

Sampling in Statistics and Research

Steve Thompson

thompson@stat.sfu.ca

Simon Fraser University

Fall 2009 Joint UBC/SFU Graduate Student Workshop

21 November 2009

Outline

1. **Examples**
2. **Sampling ideas**
3. **Designs etc.**

Sampling

The problem: We observe only a sample, but are interested in characteristics of the wider population or distribution.

Population of units: $1, 2, \dots, N$

Variables of interest: y_1, y_2, \dots, y_N

Sample: \mathcal{S} , a subset or sequence of units from the population.

Sampling in networks

Population of units or nodes: $1, 2, \dots, N$

Node variables of interest: y_1, y_2, \dots, y_N

Link-indicators or weights: $w_{ij}, \quad i, j = 1, \dots, N$

(Variables of interest associated with **pairs** of nodes)

Sample: A subset or sequence \mathcal{S} of *units* and *pairs of units*
from the population: $\mathcal{S} = (\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$

y is observed in $\mathcal{S}^{(1)}$.

w is observed in $\mathcal{S}^{(2)}$.

Types of sampling designs

The procedure by which we select the sample.

Conventional design: $p(s)$

Procedure for selecting the sample does not depend on values of variables of interest observed during the survey.

Adaptive design: $p(s \mid y)$

Procedure for selecting sample can depend on values of variables of interest.

(Design can also depend on auxiliary variables x .)

Approaches to inference from samples

Design based approach:

The values of the variables of interest in the population are fixed, unknown constants.

$$\mathbf{y} = (y_1, \dots, y_N)$$

$$\mathbf{w} = \{w_{ij}\}, i, j \in \{1, \dots, N\}$$

Probability enters only through the design

Model based approach:

The population values are random variables, which we try to model.

$Y_1, \dots, Y_N, W_{11}, \dots, W_{NN}$ have some joint probability distribution, described by a stochastic graph model

Optimal sampling strategies

Find the **design** $p(s | \mathbf{y})$ and **estimator** \hat{Z} of population quantity Z to minimize the mean square error

$$E(\hat{Z} - Z)^2$$

subject to unbiasedness, $E(\hat{Z}) = E(Z)$

The optimal strategy is in most cases an adaptive one.

Reasoning:

1. Stop part way through the survey and look at what has been observed so far:

initial sample and values (s_1, \mathbf{y}_{s_1})

2. Choose the rest of the sample s_2 to minimize the mean square error of the estimate **given** what has been observed so far.

$$\min E \left[(\hat{Z} - Z)^2 \mid s_1, \mathbf{y}_{s_1} \right]$$

(Zacks 1969, Thompson and Seber 1996, Chao and Thompson 2000)

Sufficiency, completeness, Rao-Blackwell

sampling **data** = (s, y_s)

sufficient statistic = set of distinct units, associated y values

Rao-Blackwell estimate = $E[\text{simple estimator} \mid \text{sufficient statistic}]$

Minimal sufficient statistic is not complete so more than one possible estimator.

Likelihood function

$$\text{Prob}(\text{data} \mid \text{parameters}) = P(s, \mathbf{y}_s \mid \theta)$$

$$\begin{aligned} L(\theta; s, \mathbf{y}_s) &= \int p(s \mid \mathbf{y}; \theta) f(\mathbf{y}; \theta) d\mathbf{y}_{\bar{s}} \\ &= \int (\mathbf{design})(\mathbf{model}) d(\text{unobserved}) \end{aligned}$$

Wrong answer without design!

“Ignorable” design

If the design depends only on values that are observed and recorded in the data, then the design disappears from likelihood-based estimates.

$$L(\theta; s, \mathbf{y}_s) = p(s \mid \mathbf{y}_s; \theta_1) \int f(\mathbf{y}; \theta_2) d\mathbf{y}_{\bar{s}}$$

(Caveat: May need to **implement** a probability design to make the design ignorable!)

Design-induced distribution

Implement a probability design $p(s)$ to select sample s .

Look at the distribution of any sample statistic induced by the design.

Avoids assumptions about population

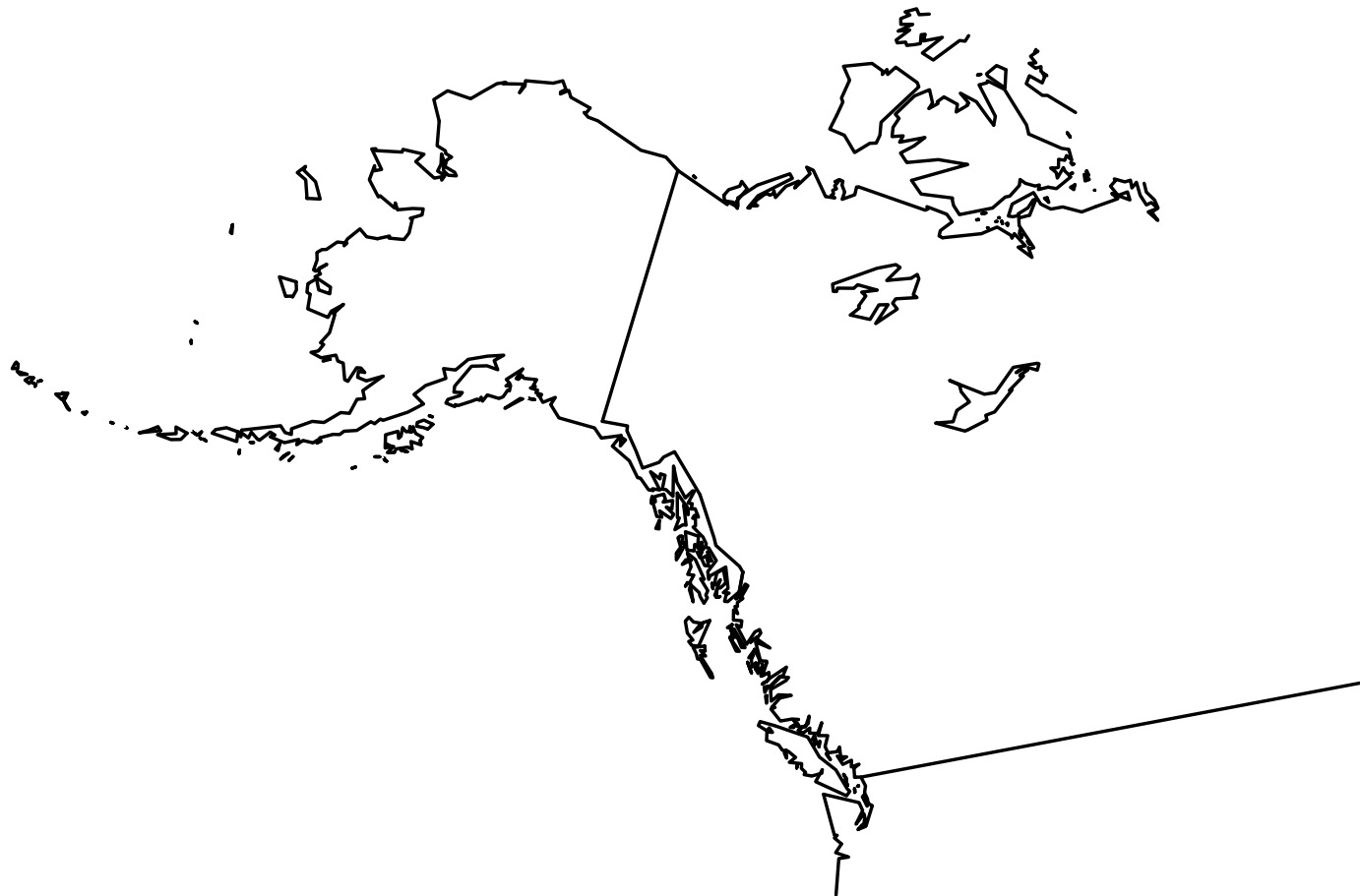
Modeling approach

assume statistical model $f(y;\theta)$ for population values

can help with design and inference

Works best when good design is implemented!

