CENTRO DE INVESTIGACIÓN EN COMPLEJIDAD SOCIAL

THE UNIVERSITY OF UTAH®

How does network structure in Agent-Based Models affect epidemiological parameters?

A large simulation study

Aníbal Olivera[1], George Vega Yon[2], Chong Zhang[2], Matthew Samore[2], Karim Khader[2], Alun Thomas[2]

1. Introduction
    a. The context
    b. The goal

2. Simulation study
    a. Methods

3. Results
    a. Peak Prevalence
    b. Peak Time
    c. Generation Time
    d. Reproductive Number

4. Conclusions

Introduction

The context

The context

Early days of epidemiology → Modelling with continuous mathematics

The context

Early days of epidemiology → Modelling with continuous mathematics

- Compartmentalization hypothesis

The context

Early days of epidemiology → Modelling with continuous mathematics

- Compartmentalization hypothesis
  - An individual can be in one of three states: Susceptible (S), Infected (I), Recovered (R), etc.
  - No overlap between states.

The context

Early days of epidemiology → Modelling with continuous mathematics

- Compartmentalization hypothesis
  - An individual can be in one of three states: Susceptible (S), Infected (I), Recovered (R), etc.
  - No overlap between states.

- Homogenous Mixing hypothesis
  - Each individual has the same chance of coming into contact with an infected individual.
  - Eliminates the need to know the precise contact network.

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SI Model

$$i(t) = I(t)/N$$

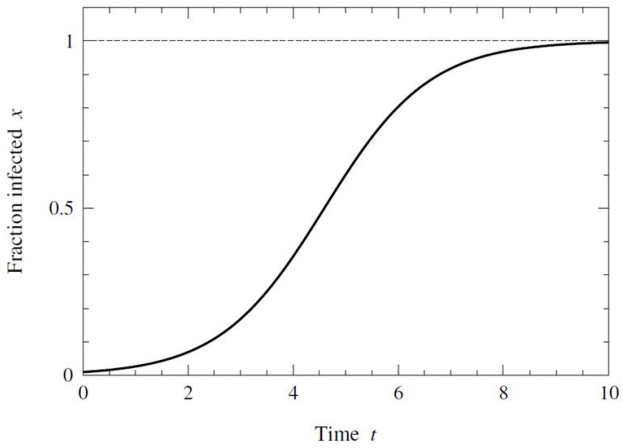
$$\frac{di}{dt} = \beta\langle k \rangle (1 - i) i$$

$$i(t) = \frac{i_0 e^{\beta\langle k \rangle t}}{1 - i_0 + i_0 e^{\beta\langle k \rangle t}}$$

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

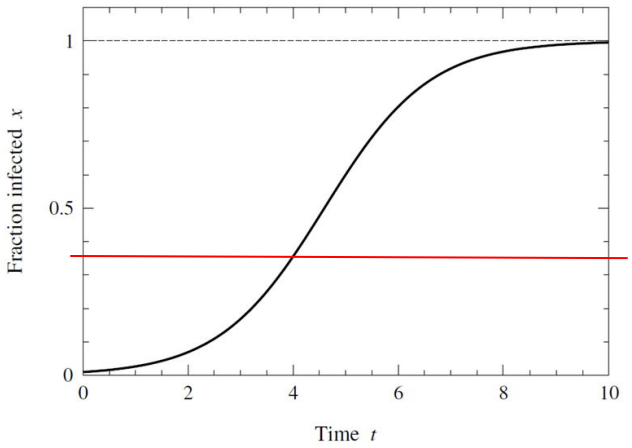
SI Model



The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SI Model



Characteristic time

$$\tau = \frac{1}{\beta \langle k \rangle}$$

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SIS Model

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SIS Model

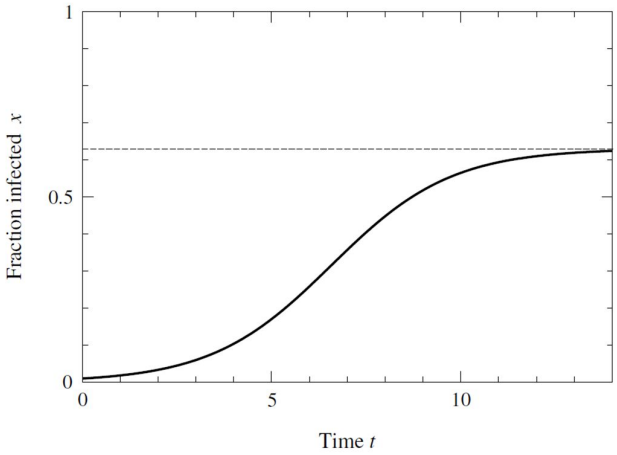
$$\frac{di}{dt} = \beta \langle k \rangle i(1-i) - \mu i$$

$$i(t) = \left[1 - \frac{\mu}{\beta \langle k \rangle}\right] \frac{Ce^{(\beta \langle k \rangle - \mu)t}}{1 + Ce^{(\beta \langle k \rangle - \mu)t}}$$

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SIS Model

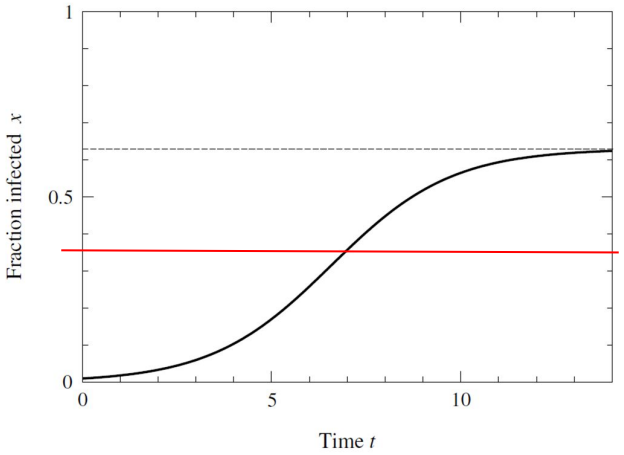


The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SIS Model

$$\tau = \frac{1}{\beta\langle k \rangle - \mu}$$

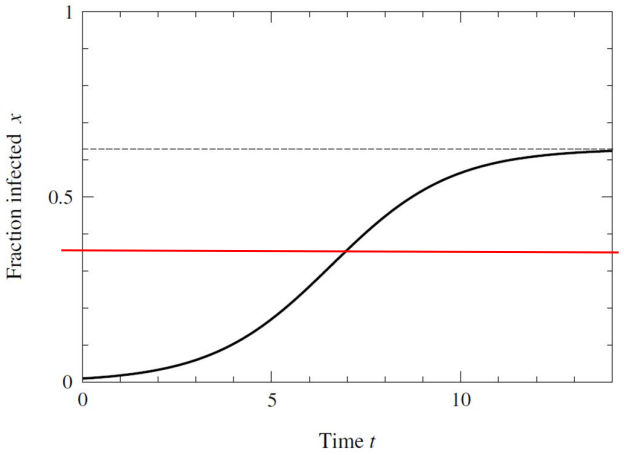


The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SIS Model

$$\tau = \frac{1}{\beta\langle k \rangle - \mu}$$



$$\tau = \frac{1}{\mu\left(R_0 - 1\right)}$$

Reproductive Number

$$R_0 = \frac{\beta\langle k \rangle}{\mu}$$

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

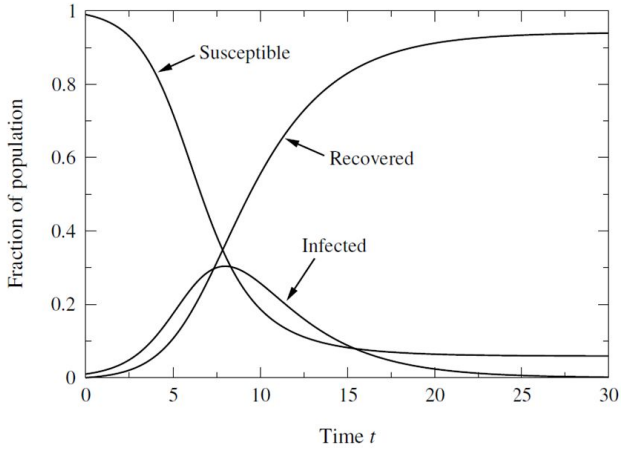
SIR Model

$$\frac{ds}{dt} = -\beta\langle k\rangle i[1 - r - i]$$
$$\frac{di}{dt} = -\mu i + \beta\langle k\rangle i[1 - r - i]$$
$$\frac{dr}{dt} = \mu i$$

The context

If we assume that a typical individual has $\langle k \rangle$ contacts, then

SIR Model



The context

The problem with $\langle k \rangle$

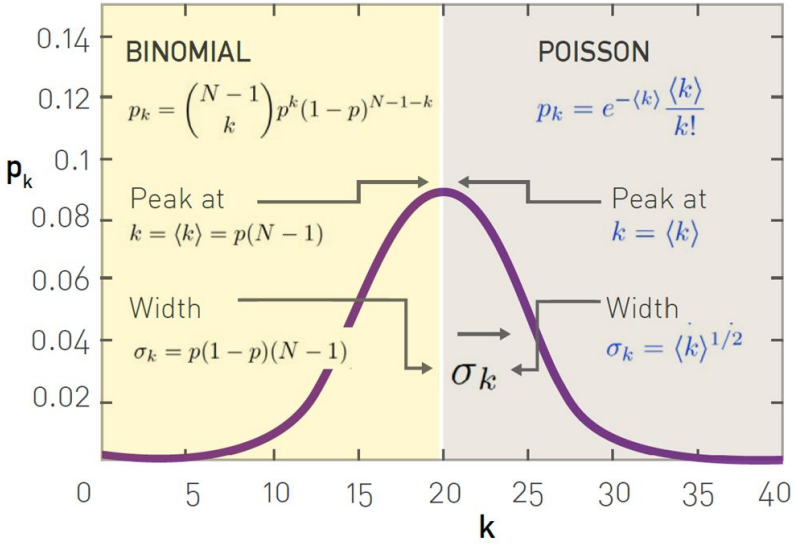
- Let's assume that all individuals interact with others randomly

The context

The problem with $\langle k \rangle$

- Let's assume that all individuals interact with others randomly

$$\rightarrow \quad p_k = \mathrm{e}^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$



