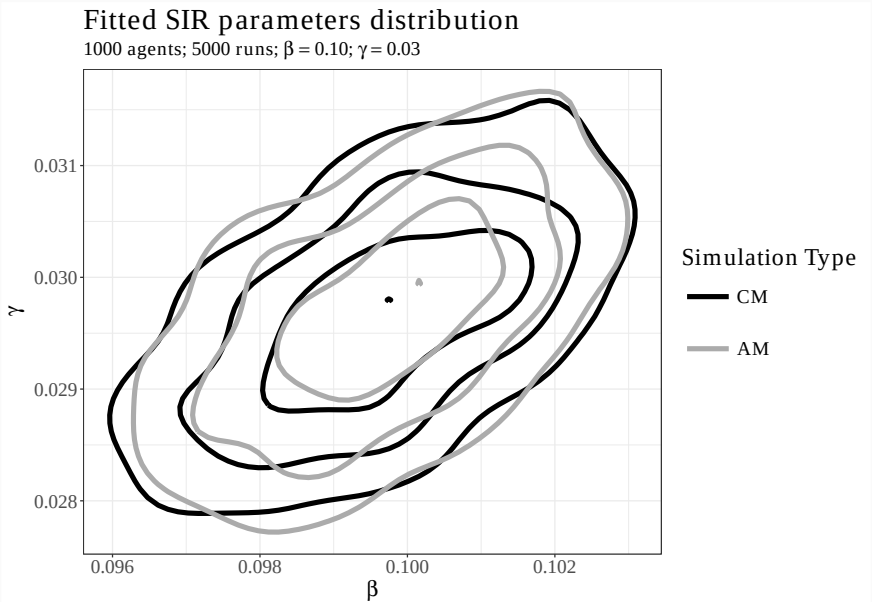
$$\left(\hat{X}_1(t), \hat{X}_2(t), \ldots, \hat{X}_K\right)^{CM} \stackrel{d}{=} \left(\hat{X}_1(t), \hat{X}_2(t), \ldots, \hat{X}_K(t)\right)^{AM}$$

Proof.

The initial conditions are designed to be the same in each model. Noting that the Multinomial draws in the CM model can be thought of a sum of independent Multinomial draws of size 1, the claim quickly follows. □

We have a foundation from which to compare CM-AM pairs!

We can compare CM/AM pairs and AM/AM pairs by fitting the underlying model



AMs are appealing because they can be run multiple times

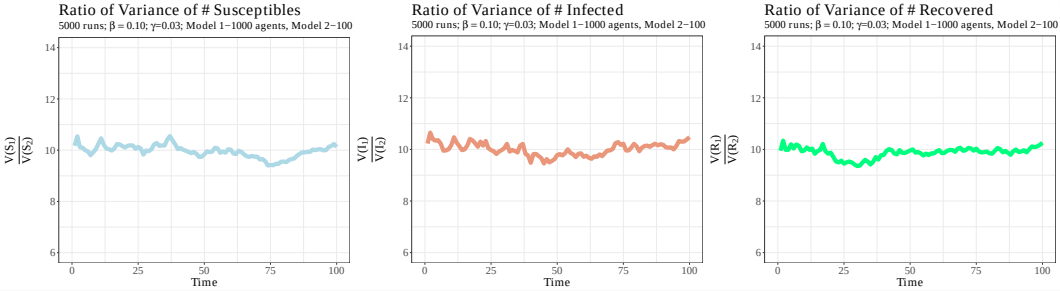
- Simulate an epidemic en masse!
- A run - same initial parameters, different random numbers
- Runs (L) are independent of one another $\implies$ **parallelization**
- Roughly, the variance of compartments $\downarrow$ when N, L $\uparrow$

AMs are appealing because they can be run multiple times

- Simulate an epidemic en masse!
- A run - same initial parameters, different random numbers
- Runs (L) are independent of one another $\implies$ **parallelization**
- Roughly, the variance of compartments $\downarrow$ when N, L $\uparrow$

Goal: Improve computation time without sacrificing statistical details

There is a tradeoff between the number of agents and number of runs



The calculations show that the variance of $\hat{S}(t)$ scales with N

- Note that for a given β and γ, if $\frac{S_1(0)}{N_1} = \frac{S_2(0)}{N_2} \implies \frac{S_1(t)}{N_1} = \frac{S_2(t)}{N_2}$

The calculations show that the variance of $\hat{S}(t)$ scales with N

- Note that for a given β and γ, if $\frac{S_1(0)}{N_1} = \frac{S_2(0)}{N_2} \implies \frac{S_1(t)}{N_1} = \frac{S_2(t)}{N_2}$
- $V\left[\hat{S}(t+1)\right] = S(t)(1-p_t)p_t + (1-p_t)^2 V\left[\hat{S}(t)\right]$

The calculations show that the variance of $\hat{S}(t)$ scales with N

- Note that for a given β and γ, if $\frac{S_1(0)}{N_1} = \frac{S_2(0)}{N_2} \implies \frac{S_1(t)}{N_1} = \frac{S_2(t)}{N_2}$
- $V\left[\hat{S}(t+1)\right] = S(t)(1-p_t)p_t + (1-p_t)^2 V\left[\hat{S}(t)\right]$
- $V[\hat{S}_2(t)] = \frac{N_2}{N_1} V[\hat{S}_1(t)]$

