Statistical estimation of network models from egocentrically sampled network data

Jeanette Birnbaum

Center for AIDS Research

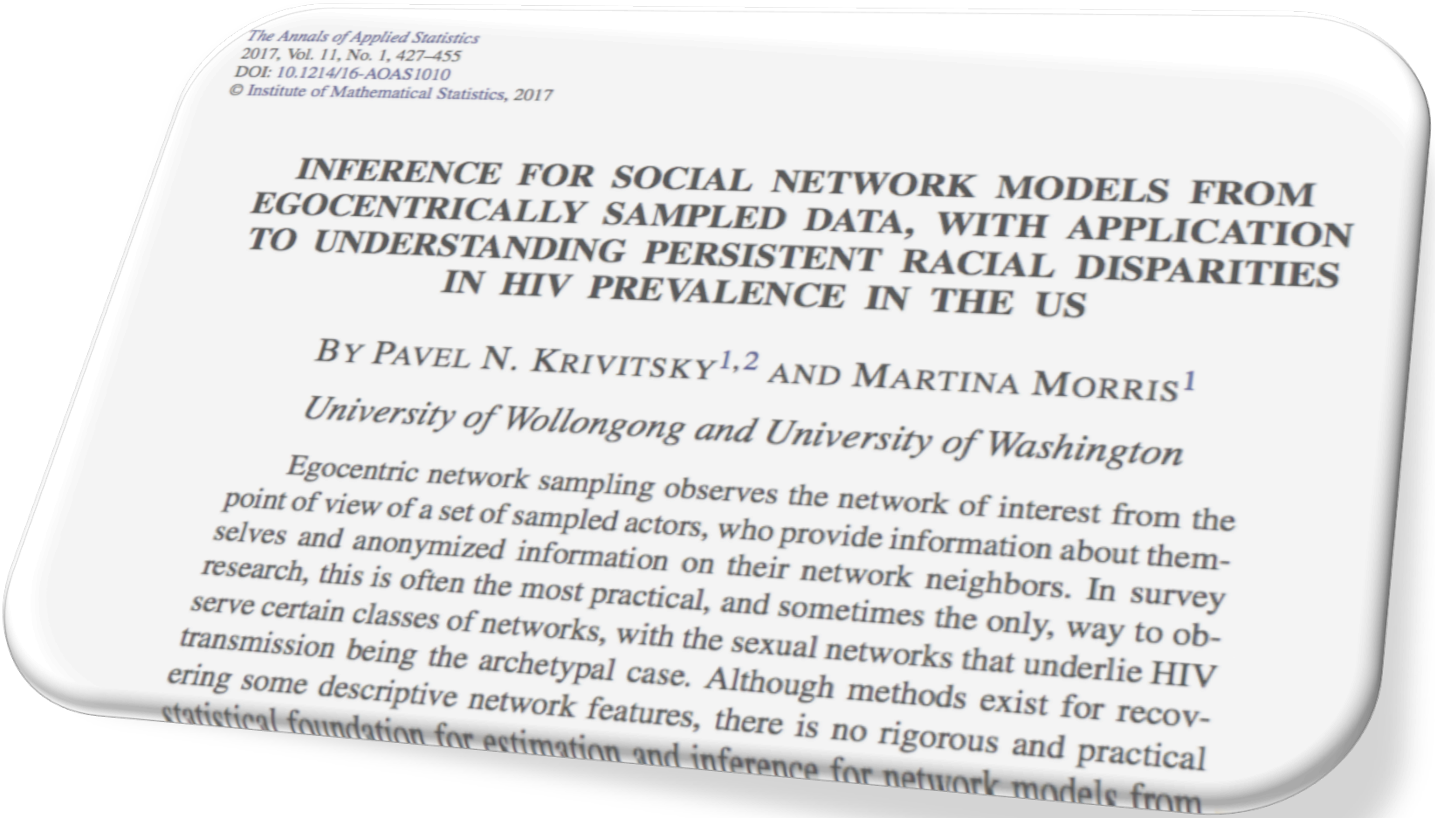
University of Washington

Recent Advances in Statistical Network Analysis

2019 Symposium on Data Science and Statistics

Pavel Krivitsky
Department of Statistics
University of Wollongong

Martina Morris
Departments of Statistics & Sociology
University of Washington



The Annals of Applied Statistics
2017, Vol. 11, No. 1, 427–455
DOI: 10.1214/16-AOAS1010

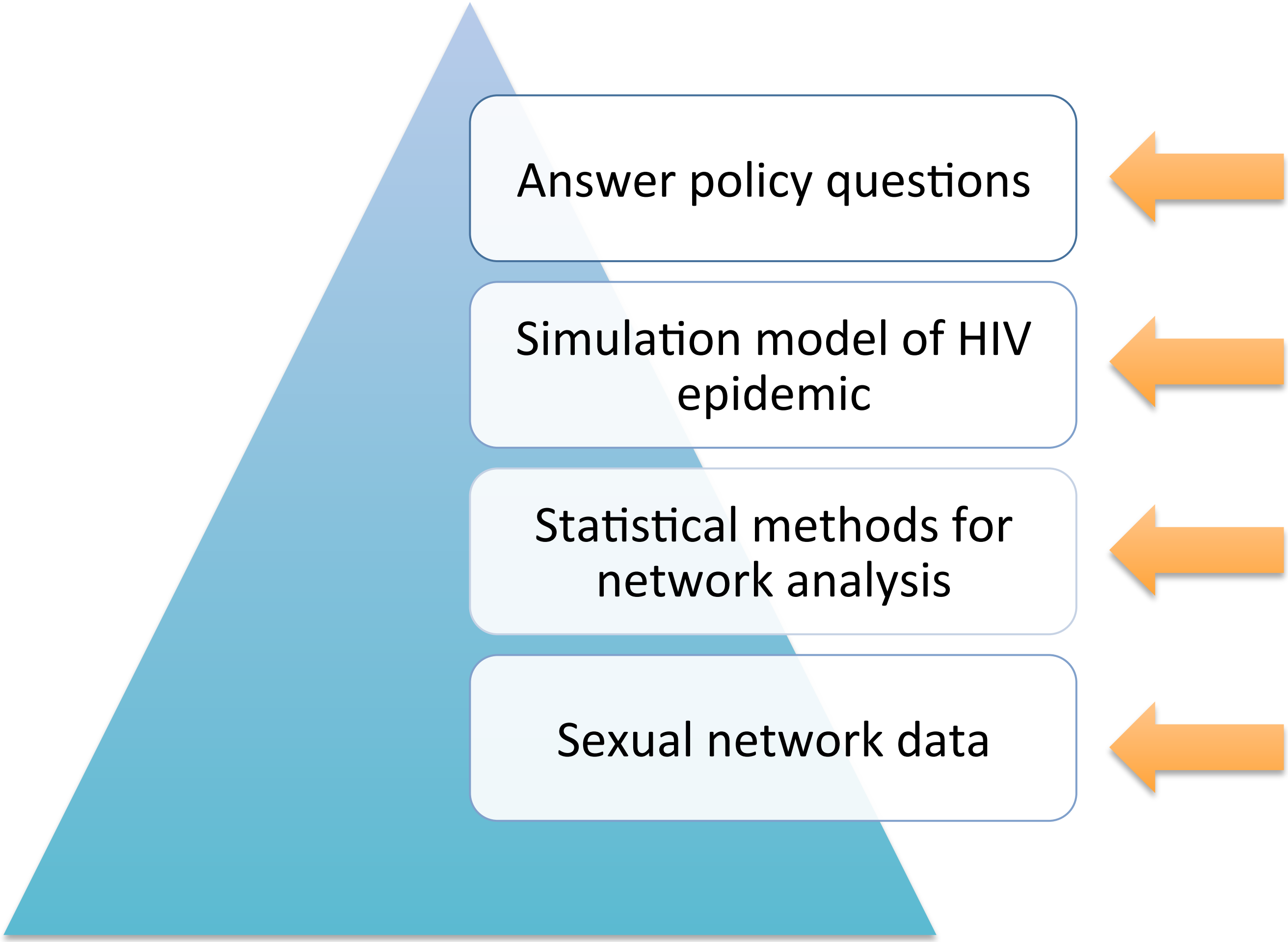
INFERENCE FOR SOCIAL NETWORK MODELS FROM EGOCENTRICALLY SAMPLED DATA, WITH APPLICATION TO UNDERSTANDING PERSISTENT RACIAL DISPARITIES IN HIV PREVALENCE IN THE US

BY PAVEL N. KRIVITSKY[1,2] AND MARTINA MORRIS[1]

University of Wollongong and University of Washington

Egocentric network sampling observes the network of interest from the point of view of a set of sampled actors, who provide information about themselves and anonymized information on their network neighbors. In survey research, this is often the most practical, and sometimes the only, way to observe certain classes of networks, with the sexual networks that underlie HIV transmission being the archetypal case. Although methods exist for recovering some descriptive network features, there is no rigorous and practical statistical foundation for estimation and inference for network models from

Expanding the scope of network analysis for HIV epidemic modeling



Answer policy questions

Simulation model of HIV epidemic

Statistical methods for network analysis

Sexual network data

Outline

Network analysis in HIV

- *Sampled* sexual network data for HIV modeling

Statistical methods for analyzing network data

- Exponential-family random graph models (ERGMs)
- Extension to sampled data
- Integration with epidemic modeling

Application

- Heterosexual HIV dynamics in Seattle/King County

Network analysis in HIV

Network structure matters to HIV

Central to population-level disease transmission, when contacts are

- Rare
- Systematically heterogeneous in probability

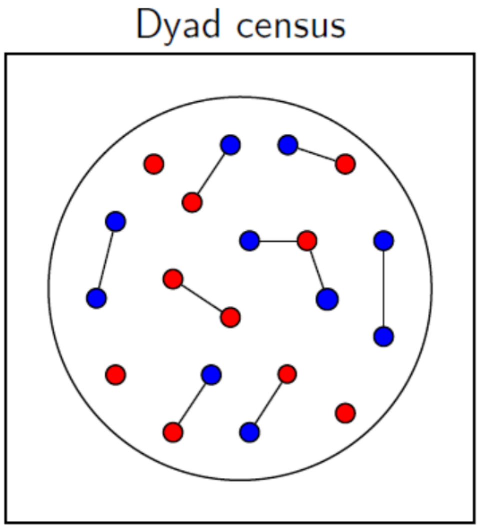
Empirical network research = methods + data

