Inferring Groups from Communication Data

Chris DuBois
Dept. of Statistics

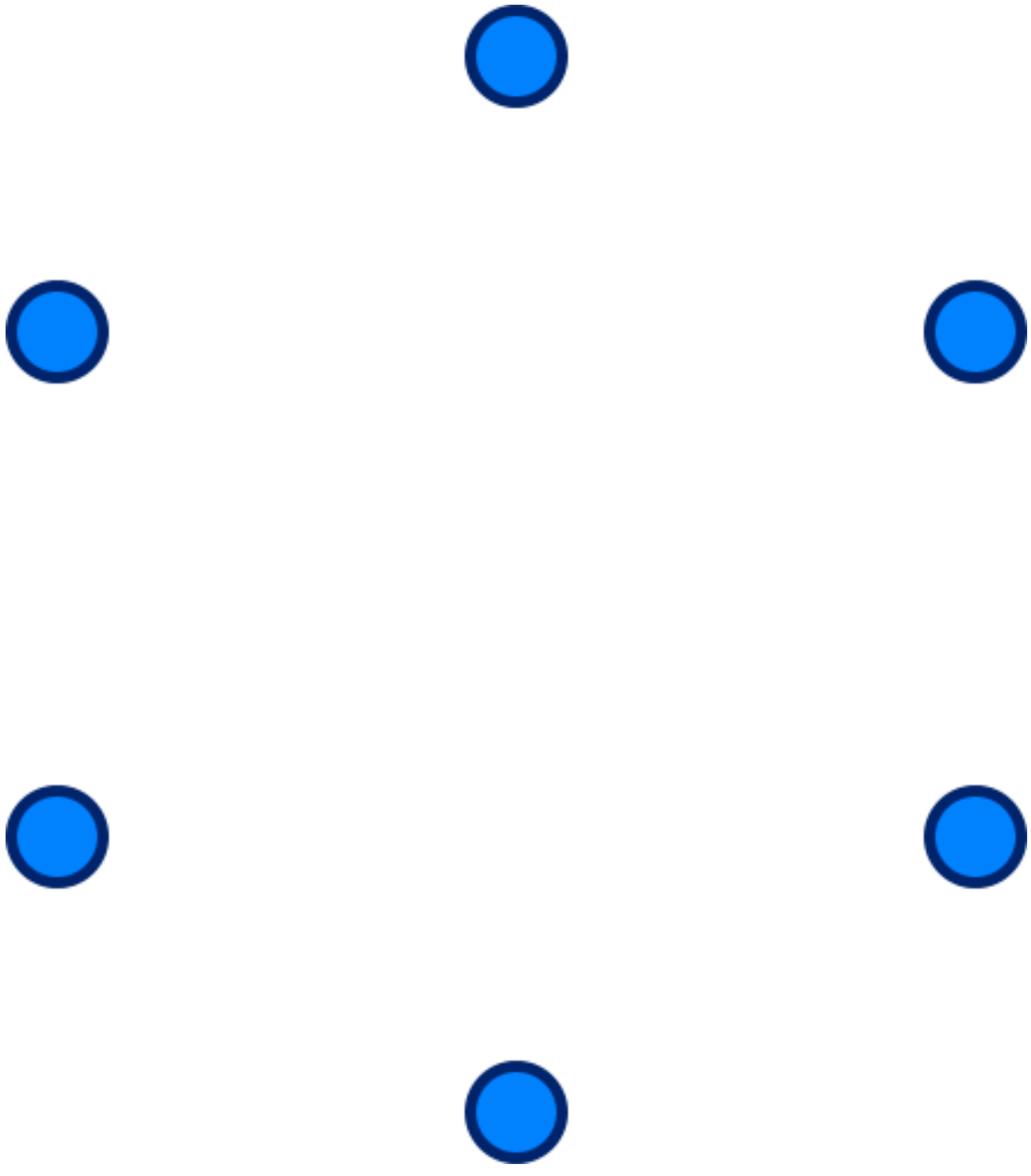
Padhraic Smyth
Dept. of Computer Science

MURI Group Meeting
November 13, 2010

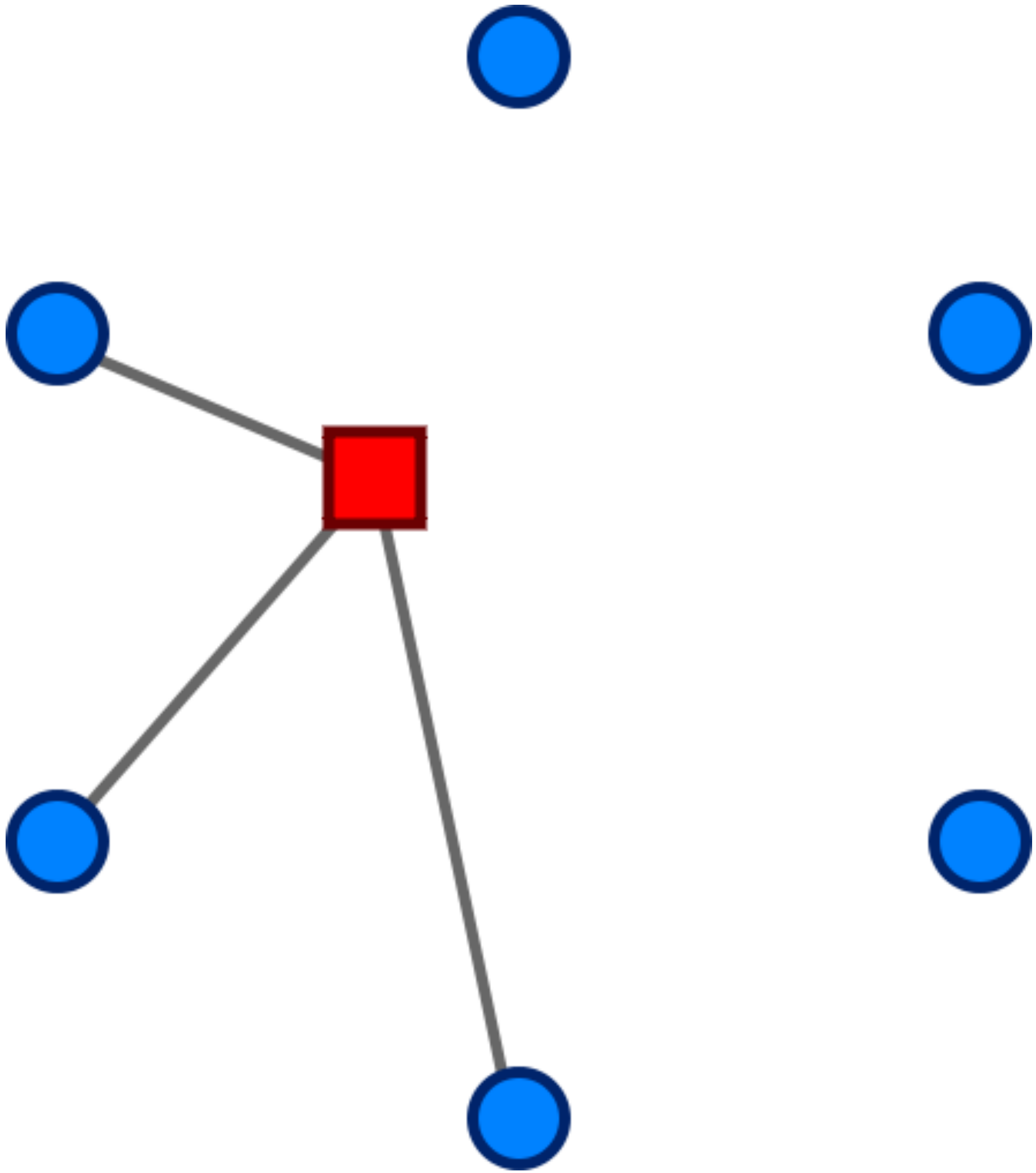
Outline

- Communication data as co-appearance data
- Inferring groups: theory and applications
- Statistical approach: latent variable modeling
- Quick illustration
- Application: large-scale email analysis

