NETWORKS

BST228

TODAY

- Review of ignorability.
- Statistical network models: Erdős–Rényi model, (soft) random geometric graphs and latent space models, stochastic block models, …
- Mechanistic network models (maybe)

IGNORABILITY FOR DATA MAR

Suppose data are MAR such that $p\left(\mathbf{I} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}}, \boldsymbol{\phi}\right) = p\left(\mathbf{I} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\phi}\right)$ and the parameters of our model $\boldsymbol{\theta}$ and of the missingness model $\boldsymbol{\phi}$ are independent a priori, i.e., $p\left(\boldsymbol{\theta}, \boldsymbol{\phi}\right) = p\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\phi}\right)$. Using the expression for the joint posterior,

$$p\left(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{I}\right) = \int d\boldsymbol{\phi} \, d\mathbf{y}_{\mathrm{mis}} \, p\left(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{y}_{\mathrm{mis}} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{I}\right)$$

$$\propto \int d\boldsymbol{\phi} \, d\mathbf{y}_{\mathrm{mis}} \, p\left(\mathbf{I} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\phi}\right) p\left(\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\phi}\right).$$

The integral with respect to $\mathbf{y}_{\mathrm{mis}}$ and $\boldsymbol{\phi}$ is separable such that

$$p\left(\boldsymbol{\theta} \mid \mathbf{y}_{\mathrm{obs}}, \mathbf{I}\right) \propto \left[\int d\boldsymbol{\phi} \, p\left(\mathbf{I} \mid \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\phi}\right) p\left(\boldsymbol{\phi}\right)\right] \left[\int d\mathbf{y}_{\mathrm{mis}} \, p\left(\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)\right]$$

$$\propto \int d\mathbf{y}_{\mathrm{mis}} \, p\left(\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)$$

$$\propto p\left(\mathbf{y}_{\mathrm{obs}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right).$$

The second line follows because the first integral does not depend on $\boldsymbol{\theta}$ and can be absorbed by the proportionality. The third line follows by the law of total probability.

The missing data model does not appear in the posterior, and we say the missing data mechanism is *ignorable*. This means we do not need to model $p\left(\mathbf{I} \mid \ldots\right)$, but it does *not* mean that we can ignore that data are missing.
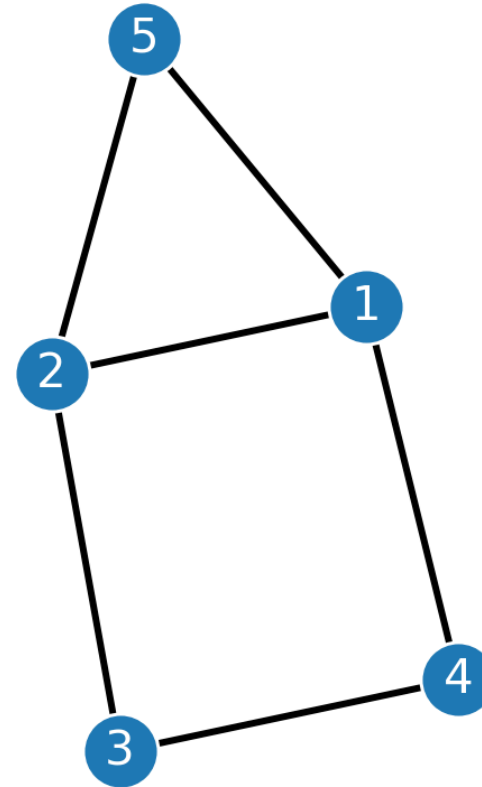
SO WHY NETWORKS?

- Social networks and isolation.
- Contact networks and infection.
- Referral networks for specialist care.
- …

WHAT IS A NETWORK?

