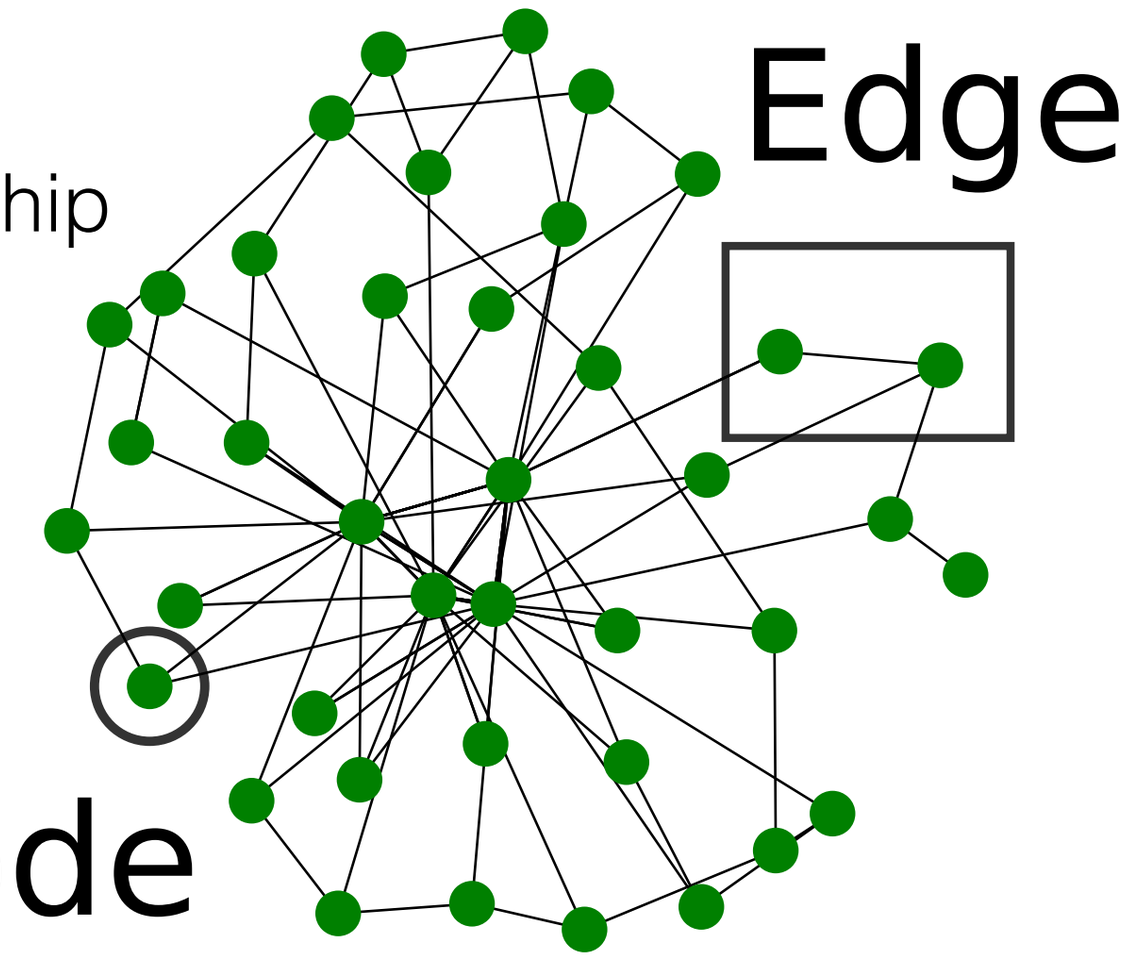
Scientific Collaboration Networks and Persistence of Endemic Disease in Heterogeneous Populations

B Exam: Daniel T Citron
Advisor: Chris Myers
July 24, 2017

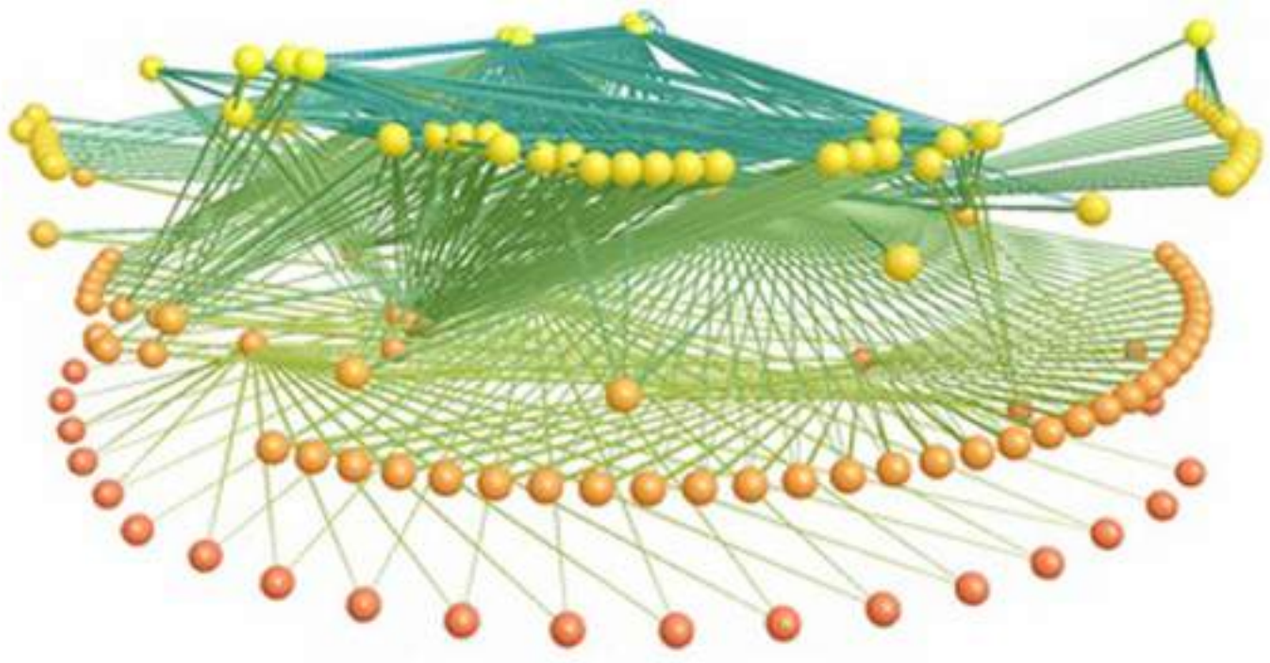
What are Networks?

- A collection of nodes connected by edges
  - Node: system component
    - Edge: interaction or relationship
- Variety of contexts



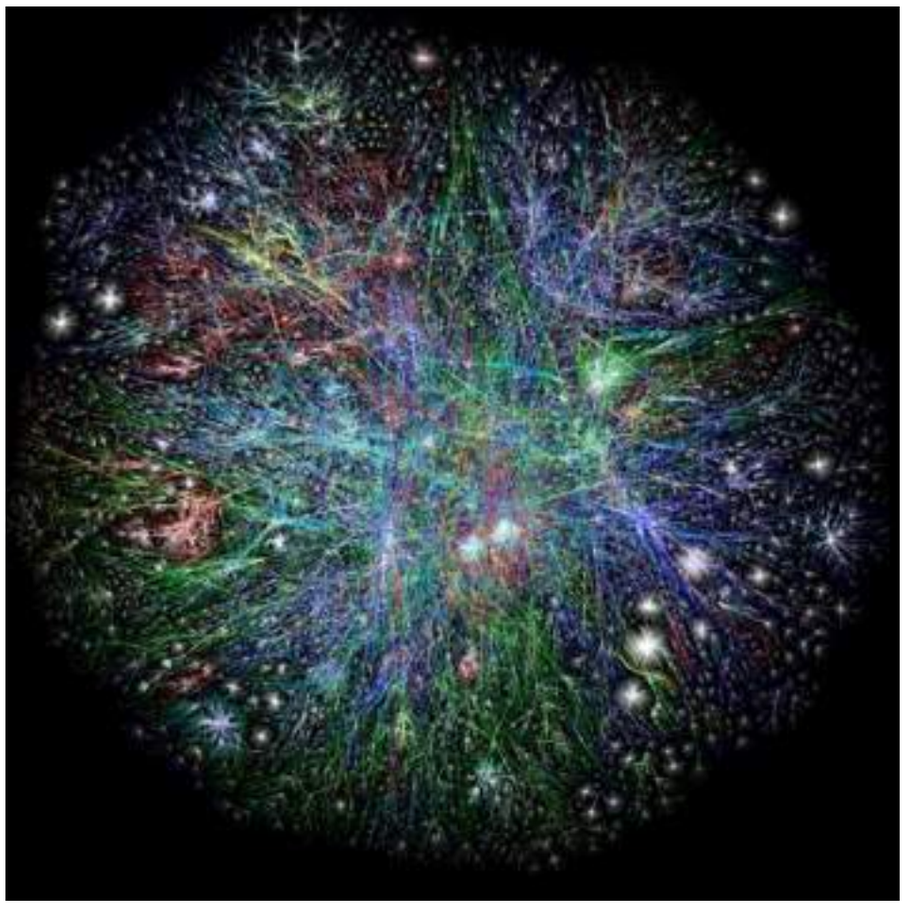
Biological Network

Food Web



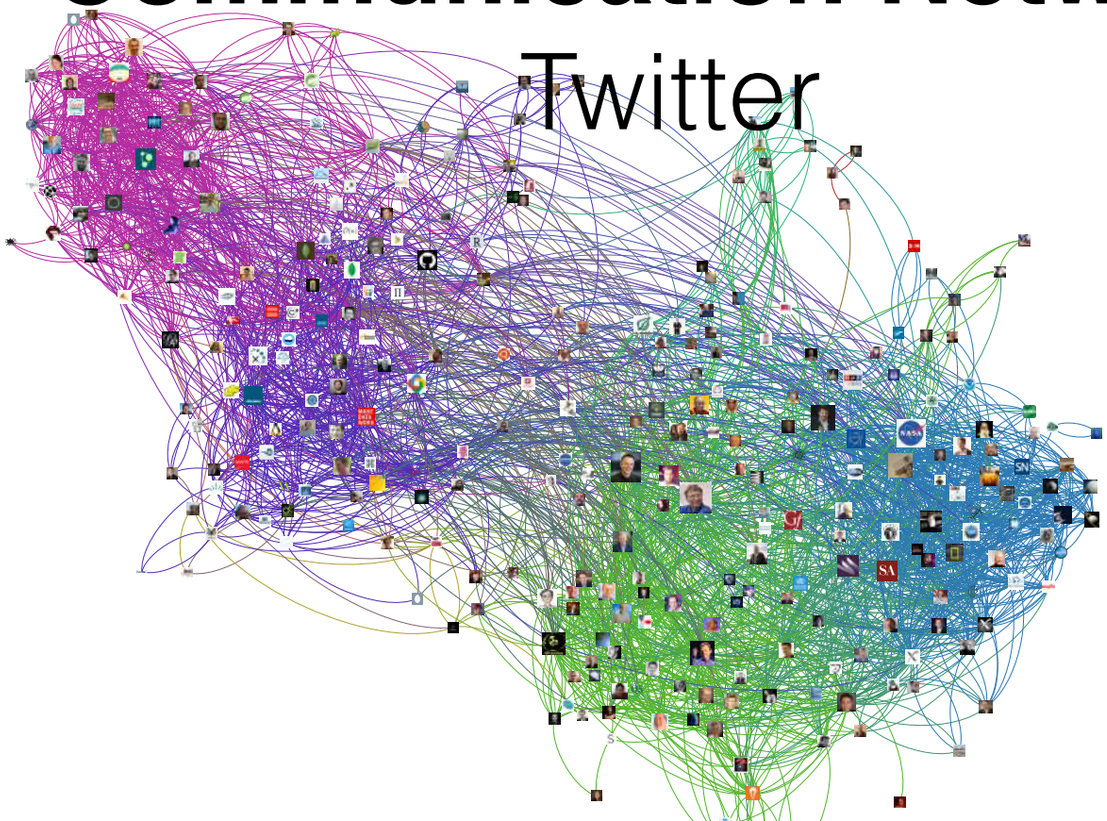
Technological Network

Internet



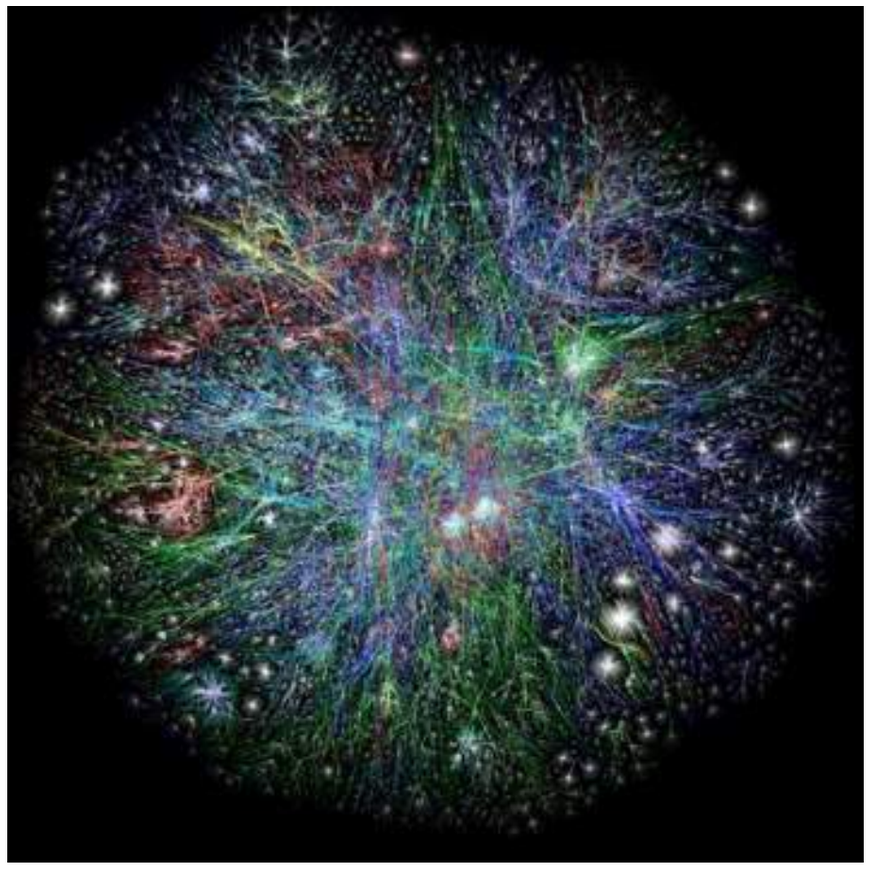
Communication Network

Twitter



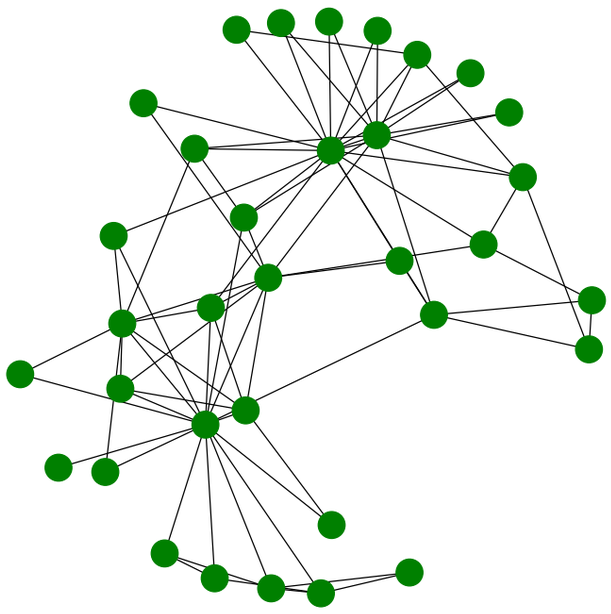
Why Networks?

- Versatile mathematical framework
- Well-suited for understanding complex systems
- Lack symmetry or regularity
- Sparsely connected
- Structure across multiple scales

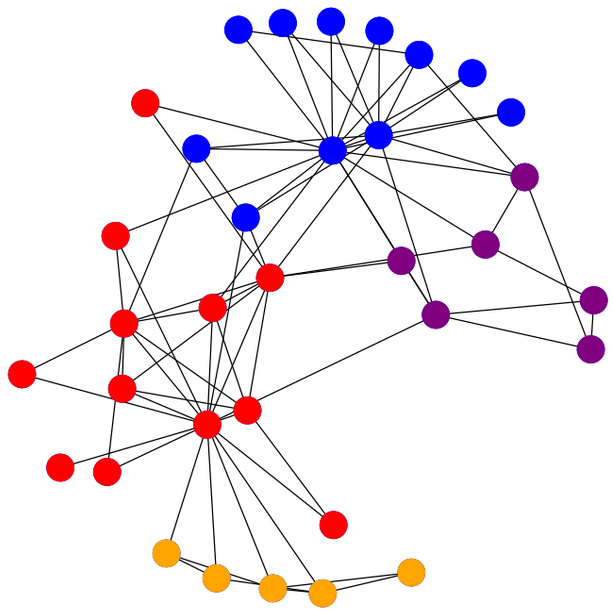


Studying Network Structure

- Exploring a structured data set
- Network represents the data
- Example: Clustering in network data

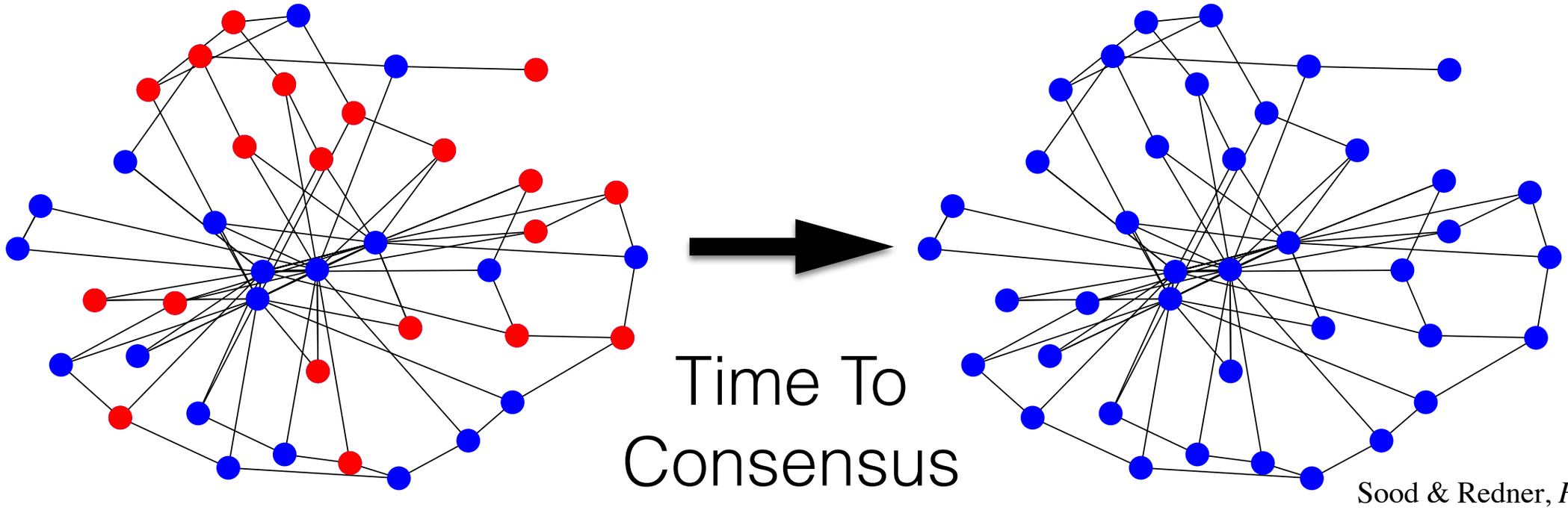


Partition based on connectivity



Dynamics on Networks

- Network acts as a substrate for a physical process
- How does the network structure change the process?
- Example: Voter model of consensus formation



Sood & Redner, *PRL* (2005).

Castellano et al., *arXiv* (2007)

Outline

- Part 1: Modeling
  - Study of persistence of endemic disease
- Part 2: Structure
  - Structural study of co-authorship network evolution

Persistence of Endemic Disease in Heterogeneous Populations

Modeling Disease Dynamics

- Basic biological and clinical research is always necessary
  - Understanding modes of transmission
  - Discovering treatments and cures
- Modeling useful for simulation-based experiments:
  - Understanding population-level dynamics
  - Forecasting
  - Evaluating treatment deployment strategies

Goal:
Modeling Endemic Disease Persistence

- Persists in a population, rather than dying out after a single outbreak

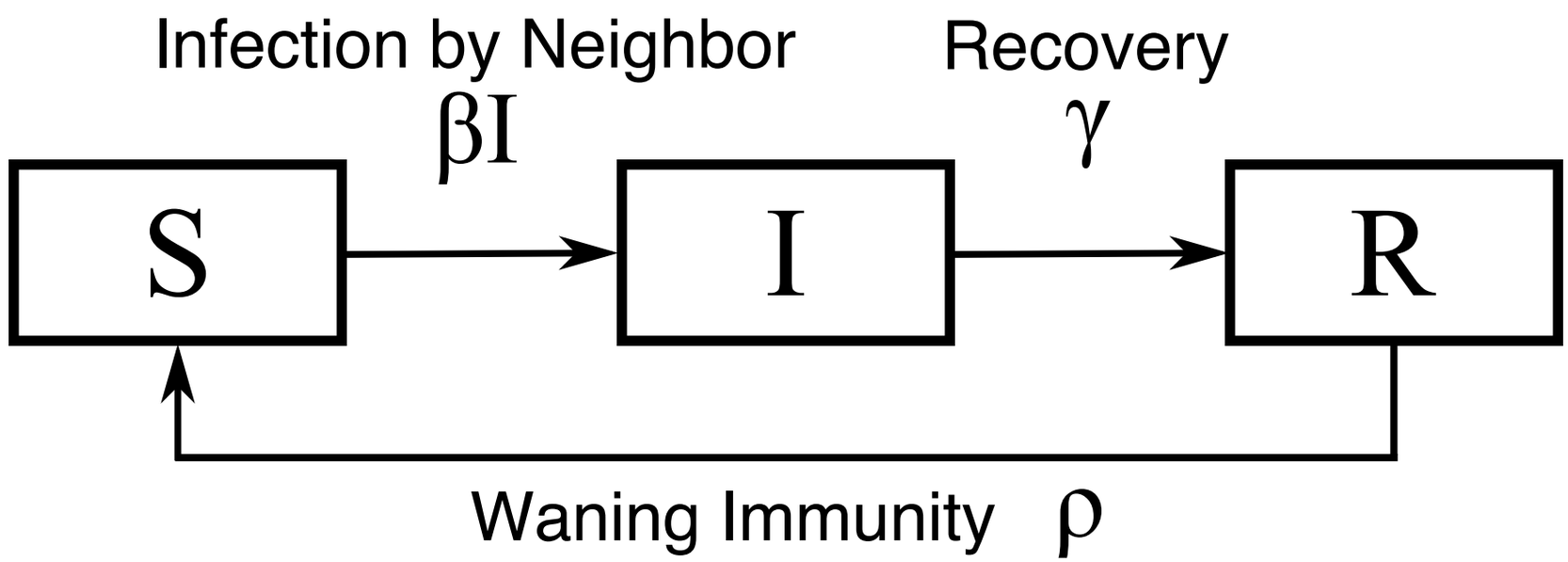
- Possible to die out spontaneously

- Our goal: use modeling to understand
  - Population-level dynamics of endemic disease
  - Times to extinction
  - How structure of contact network affects outcomes

Compartmental Models

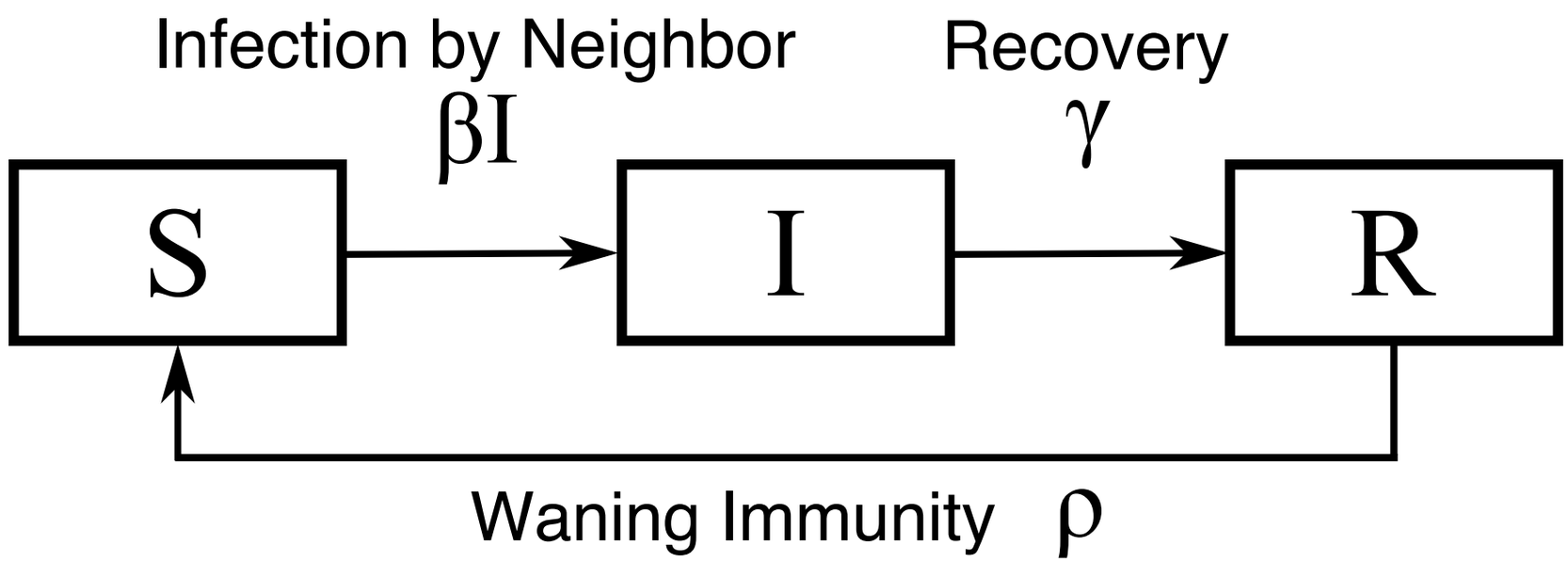
- Consider a population of hosts
- Each host may be in one of these states:
  - Susceptible - not yet infected
  - Infected - has disease, and can spread disease to others
  - Recovered - no longer has disease, no longer susceptible