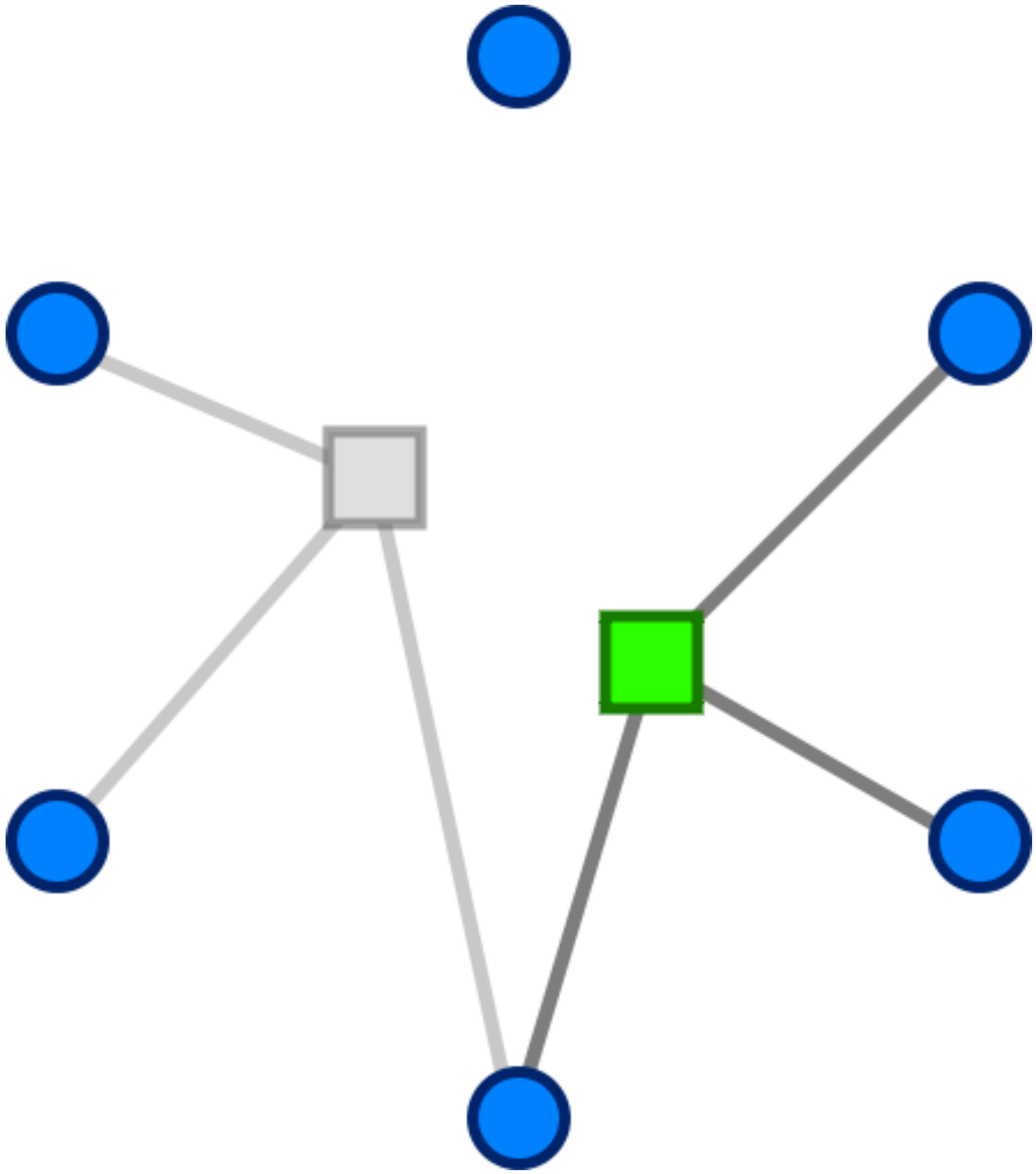
Co-appearance Data



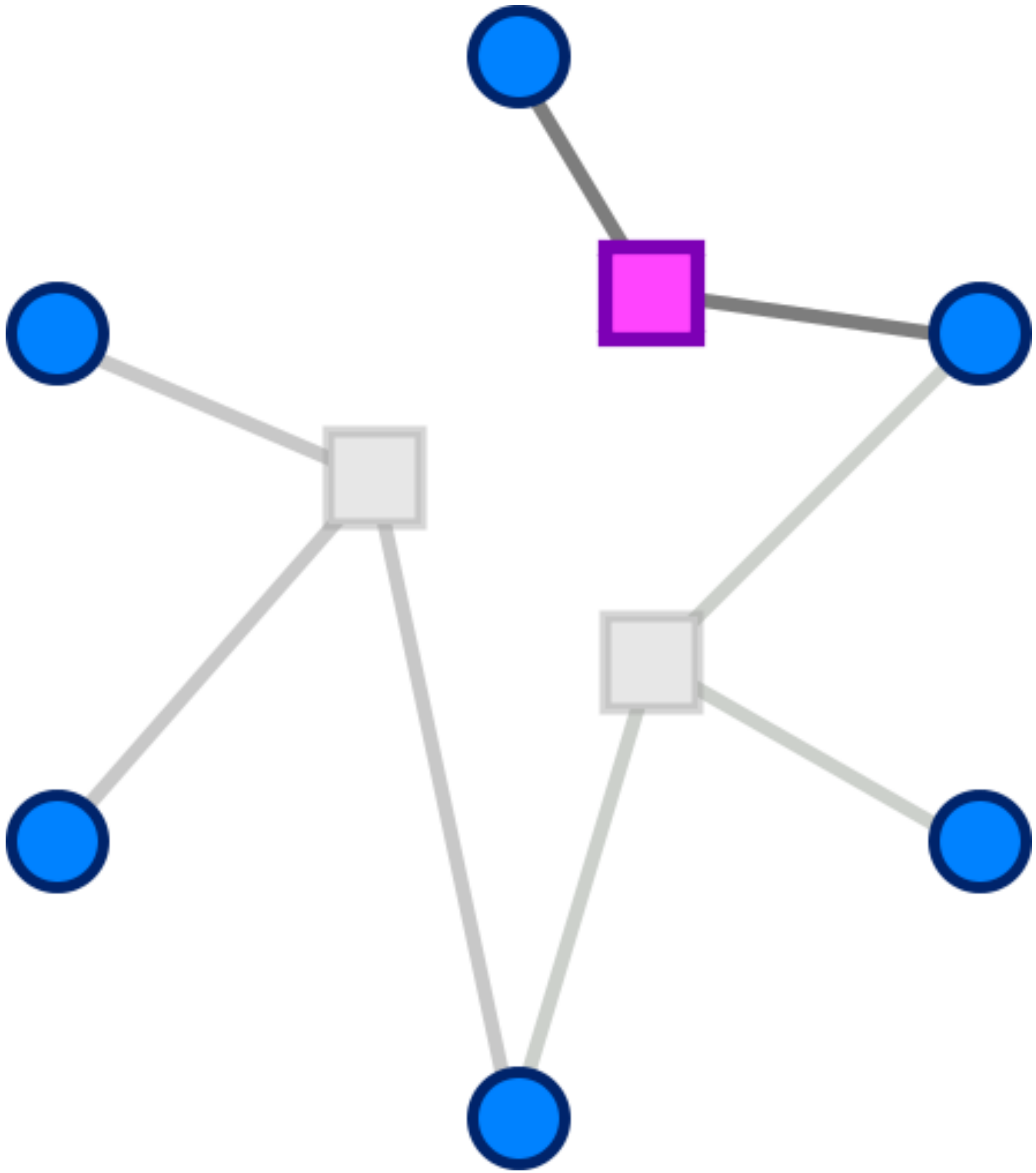
Co-appearance Data



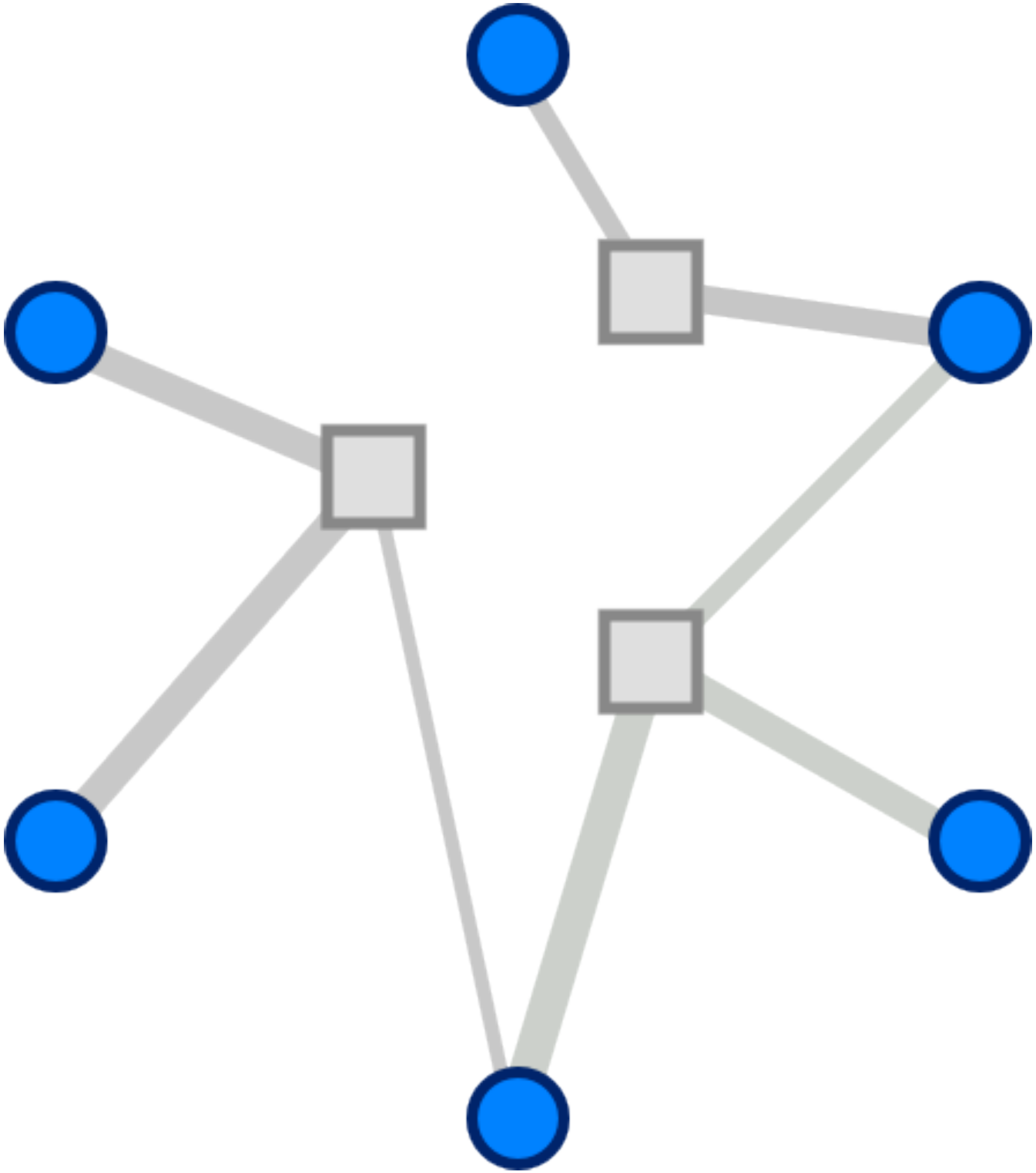
Co-appearance Data



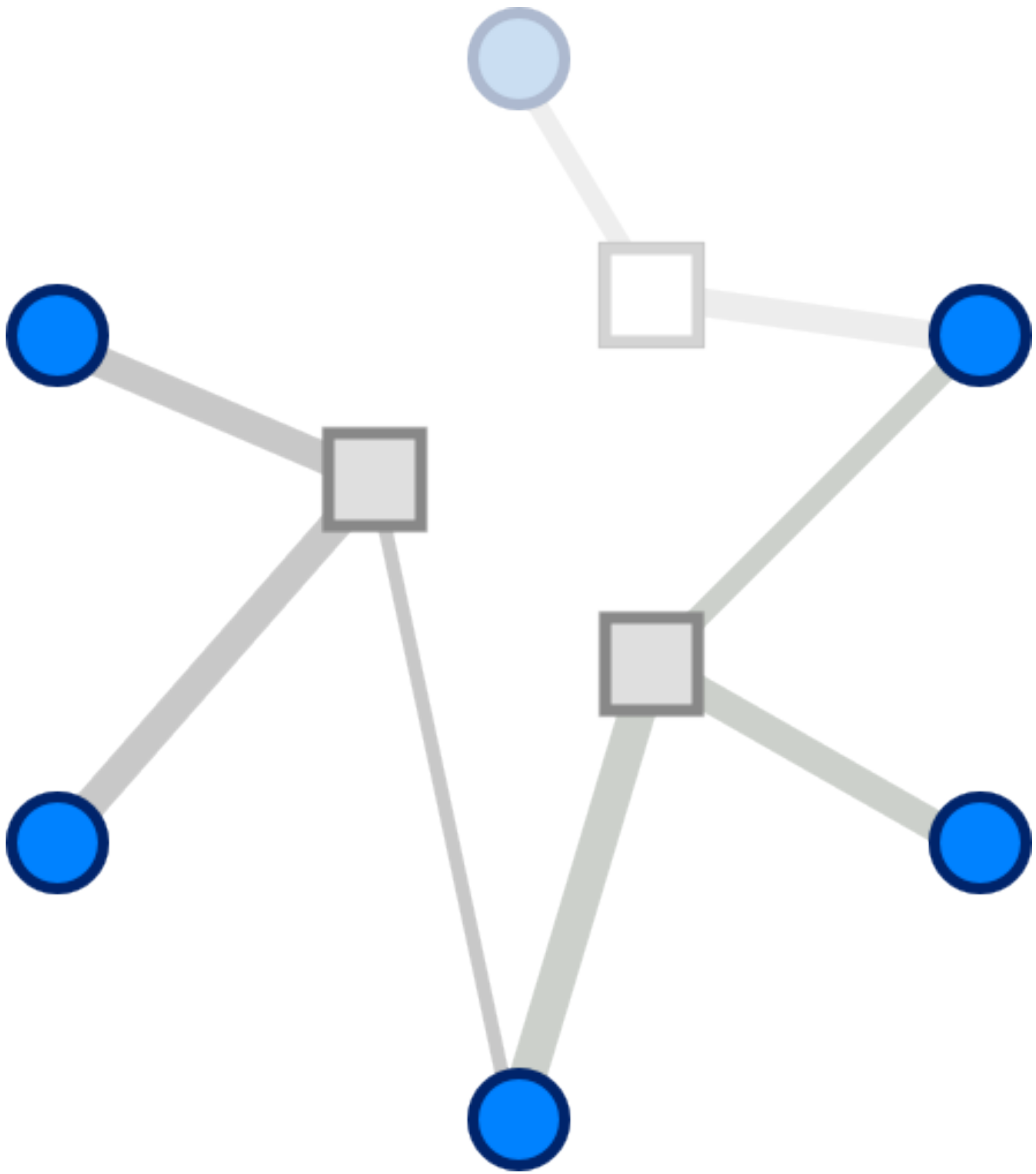
Co-appearance Data



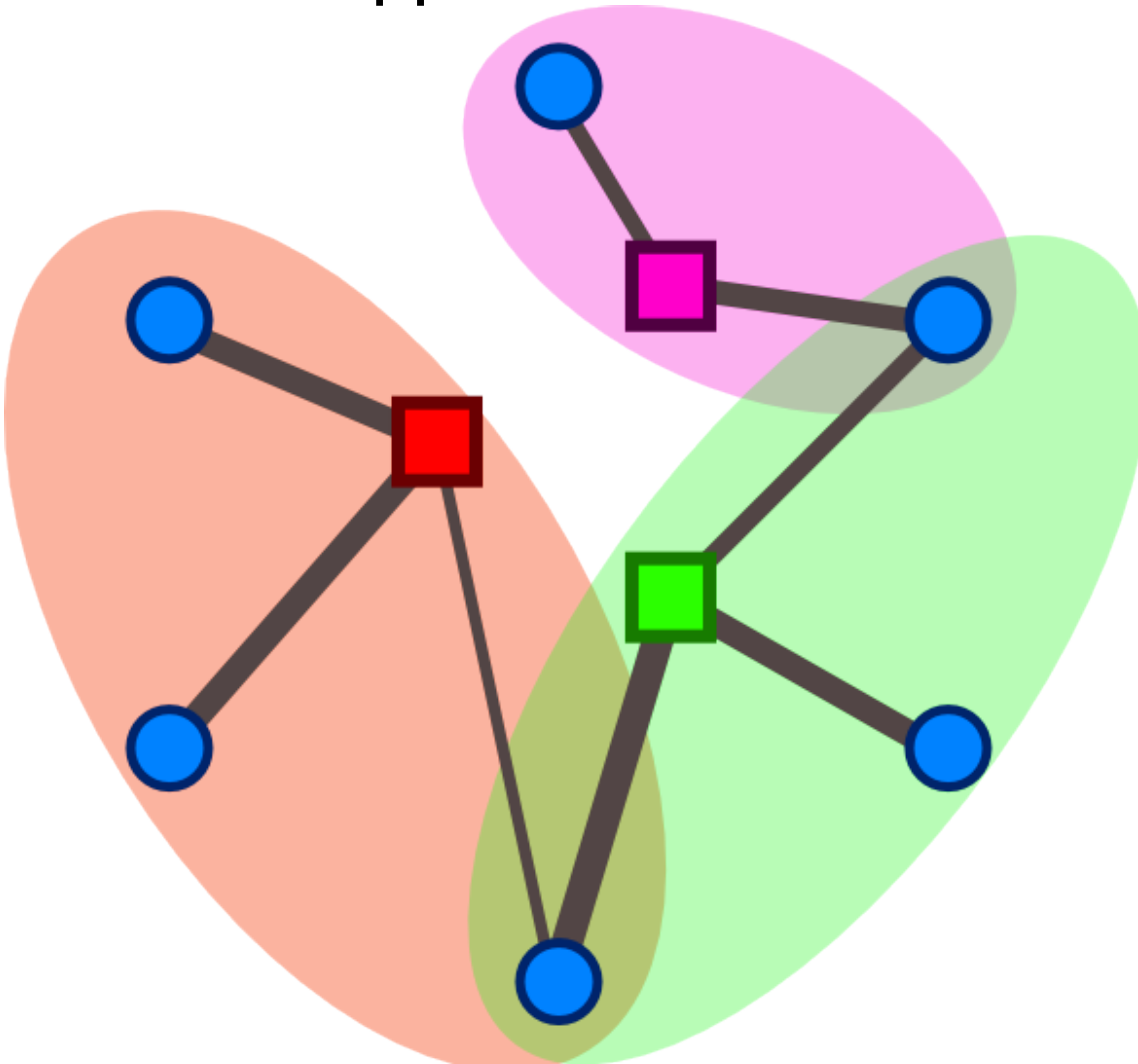
Co-appearance Data



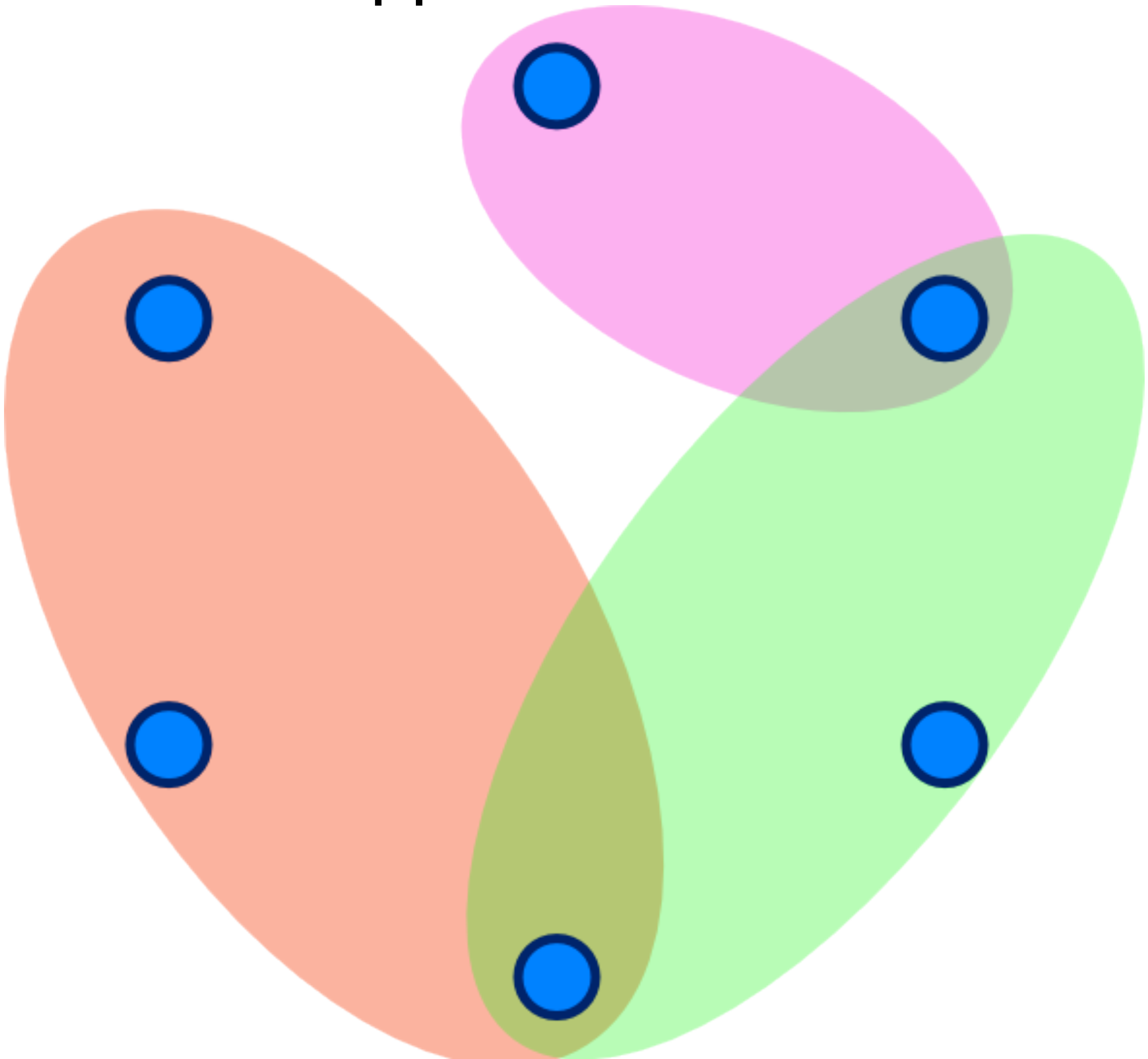
Co-appearance Data



Co-appearance Data



Co-appearance Data



Sociological motivation for latent sets

Theoretical foundations:

- **Simmel**: people's **social identities** defined by their membership to various groups (e.g. family, occupation, neighborhood, other organizations)

- **Feld**: shared **foci** help explain dyadic interactions among actors (e.g. activities and interests, either known or unknown)