- A graph has n nodes with labels $i \in \{1, \ldots, n\}$.
- If two nodes i and j are connected, $A_{ij} = 1$ in the adjacency matrix $\mathbf{A}$. If not connected, $A_{ij} = 0$.
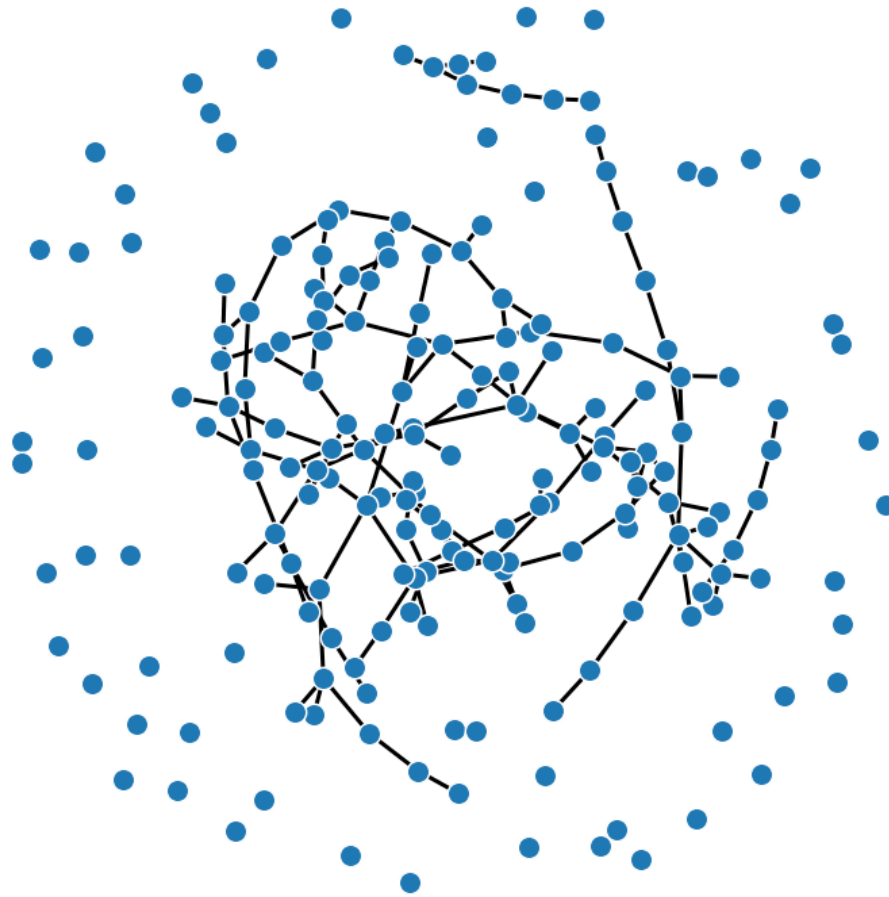
THE SIMPLEST OF MODELS: ERDŐS–RÉNYI

In an Erdős–Rényi model, all edges are independent and present with probability λ such that

$$p\left(\mathbf{A} \mid \lambda\right) = \prod_{i,j=1}^{n} \lambda^{A_{ij}} \left(1 - \lambda\right)^{1 - A_{ij}} .$$

- An illustration of an ER graph with 200 nodes and connection probability just high-enough for a giant connected component to form.
- At this still relatively low connection probability, a non-negligible number of nodes remain isolated.

