Simulation of Spatially-Embedded Networks

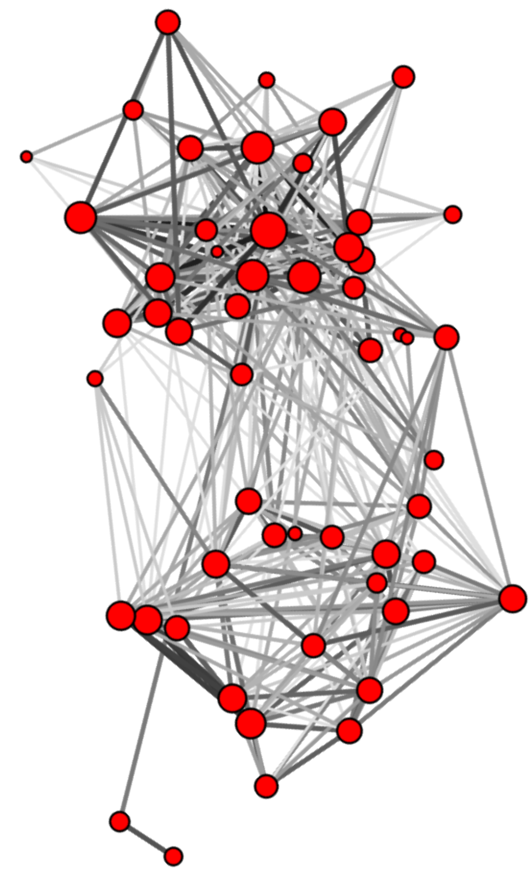
Carter T. Butts[1,2]
Ryan Acton[1], Zack Almquist[1]

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

Prepared for the August 25, 2009 UCI MURI AHM. This research was supported by ONR award N00014-08-1-1015.

Spatially Embedded Networks

- **Simple idea: assign vertices to spatial locations**
- **Spatial embedding of $G=(V,E)$**
  - Location function $\ell: V \rightarrow S$, where S is an abstract space
  - Properties of S
    - Admits some distance, d
    - May or may not be continuous
    - May or not be metric
    - May contain social dimensions (“Blau” space) as well as physical ones
  - For present purposes, take ℓ as given, fixed
    - Useful, but can be relaxed



(Data from Freeman et al., 1988)

An Inhomogeneous Bernoulli Family for Spatially Embedded Networks

- **A simple family of models for spatially embedded social networks:**

$$\Pr\left(Y=y\middle|d\right)=\prod_{\{i,j\}} B\left(Y_{ij}=y_{ij}\middle|\mathscr{F}\left(d_{ij}\right)\right)$$

  - where $Y\in\{0,1\}^{NxN}$, $d\in[0,\infty)^{NxN}$, $\mathscr{F}$: $[0,\infty)\rightarrow[0,1]$, B Bernoulli pmf
- **Special case of the inhomogeneous Bernoulli graph family with parameter matrix** $\Phi_{ij}=\mathscr{F}(d_{ij})$
  - Assumes that dependence among edges absorbed by distance structure – edges conditionally independent
- **Related to the *gravity models*, i.e.**

$$\mathbf{E}Y_{ij}=P(i)P(j)F(d_{ij})$$

  - where P is an interaction potential, and F is an impedance or spatial interaction function

Generalization to Curved Exponential Random Graph Models

- **Increasingly widely used approach – ERG form**
- **Our likelihood can be rewritten as a curved ERG**

$$\Pr\left(Y=y\middle|\theta,d\right)\propto\exp\left(\sum_{\{i,j\}}\eta\left(\theta,d\right)y_{ij}\right),\quad \eta\left(\theta,d\right)=\operatorname{logit}\mathscr{F}\left(\theta,d\right)$$

  - Sufficient statistics are the edge indicators of Y; canonical parameters (η) are logits of marginal edge probabilities
    - $O(N^2)$ canonical parameters – computational savvy advised

- **General curved model: space + other effects**

$$\Pr\left(Y=y\middle|\theta_{\mathscr{F}},\theta,d\right)\propto\exp\left(\eta\left(\theta\right)^T t\left(Y\right)+\sum_{\{i,j\}}\eta_{\mathscr{F}}\left(\theta_{\mathscr{F}},d\right)y_{ij}\right)$$

  - Allows for integration of complex edge dependence, degree distribution constraints, other covariate effects, etc. (through t)
  - Can use to control for social mechanisms when seeking spatial effects, or spatial effects when seeking social mechanisms

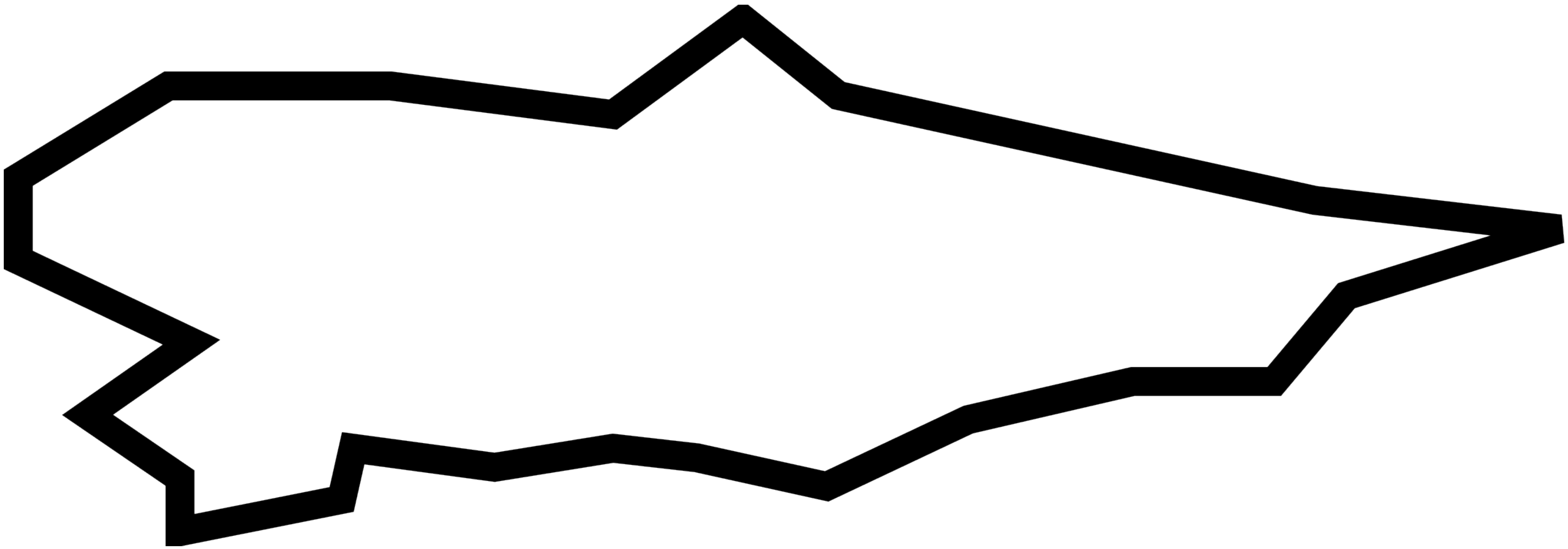
Using Spatial Models for Detailed Network Simulation

- **Start with GIS data on populations in space**
  - Here, block-level information on metropolitan/micropolitan areas, as defined by US census
- **Draw individual positions from a point process, given GIS constraints**
  - Attempt to approximate distribution of individual residences within blocks
- **Draw network from spatial Bernoulli graph model given individual positions**
  - Requires a fitted SIF (obtained from prior data, or first principles)
  - In practice, need to do clever things to make this scale
    - Subdivide space into regions, avoid simulating all pairs for regions with very low probability of positive tie volume

Placing People Within Blocks

- **Challenge: placing individuals within census blocks**
  - Want simple, reasonably fast model which captures basic properties of residential settlement in a plausible way
    - Constraints: fixed total population, household size distribution
  - Needs to work with arbitrary regions, without requiring additional data
- **Two simple approaches used here for planar coordinates**
  - Uniform placement: household coordinates drawn as Poisson process with constant intensity within each block; individual coordinates w/in household drawn from circular distribution centered on household location with maximum radius 5m
  - Quasi-random placement: as per uniform placement, but household coordinates drawn from a two-dimensional Halton sequence (bases 2 and 3)

