Large-Scale Social Network Analysis of Facebook Data

Emma S. Spiro[1] Zack W. Almquist[1] Carter T. Butts[1,2]

[1]Department of Sociology
[2]Institute for Mathematical Behavioral Sciences
University of California – Irvine

Presented at MURI All Hands Meeting January 10, 2012

This material is based on research supported by the Office of Naval Research under award N00014-08-1-1015.

As well as the National Science Foundation under awards BCS-0827027 and OIA-1028394.

Scalable Methods for the Analysis of Network-Based Data

MURI Themes and Goals

- Large-scale social networks
- Spatially embedded networks
- Rich models with complex covariates
- Scalable methods and models

Spatially Embedded Networks

- Social interaction occurs within a spatial context
  - Opportunities for, costs of interaction strongly influenced by spatial factors
  - Interest in spatial factors per se (e.g., neighborhood research)
  - Propinquity known to be a powerful determinant of tie probability
- Extension to attribute spaces (Blau space)
  - Useful way to parameterize homophily, clustering effects
- Simple idea: assign vertices to spatial locations
- Location function: $\ell : V \Rightarrow S$ where S is an abstract space.
- Take ℓ as given fixed, e.g. latitude/longitude coordinates

Spatial Bernoulli Graphs, (Butts 2002)

- A simple family of models for spatially embedded social networks

$$\Pr(\mathbf{Y} = \mathbf{y} | \mathbf{D}) = \prod_{\{i,j\}} B\left(Y_{ij} = y_{ij} | \mathcal{F}_d\left(D_{ij}\right)\right) \tag{1}$$

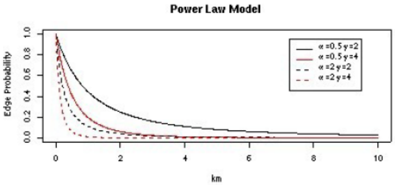
  - $\mathbf{Y} \in \{0,1\}^{N \times N}$
  - $\mathbf{D} \in [0, \infty)^{N \times N}$
  - $\mathcal{F}_d : [0, \infty) \mapsto [0, 1]$
- Assumes that dependence among edges is absorbed by the distance structure – edges conditionally independent.
- Related to gravity model from geography.
- Advantage: Estimable under sampling and scalable
- How does distance effect tie probability?

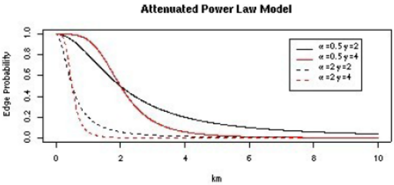
Spatial Interaction Function

- Decay as a power law in distance

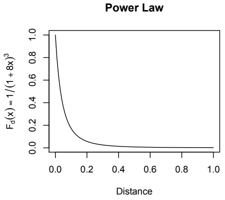
$$\mathcal{F}_d(x) = \frac{p_b}{(1+\alpha x)^\gamma}$$

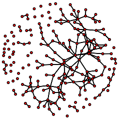
  where $0 \leq p_b \leq 1$ is a baseline tie probability, $\alpha \geq 0$ is a scaling parameter, and $\gamma > 0$ is the exponent which controls the distance effect
- Attenuated power law, arctangent decay, etc.

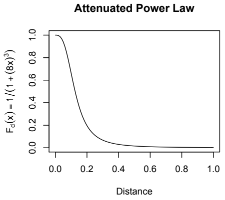


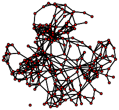


Spatial Interaction Function









- Small changes in the SIF can make big differences in the underlying network
- Changes in the functional form of the SIF can also make a big difference
- Notice that the difference between the APL and the PL is not visually striking but the resulting networks are quite different

Theories of the Distance Effect

- How does distance effect tie probability?
- Is the way in which distance matters homogeneous?
  - Vary along lines of status or prestige
  - Want to allow for inhomogeneity in the relationship between distance and tie probability
  - How to extend the spatial Bernoulli models

Spatial Bernoulli Models with Covariates

- We can extend the model in a simple way to include tie covariates
- Add GLM structure to the parameters of the SIF, $\mathcal{F}_d$

$$\Pr(Y_{ij} = 1) = \frac{p_{b_{ij}}}{(1 + \alpha_{ij} d_{ij})^{\gamma_{ij}}}$$

where

$$p_{b_{ij}} = ilogit(\theta * X_{ij})$$

$$\alpha_{ij} = exp(\psi * W_{ij})$$

$$\gamma_{ij} = exp(\phi * U_{ij})$$

and where θ, ψ, and ϕ are parameter vectors, and **X**, **W**, and **U** are covariate matrices.