DEGREE DISTRIBUTION OF THE ERDŐS–RÉNYI MODEL

The degree of node i is its number of connections

$$k_i = \sum_{j=1}^{n} A_{ij}.$$
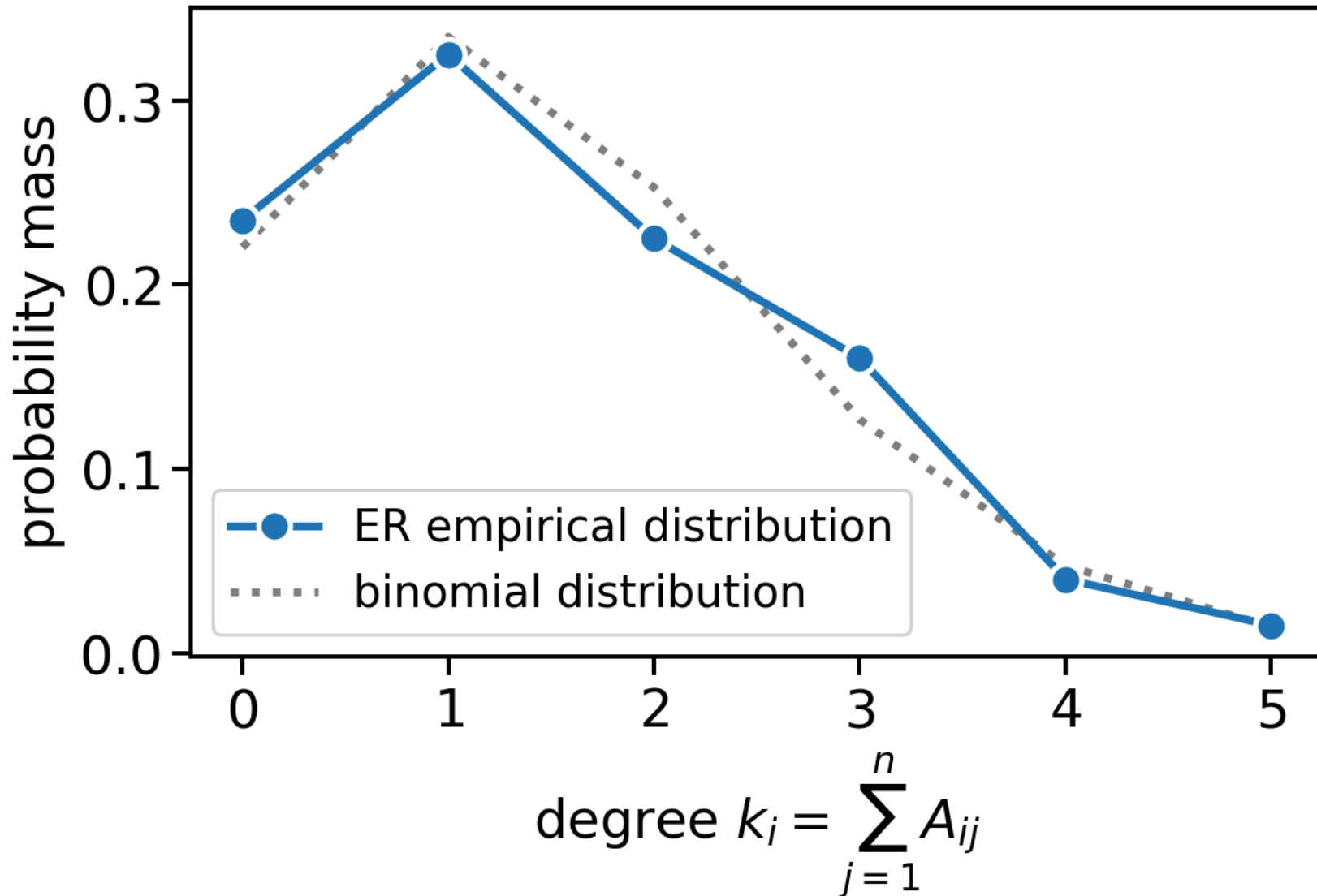
- The empirical degree distribution of the network in the previous figure is shown in blue.
- It agrees with our expectation of a binomial degree distribution with appropriate parameters.
- This degree distribution is not realistic because observed degree distributions are often very heavy-tailed. E.g., celebrities on social media sometimes have millions of followers whereas most accounts only have a few.
- Heavy-tailed degree distributions are particularly important for sexually transmitted infections where "hubs" can lead to large-scale spreading events.
- While amenable to theoretical analysis, the ER model is not particularly useful in public health applications.
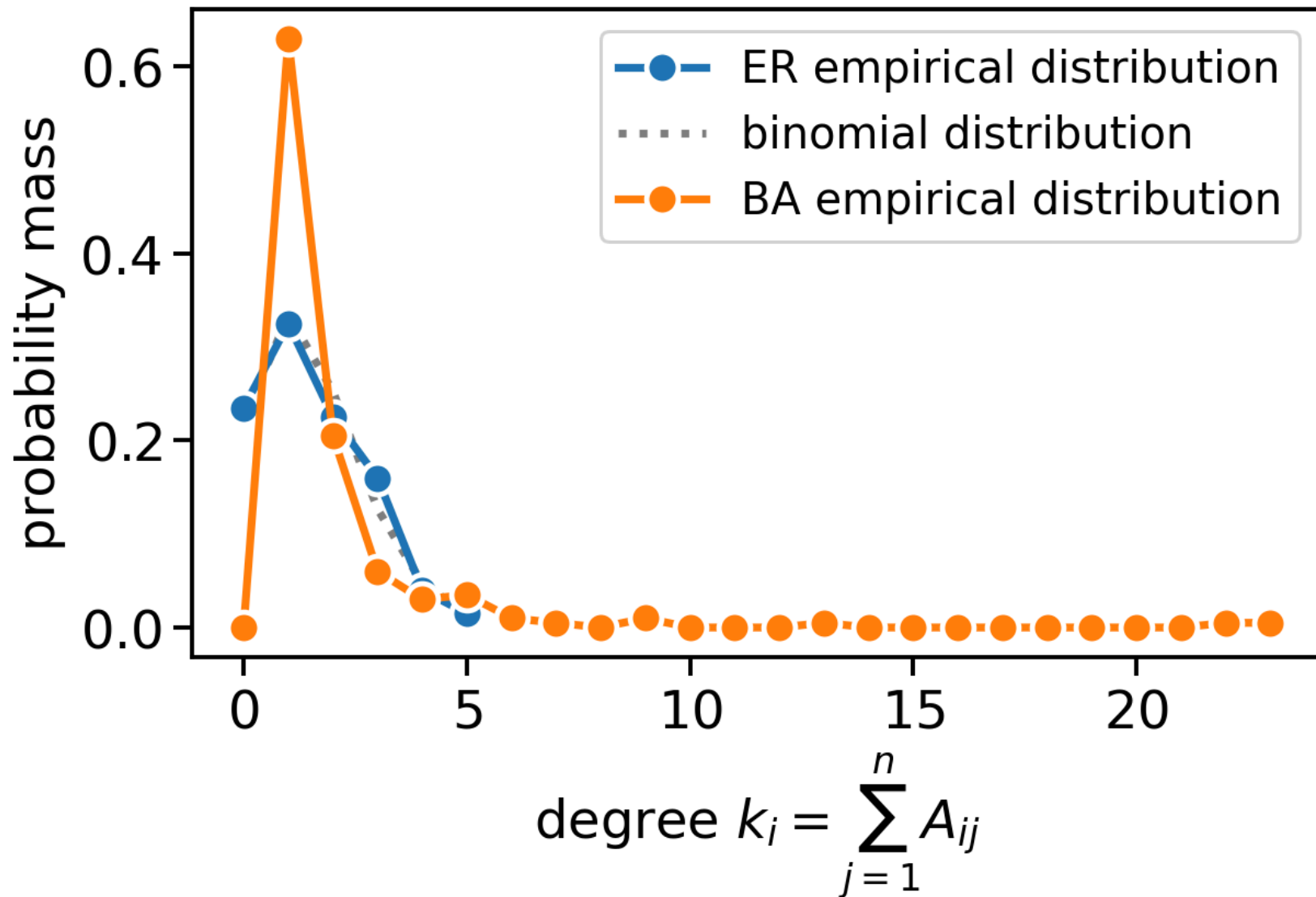
Speaker notes