Modeling Endemic Disease



- “SIRS” Model
- Replenish susceptibles
- Loss of immunity

Modeling Endemic Disease



- "SIRS" Model

$$S + I + R = 1$$

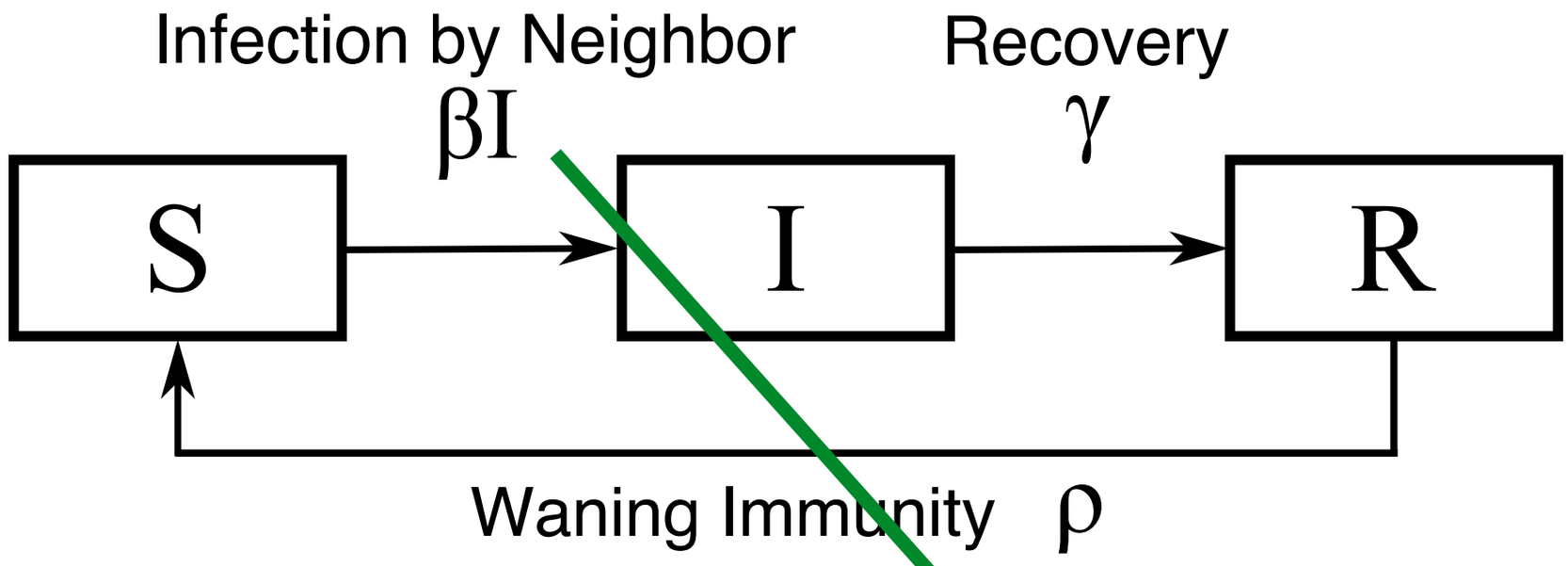
- Replenish susceptibles

$$\frac{dS}{dt} = -\beta SI + \rho\left(1 - S - I\right)$$

- Loss of immunity

$$\frac{dI}{dt} = \beta SI - \gamma I$$

Modeling Endemic Disease



- "SIRS" Model

$$S + I + R = 1$$

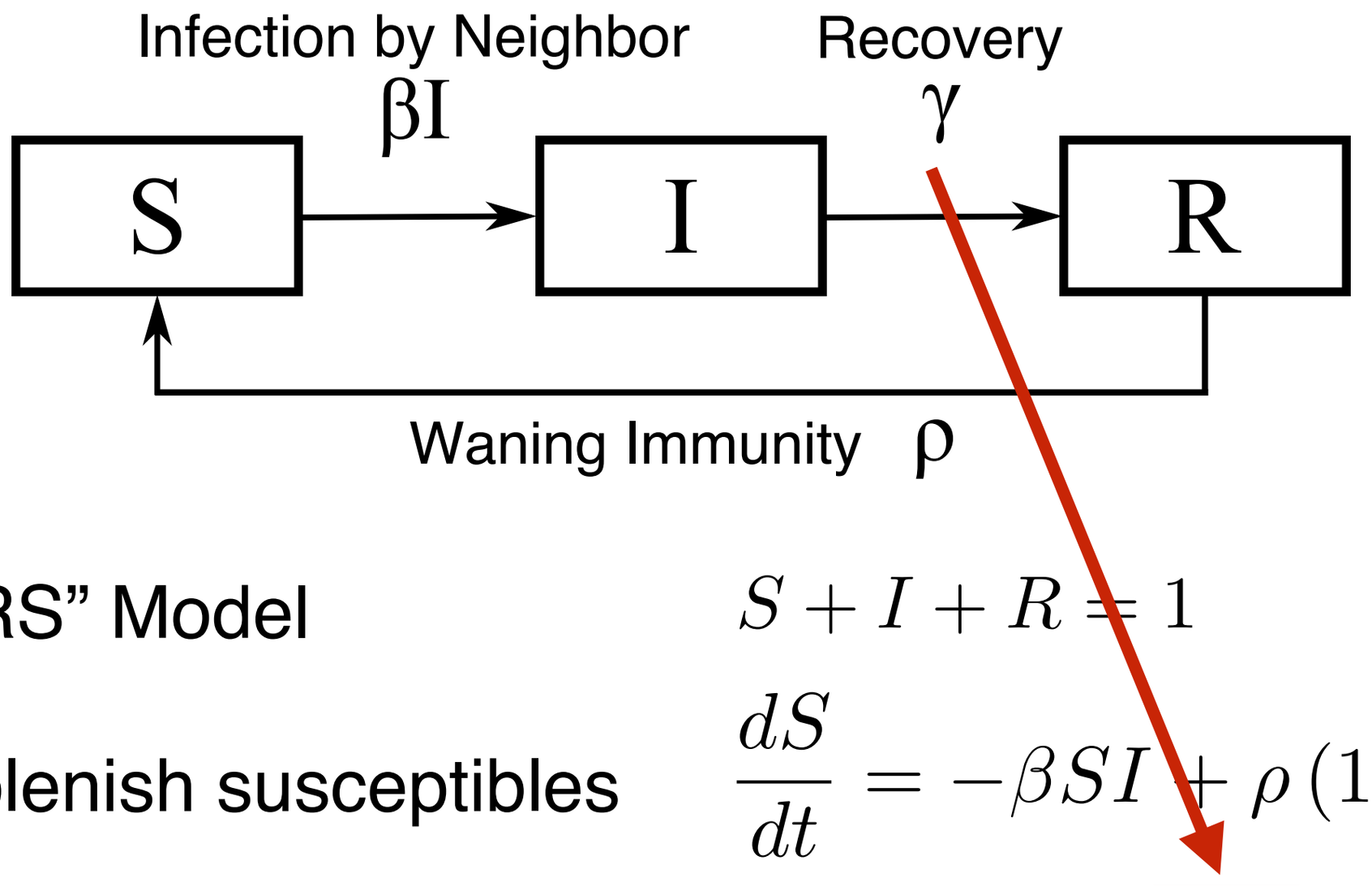
- Replenish susceptibles

$$\frac{dS}{dt} = -\beta SI + \rho\,(1 - S - I)$$

- Loss of immunity

$$\frac{dI}{dt} = \beta SI - \gamma I$$

Modeling Endemic Disease



- "SIRS" Model

$$S + I + R = 1$$

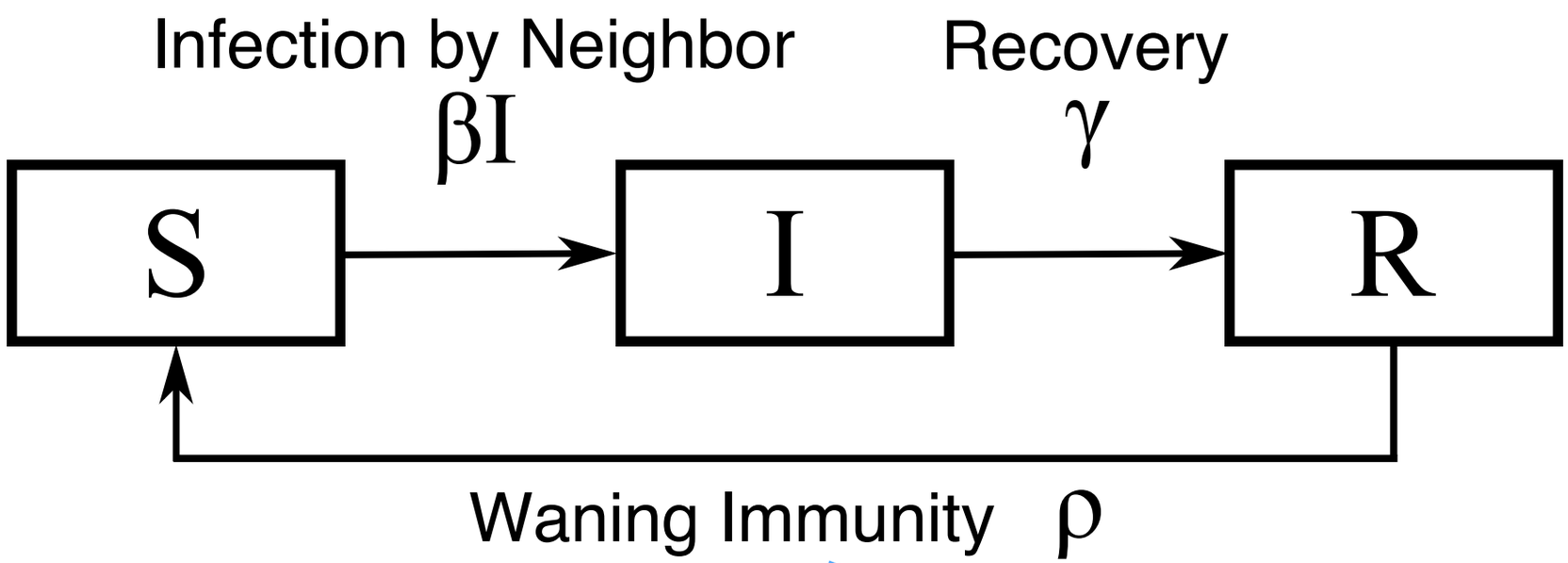
- Replenish susceptibles

$$\frac{dS}{dt} = -\beta SI + \rho\left(1 - S - I\right)$$

- Loss of immunity

$$\frac{dI}{dt} = \beta SI - \gamma I$$

Modeling Endemic Disease



- "SIRS" Model
- Replenish susceptibles
- Loss of immunity

$$S + I + R = 1$$

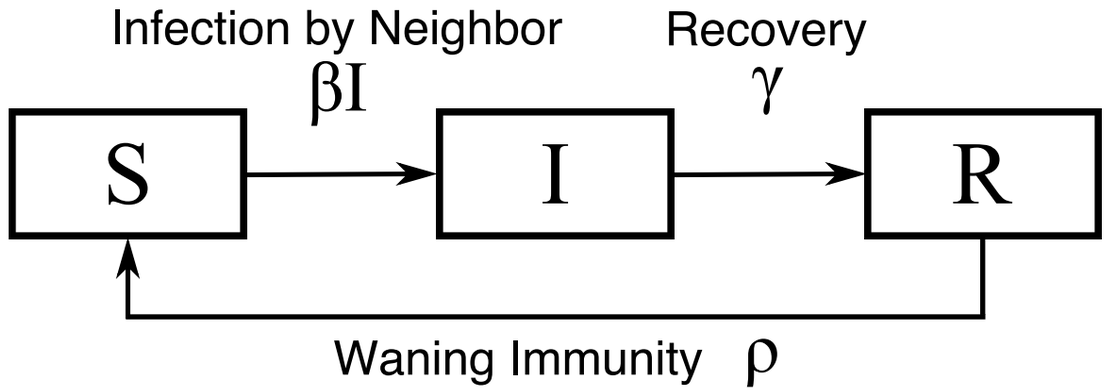
$$\frac{dS}{dt} = -\beta SI + \rho(1 - S - I)$$

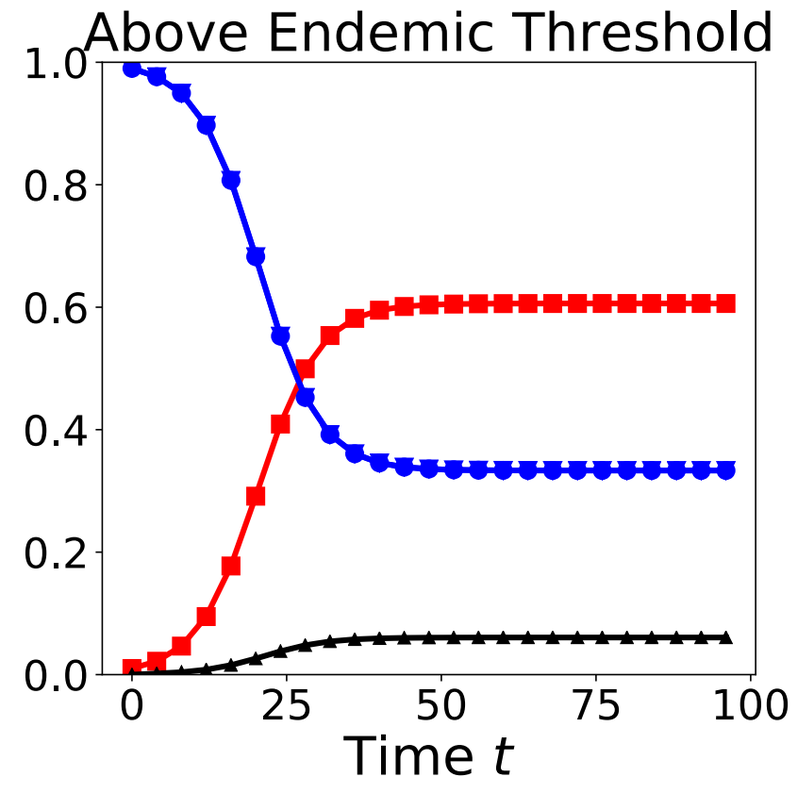
$$\frac{dI}{dt} = \beta SI - \gamma I$$

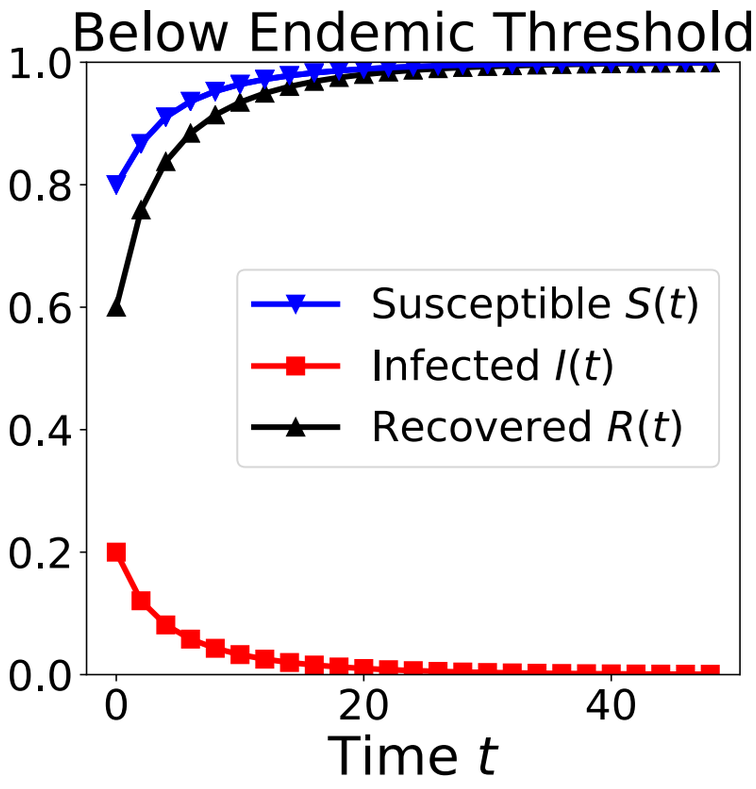
SIRS Endemic State

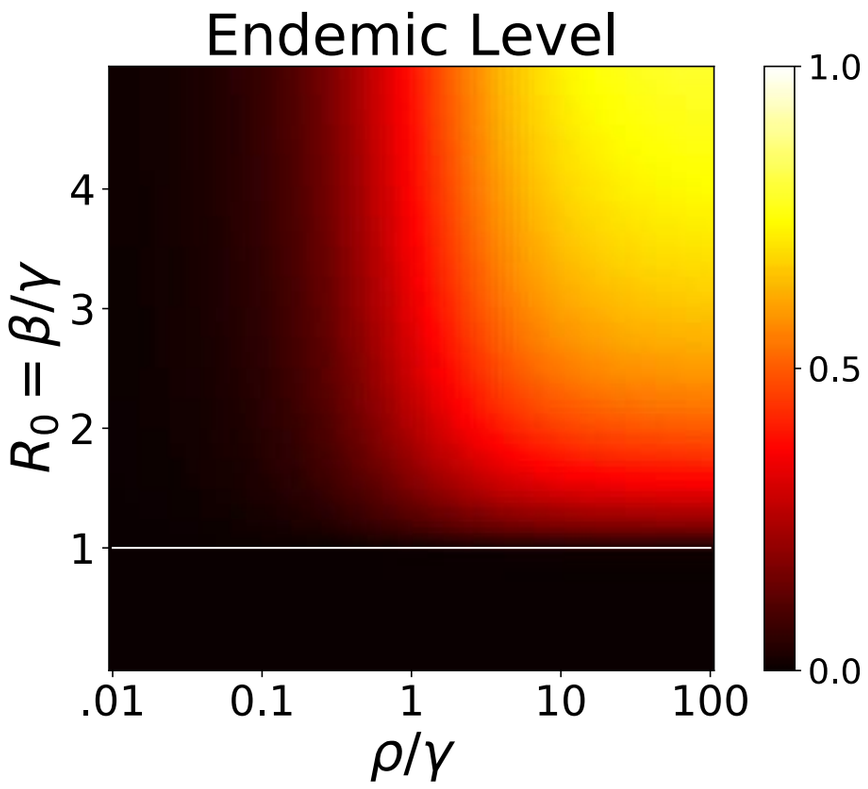
Steady State Endemic Level:

$$I^* = \frac{\rho}{\rho + \gamma}\left(1 - \frac{\gamma}{\beta}\right), \qquad \frac{\beta}{\gamma} > 1$$



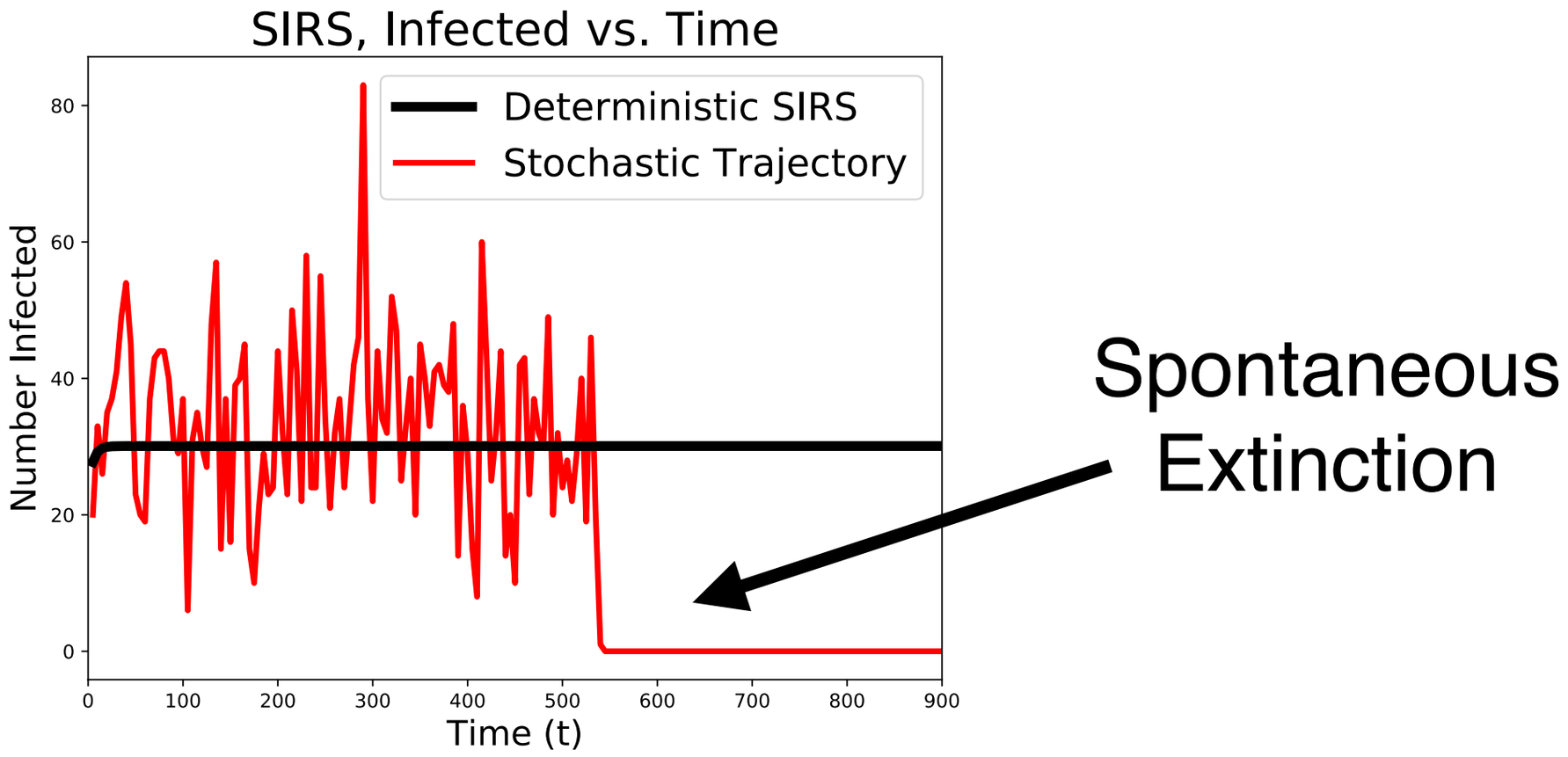






Spontaneous Extinctions

- Deterministic models: endemic disease persists forever
- Real-world endemic disease dies out spontaneously



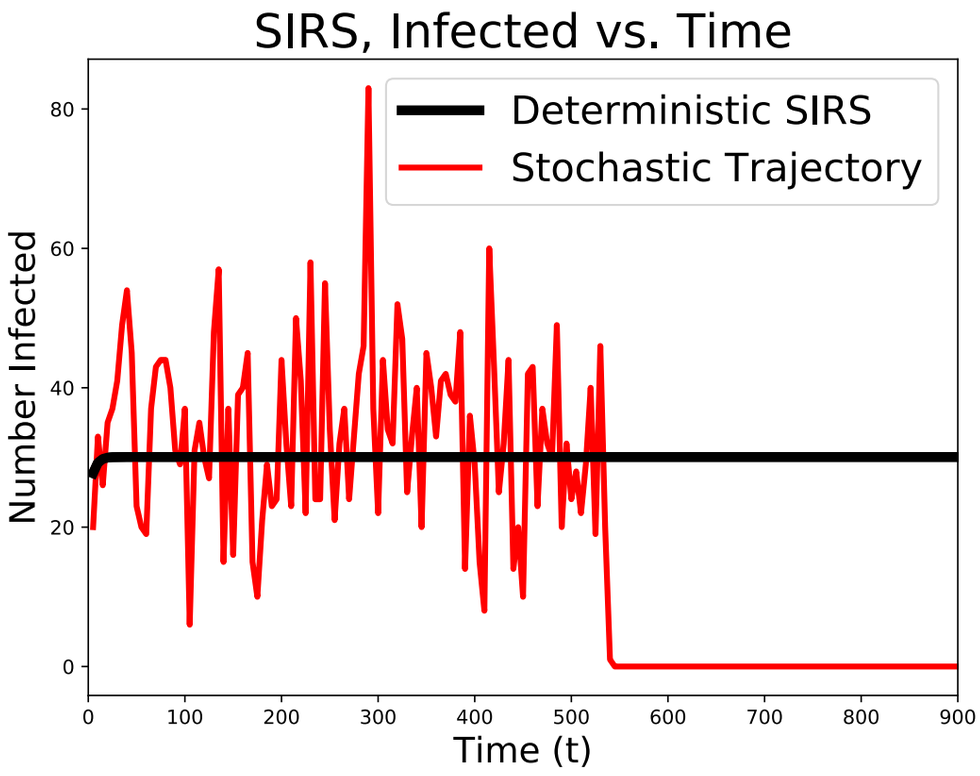
Spontaneous Extinctions

- Fluctuations matter:

  Larger fluctuations/mean => more likely

  Smaller fluctuations/mean => less likely

- Distribution of fluctuations
- Mean time to extinction (τ)
- (Contact network structure?)



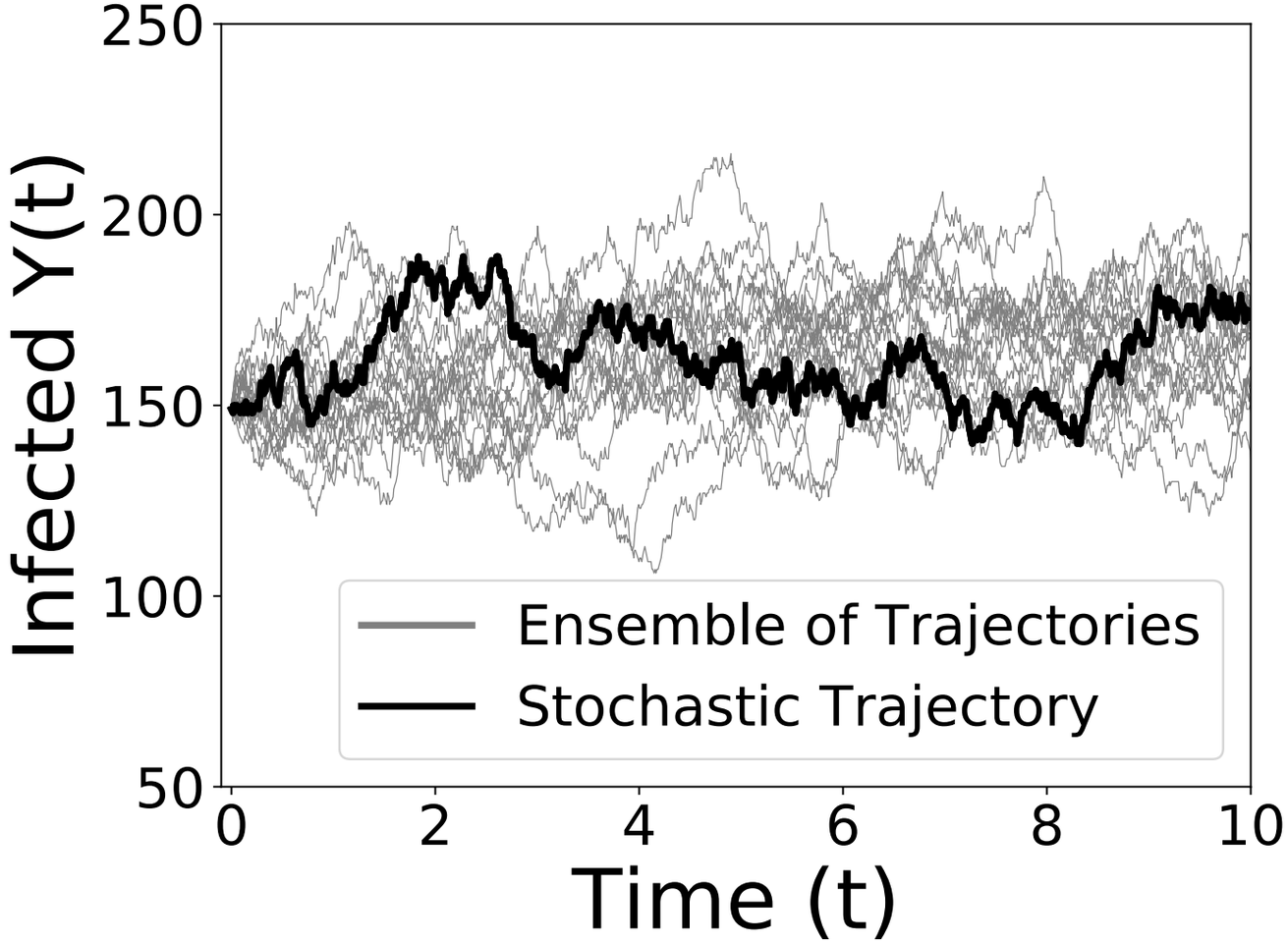
Stochastic Modeling

- Continuous variables (S,I) -> discrete variables (X,Y)
- Probability distribution **P(X=m, Y=n, t)**
  - probability of finding m susceptible, n infected
  - continuous time Markov Chain
  - described by Master Equation

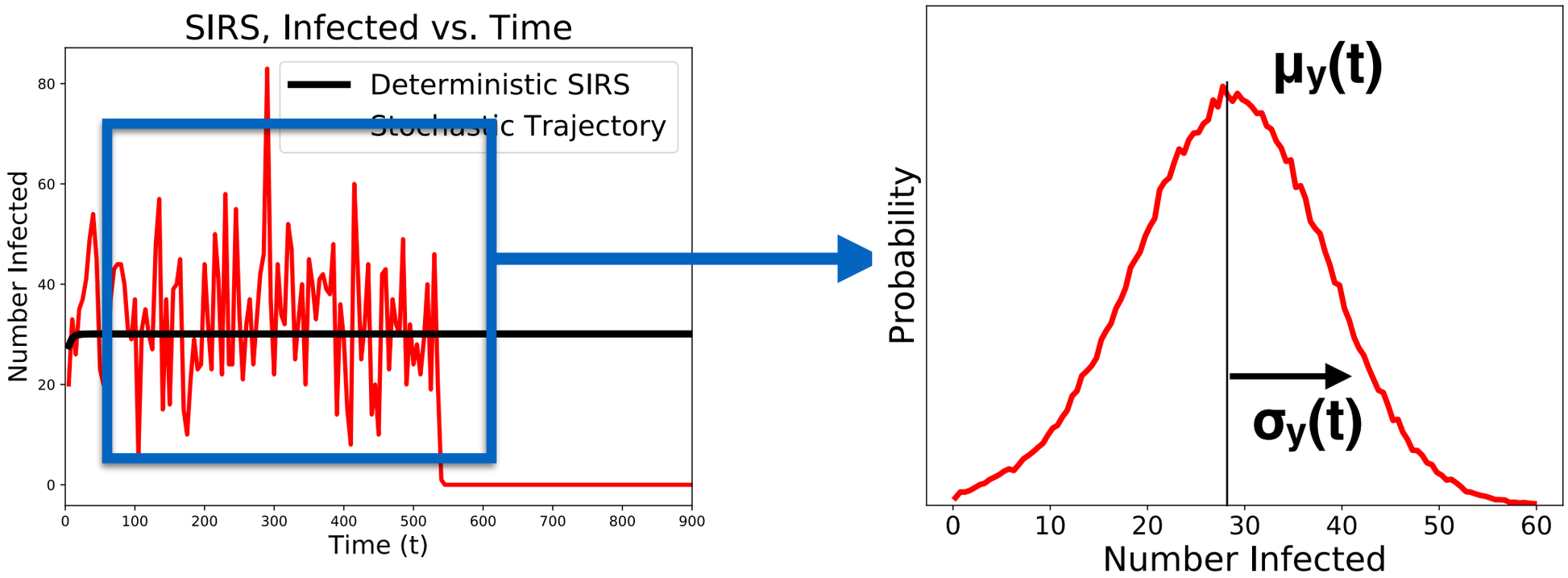
Event	Transition	Rate
Infection	$(m,n) \longrightarrow (m-1, n+1)$	$\beta mn/N$
Recovery	$(m,n) \longrightarrow (m, n-1)$	γm
Loss of Immunity	$(m,n) \longrightarrow (m+1, n)$	$\rho\,(N-m-n)$

Stochastic Trajectories

- **P(X=m,Y=n, t)** describes ensemble of trajectories
- Want to describe the properties of this ensemble



Capturing Fluctuations



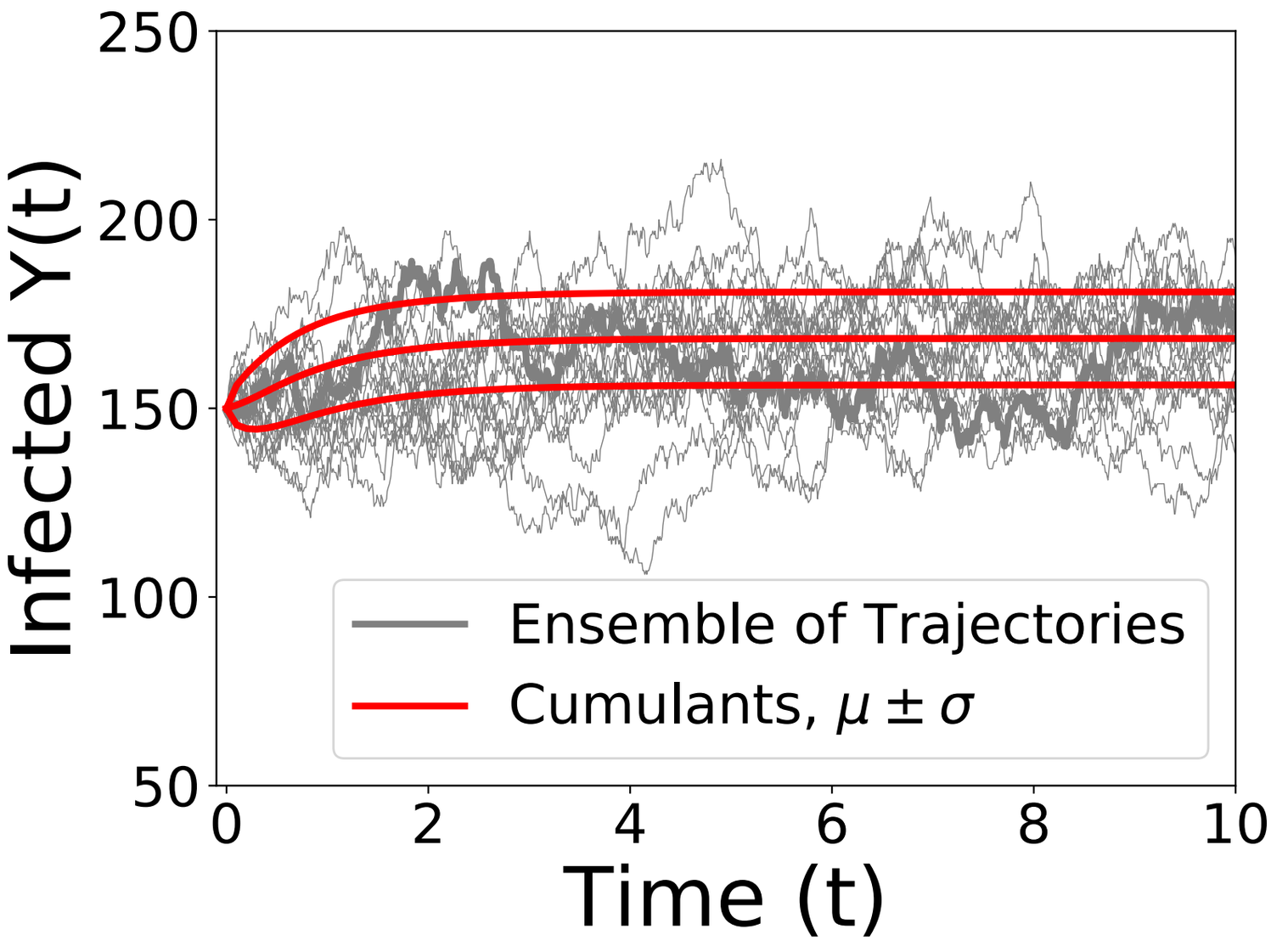
- Fluctuations about the mean have a typical size
  - Mean number susceptible/infected (**$\mu_x(t)$**, **$\mu_y(t)$**)
  - Fluctuation sizes (**$\sigma_x(t)$**, **$\sigma_y(t)$**, **$\sigma_{x,y}(t)$**)

