Stockholms universitet

L7, Networks (cont'd) and inference from big outbreaks

Tom Britton

July, 2023

Stockholms universitet

A model for an STI in a heterosexual community

The model (Britton, Nordvik and Liljeros, 2007)

- $D = \#$ sex-partners (e.g. during a year)
- $p = P(\text{transmission in a relationship})$

Stockholms universitet

A model for an STI in a heterosexual community

The model (Britton, Nordvik and Liljeros, 2007)

- $D = \#$ sex-partners (e.g. during a year)
- $p = P(\text{transmission in a relationship})$
- Heterosexual community: D_f, D_m, p_f, p_m

Stockholms universitet

A model for an STI in a heterosexual community

The model (Britton, Nordvik and Liljeros, 2007)

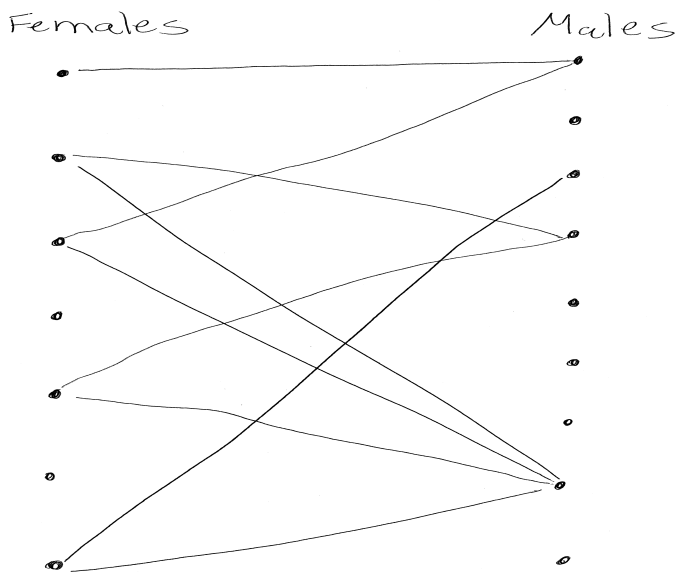
- $D = \#$ sex-partners (e.g. during a year)
- $p = P(\text{transmission in a relationship})$
- Heterosexual community: D_f, D_m, p_f, p_m
  $\Longrightarrow$ bipartite graph

Stockholms universitet

A model for an STI in a heterosexual community

The model (Britton, Nordvik and Liljeros, 2007)

- $D = \#$ sex-partners (e.g. during a year)
- $p = P(\text{transmission in a relationship})$
- Heterosexual community: D_f, D_m, p_f, p_m
  $\Longrightarrow$ bipartite graph



It can be shown that

$$R_0 = \sqrt{p_f \left(E(D_f) + \frac{V(D_f) - E(D_f)}{E(D_f)} \right)} \times \sqrt{p_m \left(E(D_m) + \frac{V(D_m) - E(D_m)}{E(D_m)} \right)}$$

Similar to before:

A heavy-tailed degree distribution makes R_0 large.

It can be shown that

$$R_0 = \sqrt{p_f \left(E(D_f) + \frac{V(D_f) - E(D_f)}{E(D_f)} \right)} \times \sqrt{p_m \left(E(D_m) + \frac{V(D_m) - E(D_m)}{E(D_m)} \right)}$$

Similar to before:

A heavy-tailed degree distribution makes R_0 large. $\Longrightarrow$

promiscuous people (super-spreaders) play an important role

Stockholms universitet

Improved analysis

However:

- P(transmission) depends on # sex-acts in relationship

Improved analysis

However:

- P(transmission) depends on # sex-acts in relationship
- Promiscuous individuals tend to have fewer sex-acts *per partner*

Improved analysis

However:

- P(transmission) depends on # sex-acts in relationship
- Promiscuous individuals tend to have fewer sex-acts *per partner*
- This should reduce R_0!