- Like the previous figure, a binomial degree distribution is shown in blue.
- The distribution in orange is the degree of a Barabási–Albert model which is a classic model to reproduce realistic degree distributions.

CONDITIONALLY INDEPENDENT EDGE MODELS

In a *conditionally* independent edge model,

$$p\left(\mathbf{A} \mid \mathbf{X}\right) = \prod_{i,j=1}^{n} p\left(A_{ij} \mid \mathbf{x}_i, \mathbf{x}_j\right),$$
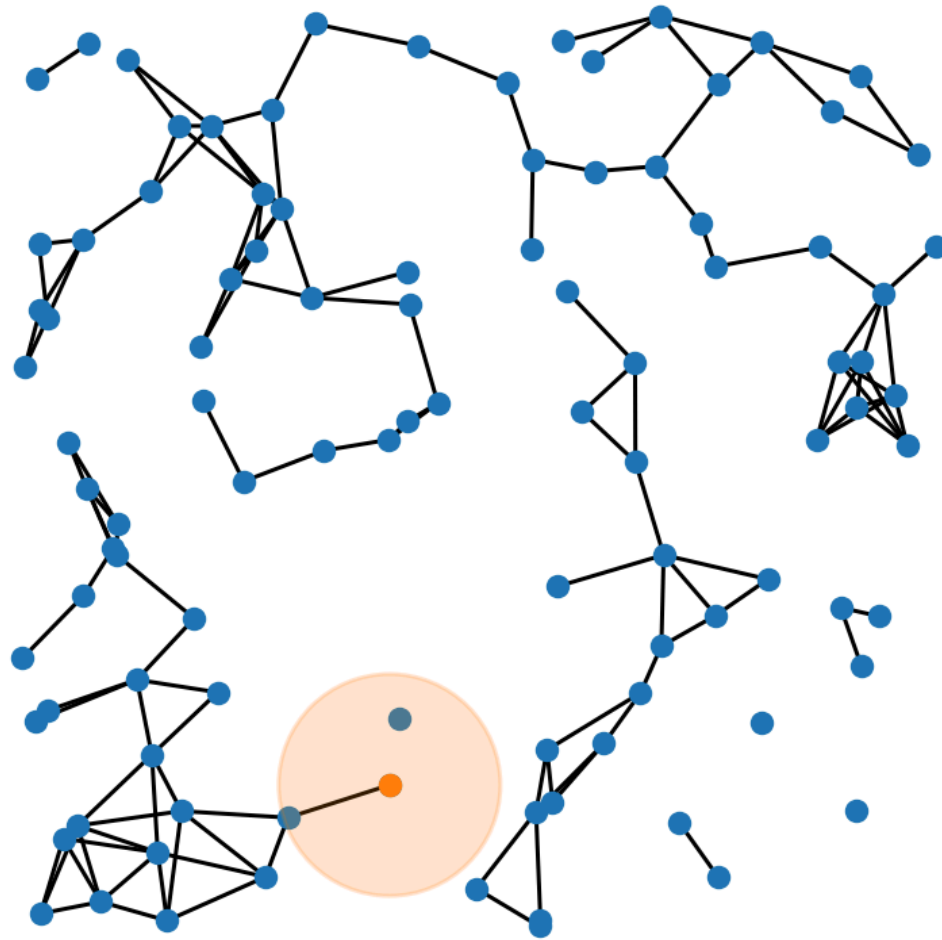
where $\mathbf{X}$ is a matrix of node features.