Vertex Placement, Illustrated

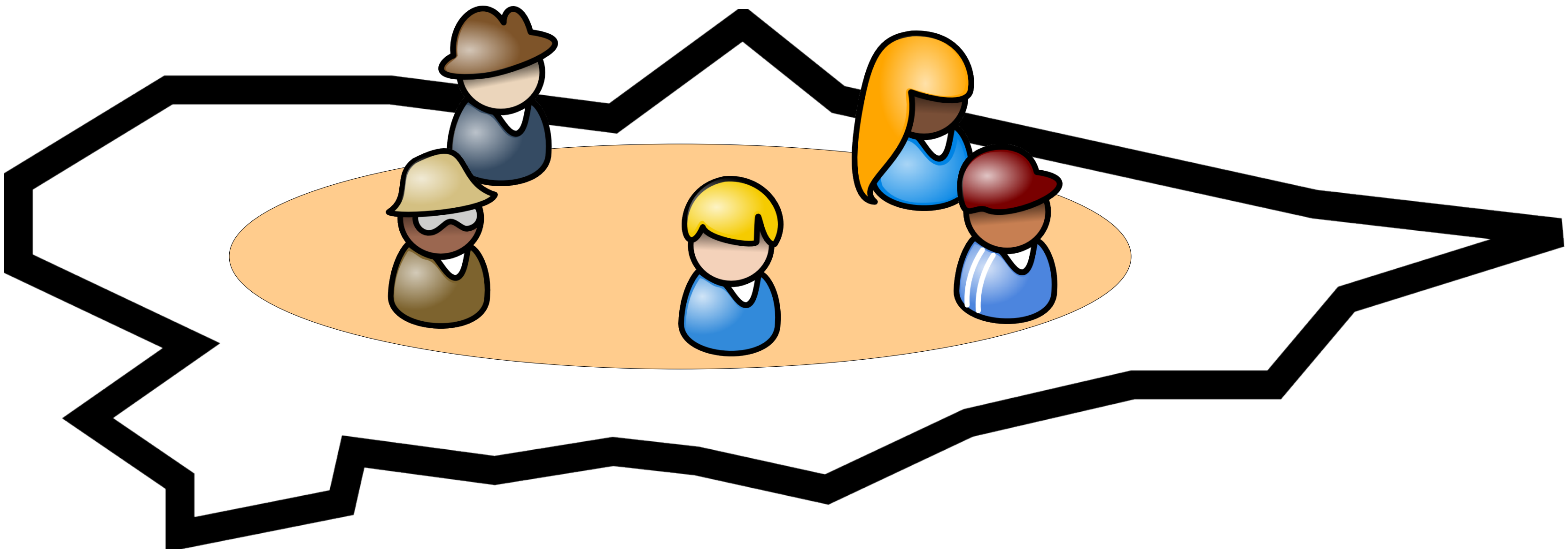


Vertex Placement, Illustrated

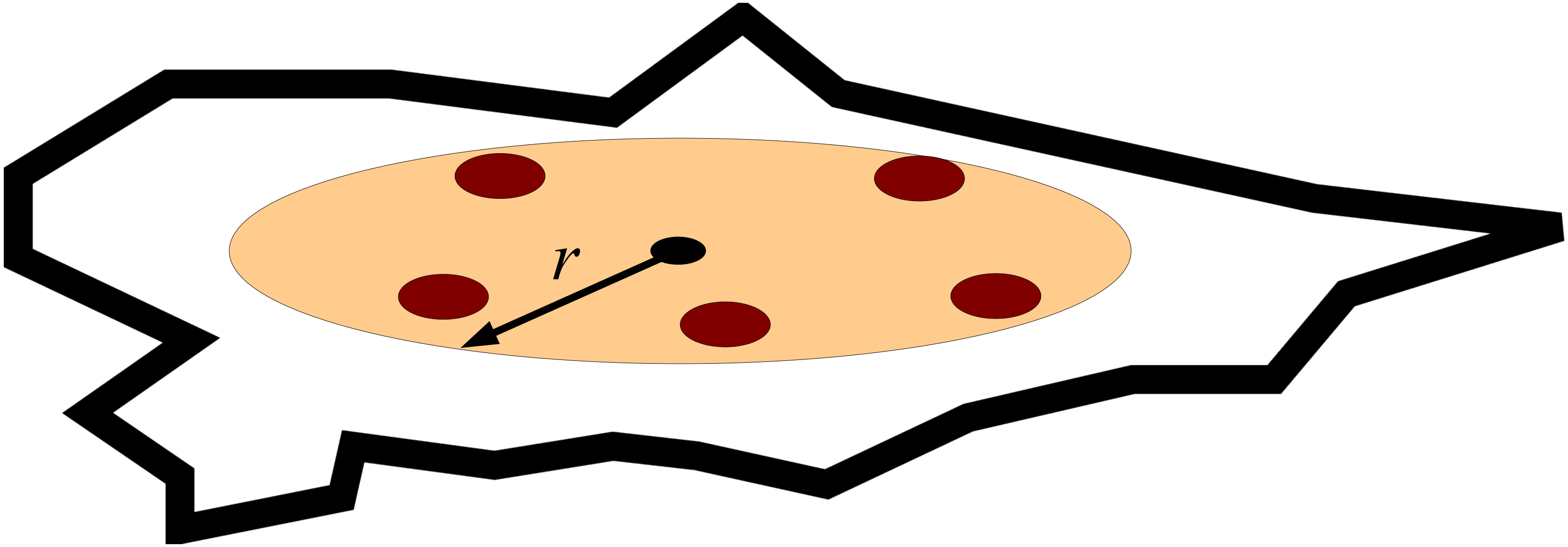
Vertex Placement, Illustrated

Vertex Placement, Illustrated

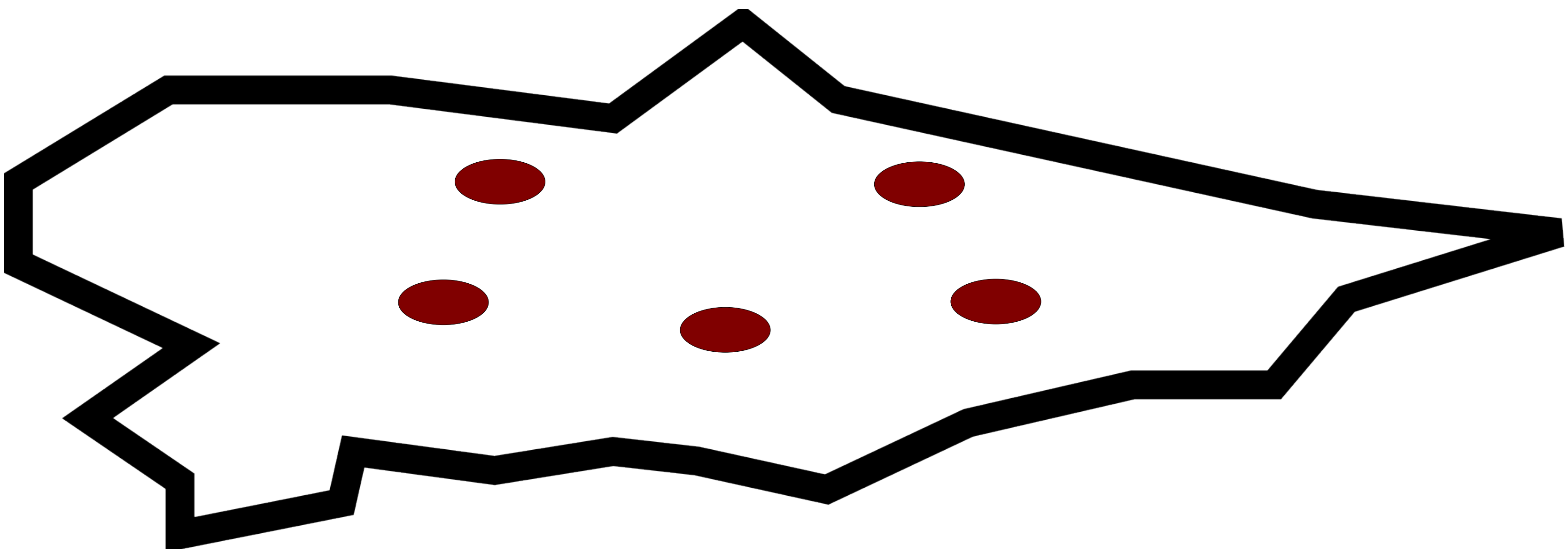
Vertex Placement, Illustrated



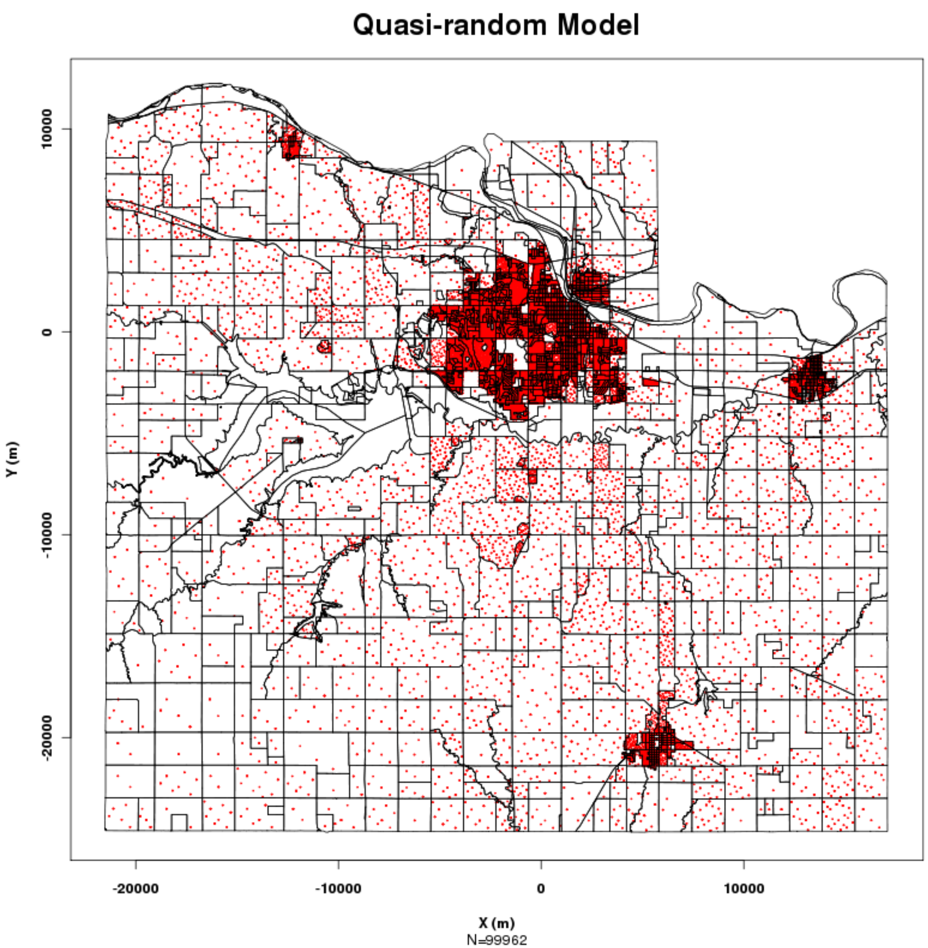
Vertex Placement, Illustrated

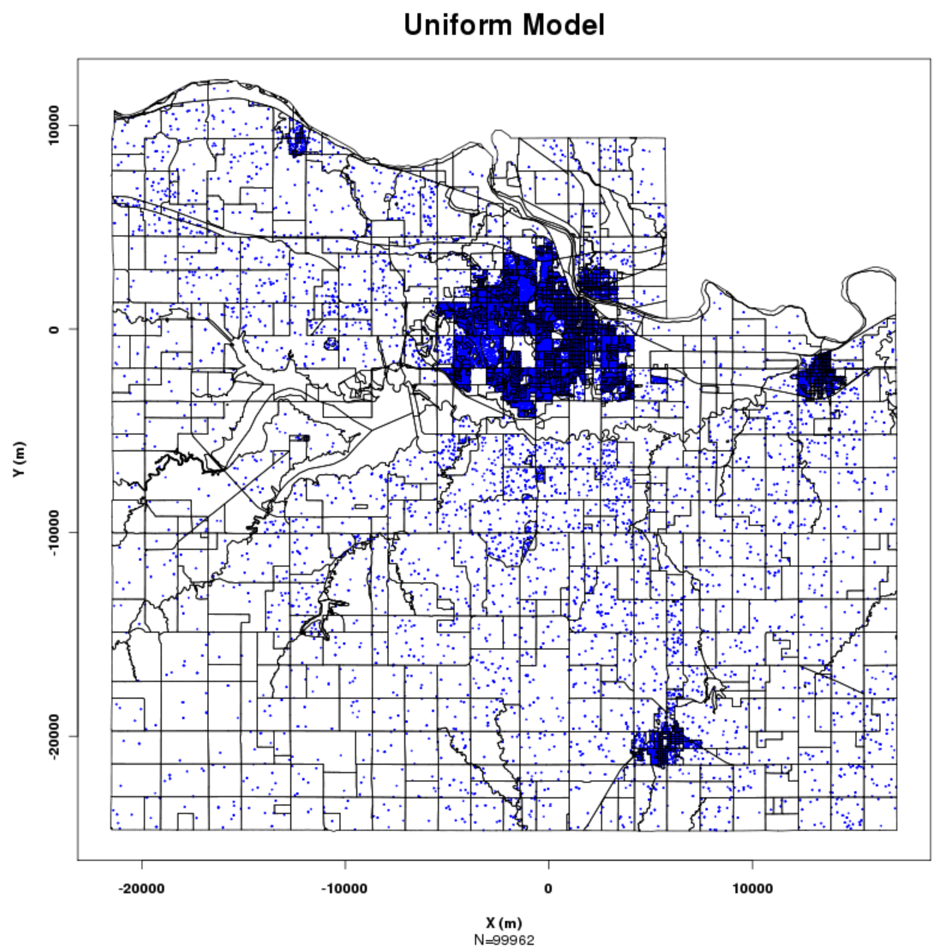


Vertex Placement, Illustrated

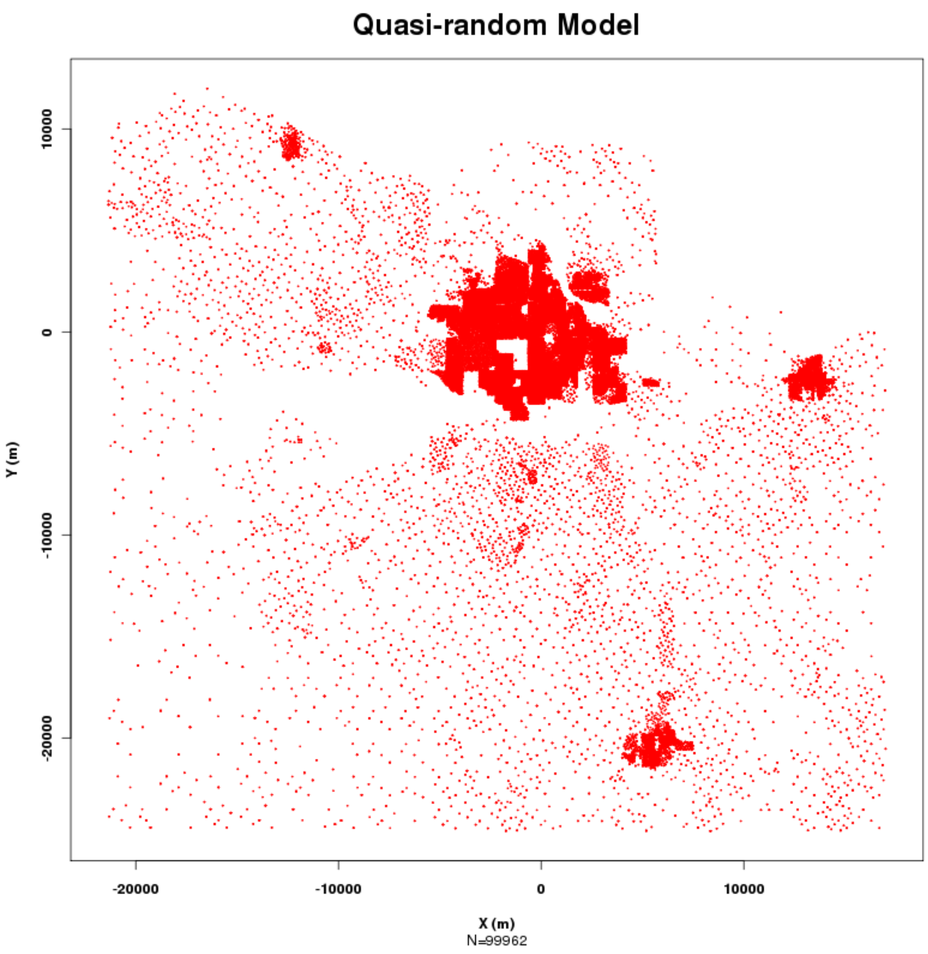


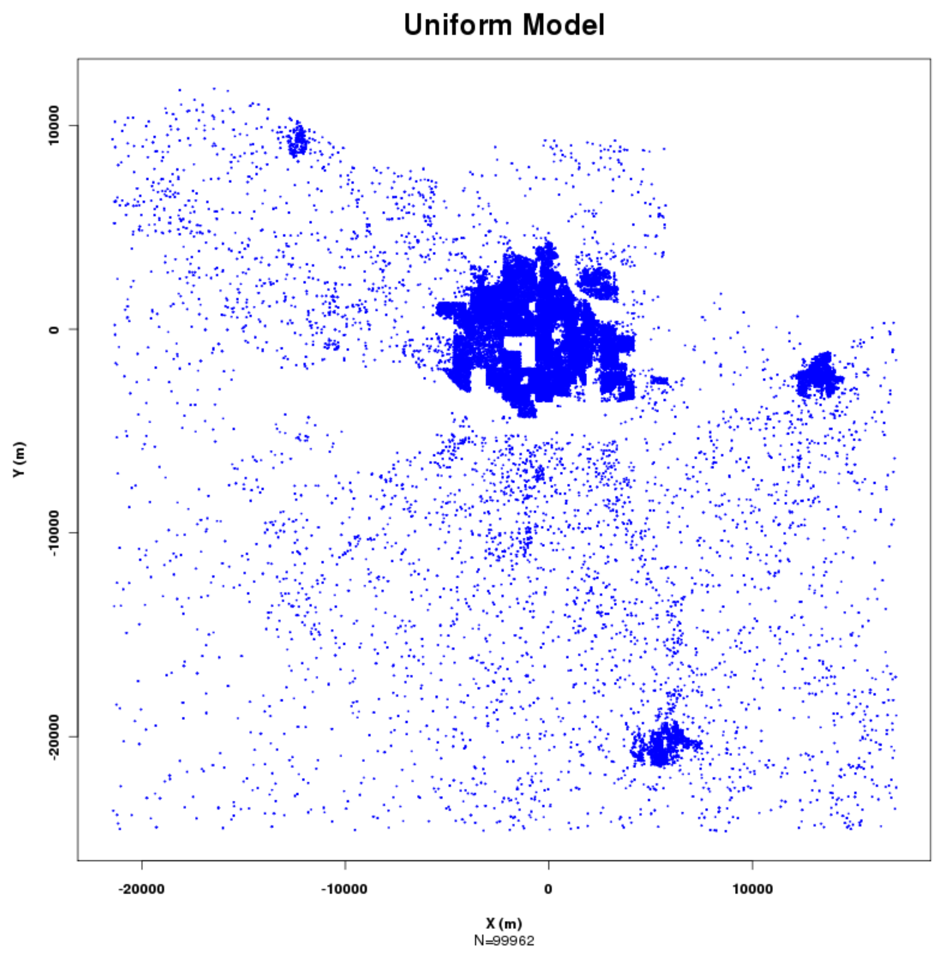
Example: Lawrence, KS





Example: Lawrence, KS

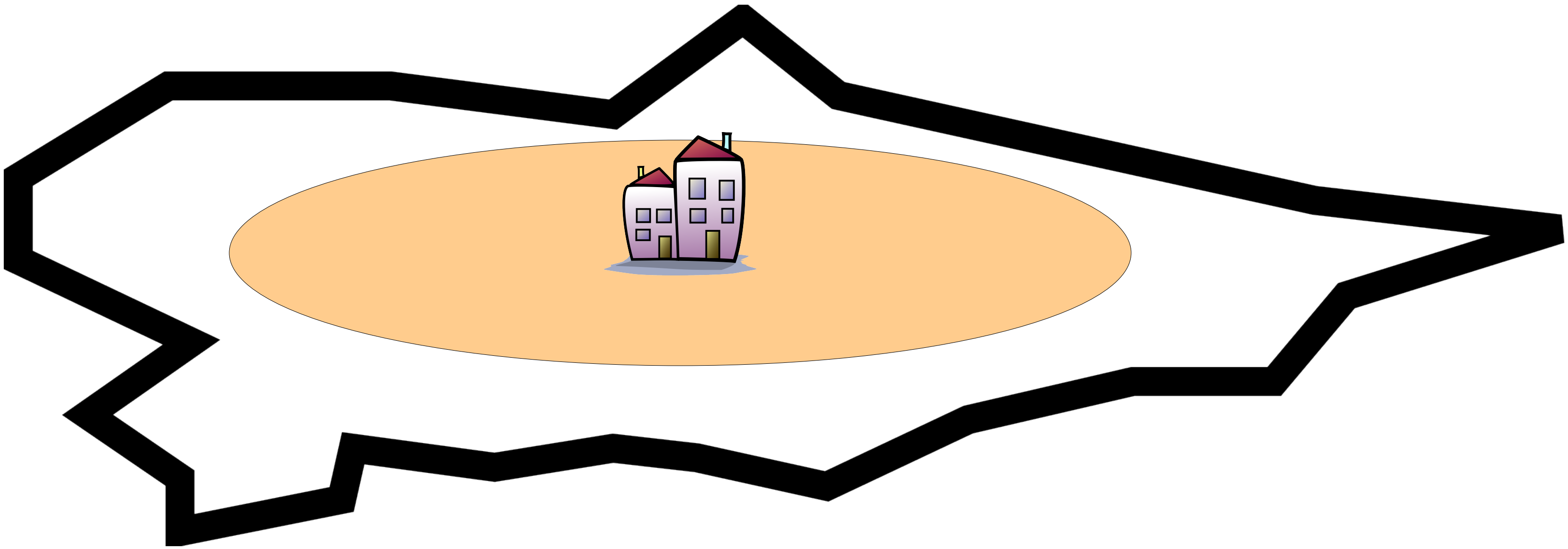




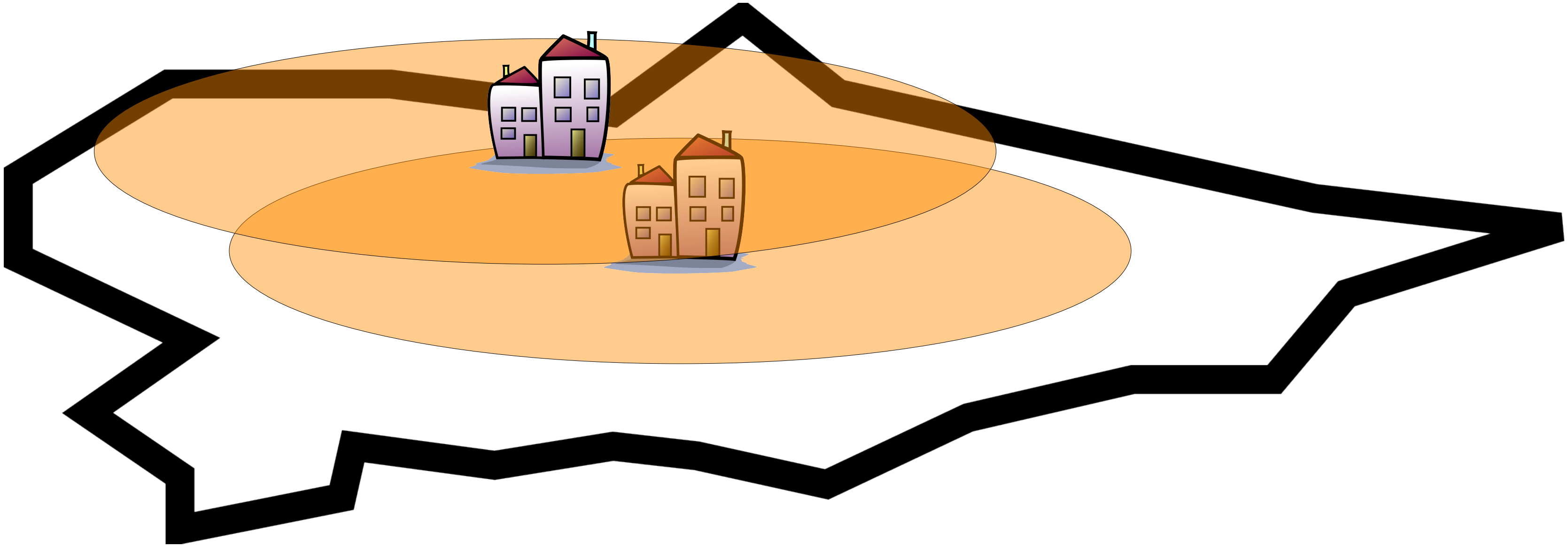
Artificial Elevation

- **Issue: actual high-density blocks feature "artificial elevation" as an essential feature of the built environment**
  - Effect: inhibits local interaction, relative to what would be observed if everyone resided in x,y plane
- **Crude solution: artificial elevation model**
  - Assign households to planar coordinates in random order
  - When placing ith household, note number of other households within planar radius r (call this k)
  - Set elevation of ith household to αk meters
- **Result: single-story housing predominates, but multi-story configurations emerge in dense areas**
  - For our purposes, arbitrarily set r=10m, α=4m