Surveys of fish and shellfish



Trawl survey, Kodiak Island

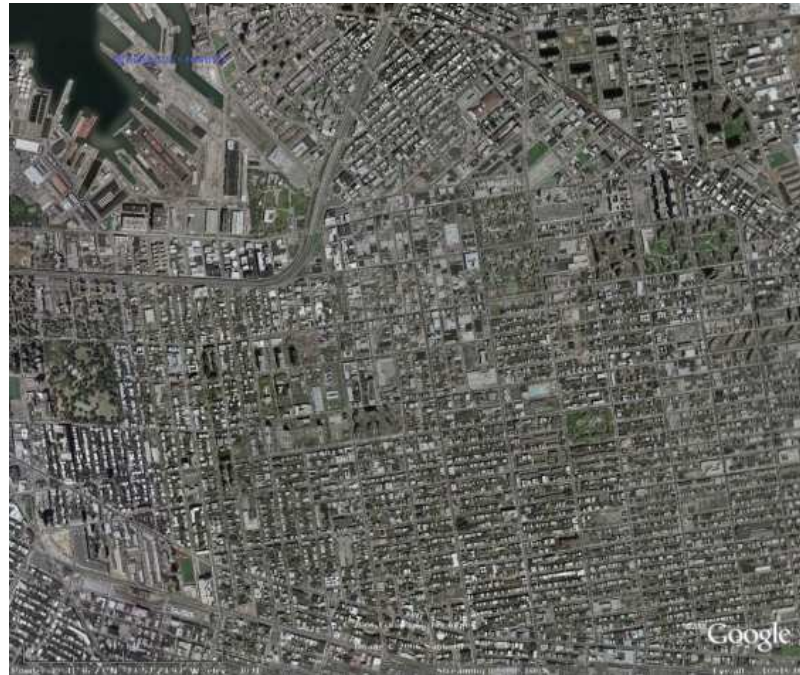


Migratory waterfowl survey



J.I. Hodges

Studies of hidden populations

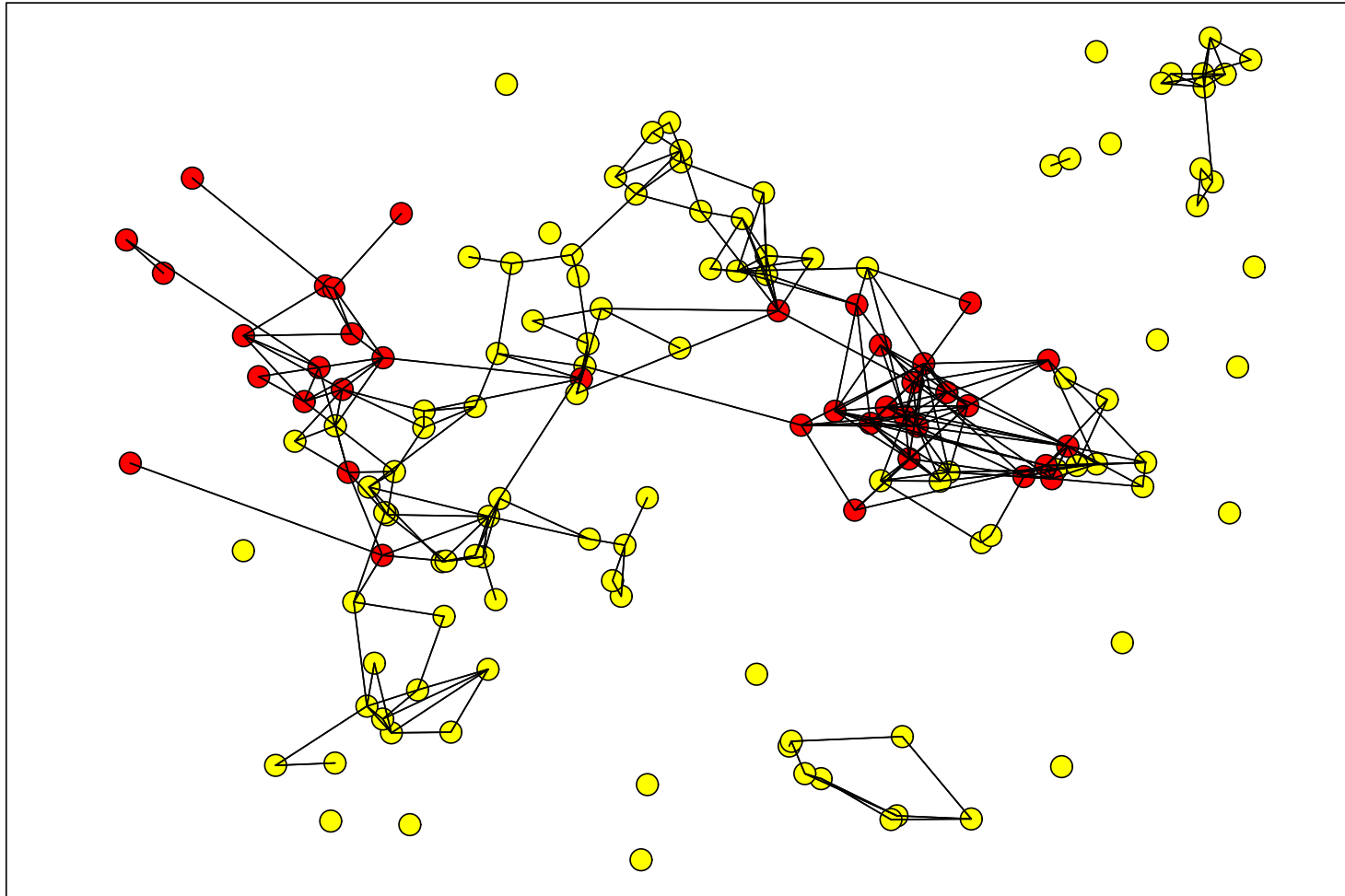


HIV/AIDS at-risk study

M. Miller

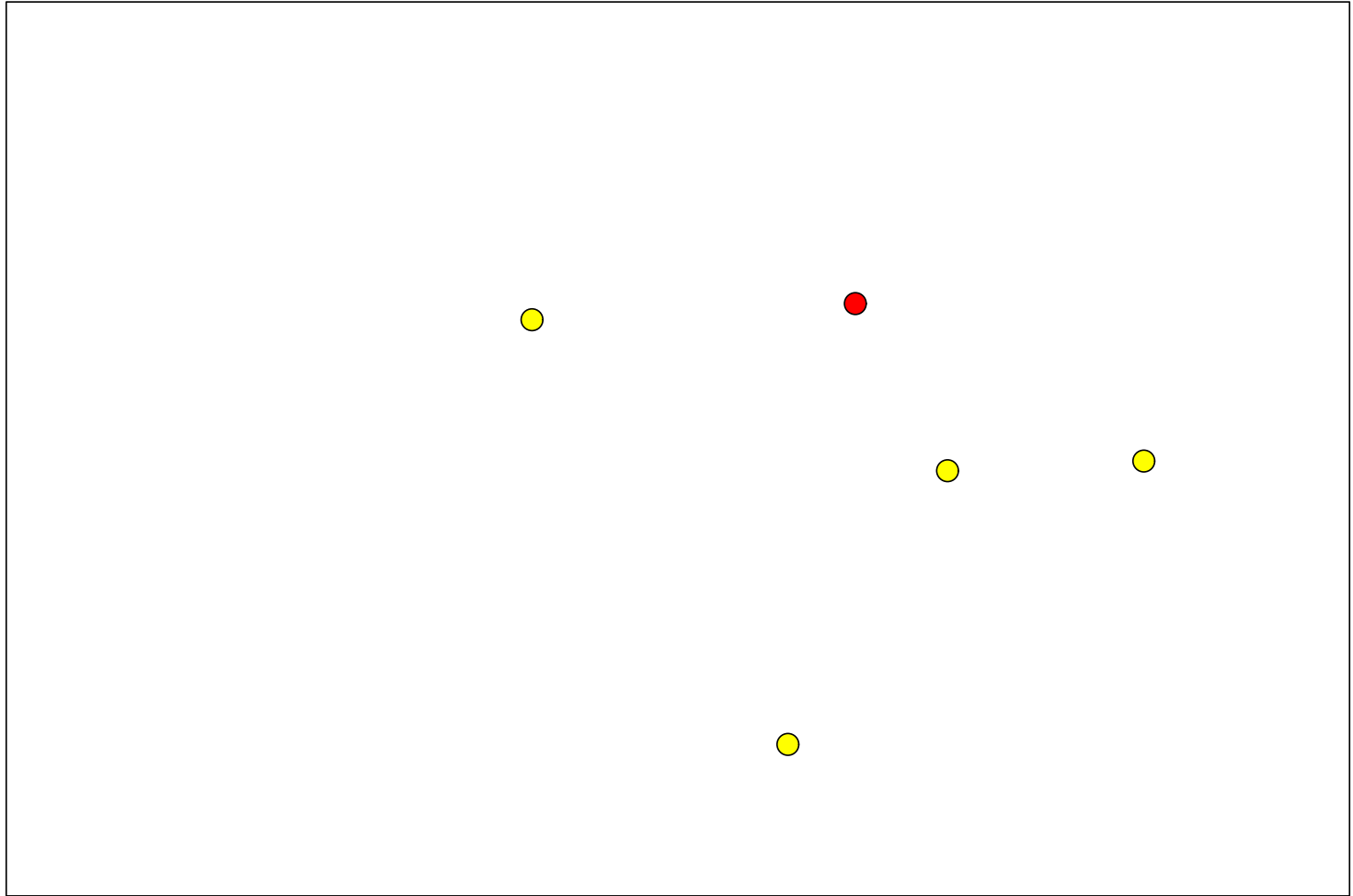
Example network population

population graph



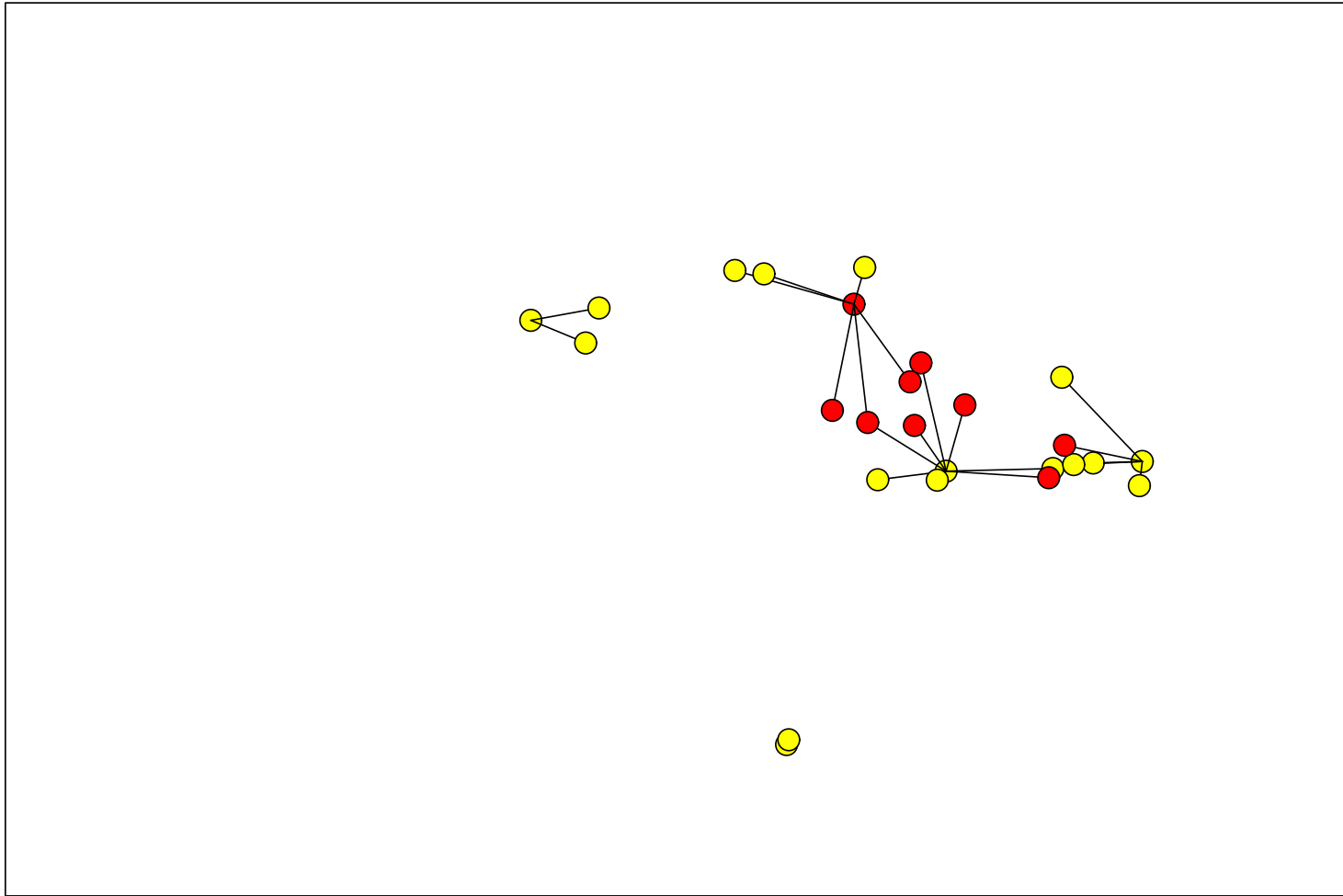
Random sample

sample



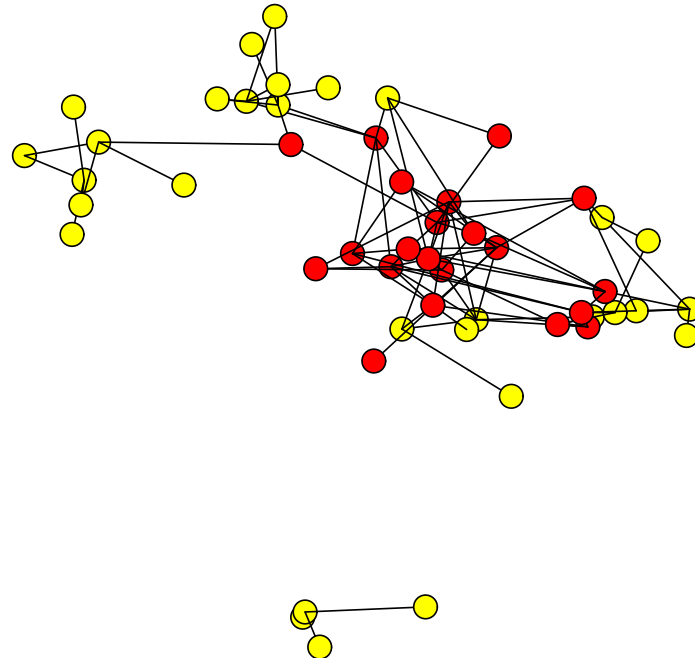
Snowball sample

sample

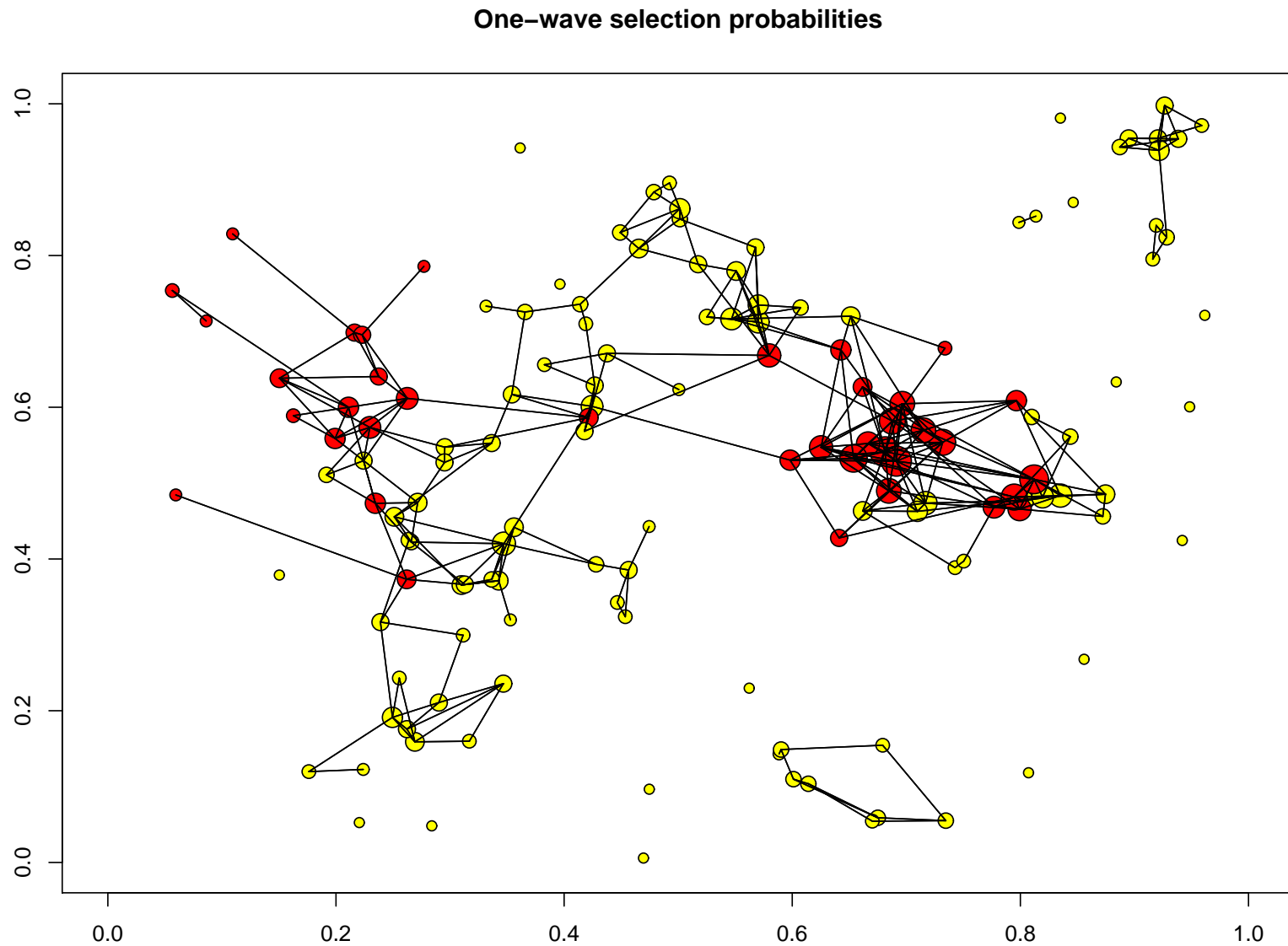


Snowball sample

sample

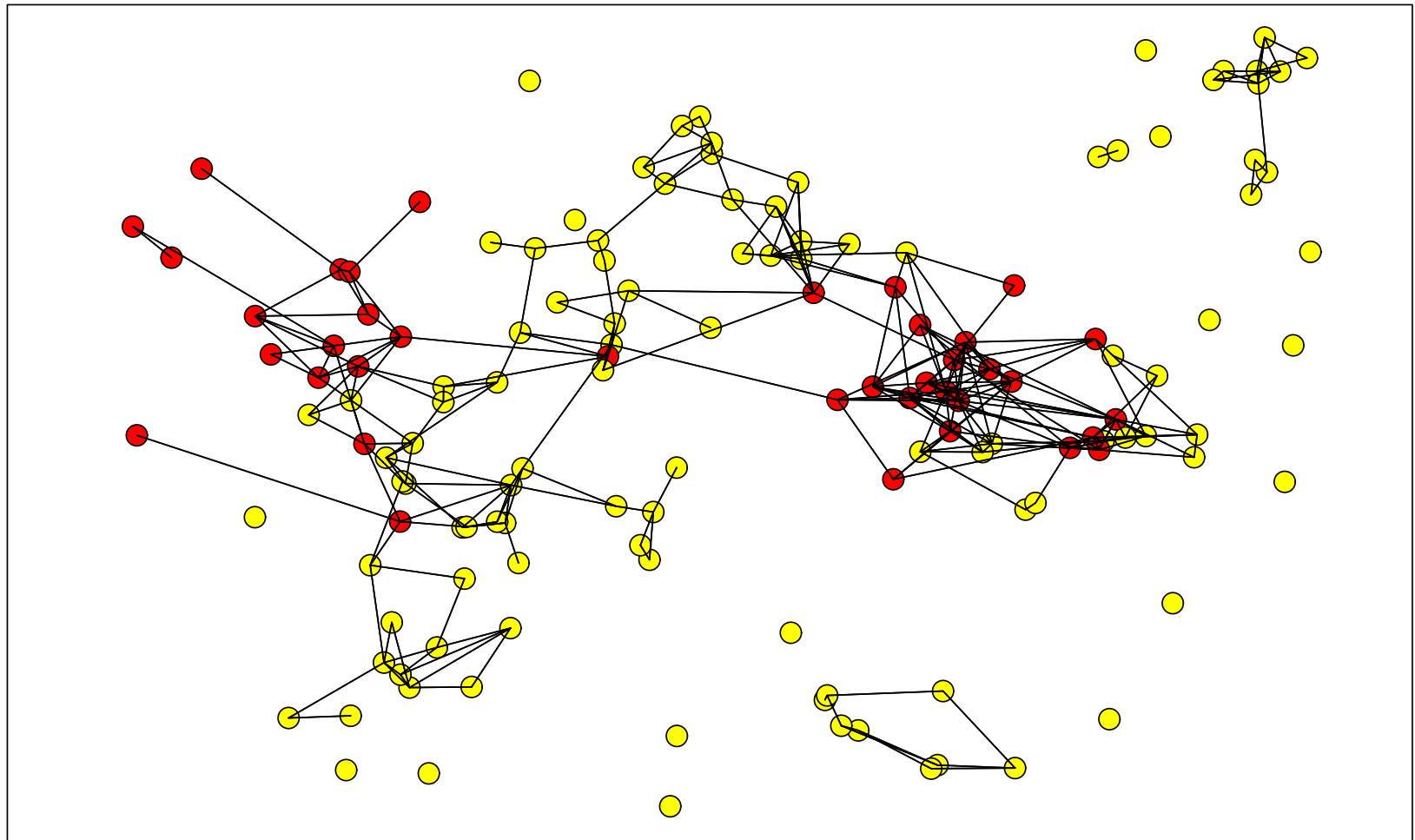


One-wave snowball selection probabilities



The population again

population graph



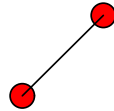
Random walk sample

walk



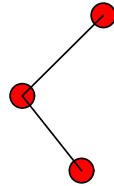
Random walk sample

walk



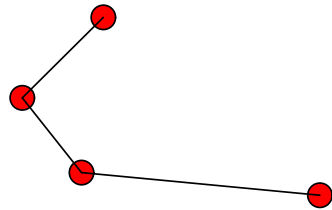
Random walk sample

walk



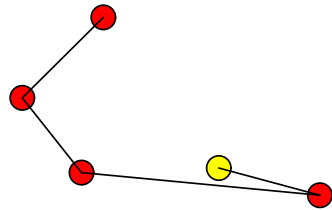
Random walk sample

walk



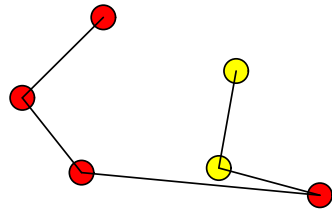
Random walk sample

walk



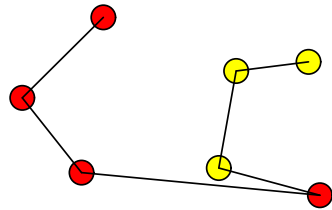
Random walk sample

walk



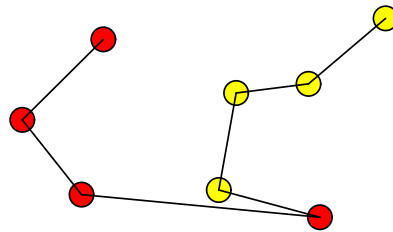