Substantial progress in statistical methods in last 20 years…

- Exponential-family random graph models (ERGMs)

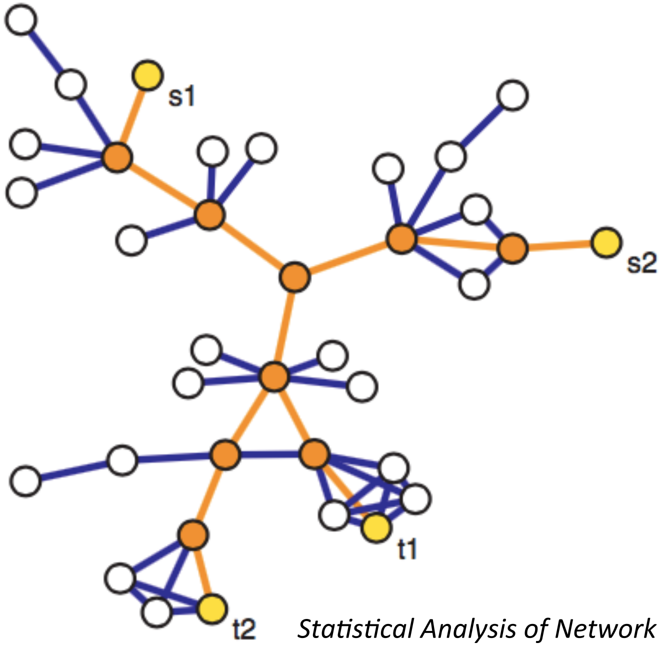
…based on network "census" data

- Often impractical
- e.g. population-level sexual networks



Two types of network sampling

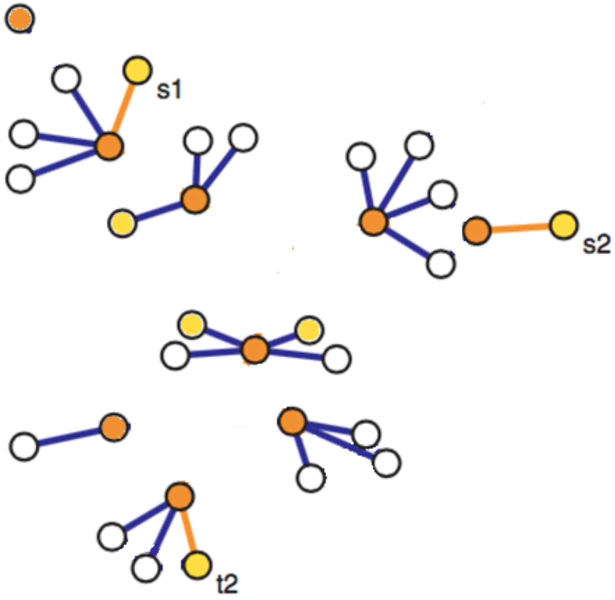
Adaptive (link trace)



Statistical Analysis of Network Data
Kolaczyk 2009

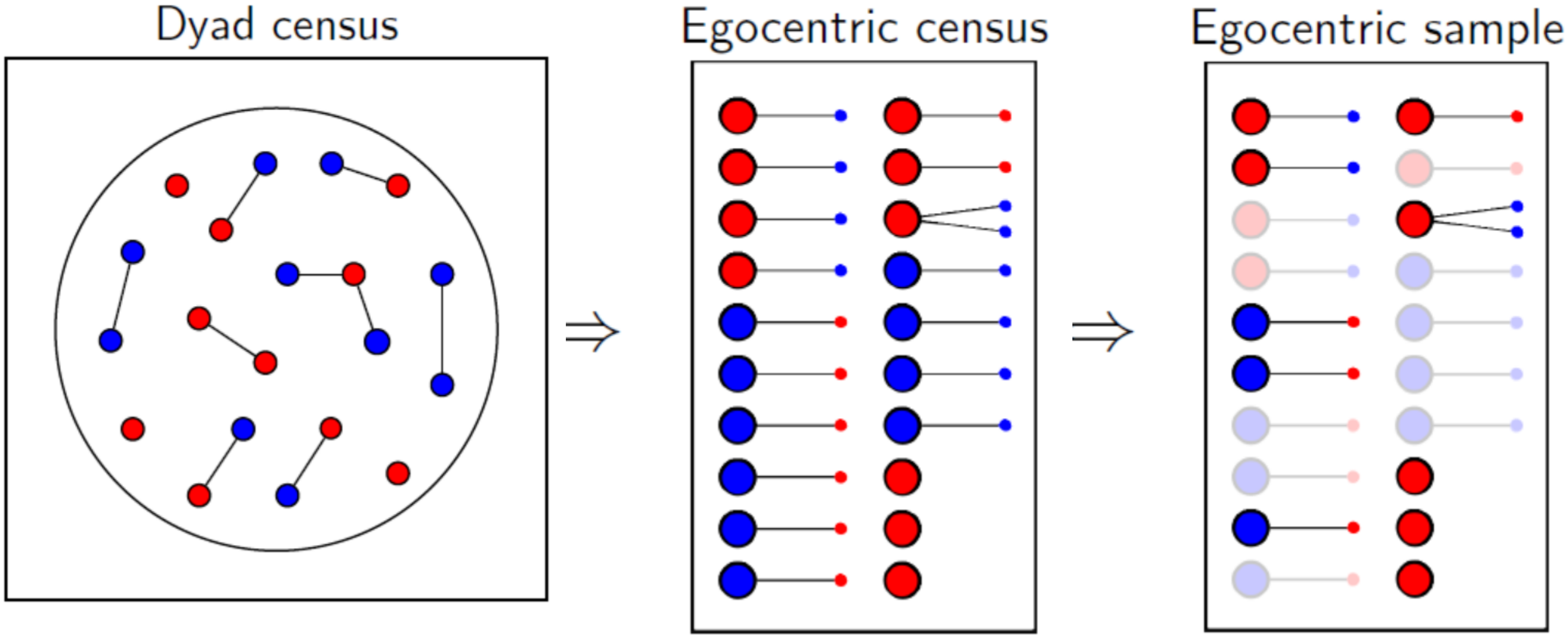
Multiple waves

Egocentric



Standard sample surveys

Egocentric sampling



Observe the
complete network

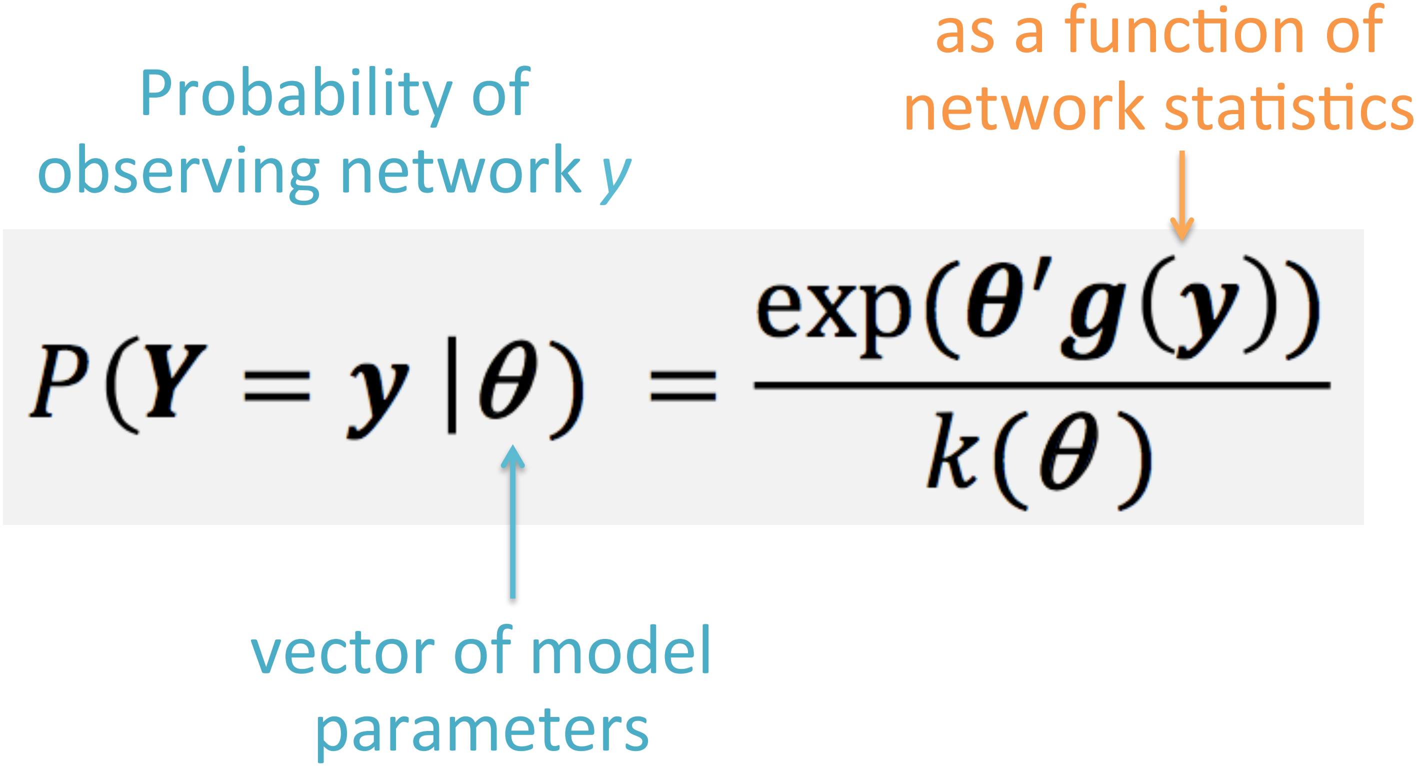
Observe all egos +
Reported info on alters