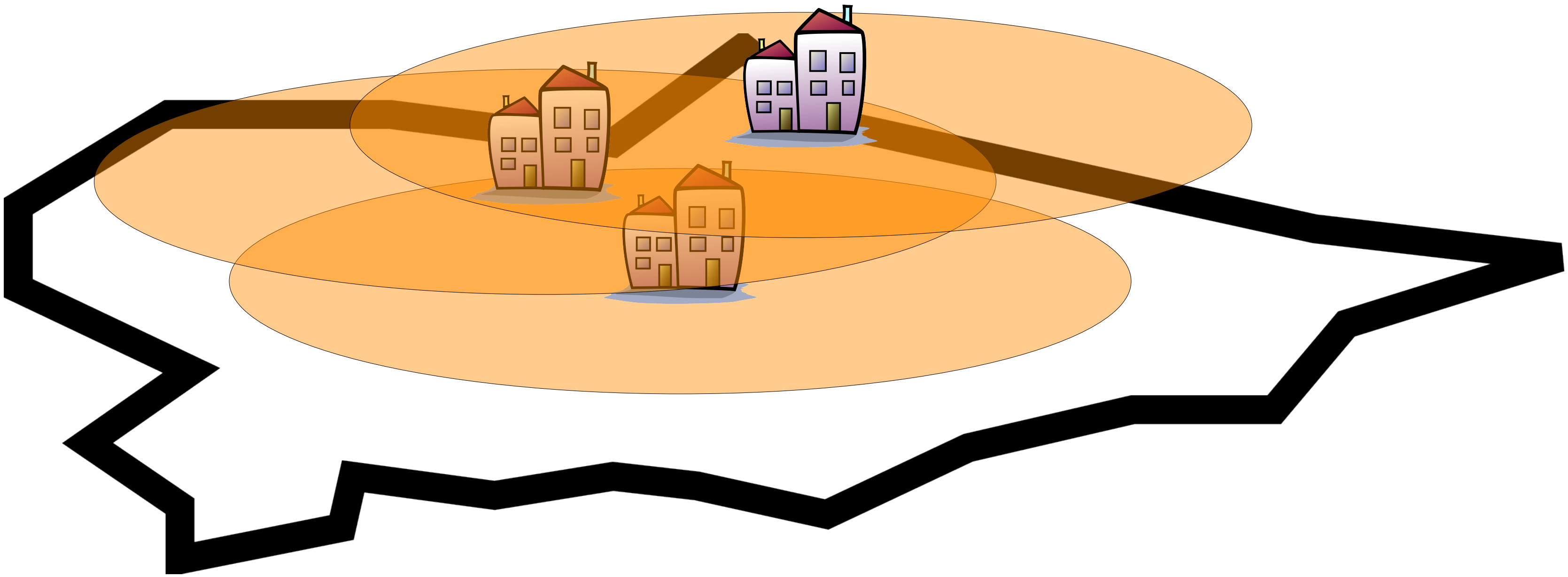
Artificial Elevation, Illustrated



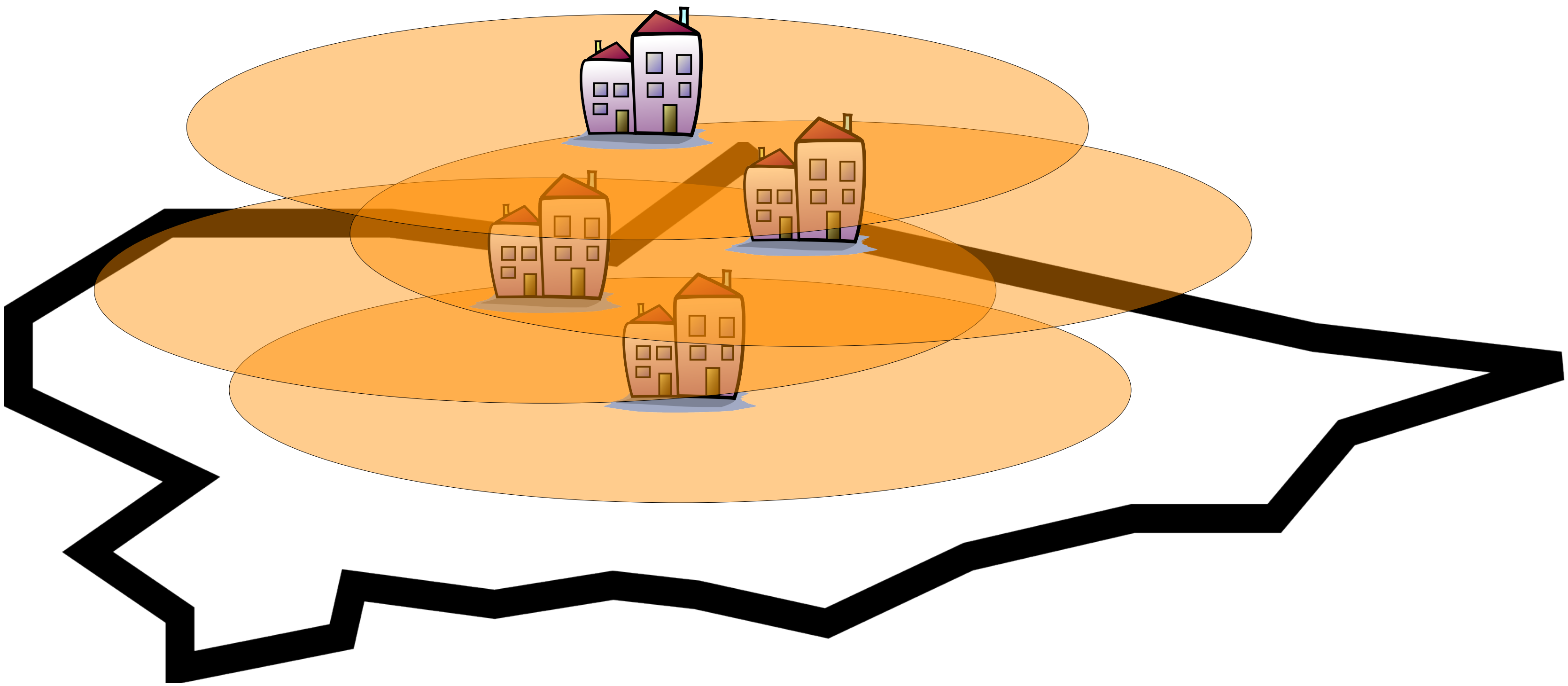
Artificial Elevation, Illustrated



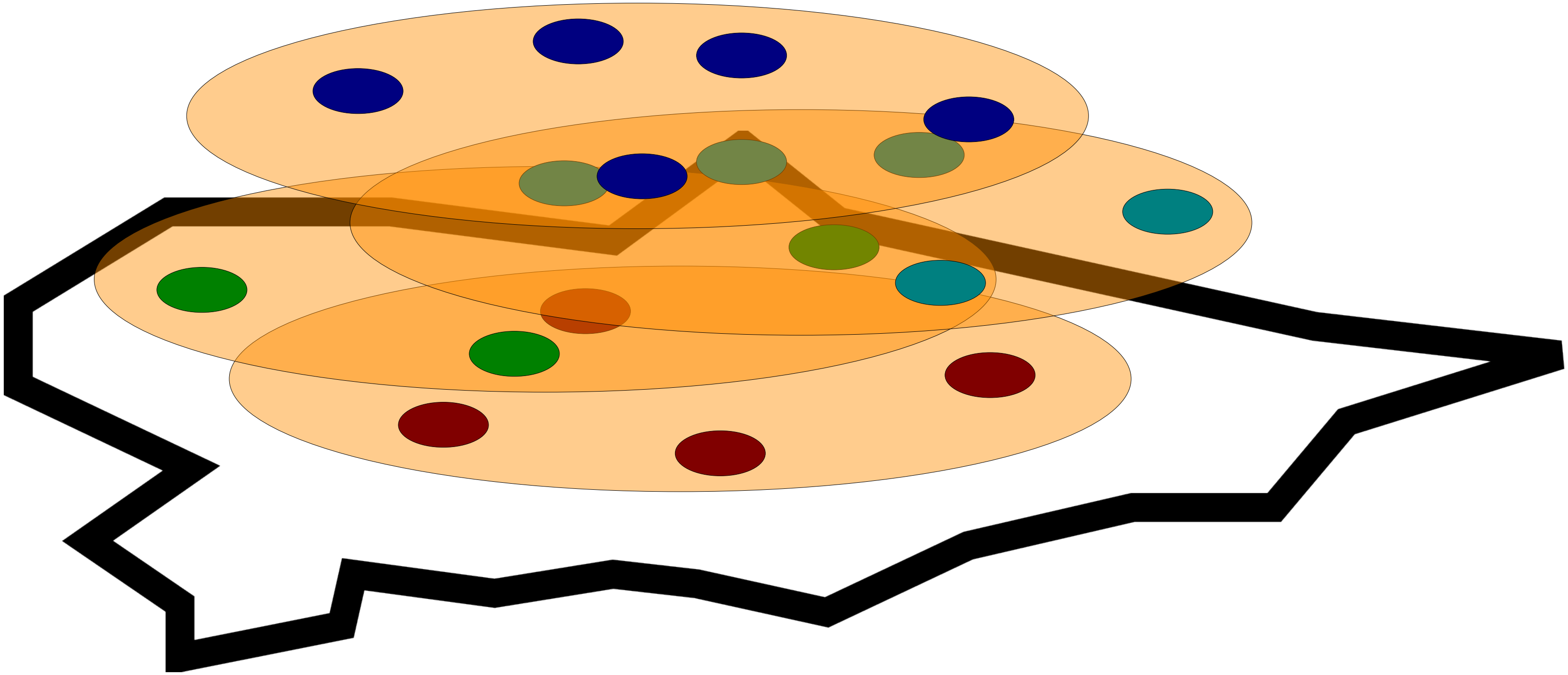
Artificial Elevation, Illustrated