Application: Selective Mixing on Facebook

- Facebook is an extremely large online social network
- Data: sample of almost 1 million egocentric networks (Gjoka et al. 2009)
- Each Facebook user may indicate a university affiliation, $< 4\%$ actually do
- Rich set of covariates at the institution level
- Online context is a best case scenario for equal mixing and "weak" distance effects

Selecting Covariates of Interest

- Institutional prestige: USNWR National University Ranking
  - Top 194 schools receive a rank, score, and selectivity measure
  - Prestige as the first principal component scores of these measures
- Public/Private
- Endowment, Tuition, Location etc.

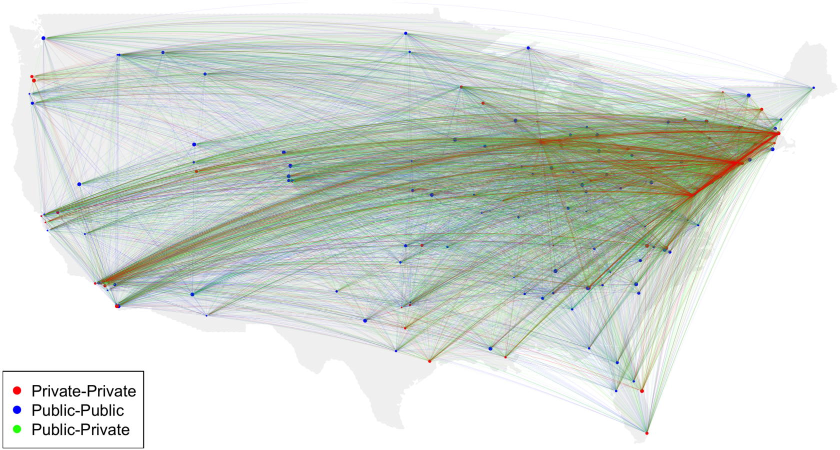
Quick Comment on Model Fitting and Computation

- Fitting these models is not an easy task
- Bayesian point estimation
- Importance sampling to fit the exponential family model
- Numerical tricks

Model Fitting and Selection

Model	p_b Effects			α Effects			γ Effects			SIF Form	BIC
Covariate	Intercept	Pub/Priv	Prestige	Intercept	Pub/Priv	Prestige	Intercept	Pub/Priv	Prestige		
Model 1	√	√	√	√	√	√	√	√		pl	24911904
Model 2	√	√	√	√	√		√	√	√	pl	24918710
Model 3	√	√	√	√	√		√	√		apl	24926060
Model 4	√	√	√	√		√	√	√	√	apl	24933741
Model 5	√		√	√	√	√	√	√		apl	24935807
Model 6	√			√			√			apl	25139114