Random walk sample

walk



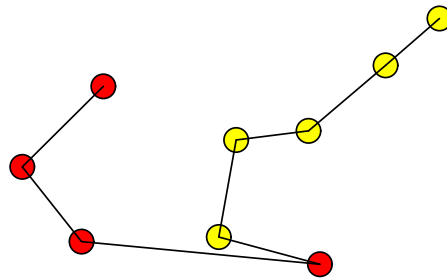
Random walk sample

walk



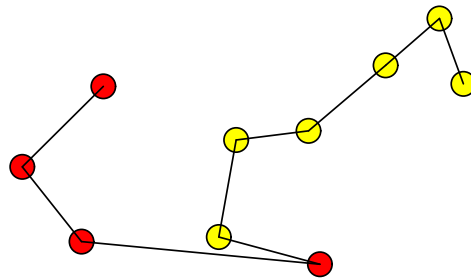
Random walk sample

walk



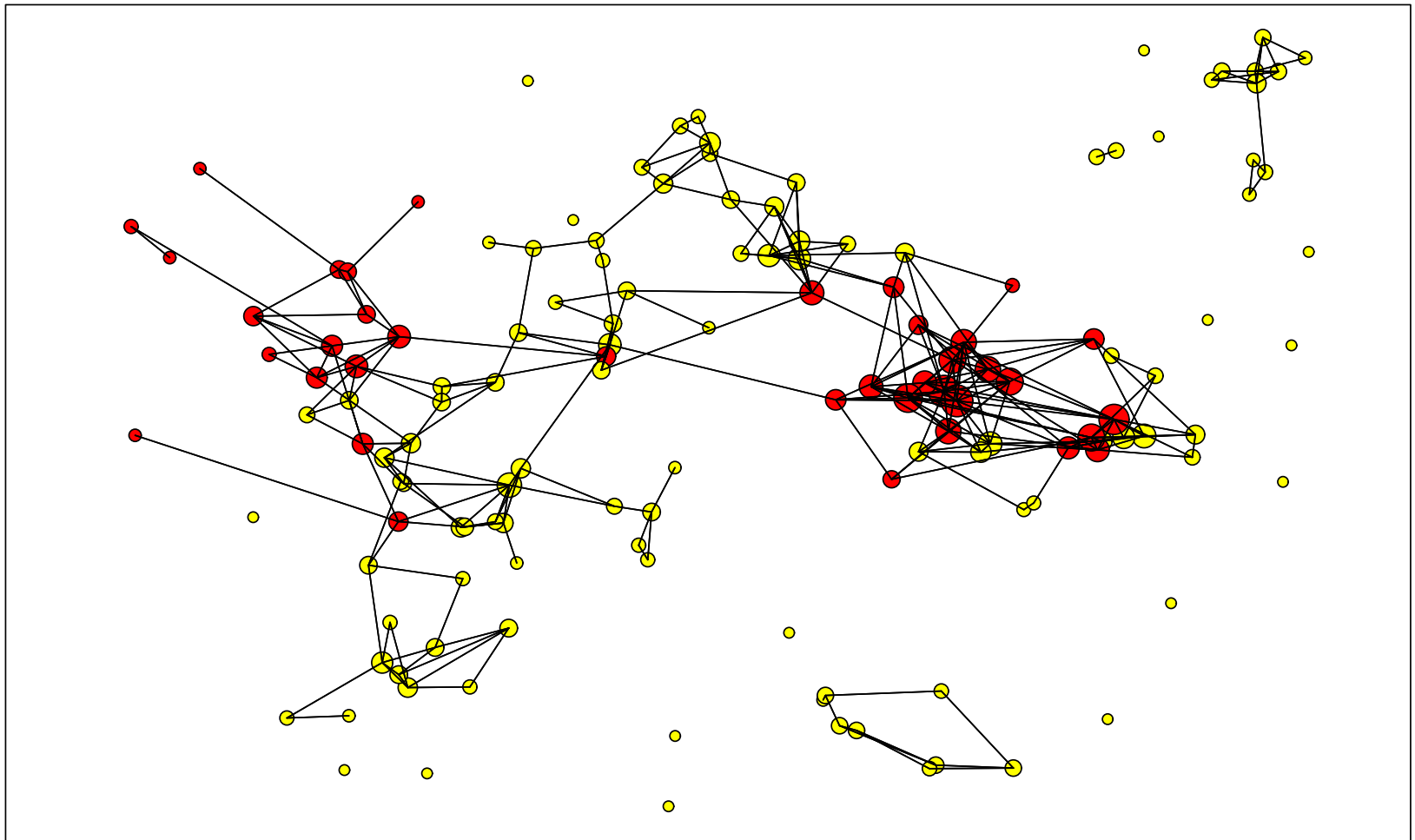
Random walk sample

walk



Random walk limit selection probabilities

Limit random walk probabilities



Random walk as Markov chain

W_k is the node of the graph selected at k th wave.

$a_{ij} = 1$ indicates a link from node i to node j .

$\{W_0, W_1, W_2, \dots\}$ is a Markov chain with

$$P(W_{k+1} = j \mid W_k = i) = a_{ij}/a_i.$$

Q is the transition matrix of the chain,

$$q_{ij} = P(W_{k+1} = j \mid W_k = i).$$

The stationary probabilities (π_1, \dots, π_N) satisfy $\pi_j = \sum \pi_i q_{ij}$
for $j = 1, \dots, N$.

Approach using limiting distribution of random walk

For **random walk** design **with-replacement** in a **single-component** network and if the links are **symmetric**, then the limiting selection probability is proportional to the person's **degree** (d_i)

Generalized ratio estimator of mean for behavioral characteristic y :

$$\hat{\mu} = \frac{\sum_s y_i / d_i}{\sum_s 1 / d_i}$$