Sample egos +
Reported info on alters

Statistical methods for analyzing network data

ERGMs

Foundation: exponential-family random graph models (ERGMs)



Probability of observing network *y*

as a function of network statistics

$$P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}' \boldsymbol{g}(\boldsymbol{y}))}{k(\boldsymbol{\theta})}$$

vector of model parameters

Re-expressed at the dyad level

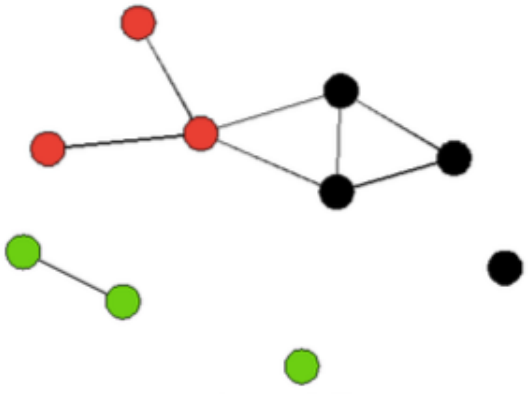
$$logit(P(Y_{ij} = 1 \,|Y_{ij}^{c})) \;=\; log\left[\frac{P(Y_{ij} = 1 \,|Y_{ij}^{c})}{P(Y_{ij} = 0 \,|Y_{ij}^{c})}\right]$$

$$= \theta' \underset{\substack{\uparrow \\ \text{“change statistic”}}}{\delta}(\mathrm{g}(\mathrm{y}))$$

θ is the per-unit change in the log odds of a tie

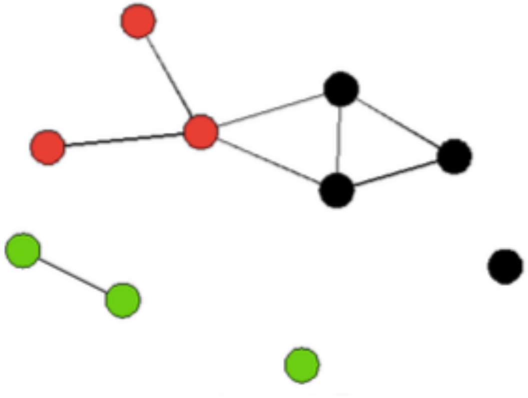
Common network statistics $\boldsymbol{g(y)}$

Edges	7 edges
Nodefactor/ nodecov	Red degree = 1 + 1 + 4
Homophily	Red-Red edges = 2
k-stars/degree(k)	Degree(0) = 2
3-cycles	# triangles = 2



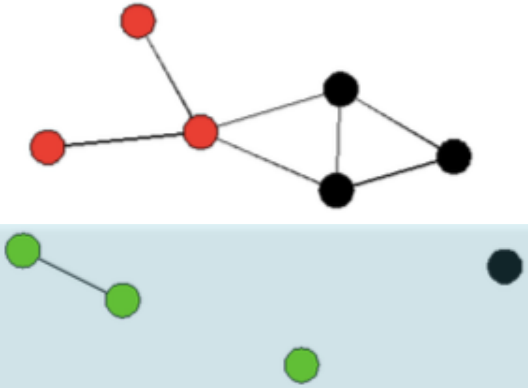
Common network statistics $g(y)$

	Dyad independent	Dyad dependent
Edges	X	
Nodefactor/ nodecov	X	
Homophily	X	
k-stars/degree(k)		X
3-cycles		X



Common network statistics $g(y)$

	Dyad independent	Dyad dependent
Edges	X	
Nodefactor/ nodecov	X	
Homophily	X	
k-stars/degree(k)		X
3-cycles		X



Differential network exposure (components >2) + homophily → HIV disparities

ERGM estimation

MCMC Maximum Likelihood Estimation

- For dyad dependent models
- Uses MPLE as initial starting value

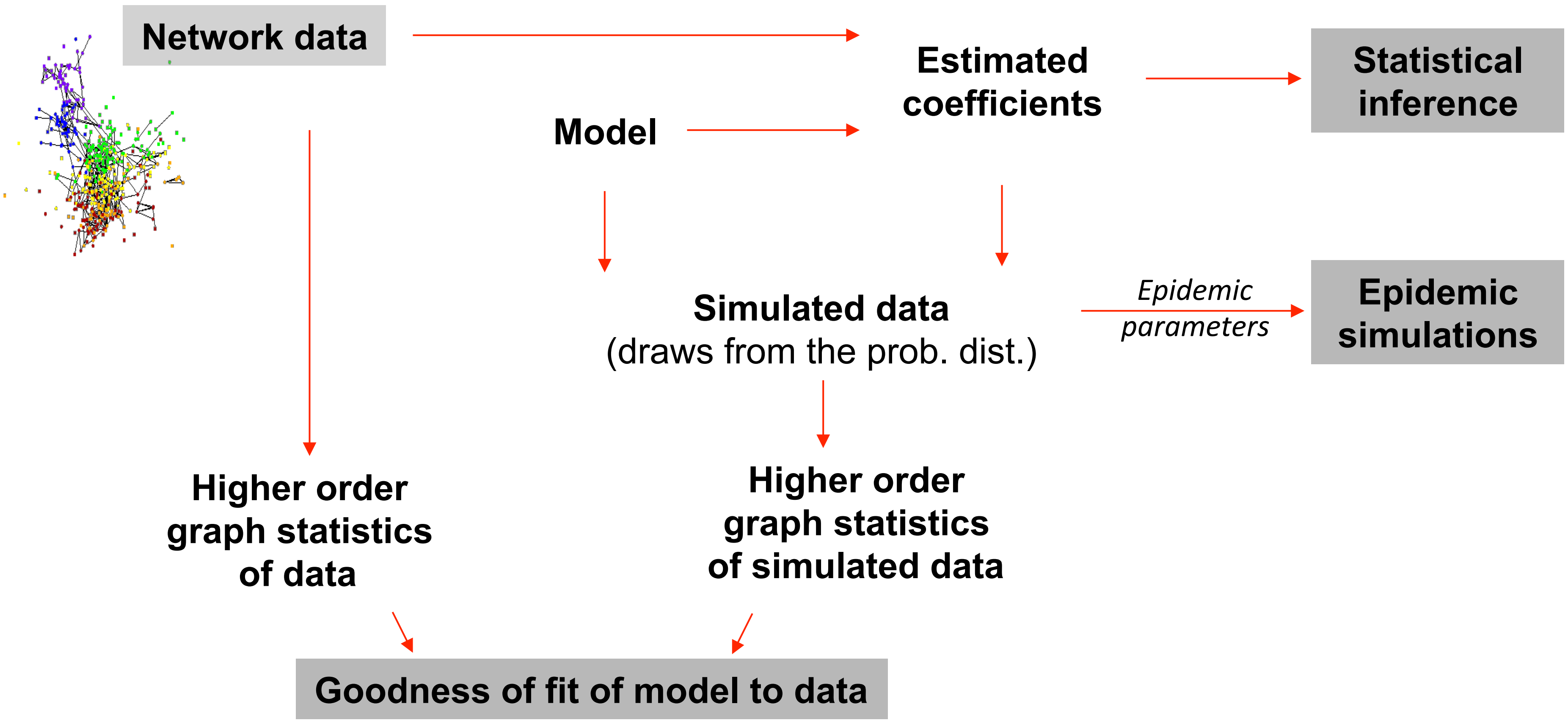
Metropolis Hastings algorithm

- Sampling from the distribution of networks at each iteration
- Given candidate θ_i

Effectively: a network simulation algorithm*

- Proposing toggles, one dyad at a time
- And selecting networks from the chain at suitable intervals

Simulating from a fitted model



Network data

Model

Estimated coefficients

Statistical inference

Simulated data
(draws from the prob. dist.)

Epidemic parameters

Epidemic simulations

Higher order graph statistics of data

Higher order graph statistics of simulated data

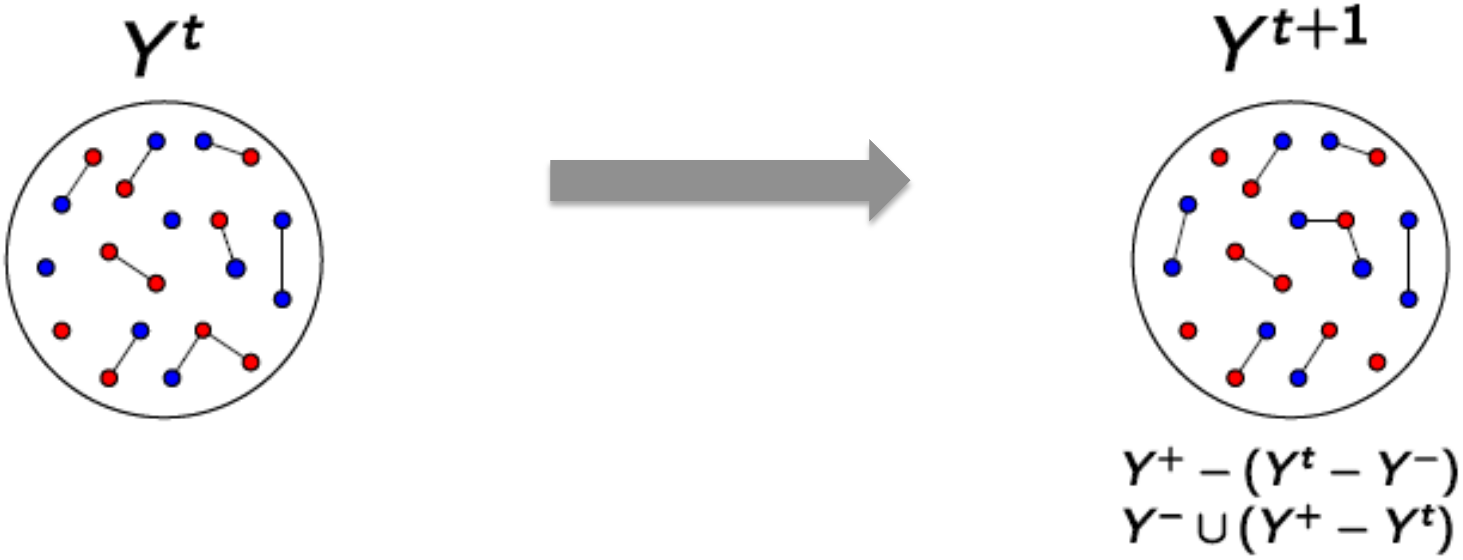
Goodness of fit of model to data

Steve Goodreau
Network Modeling for Epidemics 2014

Temporal ERGMs

Model link formation and dissolution over time

Krivitsky & Handcock 2014, "A Separable Model for Dynamic Networks"