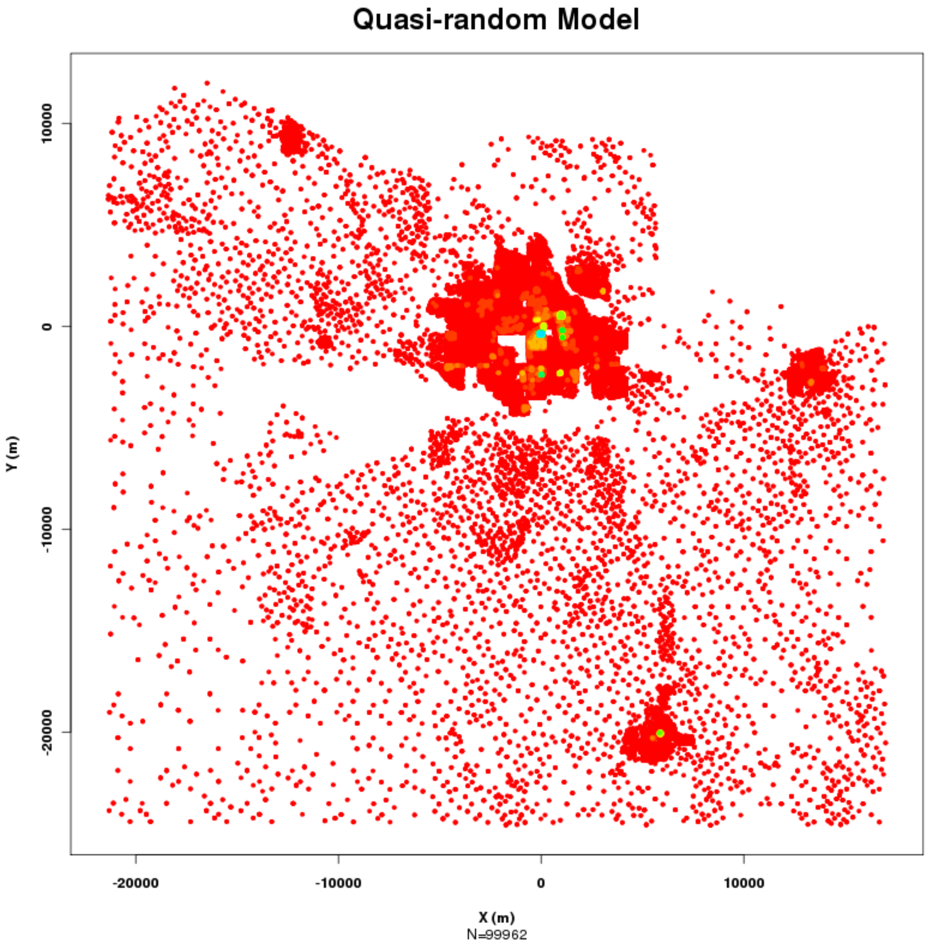
Artificial Elevation, Illustrated

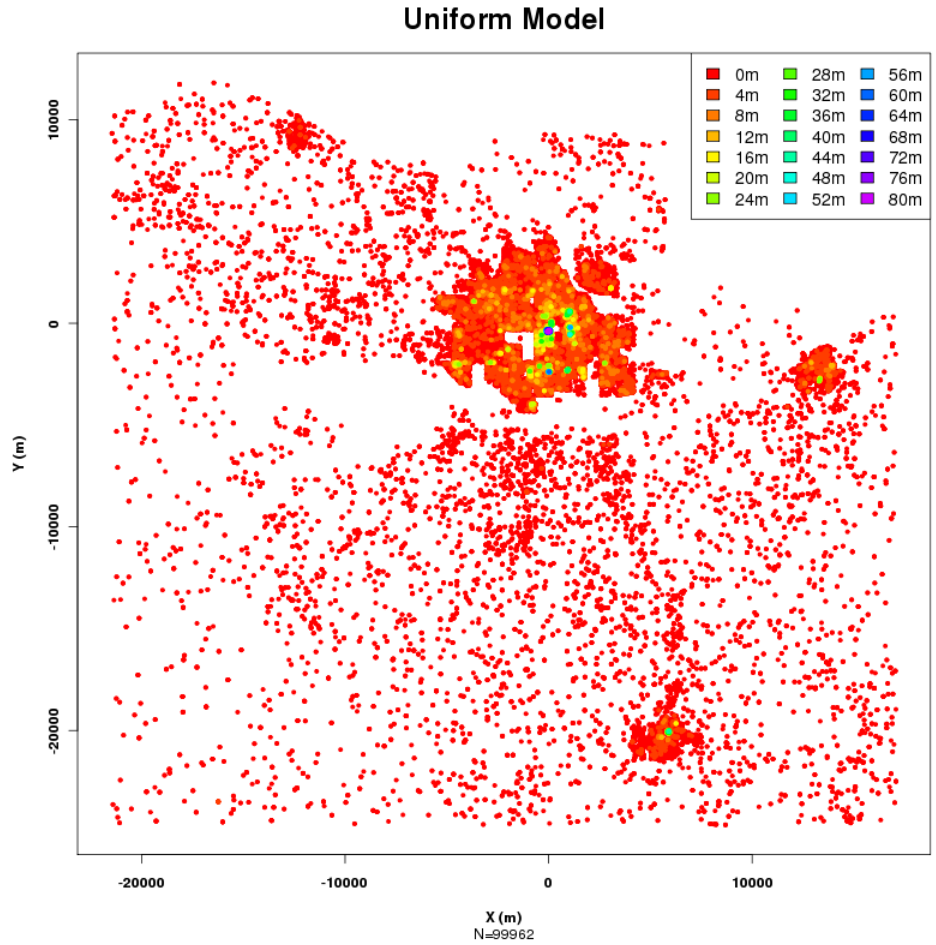


Artificial Elevation, Illustrated



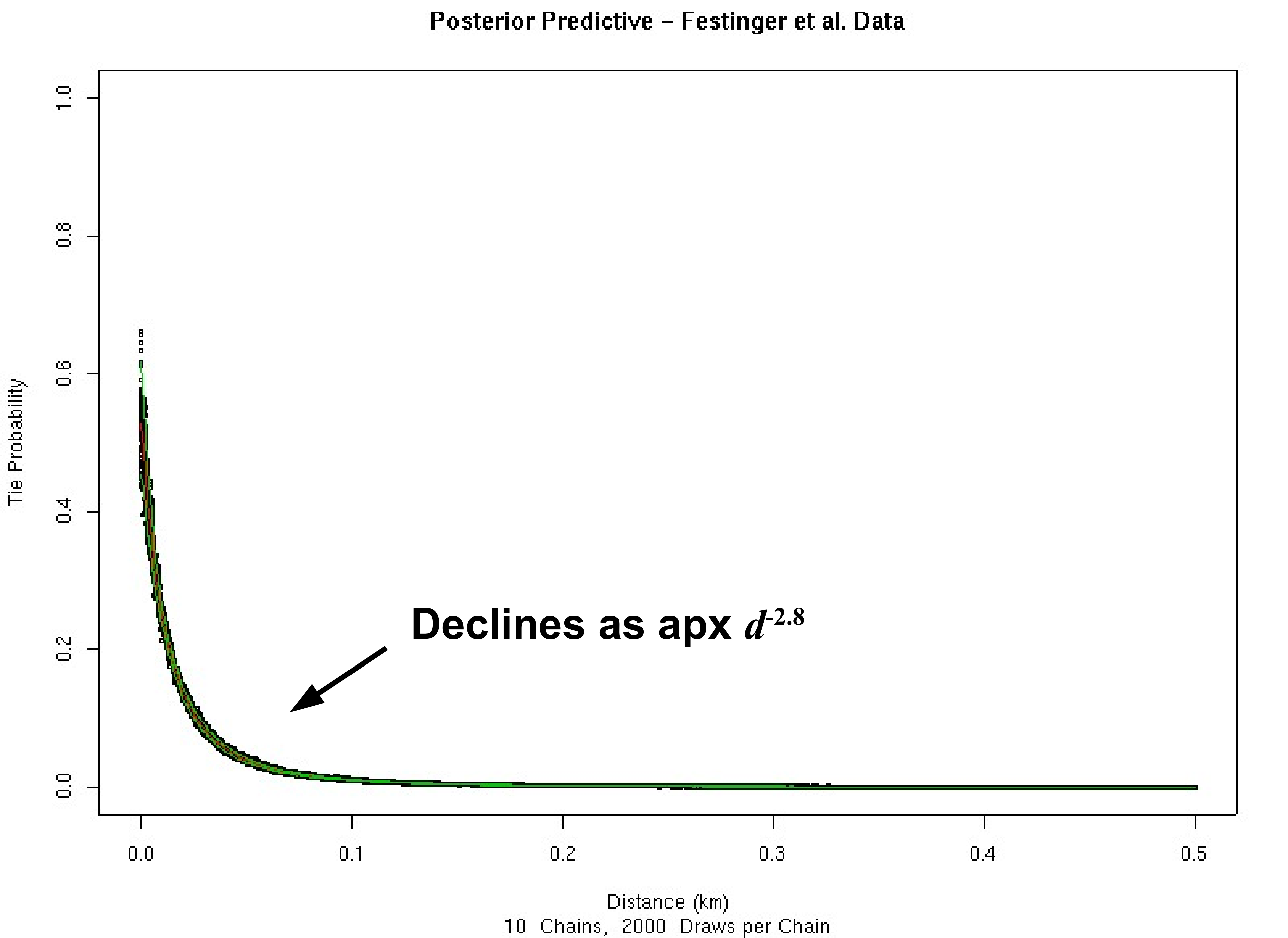
Artificial Elevation in Lawrence, KS Models