4. Targeted random walk designs

1. Uniform random walk
2. More general targetting

Targeted walk designs

Let $\pi_i(y)$ denote the desired stationary selection probability for the i th node as a function of its value or degree.

The transition probabilities for the targeted walk are

$$P_{ij} = q_{ij}\alpha_{ij} \quad \text{for } i \neq j$$

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min \left\{ \frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1 \right\}$$

Adaptive web sampling

At any point in the sampling,

- the next unit or set of units is selected from a distribution that depends on the values of variables of interest in an **active set** of units already selected. (**follow a link**)

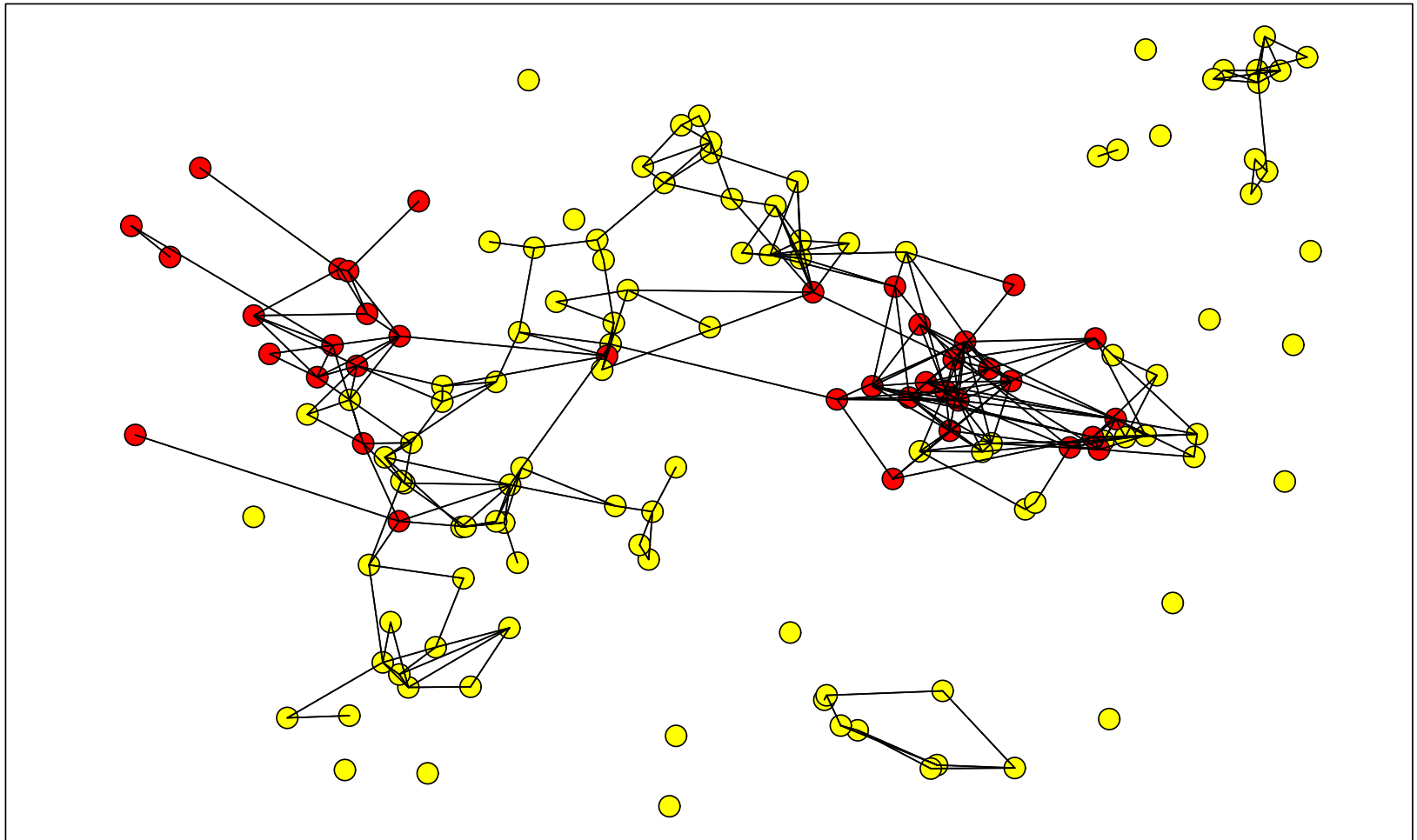
Adaptive web sampling

At any point in the sampling,

- the next unit or set of units is selected from a distribution that depends on the values of variables of interest in an **active set** of units already selected. (**follow a link**)
- With some probability, however, the selection may be made from a distribution not dependent on those values. (**random jump**)

