Topic-Conditioned Hierarchical Latent Space Models for Text-Valued Networks

Matthew J. Denny[†1], James ben-Aaron[†2], Hanna Wallach[†‡3], Bruce A. Desmarais[†4]

[†]University of Massachusetts Amherst; [‡]Microsoft Research | [1]mdenny@polsci.umass.edu; [2]jbenaaro@acad.umass.edu; [3]hanna@dirichlet.net; [4]desmarais@polsci.umass.edu

Research Objectives

- Jointly model content and structure of communication.
- Examine communication patterns in local government.
- Introduce local government email dataset and R package.

Motivation

Structure Matters | Identify Bias

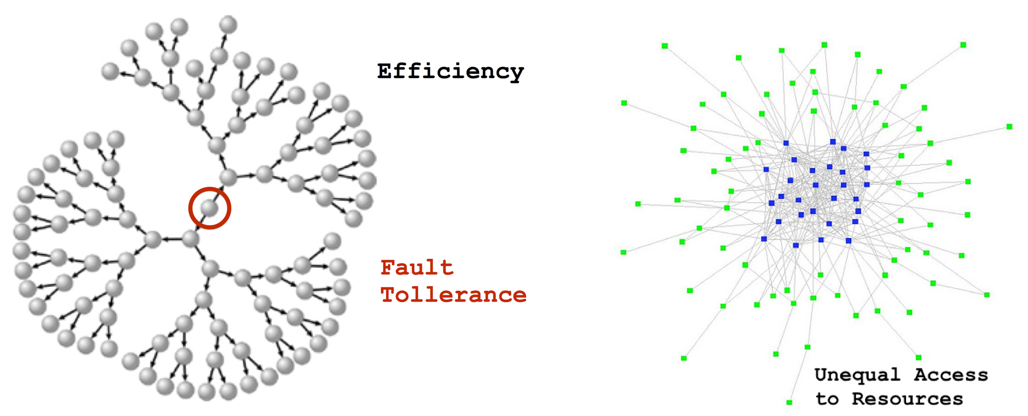
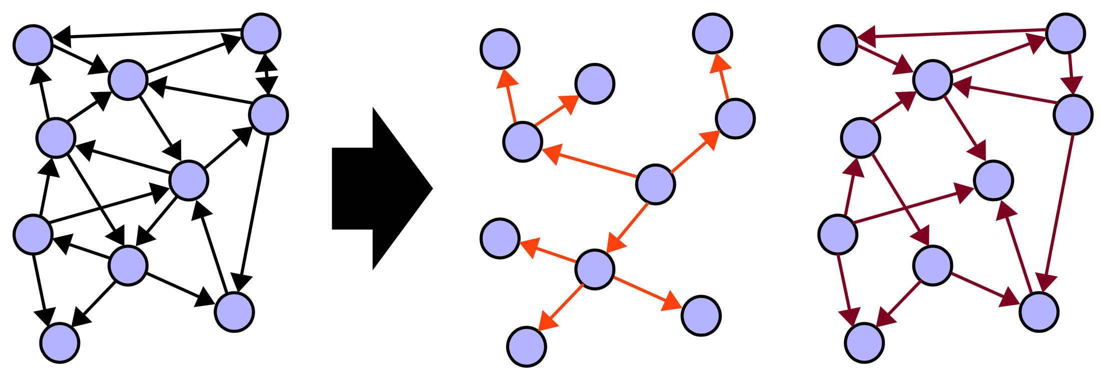


Figure: Different content-conditional structures may underlay a communication network.



Existing Approaches

Latent Space Models:

- Each node represented by a k-vector $\mathbf{s} = \{s_1, s_2, \ldots, s_k\}$
- d_{ij} is the distance between the attributes of nodes i and j.