RANDOM GEOMETRIC GRAPHS

Each node i has coordinates $\mathbf{x}_i$ in a q-dimensional space, and nodes are connected with probability λ if $|\mathbf{x}_i - \mathbf{x}_j| < \theta$ for some threshold θ.

- The figure shows a realization of a soft random geometric graph with $\lambda = 0.8$. Nodes are not connected if they are too far apart. If they are close enough (within the distance indicated by the orange circle), they may connect.

LOGISTIC REGRESSION FOR NETWORKS

$$A_{ij} \mid \mathbf{x}_i, \mathbf{x}_j \sim \text{Bernoulli}\left(\lambda_{ij}\right)$$

$$\text{logit}\left(\lambda_{ij}\right) = a + \mathbf{b}^\mathsf{T}\mathbf{f}\left(\mathbf{x}_i, \mathbf{x}_j\right),$$

where the intercept a controls overall edge density, $\mathbf{b}$ are regression coefficients, and $\mathbf{f}$ maps node attributes $\mathbf{x}$ to features.

SOCIAL SEPARATION

Consider the social separation

$$\psi\left(\mathbf{x}_i, \mathbf{x}_j\right) = \operatorname{logit}\left(\lambda_{jj}\right) - \operatorname{logit}\left(\lambda_{ij}\right),$$

which measures how isolated i is from j.

Speaker notes

- This measure of social separation is based on Hoffmann and Jones (2020).
- If the probability for connection between nodes at different locations in the space is low (second term), the separation is large (remember that $\operatorname{logit}\left(\text{small probability}\right) = \text{large negative number}$).
- If there is no dependence on attributes, $\psi = 0$.
- If there is heterophily, i.e., a tendency for nodes to connect with others who are different, there is negative separation.

SOCIAL ISOLATION

Consider the social isolation

$$\phi\left(\mathbf{x}_i\right) = \int d\mathbf{x}_j \; \psi\left(\mathbf{x}_i, \mathbf{x}_j\right) p\left(\mathbf{x}_j\right)$$

which measures how separated i is from others on average.

SOCIAL STRAIN

Consider the social strain

$$\Phi = \int d\mathbf{x}_i \, p\left(\mathbf{x}_i\right) \phi\left(\mathbf{x}_i\right)$$

which measures how isolated people are on average.

- This is a utilitarian definition of social strain because we focus on average social isolation.
- One might argue that other measures of social strain are more appropriate.
- Maybe we instead want to minimize the isolation of the most isolated person.
- Maybe we want to instead make sure the 75th percentile of isolation is below some threshold.
- We would have to define another measure in this case.

INFERENCE FROM SURVEY DATA

- We ask respondents about their demographics and also about the demographics of their "friends."
- We treat these pairs as "cases" and randomly pair respondents as "controls."
- If we properly account for the missing data (all the present or absent edges we didn't observe), we can infer the parameters of the logistic regression model.

MODEL-BASED UNCERTAINTY QUANTIFICATION

- Because ψ, ϕ and Φ are defined in terms of the model, we automatically get uncertainty quantification.
- Unlike regression coefficients, these quantities have real-world implications and interpretation.
- Even for non-parametric logistic regression, the notion of social separation, isolation, and strain remain valid and useful.