Population graph

population graph



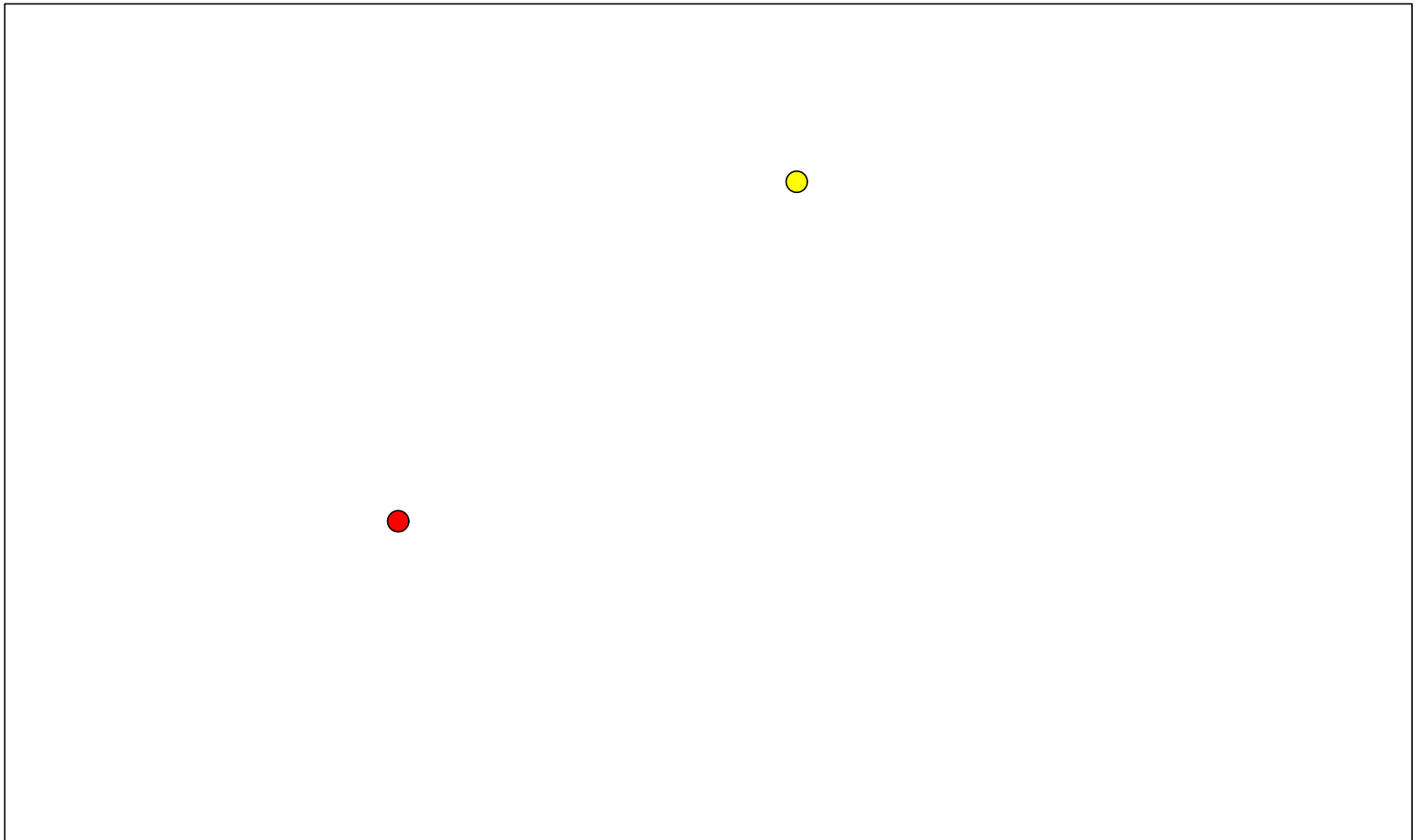
Adaptive web design

weighted links



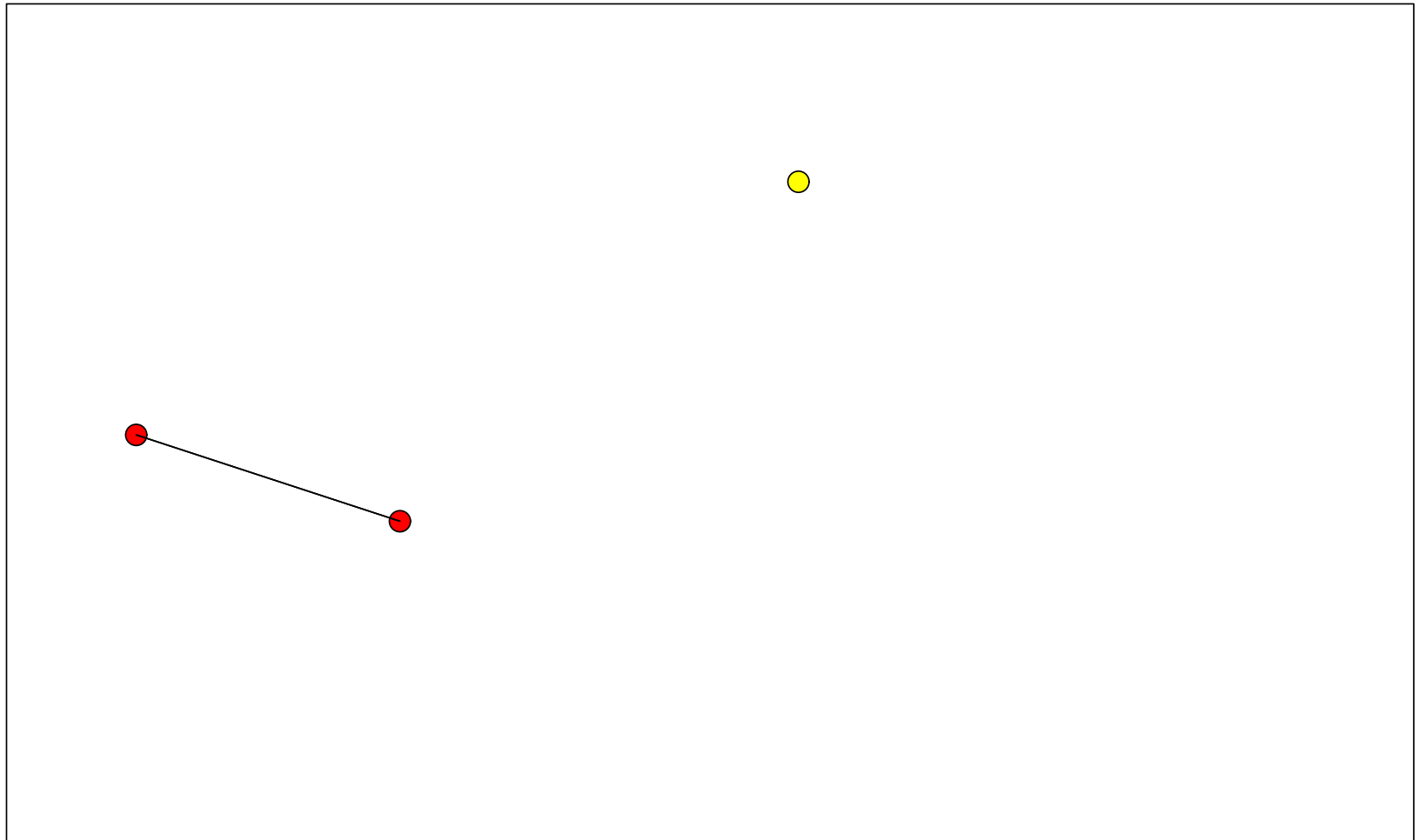
Adaptive web design

weighted links



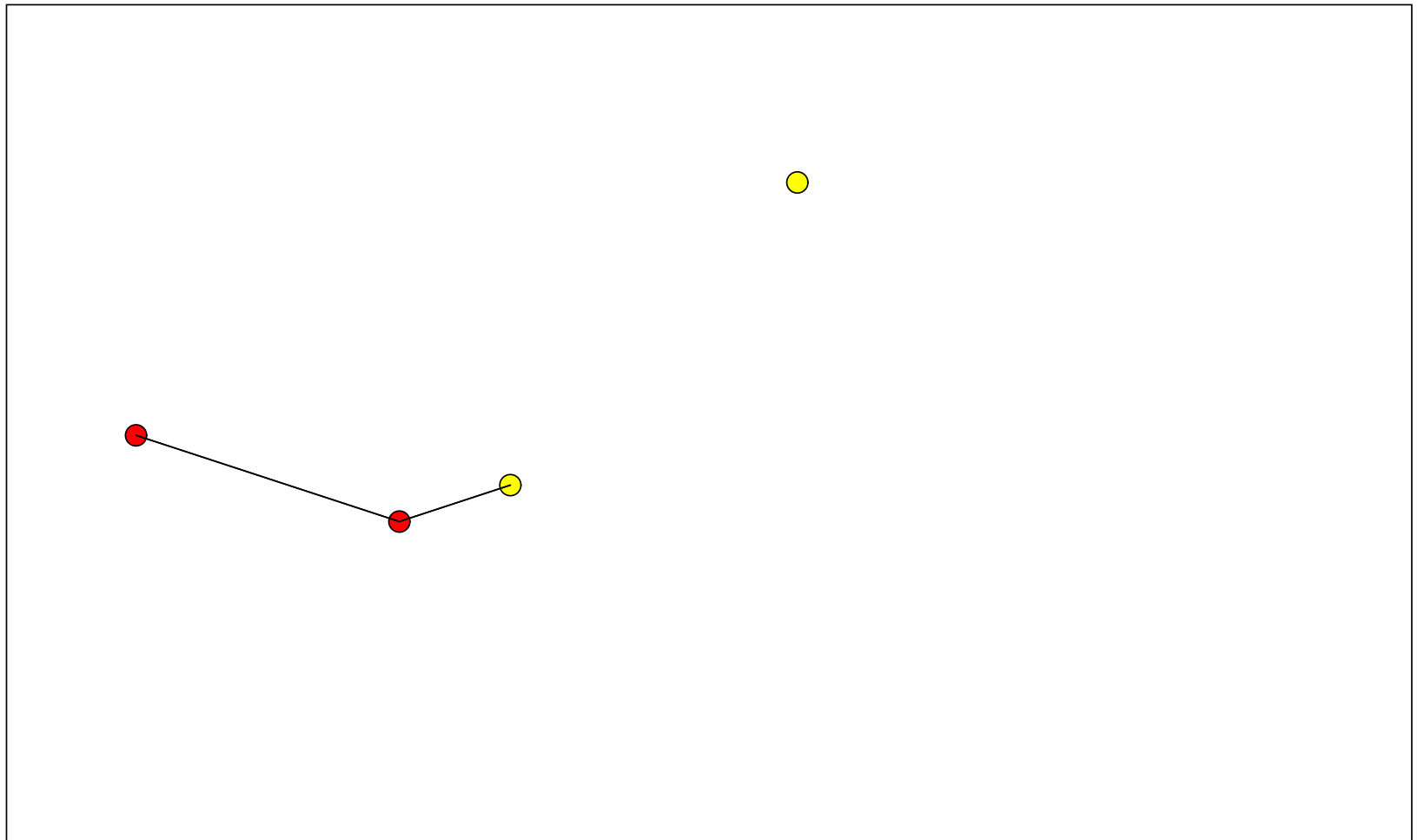