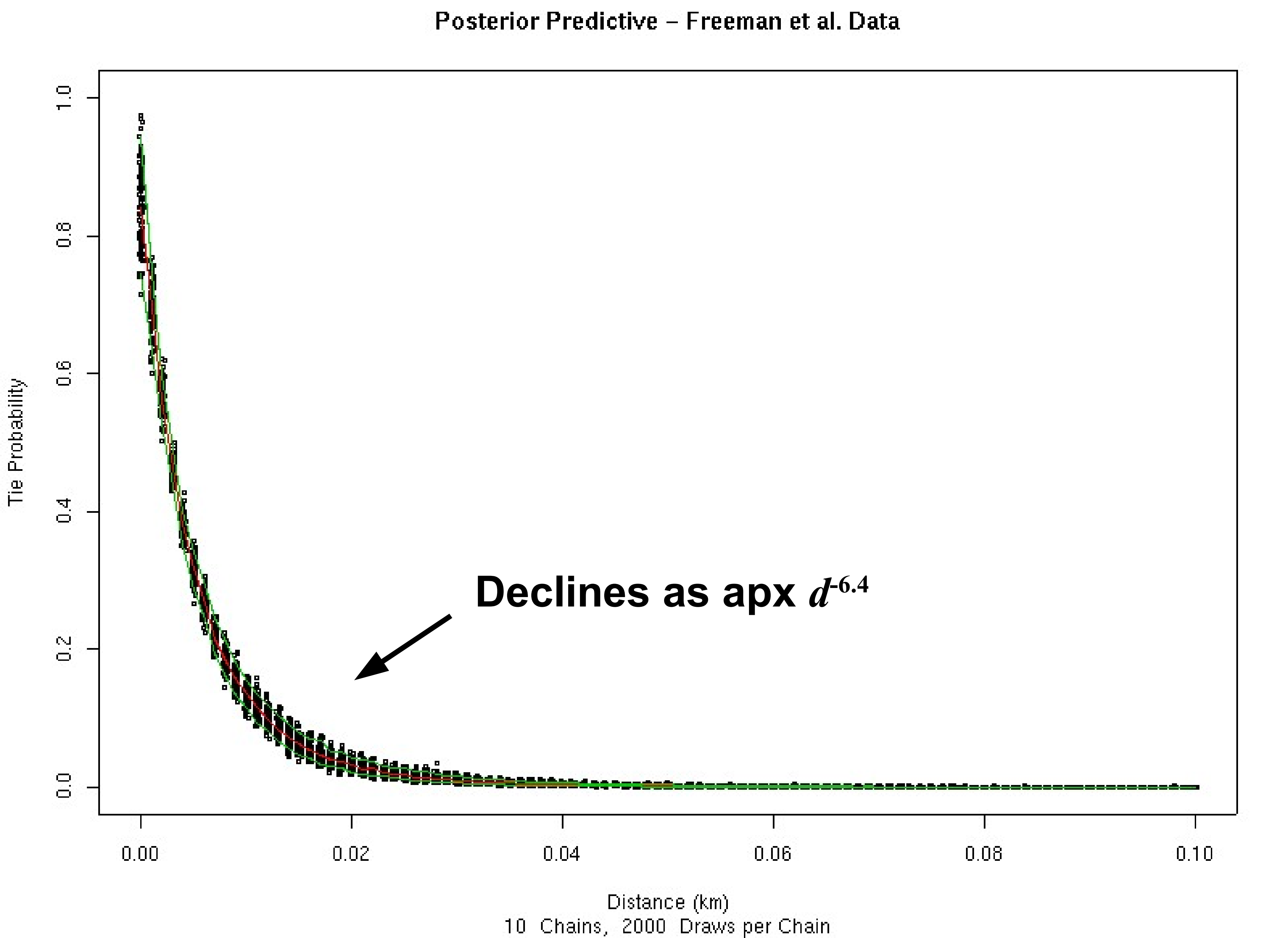




Choosing an SIF: Two "Test" Relations

- **Festinger et al. (1950) - "Social Friendship"**
  - Collected in post WW-II housing project during 1946-47
  - Subjects asked to provide three people "you most see socially"
- **Freeman et al. (1988) - "Face-to-Face Interaction"**
  - Collected on a southern California beach
  - 54 subjects observed for 60 hours over a 30 day period
  - Mean distance between actors and minutes interacting reported
- **For each data set, numbers of possible, observed edges at each distance used to infer SIF**
  - Bayesian inhomogeneous Bernoulli graph model, SIF selected by Bayes Factor

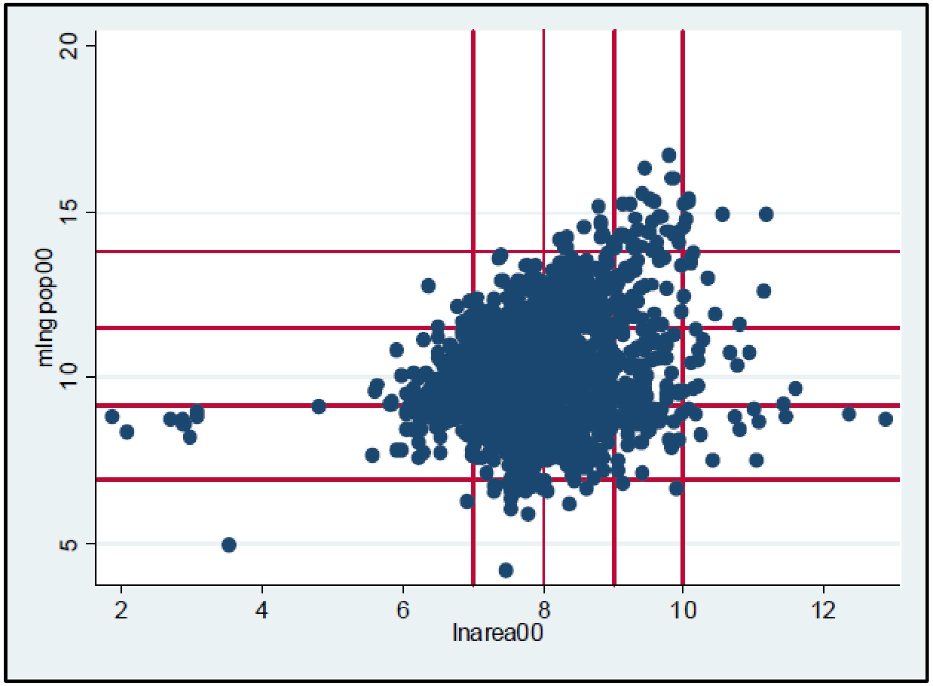

Posterior Predictive – Festinger et al. Data
Tie Probability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
Declines as apx $d^{-2.8}$
Distance (km)
10 Chains, 2000 Draws per Chain



Posterior Predictive – Freeman et al. Data
Tie Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
Distance (km)
10 Chains, 2000 Draws per Chain
Declines as apx $d^{-6.4}$

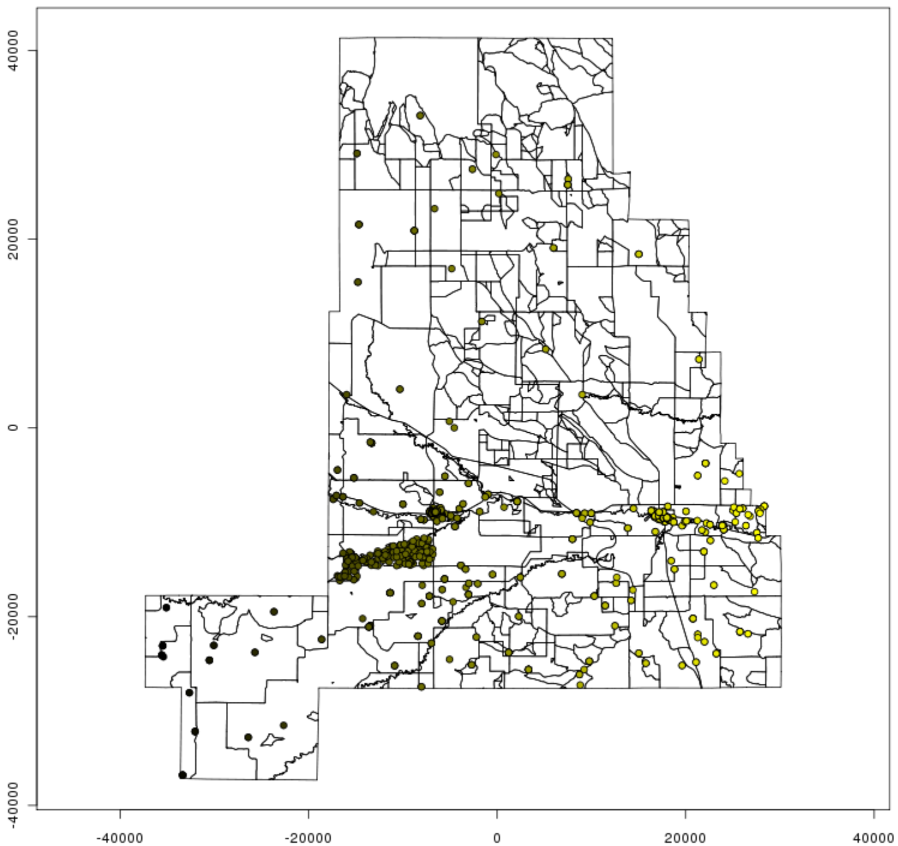
Geographical Cases

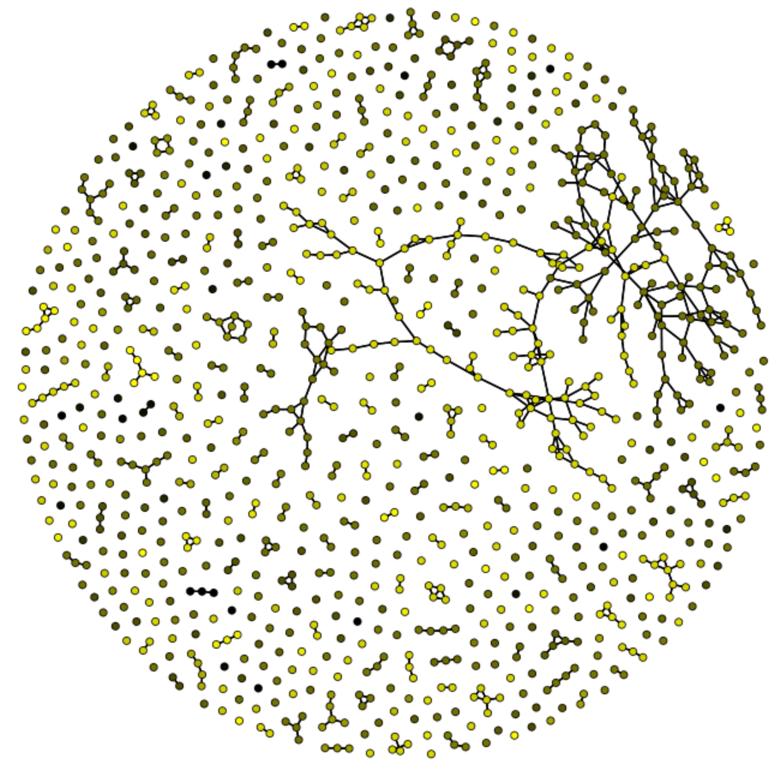
- **In progress: examine a range of functional settlements, stratified by total population and spatial area**
  - Started with all US micropolitan and metropolitan areas (as defined by the US census
  - For population, consider approximate size strata of 1,000, 10,000, 100,000, and 1,000,000
  - For area, consider log area strata of 9, 10, 11, and 12
  - For each combination of conditions, sought settlement with minimum least squares deviation from desired population/land area
- **Network simulation performed on each settlement, using each coordinate generation model and each relation**