$$d_{ij} = \sqrt{\sum_{h=1}^{k} \left(s_h^{(i)} - s_h^{(j)}\right)^2}$$

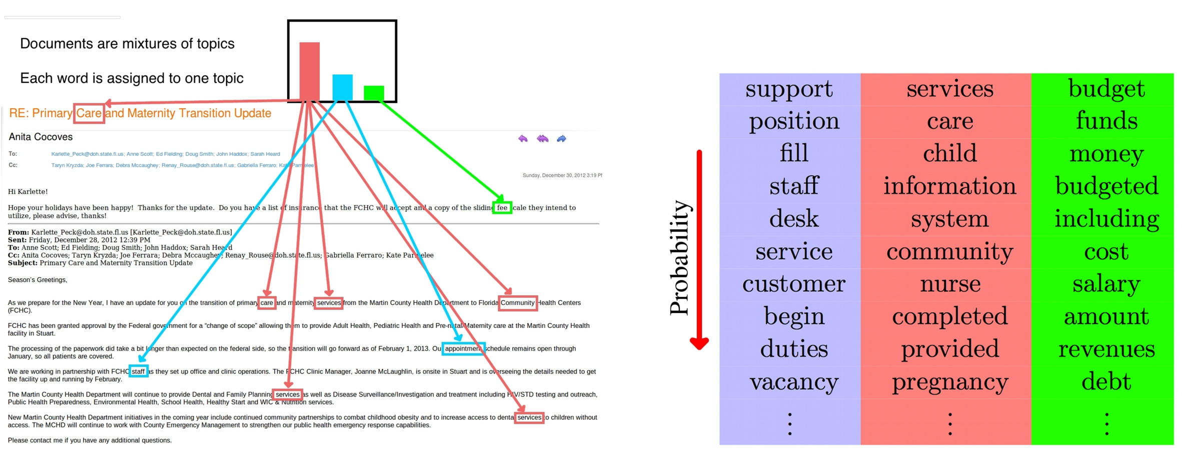
- Then the probability of an edge from i to j is

$$p_{ij} = \text{logit}^{-1}(b_0 - d_{ij}), \quad \text{logit}^{-1}(x) = \frac{1}{1+\exp(-x)}$$

- b_0 controls density and the $\mathbf{s}$ models specific connections.
- Adding in covariates.

$$p_{ij} = \text{logit}^{-1}(b_0 + b_1 x_{ij} - d_{ij})$$

Latent Dirichlet Allocation:

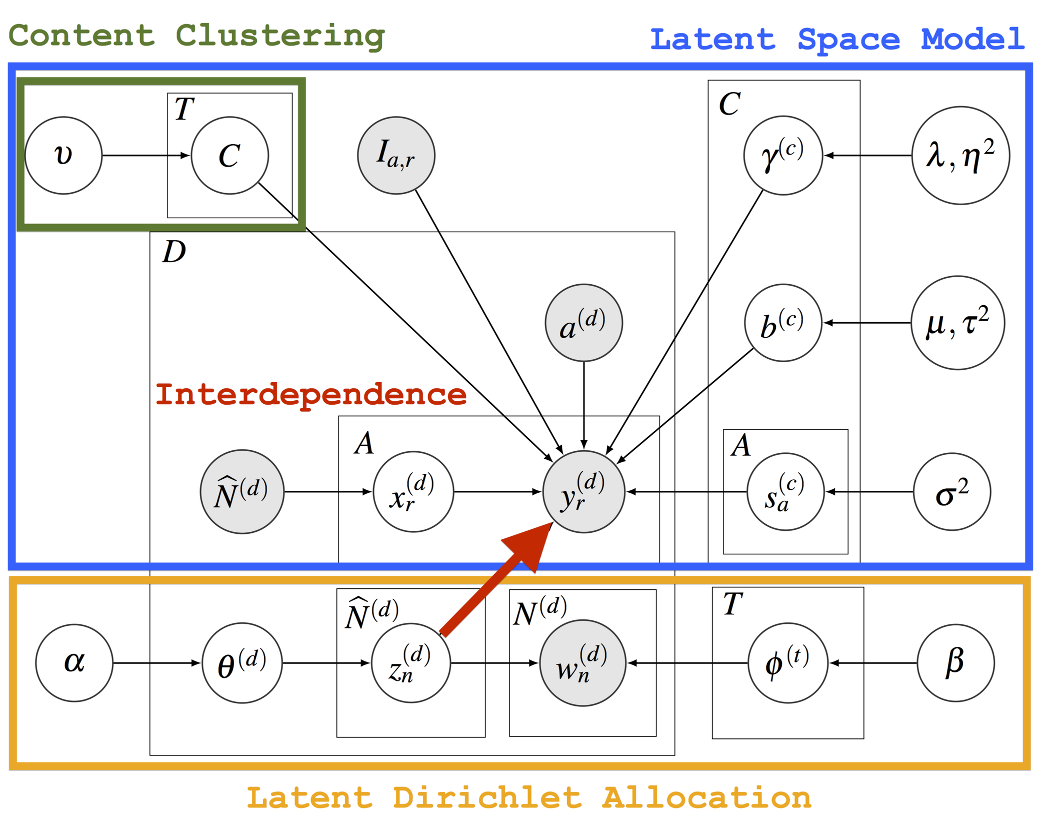


- Corpus consists of d documents and k topics.
- Topic distribution for a document (d) is given by $\theta \sim \text{Dir}(\alpha)$
- Topic $z_n \sim \text{Multinomial}(\theta)$ for word w_n
- Each word is drawn from a multinomial, parameterized by ϕ_z where $\phi \sim \text{Dir}(\beta)$.