Adaptive web design

weighted links



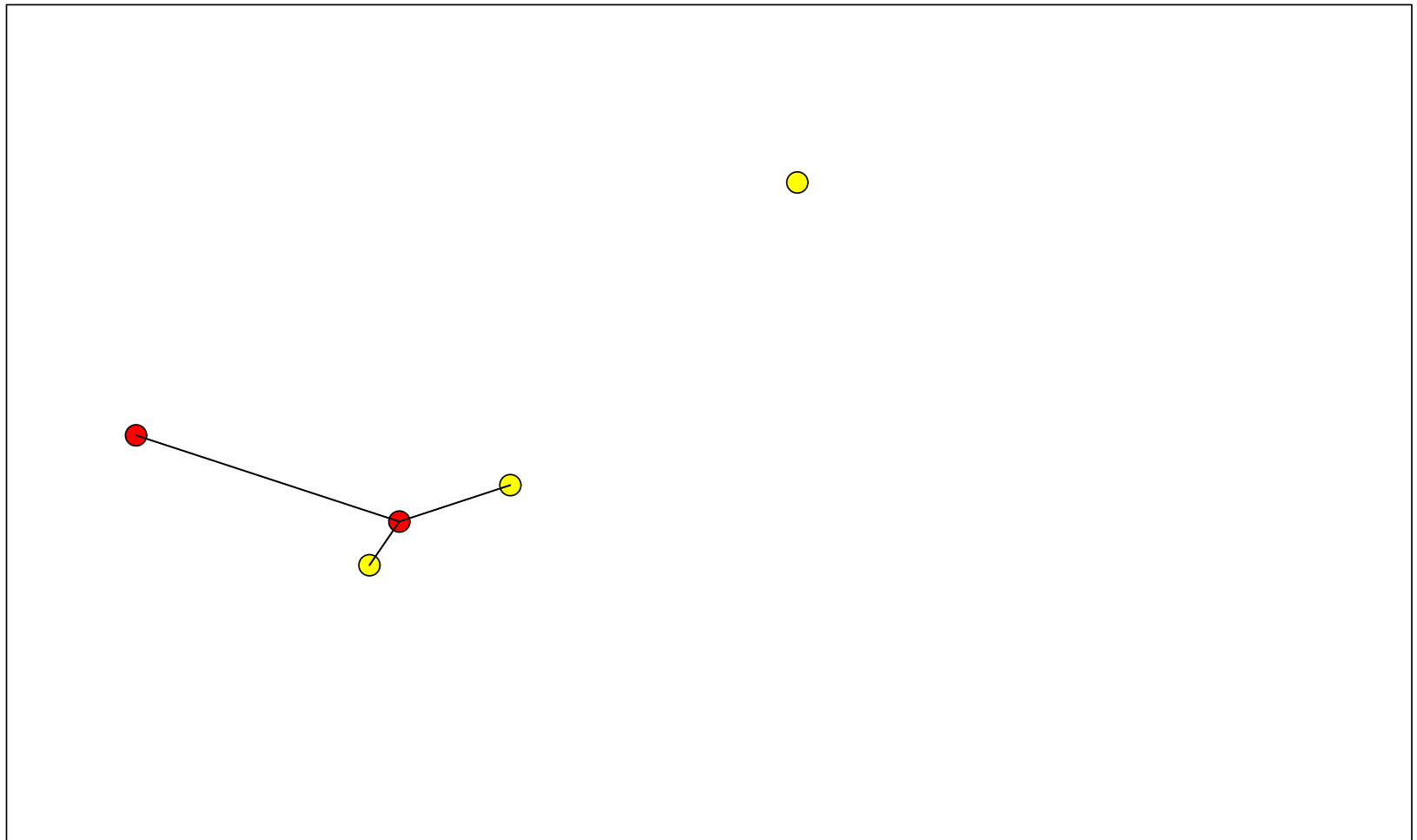
Adaptive web design

weighted links



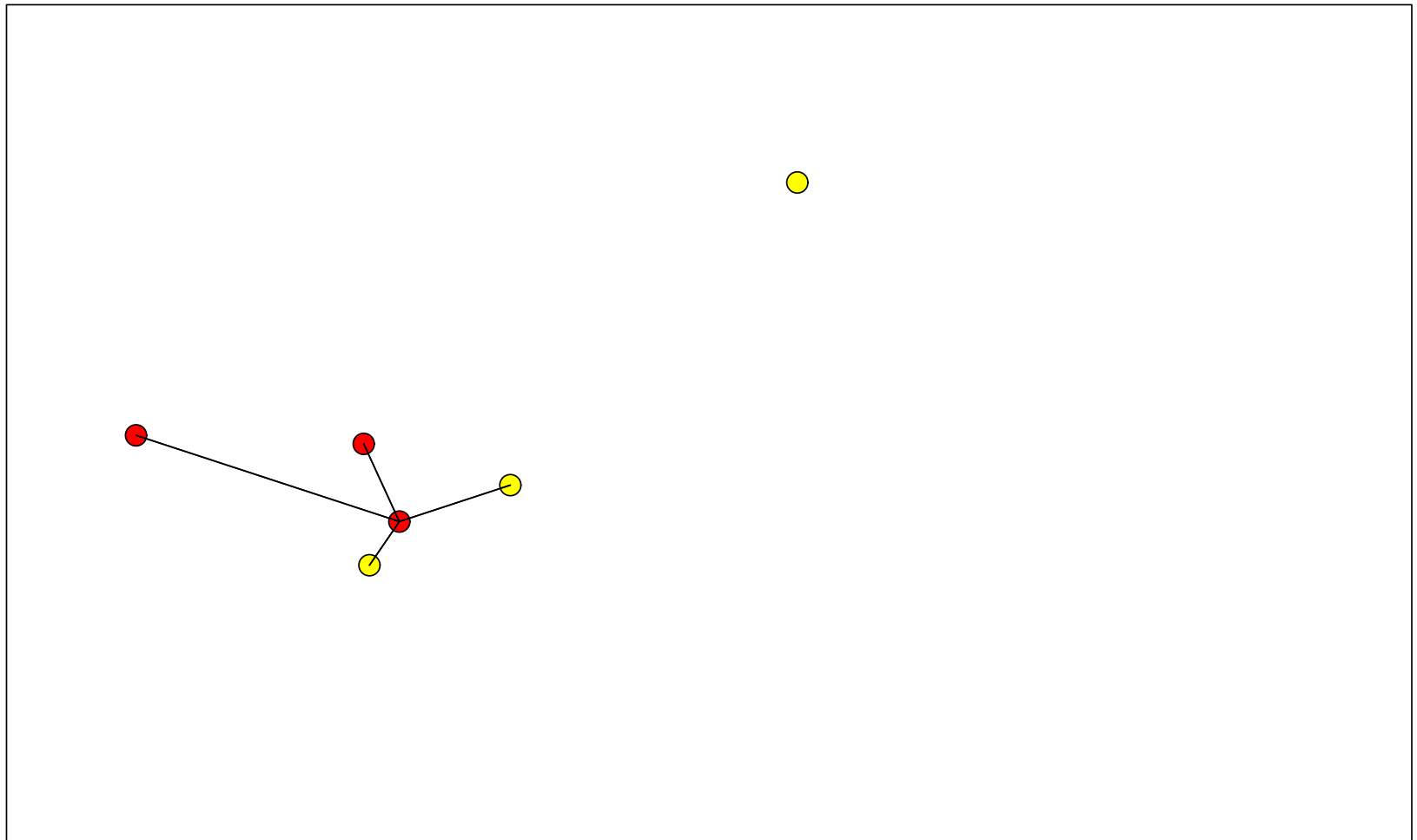
Adaptive web design

weighted links



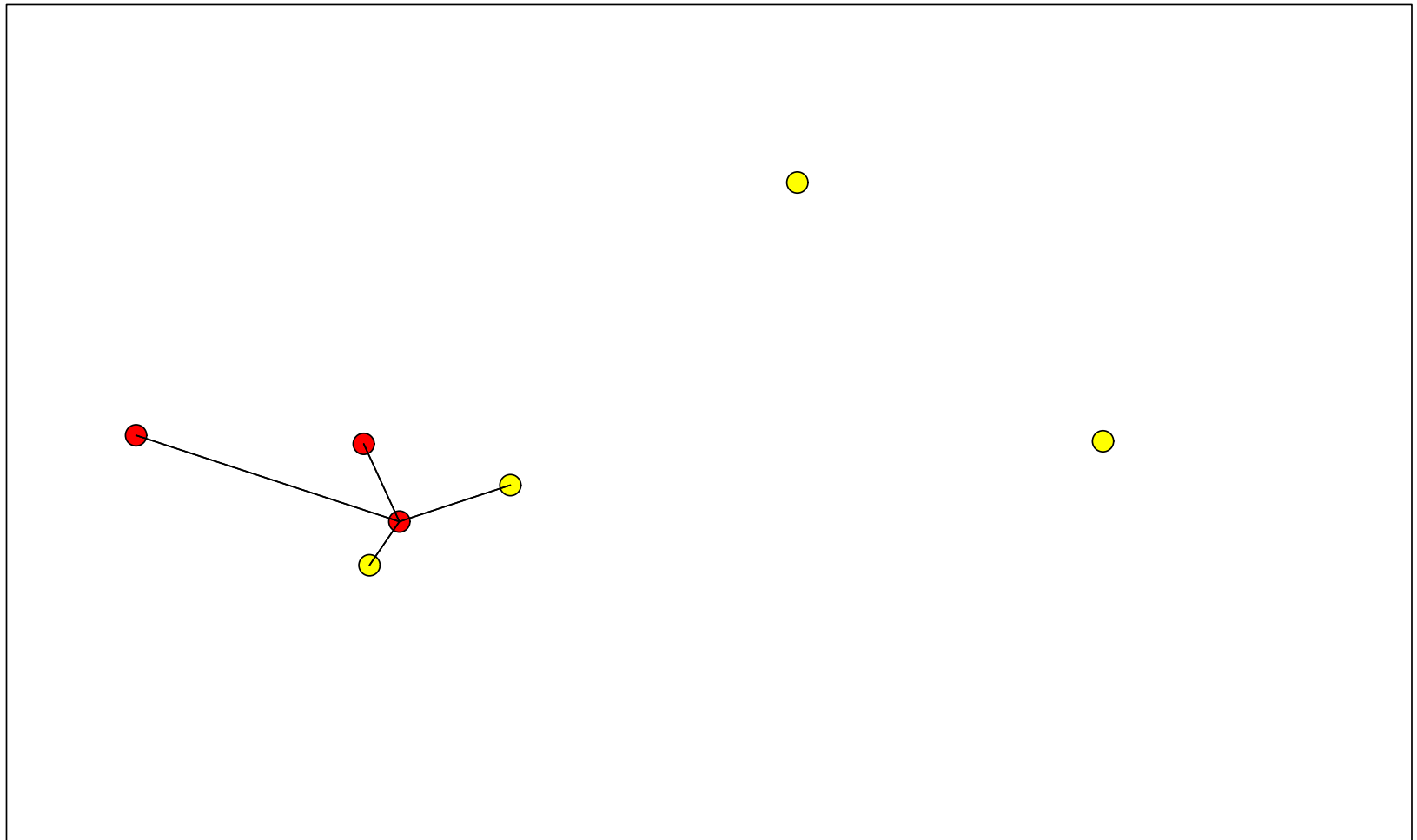