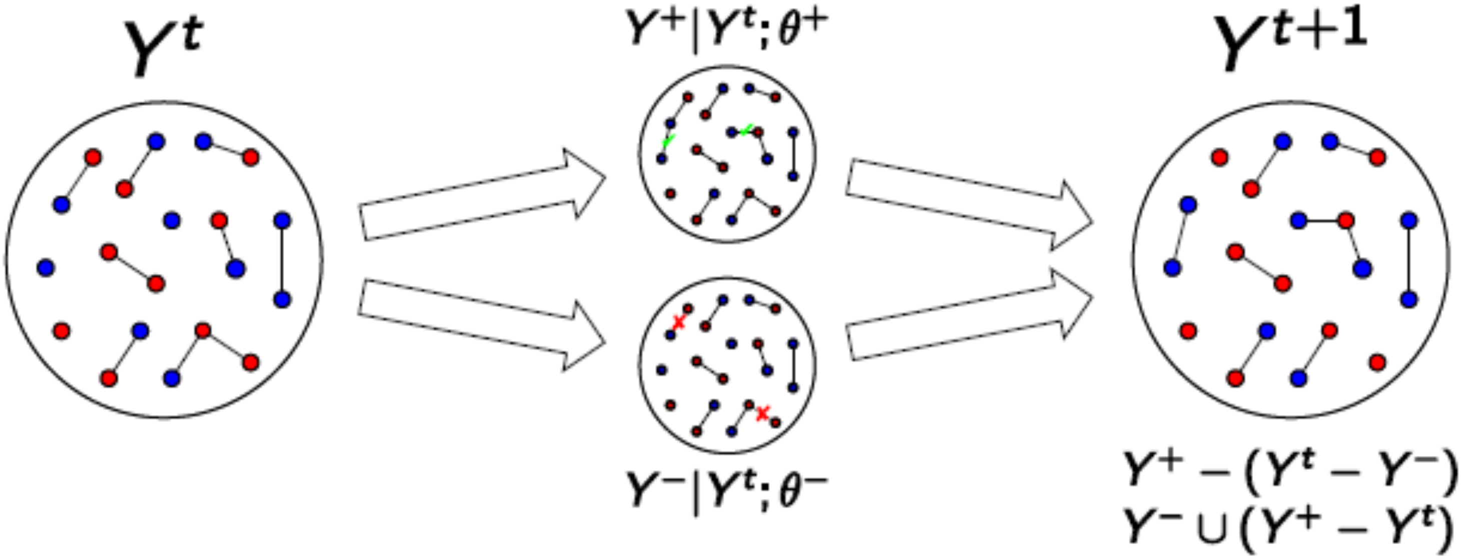


Temporal ERGMs

STERGMs = Separable Temporal ERGMs

- Independent within a time step
- Markov dependent between time steps

a formation ERGM

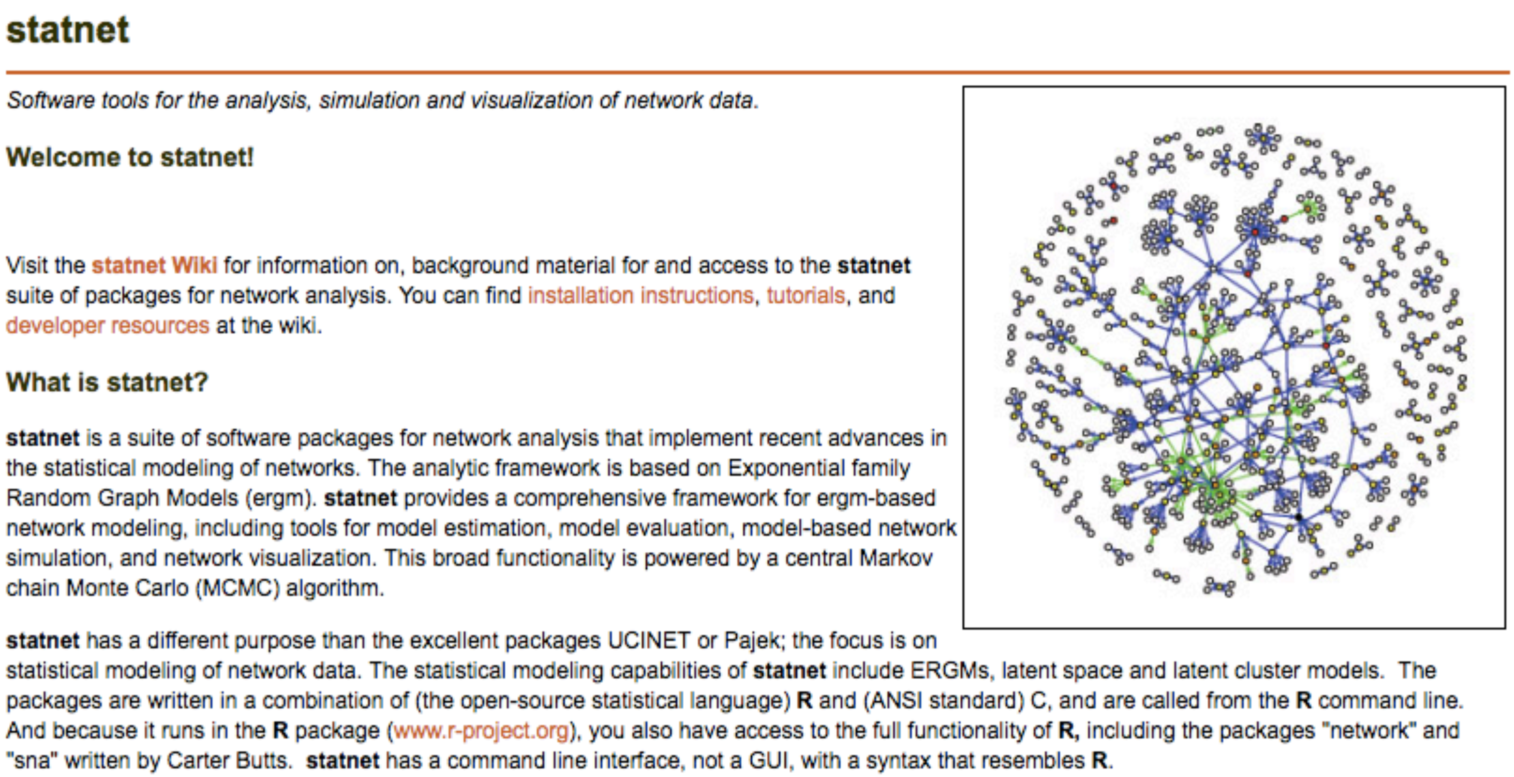


and a dissolution ERGM

In R: *statnet* suite

ergm and *tergm*

www.statnet.org



statnet

Software tools for the analysis, simulation and visualization of network data.

Welcome to statnet!

Visit the **statnet Wiki** for information on, background material for and access to the **statnet** suite of packages for network analysis. You can find installation instructions, tutorials, and developer resources at the wiki.

What is statnet?

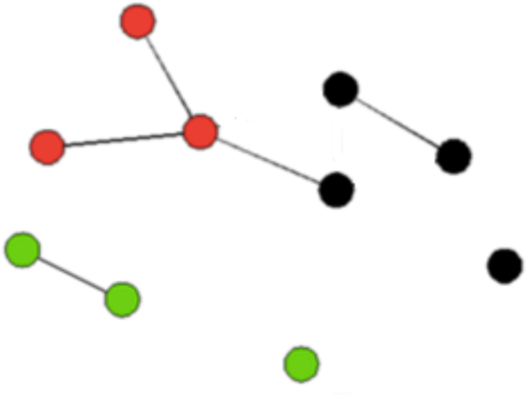
statnet is a suite of software packages for network analysis that implement recent advances in the statistical modeling of networks. The analytic framework is based on Exponential family Random Graph Models (ergm). **statnet** provides a comprehensive framework for ergm-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm.

statnet has a different purpose than the excellent packages UCINET or Pajek; the focus is on statistical modeling of network data. The statistical modeling capabilities of **statnet** include ERGMs, latent space and latent cluster models. The packages are written in a combination of (the open-source statistical language) **R** and (ANSI standard) C, and are called from the **R** command line. And because it runs in the **R** package (www.r-project.org), you also have access to the full functionality of **R,** including the packages "network" and "sna" written by Carter Butts. **statnet** has a command line interface, not a GUI, with a syntax that resembles **R**.