USS '14   GSS '04
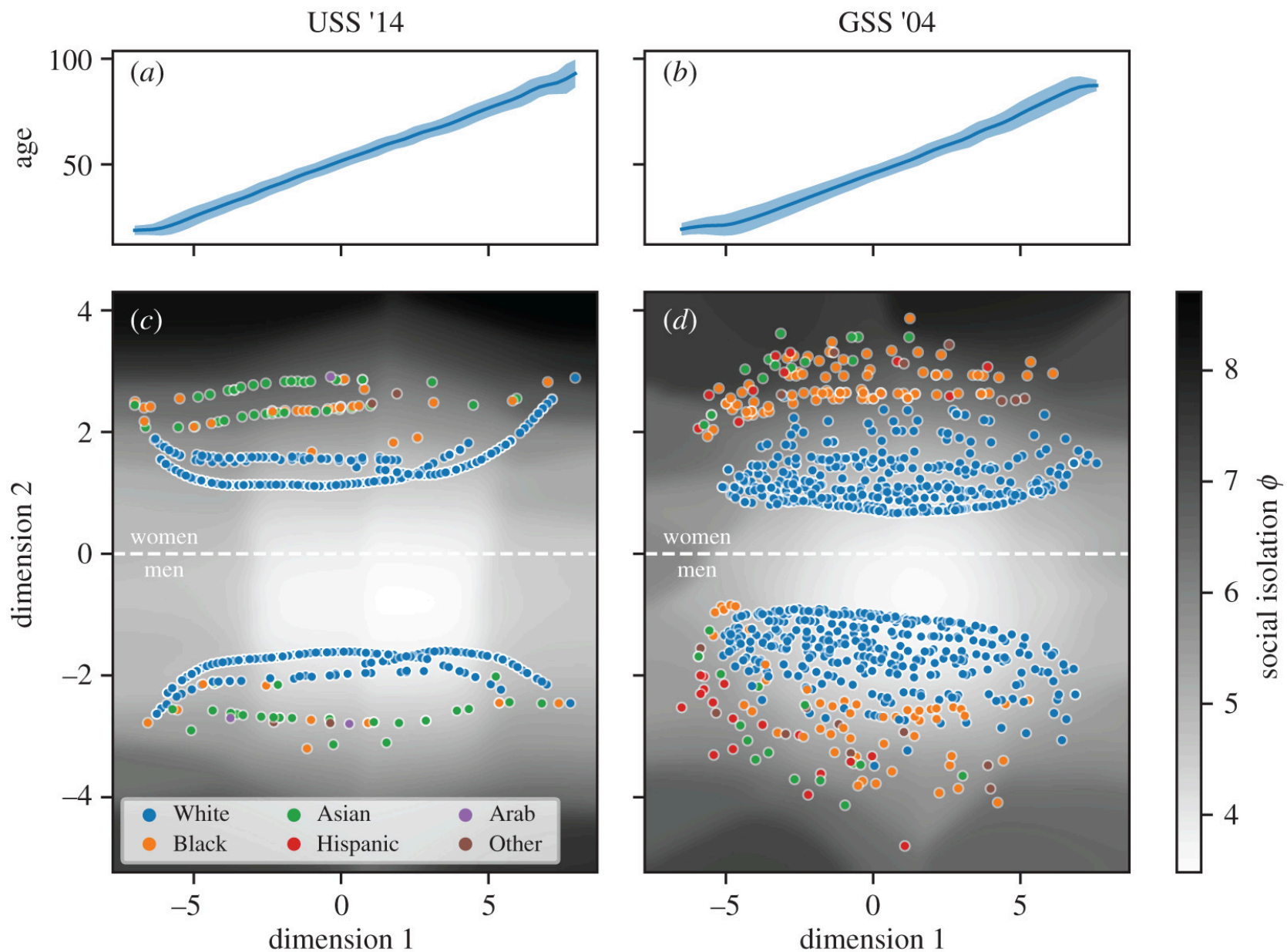
STOCHASTIC BLOCK MODELS (SBMS)

For a stochastic block model, features $x \in \{1, \dots, q\}$ indicate block membership and

$$A_{ij} \sim \text{Bernoulli}\left(\lambda_{x_i x_j}\right),$$

where $\boldsymbol{\lambda}$ is a $q \times q$ matrix of connection probabilities.
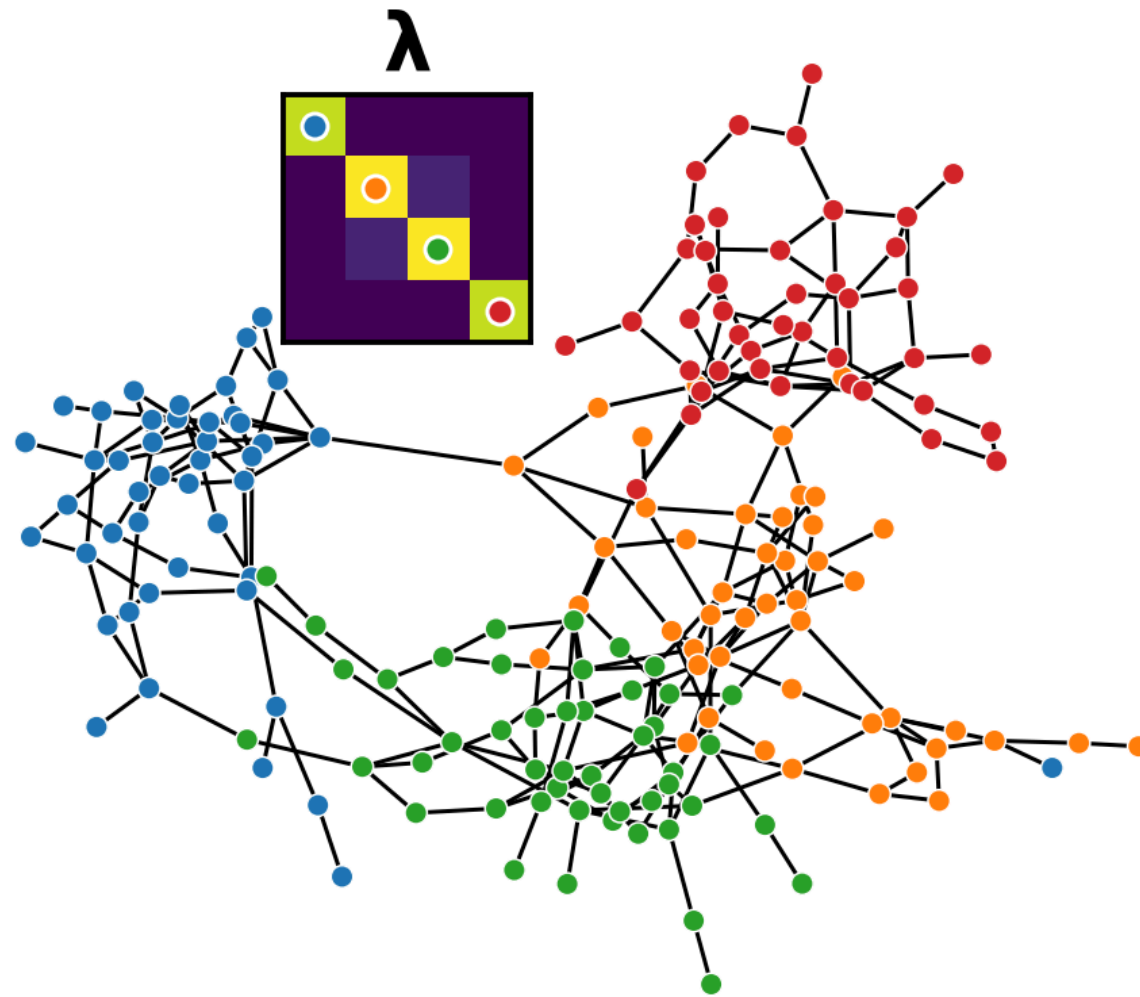
Speaker notes