The context

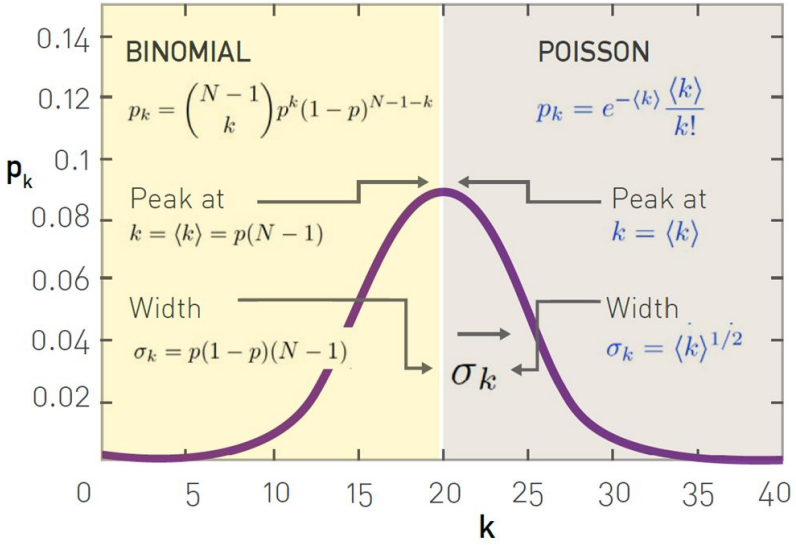
The problem with $\langle k \rangle$

- Let's assume that all individuals interact with others randomly

$$\rightarrow \quad p_k = \mathrm{e}^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

- How many individuals do we expect to interact with?

$$\langle k \rangle = 1,000$$



The context

The problem with $\langle k \rangle$

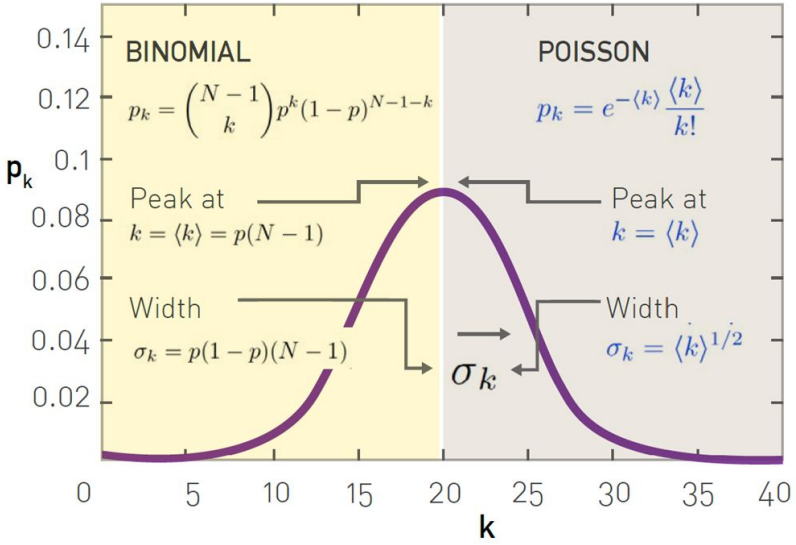
- Let's assume that all individuals interact with others randomly

$$\rightarrow \quad p_k = \mathrm{e}^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

- How many individuals do we expect to interact with?

$$\langle k \rangle = 1,000$$

$$k \approx \langle k \rangle \pm \langle k \rangle^{1/2}$$



The context

The problem with $\langle k \rangle$

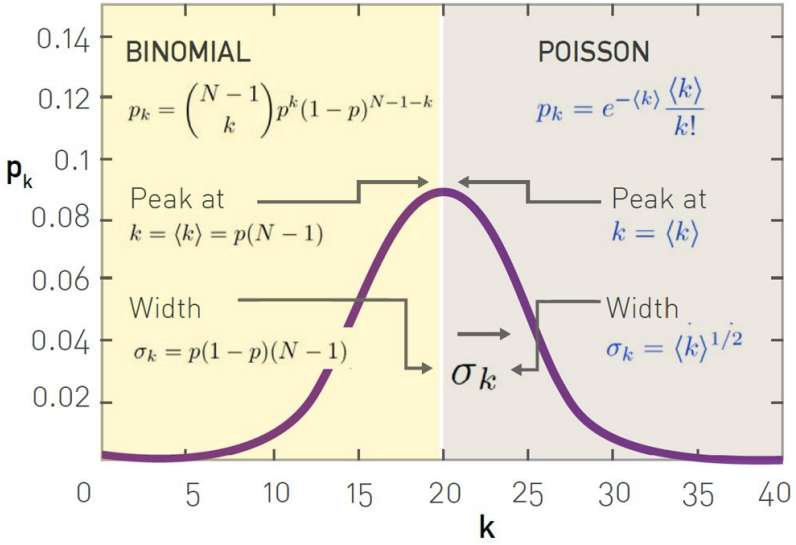
- Let's assume that all individuals interact with others randomly

$$\rightarrow \quad p_k = \mathrm{e}^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

- How many individuals do we expect to interact with?

$$\langle k \rangle = 1,000 \quad \rightarrow \quad \langle k \rangle^{1/2} = 31.62$$

$$k \approx \langle k \rangle \pm \langle k \rangle^{1/2}$$



The context

The problem with $\langle k \rangle$

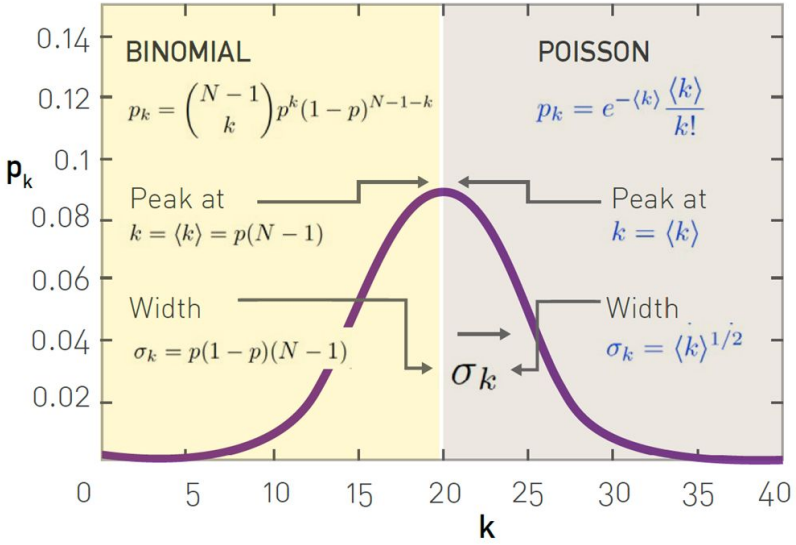
- Let's assume that all individuals interact with others randomly

$$\rightarrow \quad p_k = \mathrm{e}^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

- How many individuals do we expect to interact with?

$$\langle k \rangle = 1,000 \quad \rightarrow \quad \langle k \rangle^{1/2} = 31.62$$

$$k \approx \langle k \rangle \pm \langle k \rangle^{1/2} \quad \rightarrow \quad k \approx 1000 \pm 31.6$$



The context

The problem with $\langle k \rangle$

- Compartmentalization hypothesis
- Homogenous Mixing hypothesis

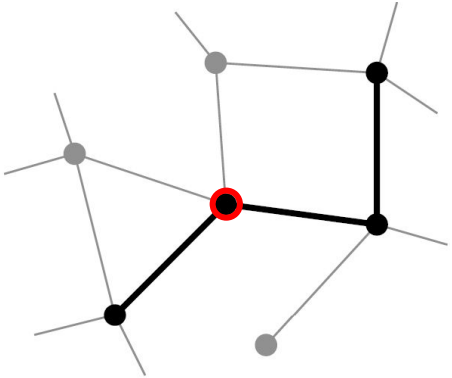
The context

The problem with $\langle k \rangle$

- Compartmentalization hypothesis
- Homogenous Mixing hypothesis

→ No realistic approach

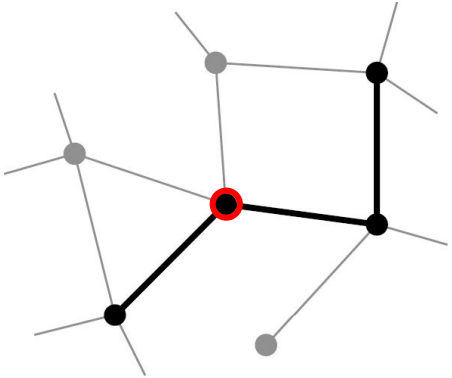
The context



In Agent Based Models, we can add the details of the ego-network of each agent

Rich network-structure

The context



In Agent Based Models, we can add the details of the ego-network of each agent

Rich network-structure

More accurate predictions

The context

The problem with $\langle k \rangle$

- Compartmentalization hypothesis
- Homogenous Mixing hypothesis

→ No realistic approach

With ABM, we have the actual k

- Compartmentalization hypothesis
- ~~Homogenous Mixing hypothesis~~ Network structure

The context

The problem with $\langle k \rangle$

- Compartmentalization hypothesis
- Homogenous Mixing hypothesis

→ No realistic approach

With ABM, we have the actual k

- Compartmentalization hypothesis
- ~~Homogenous Mixing hypothesis~~ Network structure
- We can compute the Prevalence and R_0 more accurately

The context

The problem with $\langle k \rangle$

- Compartmentalization hypothesis
- Homogenous Mixing hypothesis

→ No realistic approach

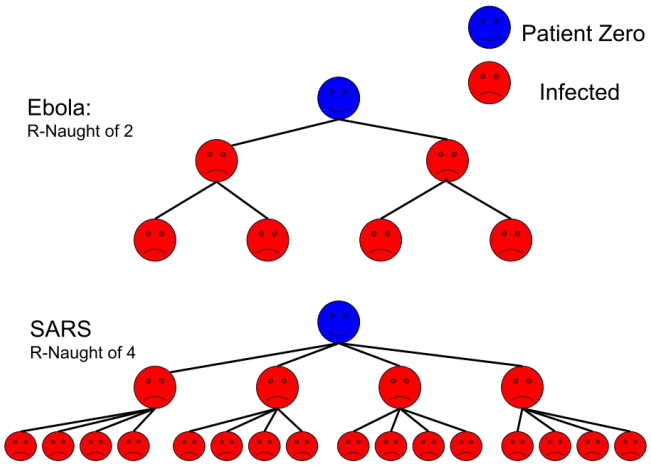
With ABM, we have the actual k

- Compartmentalization hypothesis
- ~~Homogenous Mixing hypothesis~~ Network structure
- We can compute the Prevalence and R_0 more accurately