Statistical methods for analyzing network data

Extension to sampled data

Egocentric design determines observable sample statistics



Observable in egocentric sample

	Dyad independent	Dyad dependent
Edges	X	
Nodefactor/ nodecov	X	
Homophily	X	
k-stars/degree(k)		X
~~3-cycles~~		~~X~~

Key ideas for egocentric estimation

Krivitsky & Morris 2017

1. Use sample statistics to estimate population statistics g(y)

   Requires a scaling assumption

   - Assume mean degree is the scale invariant property
   - Use inverse probability weighted *Hájek estimator* of g(y)
   - These are the sufficient statistics for estimating the ERGM

2. Use estimated population statistics to estimate PMLE of θ (Binder 1983)

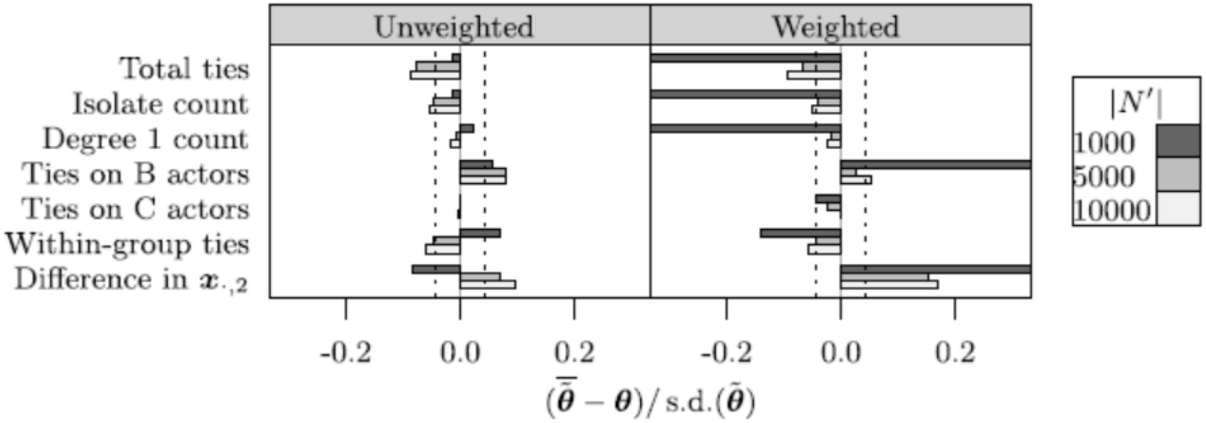
   - And variance of this estimate

Simulation studies indicate good properties

Krivitsky & Morris 2017

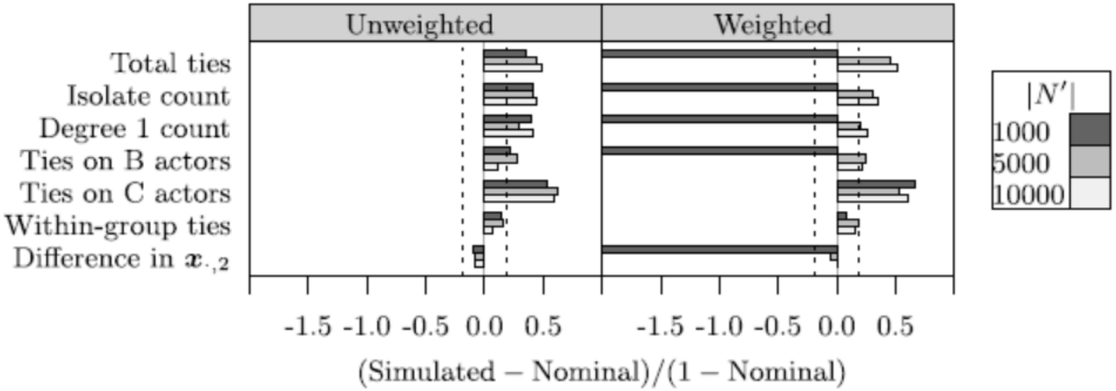
Bias

- Sampling weights require larger N to minimize bias



Coverage

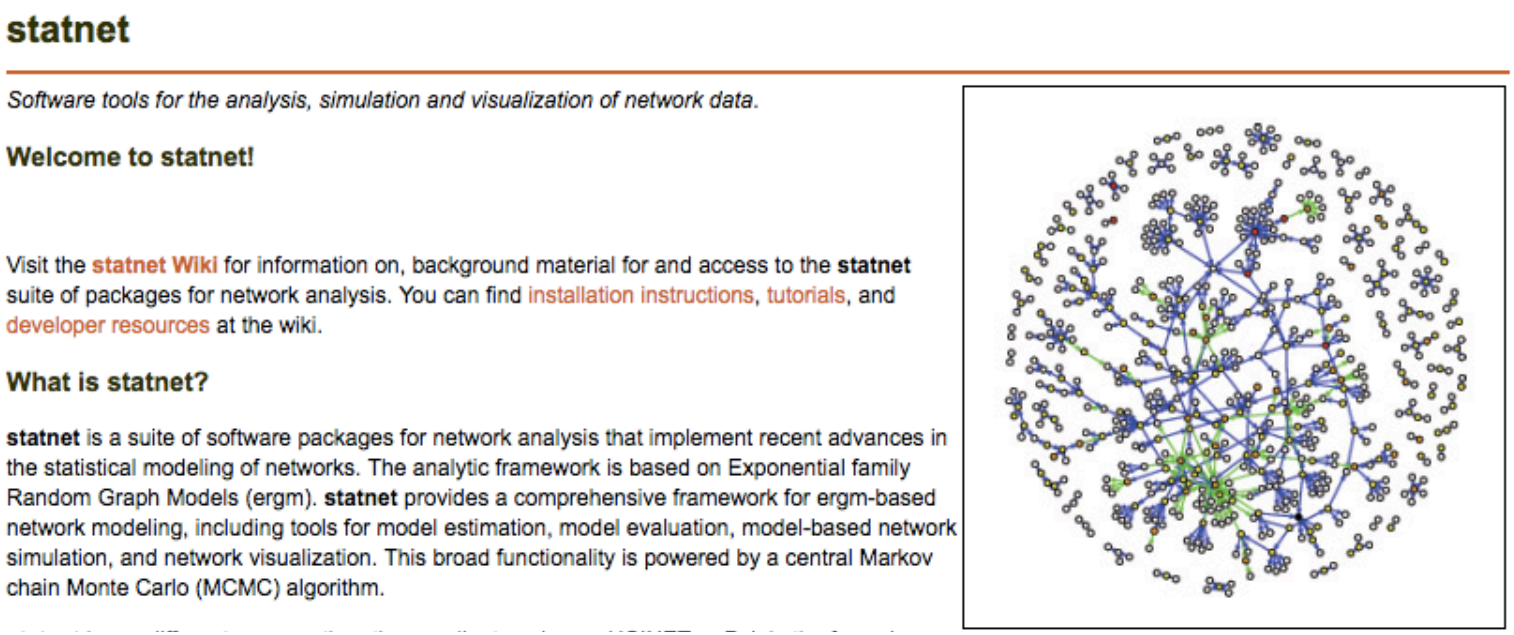
- Estimated standard errors appear slightly conservative



In R: *ergm.ego*

Also part of the statnet suite

www.statnet.org



statnet

Software tools for the analysis, simulation and visualization of network data.

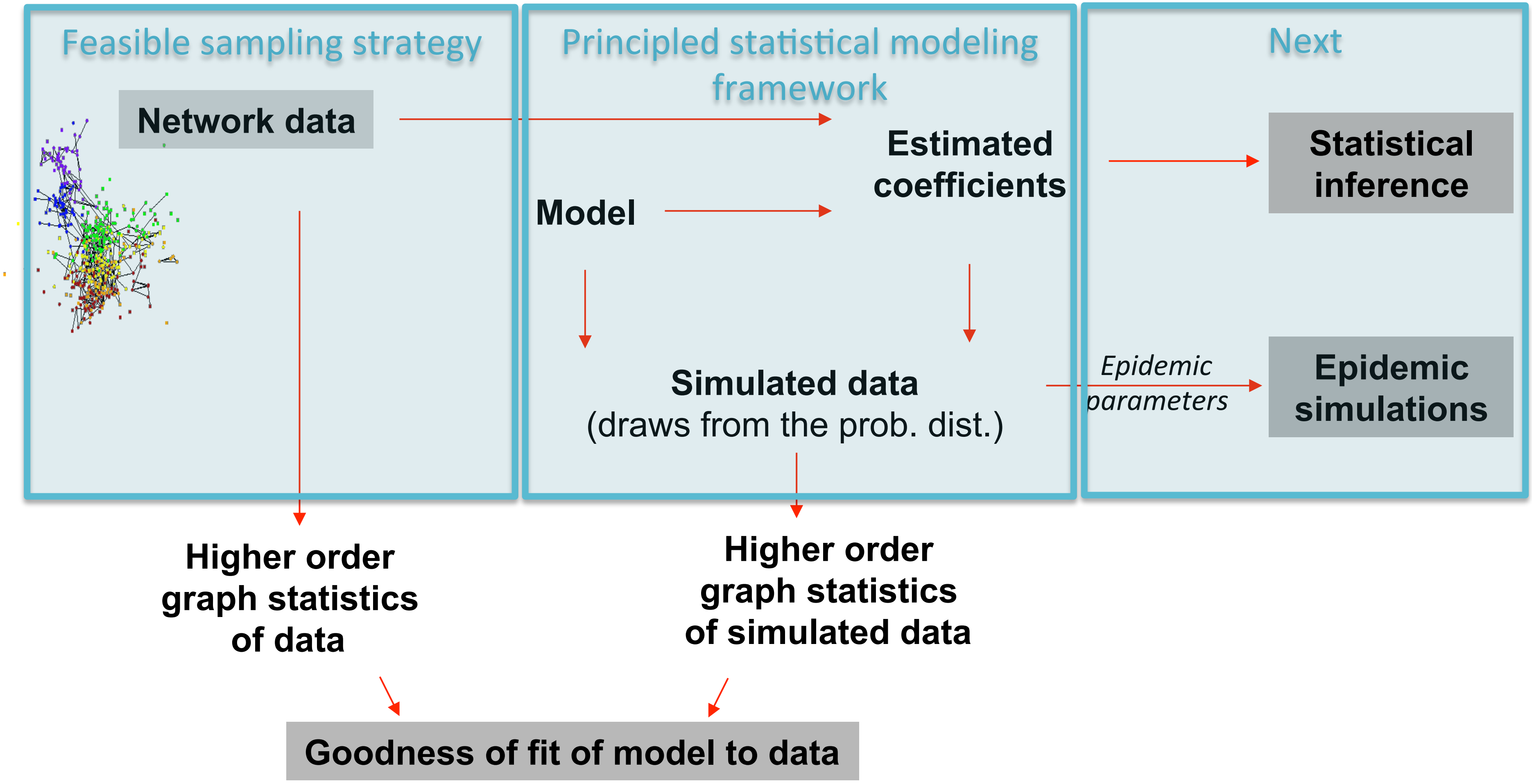
Welcome to statnet!