Improved analysis: continued

Extended model: short and long term relationships

Improved analysis: continued

Extended model: short and long term relationships

$\Longrightarrow$ two types of edges (with different trans prob)

New (complicated) expression for R_0

The effect of different transmission probabilities depends on calibration

Stockholms universitet

Calibration using survey on sexual habits

Data:

- (Anonymous) study of sexual habits in Gotland
- $\approx$ 800 people (17-28 yrs)
- Among other things: How many sex-partners during last year and how many sex-acts in each relationship

Stockholms universitet

Calibration using survey on sexual habits

Data:

- (Anonymous) study of sexual habits in Gotland
- $\approx$ 800 people (17-28 yrs)
- Among other things: How many sex-partners during last year and how many sex-acts in each relationship

$P(\text{transmission}|p)$ for short/long relationship estimated as cohort mean of:

$$P(\text{transmission}) = 1-(1-p)^{\#\text{ sex-acts}}, \quad p = \text{per sex-act trans prob}$$

Stockholms universitet

Calibration using survey on sexual habits

Data:

- (Anonymous) study of sexual habits in Gotland
- $\approx$ 800 people (17-28 yrs)
- Among other things: How many sex-partners during last year and how many sex-acts in each relationship

$P(\text{transmission}|p)$ for short/long relationship estimated as cohort mean of:

$$P(\text{transmission}) = 1-(1-p)^{\#\text{ sex-acts}}, \quad p = \text{per sex-act trans prob}$$

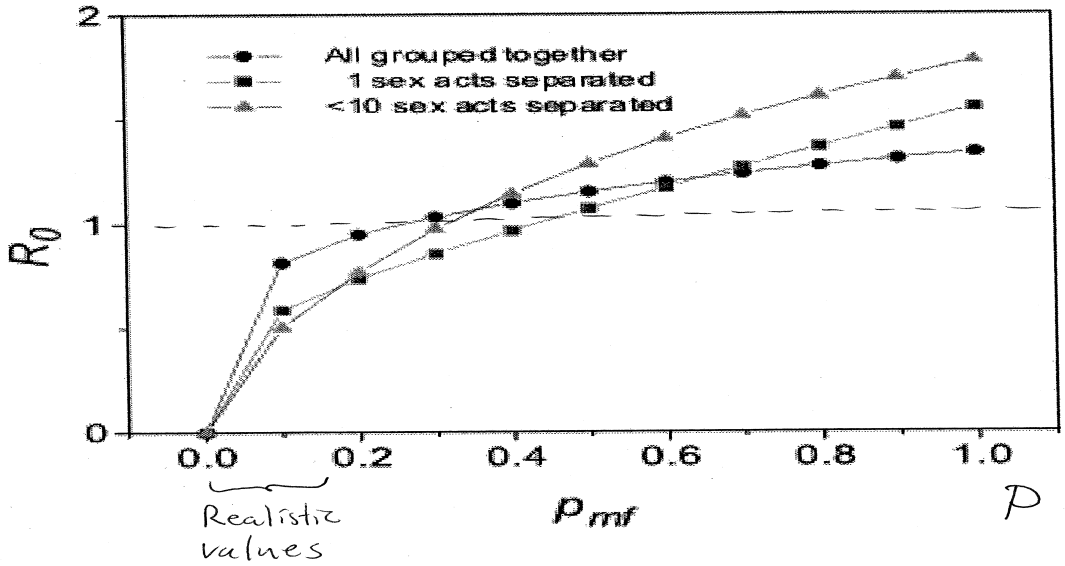
R_0 fitted to data and computed as a function of p: for one type of relationship, and two separations of short vs long

R_0 as function of p (fitted to Gotland data)

R_0 as function of $p = P(\text{transmission per contact})$

$P_{longterm}$, $P_{short\ term}$ } = mean values of $1-(1-p)^{\#\text{sex-acts}}$

Assumption $P_m = P_f = P$



Conclusions:

1. Heavy-tailed degree distribution (promiscuity) increases R_0
2. Acknowledging short and long-term relationships **reduces** this effect
3. Endemicity not possible (for realistic p's)

Conclusions:

1. Heavy-tailed degree distribution (promiscuity) increases R_0
2. Acknowledging short and long-term relationships **reduces** this effect
3. Endemicity not possible (for realistic p's) but maybe in sub-communities ...

Stockholms universitet

Homogeneous vs Heterogeneous: qualitative results

We now illustrate a general conclusion with an example (from the network model defined earlier)

Recall that $R_0 = p\left(E(D) + \frac{V(D)-E(D)}{E(D)}\right)$

Consider **two networks** with the same mean degree $E(D) = 4$

Network 1: $D \equiv 4$, so $V(D) = 0$ and $E(D) + \frac{V(D)-E(D)}{E(D)} = 3$

Network 2: $P(D = 1) = P(D = 7) = 0.5$, so $V(D) = 9$ and $E(D) + \frac{V(D)-E(D)}{E(D)} = 5.25$

Infectious Disease 1: $p = 0.25$

Network 1: $R_0 = 3/4 = 0.75$, Network 2: $R_0 = 5.25/4 = 1.31$

$\Longrightarrow$ R_0 larger for Network 2. Outbreak not possible in Network 1 but possible for Network 2

Stockholms universitet

Homogeneous vs Heterogeneous: qualitative results, cont'd

Infectious Disease 2: p=0.75

Network 1: $R_0 = 3 \cdot 0.75 = 2.25$,
Network 2: $R_0 = 5.25 \cdot 0.75 = 3.93$

$\Longrightarrow$ R_0 larger for Network 2. Outbreak possible in both networks

Which outbreak will be bigger?