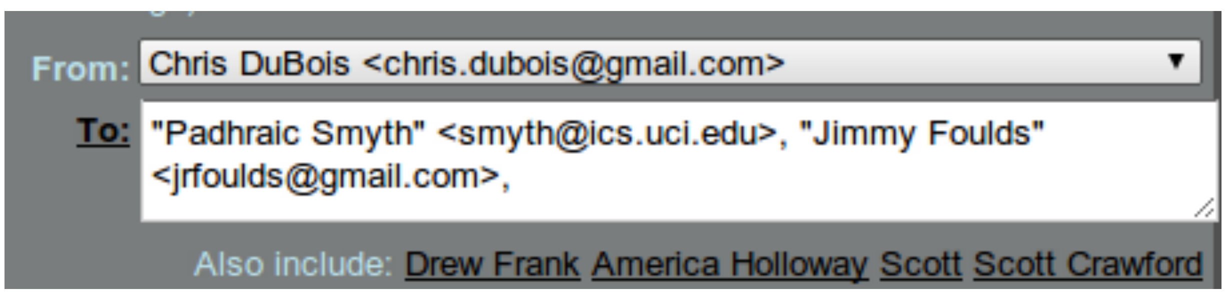
- **Homans**: **groups** of people (partially) defined by interactions

Takeaway: a fair amount of intuition behind the idea of (possibly overlapping) latent sets

Practical application: email services

Prediction of other possible recipients on an email

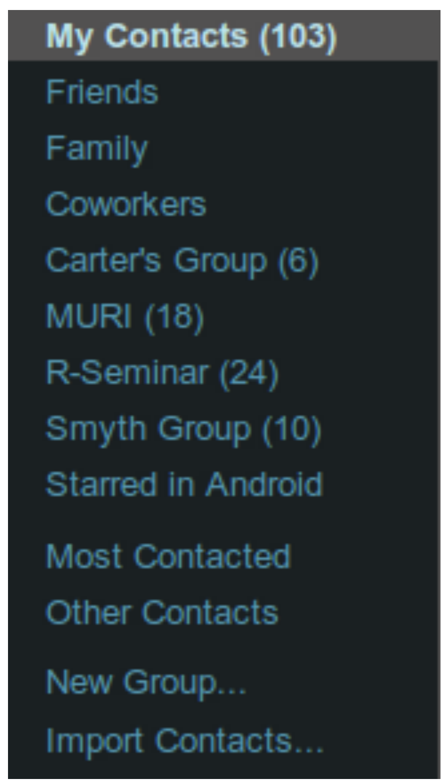
- Favorable response to Gmail's experimental tools, "What about Bob?" and "Wrong Bob?"



Practical application: email services

Automatic group detection

- People are unwilling to manually create groups
- People prefer to interact differently with separate social groups (e.g. work / family)



Statistical models for network data

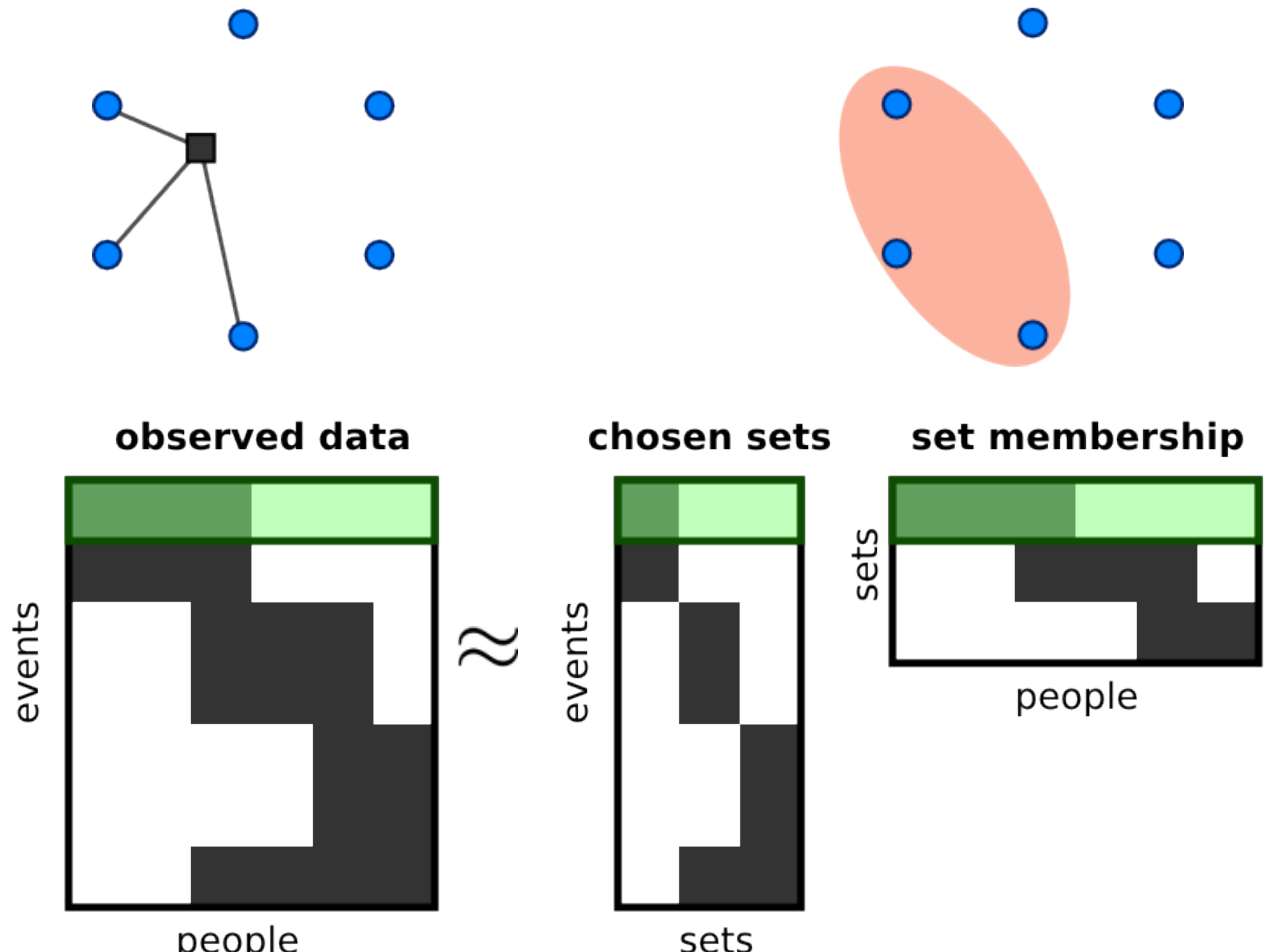
Goals:

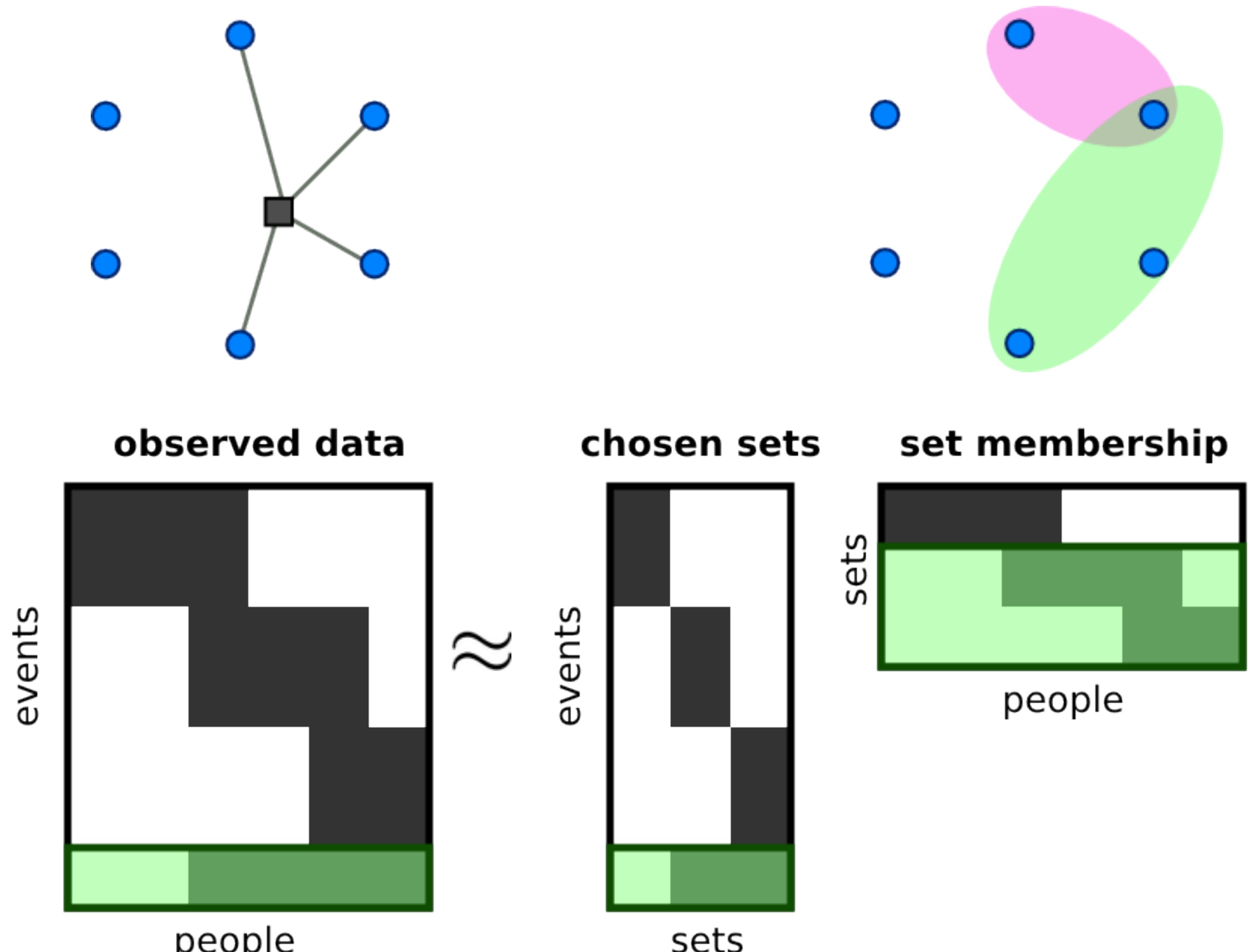
- Make predictions about missing or future data
- Explore scientific hypotheses
- Do the above in a general and principled framework

... even if we have ...