$$P(z_i = j \mid z_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta/W}{n_{-i,j}^{(\cdot)} + \beta} \left(\frac{n_{-i,j}^{(d_i)} + \alpha/T}{n_{-i}^{(d_i)} + \alpha}\right) \tag{1}$$

Generative Model

The generative process is represented below as a graphical model where plates represent repeated subgraphs, arrows represent dependencies, and shaded nodes are observed variables.



Inference

Figure: Variable Definitions (observed data are highlighted in grey).

- Message data $\mathcal{D} = \{w^{(d)}, a^{(d)}, y^{(d)}, I_{i,j}\}_d^D$
- Tokens $\mathcal{W}$.
- Message authors $\mathcal{A}$.
- Message recipients $\mathcal{Y}$.
- Edge types $\mathcal{I}$.
- Topic-word distribution Φ.
- Document-topic distribution Θ.
- Topic-cluster assignments C_t.
- Node latent positions $\mathcal{S} = \{S^{(c)}\}_{c=1}^C$.
- Cluster scalar bias terms $\mathcal{B} = \{b^{(c)}\}_{c=1}^C$
- Mixing parameters $\Gamma = \{\gamma^{(c)}\}_{c=1}^C$
- Token topic assignments $\mathcal{Z} = \{z^{(d)}\}_{d=1}^D$
- Edge topic assignments $\mathcal{X} = \{X^{(d)}\}_{d=1}^D$

Sampling Equations:

Token Topic Assignments $\mathcal{Z}$:

$$P(z_n^{(d)} = t \mid w_n^{(d)} = \nu, \mathcal{W}_{\backslash d,n}, \mathcal{A}, \mathcal{Y}, C_t, \mathcal{S}, \Gamma, \mathcal{I}, \mathcal{Z}_{\backslash d,n}, \mathcal{X}, \alpha, \beta)$$

$$\propto \begin{cases} \left(N_{\backslash d,n}^{(t|d)} + \frac{\alpha}{T}\right) \frac{N_{\backslash d,n}^{(\nu|t)} + \frac{\beta}{V}}{N_{\backslash d,n}^{(t)} + \beta} \Pi_{r:x_r^{(d)}=n} \left(p_{a^{(d)}r}^{(c)}\right)^{y_r^{(d)}} \left(1 - p_{a^{(d)}r}^{(c)}\right)^{1-y_r^{(d)}} & \text{for } N^{(d)} > 0 \\ \Pi_{r:r \neq a^{(d)}} \left(p_{a^{(d)}r}^{(c)}\right)^{y_r^{(d)}} \left(1 - p_{a^{(d)}r}^{(c)}\right)^{1-y_r^{(d)}} & \text{otherwise} \end{cases} \tag{2}$$

Edge topic assignments $\mathcal{X}$:

$$P(x_r^{(d)} = n \mid \mathcal{A}, \mathcal{Y}, \mathcal{S}, C_t, \Gamma, \mathcal{I}, z_n^{(d)} = t, \mathcal{Z}_{\backslash d,n}) \propto \left(p_{a^{(d)}r}^{(c)}\right)^{y_r^{(d)}} \left(1 - p_{a^{(d)}r}^{(c)}\right)^{1-y_r^{(d)}} \tag{3}$$