- The figure shows a realization of a stochastic block model with strong community structure (i.e., connections within communities are more likely than between communities).
- The corresponding matrix of connection probabilities λ is shown as an inset heatmap.
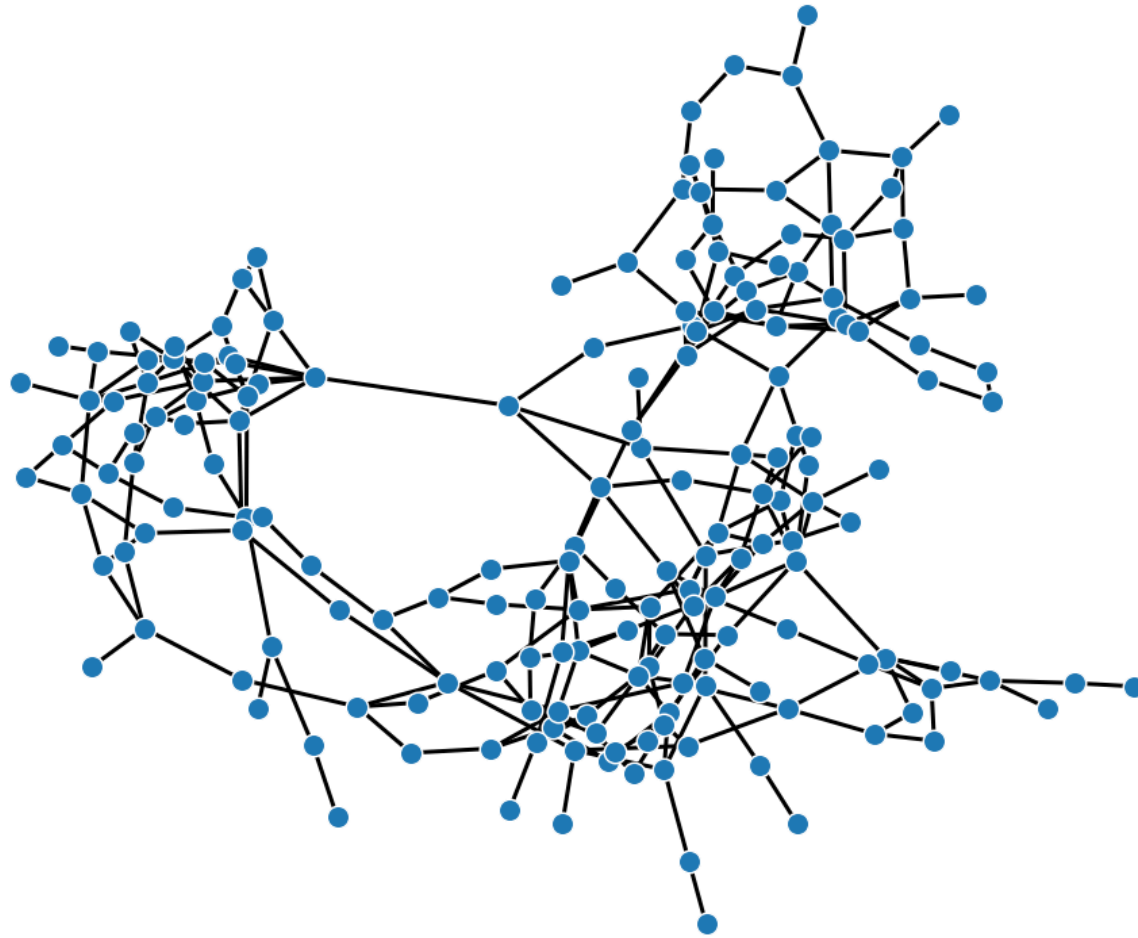
- Empirical network data often do not include ground truth community labels.
- We may want to find clusters—a task known as community detection. E.g., proteins that interact, genes that are co-expressed, clusters of financial transactions, or asset classes in financial markets can form communities.