- missing data
- sparse data
- either egocentric or global data
- additional covariates about actors and/or events
- large, dynamic datasets

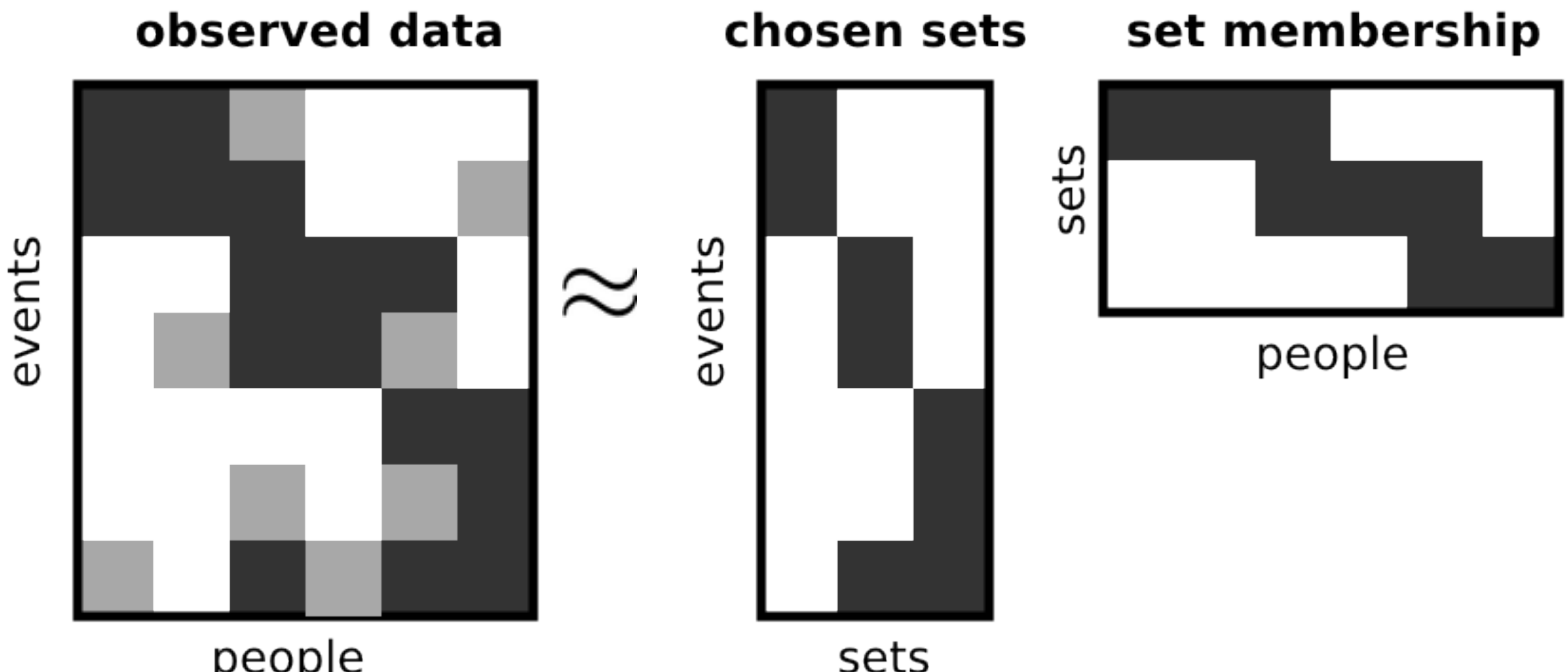
Model Development


observed data
events
people
≈
chosen sets
events
sets
set membership
sets
people


observed data
chosen sets
set membership
events
people
≈
events
sets
sets
people

Probabilistic Model

$$\Pr(y_{ij} = 1) = 1 - \prod_{k=1}^{K} (1 - \omega_k)^{w_{ik} z_{jk}}$$



Probabilistic Model

$$\Pr(y_{ij} = 1) = 1 - \prod_{k=1}^{K} (1 - \omega_k)^{w_{ik} z_{jk}}$$