The calculations show that the variance of $\hat{S}(t)$ scales with N

- Note that for a given β and γ, if $\frac{S_1(0)}{N_1} = \frac{S_2(0)}{N_2} \implies \frac{S_1(t)}{N_1} = \frac{S_2(t)}{N_2}$
- $V\left[\hat{S}(t+1)\right] = S(t)(1-p_t)p_t + (1-p_t)^2 V\left[\hat{S}(t)\right]$
- $V[\hat{S}_2(t)] = \frac{N_2}{N_1} V[\hat{S}_1(t)]$

$$\frac{V\left[\frac{1}{L_1}\sum_{\text{runs } \ell} \frac{\hat{S}_1(t)}{N_1}\right]}{V\left[\frac{1}{L_2}\sum_{\text{runs } \ell} \frac{\hat{S}_2(t)}{N_2}\right]} = \frac{L_2 N_2^2}{L_1 N_1^2} \cdot \frac{V[\hat{S}_1(t)]}{V[\hat{S}_2(t)]}$$

$$= \frac{L_2 N_2}{L_1 N_1}.$$

The calculations show that the variance of $\hat{S}(t)$ scales with N

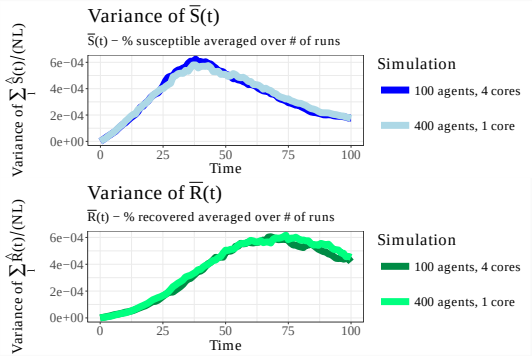
- Note that for a given β and γ, if $\frac{S_1(0)}{N_1} = \frac{S_2(0)}{N_2} \implies \frac{S_1(t)}{N_1} = \frac{S_2(t)}{N_2}$
- $V\left[\hat{S}(t+1)\right] = S(t)(1-p_t)p_t + (1-p_t)^2 V\left[\hat{S}(t)\right]$
- $V[\hat{S}_2(t)] = \frac{N_2}{N_1} V[\hat{S}_1(t)]$

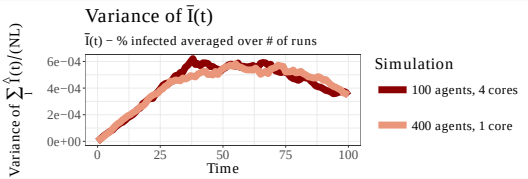
$$\frac{V\left[\frac{1}{L_1}\sum_{\text{runs } \ell} \frac{\hat{S}_1(t)}{N_1}\right]}{V\left[\frac{1}{L_2}\sum_{\text{runs } \ell} \frac{\hat{S}_2(t)}{N_2}\right]} = \frac{L_2 N_2^2}{L_1 N_1^2} \cdot \frac{V[\hat{S}_1(t)]}{V[\hat{S}_2(t)]}$$

$$= \frac{L_2 N_2}{L_1 N_1}.$$

We can replace agents with runs!

Through paralellization, we can get a speed-up without losing statistical information





Simulation 1 (100 agents, 4 cores, 100 times): 3:30 minutes

Simulation 2 (400 agents, 1 core, 100 times): 4:05 minutes

Future work

There is more work to be done: short-term

- Implementation of current methods in FRED
  - FRED - an open source, supported, flexible AM
  - Incorporate different levels of homogeneity
    1. Independent agents
    2. Agents go to one other activity (school, work, neighborhood)
    3. Multiple activities
  - Compare CM and AM parameters empirically
- Empirically determine when different regions can be combined

There is more work to be done: long-term

There is more work to be done: long-term

- Statistical tests to compare AMs to AMs
  - Fit a deterministic CM and compare parameters
  - Higher dimensional summary statistics
  - Incorporation of prevention strategies

There is more work to be done: long-term

- Statistical tests to compare AMs to AMs
  - Fit a deterministic CM and compare parameters
  - Higher dimensional summary statistics
  - Incorporation of prevention strategies
- Practical methods for speeding up AMs
  - Reduce number of agents
  - Parallelization
  - Relaxing the independence assumption

There is more work to be done: long-term

- Statistical tests to compare AMs to AMs
  - Fit a deterministic CM and compare parameters
  - Higher dimensional summary statistics
  - Incorporation of prevention strategies
- Practical methods for speeding up AMs
  - Reduce number of agents
  - Parallelization
  - Relaxing the independence assumption
- Software
  - Build a hybrid CM/AM
  - Simulate a large epidemic

There is more work to be done: long-term

- Statistical tests to compare AMs to AMs
  - Fit a deterministic CM and compare parameters
  - Higher dimensional summary statistics
  - Incorporation of prevention strategies
- Practical methods for speeding up AMs
  - Reduce number of agents
  - Parallelization
  - Relaxing the independence assumption
- Software
  - Build a hybrid CM/AM
  - Simulate a large epidemic
- Statistical analysis of compartments
  - Distribution of time within compartments
  - Distribution of parameters such as $\mathcal{R}_0$

Thank you!

Questions?