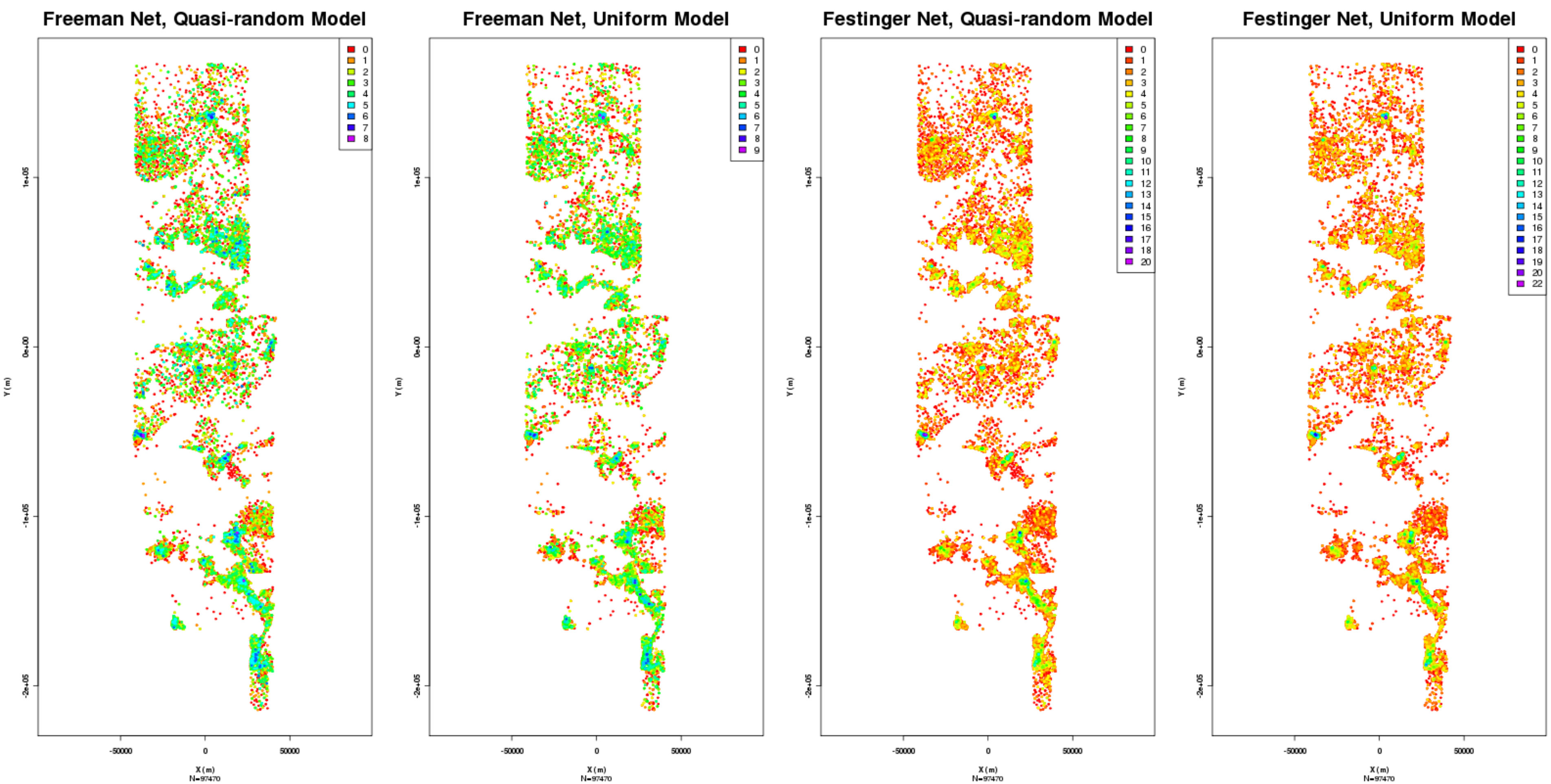
US Settlement Distribution by Log Population, Land Area

Golden Valley, MT: Festinger Net, Uniform Model

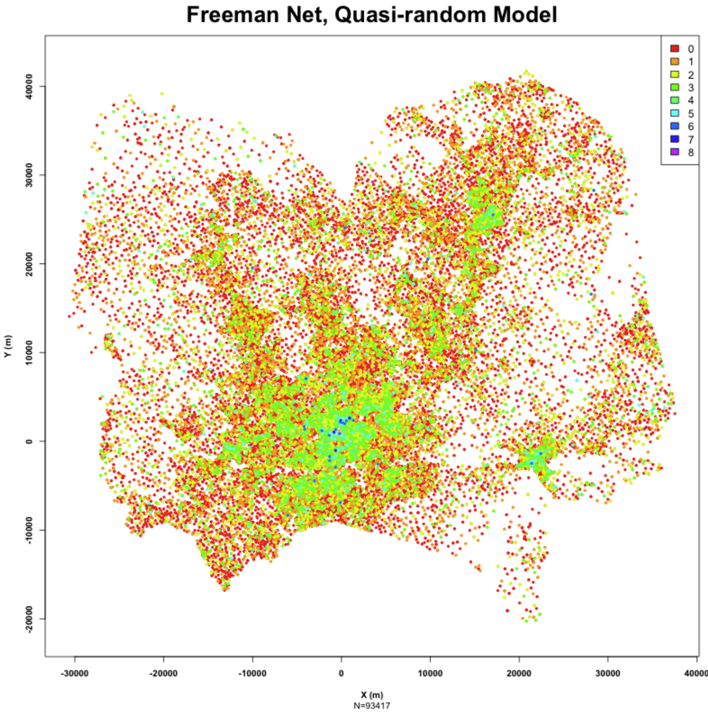


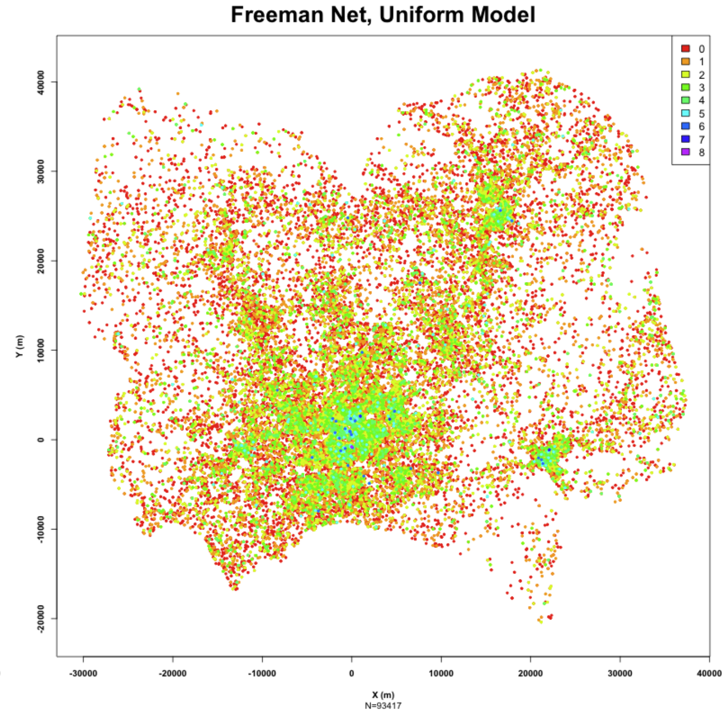


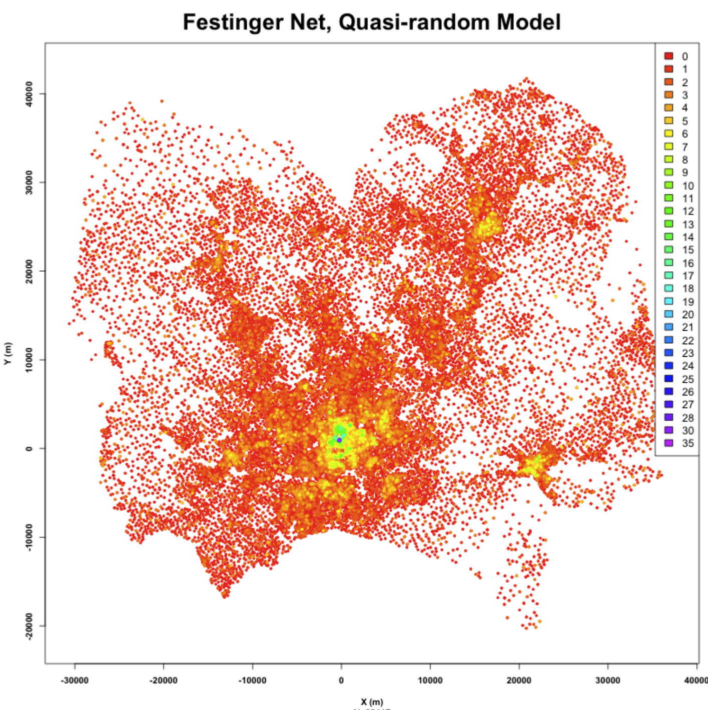
Spatial Distribution of Degree: Navajo, AZ