SBMS FOR COMMUNITY DETECTION

We treat both the block memberships $\mathbf{x}$ and connection probabilities $\boldsymbol{\lambda}$ as parameters of the model:

$$A_{ij} \sim \text{Bernoulli}\left(\lambda_{x_i x_j}\right)$$
$$\lambda_{rs} \sim \text{Beta}\left(1, 1\right)$$
$$x_i \sim \text{Uniform}\left(1, q\right)$$

GIBBS SAMPLER FOR COMMUNITY DETECTION

We can sample from the conditional distributions

$$\mathbf{x} \mid \mathbf{A}, \boldsymbol{\lambda} \sim \text{Categorical} (\dots)$$

$$\boldsymbol{\lambda} \mid \mathbf{x}, \mathbf{A} \sim \text{Beta} (\dots)$$

to obtain a Gibbs sampler for community detection.

APPLICATIONS OF NETWORK MODELS

- Answering science questions using domain-specific models, e.g., studying gene co-expression.
- Predict links using $p\left(\mathbf{A}^{\mathrm{pred}} \mid \mathbf{A}\right)$. Or identify surprising edges.
- Predict attributes using $p\left(\mathbf{x}^{\mathrm{pred}} \mid \mathbf{A}, \mathbf{x}\right)$.
- Sensitivity analysis for infectious disease transmission dynamics by replicating networks using $p\left(\mathbf{A}^{\mathrm{repl}} \mid \mathbf{A}\right)$

MECHANISTIC NETWORK MODELS

- All of the previous models are *statistical* network models with tractable likelihood.
- But they cannot (easily) reproduce a number of features of real-world data (like degree distributions or temporal effects).
- They also pose a challenge for assessing policy interventions.
- Another approach is to use *mechanistic* network models whose evolution is described by a set of rules or mechanisms.

SEXUAL CONTACT NETWORK MODELS

```
1  n <- 1000  # Expected number of nodes.
2  mu <- 0.001  # Probability of leaving population.
3  rho <- 0.1  # Prob. to form relationship.
4  sigma <- 0.05  # Prob. for relationship to dissolve.
5  omega0 <- 0.2  # Prob. for casual contact for singles.
6  omega1 <- 0.1  # Prob. for casual contact for paired.
7  T <- 10000 # Number of simulation steps.
```

- This code block defines parameters for a mechanistic network model for men who have sex with men from 10.1016/j.epidem.2019.02.001.

SEXUAL CONTACT NETWORK MODELS

```
1   graph <- empty()
2   for (t in 1:T) {
3       graph <- remove_nodes_with_prob(graph, mu)
4       graph <- remove_steady_with_prob(graph, sigma)
5       graph <- remove_casual(graph)
6
7       n_new_nodes <- rpoisson(1, mu * n)
8       graph <- add_new_nodes(graph, n_new_nodes)
9       singles <- get_nodes_with_zero_degree(graph)
10      graph <- add_steady_with_prob(graph, singles, rho)
11      singles <- get_nodes_with_zero_degree(graph)
12      graph <- add_casual_with_prob(
13          graph, singles, omega0)
14      paired <- get_nodes_with_nonzero_degree(graph)
15      graph <- add_casual_with_prob(paired, omega1)
16
17      # Disease dynamics ...
18  }
```

BENEFITS

- We can easily run sensitivity analyses.
- We can simulate intervention policies.
- The model is easily interpretable.

INFERENCE CHALLENGES

- Even though we use simple building blocks, the likelihood of the model is not tractable.
- Markov chain Monte Carlo methods are not applicable because we cannot evaluate the posterior density.
- Inference for these models is an open challenge and the topic of research in simulation-based inference.

RECAP AND FURTHER TOPICS

- What are networks and why should we care?
- Statistical network models: Erdős–Rényi, (soft) random geometric graphs, conditionally independent edge models, logistic regression, stochastic block models.
- Mechanistic models, their advantages, and challenges.
- Further topics: Directed vs undirected networks, temporal networks, multilayer networks, hypergraphs, …