→ ABMs are an important research and policy tool in epidemiology

The context

The idea behind R_0

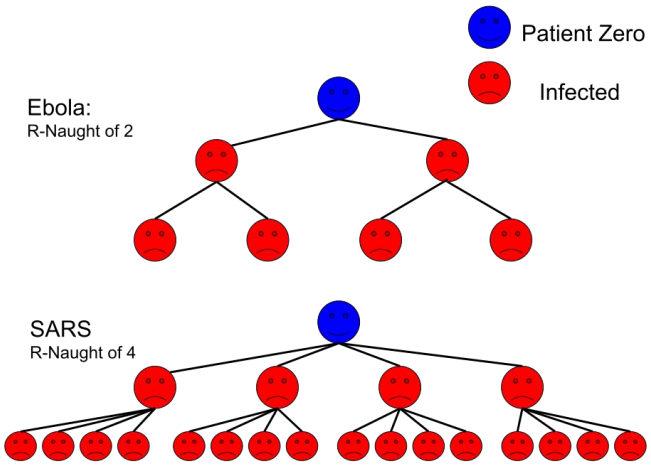


$R_0 > 1$: Epidemic grows

$R_0 < 1$: Epidemic declines

The context

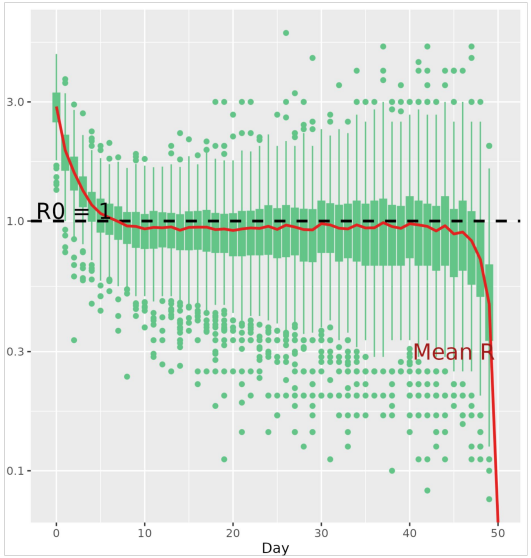
The idea behind R_0



$R_0 > 1$: Epidemic grows

$R_0 < 1$: Epidemic declines

but…



Actually $R_0 < 1$ and still feature full transmission, both in SIR and SEIR! [1, 2]

The goal

- We know that real-networks are not random.

The goal

- We know that real-networks are not random.
- Most studies uses particular network models to simulate outbreaks.

The goal

- We know that real-networks are not random.
- Most studies uses particular network models to simulate outbreaks.
- Each network topology is characterized by different levels of local-structures.

The goal

- We know that real-networks are not random.
- Most studies uses particular network models to simulate outbreaks.
- Each network topology is characterized by different levels of local-structures.

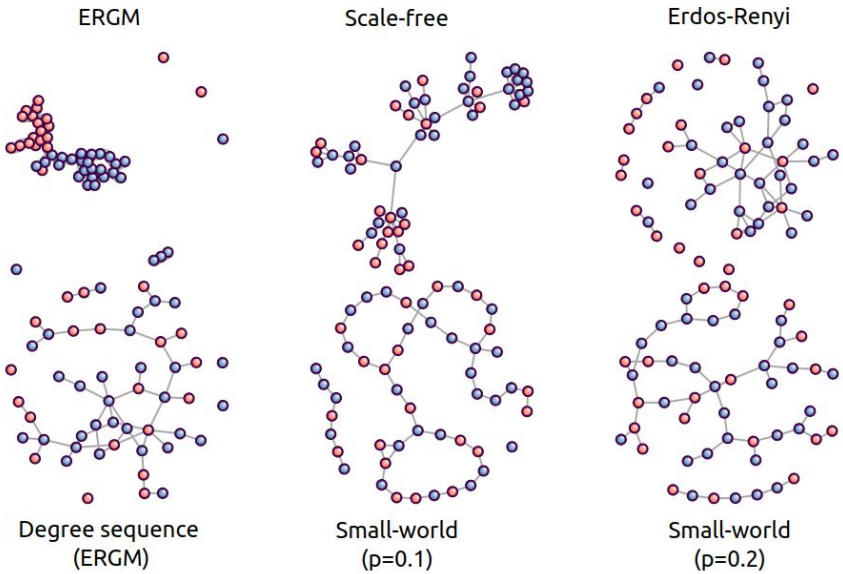
How much do these local structures affect different epidemiological indicators?

Simulation Study

Methods

Based on real-network data:

- We constructed **1,000 networks** of six topologies, all with similar density.



$$N = 534$$

$$\rho \sim 0.029$$

Methods

Based on real-network data:

- We constructed **1,000 networks** of six topologies, all with similar density.
- We simulated **20,000 SEIR** outbreaks, using the epiworldR package [3].

Methods

Based on real-network data:

- We constructed **1,000 networks** of six topologies, all with similar density.
- We simulated **20,000 SEIR** outbreaks, using the epiworldR package [3].

$$\beta = 1/43$$

Transmission rate
[4]

Methods

Based on real-network data:

- We constructed **1,000 networks** of six topologies, all with similar density.
- We simulated **20,000 SEIR** outbreaks, using the epiworldR package [3].

$$\beta = 1/43 \qquad a = 1/7$$

Transmission rate [4]

7 Days from Exposed to Infected

Methods

Based on real-network data:

- We constructed **1,000 networks** of six topologies, all with similar density.
- We simulated **20,000 SEIR** outbreaks, using the epiworldR package [3].

$\beta = 1/43$ — Transmission rate [4]

$a = 1/7$ — 7 Days from Exposed to Infected

$\gamma = 1/7$ — 7 Days from Infected to Recovered

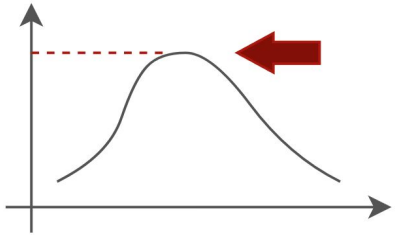
$T = 100$

Methods

From each outbreak, we get:

- Four epidemiological indicators

Peak Prevalence

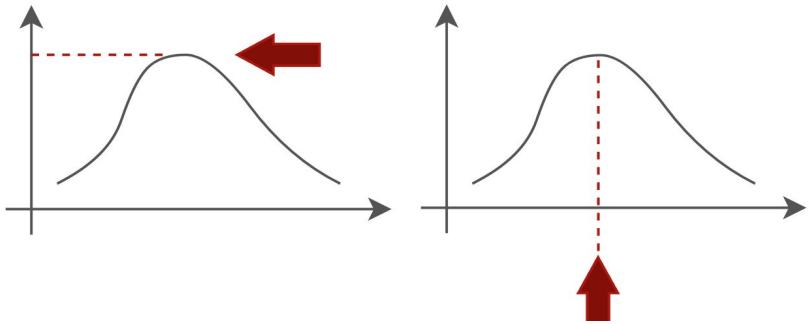


Methods

From each outbreak, we get:

- Four epidemiological indicators

Peak Prevalence - Peak time

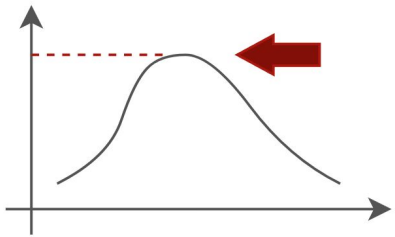


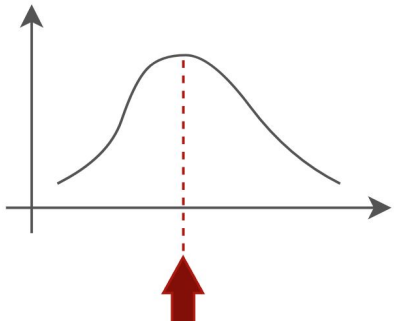
Methods

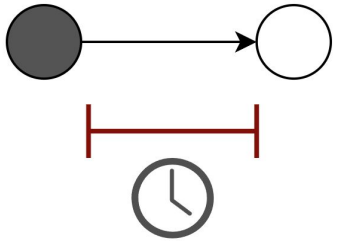
From each outbreak, we get:

- Four epidemiological indicators

Peak Prevalence - Peak time - Generation time





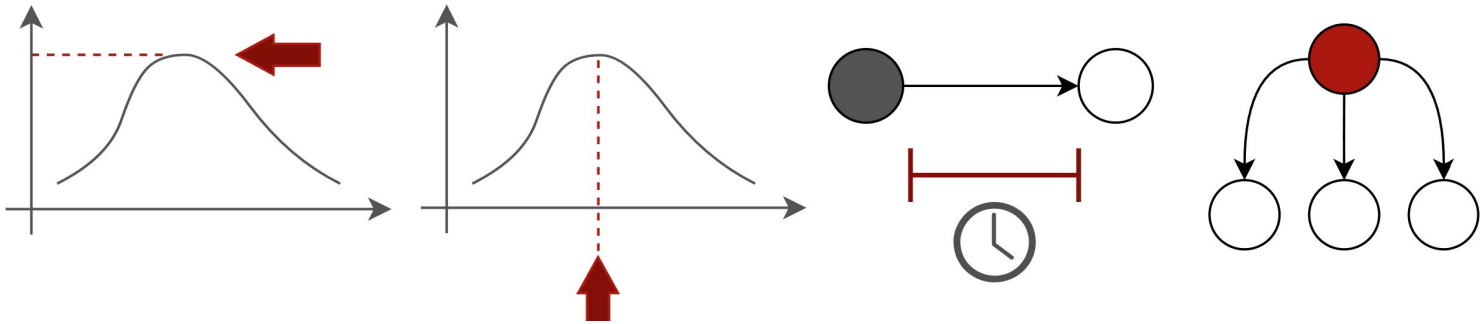


Methods

From each outbreak, we get:

- Four epidemiological indicators

Peak Prevalence - Peak time - Generation time - Reproductive number



Methods

From each outbreak, we get:

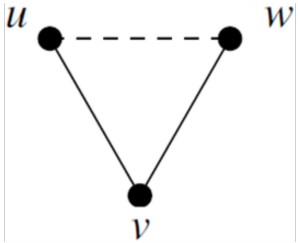
- Four epidemiological indicators

  Peak Prevalence - Peak time - Generation time - Reproductive number

- From the underlying network

  # Two paths - # Balance - # Triangles

  Avg. degree - Avg. path length - Transitivity - Modularity



Methods

- So now we can write regressions models like

$$R_0 \sim \beta_0 + \text{ Net-type } + \beta_2 \text{ Two paths } + \beta_3 \text{ Modularity}$$

Methods

- So now we can write regressions models like

$$R_0 \sim \beta_0 + \text{ Net-type } + \beta_2 \text{ Two paths } + \beta_3 \text{ Modularity}$$

- But what about…

$$R_0 \sim \beta_0 + \text{ Net-type } + \beta_2 \text{ Two paths } + \beta_3 \text{ Modularity } + \beta_4 \text{ Modularity }^2$$

Methods

- So now we can write regressions models like

$$R_0 \sim \beta_0 + \text{ Net-type } + \beta_2 \text{ Two paths } + \beta_3 \text{ Modularity}$$

- But what about…

$$R_0 \sim \beta_0 + \text{ Net-type } + \beta_2 \text{ Two paths } + \beta_3 \text{ Modularity } + \beta_4 \text{ Modularity }^2$$

- …or

$$R_0 \sim \beta_0 + \text{ Net-type } + \beta_2 \text{ Two paths } + \beta_3 \text{ log(Modularity)}$$

Methods

- We write all possible combinations of linear, logarithmic, and quadratic form of each variable.

Methods

- We write all possible combinations of linear, logarithmic, and quadratic form of each variable.
- How many?

$$\text{Total formulas} = \sum_{k=1}^{n} (1 + 2k) \binom{n}{k}$$

Quadratic + Log forms of each variable

Possible ways to select *k* variables from *n* variables

Methods

- We write all possible combinations of linear, logarithmic, and quadratic form of each variable.
- How many?

$$\text{Total formulas } = \sum_{k=1}^{n} (1 + 2k) \binom{n}{k}$$

Quadratic + Log forms of each variable

Possible ways to select *k* variables from *n* variables

$$n = 7 \implies \text{Total formulas } = 1,023$$

Methods

- We write all possible combinations of linear, logarithmic, and quadratic form of each variable.
- How many?

$$\text{Total formulas } = \sum_{k=1}^{n} (1 + 2k) \binom{n}{k}$$

Quadratic + Log forms of each variable

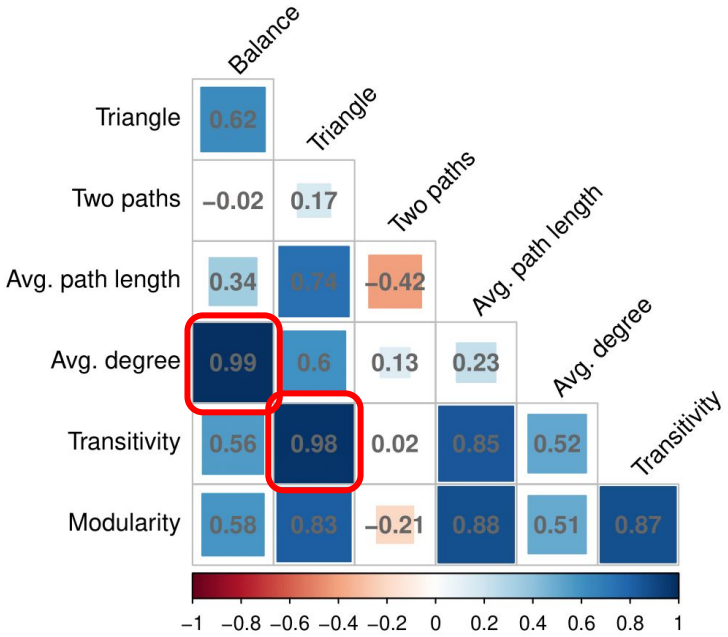
Possible ways to select *k* variables from *n* variables

$$n = 7 \implies \text{Total formulas } = 1,023$$

4 Dependent Variables $\implies$ $4 \times \text{ Total formulas } = 4,092$

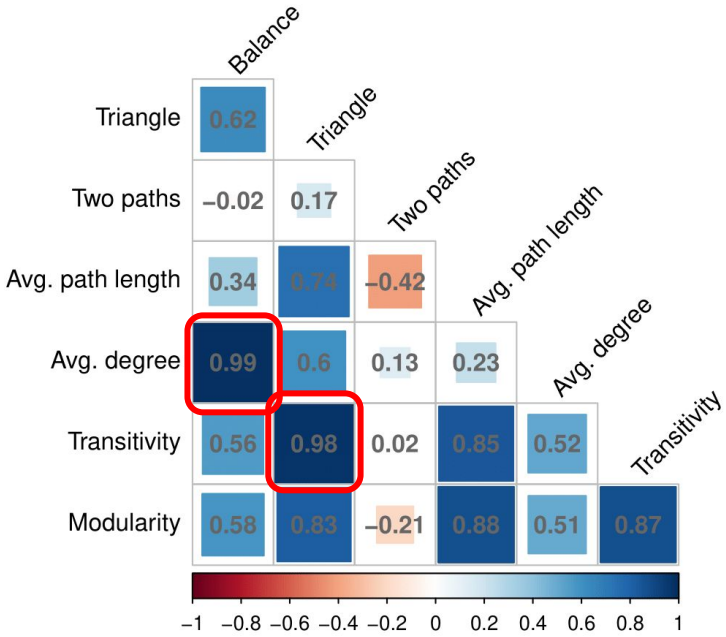
Methods

- Unfortunately, there is high correlation between the variables



Methods

- Unfortunately, there is high correlation between the variables



- To identify the problematic models:

  Generalized Variance Inflation Factor (GVIF)

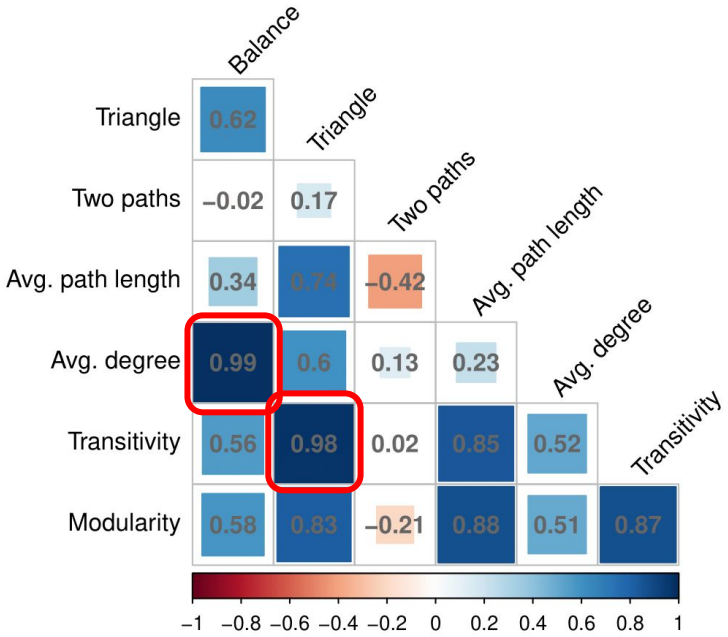
- If..

$$\mathrm{GVIF}^{(1/2\mathrm{Df})} > 10$$

..the model was dropped

Methods

- Unfortunately, there is high correlation between the variables



- To identify the problematic models:

  Generalized Variance Inflation Factor (GVIF)

- If..

$$\mathrm{GVIF}^{(1/2\mathrm{Df})} > 10$$

..the model was dropped

- This left us with 208 models in total

Methods

- Our goal is to find a well-performing and general model for each epidemiological indicator.

Methods

- Our goal is to find a well-performing and general model for each epidemiological indicator.

- To find best performing models, we:

  1. Got their AIC and BIC values

Methods

- Our goal is to find a well-performing and general model for each epidemiological indicator.

- To find best performing models, we:

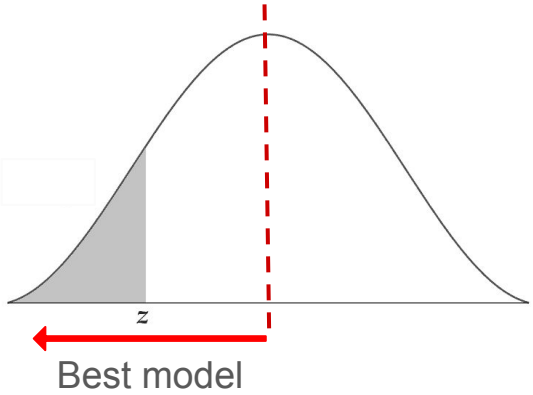
  1. Got their AIC and BIC values
  2. Normalized those values

Methods

- Our goal is to find a well-performing and general model for each epidemiological indicator.

- To find best performing models, we:

  1. Got their AIC and BIC values
  2. Normalized those values
  3. Computed the mean of the z-scores for AIC and BIC

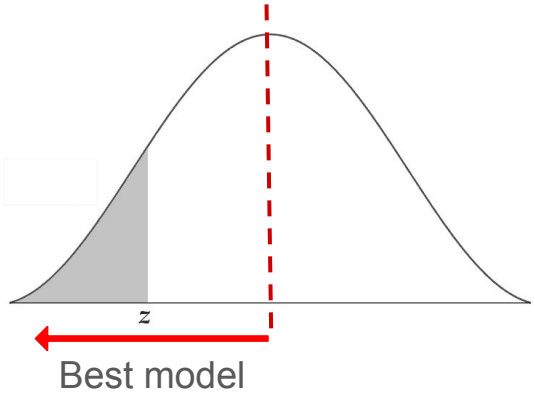


Methods

- Our goal is to find a well-performing and general model for each epidemiological indicator.