Topic-cluster assignments C_t:

$$P(c_t = c \mid \mathcal{A}, \mathcal{Y}, \mathcal{S}, C_t, \Gamma, \mathcal{I}, \mathcal{X}) \propto \Pi_{r:x_r^{(d)}=n} \left(p_{a^{(d)}r}^{(c_t)}\right)^{y_r^{(d)}} \left(1 - p_{a^{(d)}r}^{(c_t)}\right)^{1-y_r^{(d)}} \tag{4}$$

```
Algorithm 1: CLUSTER-PARTITIONED MULTINETWORK EMBEDDINGS
Data: Message-Word Matrix, Dictionary, Edge Matrix, Number of Topics, Node
      Covariates, Latent Dimensions, Number of Clusters, α, β
INITIALIZE ALL VARIABLES
for i= 0;i < Number of Outer Iterations; i++ do
    for n= 0; n < Topic Step Iterations; n++ do
        SAMPLE TOKEN TOPIC ASSIGNMENTS
        SAMPLE EDGE TOPIC ASSIGNMENTS
        SAMPLE TOPIC CLUSTER ASSIGNMENTS
    end
    SLICE SAMPLE α
    ADJUST METROPOLIS HASTINGS CONTROL PARAMETERS
    for m= 0; m < Latent Space Step Iterations; m++ do
        for c= 0; c < Clusters; c++ do
            SAMPLE CLUSTER LATENT SPACE PARAMETERS
        end
    end
end
```

Assortative Mixing

Application to gender mixing in local government communication networks.

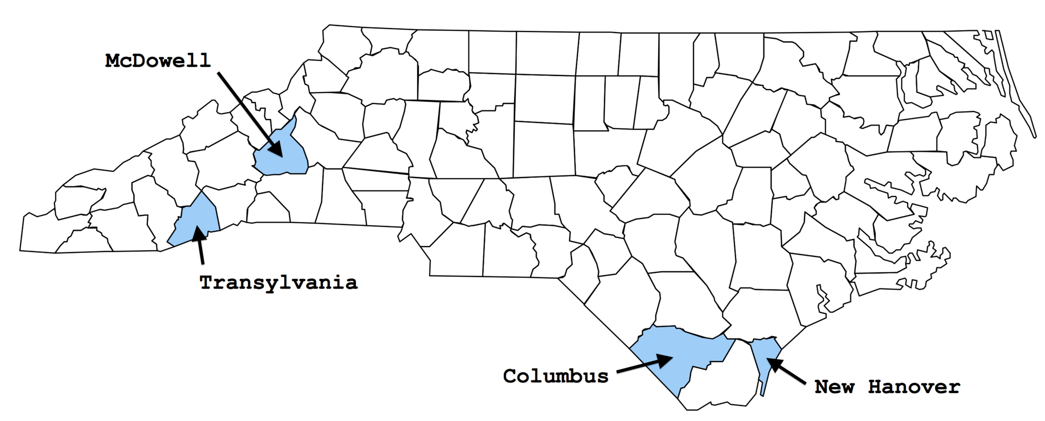
Expect to find **gender homophily**. Women tend to occupy **disadvantaged positions** in organizational networks.

- Less central in **dominant coalition**.
- More gender homophily between men – reduces access to information.

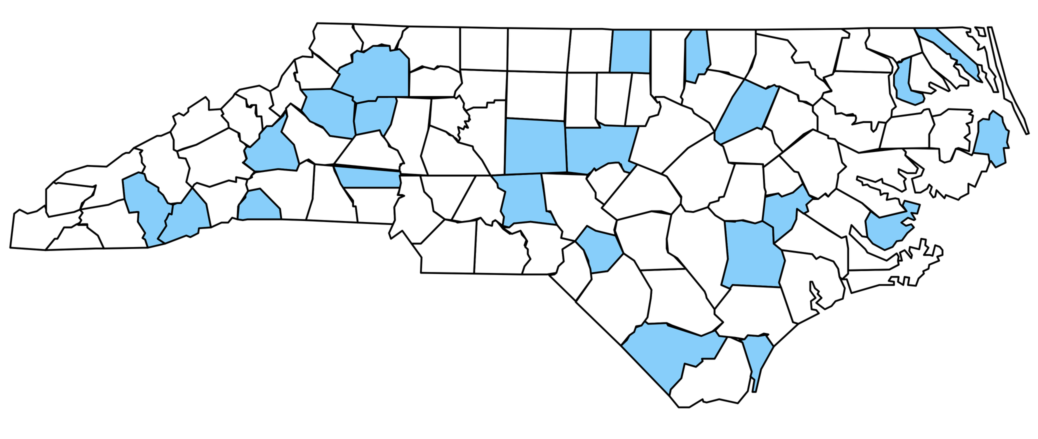
Data

Department manager email data collected from North Carolina county governments through FOIA requests.

