

Testing the Theory of Structural Holes: Modeling the Microbehavior of an Email Network



Christopher L. DuBois¹, Carter T. Butts², Padhraic Smyth¹. 1:Information and Computer Science, University of California - Irvine. 2:Department of Sociology, University of California - Irvine

Introduction

- Burt's hypothesis of *structural holes*: for a given budget of time, energy, and resources, actors will choose to minimize connections to those for which a relationship provides no extra benefit [1].
- As a network evolves, we should be able to detect whether actors are tending to fill structural holes - those potential relationships that would connect groups that lack abundant connections.
- This research aims to quantify how actors' tendency to fill structural holes compares with competing network formation mechanisms.

Metrics for Structural Holes

Proportion of i 's network time and energy invested in q $p_{iq} = \frac{z_{iq} + z_{qi}}{\sum_j z_{ij} + z_{ji}}, i \neq j$

Marginal strength of contact j 's relation with contact q $m_{jq} = \frac{z_{iq} + z_{qi}}{\max(z_{jk} + z_{kj})}, i \neq j$

Constraint Measures the extent to which an actor, $CS_i = p_{ij} + \sum_q p_{iq} p_{qj}$ $q \neq i, j$ i , is invested in people who are invested in i 's neighbors.

Effective Size For a given actor i , effective network size is a sum of the non-redundant portions of current contacts. $ES_i = \sum_j [1 - \sum_q p_{iq} m_{jq}]$ $q \neq i, j$

Betweenness Let g_{ijv} be the number of geodesics $B_v = \sum_{\{i,j:l=j,i,l=v,j,l=v\}} \frac{g_{ijv}}{g_{ij}}$ from i to j passing through v .