- To find best performing models, we:

  1. Got their AIC and BIC values
  2. Normalized those values
  3. Computed the mean of the z-scores for AIC and BIC
  4. Sort the models according to this number



Results

Results

- Since we have several regression models, we can search for patterns among the best models.

Results

- Since we have several regression models, we can search for patterns among the best models.

- So we:

  1. Select the $Q\%$ **best performing** models, with $Q \in (10, 25, 50)$

Results

- Since we have several regression models, we can search for patterns among the best models.

- So we:

  1. Select the $Q\%$ **best performing** models, with $Q \in (10, 25, 50)$

  2. Looked to the **frequency of each variable** among those models..

Results

- Since we have several regression models, we can search for patterns among the best models.

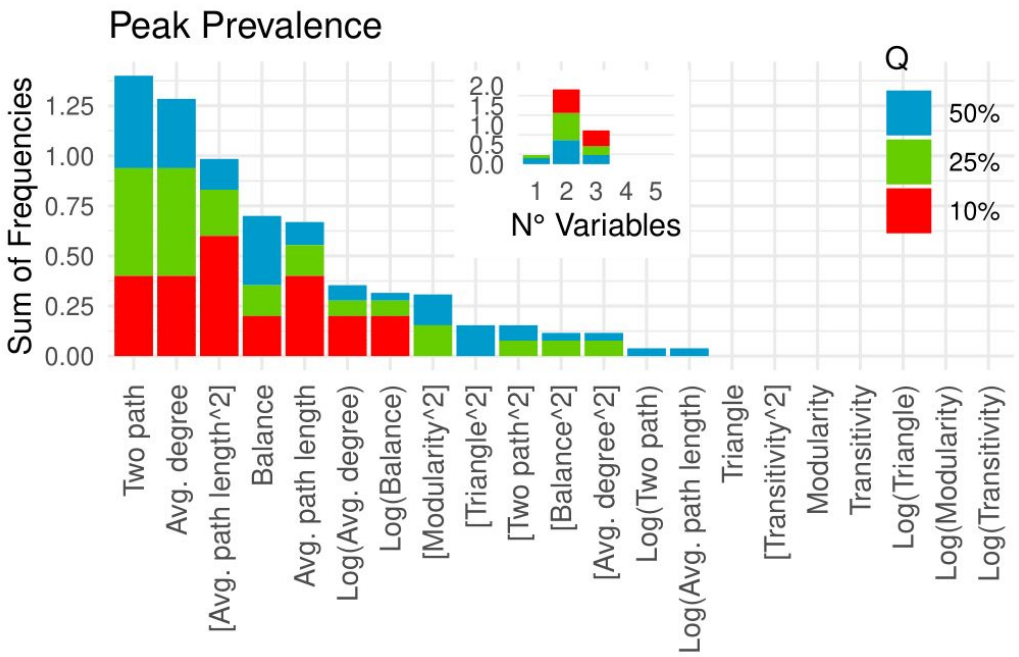
- So we:

  1. Select the $Q\%$ **best performing** models, with $Q \in (10, 25, 50)$

  2. Looked to the **frequency of each variable** among those models..

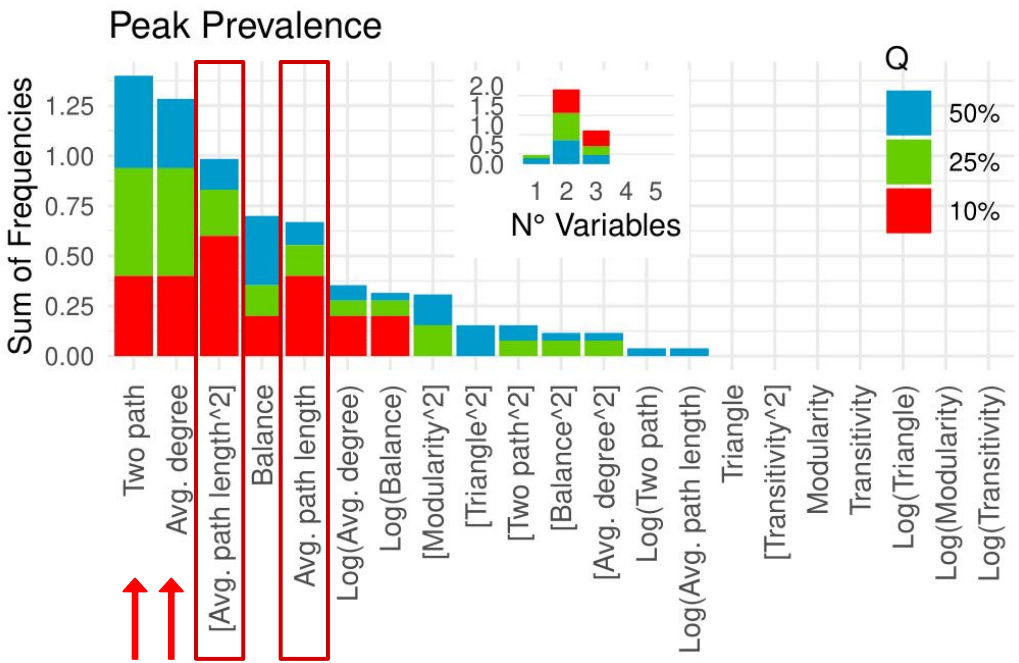
  3. Looked to the **total number of variables** of those models

Results - Peak Prevalence



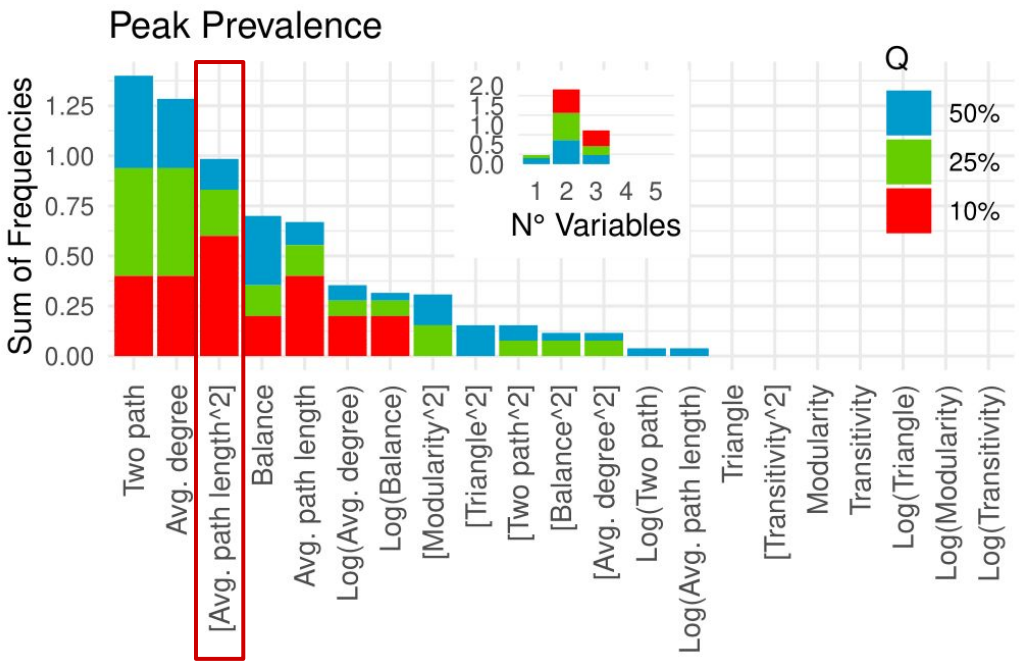
- The best 10% have only **2** or **3** *variables*

Results - Peak Prevalence



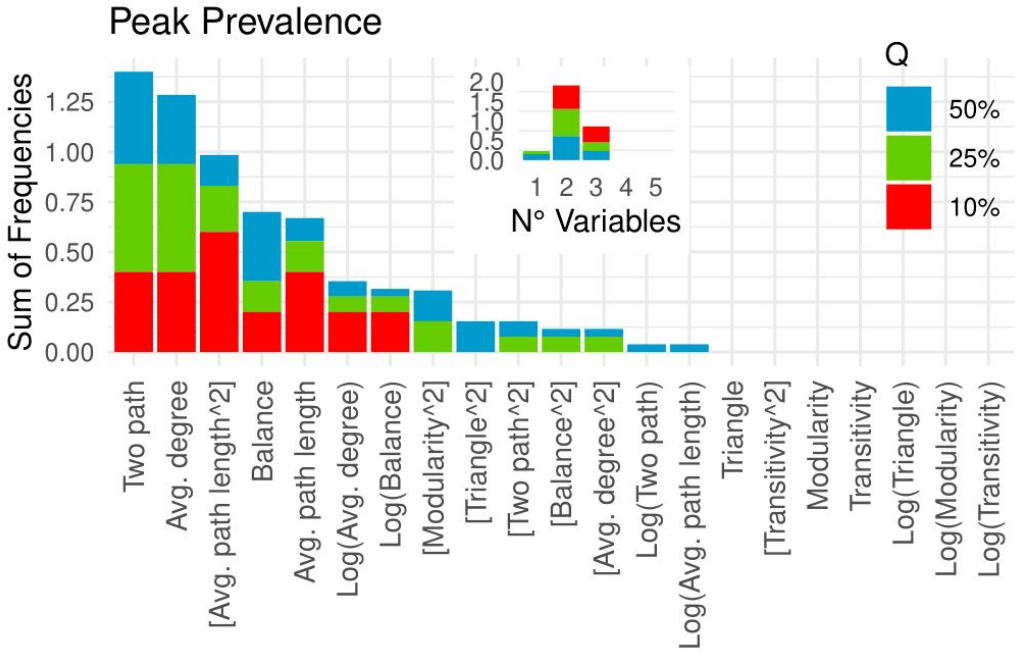
- The best 10% have only **2** or **3** *variables*
- The most prominent variable is **Avg. path length** (linear and quadratic form), along with Two path - Avg. degree