Adaptive web design

weighted links



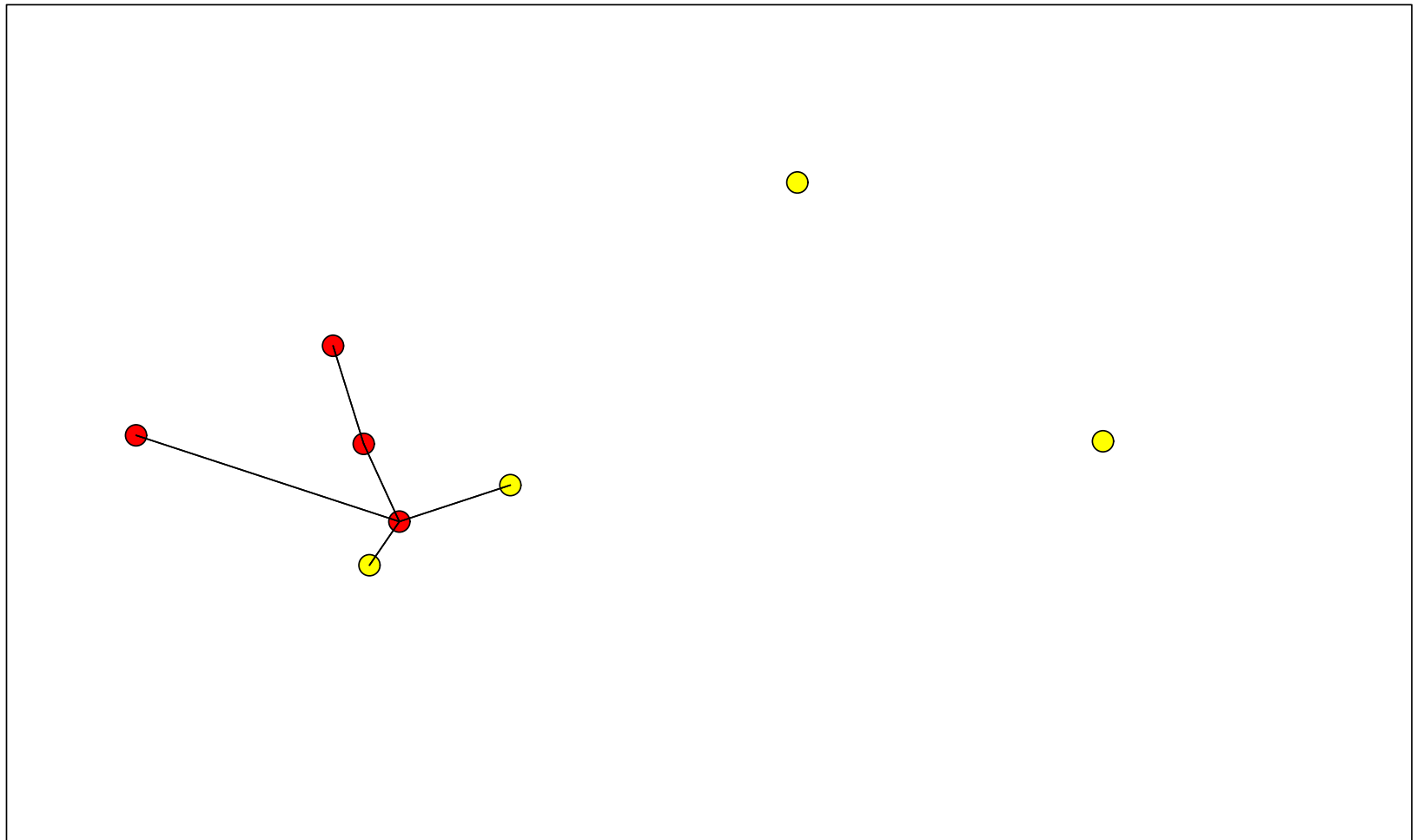
Adaptive web design

weighted links



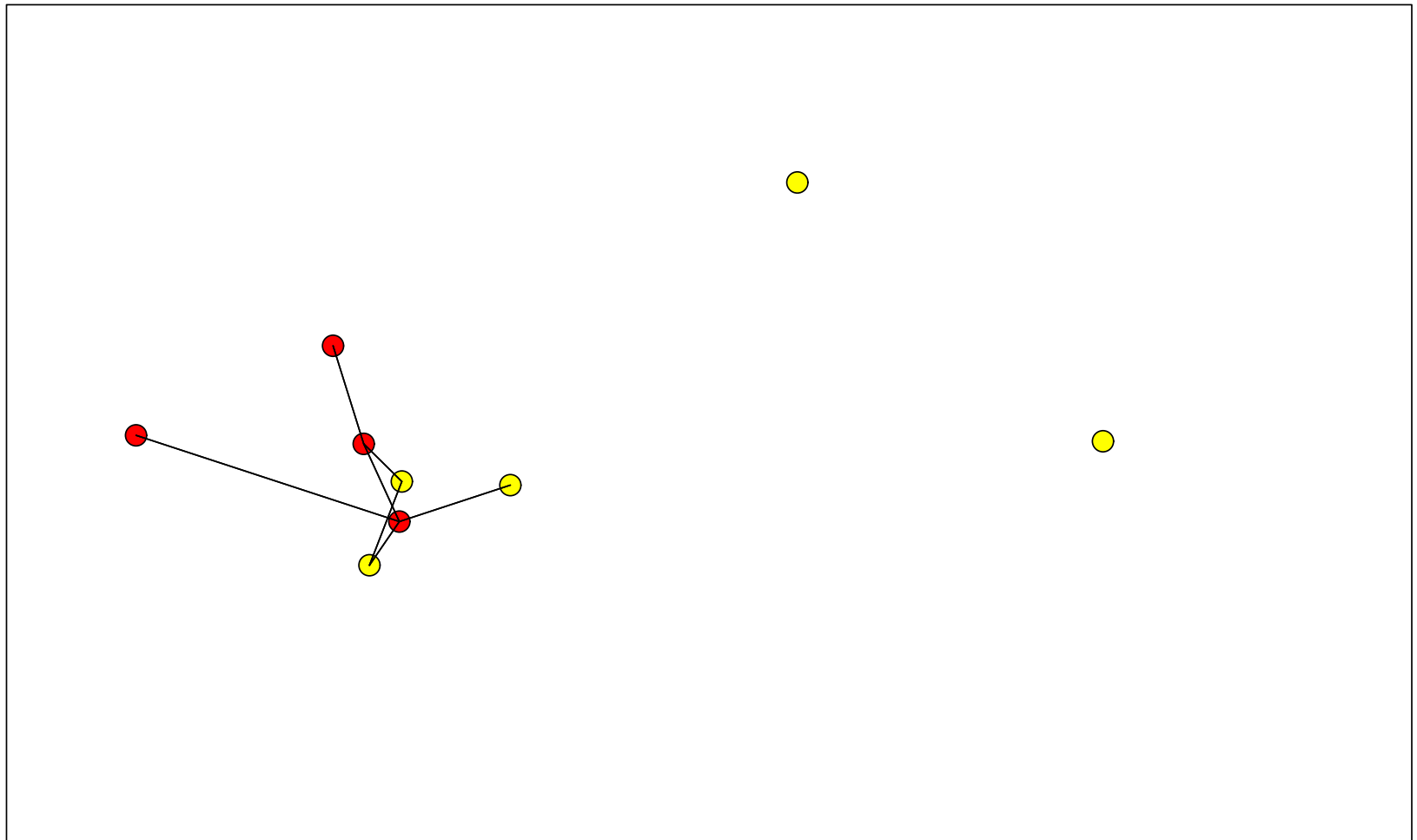
Adaptive web design

weighted links