2011 Data: 4 counties included in this study; $\sim$ 40,000 emails.



2013 Data: 23 counties; $\sim$ 500,000+ emails.



Model Specification

Using county government email data between department managers we can infer model parameters using block Metropolis-within-Gibbs:

- 100 topics and 8 topic clusters.
- 2,000 Metropolis-within-Gibbs iterations with 1,000 Metropolis Iterations per Gibbs iteration.
- Final latent position sample step – 10,000,000 MH iterations.
- α hyper-parameter slice sampled every 5 iterations.

Analysis – New Hannover County

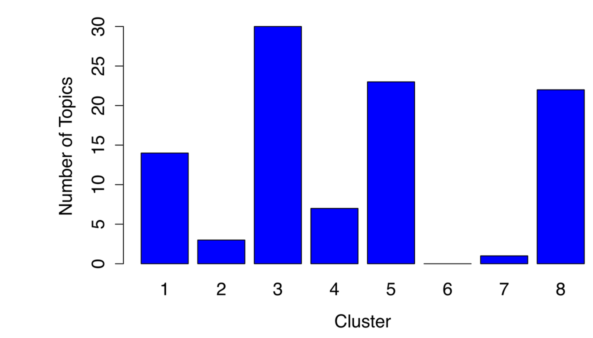
Time Frame: February 2011

Sources: All Department Heads – 27 Departments – 30 Managers

Scope: Inbox and Outbox Contents – 1,739 Internal Messages

Female Managers: 11

Figure: Histogram of topic-cluster assignments for New Hannover County. Cluster 6 has no topics assigned.



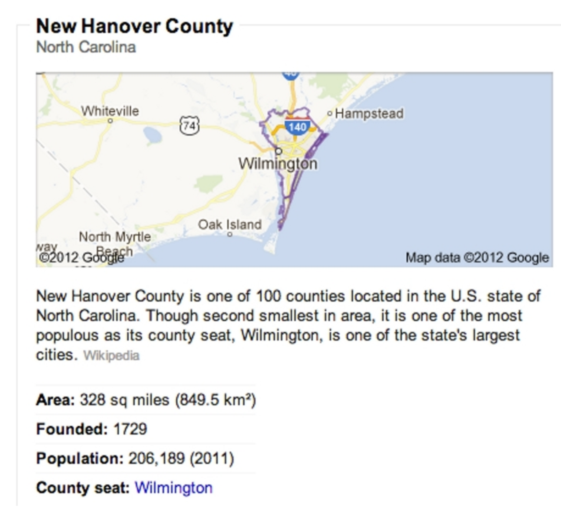


Figure: Topic top words from a sample of five topics in the two largest clusters identified for New Hannover county (numbers 3 and 5 respectively). Email networks for the finance and logistics and locus of control content partitions with nodes positioned based on inferred latent space positions and colored according to gender (female, male). Gender Asortative mixing parameter estimates with 95% confidence bars are plotted below their associated networks (note Male-Male mixing parameters were fixed at zero to allow for direct interpretation of other parameters against them).

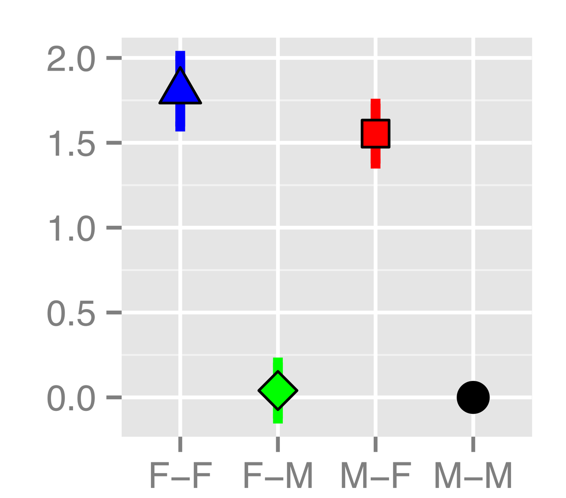
Finance and Logistics

- **support**, department, customer, employee
- attended, **training**, information, webinar
- service, **services, contract**, provide
- attached, **salary**, cam, scan
- **budget, funds**, county, information

Locus of Control

- process, **plan**, review, proposal
- community, impact, **resources**, request
- **planning, strategic**, inspections, **review**
- staff, **information**, meeting, week
- **report**, minutes, audit, feb