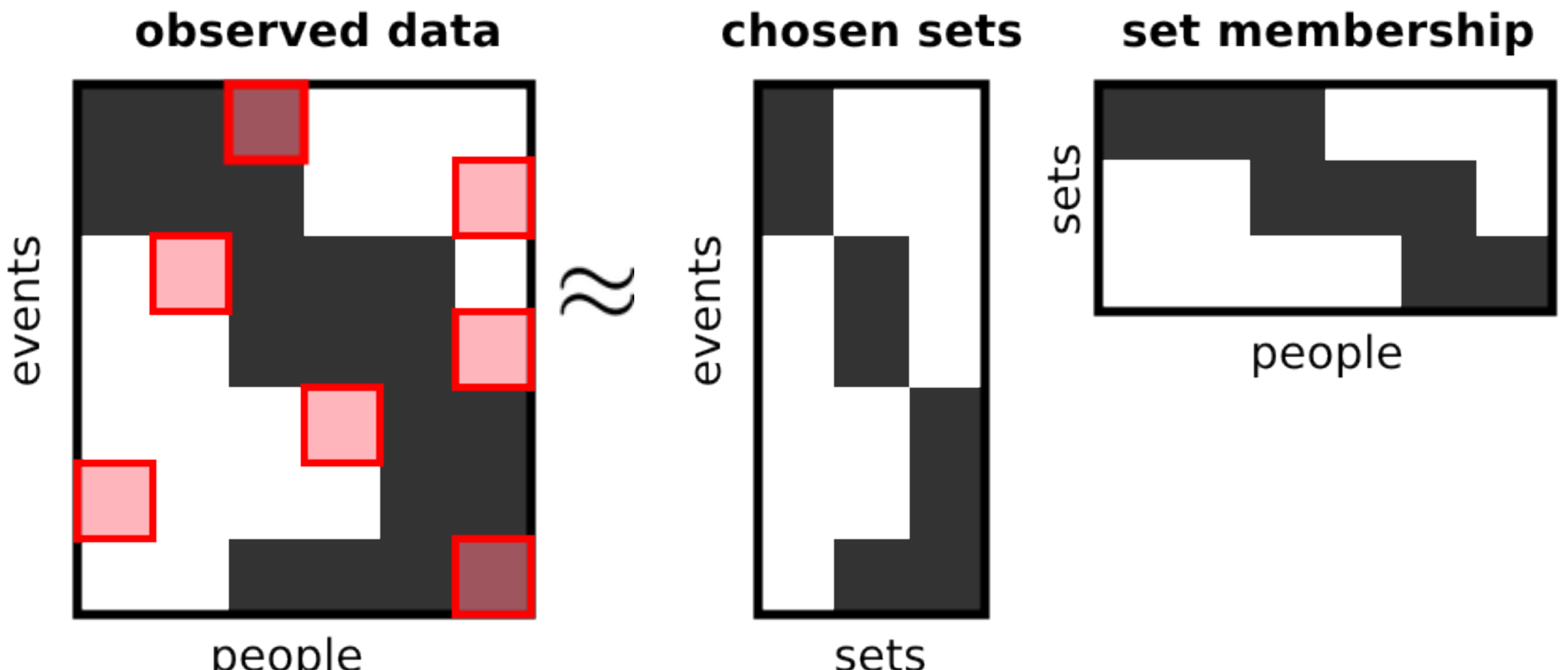


Illustration: Davis' Southern Women

- Perfect for exploring the utility of a new method aimed at two-mode data

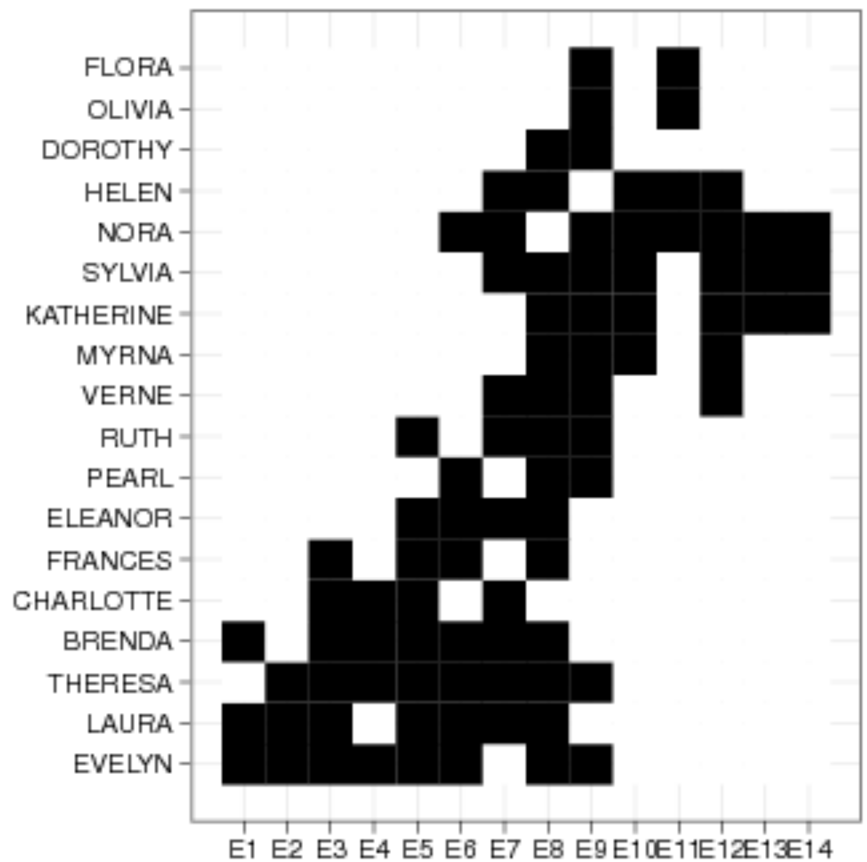
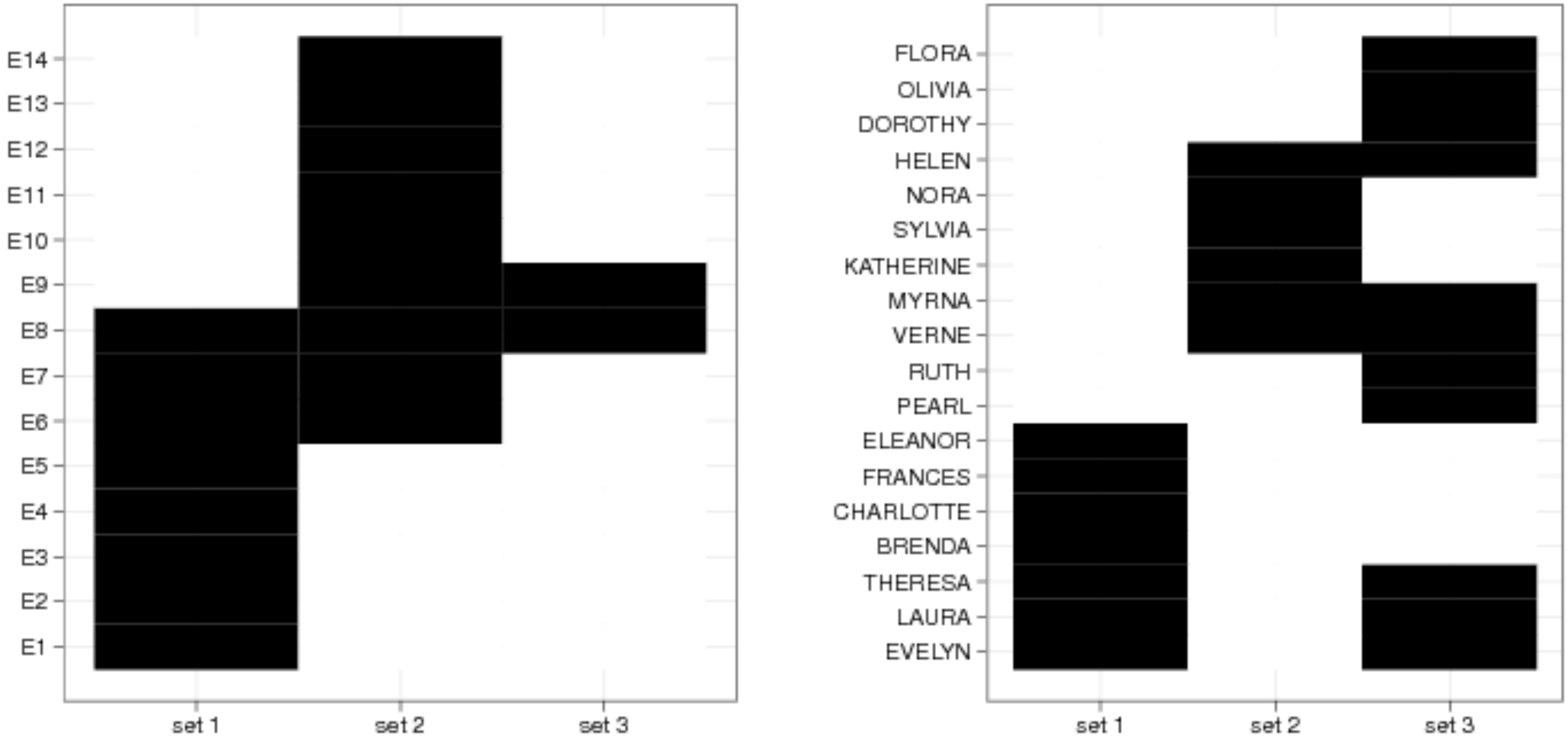
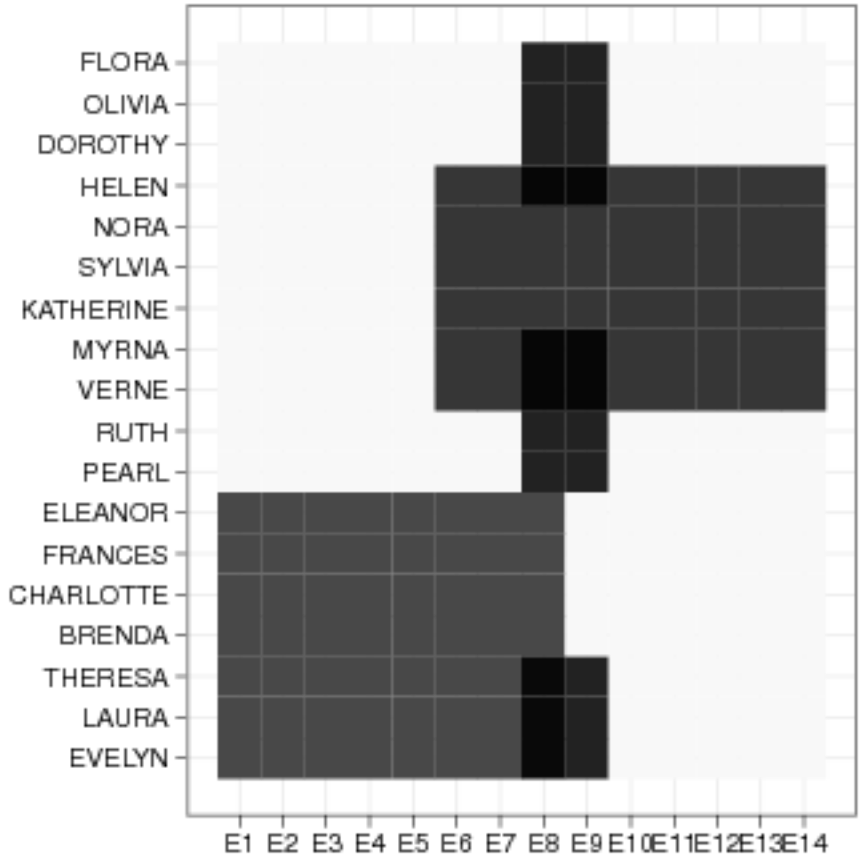


Illustration: Davis' Southern Women

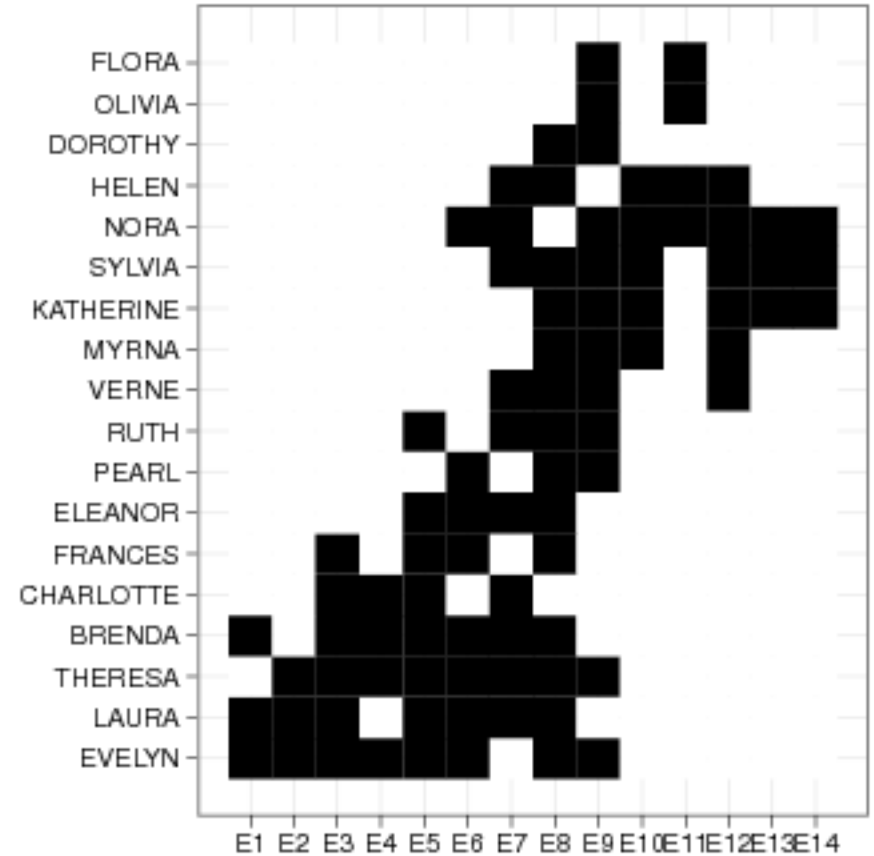


A single sample of W (left) and Z (right).

Illustration: Davis' Southern Women

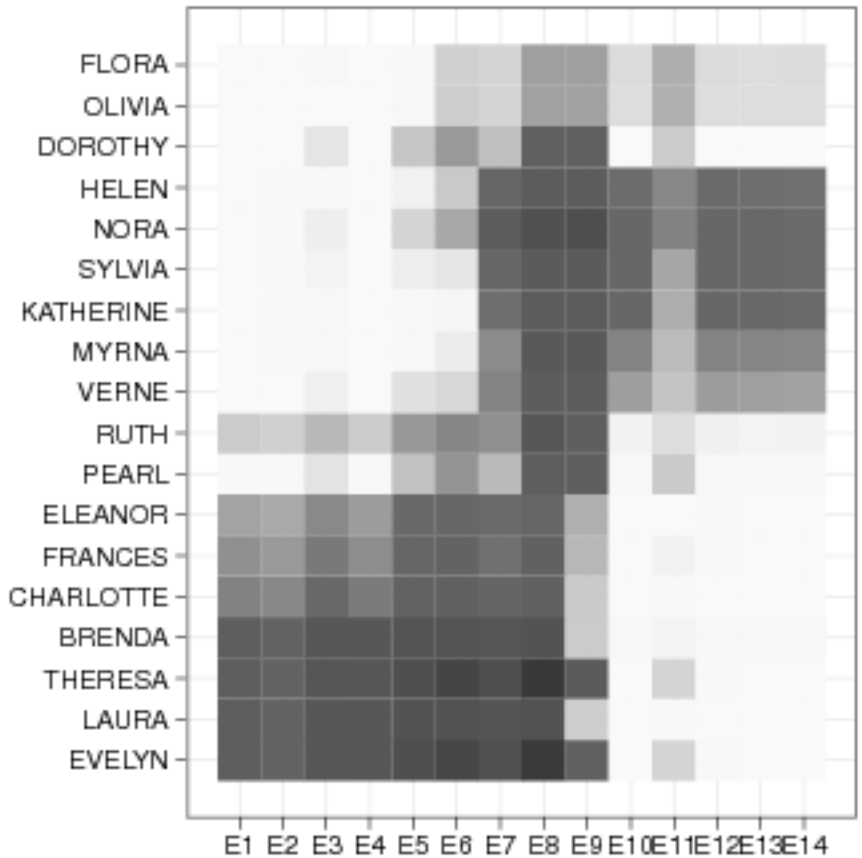


P(Y | W, Z, omega)