Results - Peak Prevalence



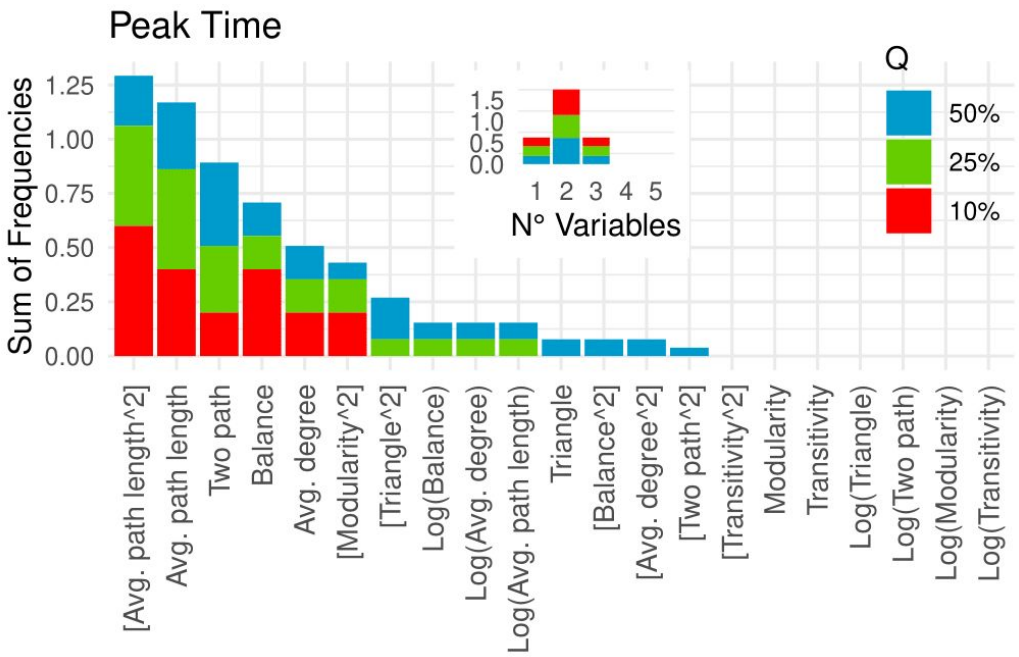
- The best 10% have only **2** or **3** *variables*
- The most prominent variable is **Avg. path length** (linear and quadratic form), along with Two path - Avg. degree
- [Avg. path length^2] has more frequency across all Q

Results - Peak Prevalence



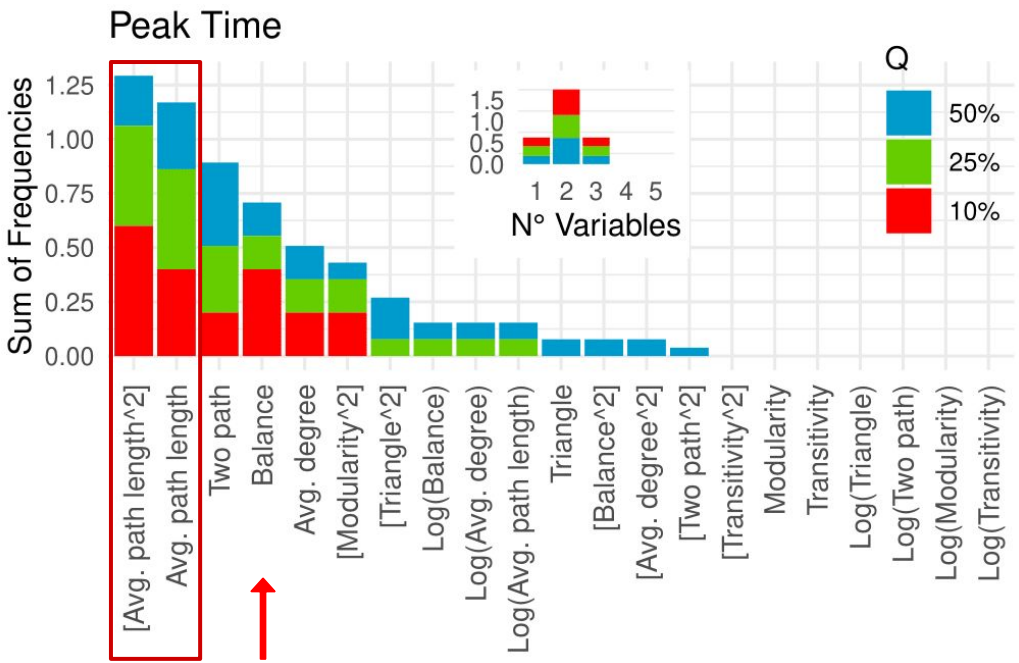
$$\begin{aligned}\text{Peak}_P = \text{Net-type} &+ \beta_0 \text{ two path} \\ &+ \beta_1 \text{ avg. degree} \\ &+ \beta_2 \left[\text{ avg. path length}^2 \right]\end{aligned}$$

Results - Peak Time



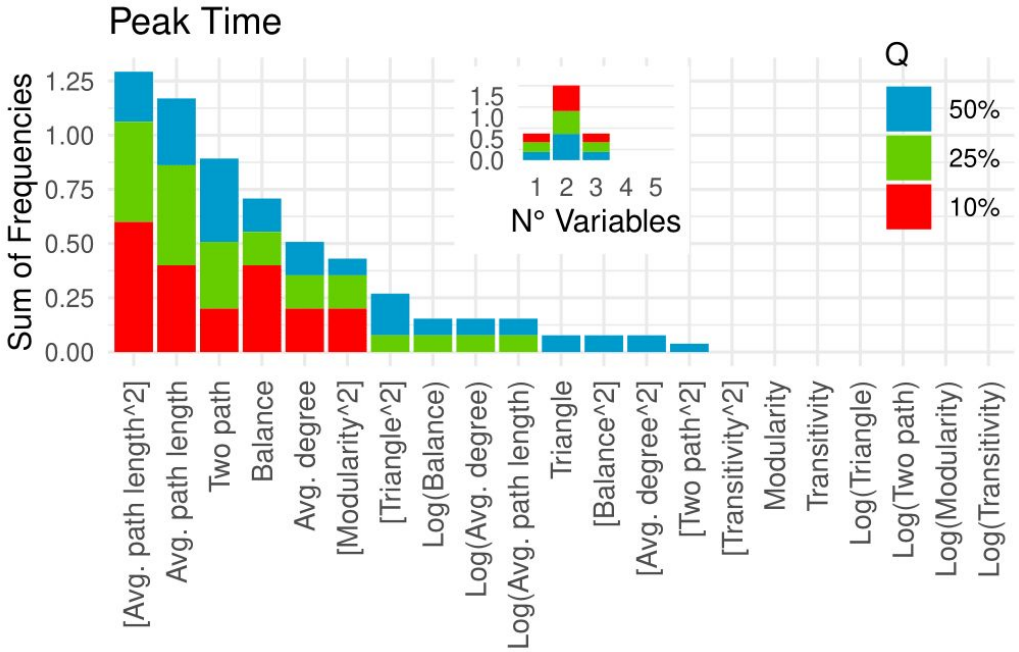
- Models with **2** *variables* performs better

Results - Peak Time



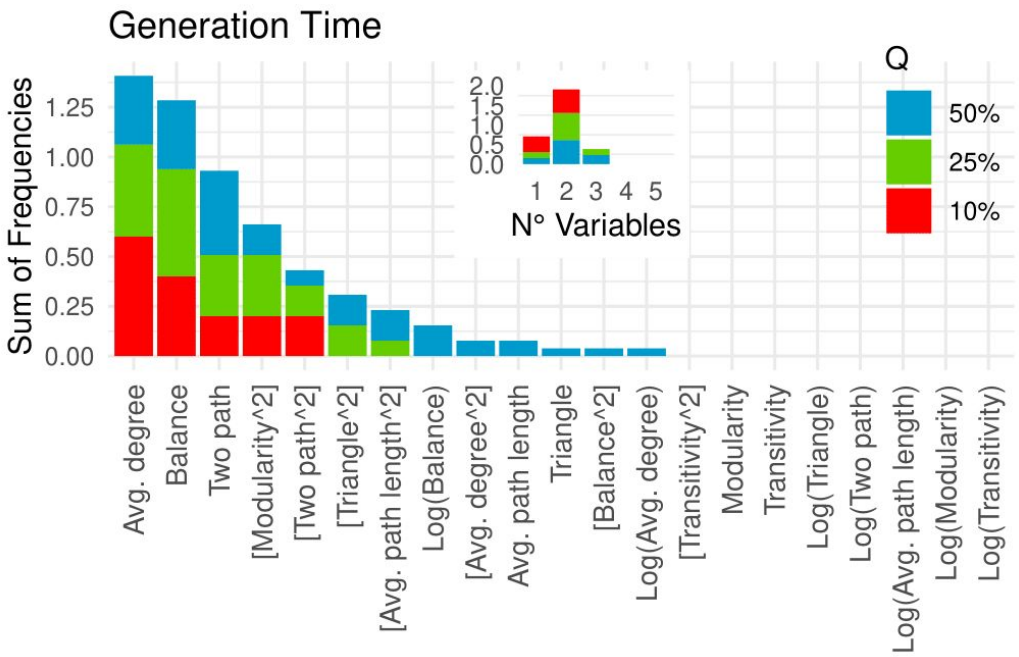
- Models with **2** *variables* performs better
- Once more **Avg. path length** is relevant (linear and quadratic form), along with **Balance**

Results - Peak Time



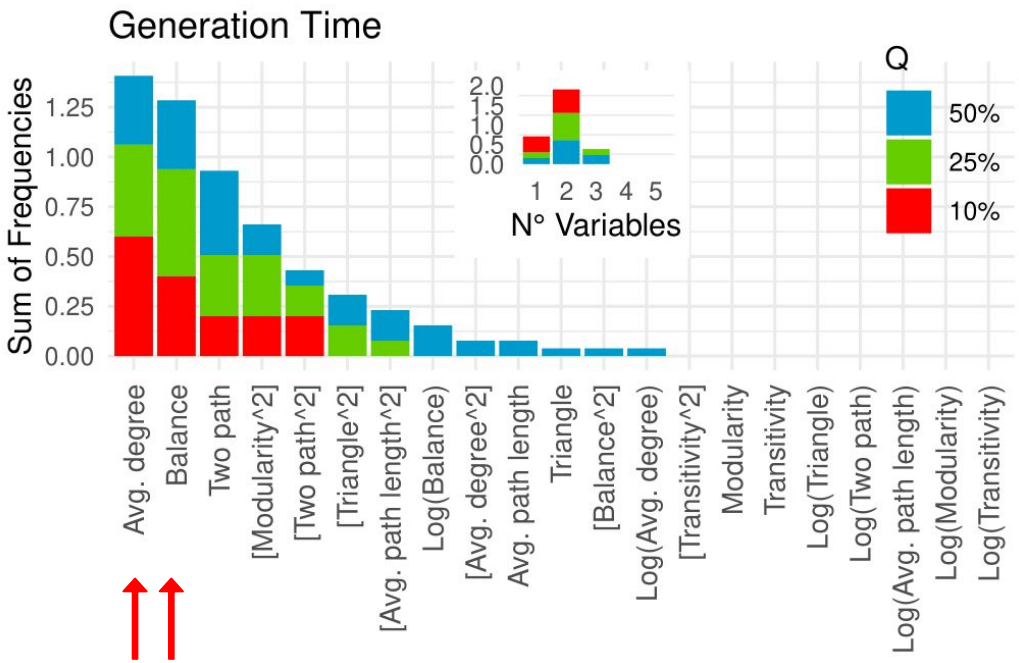
$$\text{Peak}_T = \text{Net-type} + \beta_0 \text{ balance} + \beta_1 [\text{ avg. path length}^2]$$