Visit the **statnet Wiki** for information on, background material for and access to the **statnet** suite of packages for network analysis. You can find installation instructions, tutorials, and developer resources at the wiki.

What is statnet?

statnet is a suite of software packages for network analysis that implement recent advances in the statistical modeling of networks. The analytic framework is based on Exponential family Random Graph Models (ergm). **statnet** provides a comprehensive framework for ergm-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm.

Ready for epidemic simulations



Feasible sampling strategy

Network data

Principled statistical modeling framework

Model

Estimated coefficients

Simulated data
(draws from the prob. dist.)

Next

Statistical inference

Epidemic parameters

Epidemic simulations

Higher order graph statistics of data

Higher order graph statistics of simulated data

Goodness of fit of model to data

Statistical methods for analyzing network data

Integration with epidemic modeling

HIV epidemic model overview

Foundation = dynamic network (STERGM)

Other processes overlaid

All interact with dynamic network

Demographics

- Sex, age, race structure
- Mortality

Behavior

- Coital frequency
- Condom use

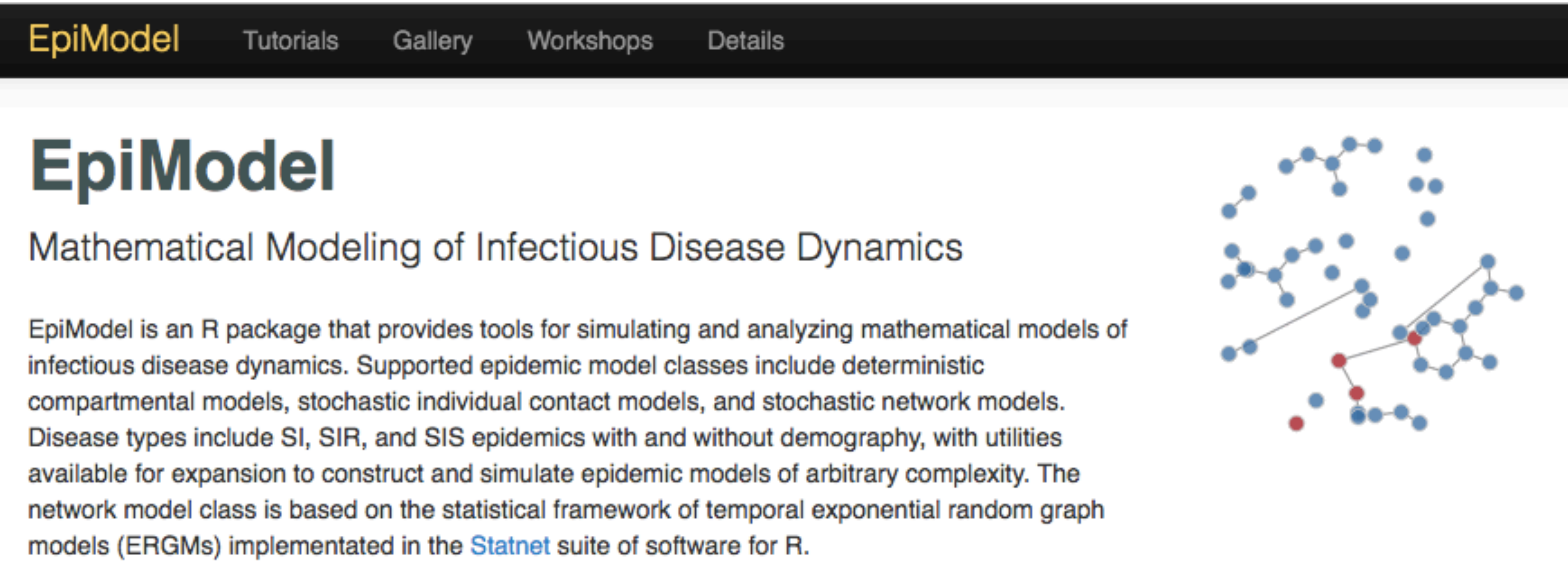
Infectivity by

- Clinical disease stage
- Diagnosis & treatment

In R: *EpiModel*

Integrates disease transmission with STERGMs

www.epimodel.org



EpiModel

Mathematical Modeling of Infectious Disease Dynamics

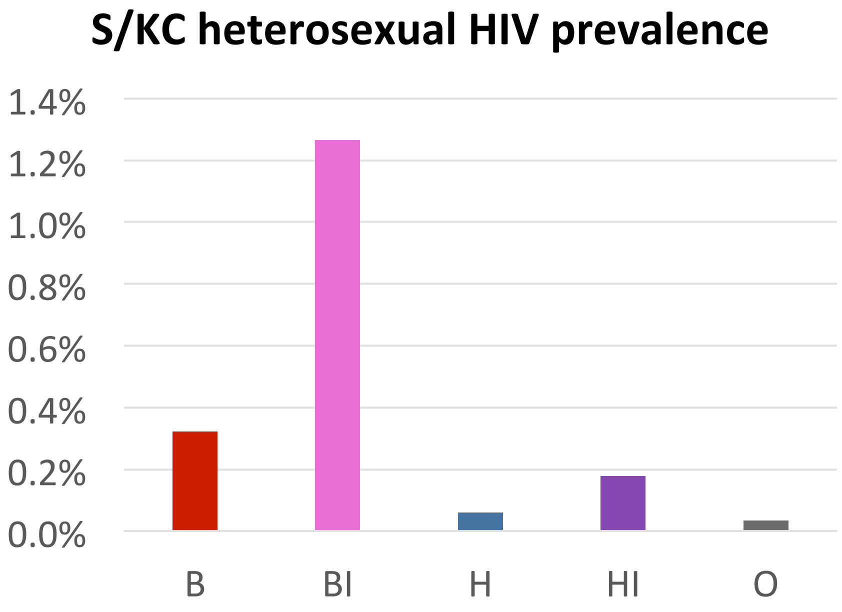
EpiModel is an R package that provides tools for simulating and analyzing mathematical models of infectious disease dynamics. Supported epidemic model classes include deterministic compartmental models, stochastic individual contact models, and stochastic network models. Disease types include SI, SIR, and SIS epidemics with and without demography, with utilities available for expansion to construct and simulate epidemic models of arbitrary complexity. The network model class is based on the statistical framework of temporal exponential random graph models (ERGMs) implementated in the Statnet suite of software for R.

Application to heterosexual networks & HIV in Seattle/King County

Local HIV dynamics & questions

Small heterosexual epidemic

- 10% of new diagnoses
  - Close to eradication?
- Racial disparities in prevalence



Local HIV dynamics & questions

1. Does the network structure contribute to observed racial disparities?

- Differential network exposure + homophily

2. Proof of concept: Can we reproduce the profile of the local epidemic?

- Disparities by race/immigration?

3. How close is heterosexual transmission to dying out (the epidemic threshold)?

- Do "bridge" contacts determine the persistence of HIV?

Potential "bridge" contacts

1. Men who have sex with men – via "MSMF"

In the Western US:

- 52% of heterosexual female HIV cases phylogenetically linked to MSM (Oster 2015)

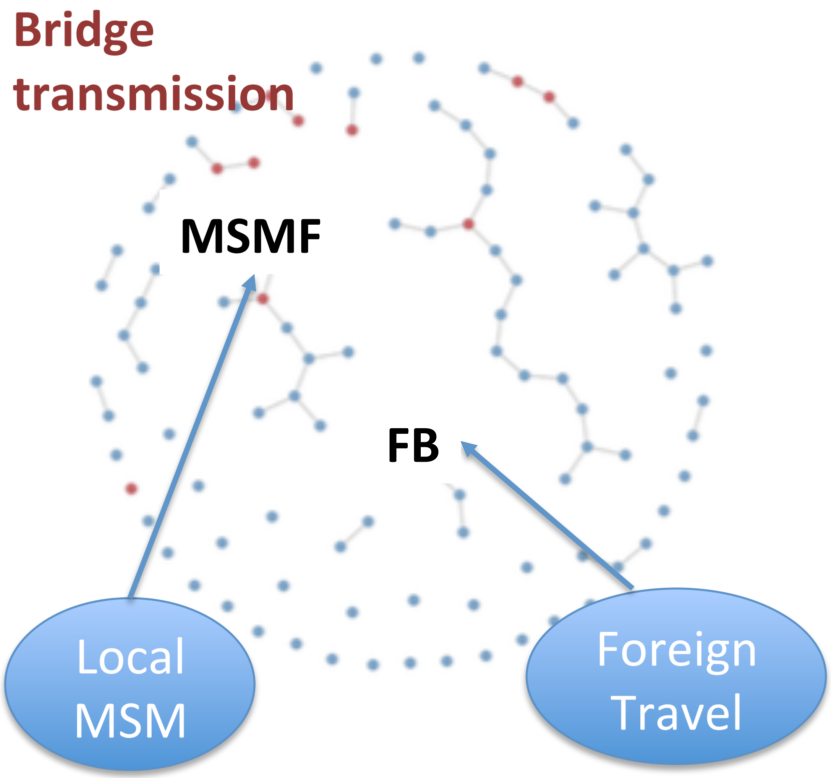
2. Foreign-acquired infections – via "FB"