Stockholms universitet

Homogeneous vs Heterogeneous: qualitative results, cont'd

Infectious Disease 2: p=0.75

Network 1: $R_0 = 3 \cdot 0.75 = 2.25$,
Network 2: $R_0 = 5.25 \cdot 0.75 = 3.93$

$\Longrightarrow$ R_0 larger for Network 2. Outbreak possible in both networks

Which outbreak will be bigger? Outbreak in Network 1 since in Network 2 individuals with degree 1 have a good chance of escaping!

General conclusion. (Starting with a homogeneous situation):
– Heterogenizing *always* increases R_0

– If original (=homogeneous case) R_0 is small, then outbreak will be bigger in heterogeneous case

– But if original R_0 is large, then heterogenizing makes outbreak **smaller**!!!

Stockholms universitet

Repetition: Inference from large outbreaks

From lecture 3: basic reproduction number R_0 and critical vaccination coverage v_c were estimated by:

$$\hat{R}_0 = -\ln(1-\tilde{\tau})/\tilde{\tau}$$

$$\hat{v}_c = 1 - \frac{\tilde{\tau}}{-\ln(1-\tilde{\tau})}$$

if outbreak takes place in a fully susceptible homogeneous community resulting in a fraction $\tilde{\tau}$ getting infected during the outbreak

How about uncertainty?

Uncertainty of previous estimate

Intuition: The larger community (and more getting infected) the less uncertainty

It was mentioned that final number infected $n\tilde{\tau} = Z$ in case of a major outbreak is normally distributed with mean $n\tau^*$ and standard deviation $\sqrt{n\sigma^2}$ where σ^2 depends on model parameters and shown two slides ahead

This result can be used to show that $\hat{R}_0$ and $\hat{v}_c$ are normally distributed with correct means (i.e. R_0 and v_c respectively) and standard errors to be derived using δ-method

Stockholms universitet

The δ-method

Suppose random variable X has mean $\mu = E(X)$ and variance $V(X)$. Suppose further that we are mainly interested in the distribution of $f(X)$ for some function $f(\cdot)$ rather than X itself

Then the δ-method gives the following approximation for the mean and variance of $f(X)$, where $f(x)$ is a "nice function":

Main idea Taylor expand X around its mean μ:
$f(X) \approx f(\mu) + (X - \mu)f'(\mu)$. This implies:

$$E(f(X)) \approx f(\mu) \qquad V(f(X)) \approx (f'(\mu))^2 V(X).$$

The approximation holds better the smaller variance X has (i.e. smaller $V(X)$).

We will use it for e.g. $f(X) = -\ln(1 - X)/X$ and with $X = \tilde{\tau}$ so that $f(\tilde{\tau}) = \hat{R}_0$

Stockholms universitet

The δ-method for $V(\hat{R}_0)$

Probabilists have proven that the asymptotic variance of $\tilde{\tau}$ equals:

$$V(\tilde{\tau}) \approx \frac{1}{n} \frac{\tau(1-\tau)}{(1-(1-\tau)R_0)^2} \left(1 + c_v^2(1-\tau)R_0^2\right)$$

where τ and R_0 are the true parameter values related by $R_0 = -\ln(1-\tau)/\tau$, and c_v is the coefficient of variation of the infectious period.

We now apply the δ-method on $\hat{R}_0 = -\ln(1-\tilde{\tau})/\tilde{\tau}$, we hence have the function $f(x) = -\ln(1-x)/x$

After some algebra we get $V(\hat{R}_0) \approx \frac{1}{n\tau(1-\tau)} \left(1 + c_v^2(1-\tau)R_0^2\right)$

For a standard error estimate we take square roots and replace unknown quantities with there estimates/observed values. The result, also for $\hat{v}_c$, is given by:

Stockholms universitet

Uncertainty of previous estimate

$$s.e.(\hat{R}_0) = \sqrt{\frac{1 + c_v^2(1-\tilde{\tau})\hat{R}_0^2}{\tilde{\tau}(1-\tilde{\tau})}/n}$$

$$s.e.(\hat{v}_c) = \sqrt{\frac{1 + c_v^2(1-\tilde{\tau})\hat{R}_0^2}{\hat{R}_0^4\tilde{\tau}(1-\tilde{\tau})}/n}$$

$c_v^2 = V(I)/(E(I))^2$= squared coefficient of variation of infectious period of individuals (variance divided by the squared mean)

Larger n gives smaller standard deviation (as expected)!