Cumulant Equations

- Transform **P(m,n,t)**: expand in terms of its cumulants
- Assume a Gaussian distribution for **P(m,n,t)**
- Truncate cumulant expansion to include only first and second cumulants:
  - $\boldsymbol{\mu_x(t)}$, $\boldsymbol{\mu_y(t)}$, $\boldsymbol{\sigma^2_x(t)}$, $\boldsymbol{\sigma^2_y(t)}$, $\boldsymbol{\sigma_{x,y}(t)}$
- Master Equation becomes set of ODE's for cumulants
- Significantly reduces problem difficulty
- Gives analytical intuition for distribution behavior

Cumulant Equations

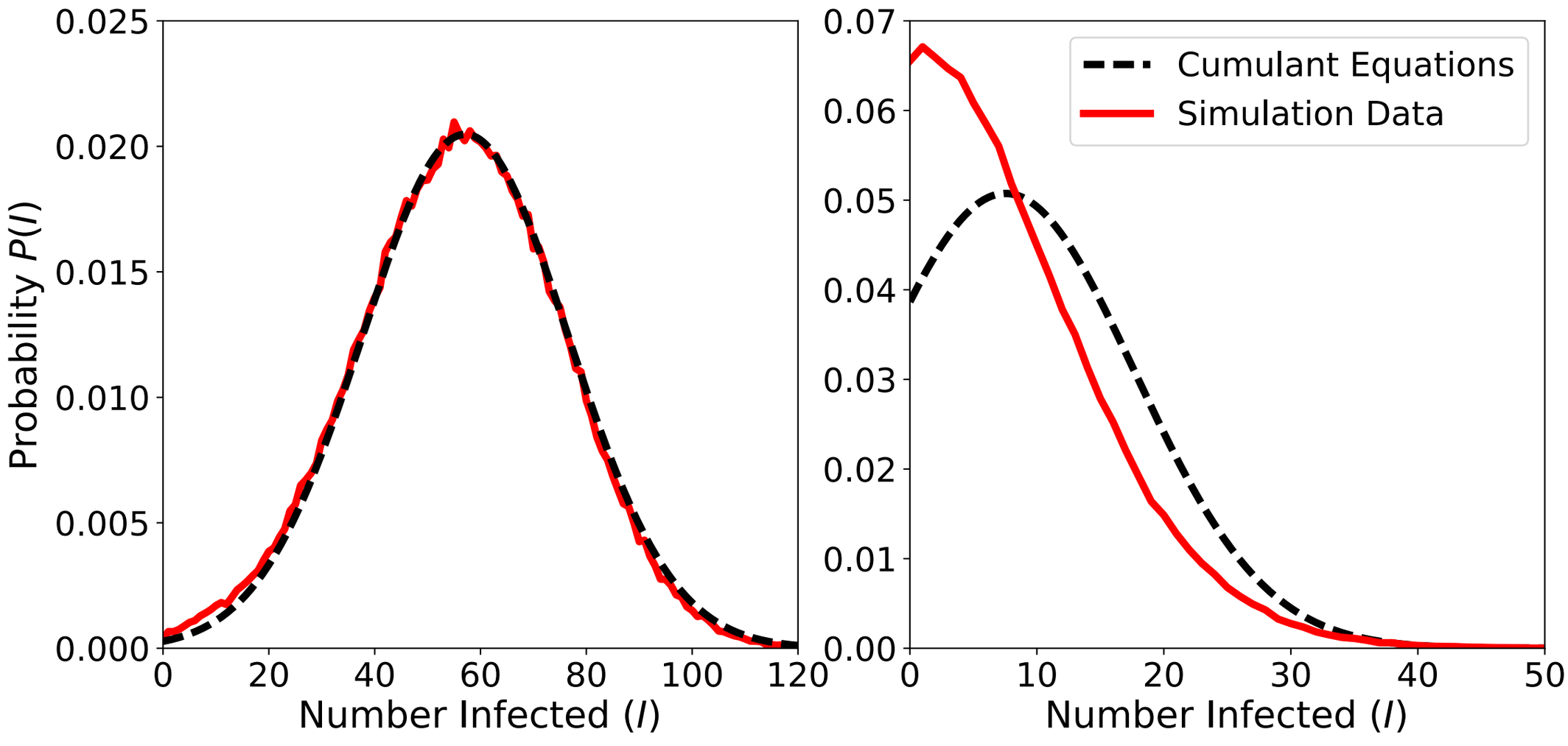
- Describe ensemble of trajectories in terms of statistical properties: $\boldsymbol{\mu_Y(t)}$,$\boldsymbol{\sigma_Y(t)}$



Comparison with Simulation

- σ_y/μ_y Small
- Distribution away from 0
- Good approximation

- σ_y/μ_y Large
- Distribution overlaps with 0
- Bad approximation

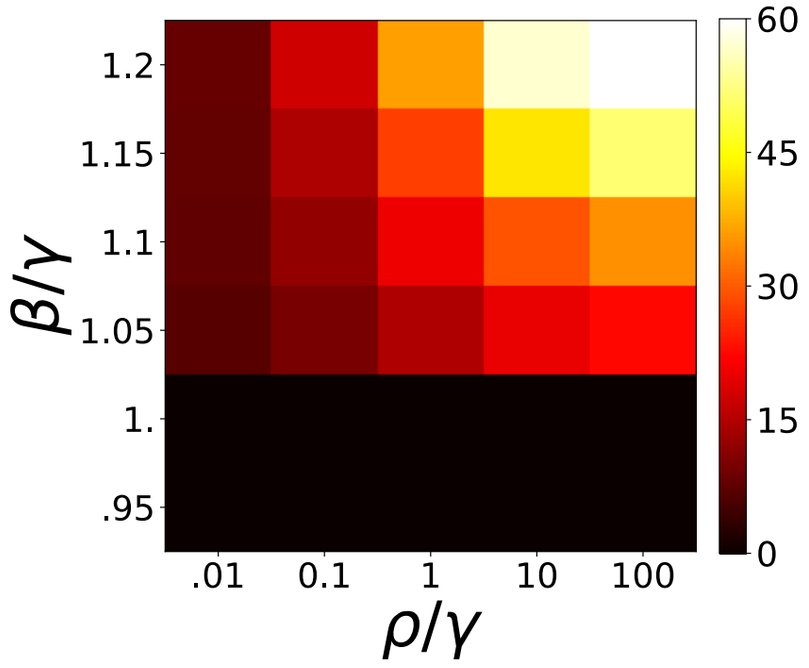


Means and Fluctuations

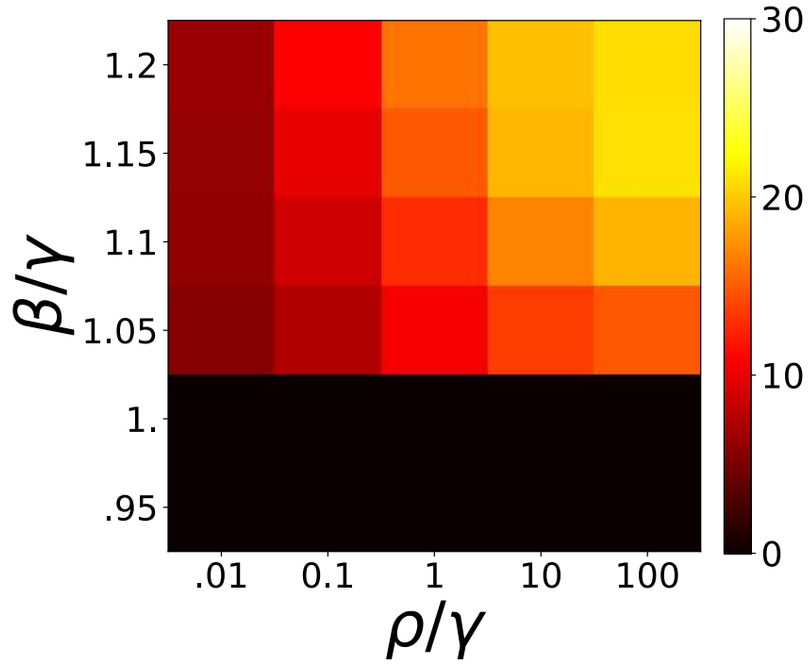
Simulations

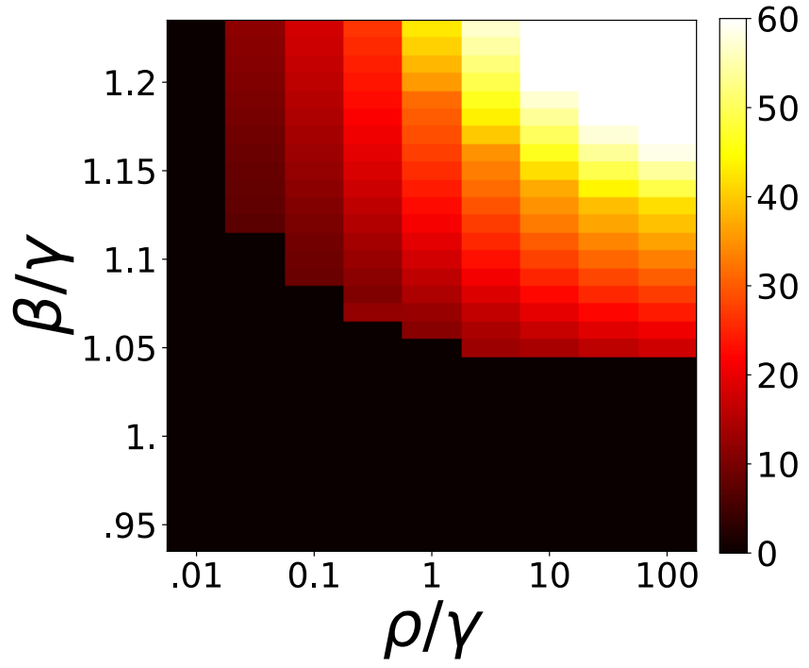
Cumulant Equations

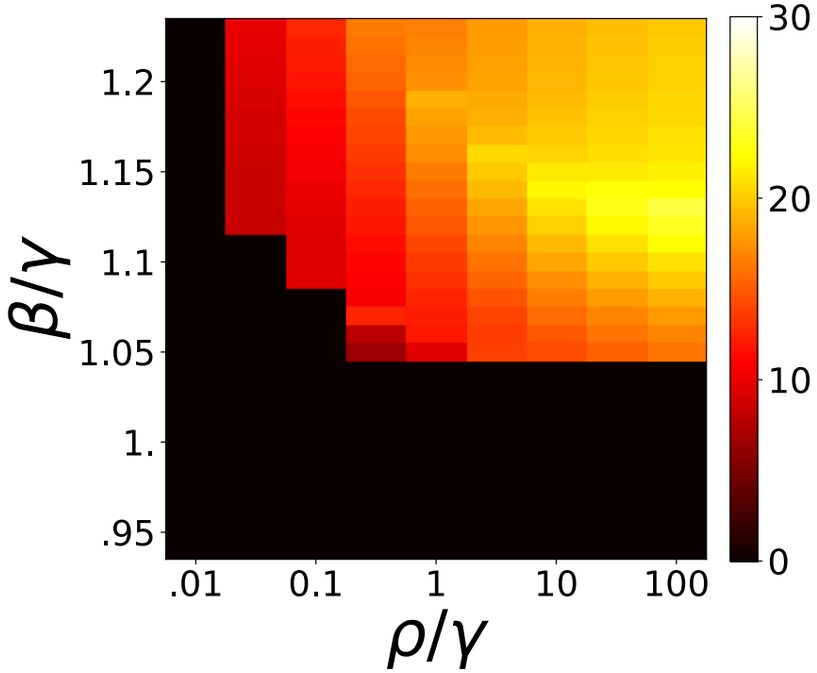
Mean Infected $\mu_y(t)$



Fluctuation $\sigma_y(t)$

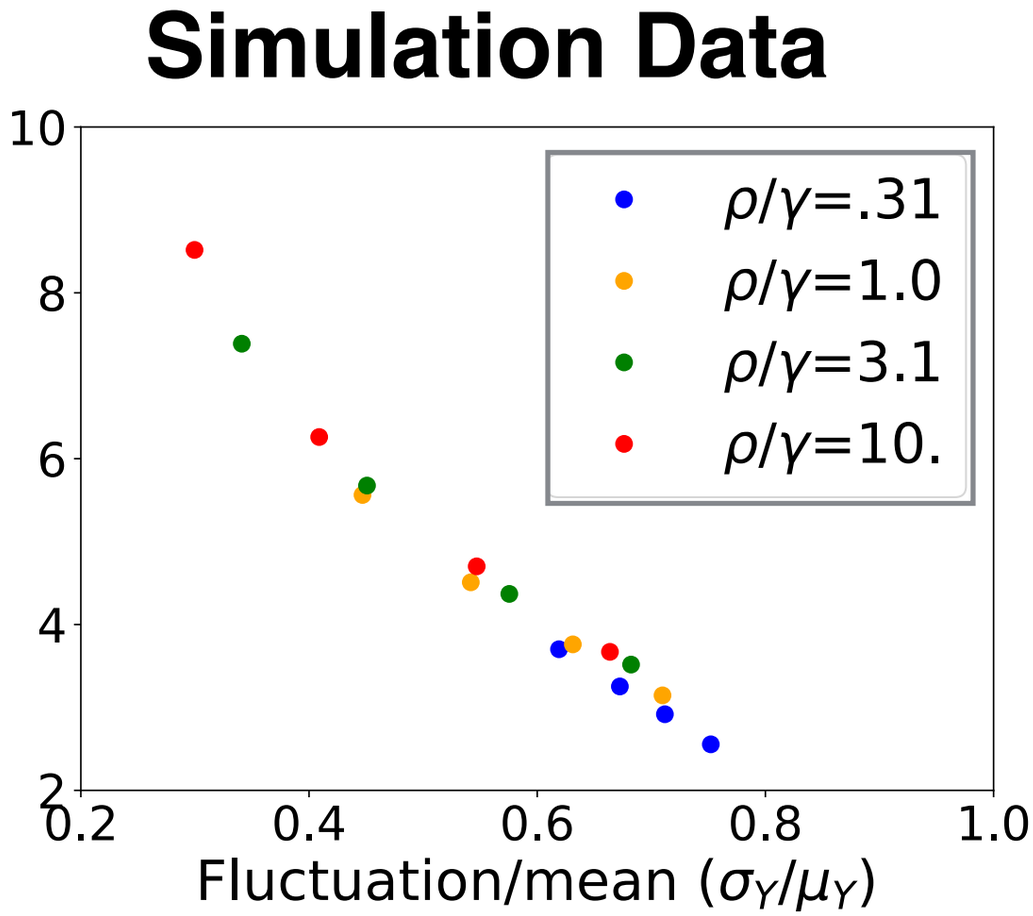






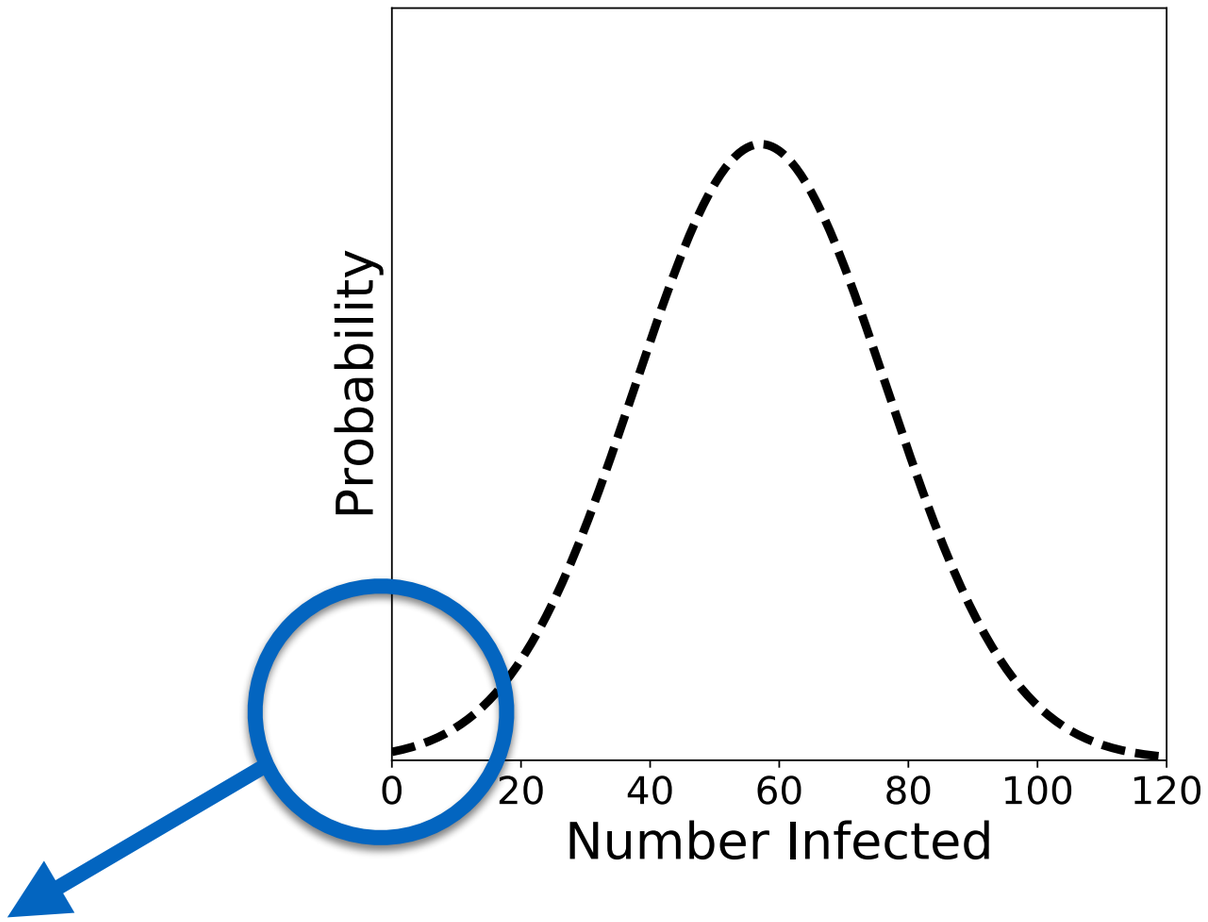
Endemic State Lifetime

Mean lifetime
log(**τ**)



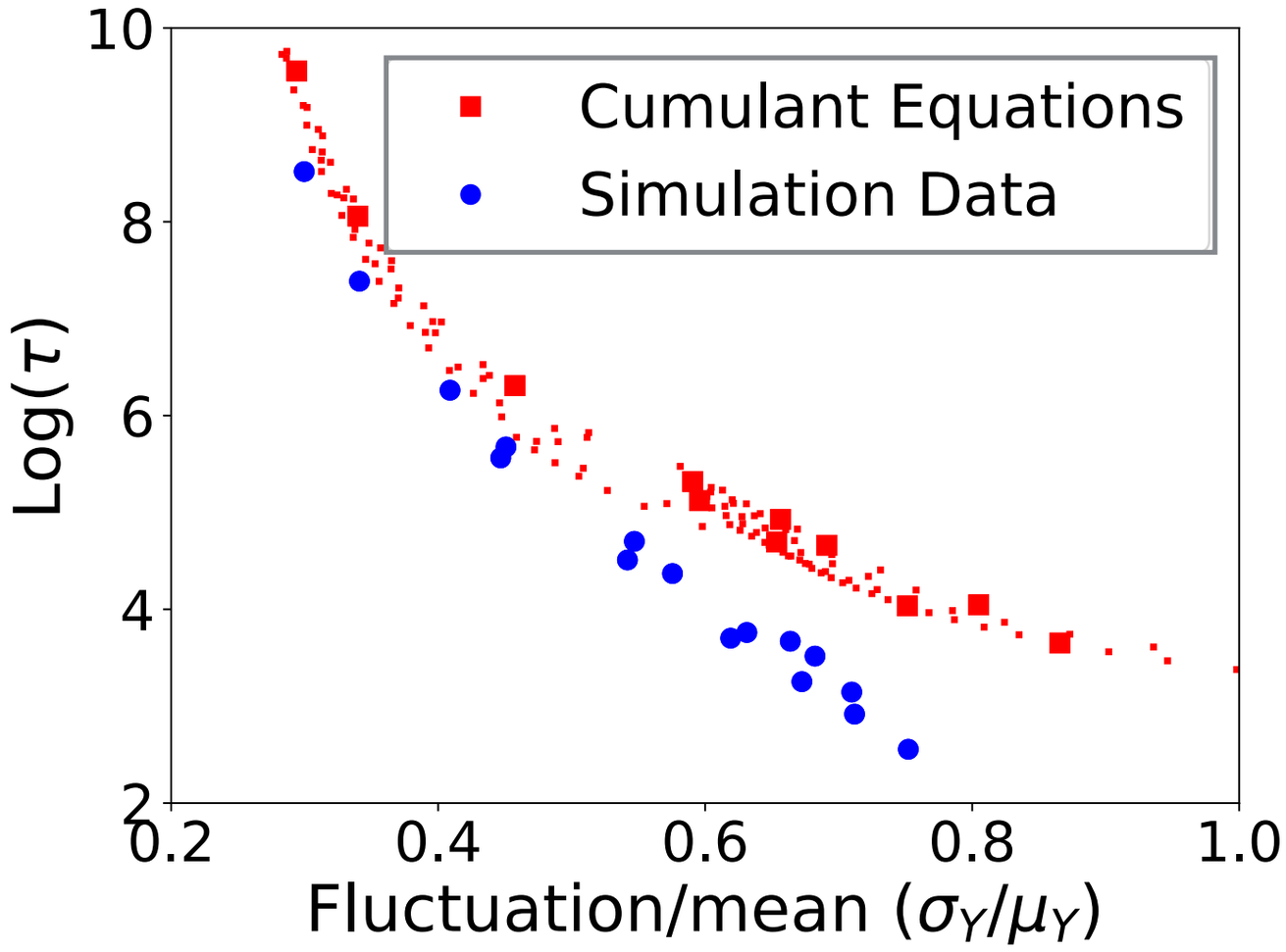
- Trajectories die out quickly for large fluctuations $\boldsymbol{\sigma_y/\mu_y}$
- Data collapse onto a low-dimensional curve
- Not sensitive to varying **ρ**

Approximate Time to Extinction



- **P(n=1)** ~ rate at which trajectories go extinct
- Can estimate using Gaussian approximation with (μ_y, σ_y)
- Naive, since Gaussian is most accurate near the mean
- Mean time to extinction $\%= ("\,P\,(n = 1))^{-1}$

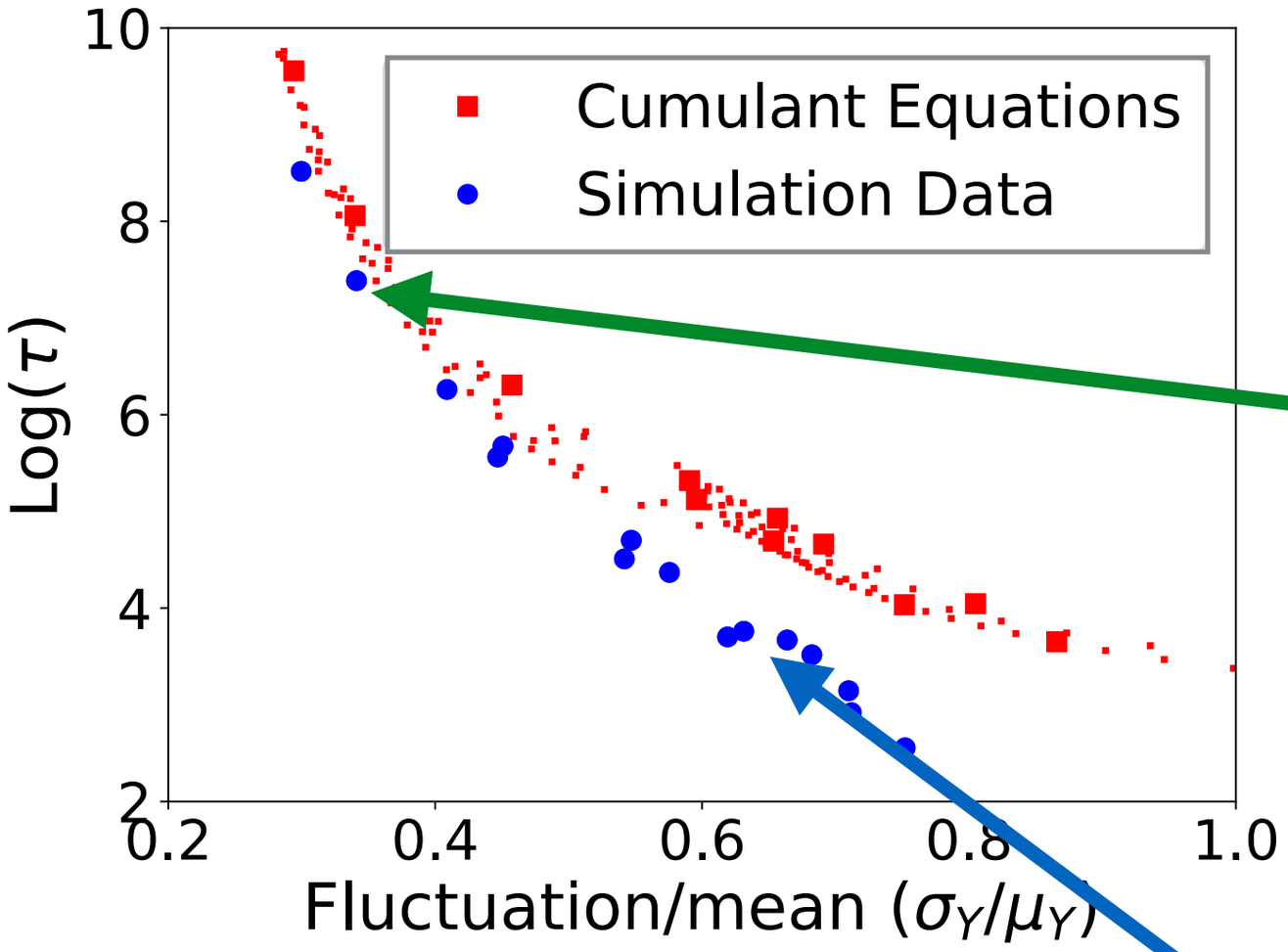
Times to Extinction

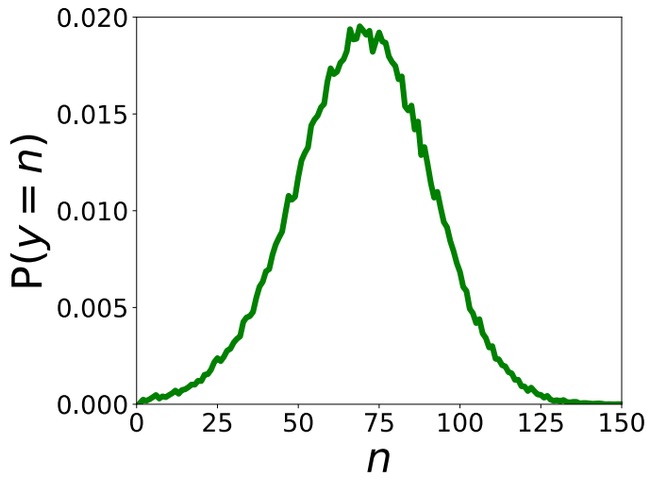


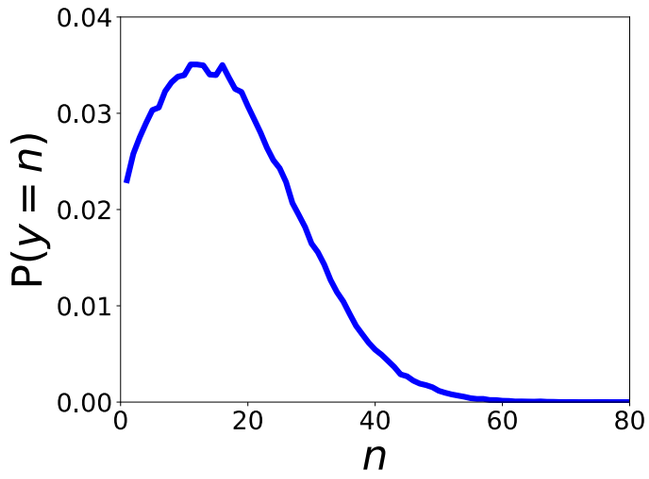
- Cumulant equations: predicted **τ** vs **σ_y/μ_y** also collapse onto a low-dimensional curve