In Seattle/KC:

- 41% of diagnoses in Blacks among foreign-born
- 52% of diagnoses in Hispanics among foreign-born

Local model structure & data

Foundation = dynamic network (STERGM)



Other processes overlaid

All interact with dynamic network

Demographics

- Sex, age, race structure
- Mortality

Behavior

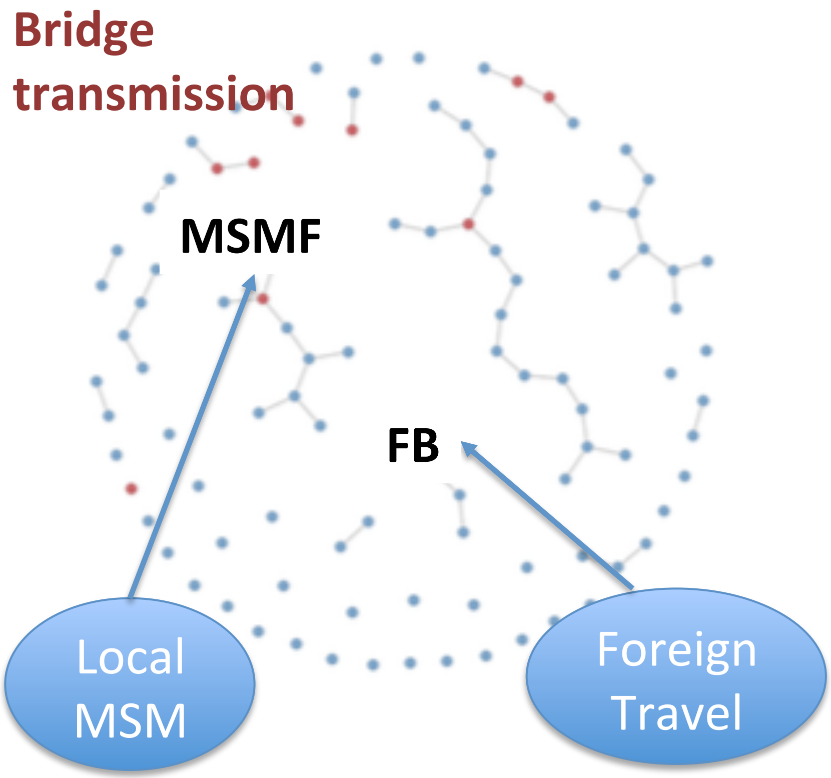
- Coital frequency
- Condom use

Infectivity by

- Clinical disease stage
- Diagnosis & treatment

Local model structure & data

Foundation = dynamic network (STERGM)



Other processes overlaid

All interact with dynamic network

Demographics

- Sex, age, race structure
- Mortality

Behavior

- Coital frequency
- Condom use

Infectivity by

- Clinical disease stage
- Diagnosis & treatment

ERGMs on egocentric data

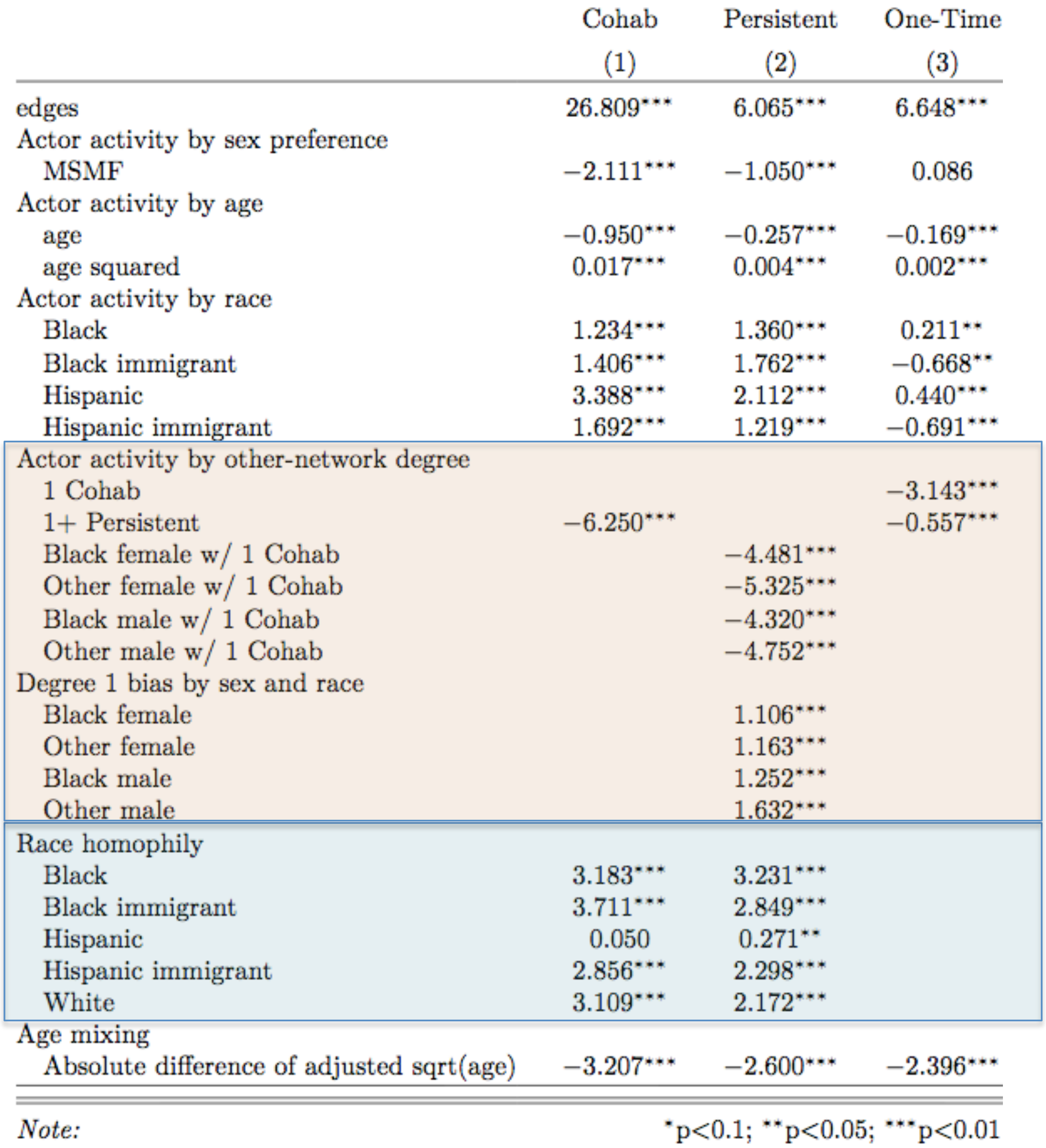
National Survey of Family Growth

- Egocentric survey of most 3 recent sexual partners
  - N = 40,000 respondents (2006-15)
- Weighted to Seattle/King County
  - By age, sex and race/immigration

Three overlapping networks

- Cohabitating
- Persistent
- One-Time

Formation models indicate differential network exposure & homophily



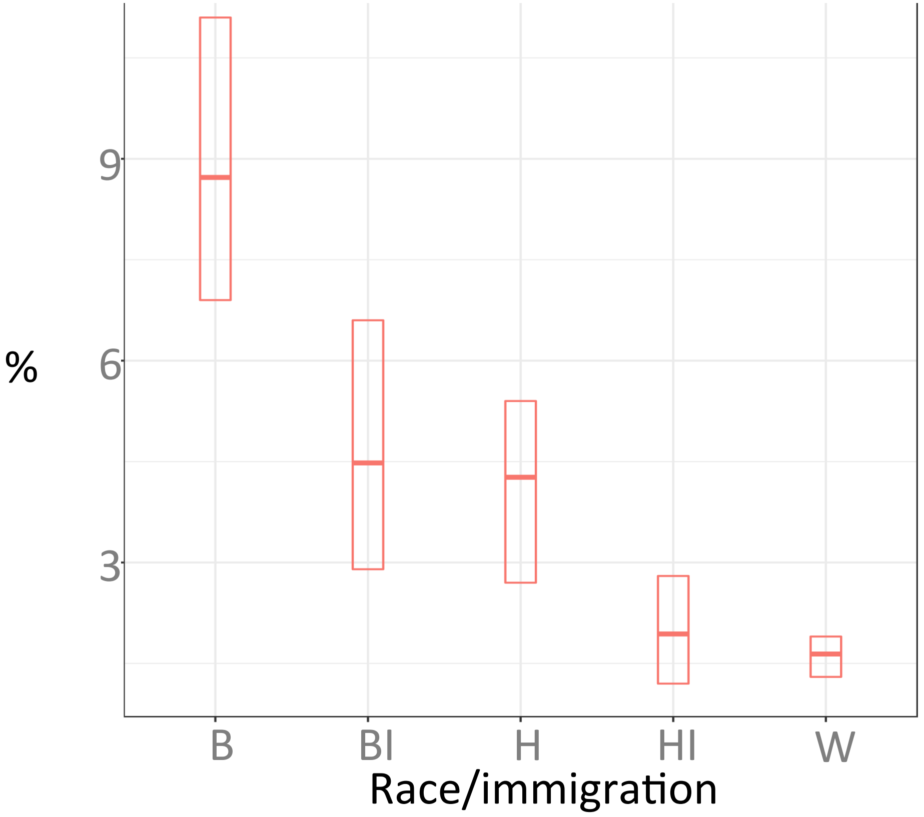
	Cohab	Persistent	One-Time
	(1)	(2)	(3)
edges	26.809***	6.065***	6.648***
Actor activity by sex preference			
MSMF	−2.111***	−1.050***	0.086
Actor activity by age			
age	−0.950***	−0.257***	−0.169***
age squared	0.017***	0.004***	0.002***
Actor activity by race			
Black	1.234***	1.360***	0.211**
Black immigrant	1.406***	1.762***	−0.668**
Hispanic	3.388***	2.112***	0.440***
Hispanic immigrant	1.692***	1.219***	−0.691***
Actor activity by other-network degree			
1 Cohab			−3.143***
1+ Persistent	−6.250***		−0.557***
Black female w/ 1 Cohab		−4.481***	
Other female w/ 1 Cohab		−5.325***	
Black male w/ 1 Cohab		−4.320***	
Other male w/ 1 Cohab		−4.752***	
Degree 1 bias by sex and race			
Black female		1.106***	
Other female		1.163***	
Black male		1.252***	
Other male		1.632***	
Race homophily			
Black	3.183***	3.231***	
Black immigrant	3.711***	2.849***	
Hispanic	0.050	0.271**	
Hispanic immigrant	2.856***	2.298***	
White	3.109***	2.172***	
Age mixing			
Absolute difference of adjusted sqrt(age)	−3.207***	−2.600***	−2.396***