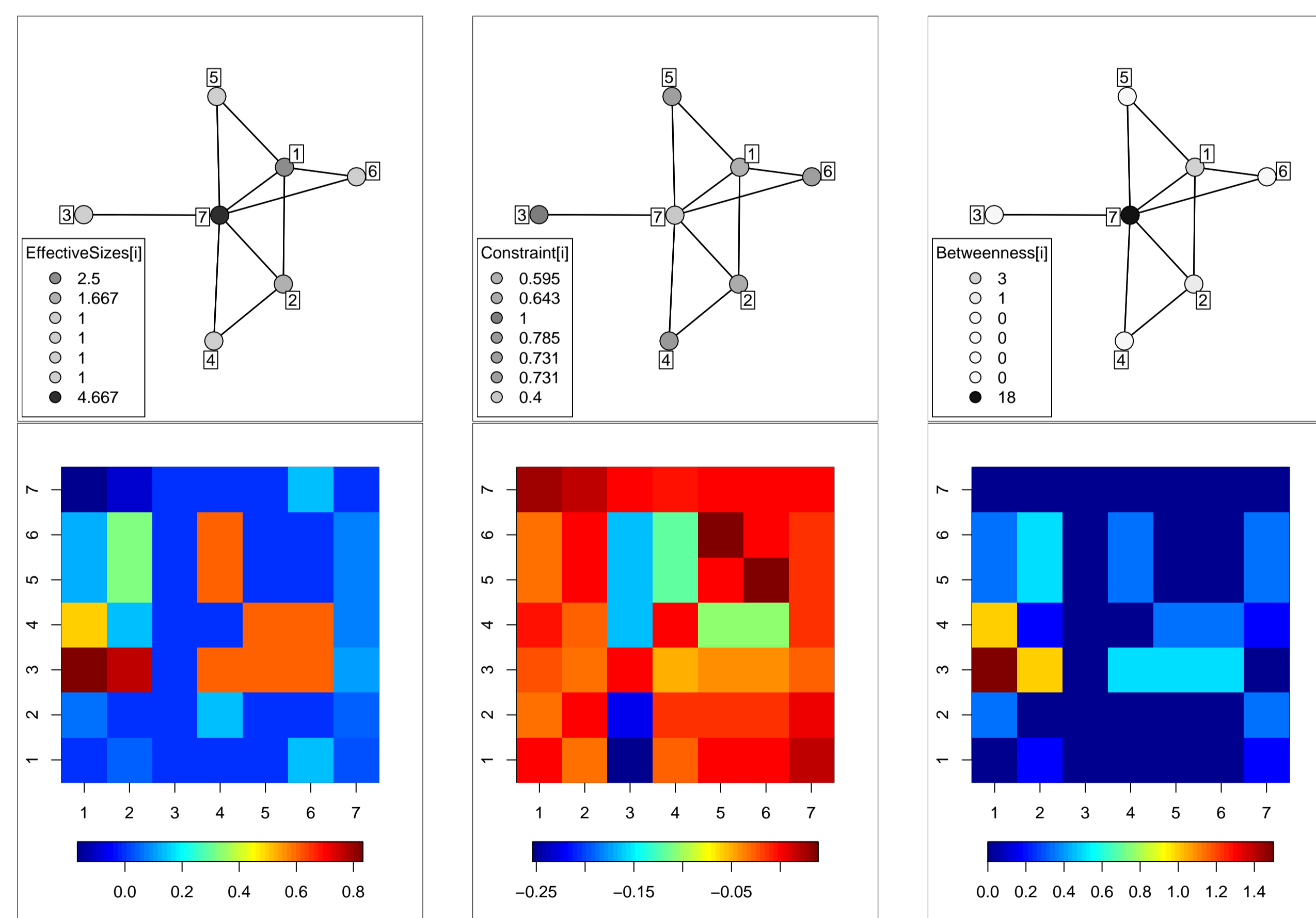


Figure 2: Toy example, illustrating effective size, constraint, and betweenness (from left to right). Top row: For each metric, the network is shown with vertices colored by the respective value. Bottom row: An image plot of the change in each metric's value for actor x (x -axis) if event (x, y) occurs. E.g. The betweenness for actor 1 will be highest if event $(1, 3)$ occurs.