Results - Generation Time



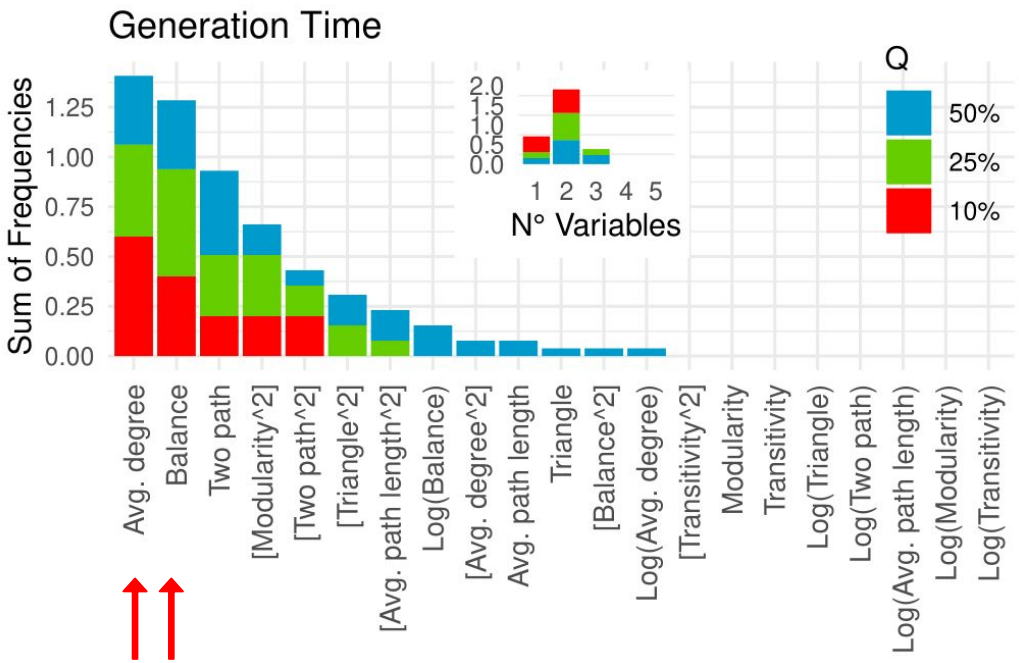
- The best models have only **1** or **2** *variables*

Results - Generation Time



- The best models have only **1** or **2** *variables*
- Clear prominence of **Avg. degree** and **Balance**…

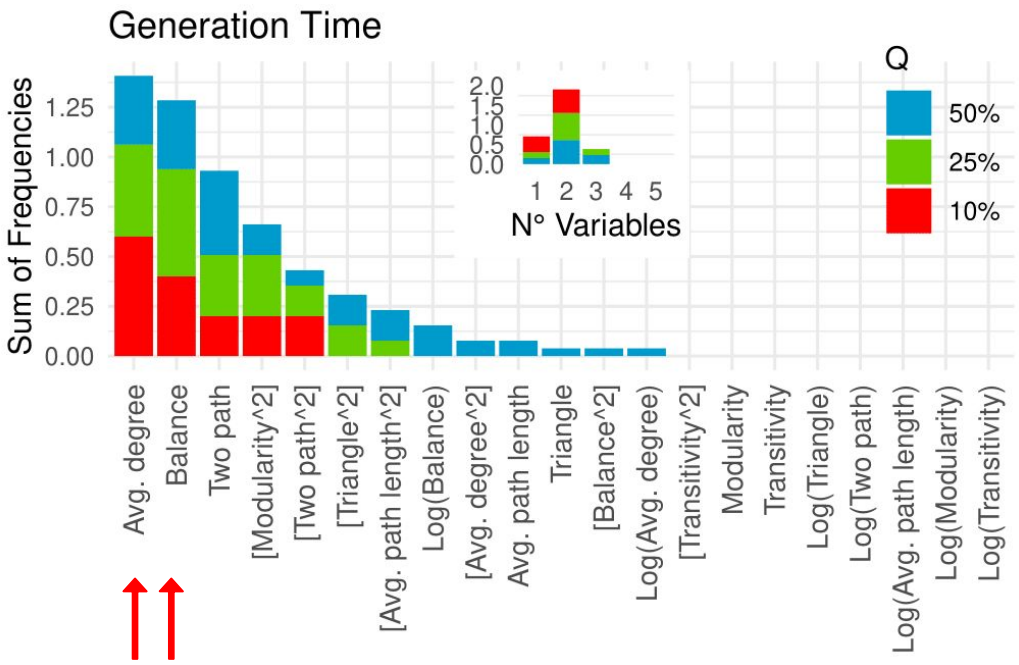
Results - Generation Time



- The best models have only **1** or **2** *variables*
- Clear prominence of **Avg. degree** and **Balance**…

  … but they shows high correlation

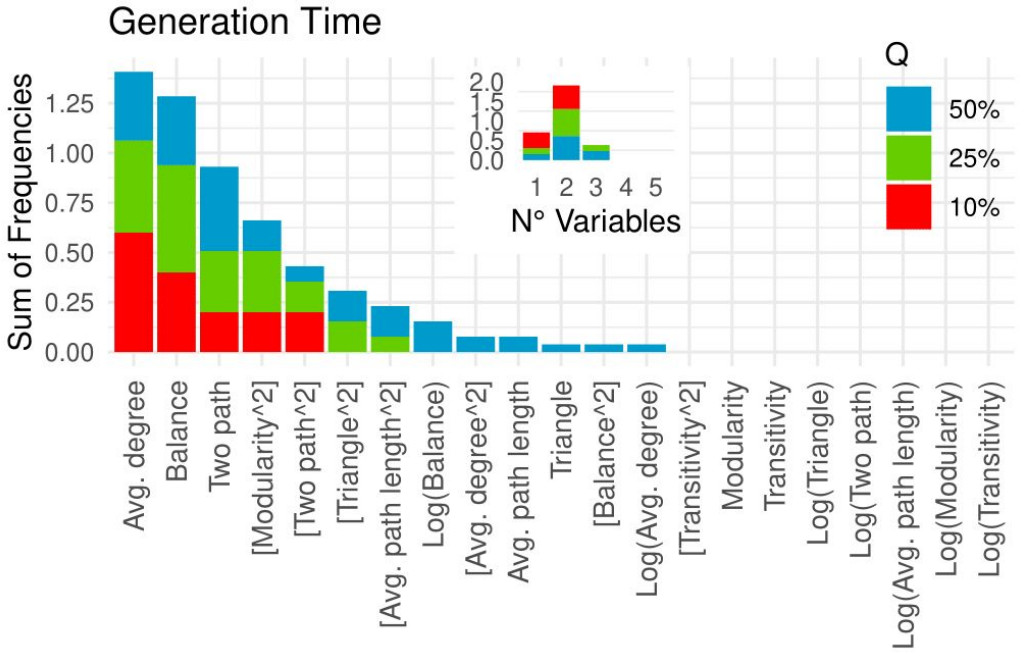
Results - Generation Time



- The best models have only **1** or **2** *variables*
- Clear prominence of **Avg. degree** and **Balance**…

  … but they shows high correlation
- So we get Balance / Avg. degree along (optional) with other relevant variable