Stockholms universitet

Uncertainty of previous estimate

c_v^2 cannot be estimated from final outbreak size – possibly known from before

If not one has to insert a "conservative" bound. E.g. $c_v^2 = 1$: very rarely is standard deviation larger than mean

Exercise 25 Suppose that 239 out of 651 individuals in an isolated village were infected during an outbreak. Estimate R_0 and v_c and give 95% confidence interval for the estimates. Consider both the case when all individuals have the same length of infectious period (so no variation) and the case where its standard deviation is equal to the mean.

Exercise 26 Do the same thing assuming 2390 out of 6510 got infected.

Stockholms universitet

More detailed data

Suppose that disease incidence is observed during outbreak – not only final number

Intuition: more detailed data should improve estimation

Answer: yes, in a couple of ways:

- estimate of R_0 and v_c becomes more complicated, but standard errors are (moderately) smaller
- enables estimation of more parameters: exponential growth rate ρ, latent and infectious period distributions, ...
- possible to detect deviations from model: changing behavior, non-homogeneity, ...

If also information about contacts are available: "transmission probability upon contact" can be estimated

Multitype epidemics

Suppose final size of a multitype epidemic observed: $\tilde{\tau}_1, \ldots, \tilde{\tau}_k$, $\tilde{\tau}_i =$ observed proportion infected among i-types

Also assumed that community fractions $\pi_1, \ldots, \pi_k$ known.

We want to estimate R_0 which is largest eigenvalue of next generation matrix M

First estimate M. Impossible!! Data has dimension k and M has dimension k^2.

$\Longrightarrow$ M and R_0 cannot be estimated consistently!

Stockholms universitet

Multitype epidemics, cont'd

Why? We can observe who was infected but not who "caused" the infections

Susceptibility easier to estimate than infectivity!

$\Longrightarrow$ only possible to obtain bounds on R_0: lower bound assuming all infections caused by least infected type – upper bound assuming all infections caused by most infected type

Stockholms universitet

Inference in networks

Inference can be performed without an outbreak: estimation of network properties: $E(D)$, $V(D)$, clustering c, ...

R_0, potential outbreak size τ and v_c can then be estimated as a function of transmission probability p

Typical conclusion: Outbreaks are only possible for a disease having higher transmission probability than $p = 0.13$

Or: An STD with $p = 0.08$ can only become endemic in core-groups with average number of partners higher than $E(D) = 4.2$ per year

Inference in more complicated models

More complicated model $\Longrightarrow$ harder inference and more detailed data need

Inference of spread of infections extra hard:

- There are strong dependencies because infections are not independent events (likelihood complicated)
- Many things unobserved: infectious contacts, latent period, infectious period, ...

Inference with more detailed data gives higher precision

Stockholms universitet

Illustration

Suppose an infected infects each susceptible independently with prob p

Data = epidemic chain: $1 \to 2 \to 2 \to 0$

Initially 1 index and 9 susceptible

Likelihood: $L(p) =$
$\binom{9}{2} p^2 (1-p)^7 \cdot \binom{7}{2} \left(1-(1-p)^2\right)^2 \left((1-p)^2\right)^5 \cdot \binom{5}{0} \left((1-p)^2\right)^5$

Maximum-likelihood (ML) estimate $\hat{p}$ maximizes $L(\cdot)$:
$\Longrightarrow$ quite easy for a computer

If we instead only know that 5 out of 10 were infected likelihood is much more complicated (a sum over all possible chains)

Stockholms universitet

Alternative approach for complicated models

Basic idea: If likelihood complicated for available data we can "pretend" as if we had more detailed data, estimate parameters under this assumption, recompute some likely more detailed data, re-estimate parameters, ...

This is underlying idea in both EM-algorithm and recently very popular *MCMC*

MCMC: here parameters are treated as outcomes of random variables (Bayesian framework) and even very complicated likelihoods (posterior probabilities) can be evaluated numerically with arbitrary high precision

MCMC: Very computer intensive. Treated specifically in other Modules