- Qualitatively consistent: Small **τ** for **σ_y/μ_y**

Times to Extinction





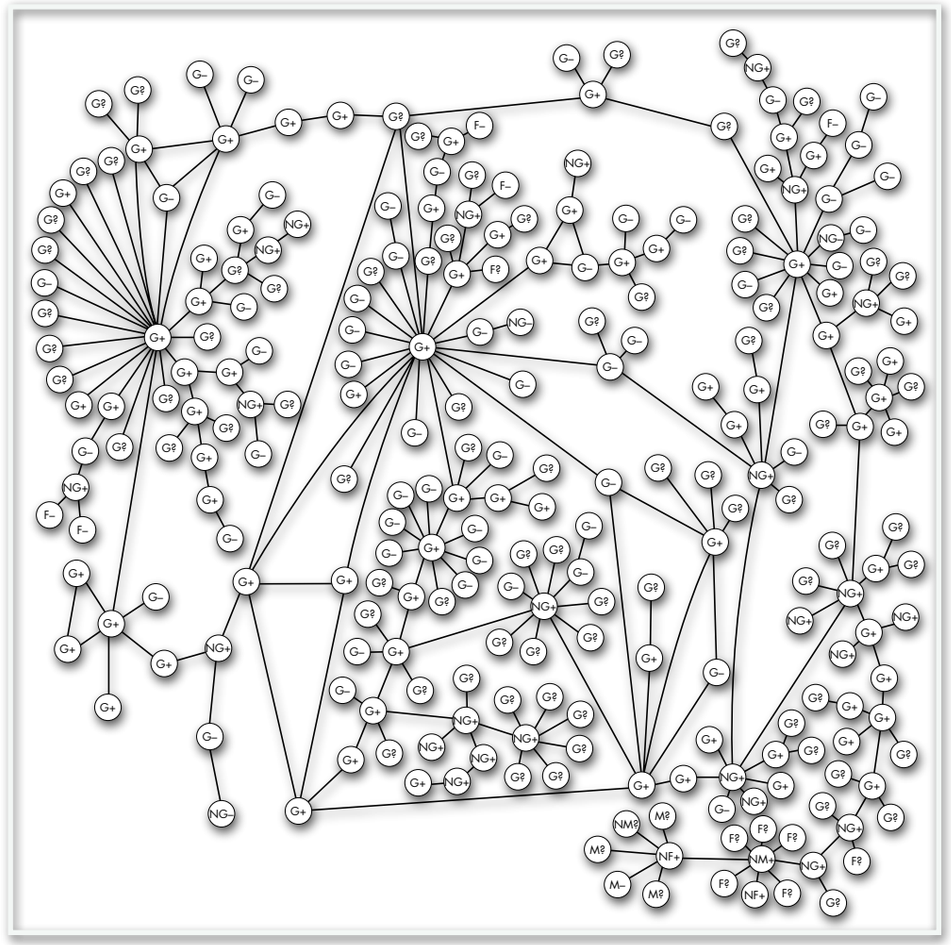


- Consistently overestimates **τ**
  - Better for small fluctuations
  - Worse for large fluctuations

Population Heterogeneity

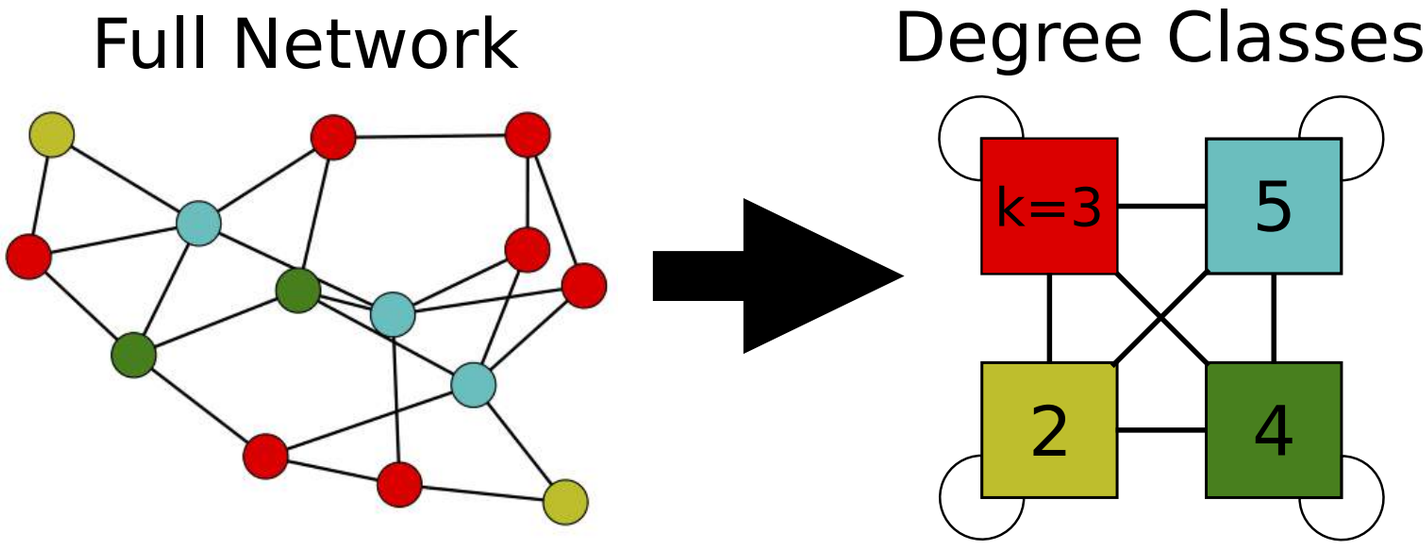
- Everything previous: homogeneous population
- What happens when not all members of a population are the same?
- Contact heterogeneity

Risk Network for Transmitting HIV



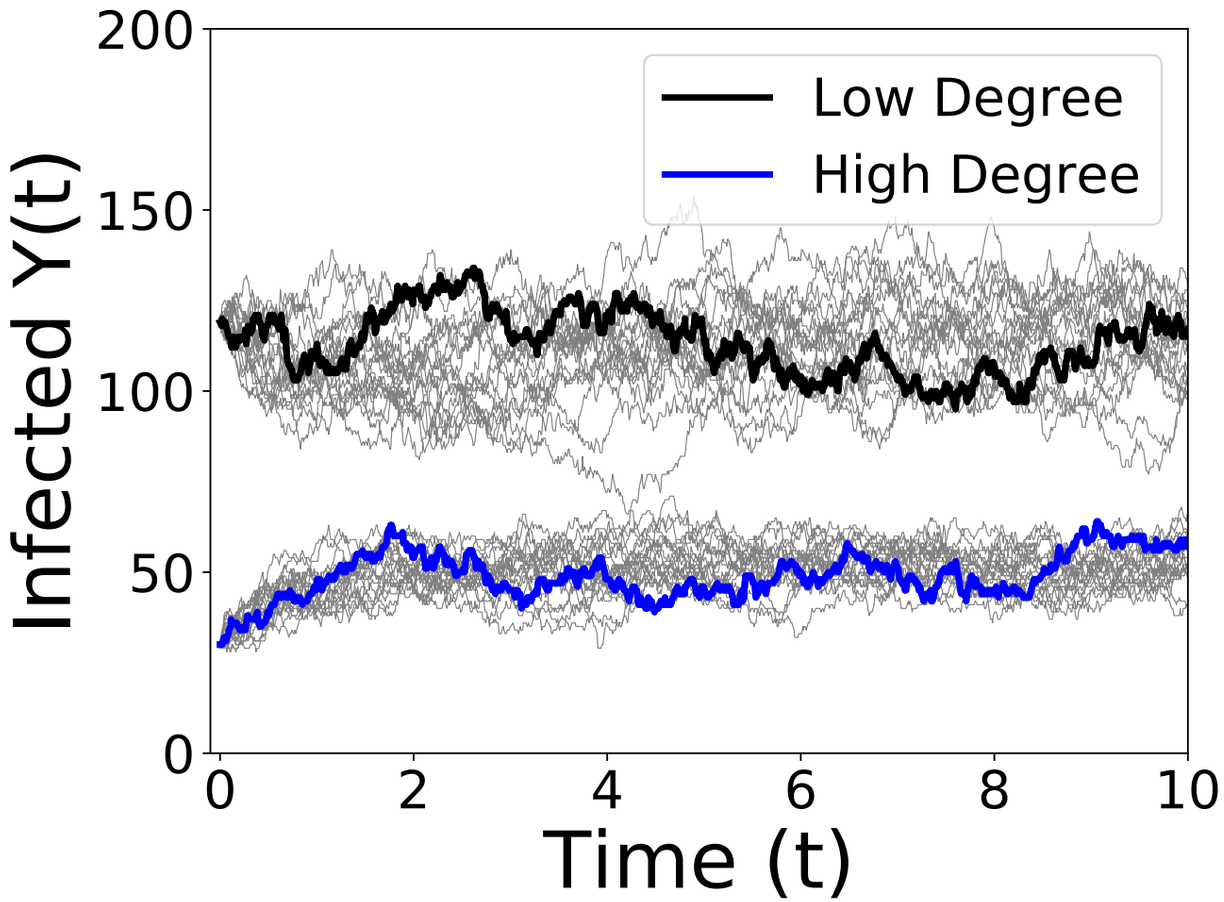
Potterat, *Sexually Transmitted Infections* (2002)

Heterogeneous Transmission



- Degree (k) = number of contacts
  - Higher degree - higher risk, transmission rate
- Rough approximation of network: compartmentalize by degree to reflect heterogeneity
- “Heterogeneous mean field”
  - Transmission proportional to node degree

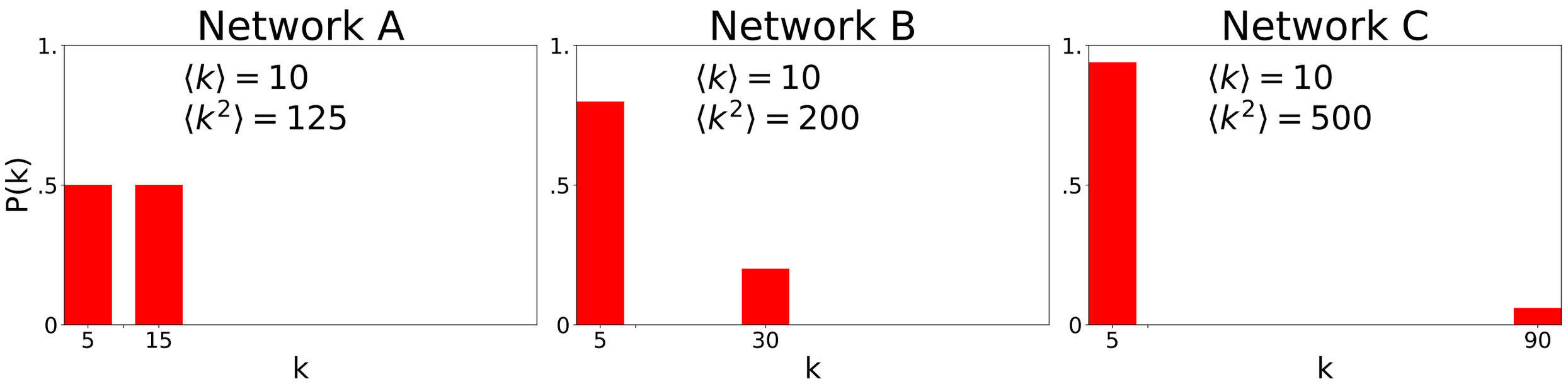
Cumulant Equations for Degree Classes



- Tracking each degree class separately
- For each degree class: $\boldsymbol{\mu_y(t)}$ and $\boldsymbol{\sigma_y(t)}$

Varying Heterogeneity

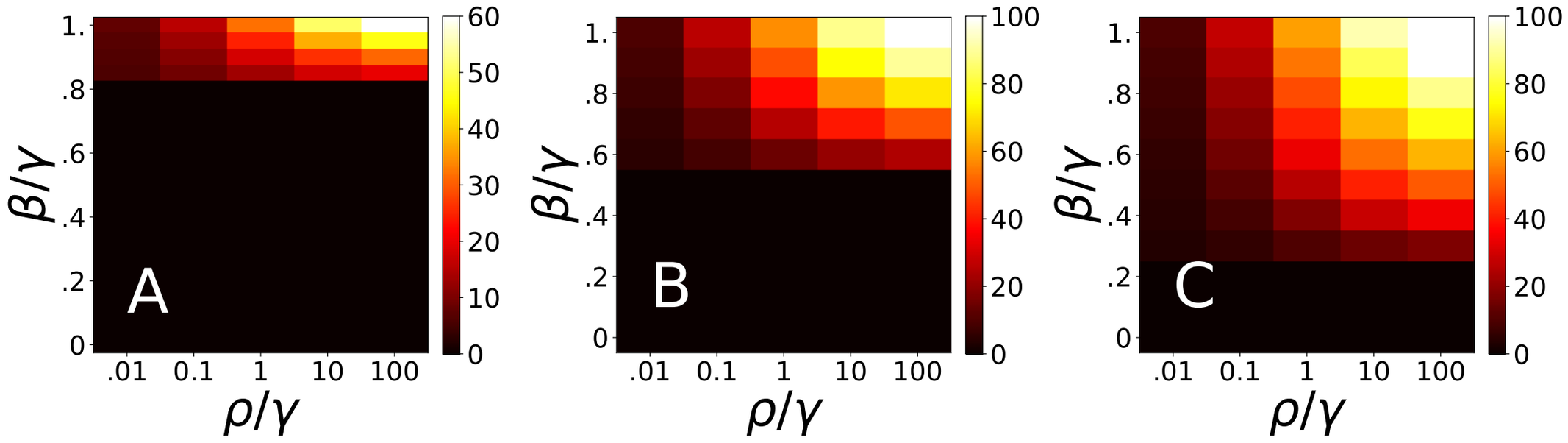
Homogeneous → Heterogeneous



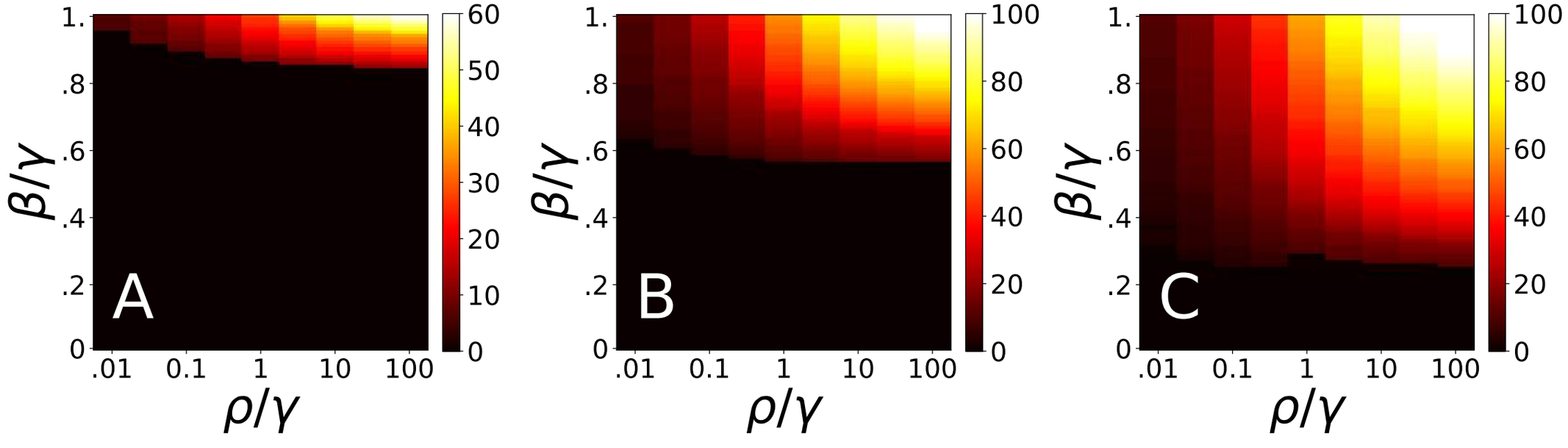
- Vary heterogeneity across populations
- 3 annealed networks with different degree distributions
  - Two degree classes each
  - Heterogeneity = Variability in degree distribution **$\langle k^2 \rangle$**