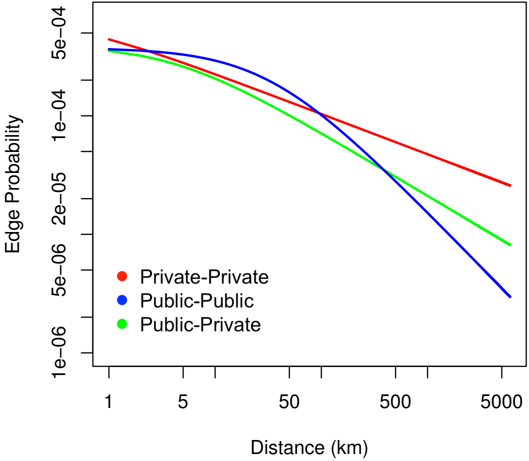
Facebook Friendship Network



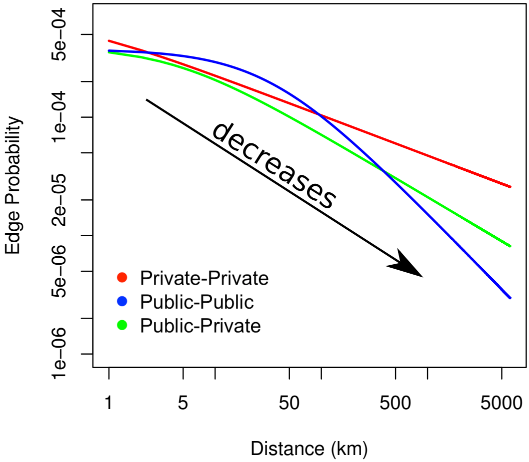
A Model of Facebook Friendship

Parameter	Component	Estimate	p.s.d.e.	
p_b	Intercept	-6.0974	0.0061	**
	Private-Public	-0.4340	0.0200	**
	Public-Public	-0.7501	0.0063	**
	Prestige	-0.0176	0.0000	**
α	Intercept	2.1687	0.0259	**
	Private-Public	-2.2169	0.0493	**
	Public-Public	-4.5387	0.0269	**
	Prestige	-0.0187	0.0001	**
γ	Intercept	-1.0789	0.0016	**
	Private-Public	0.4523	0.0026	**
	Public-Public	1.0009	0.0023	**

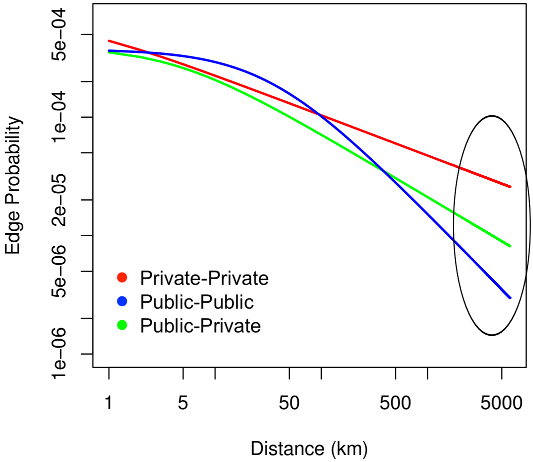
A Model of Facebook Friendship



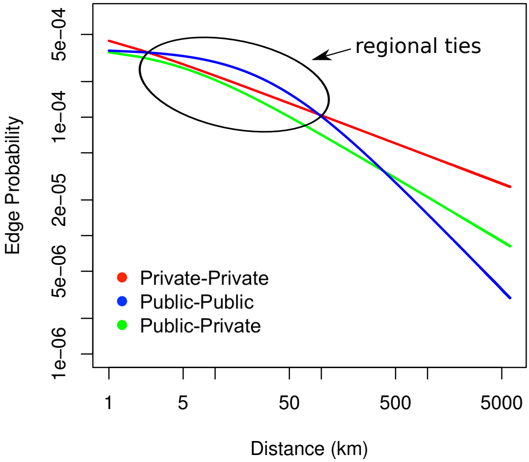
A Model of Facebook Friendship



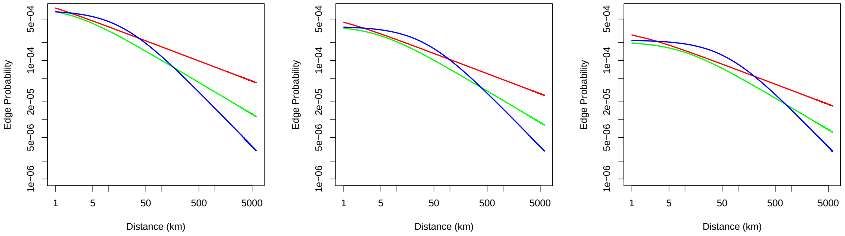
A Model of Facebook Friendship



A Model of Facebook Friendship



Effects of Difference in Prestige



Summary

- Spatial mixing models to sampled data from Facebook
- Model extension to include covariates
- Non-trivial model fitting procedure
- Inhomogeneous relationship between distance and tie probability
- Scalable models for large-scale social networks