Note: *p<0.1; **p<0.05; ***p<0.01

Key finding 1:

Differential network exposure by race confirmed in simulated networks

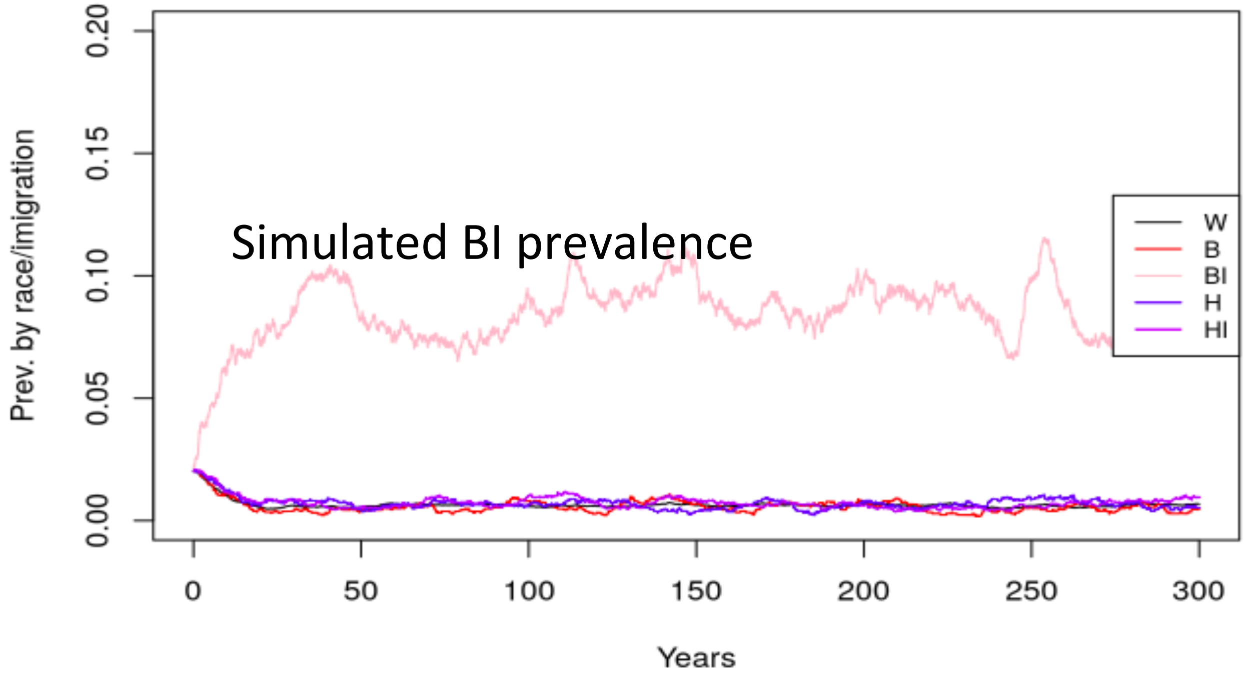
% of nodes in components > 2, by race



Key finding 2:
Disparity profiles are reproduced

Correct rank ordering of prevalence

Simulated prevalence by race/immigration

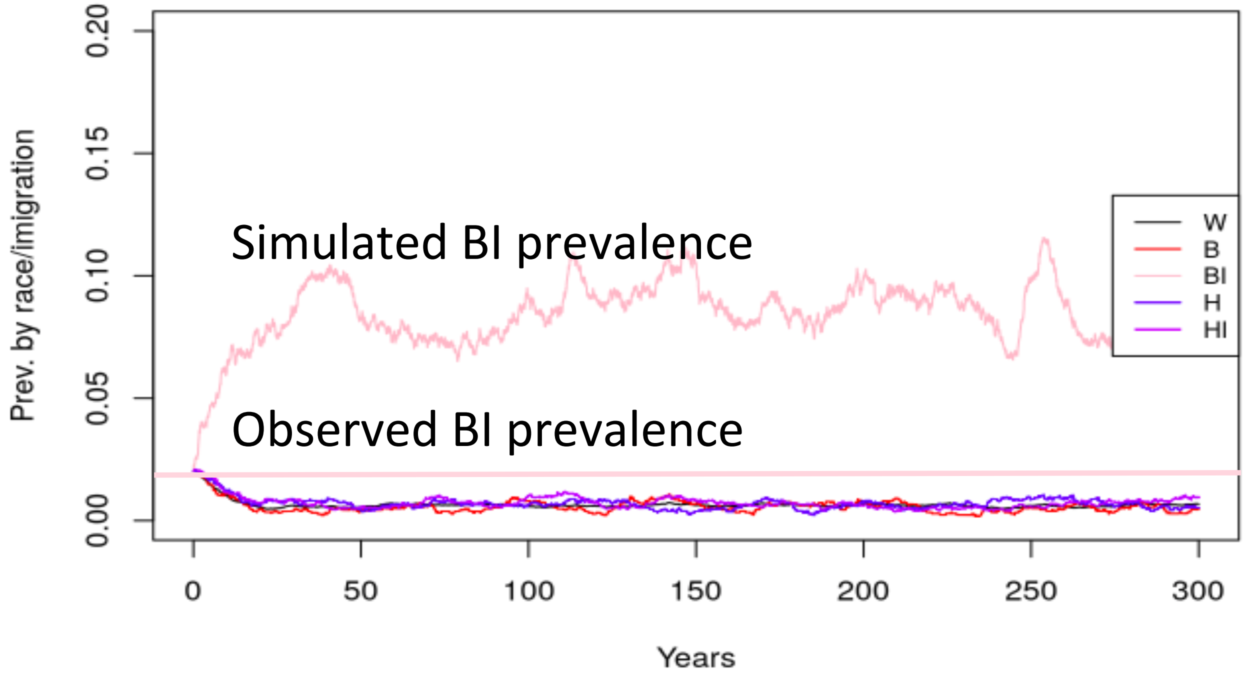


Key finding 2:

Disparity profiles are reproduced

“Bridge” group transmission too high

Simulated prevalence by race/immigration



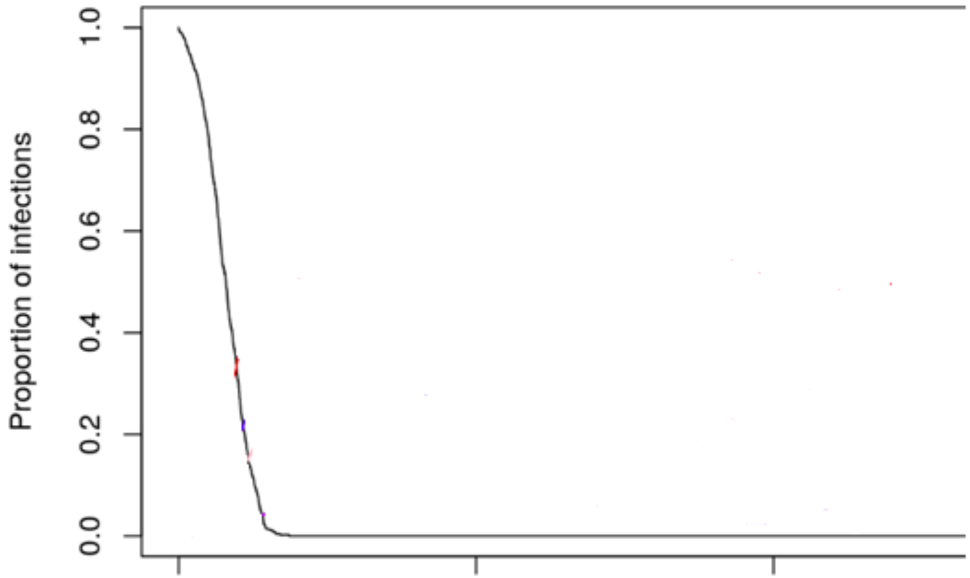
Simulated BI prevalence

Observed BI prevalence

Key finding 3:

Bridging is key to HIV persistence

The epidemic could not be sustained in the local heterosexual population alone



Initial 1% local prevalence → epidemic extinction in ~30 years.

Conclusions

ERGM framework + egocentric survey data = feasible network analysis and epidemic modeling

- Test network-structure hypotheses via statistical inference
- Strong basis for statistically-principled epidemic modeling

Supports *local* epidemic modeling

- Model structure based in local features
- Data drawn from local populations
- Investigate locally tailored policies to end HIV

Network Modeling Group

- ***FACULTY:*** Martina Morris (UW) Steve Goodreau (UW), Carter Butts (UCI), Mark Handcock (UCLA), Dave Hunter (PSU), Pavel Krivitsky (UNSW), Skye Bender-deMoll (at large), Sam Jenness (Emory)
- ***UW RESEARCH SCIENTISTS AND STUDENTS:*** **Deven Hamilton** (Soc), Sara Stansfield (Anthro), Emily Pollock (Anthro), Darcy Rao (Epi), Sara Khan (Informatics), Chad Klumb (Math), Adam Elder (Biostat)

Resources

statnet

www.statnet.org

EpiModel

www.epimodel.org

“Network Modeling for Epidemics” 5-day course