Mean Number Infected

Simulations



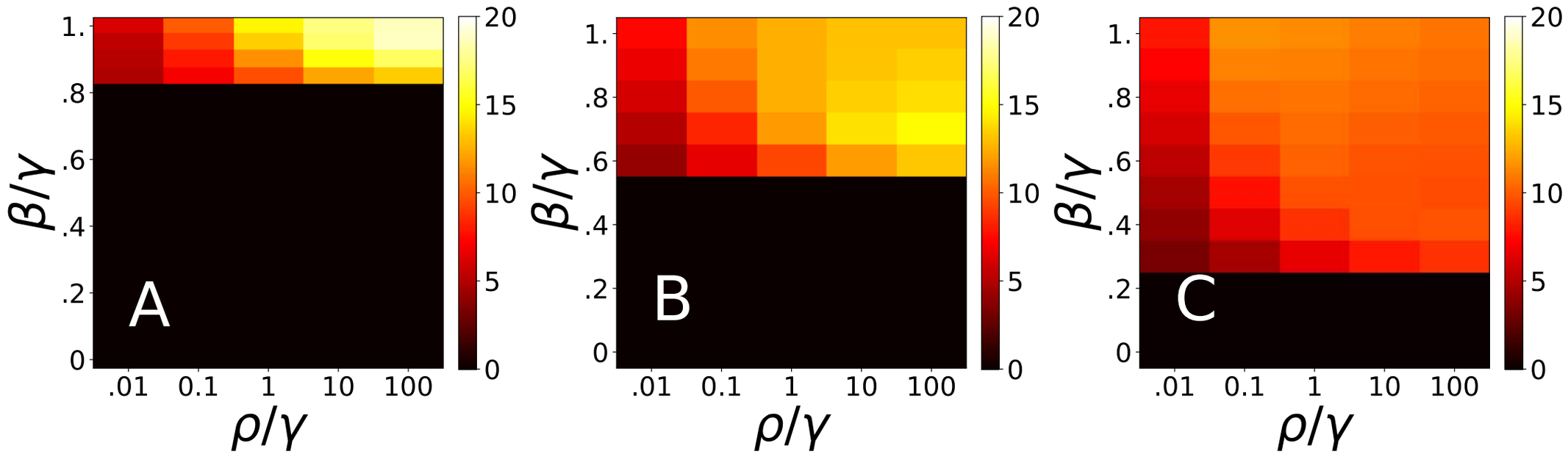
Cumulant Equations



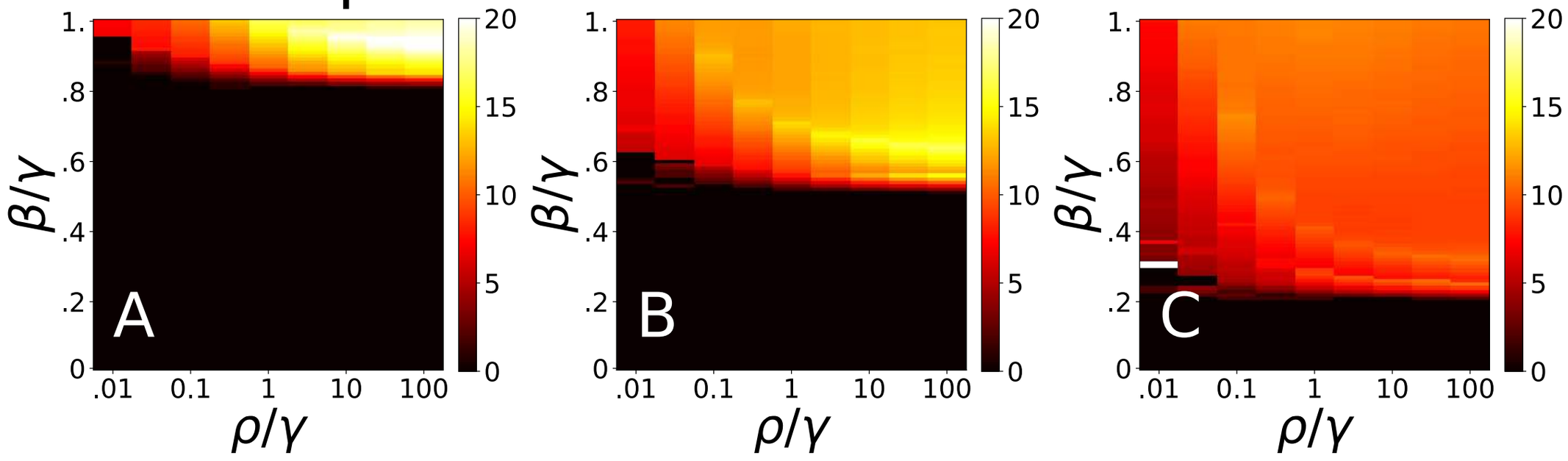
Homogeneous → Heterogeneous

Fluctuation sizes

Simulations



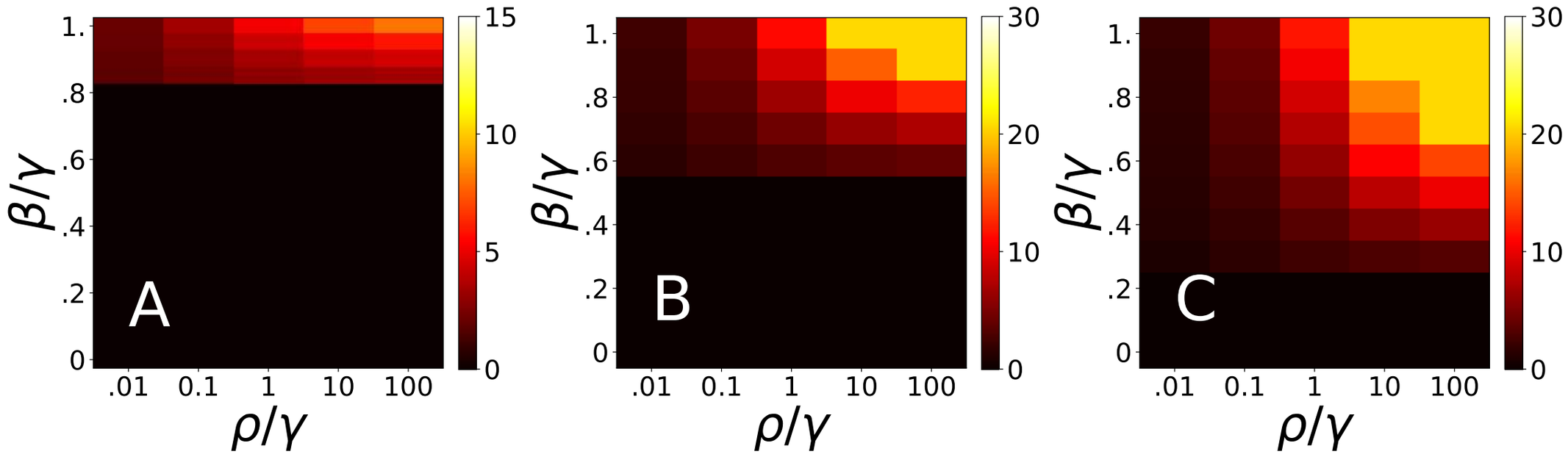
Cumulant Equations



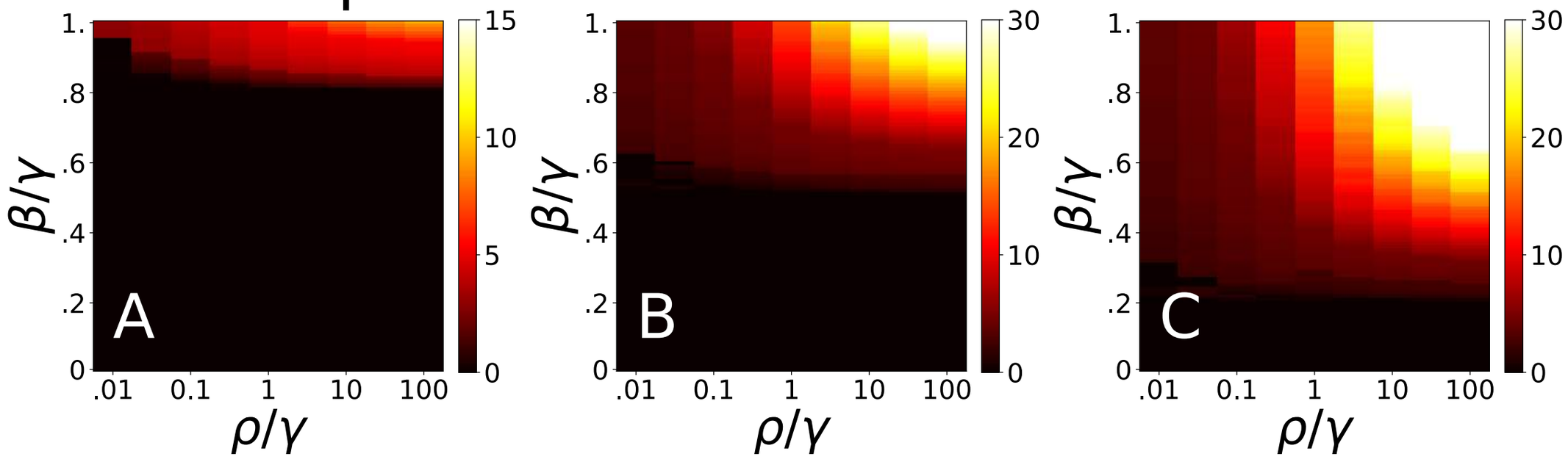
Homogeneous → Heterogeneous

Extinction Times

Simulations

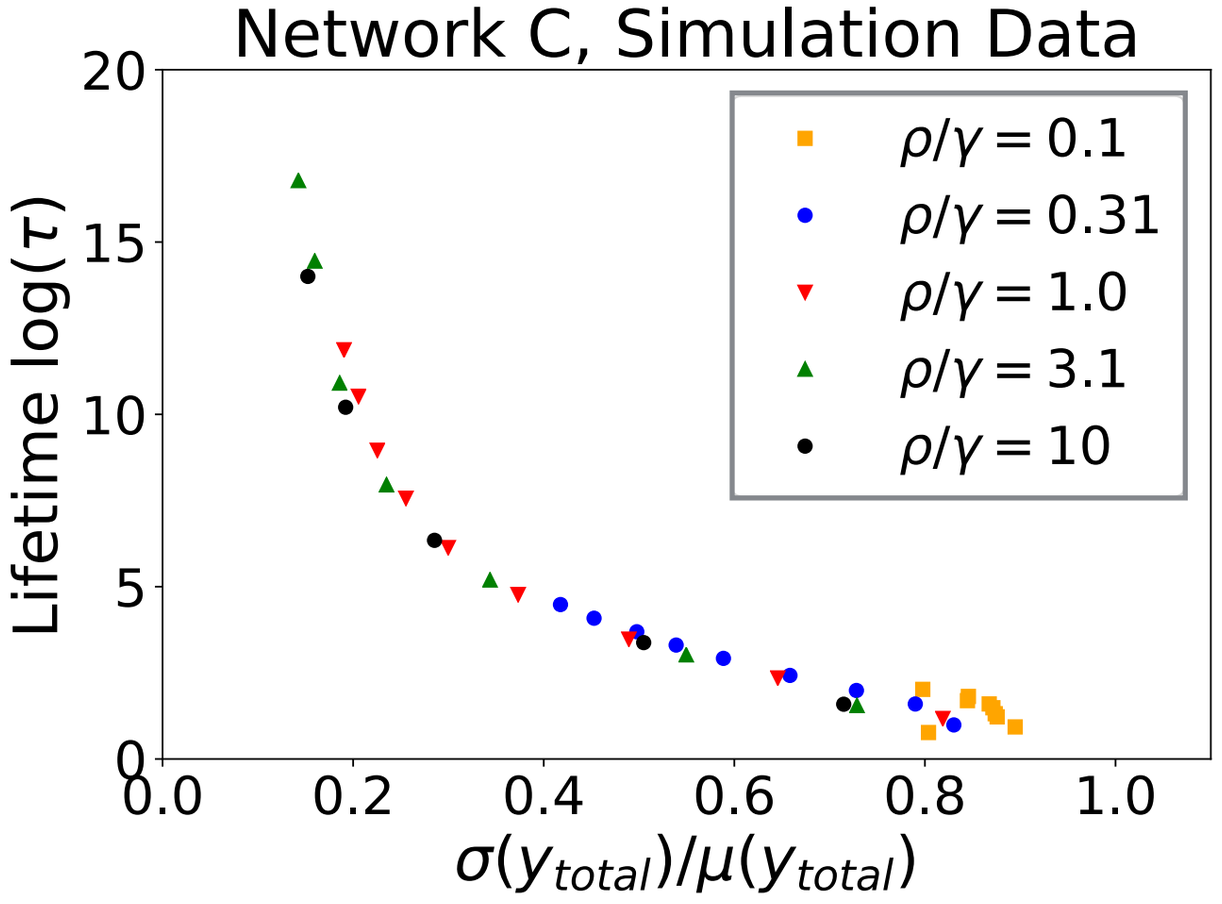


Cumulant Equations



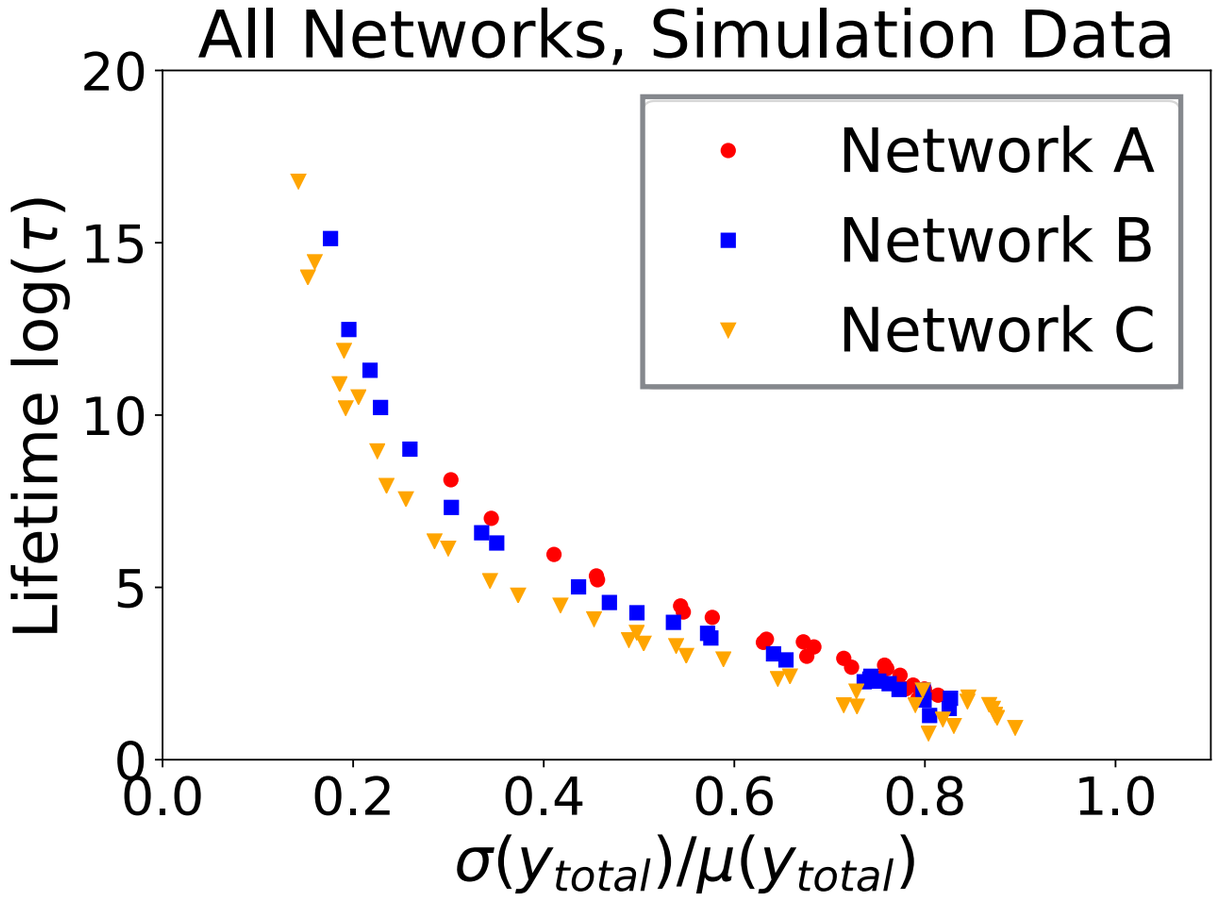
Homogeneous → Heterogeneous

Extinction Times



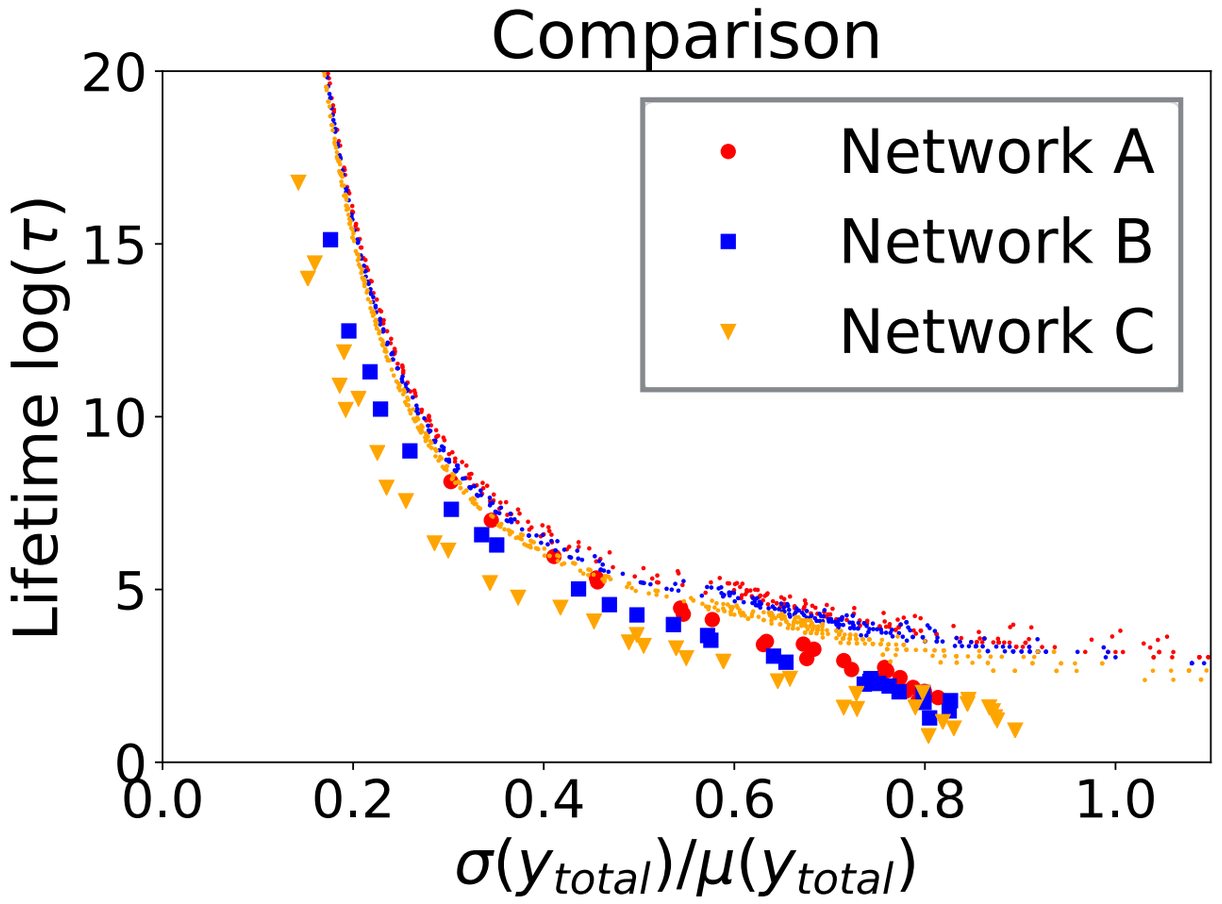
- Trajectories die out quickly for large fluctuations $\boldsymbol{\sigma_y/\mu_y}$
- Data collapse onto a low-dimensional curve

Comparing Across Networks



- Varying heterogeneity
  - No qualitative effect on **τ** vs $\boldsymbol{\sigma_y/\mu_y}$
  - Small offset: heterogeneous graphs have shorter **τ**

Cumulant Equations



- Same qualitative relationship in **τ** vs $\boldsymbol{\sigma_y/\mu_y}$
- Predicted values of **τ** are worse for heterogeneous graphs

Time

High Degree Nodes Infected

$[t_{ext} - 9.6, t_{ext} - 7.2]$ $[t_{ext} - 7.2, t_{ext} - 4.8]$ $[t_{ext} - 4.8, t_{ext} - 2.4]$ $[t_{ext} - 2.4, t_{ext}]$

A

Number Infected, High Degree

$[t_{ext} - 3.2, t_{ext} - 2.4]$ $[t_{ext} - 2.4, t_{ext} - 1.6]$ $[t_{ext} - 1.6, t_{ext} - 0.8]$ $[t_{ext} - 0.8, t_{ext}]$

B

Number Infected, High Degree

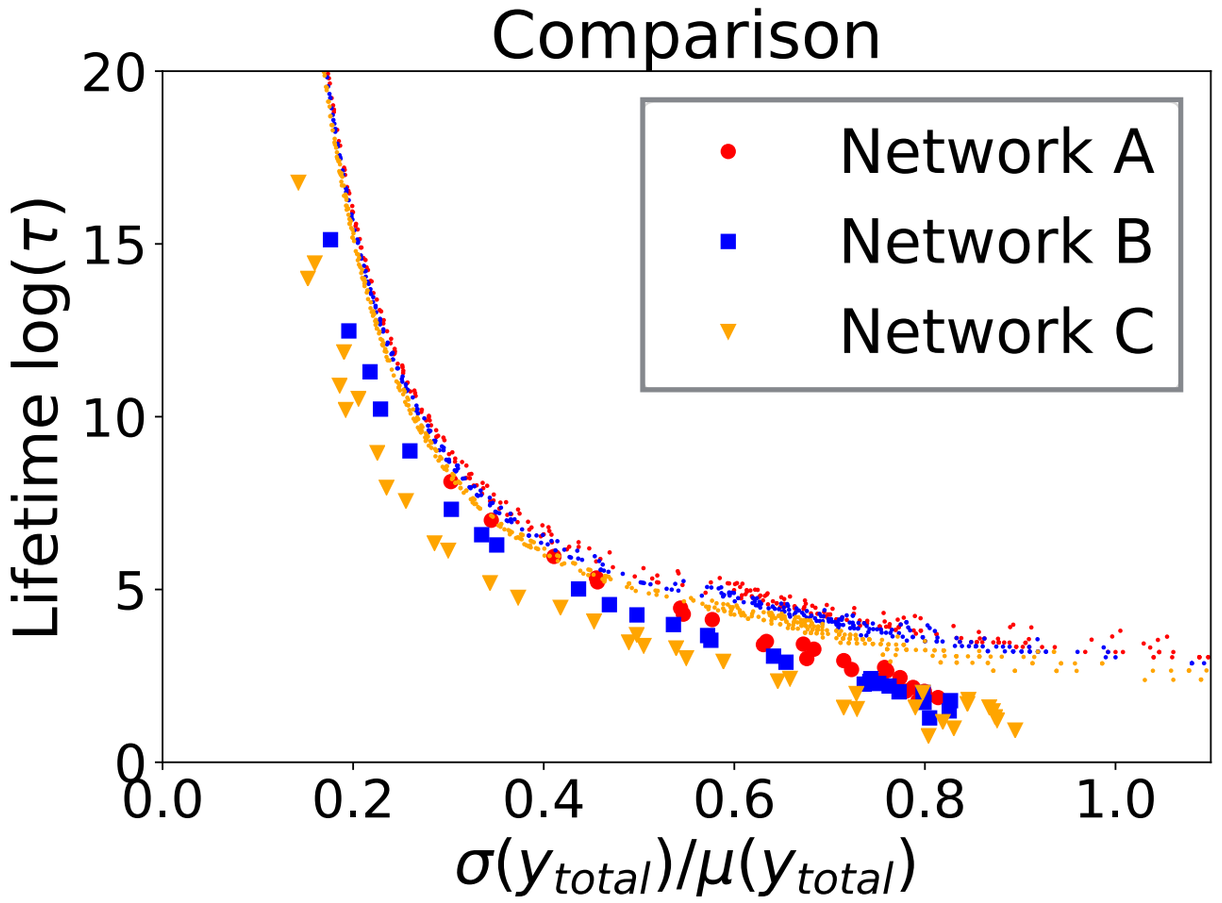
$[t_{ext} - 9.6, t_{ext} - 7.2]$ $[t_{ext} - 7.2, t_{ext} - 4.8]$ $[t_{ext} - 4.8, t_{ext} - 2.4]$ $[t_{ext} - 2.4, t_{ext}]$

C

Number Infected, Hight Degree

Low Degree Nodes Infected

Cumulant Equations



- High-degree and low-degree nodes play different roles on the path to extinction
- Naive prediction of $\boldsymbol{\tau}$ insufficient here

Summary

- Understand persistence by understanding statistical properties of the endemic state
- Cumulant equations used to approximate trajectory ensemble
- Low-dimensional relationship between τ vs σ_y/μ_y

Network Effects

- Adding contact heterogeneity:
  - Smaller fluctuations
  - Shorter endemic lifetimes
- Does not change the relationship between τ vs. σ_y/μ_y

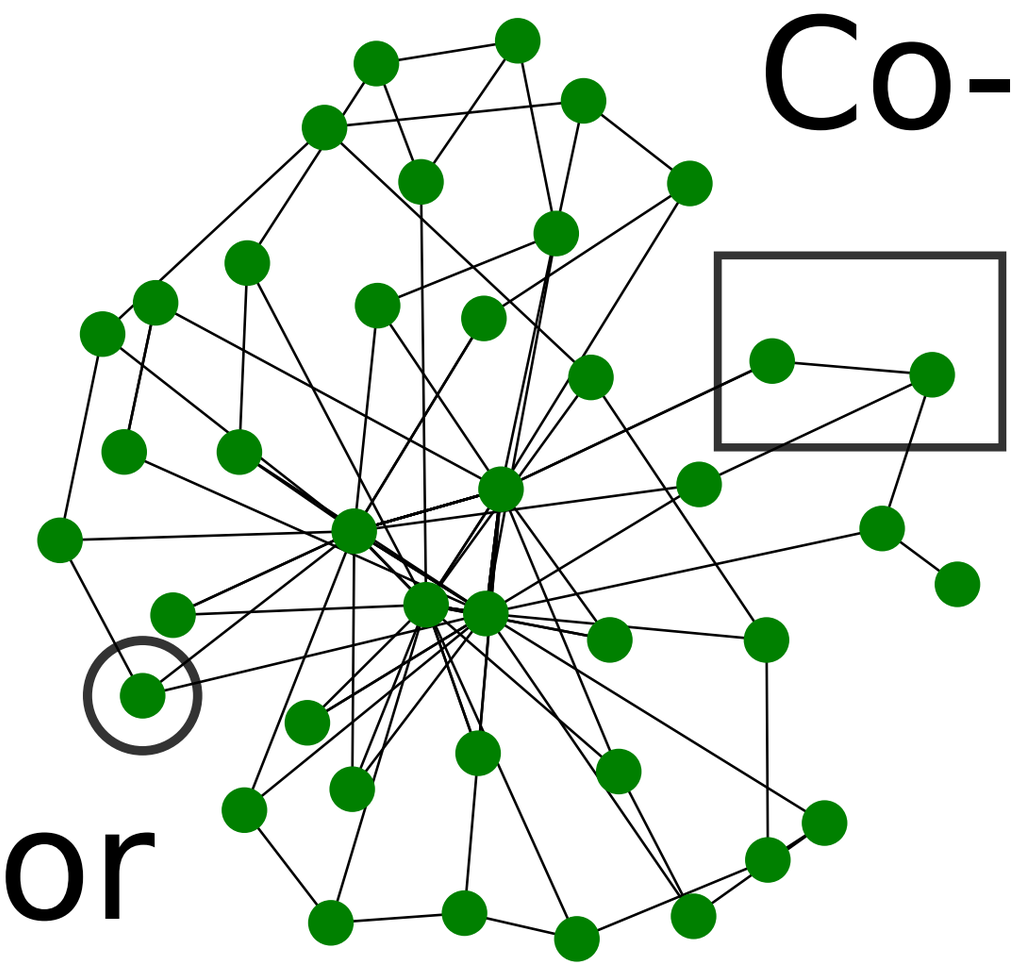
Network Assembly of Scientific Communities

Measuring Scientific Communities

- Scientific fields are about relationships
  - Transfer of ideas and skills
  - Cooperation to answer complicated questions
- Sociological standpoint:
  - How do scientific communities operate?
  - When do they succeed or fail?

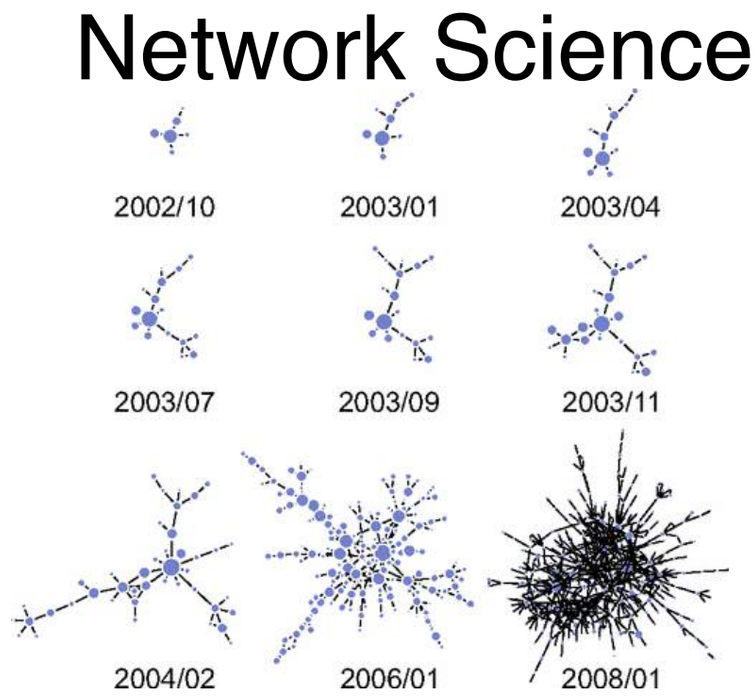
Co-Authorship Networks

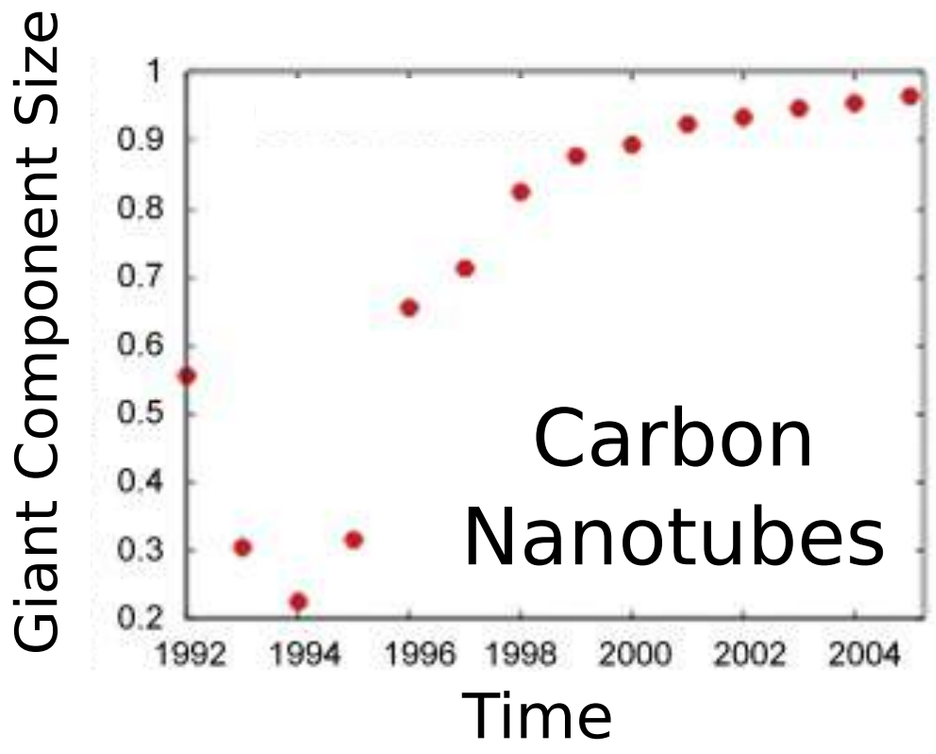
- Empirical representation of relationships
- Can measure structural properties



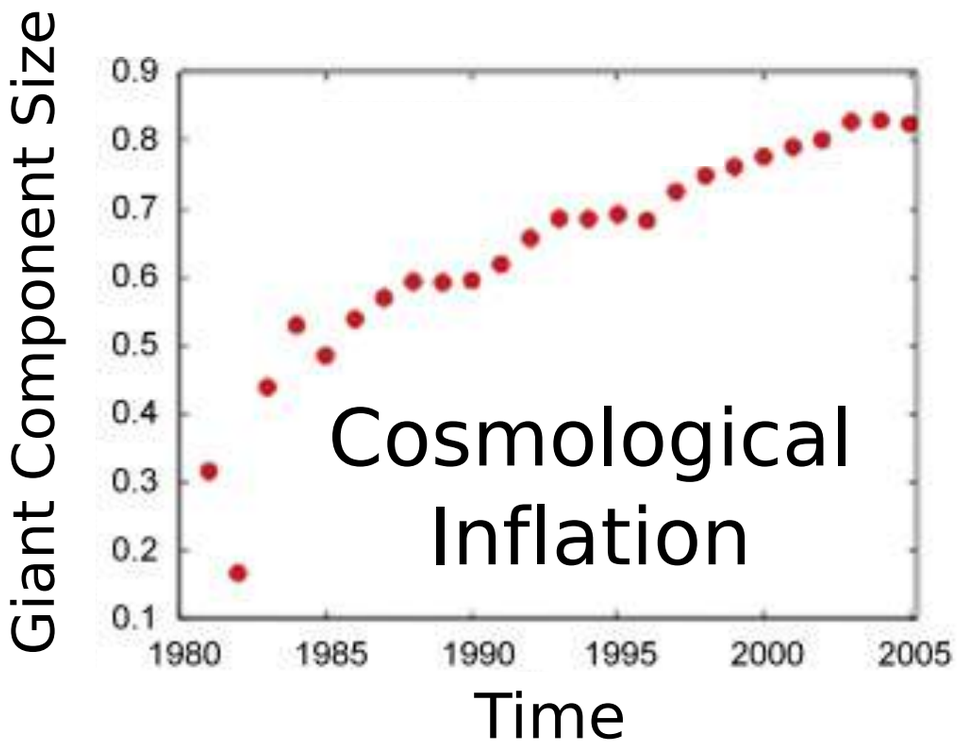
Previous Work

- Co-authorship networks in specific subfields
- Community coalesces
- Forms dense giant component





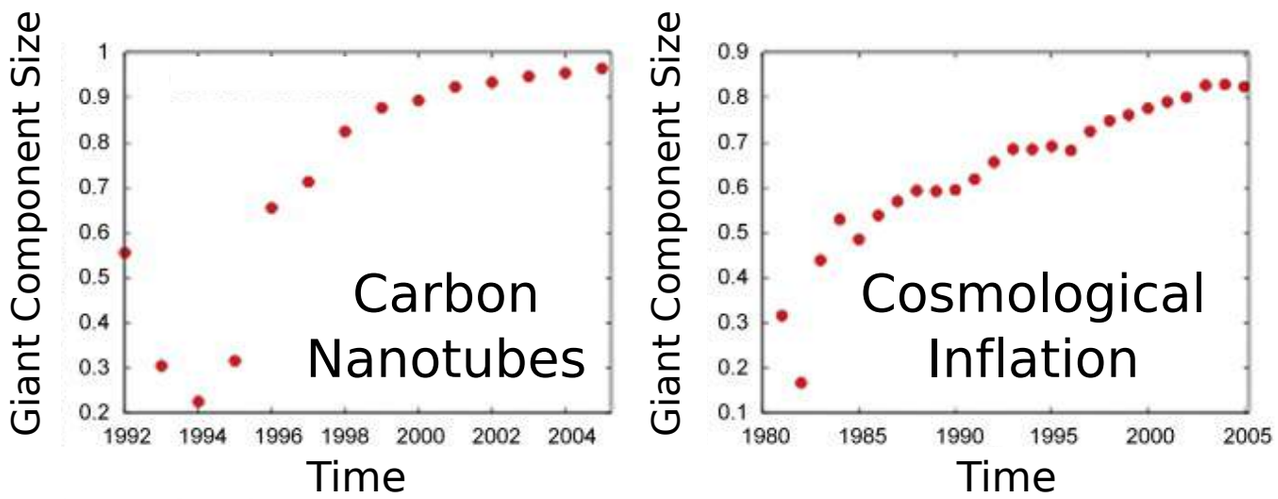
Bettencourt et al., J of Informetrics (2009)



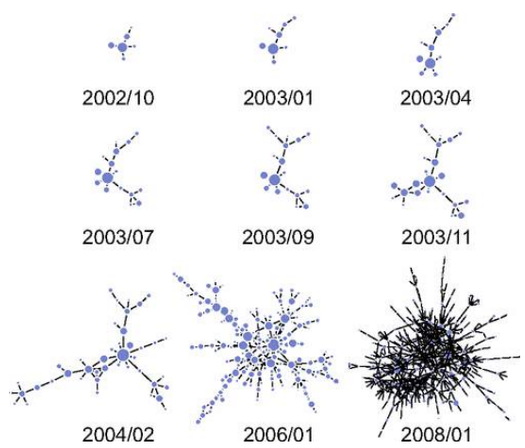
Lee et al., Phys Rev E (2010)

Previous Work Limitations

- Data curated and sorted by experts
- Limits number of fields surveyed
- Hard to compare across many different fields
- Difficult to argue that the patterns are general



Bettencourt et al., J of Informetrics (2009)



Lee et al., Phys Rev E (2010)

Build upon Previous Studies

- Comparing co-authorship networks across a large variety of scientific topics
- Use large publication data set: the arXiv
- Algorithmic generation of population of topics
- See whether topological transition is general

The Data Set

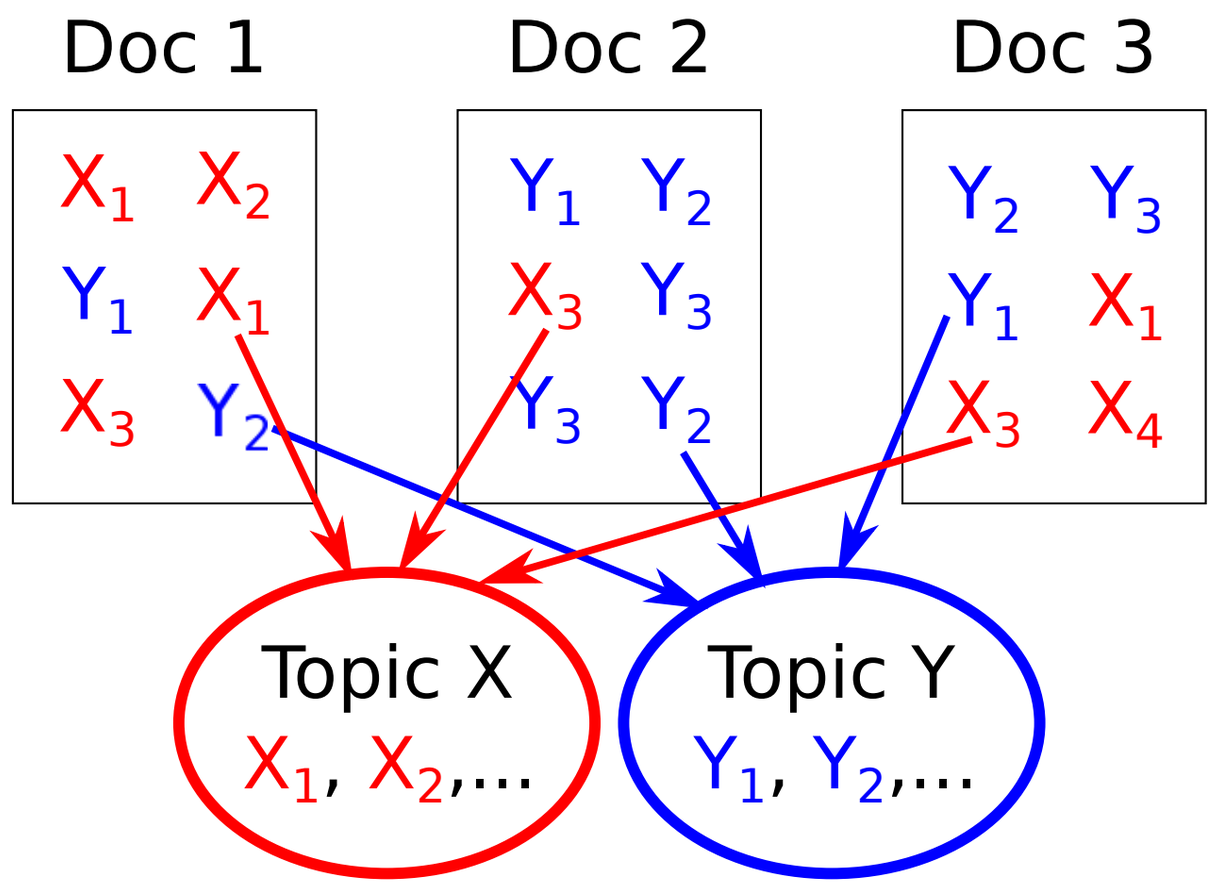
- The arXiv - preprint repository at www.arXiv.org
- Focus on Condensed Matter (cond-mat)
- Time period: April 1992 through June 2015
- 189,000 articles
- 95,000 authors

Topic Modeling

- Documents contain terms
- Assume that document set contains common topics
- Simultaneously:
  - Discover topics, based on document co-occurrence of terms
  - Assign topics to each of the documents based on content

Infer Topics for Documents

Clustering Terms into Topics



Implementation

- Applying LDA Topic Modeling, using MALLET
- Document set: titles and abstracts from arXiv's cond-mat articles
- ***k*=50 Topics**
  - Sufficient for resolving coherent clusters of articles
- Find articles strongly associated with each topic
- Common terms proxy for common scientific content

Example Result: Topic 5

- Keywords:
  - quantum
  - state
  - qubit
  - entanglement
  - spin
  - decoherence
  - coupling
  - single
  - control
  - gate
  - coupled
  - information
- Examples Titles:
  - “Demonstration of Two-Qubit Algorithms with a Superconducting Quantum Processor” (0903.2030)
  - “Strategy for implementing stabilizer-based codes on solid-state qubits” (1301.4796)
  - “Quantum Zeno effect with a superconducting qubit” (1006.2133)

Example Result: Topic 5

- Keywords:
  - quantum
  - state
  - qubit
  - entanglement
  - spin
  - decoherence
  - coupling
  - single
  - control
  - gate
  - coupled
  - information
- Examples Titles:
  - “Demonstration of Two-Qubit Algorithms with a Superconducting Quantum Processor” (0903.2030)
  - “Strategy for implementing stabilizer-based codes on solid-state qubits” (1301.4796)
  - “Quantum Zeno effect with a superconducting qubit” (1006.2133)

“Quantum Computing”

Example Result: Topic 39

- Keywords:
  - band
  - surface
  - fermi
  - electronic
  - gap
  - electron
  - photoemission
  - spectroscopy
  - spectrum
  - density
  - hole
  - quasiparticle

- Examples Titles:
  - “Electronic Structure of the Topological Insulator Bi_2Se_3 Using Angle-Resolved Photoemission Spectroscopy” (1101.5615)
  - “Electronic structure of Co_xTiSe_2 and Cr_xTiSe_2” (cond-mat/0005153)
  - “Three-Dimensional Dirac Electrons at the Fermi Energy in Cubic Inverse Perovskites: Ca_3PbO and its Family” (1106.0477)

Example Result: Topic 39

- Keywords:
  - band
  - surface
  - fermi
  - electronic
  - gap
  - electron
  - photoemission
  - spectroscopy
  - spectrum
  - density
  - hole
  - quasiparticle

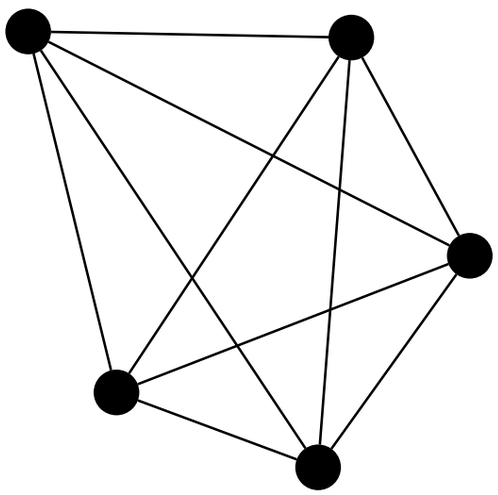
- Examples Titles:
  - “Electronic Structure of the Topological Insulator Bi_2Se_3 Using Angle-Resolved Photoemission Spectroscopy” (1101.5615)
  - “Electronic structure of Co_xTiSe_2 and Cr_xTiSe_2” (cond-mat/0005153)
  - “Three-Dimensional Dirac Electrons at the Fermi Energy in Cubic Inverse Perovskites: Ca_3PbO and its Family” (1106.0477)