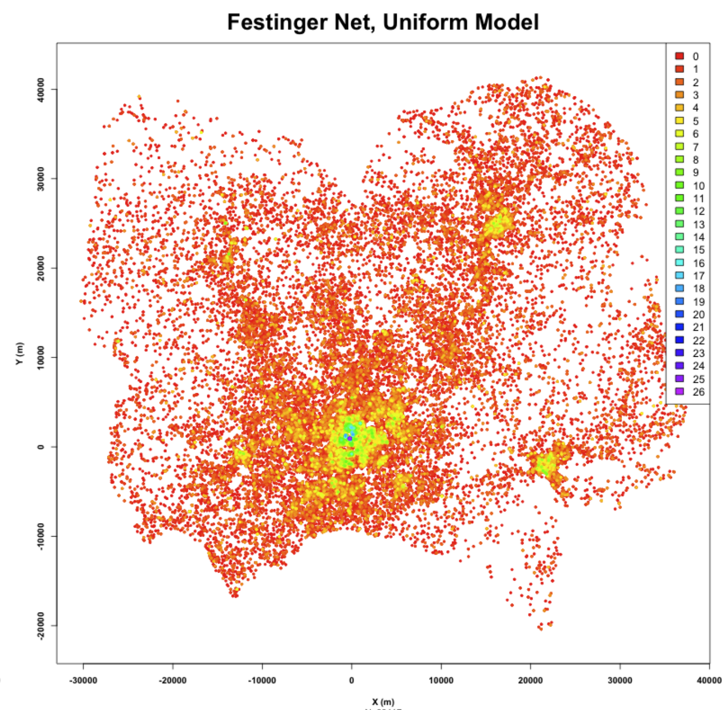


Spatial Distribution of Degree: Cookeville, TN

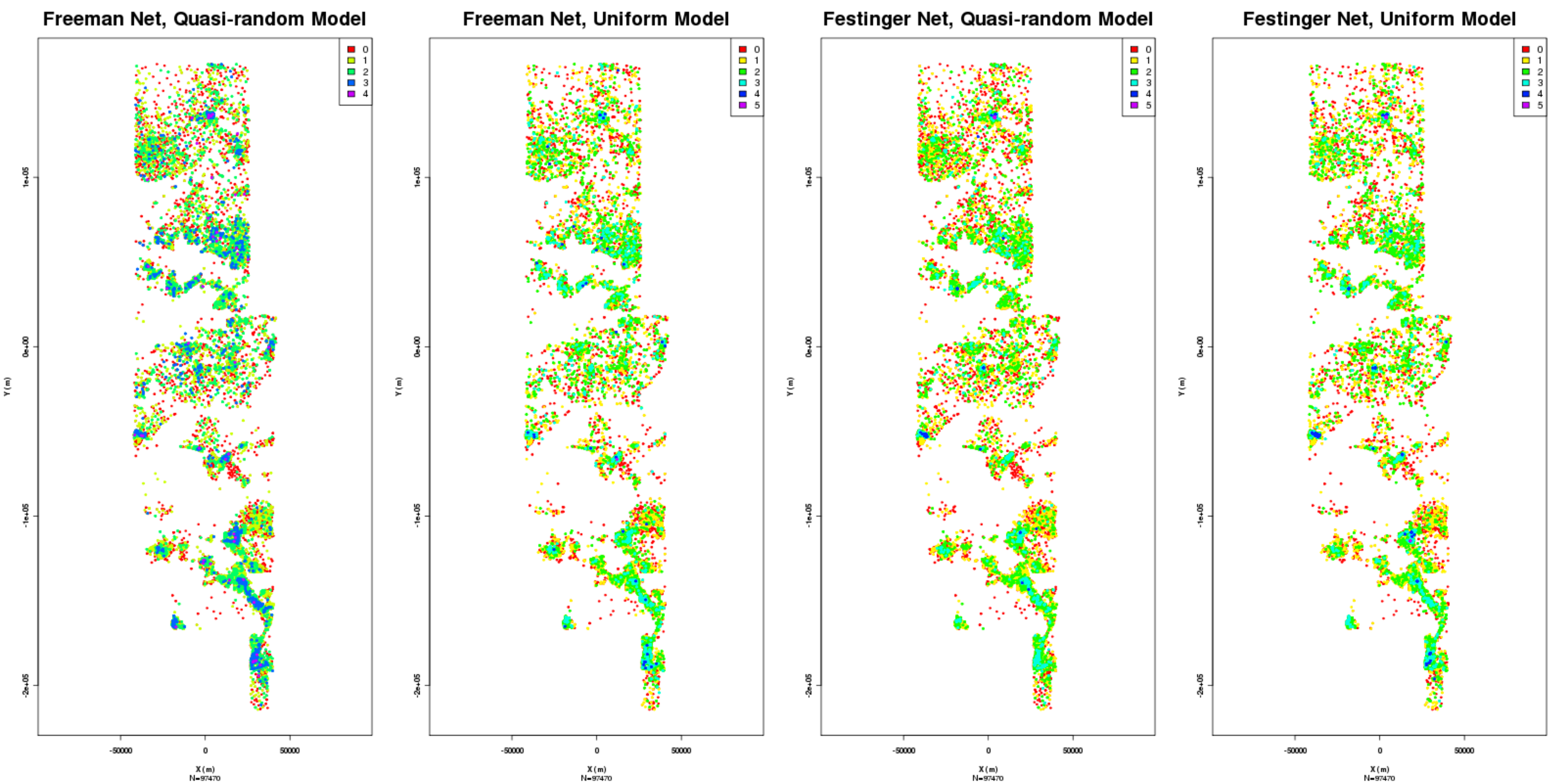




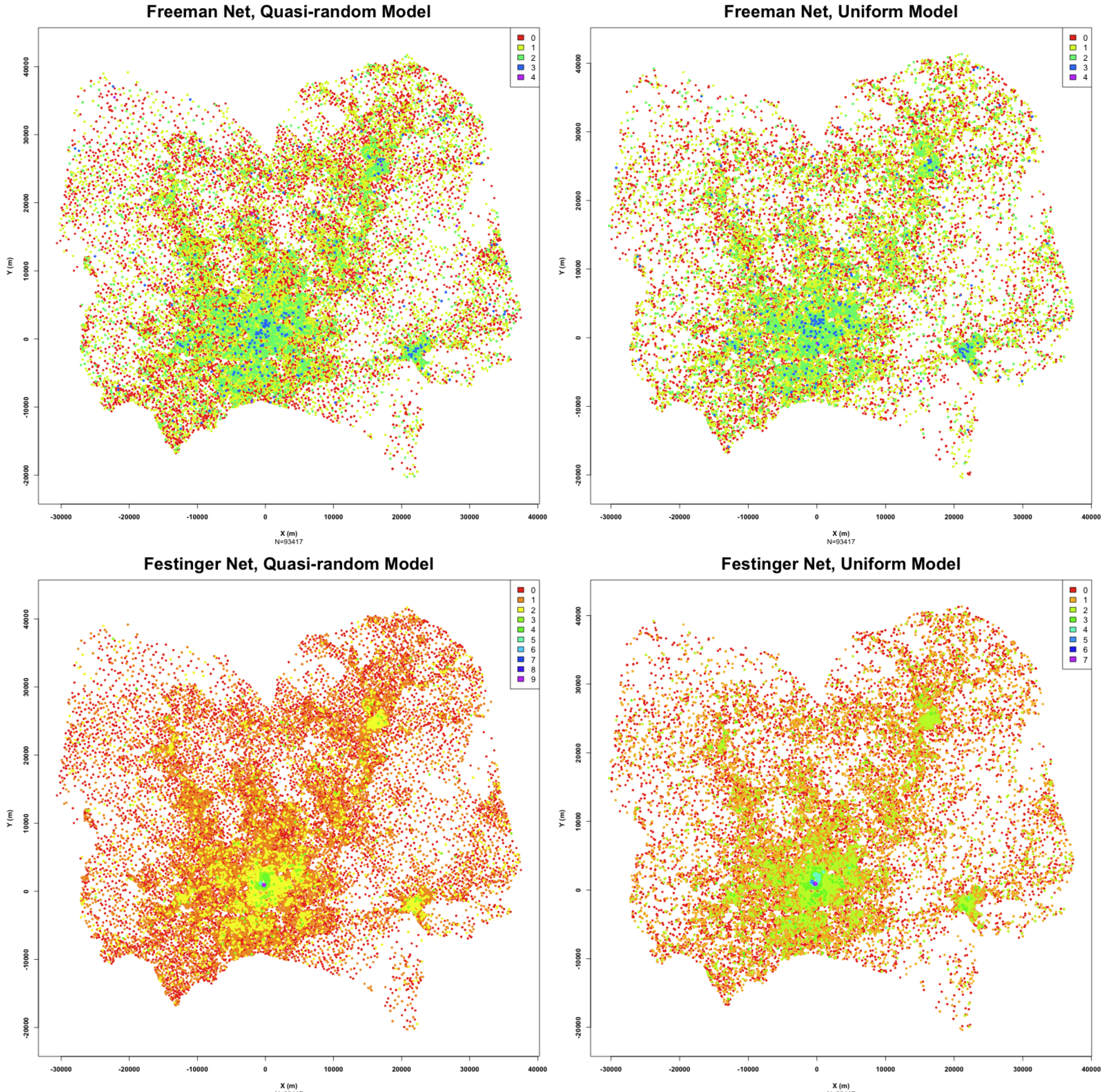




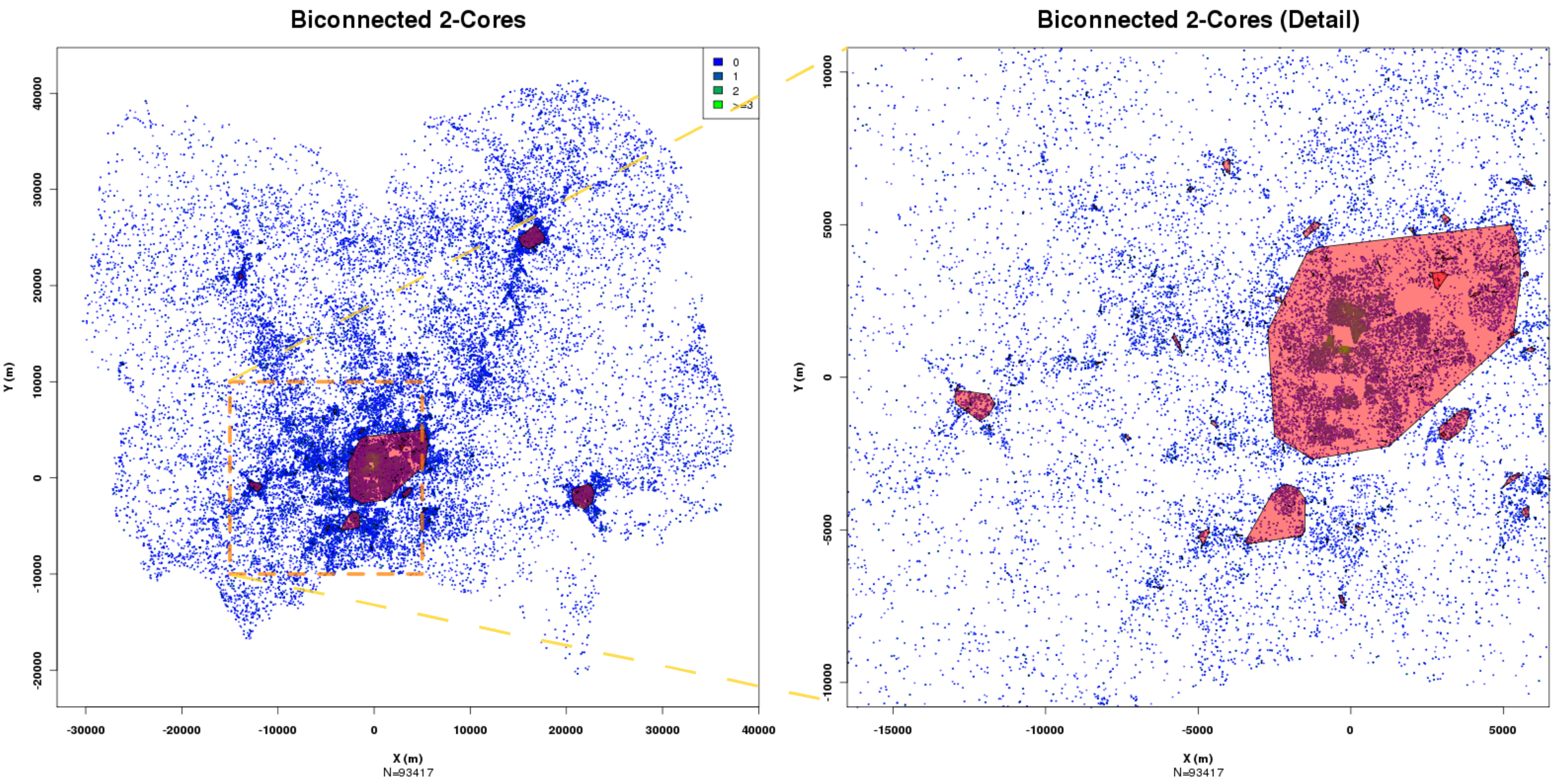
Spatial Distribution of Core Number: Navajo, AZ



Spatial Distribution of Core Number: Cookeville, TN



Spatial Coverage of Robustly Connected Groups



Initial Impressions from the Simulation Studies (So Far)

- **Clear impact of spatial clustering on network structure**
  - Spatial clustering raises density only slightly, but small changes have large threshold effects
  - Cluster spatial extent, proximity is also important (not just local intensity): cross-cluster ties can sustain connectivity
  - Spatial heterogeneity has real consequences
- **Choice of point placement algorithm does make a difference**
  - Halton sequence tends to suppress clustering, degree, by maximizing the minimum distance between households; uniform generates "clumps" that elevate local tie volume
  - Effect should be largest for short-range ties, light-tailed SIFs, or properties based on local clustering; weakest for long-range interaction

Some Ongoing Questions

- **How to further improve performance?**
  - Gridding system is often good, but requires tuning (and fails in places like Honolulu)
  - Going to try R-trees - other suggestions?
  - Network objects getting very big; may need to think of ways to segment and partially load on use (not so easy)
- **Closely related problem (much work done here, but not shown!): estimating tie volumes between regions**
  - MC quadrature using Halton sequences turned out to be expensive - a fast quasi-random strategy would be nice
    - Point-in-polygon calculations are the current bottleneck
  - Now using R-trees to speed performance; seems to work well, but is there a better way?

Summary

- **We now have tools for simulating spatially embedded networks on fairly large scales**
  - Currently works up to around 1e6 nodes
  - Can use real geography; limited to projections for exact network (for now), but tie volumes are on WGS ellipsoid
  - Assembling a large collection of test locations, networks
- **Interesting opportunities going forward**
  - More performance enhancements
  - Using testbed sims to evaluate algorithm performance
  - Applying modeling strategy to simulation of latent space models, fast approximate prediction (e.g., tie volumes between groups for block models), etc.