Results - Generation Time

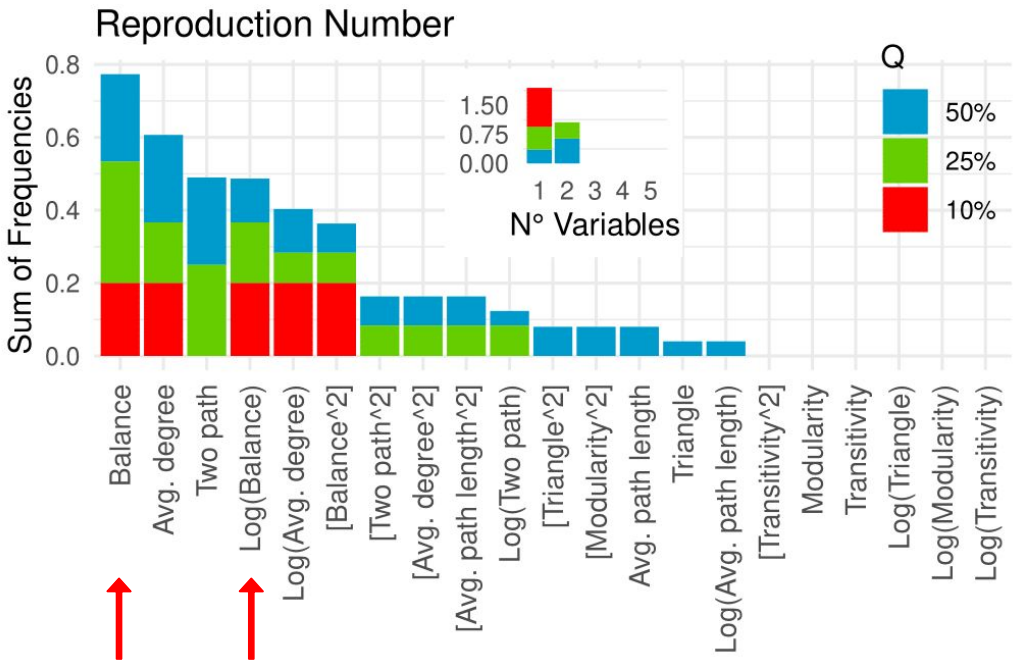


$$\text{Gen. Time} = \text{Net-type} + \beta_0 \text{ balance}$$

or

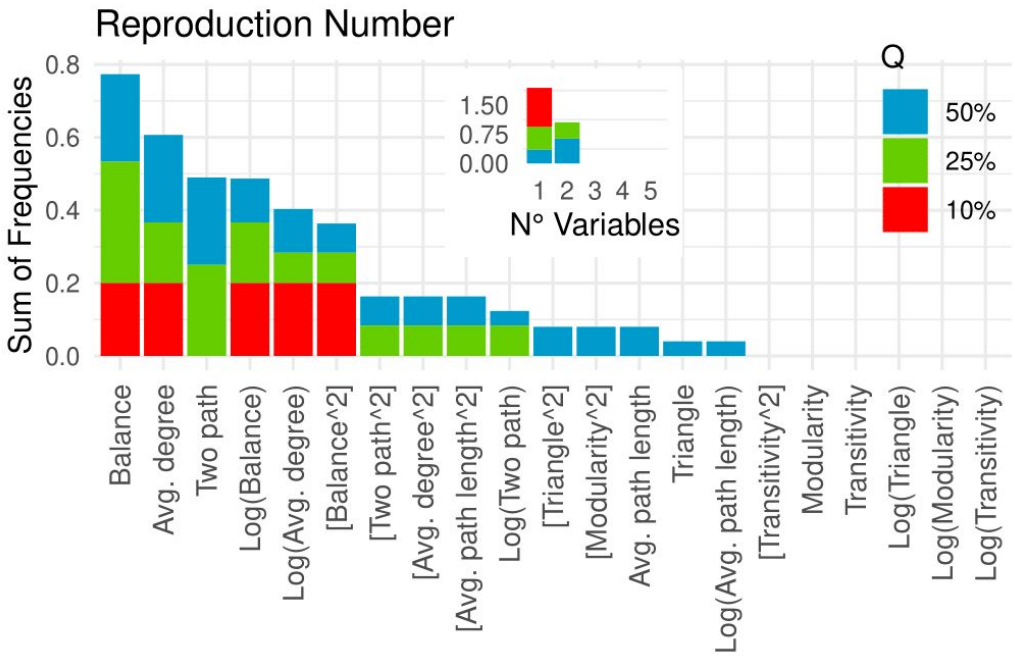
$$\text{Gen. Time} = \text{Net-type} + \beta_0 \text{ balance} + \beta_2 \left[\text{modularity}^2 \right]$$

Results - Reproductive Number



- The best models have only **1** *variable*
- That variable has to be **Balance** or **Avg. degree** (in his logarithmic or quadratic form)

Results - Reproductive Number



$$R_0 = \text{Net-type} + \beta_1 \log(\text{ balance })$$

Results

$$R_0 = \text{Net-type} + \beta_1 \log(\text{ balance })$$

$$\begin{aligned}\text{Gen. Time} = \text{Net-type} &+ \beta_0 \text{ balance} \\ &+ \beta_2 \left[\text{ modularity}^2 \right]\end{aligned}$$

$$\begin{aligned}\text{Peak}_T = \text{Net-type} &+ \beta_0 \text{ balance} \\ &+ \beta_1 [\text{ avg. path length}^2]\end{aligned}$$

$$\begin{aligned}\text{Peak}_P = \text{Net-type} &+ \beta_0 \text{ two path} \\ &+ \beta_1 \text{ avg. degree} \\ &+ \beta_2 \left[\text{ avg. path length}^2 \right]\end{aligned}$$

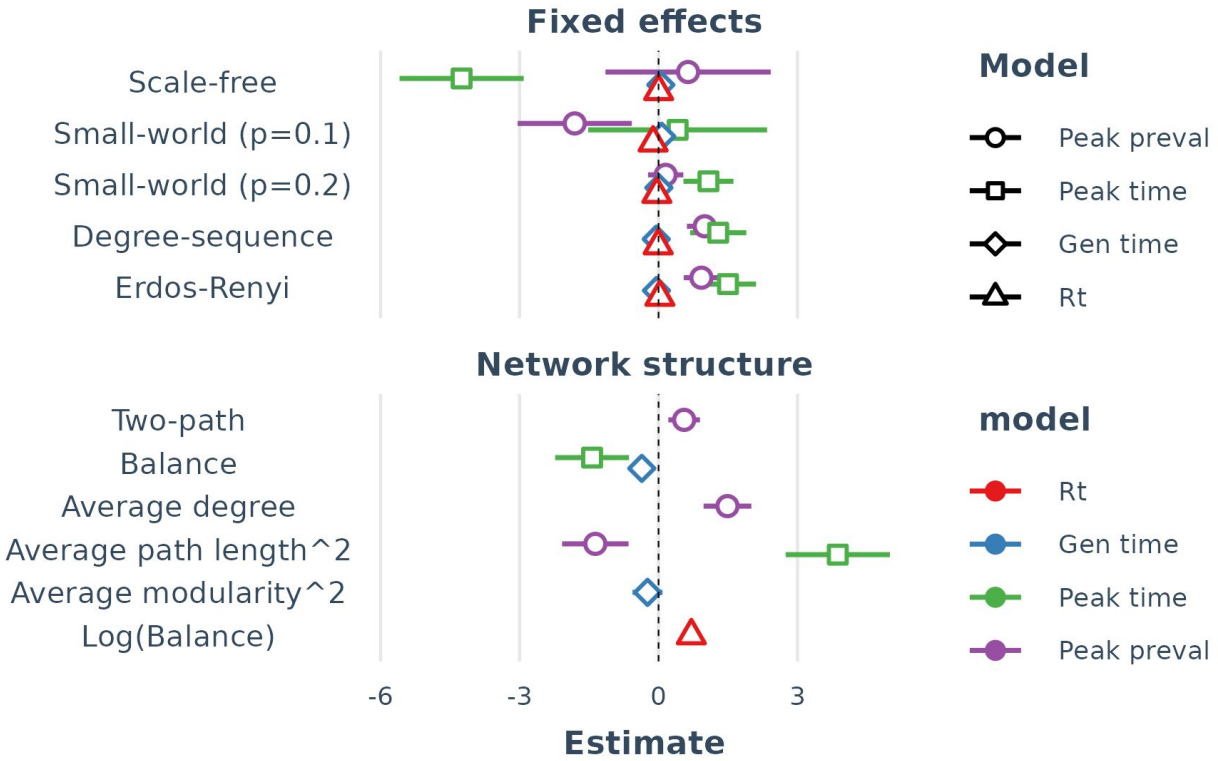
Results

$$R_0 = \text{Net-type} + \beta_1 \log(\text{ balance })$$

$$\text{Gen. Time} = \text{Net-type} + \beta_0 \text{ balance} + \beta_2 [\text{ modularity}^2]$$

$$\text{Peak}_T = \text{Net-type} + \beta_0 \text{ balance} + \beta_1 [\text{ avg. path length}^2]$$

$$\text{Peak}_P = \text{Net-type} + \beta_0 \text{ two path} + \beta_1 \text{ avg. degree} + \beta_2 [\text{ avg. path length}^2]$$



Conclusions

Conclusions

- Our recommended models:

  1. Consists of **few variables** → between 1 - 3 variables

Conclusions

- Our recommended models:

  1. Consists of few variables $\rightarrow$ between 1 - 3 variables

  2. Have **non-trivial forms**

Conclusions

- Our recommended models:

  1. Consists of few variables $\rightarrow$ between 1 - 3 variables

  2. Have non-trivial forms

  3. Relevant role of **Balance** and **Avg. path length**

Conclusions

- Our recommended models:

  1. Consists of few variables $\rightarrow$ between 1 - 3 variables

  2. Have non-trivial forms

  3. Relevant role of Balance and Avg. path length

- Add more variables to **refine the work** $\rightarrow$ $\langle k^2 \rangle$

Thank you!

Reference

[1] C.-H. Li and A. M. Yousef, Bifurcation analysis of a network-based sir epidemic model with saturated treatment function, Chaos: An Interdisciplinary Journal of Nonlinear Science 29, 033129 (2019).

[2] J. Zhou, Y. Zhao, and Y. Ye, Complex dynamics and control strategies of seir heterogeneous network model with saturated treatment 10.2139/ssrn.4151291 (2022).

[4] Derek Meyer and George Vega Yon. "epiworldR: Fast Agent-Based Epi Models". In: The Journal of Open Source Software 8.90 (Oct. 2023). DOI: 10.21105/joss.05781. URL: https://joss.theoj.org/papers/10.21105/joss.05781.

[3] To calculate the transmission rate, we use the following formula: $\beta = \frac{\gamma}{(C/R + \gamma - 1)}$, with contact rate $C = 14$, reproductive $R = 2$, and $\gamma = 1/7$.

	Peak preval	Peak time	Gen time	Rt
Scale-free	0.638	-4.251***	0.050***	0.003
	(0.910)	(0.684)	(0.021)	(0.018)
Small-world (p=0.1)	-1.809***	0.413	0.069***	-0.117***
	(0.628)	(0.985)	(0.028)	(0.017)
Small-world (p=0.2)	0.154	1.078***	0.008	-0.036***
	(0.195)	(0.276)	(0.012)	(0.017)
Degree-sequence	0.997***	1.288***	-0.049***	-0.007
	(0.196)	(0.309)	(0.020)	(0.015)
Erdos-Renyi	0.925***	1.500***	-0.045***	0.029**
	(0.195)	(0.310)	(0.020)	(0.015)
Two-path	0.553***			
	(0.174)			
Balance		-1.434***	-0.361***	
		(0.407)	(0.006)	
Average degree	1.488***			
	(0.263)			
Average path length^2	-1.364***	3.870***		
	(0.366)	(0.574)		
Average modularity^2			-0.238*	
			(0.165)	
Log(Balance)				0.708***
				(0.097)
Num.Obs.	18011	18011	18009	18011
AIC	97940.8	114131.5	1600.7	77572.4
BIC	98018.8	114201.7	1670.9	77634.8

* p < 0.15, ** p < 0.1, *** p < 0.05