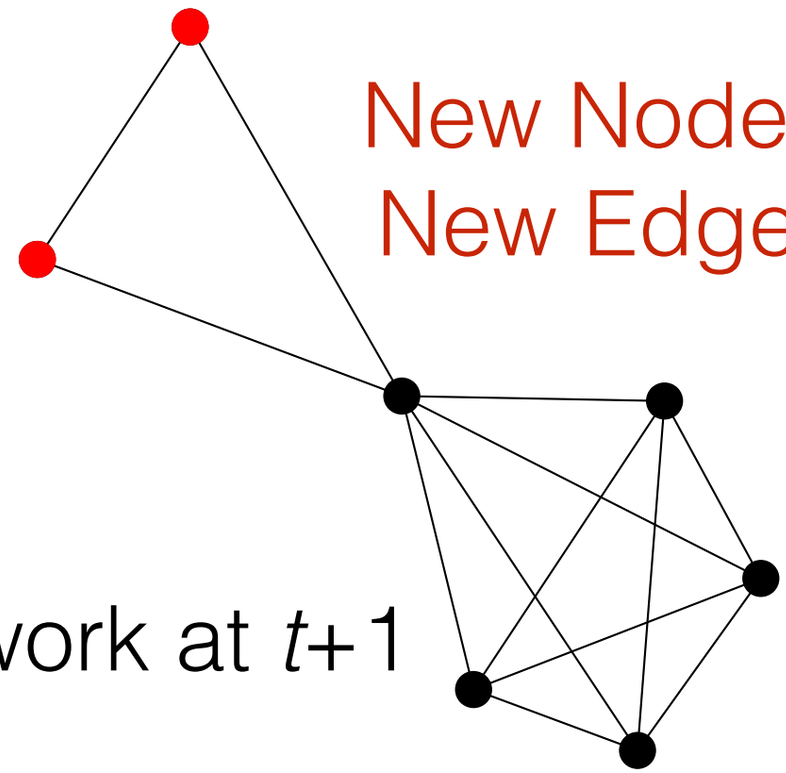
“Electronic Structure and Spectroscopy”

Constructing Co-Authorship Networks

- Each node is an author, each edge is co-authorship
- Networks grow over time
  - New nodes and edges are added
  - Time step = 1 month