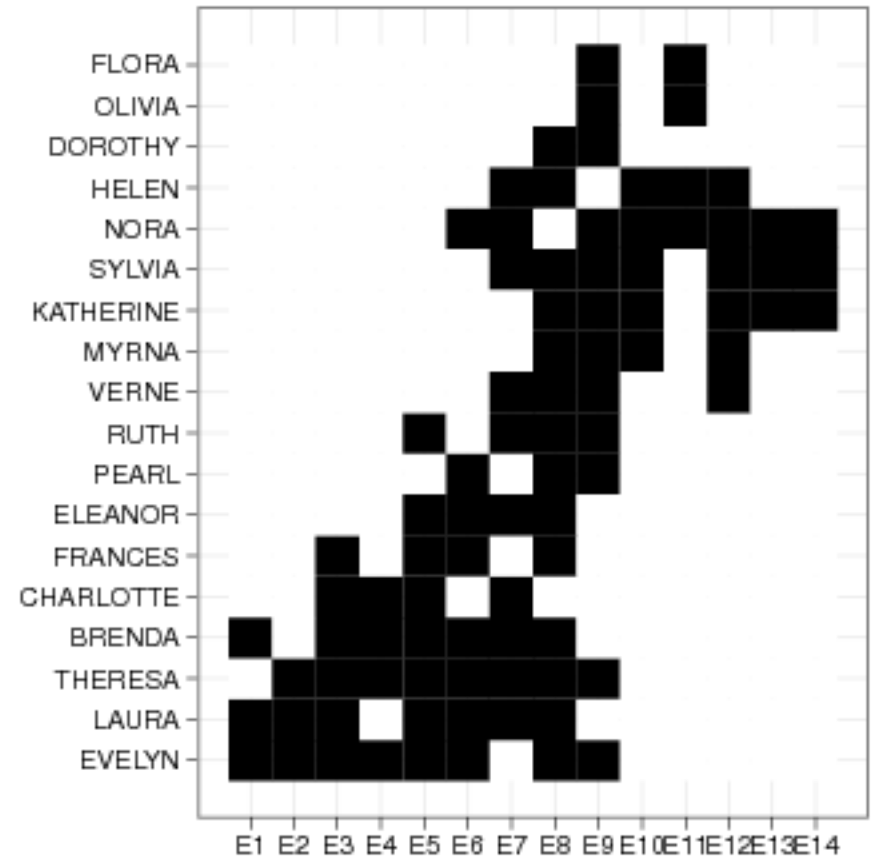


Observed Data

Illustration: Davis' Southern Women

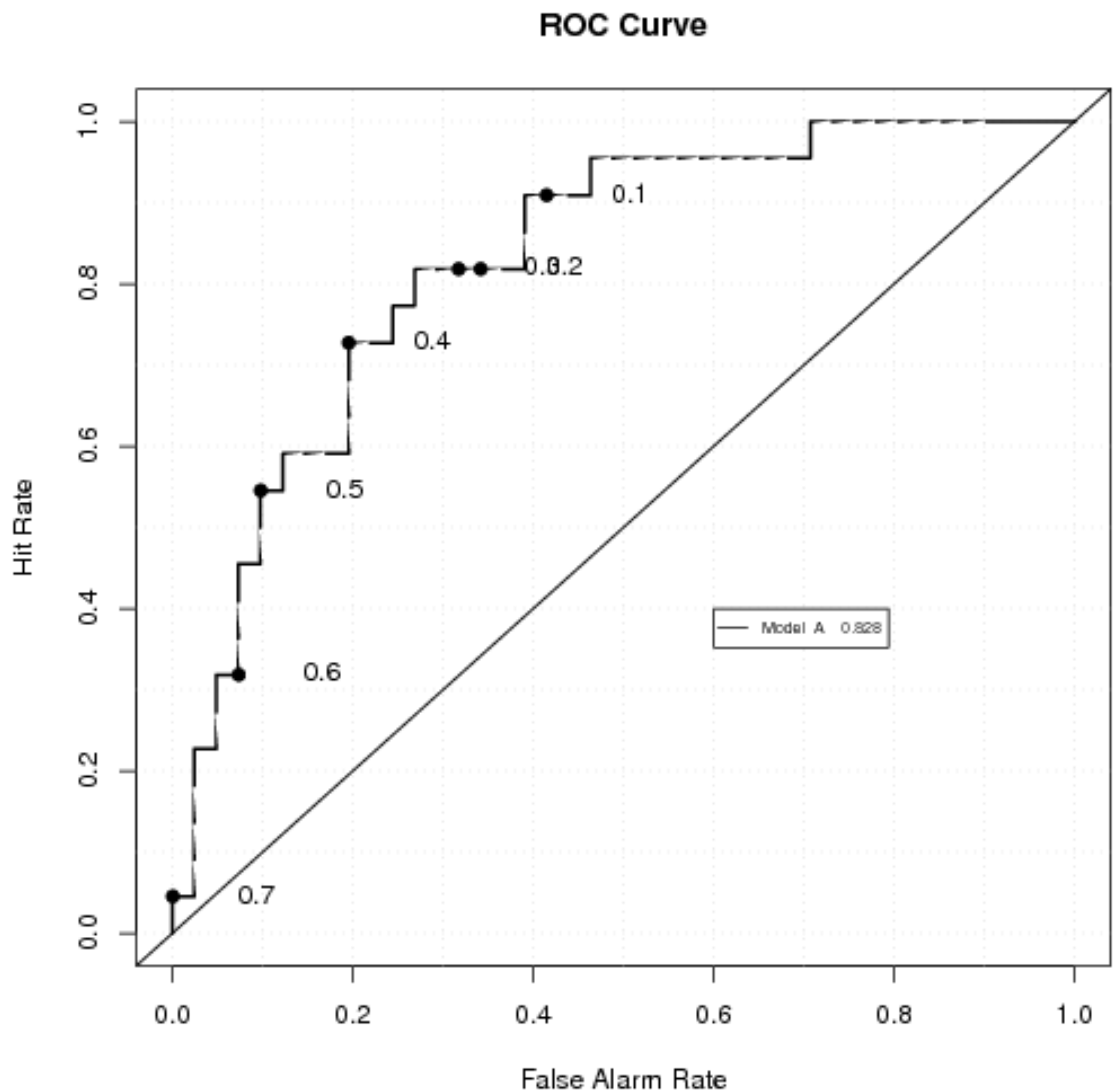


Estimate of posterior predictive distribution



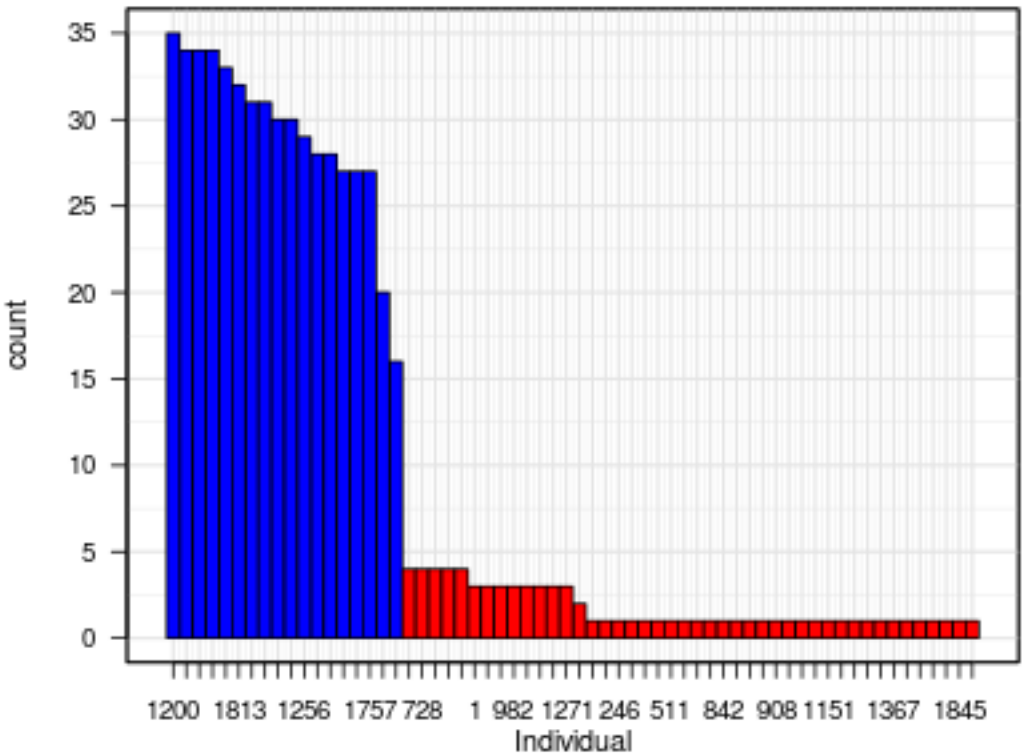
Observed Data

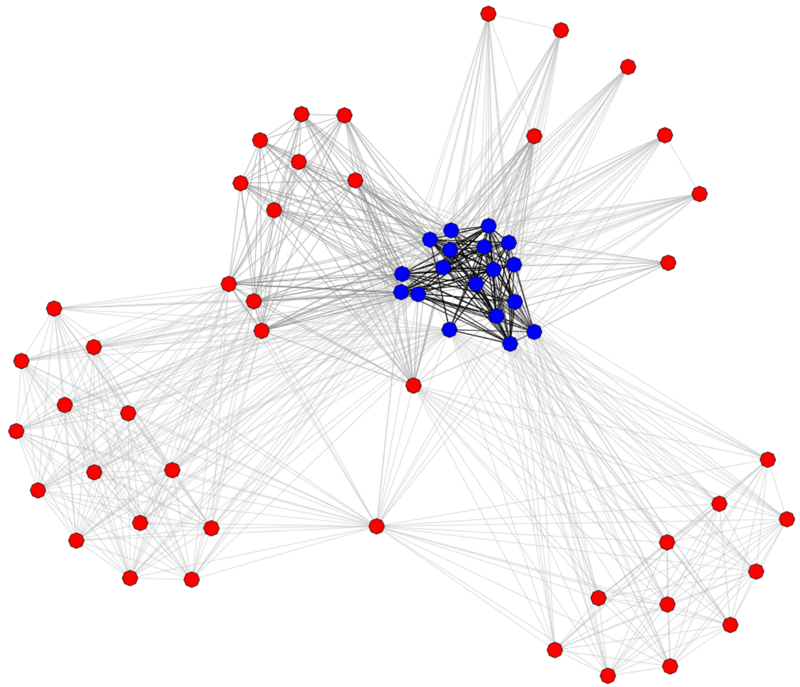
Missing data experiment on Davis



Prediction performance with 25% of dyads missing

Groups in the Eckmann Email Data

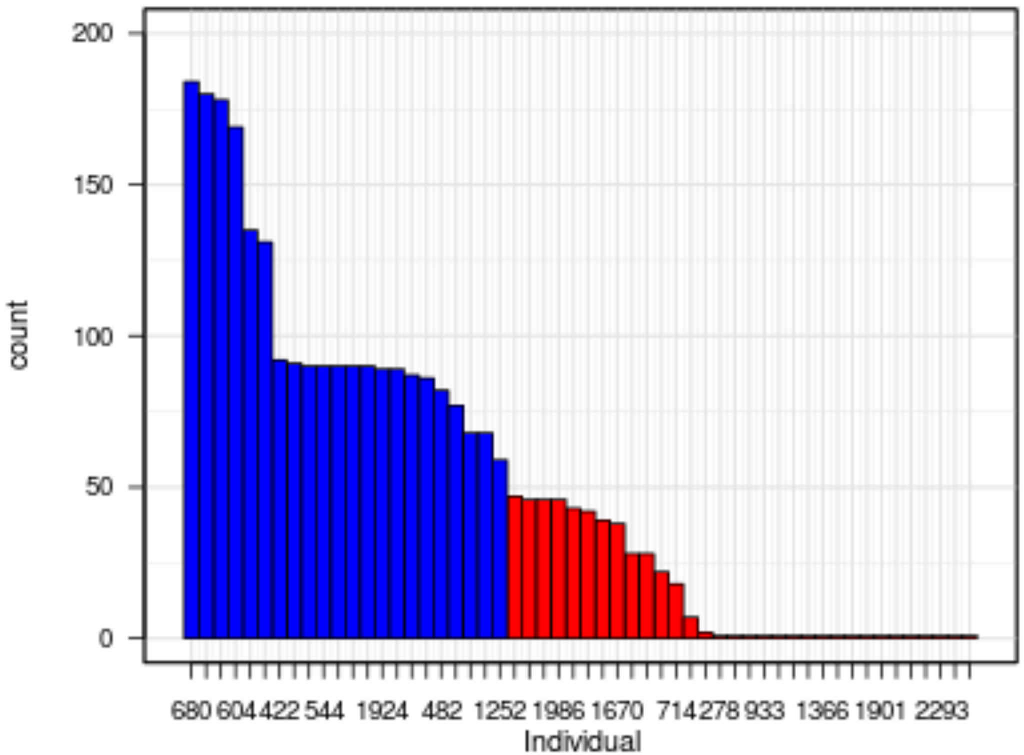


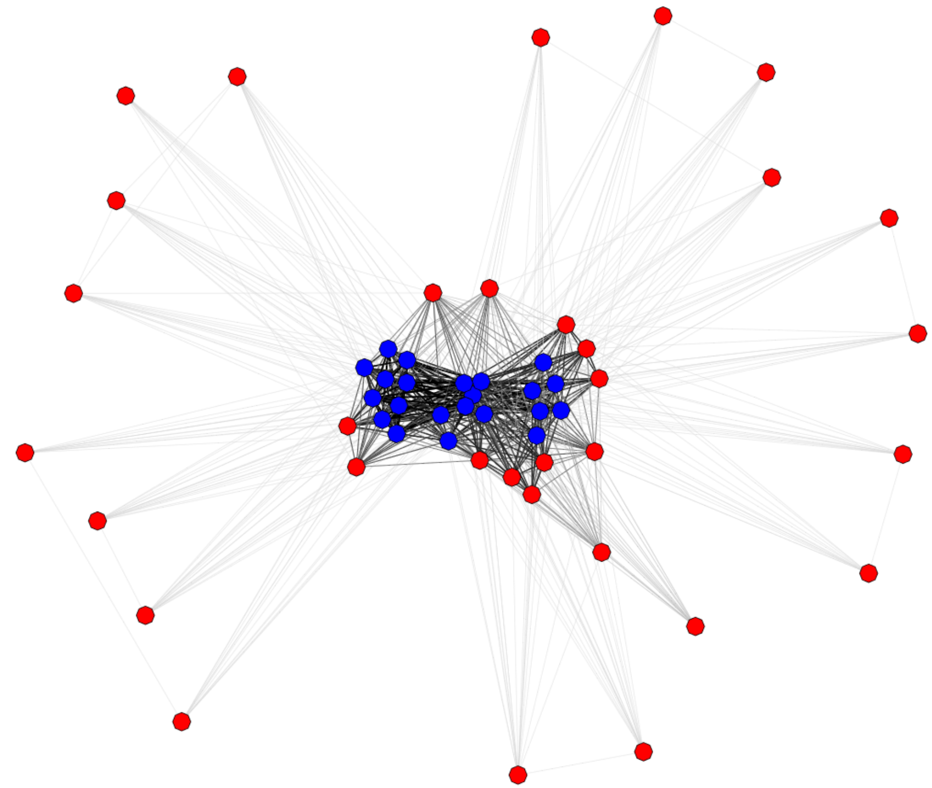


- Number of emails per person where set k is "active".
- Members of set k colored blue.

- Dark grey edges indicate higher counts (log scale).
- Members of set k colored blue.

Groups in the Eckmann Email Data





- Number of emails per person where set k is "active".
- Members of set k colored blue.

- Dark grey edges indicate higher counts (log scale).
- Members of set k colored blue.

Advantages of this approach

- Latent set models a natural choice for co-appearance data

- Validate predictively

- Allows missing data and egocentric data

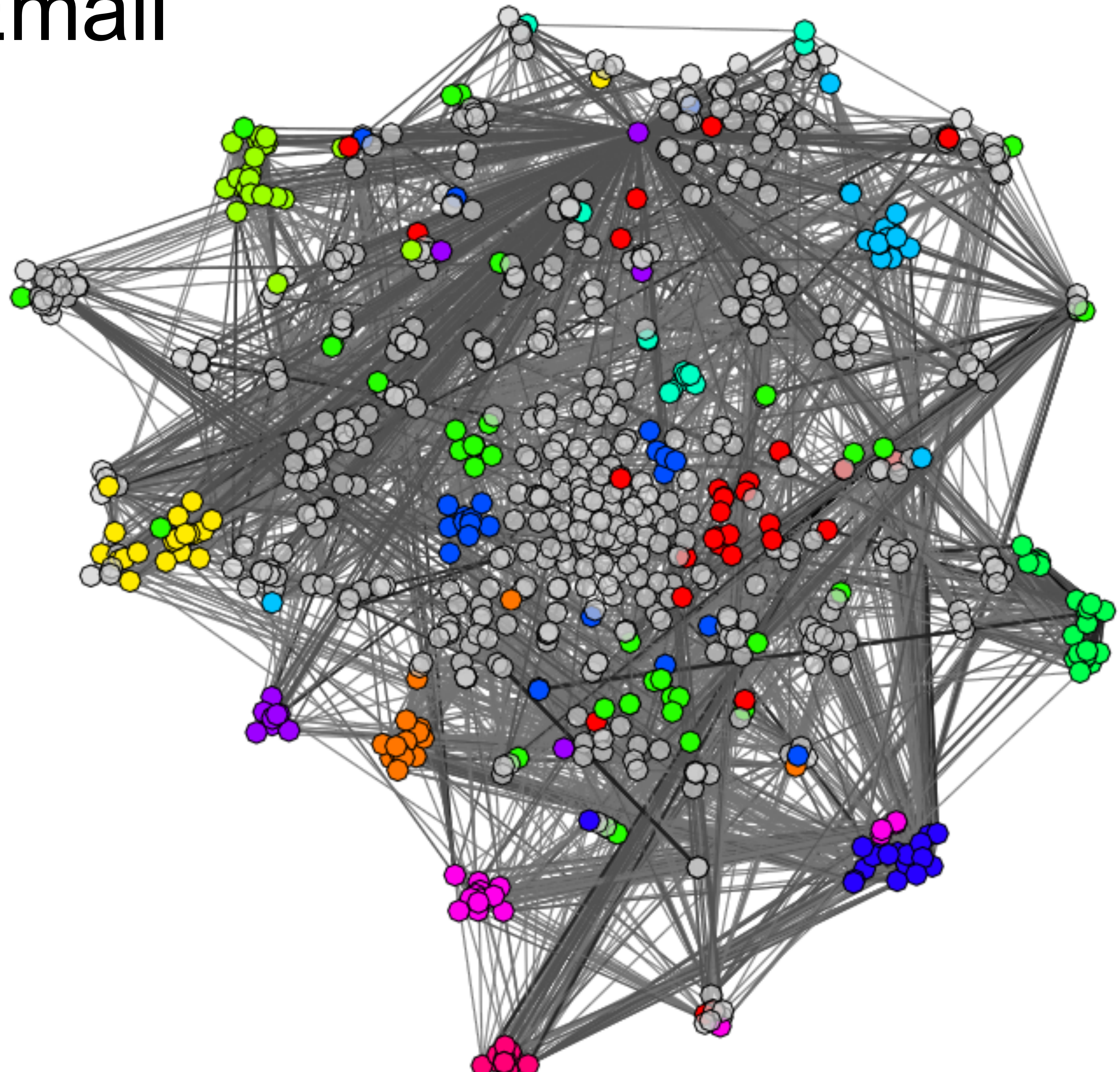
- Interpretable model estimates
    - Inferred groups of actors
    - Actors within each set likely to appear together

- Scalable

Thanks

Extra slides

Eckmann Email



Model Development

Assume T events, N actors, K latent sets

Unknown variables

- Z: binary NxK matrix indicates set memberships
- W: binary TxK matrix indicates each event's "active" sets
- omega: vector of K reals.

Noisy OR:

$$\Pr(y_{ij} = 1) = 1 - \prod_{k=1}^{K} (1 - \omega_k)^{w_{ik} z_{jk}}$$

Interpretation of omega:

- probability actor j is present for event i when j is in set k and only set k is active

Inference

- Data augmentation
- EM: tough to analytically compute expectation step because W and Z depend on each other
- Markov chain Monte Carlo
  - Gibbs sampling: sample a variable conditioned on everything else (NB: can integrate out a few things)
  - Iteratively sample W matrix, Z matrix, and omegas
  - Make predictions by averaging over samples
- Beware of local modes!
  - Initializing with hierarchical clustering or kmeans seems to work well in practice