The Enron Data Set

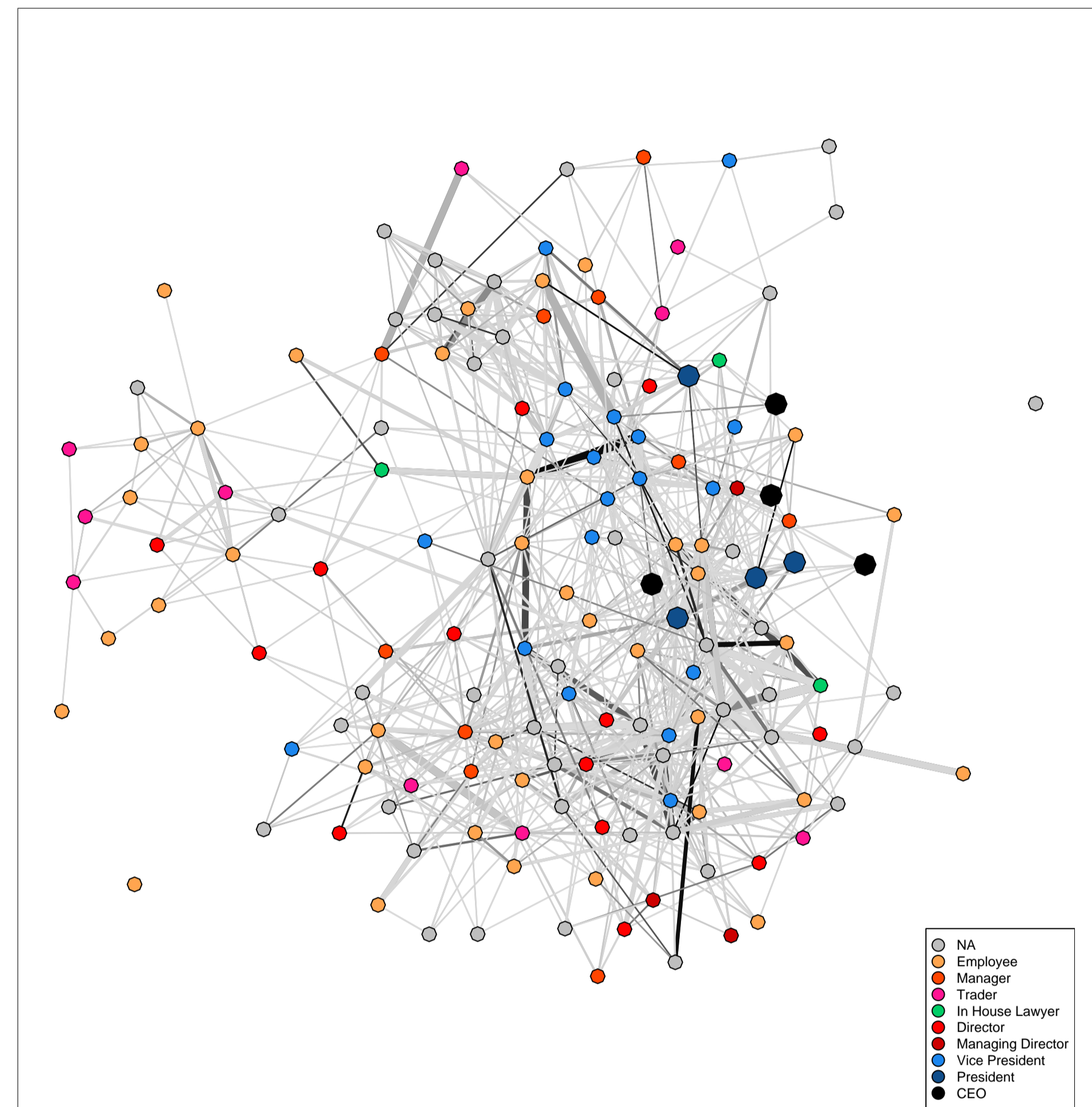


Figure 1: The accumulated network of Enron email correspondence. Edges are weighted and colored by the count of accumulated events between those actors. Note also that Vice Presidents and CEOs tend to be more centrally located.

Preliminary Results

BIC/M Statistics for the Fitted Relational Event Models

Network	Sample 1	Sample 2	Sample 3	Sample 4
N	50	50	50	50
M	1067	1056	584	614
FE	12.45	12.92	13.48	14.33
SH	10.24	11.09	10.26	N/A
R	9.91	10.42	10.02	11.21
FE+SH	9.10	9.29	8.07	10.74
SH+R	9.07	9.76	9.26	10.11
FE+R+SH	8.44	8.12	5.91	10.02

Table 1: Above models include varying combinations of fixed effects terms (FE), structural holes covariates (SH), recency terms (R), and participation shift terms (PA). See [2].

Parameter Estimates for SH+R Model for Sample 1

	$\hat{\theta}$	s.e.	z	$\Pr(> z)$
RSndSnd	2.932	0.067	44.076	<2e-16 ***
Constraint	0.186	0.245	0.759	0.448
Effective Size	-4.219	0.100	-42.327	<2e-16 ***

Table 2: Goodness of fit for one of the models considered above. Notice the negative coefficient for effective size.

Modeling Relational Events [2]

Assumption

- Each possible action is an outcome of an independent process with piecewise constant rate, where rates can change once an action occurs.
- In other words, events are modeled as homogeneous Poisson processes that are conditionally independent given the previous history.

Overview of the Model

$a_i = (s, r, t)$ = represents an action from actor s to actor r at time t
 $A_t = \{a_i\}$ = all actions before time t , letting $i \in \{1, \dots, t-1\}$
 $\lambda_a(\theta, A_t) = e^{\theta^T u}$ = the rate of action a , where u is a vector of statistics

$t \sim e^{\lambda_a(\theta, A_t)}$ = waiting time for a to occur
 $h_a(t) = \lambda_a$ = the instantaneous probability of event a occurring at time t
 $S_a(t) = e^{-t\lambda_a}$ = is the probability of action a happening sometime after time t

$$\begin{aligned} Pr(a = (s, r, t)) &= h_a \prod_{b \neq a} S_b(t) \\ &= \lambda_a(\theta, A_t) \prod_{b \neq a} e^{-t\lambda_b(\theta, A_t)} \\ &= \text{probability of event } a \text{ happening at time } t, \text{ given no previous history.} \end{aligned}$$

$$\begin{aligned} Pr(A_M) &= \prod_{i=1}^M Pr(a_i = (s, r, i)) = \text{The probability of the history of events.} \\ Pr(A_M | \theta) &= \prod_{i=1}^M \frac{\lambda_{a_i}(\theta, A_i)}{\sum_{b \neq a_i} \lambda_b(\theta, A_i)} = \text{Probability of } A_M \text{ with only ordinal information.} \end{aligned}$$

Big Picture

- This approach models the full likelihood of a sequence of relational events.
- We can parameterize the rates using a variety network statistics, including those that we hypothesize will be associated with actions occurring more often.

Future Directions

- Obtain fits using betweenness, shared partner covariates
- Scale implementation to larger networks

References

- [1] Ronald S. Burt. *Structural holes : the social structure of competition*. Harvard University Press, Cambridge, Mass., 1992.
- [2] Carter T. Butts. A relational event framework for social action. *Sociological Methodology*, 38(1):155–200, December 2008.