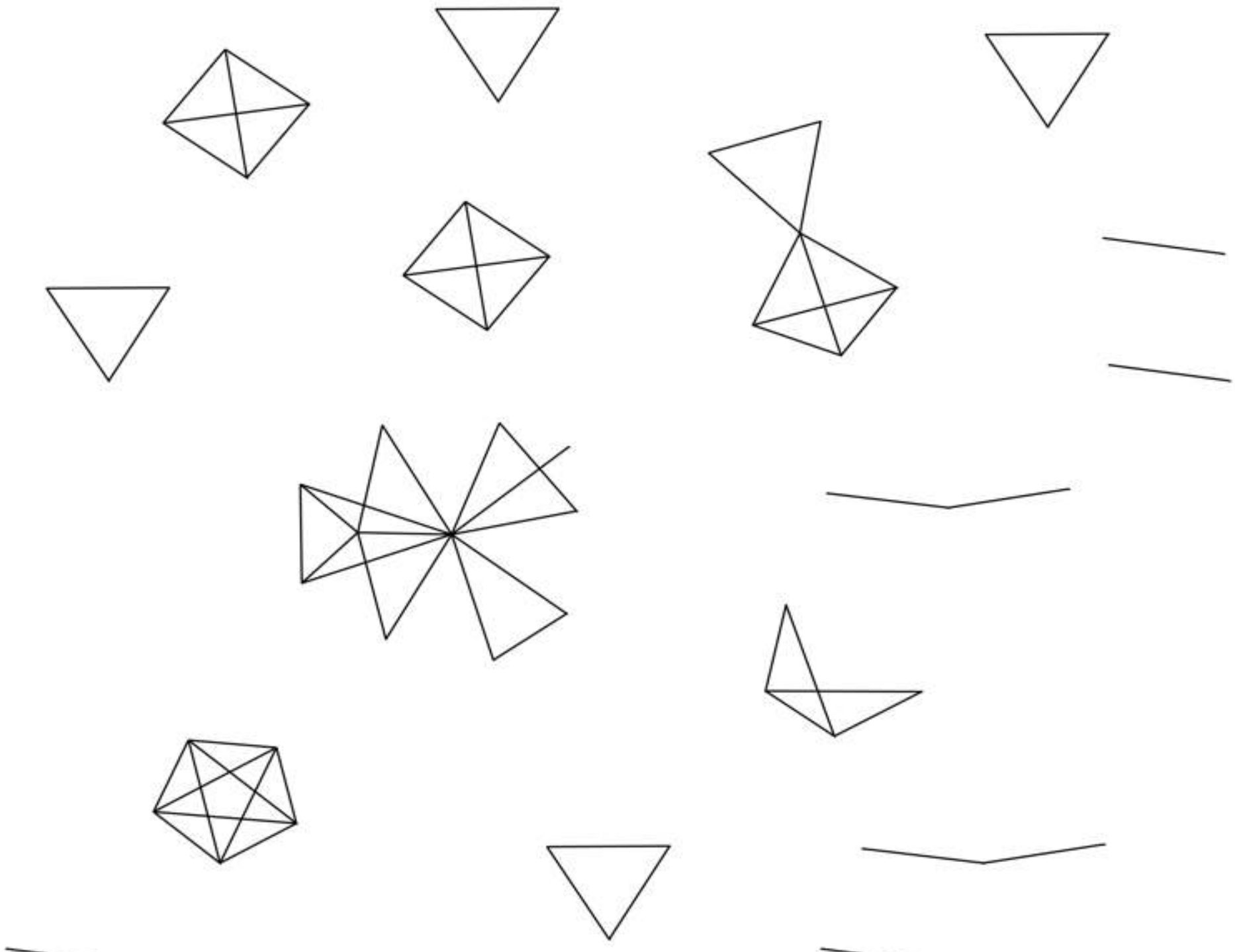
Network at *t*



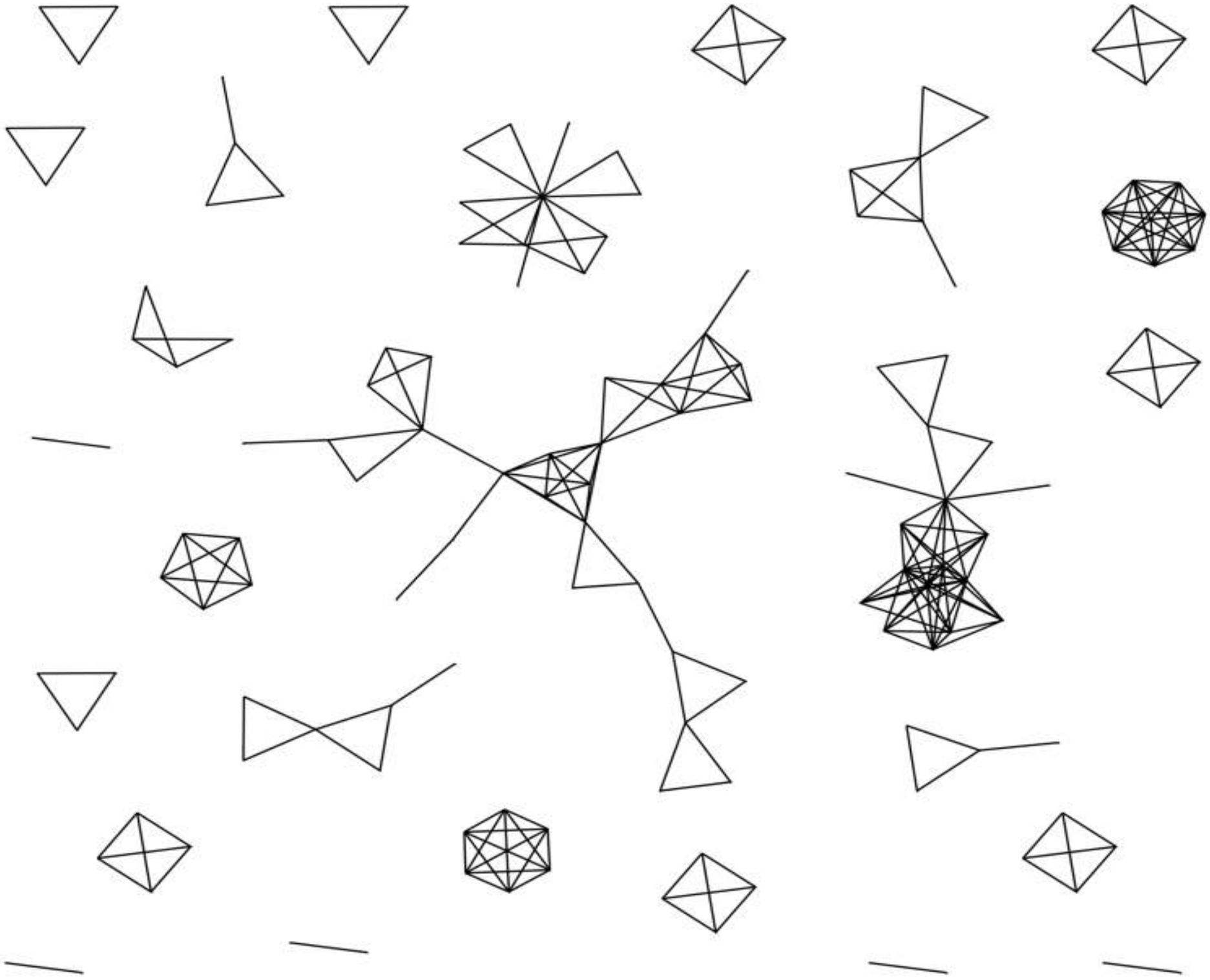


Network at *t*+1

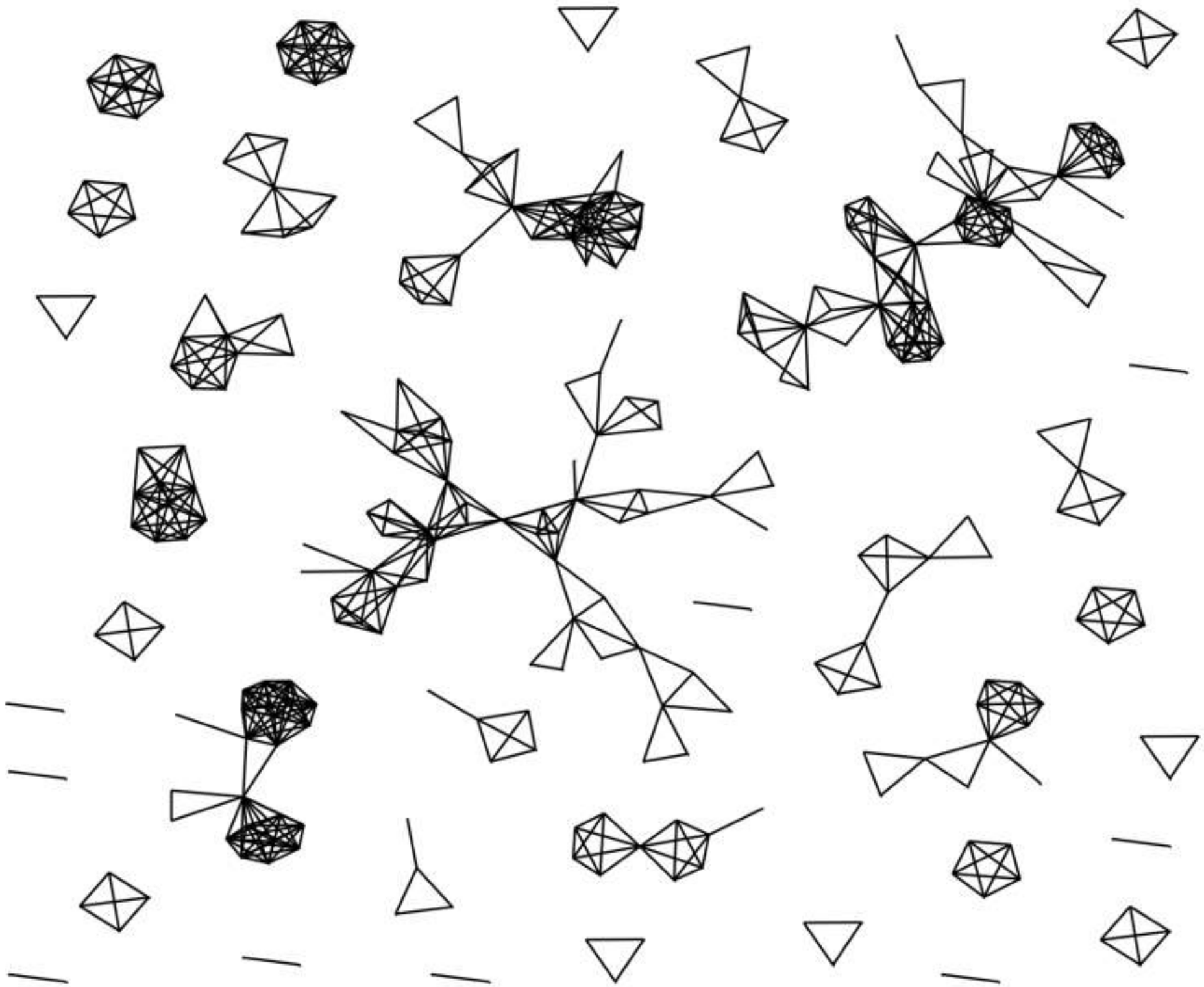
2001



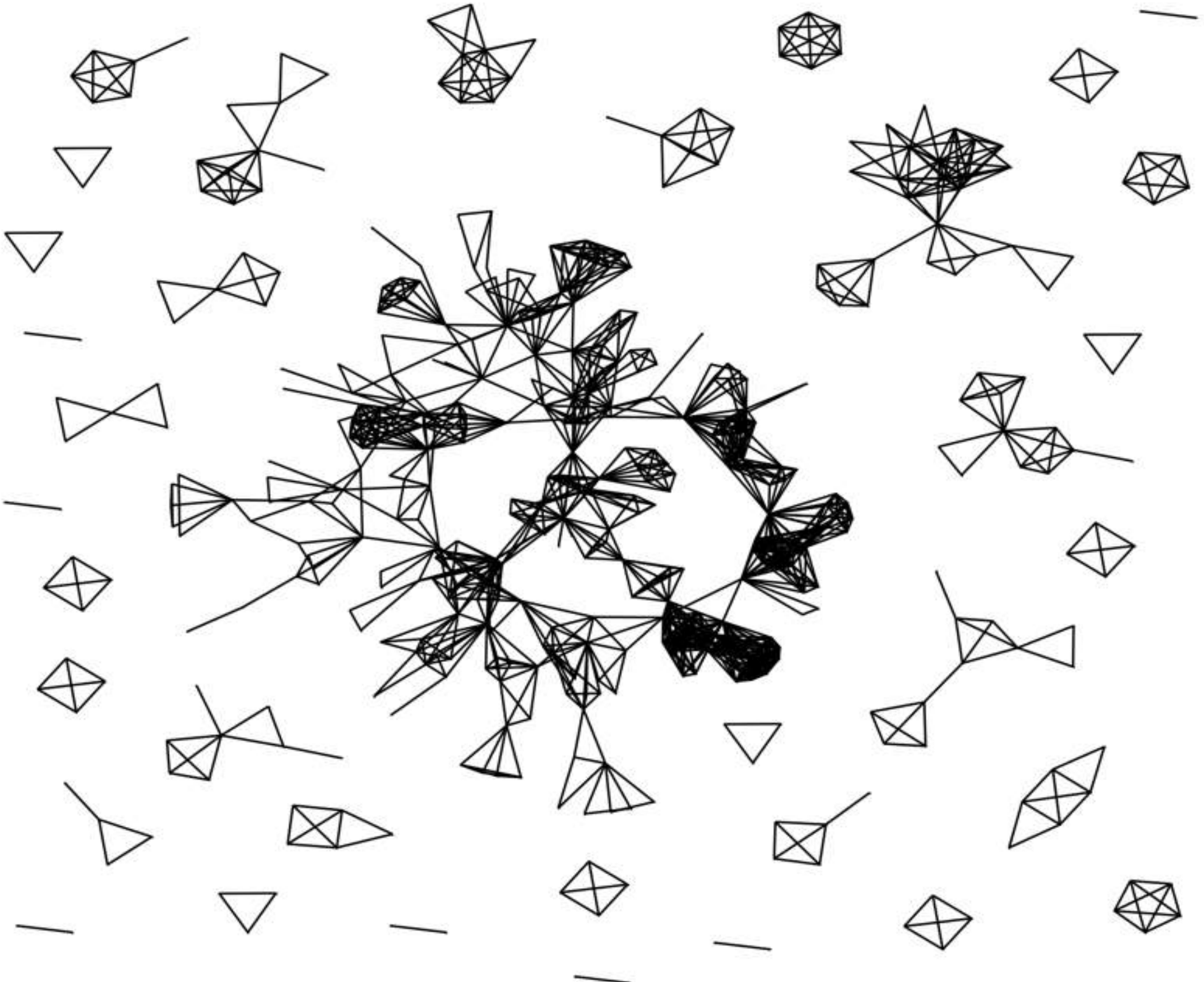
2003



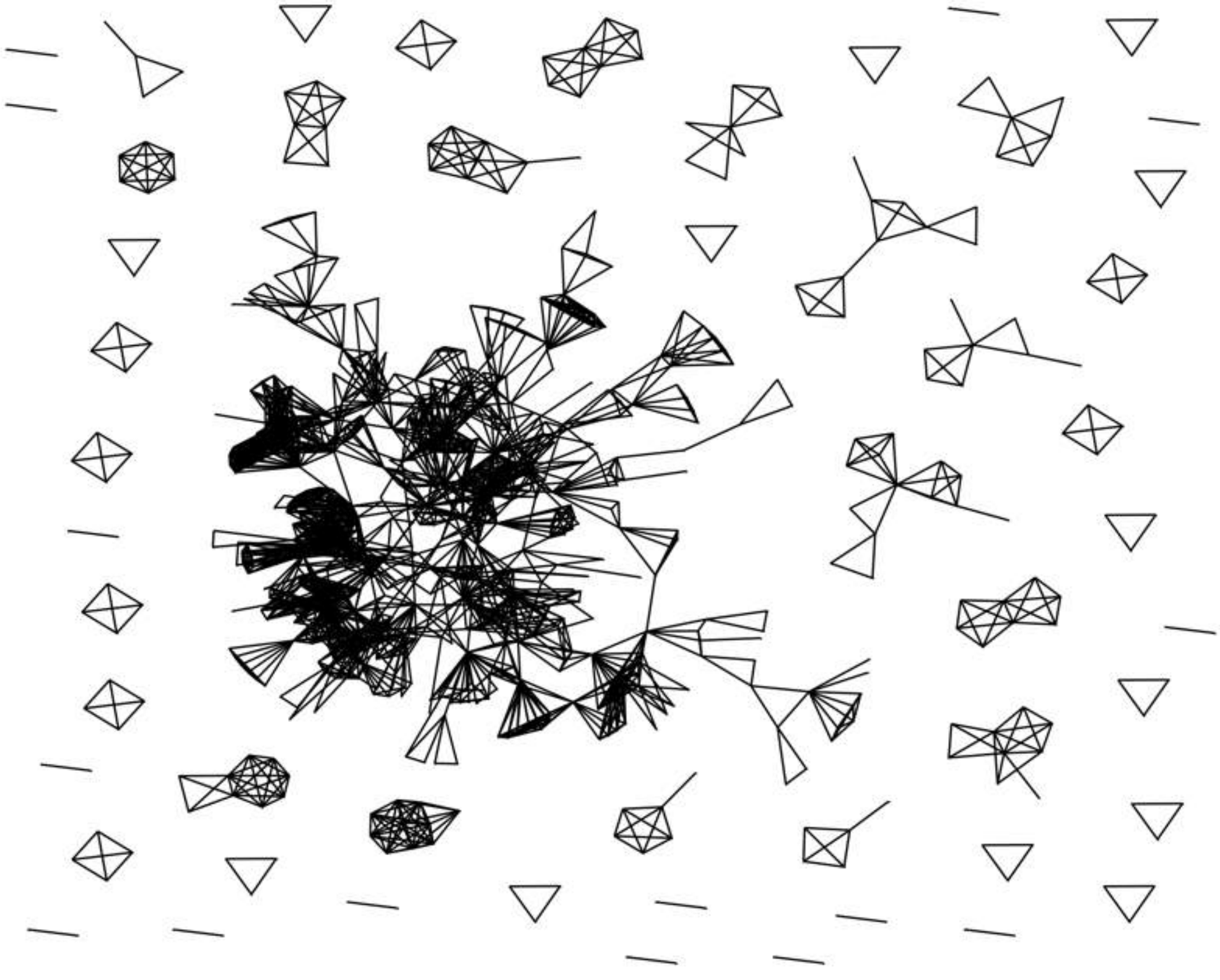
2005



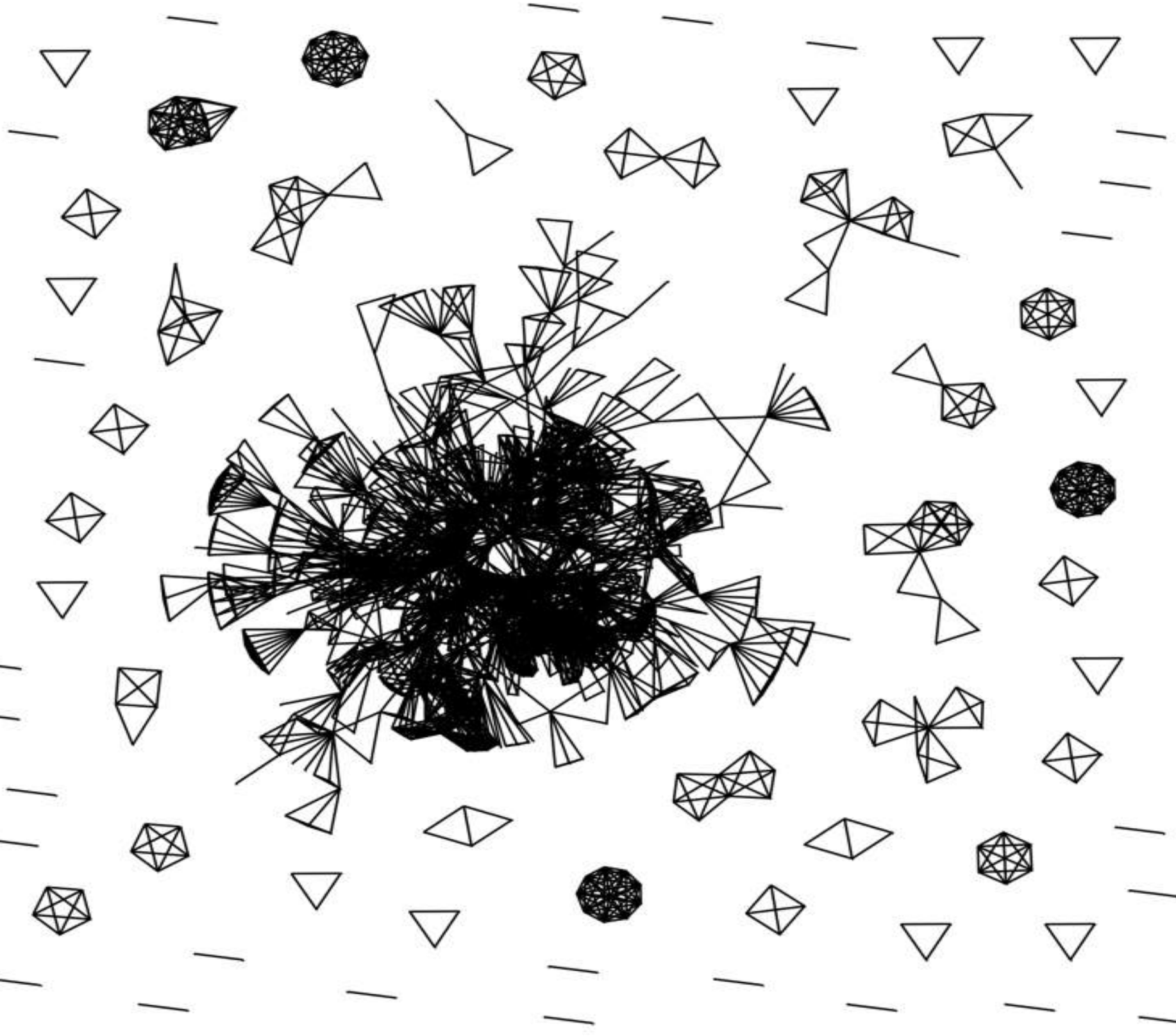
2007



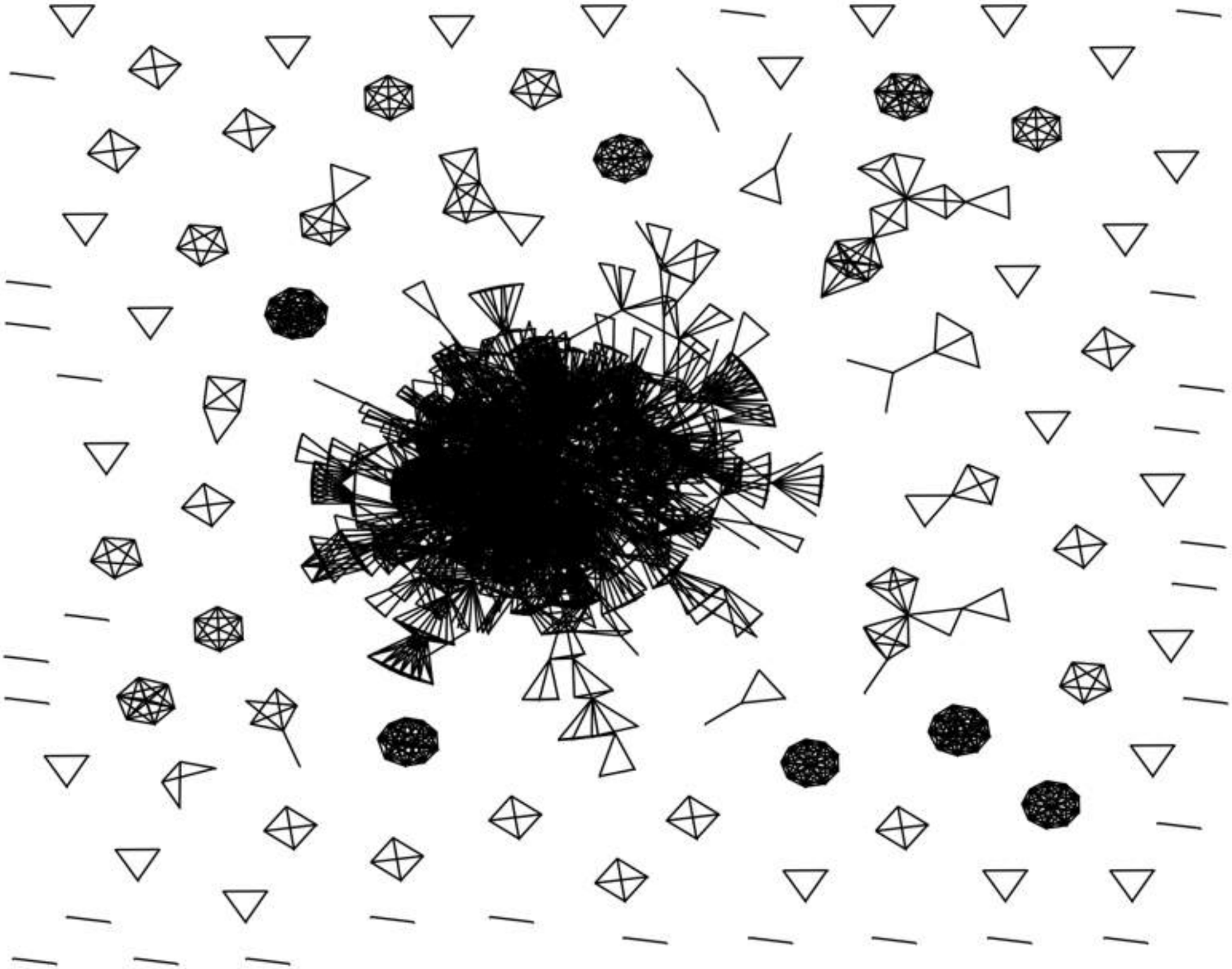
2009



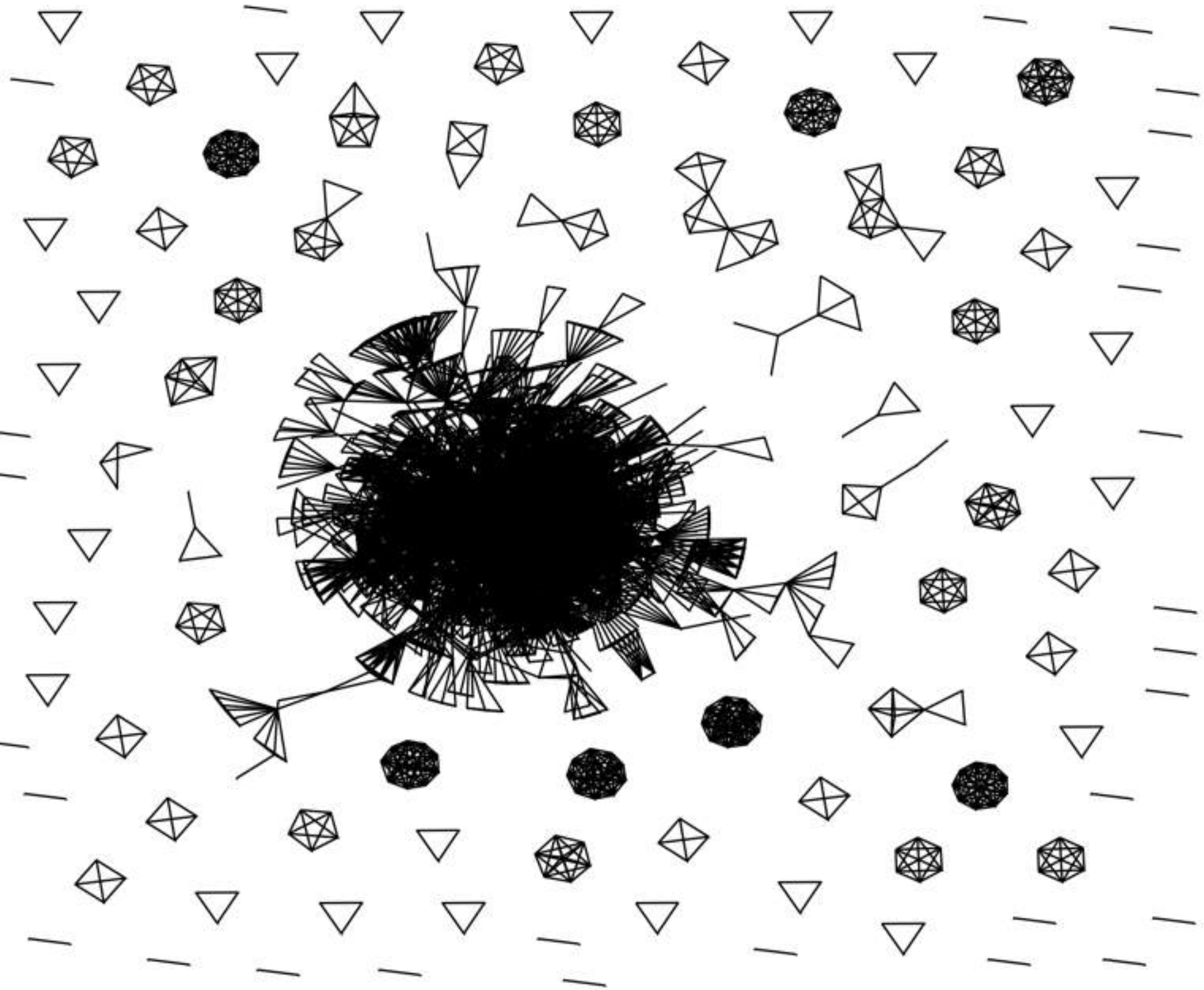
2011



2013



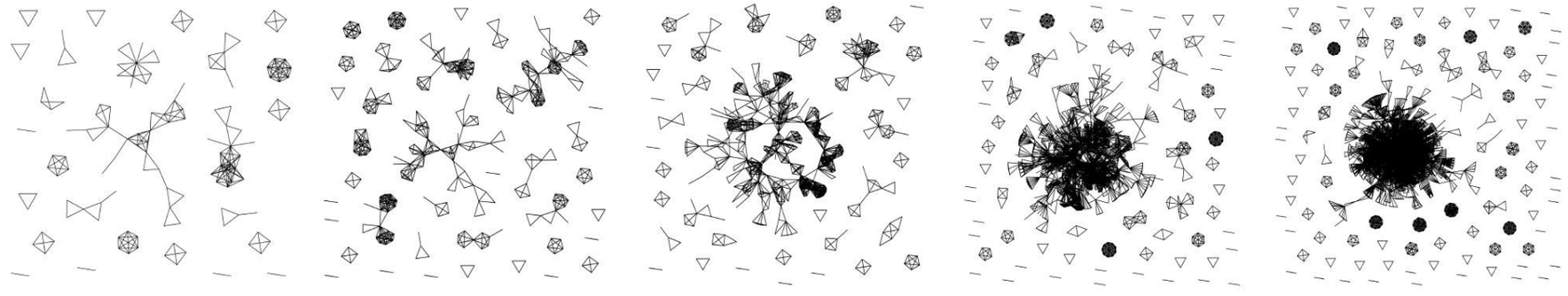
2015



Comparing Many Topics

Time

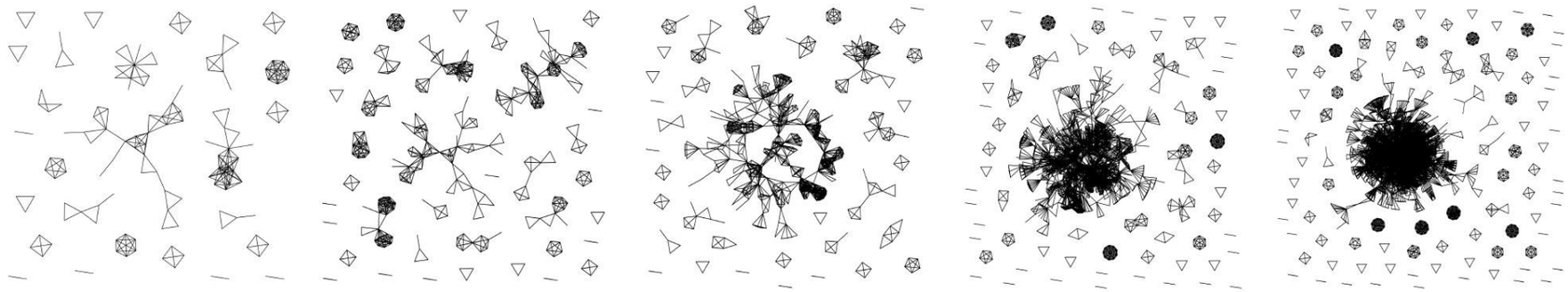
Topic 5:
Quantum Computing



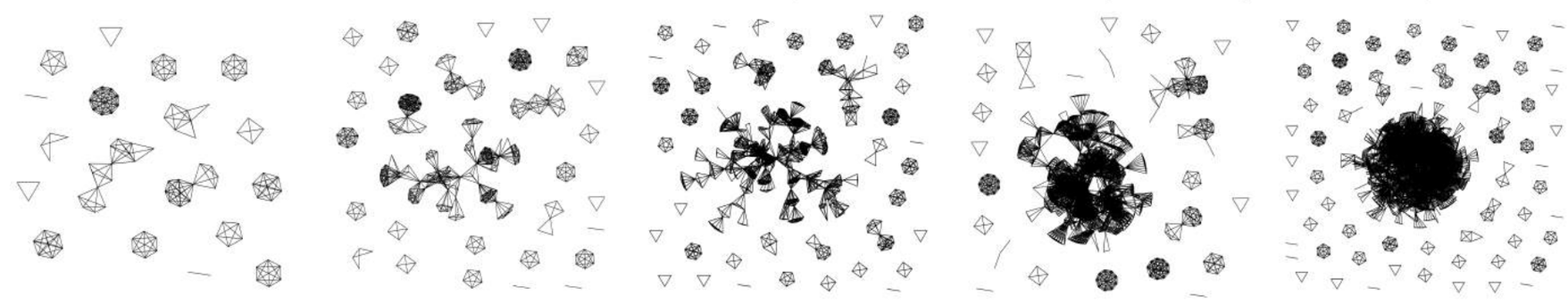
Comparing Many Topics

Time →

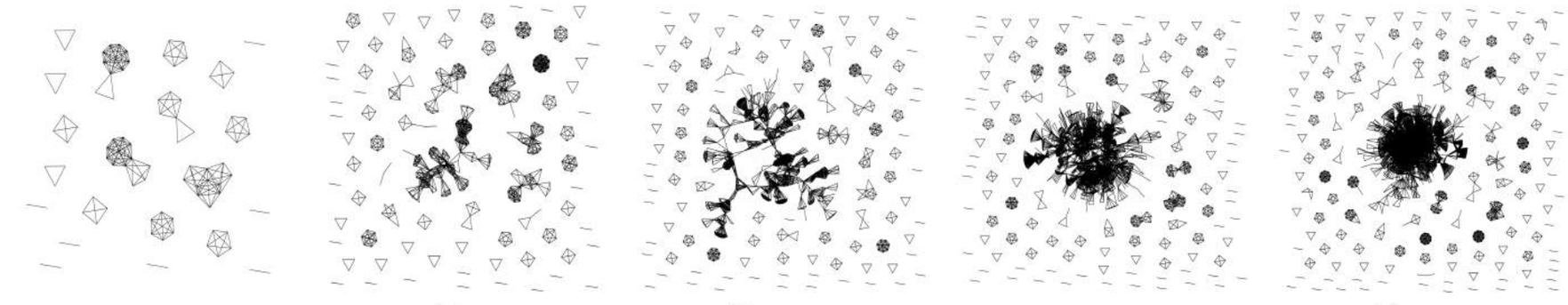
Topic 5:
Quantum Computing



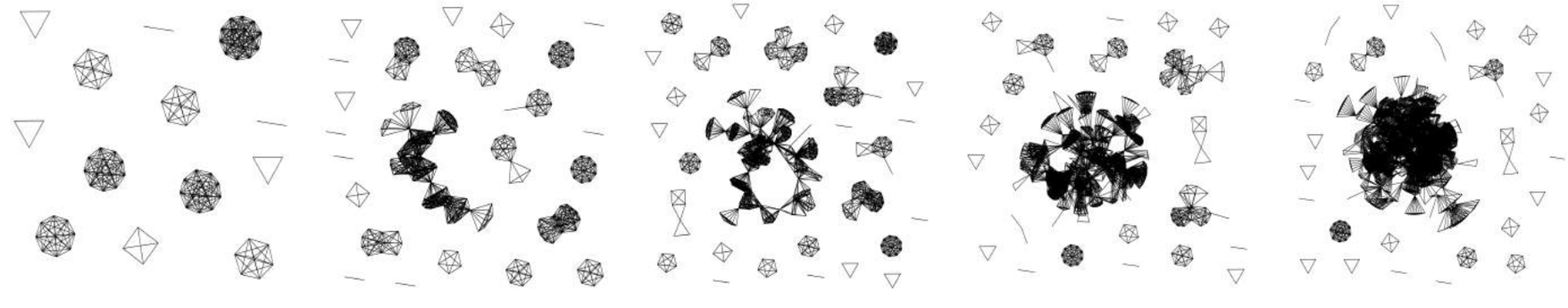
Topic 19:
Magnetic Materials



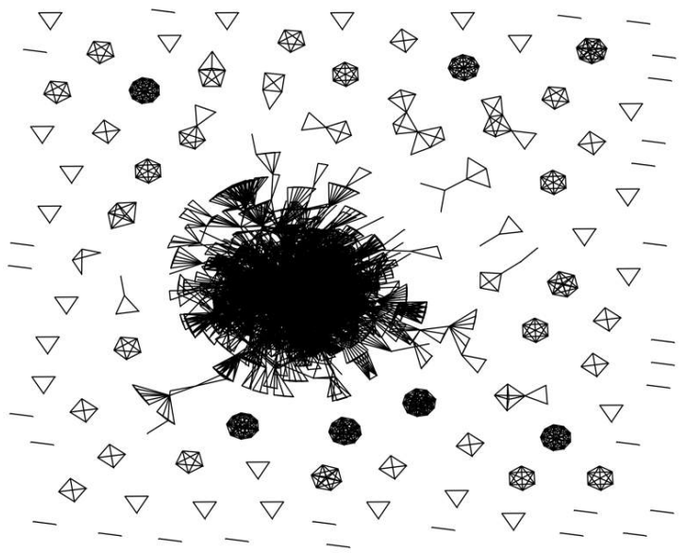
Topic 29:
Graphene

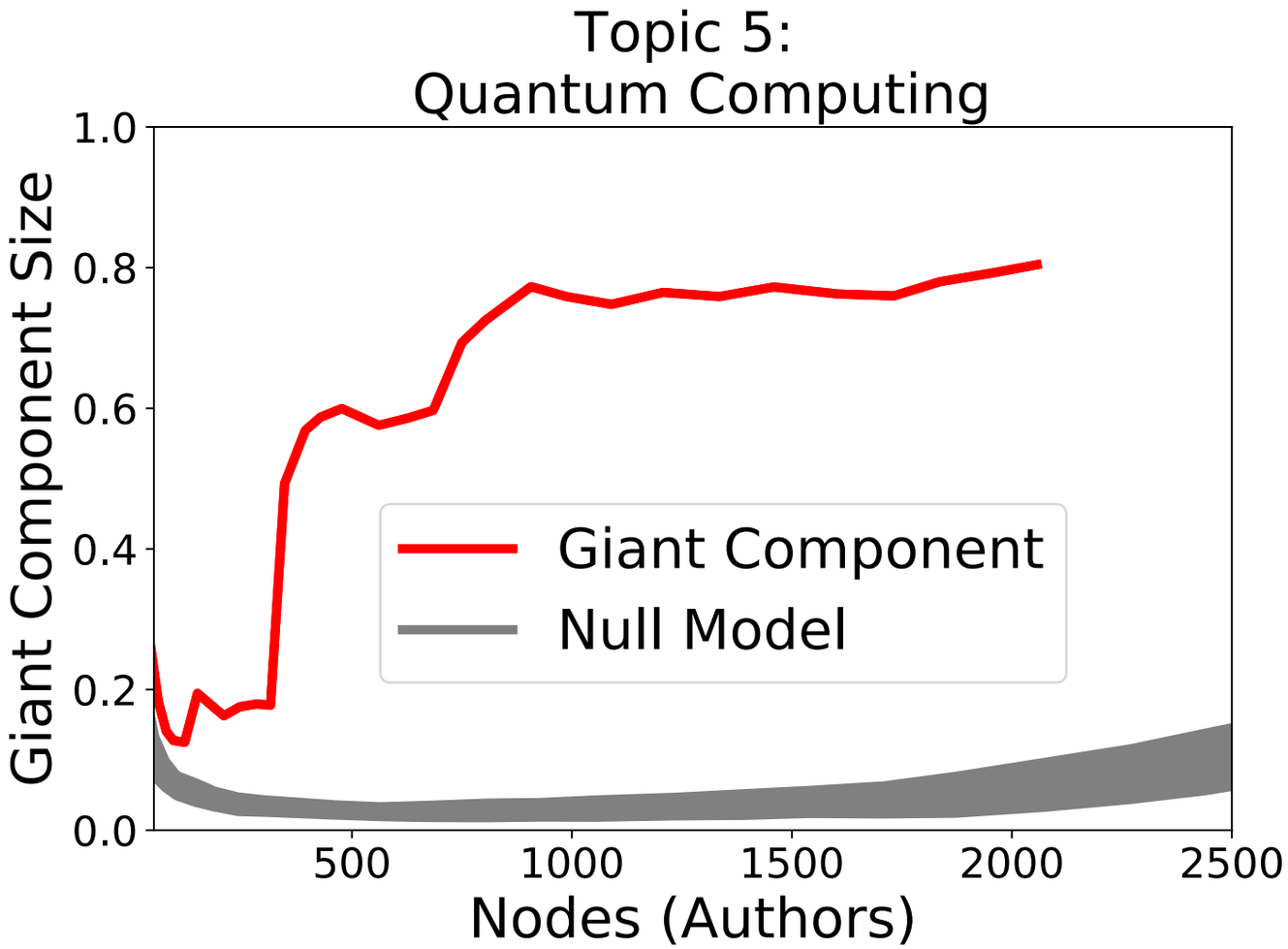


Topic 39:
Electron Spectroscopy



Giant Component Size

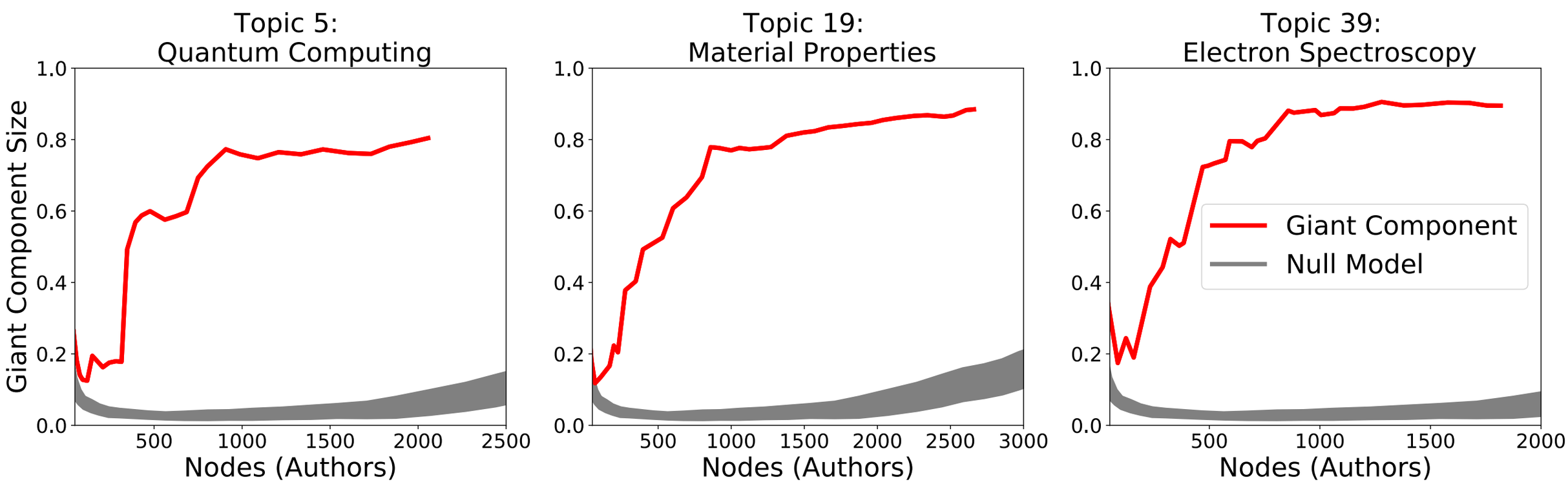




- Fraction of nodes in largest component
- Grows to dominate entire network
- Null model: articles chosen at random

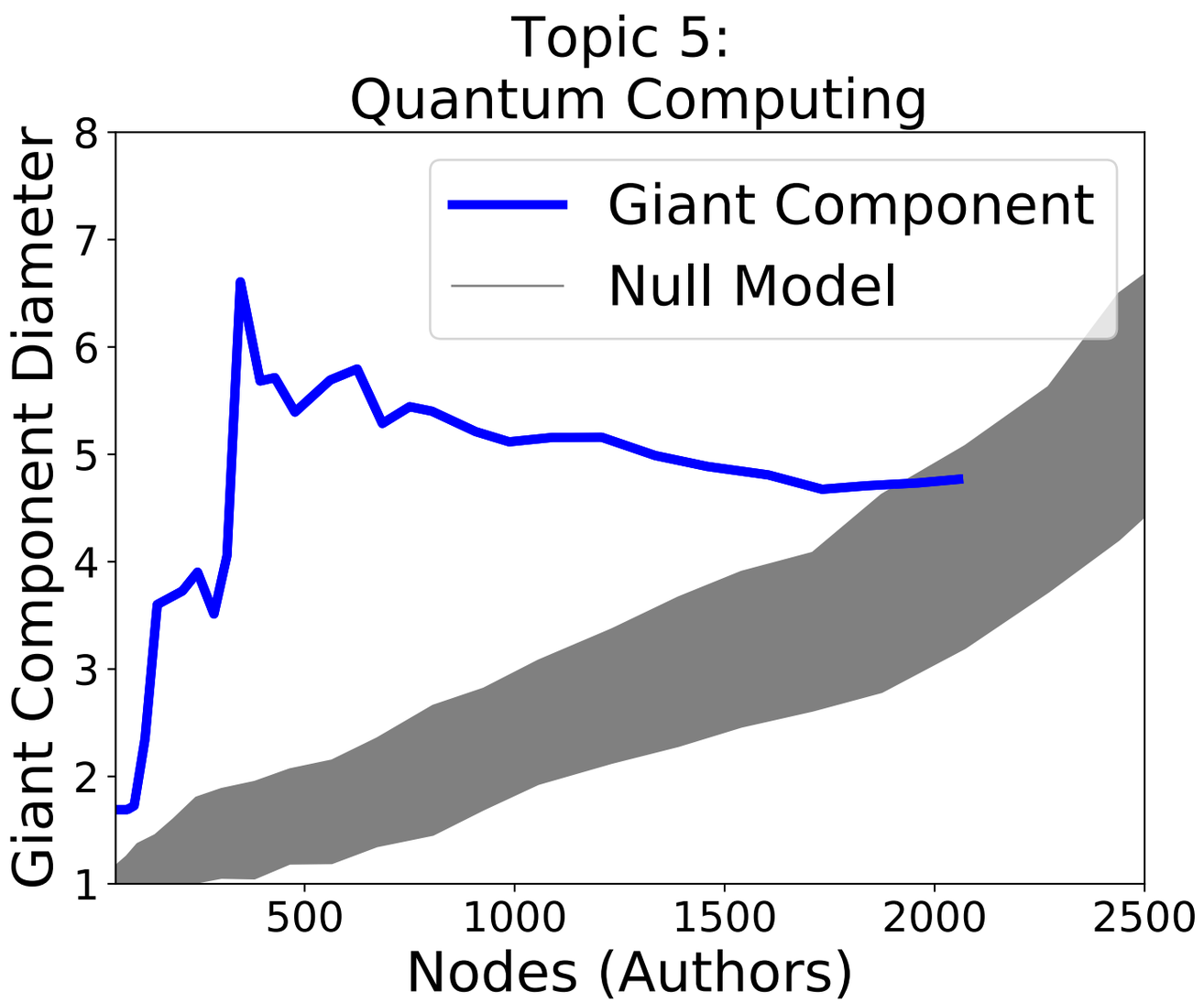
Giant Component Size

- Formation and Growth
- Consistent across different topics



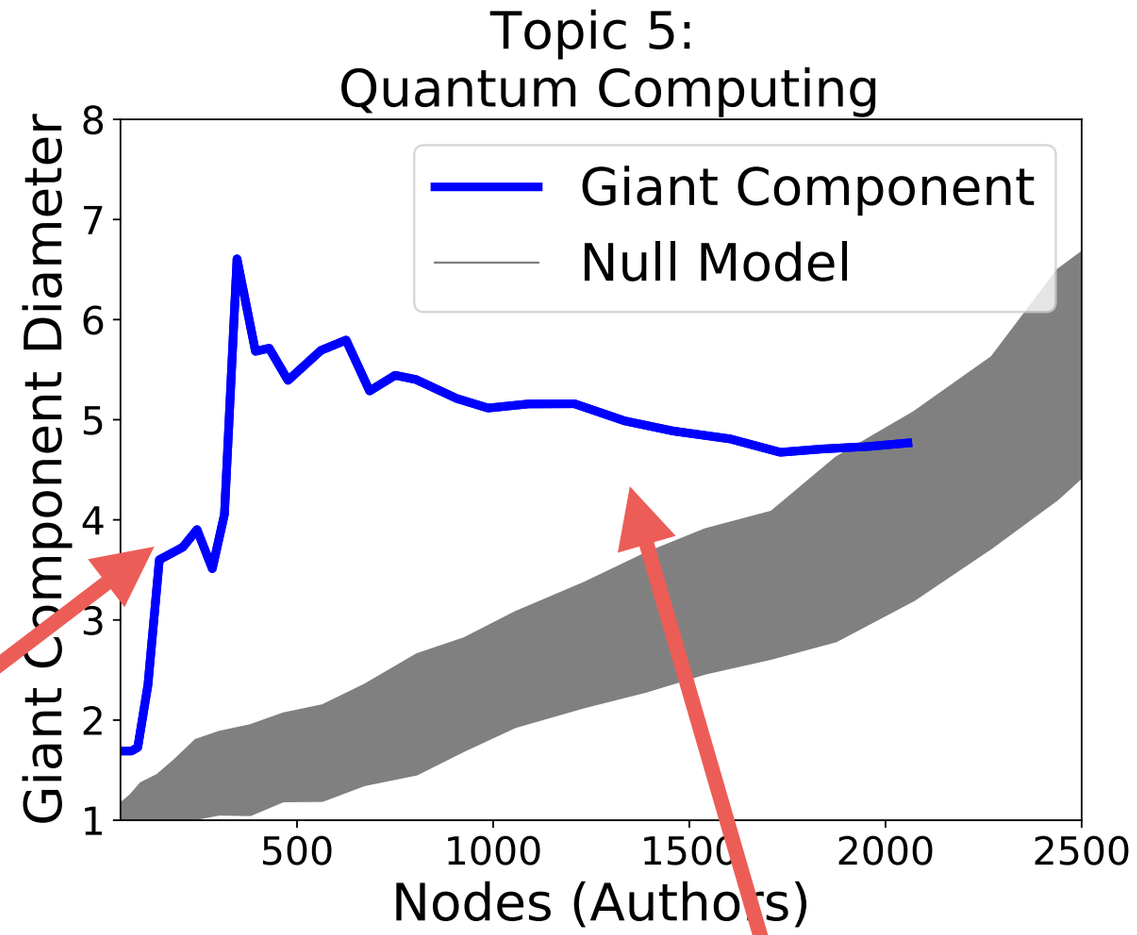
Giant Component Diameter

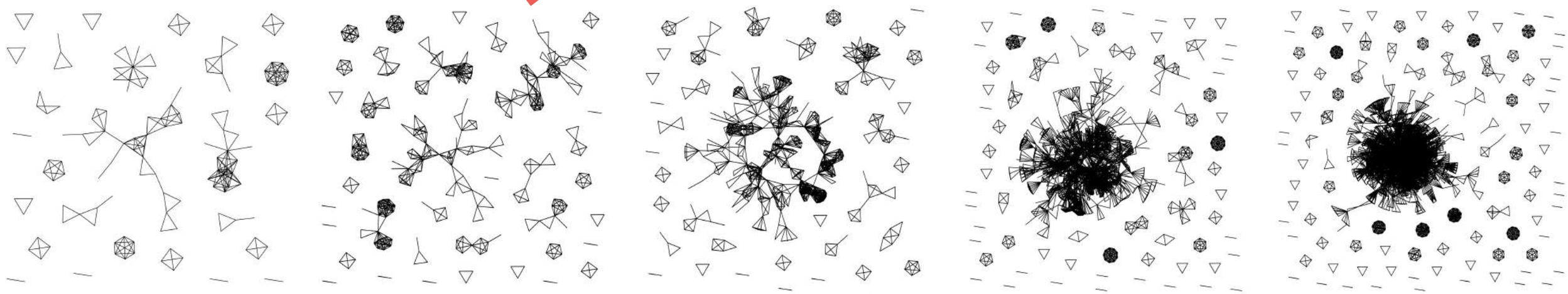
- Diameter: Mean Path Length in giant component
- Characteristic distance inside giant component
- Two stages:
  - Grows
  - Shrinks
- Null Model



Two Stages of Network Assembly

- Tree-like growth
  - Diameter increases
- Densification
  - Diameter shrinks

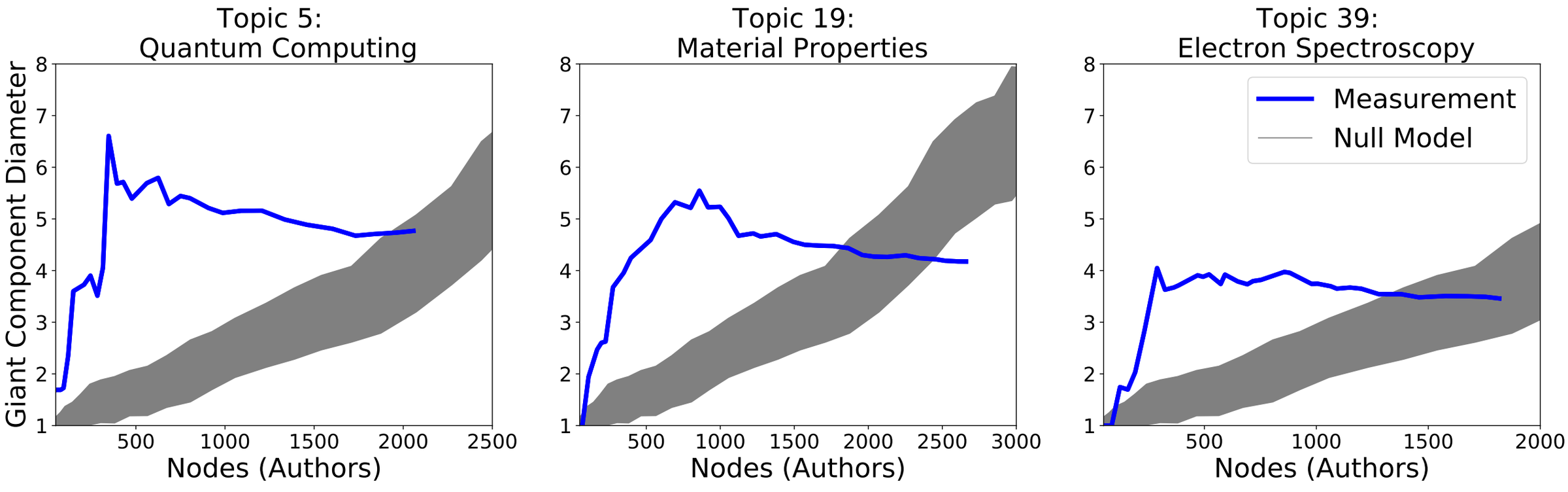




Time

Giant Component Diameter

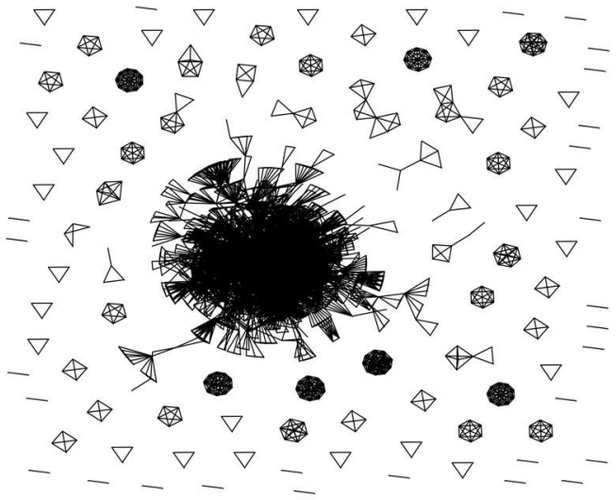
- Rapid initial growth, followed by steady decline
- Consistent across different topics



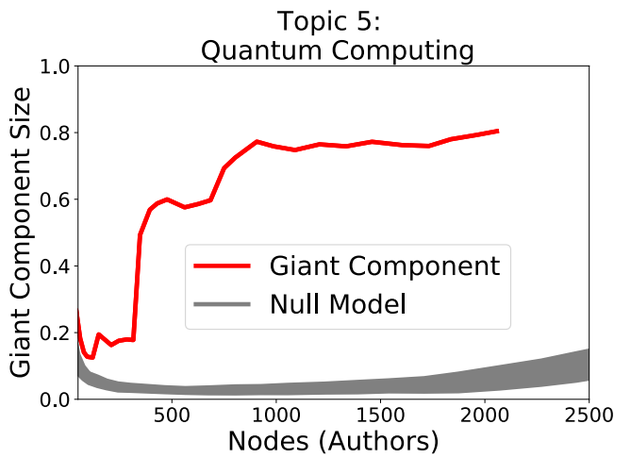
Validations

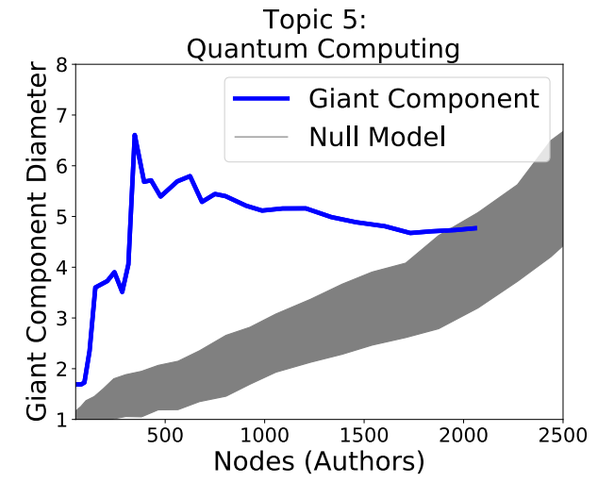
- Results are robust across corpora
  - Inferred topics on Web of Science cond-mat articles
  - More extensive coverage of experimental articles
- Robust to adding temporal noise
- Robust to different thresholding criteria
- Robust to different number of topics