Mixing Parameter Estimates



Mixing Parameter Estimates

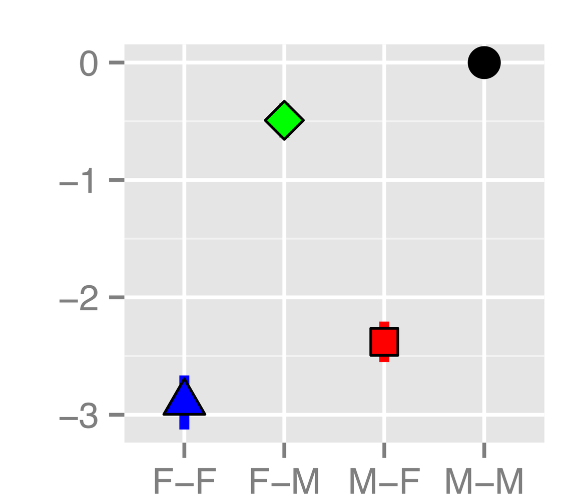


Figure: Finance and Logistics: 827 Edges, Cluster 3, fit on 1459 emails.

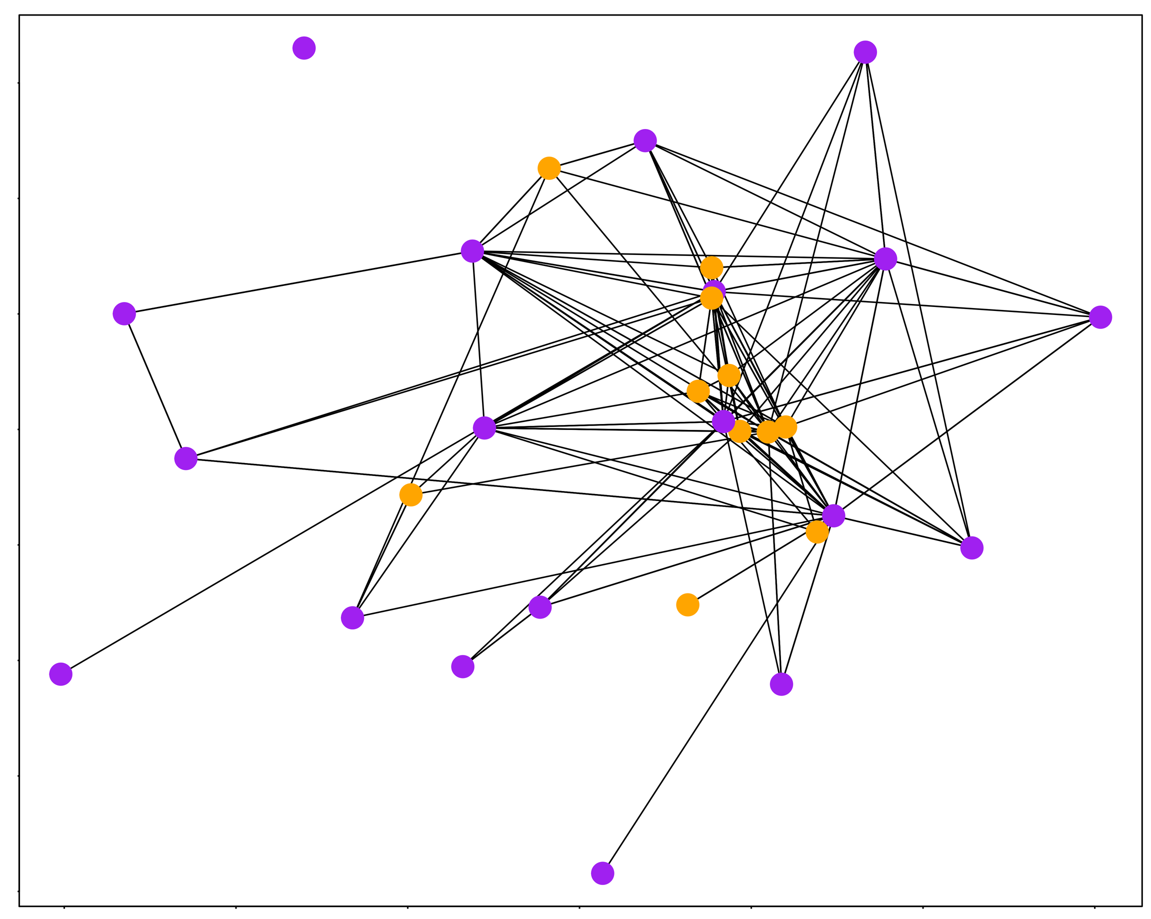
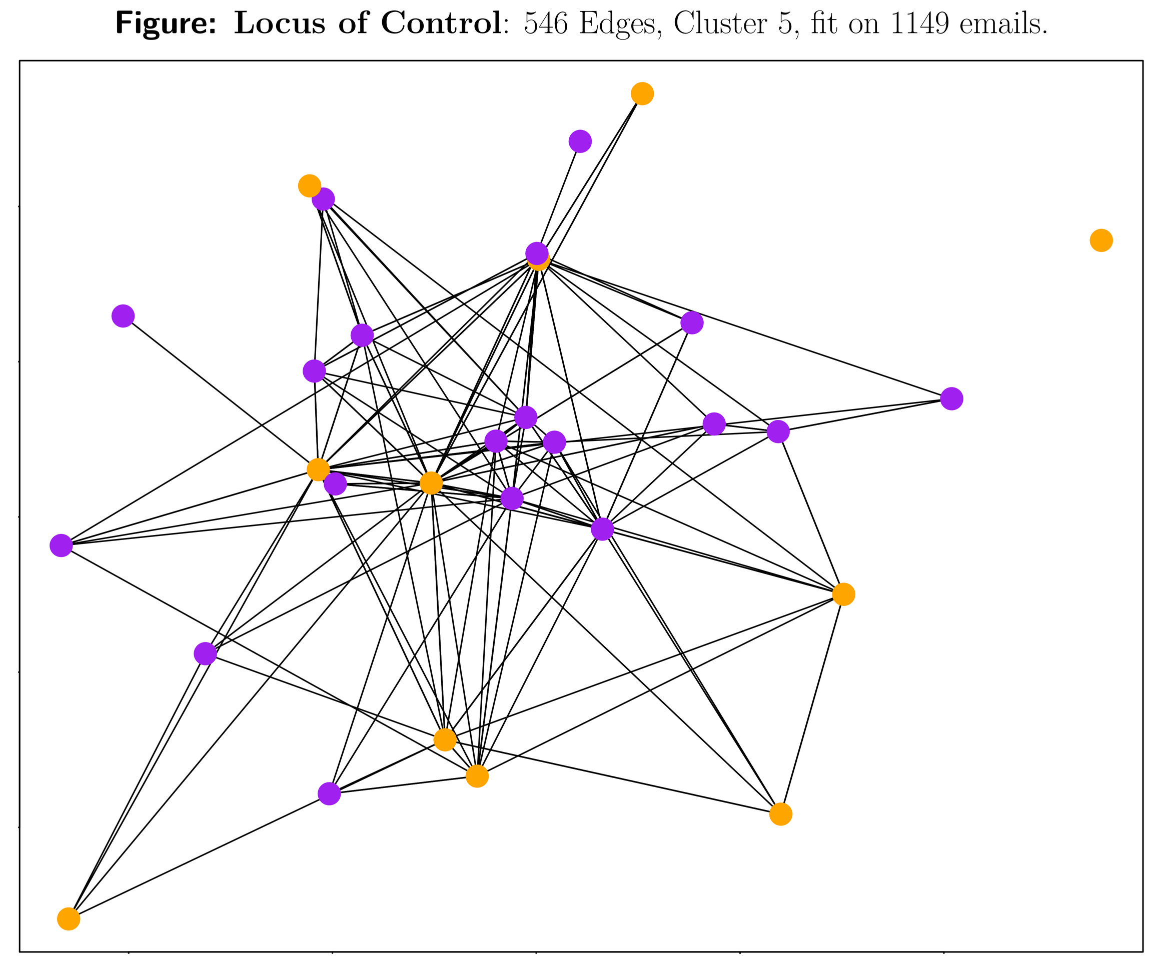


Figure: Locus of Control: 546 Edges, Cluster 5, fit on 1149 emails.



UMass Amherst

Previous Work:
Krafft, P., Moore, J., Desmarais, B. A., & Wallach, H. (2012). Topic-partitioned multinetwork embeddings. In *Advances in Neural Information Processing Systems 25*.

Work supported by NSF grant CISE-1320219 (Wallach and Desmarais, PIs)

INSTITUTE FOR SOCIAL SCIENCE RESEARCH
UNIVERSITY OF MASSACHUSETTS AMHERST