Summary



- Measured properties of large population of algorithmically-generated co-authorship networks
- Topological transition
- Two stages of network assembly: growth and densification
- General property of co-authorship networks
- Framework for performing larger-scale comparisons of different scientific fields





Acknowledgments

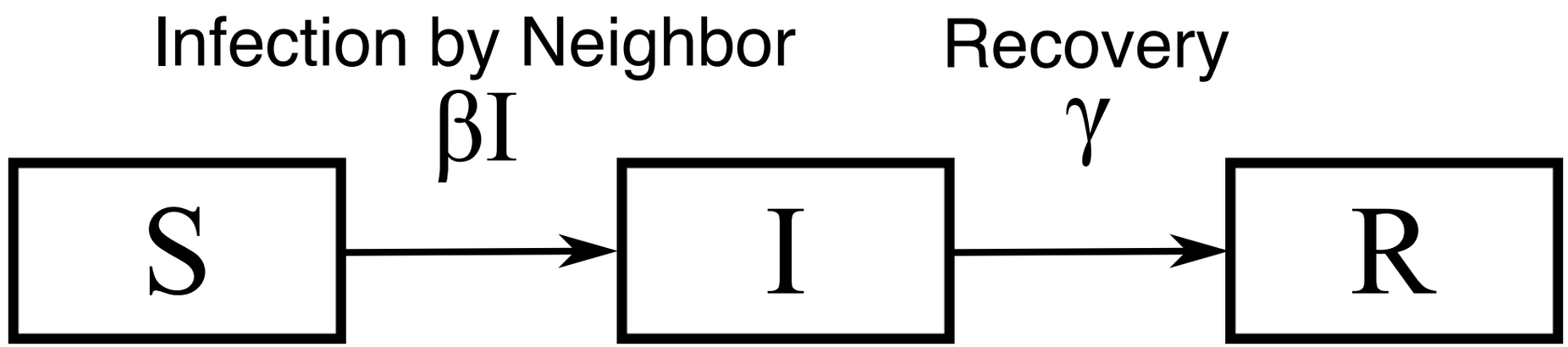
arXiv Project

- Samuel F. Way (co-author)
- Santa Fe Institute CSSS 2015 attendees:
  - Laurence Brandenberger
    - Brent Schneemann
    - Richard Barnes
- Michael Macy
- Paul Ginsparg
- Alexandra Schofield
- Haofei Wei

SIRS Project

- Chris Myers (Advisor)
- Sarabjeet Singh
- Drew Dolgert
- AFIDD group members:
  - Dave Schneider
  - Jason Hindes
  - Oleg Kogan

Mathematical Representation



$$S + I + R = 1$$

$$\frac{dS}{dt} = -\beta SI$$

- Deterministic mean-field model

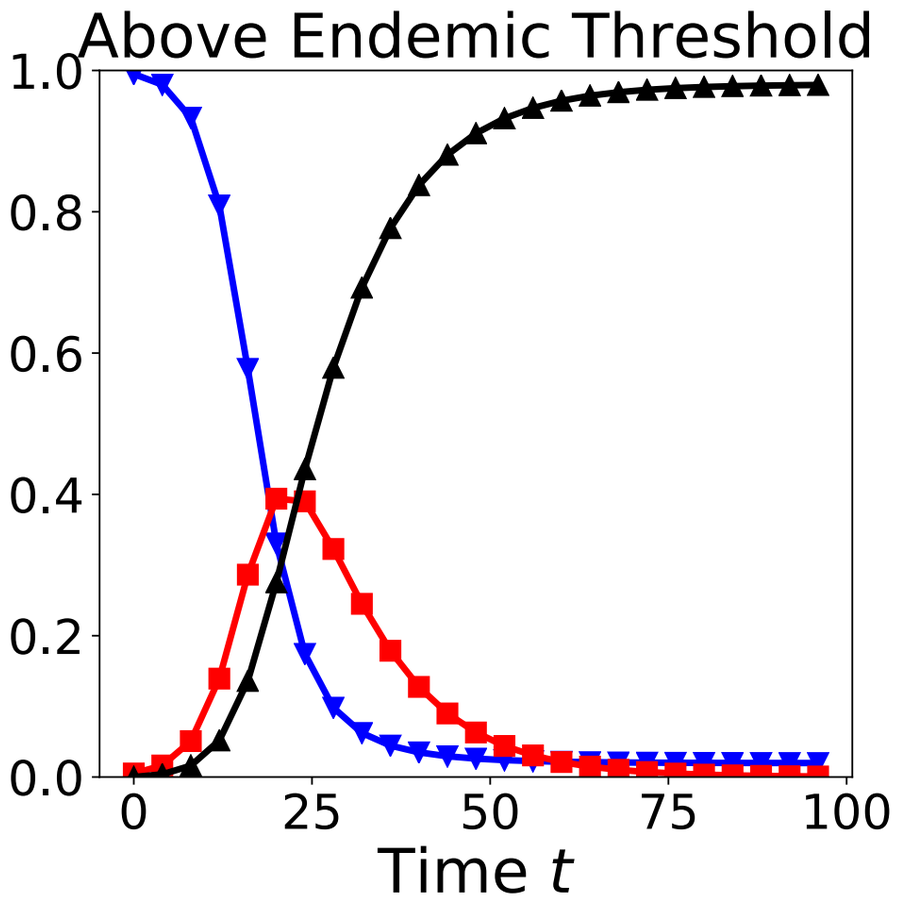
$$\frac{dI}{dt} = \beta SI - \gamma I$$

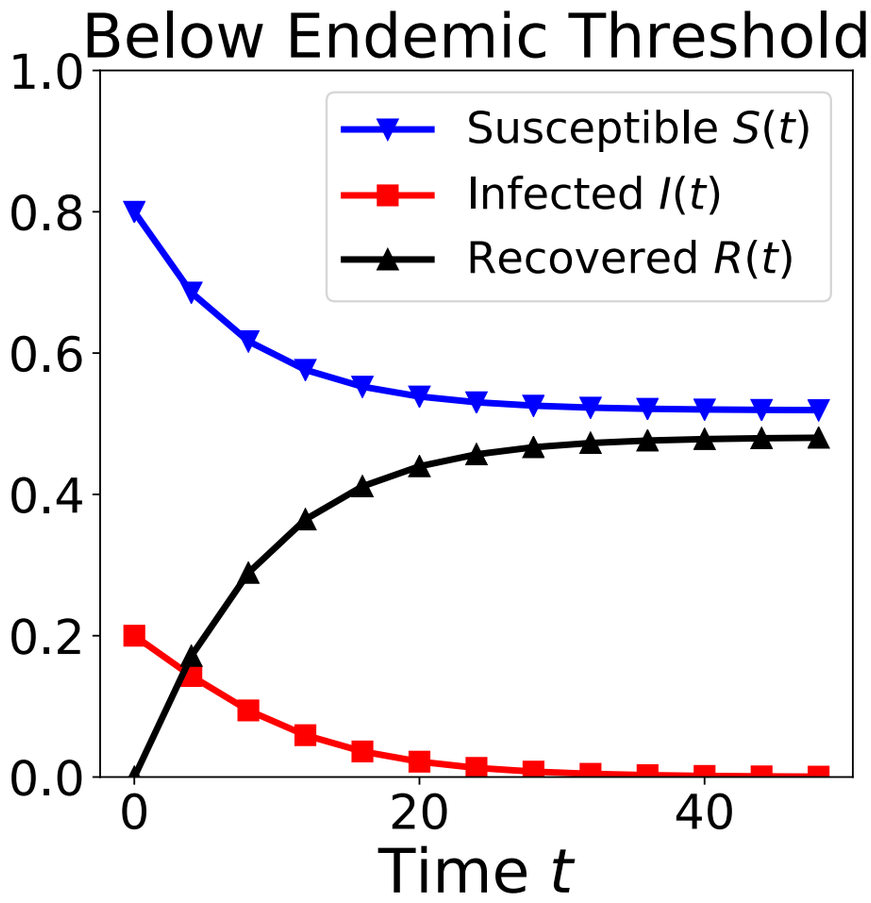
- Homogeneous mixed population

$$\frac{dR}{dt} = \gamma I$$

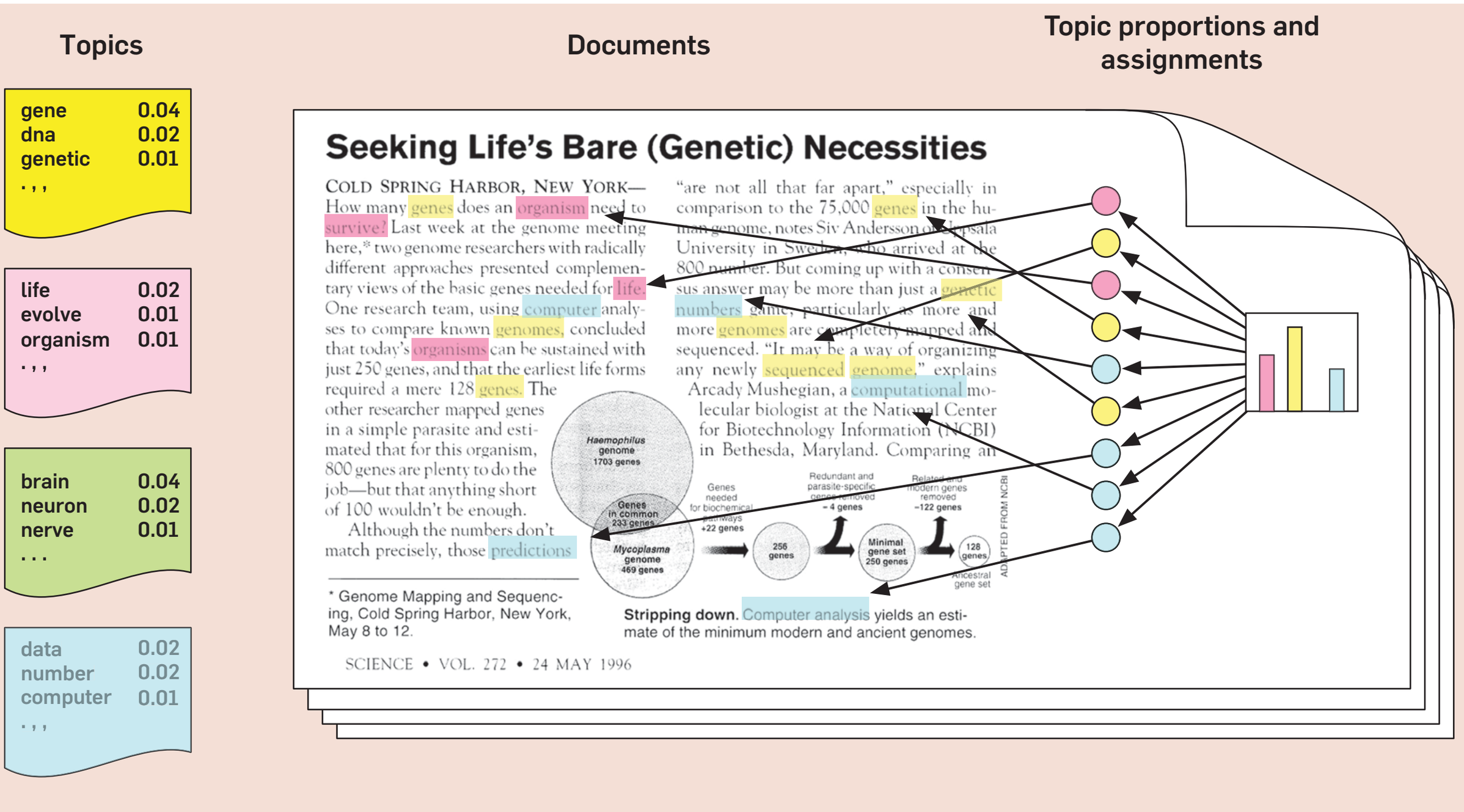
SIR Solution

- “Epidemic Threshold” - when disease transmission is fast enough to affect population
- Single outbreak above $\mathbf{R_0}= \beta/\gamma > 1$
- No outbreak below threshold





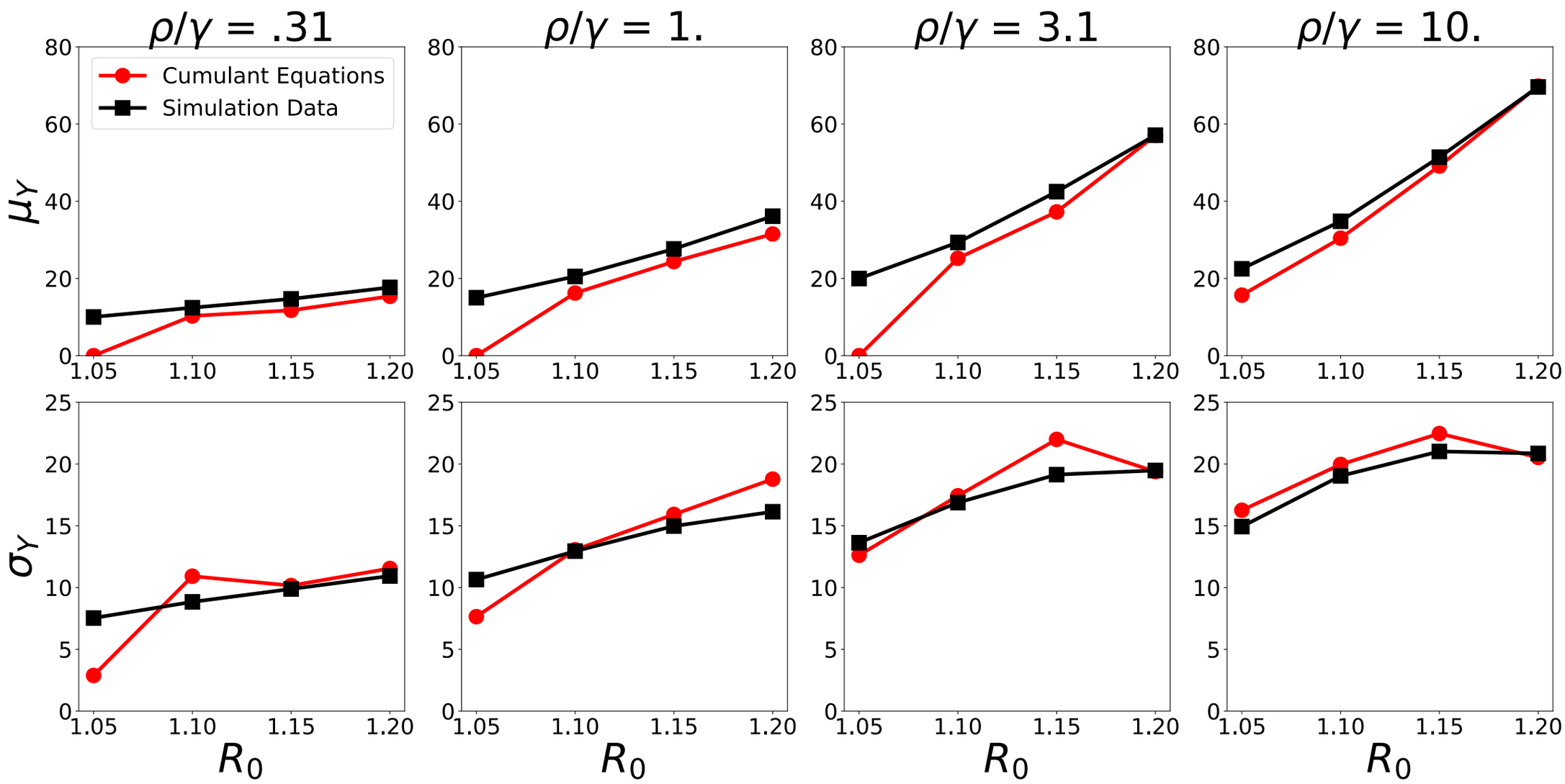
Topic Modeling



Blei, *Comm. of the ACM* (2002)

Means and Fluctuations

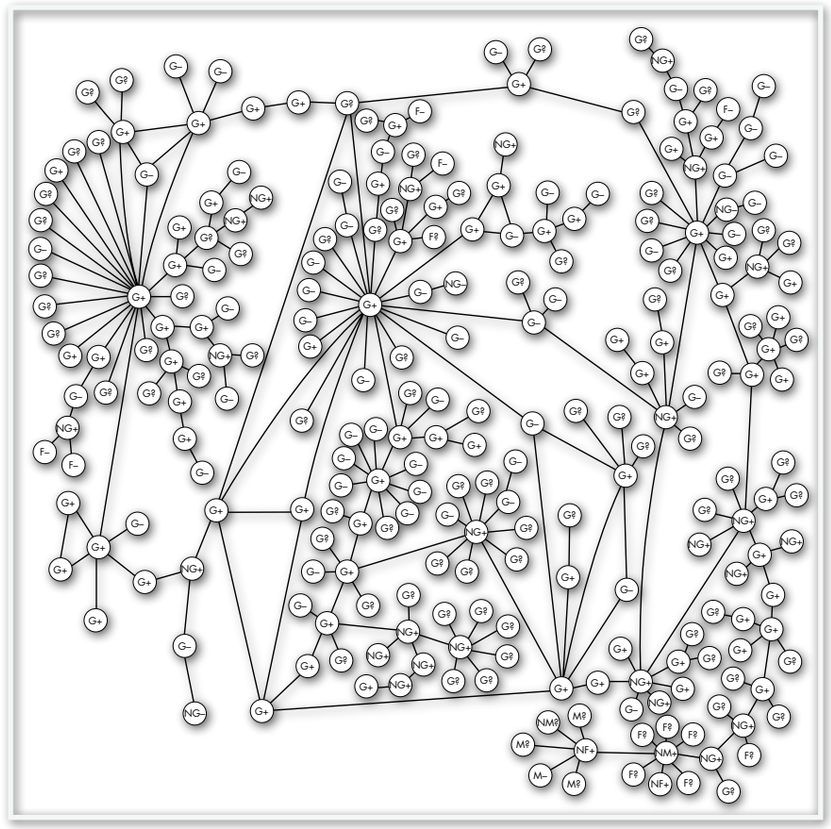
- Closest agreement :
  - large $R_0 = \beta/\gamma$, large ρ/γ
  - large μ_y, small σ_y/μ_y



Network Heterogeneity

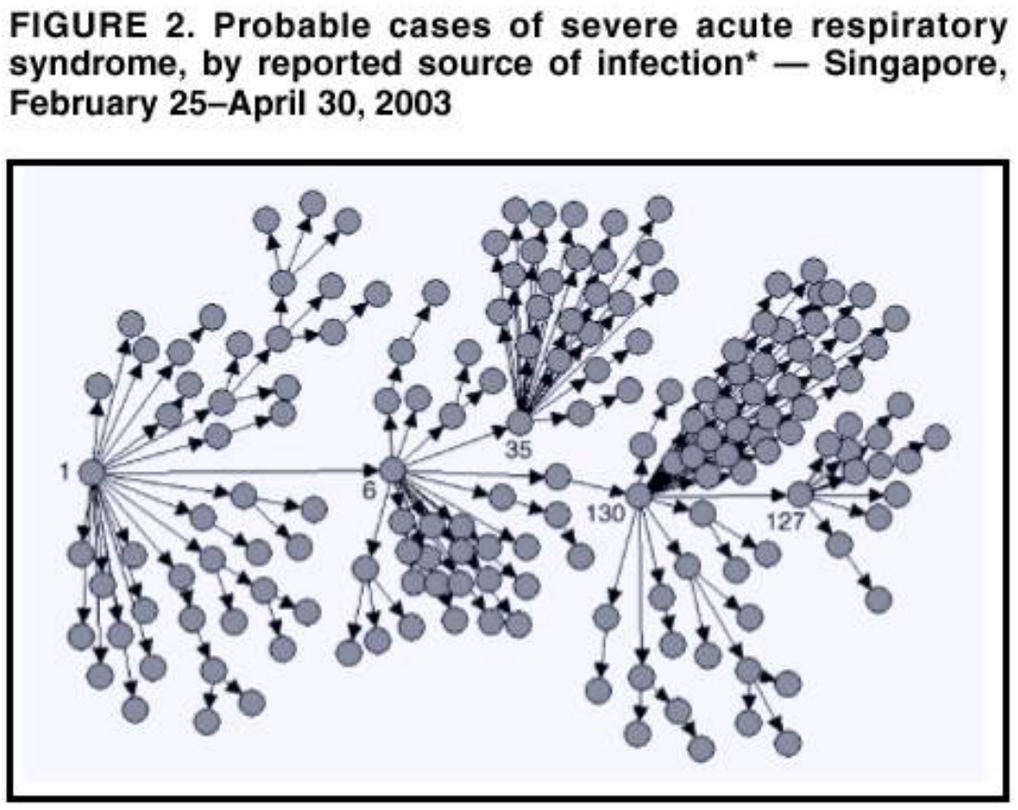
- Networks naturally describe heterogenous connections, or in transmission, between hosts

Heterogeneous Contact Network: HIV



Potterat, *Sexually Transmitted Infections* (2002)

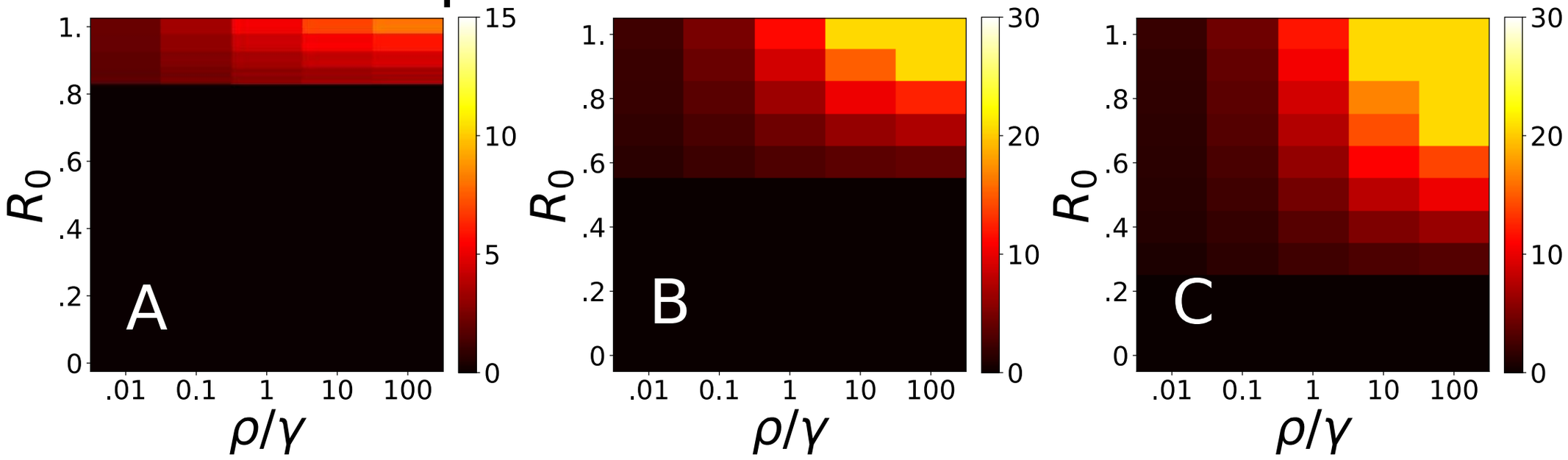
Heterogeneous Transmission Network: SARS



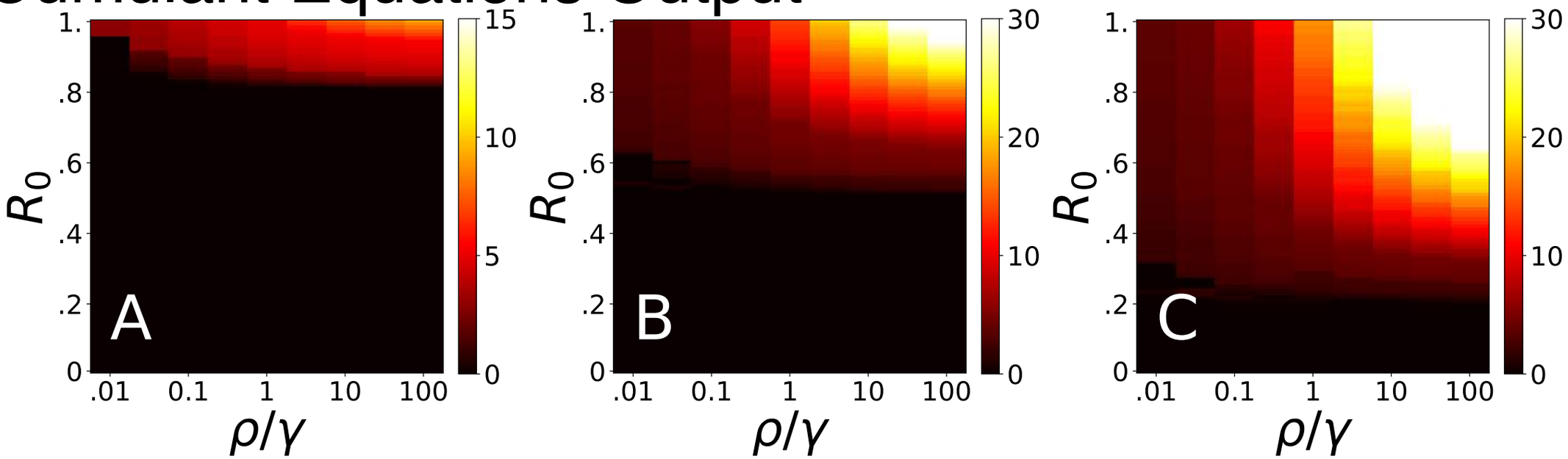
MMWR 2003;52:241–8. from www.cdc.gov

Extinction Times

Simulation Output

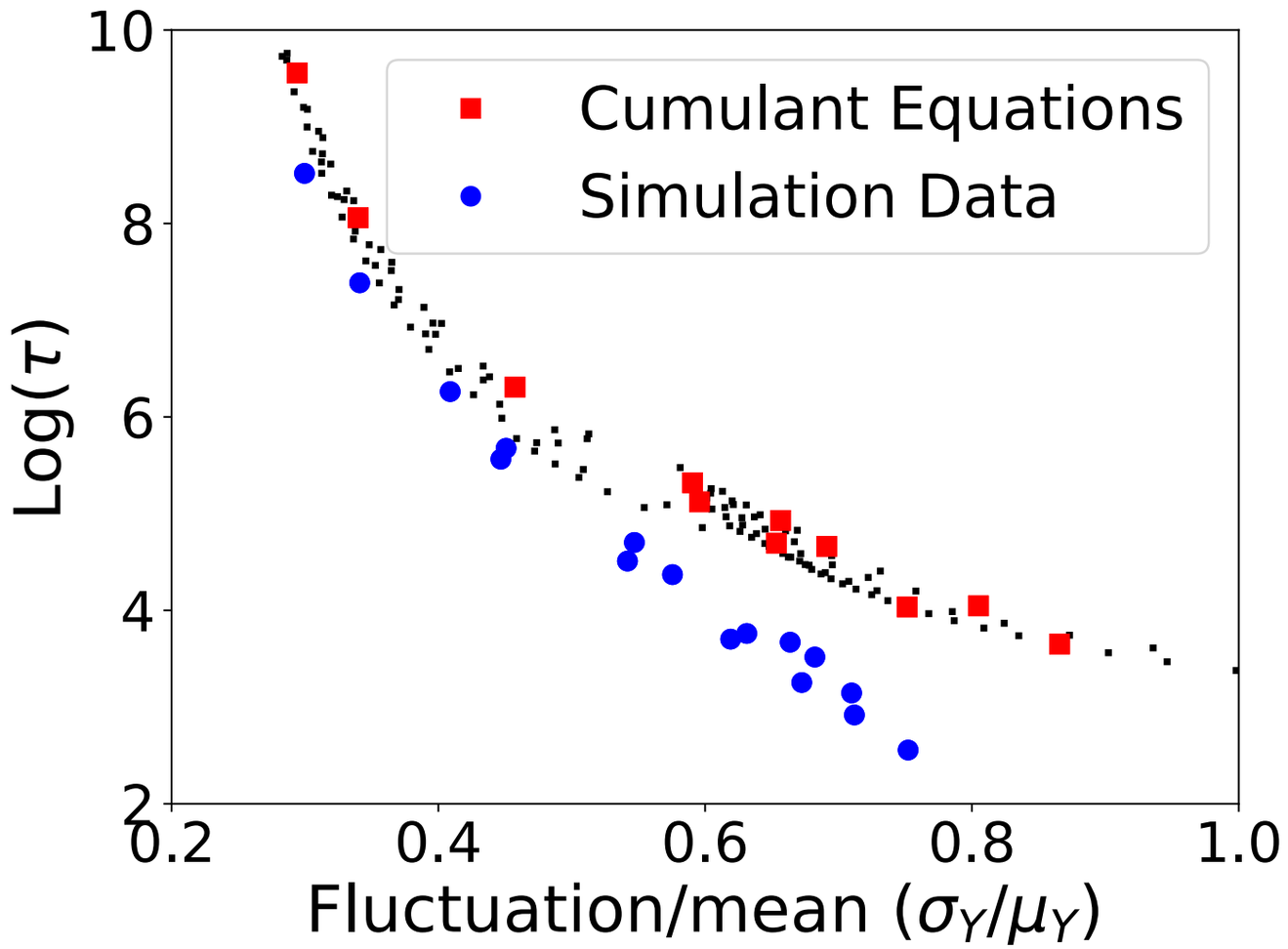


Cumulant Equations Output



Homogeneous → Heterogeneous

Times to Extinction



- Cumulant equations: data also collapse onto a low-dimensional curve
- Consistently overestimates τ, but qualitatively consistent
- Across all different parts of parameter space (β,ρ)