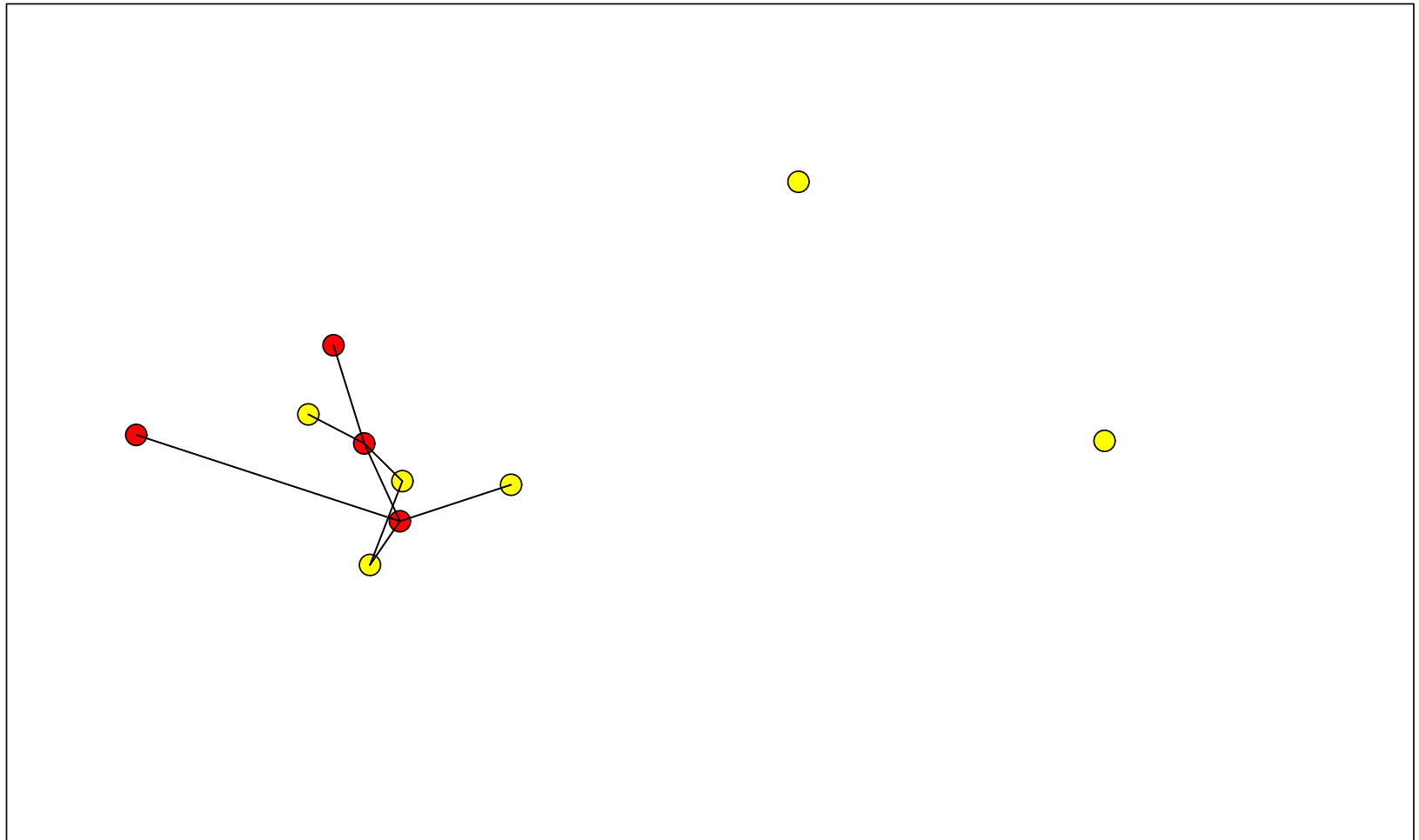
Adaptive web design

weighted links



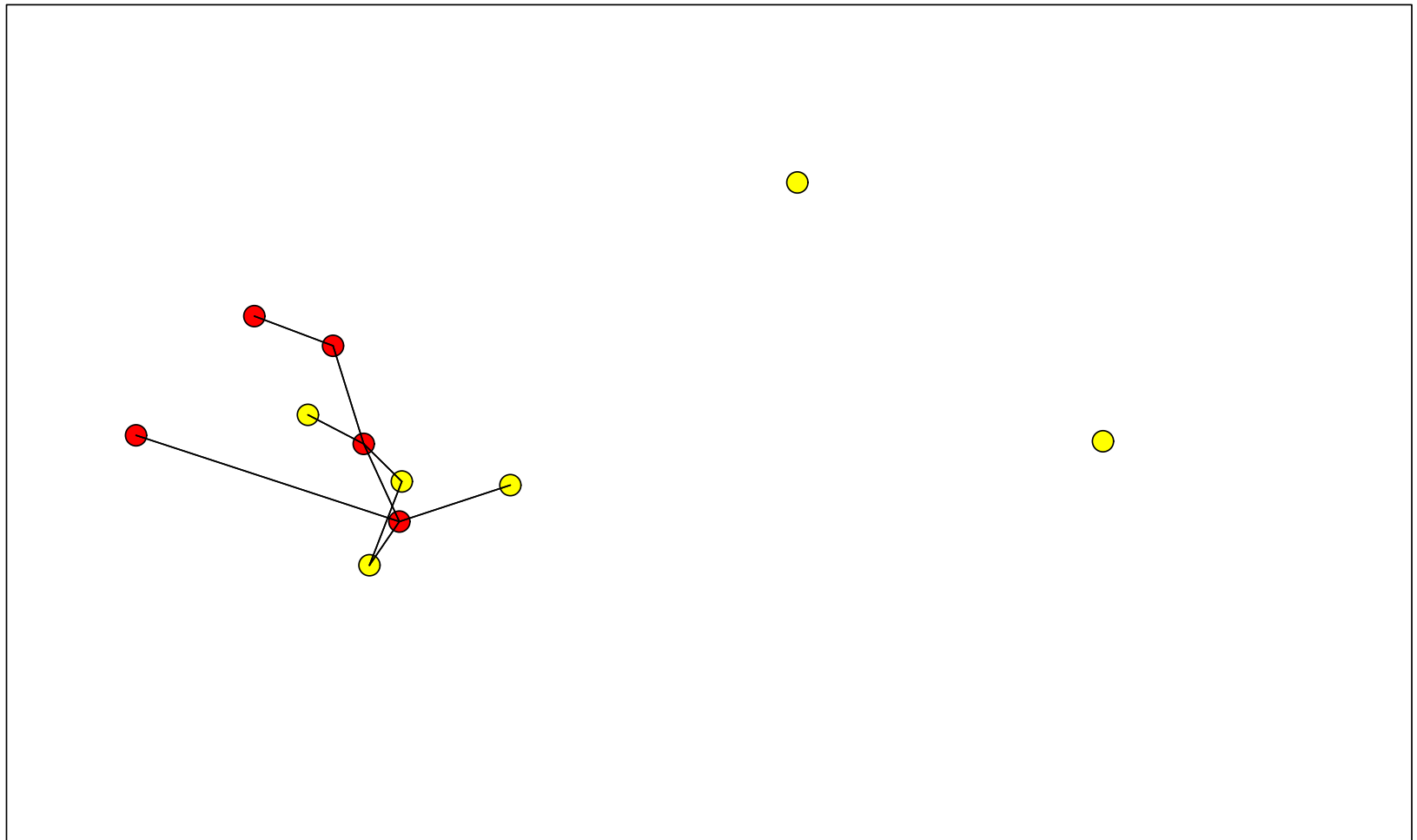
Adaptive web design

weighted links



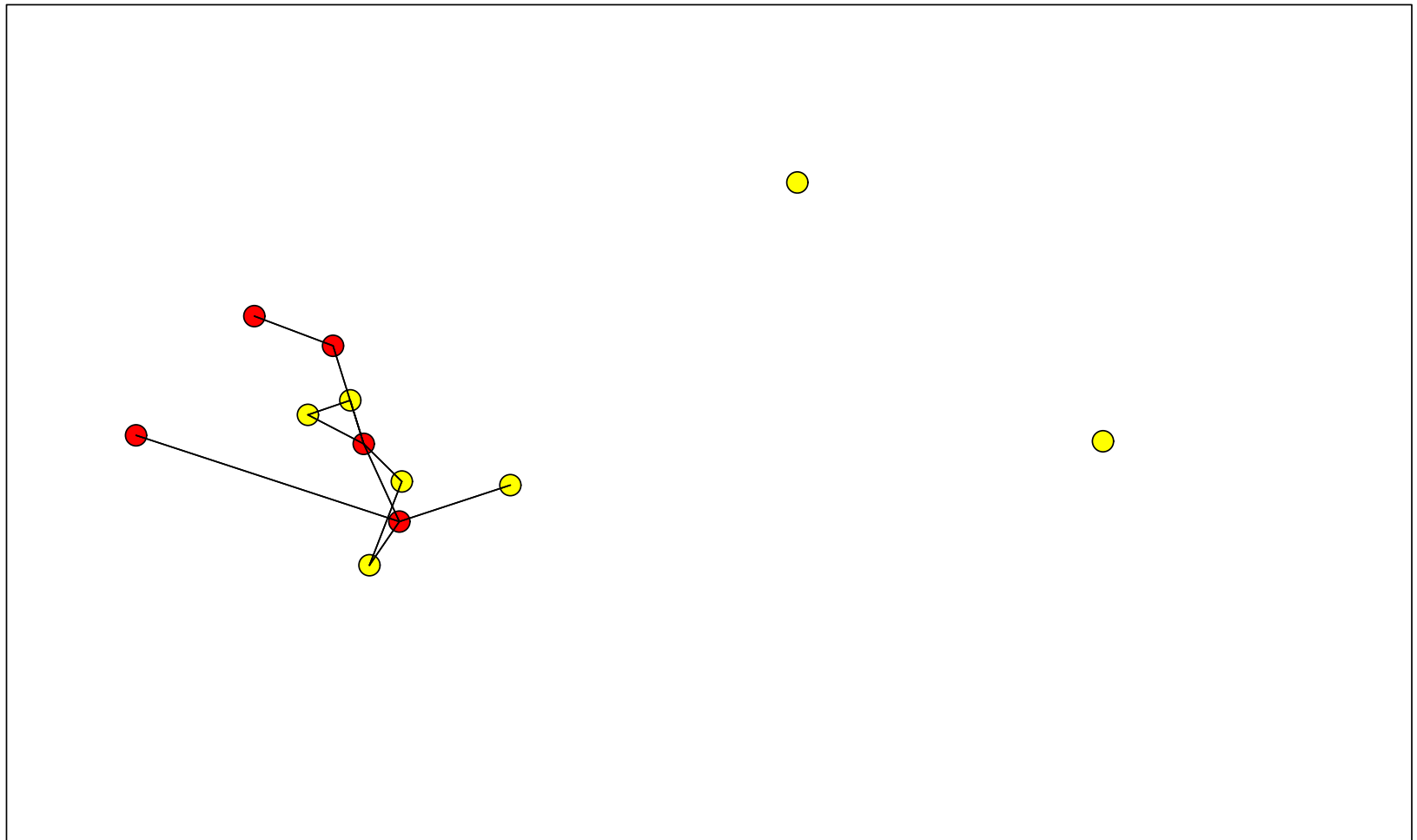
Adaptive web design

weighted links



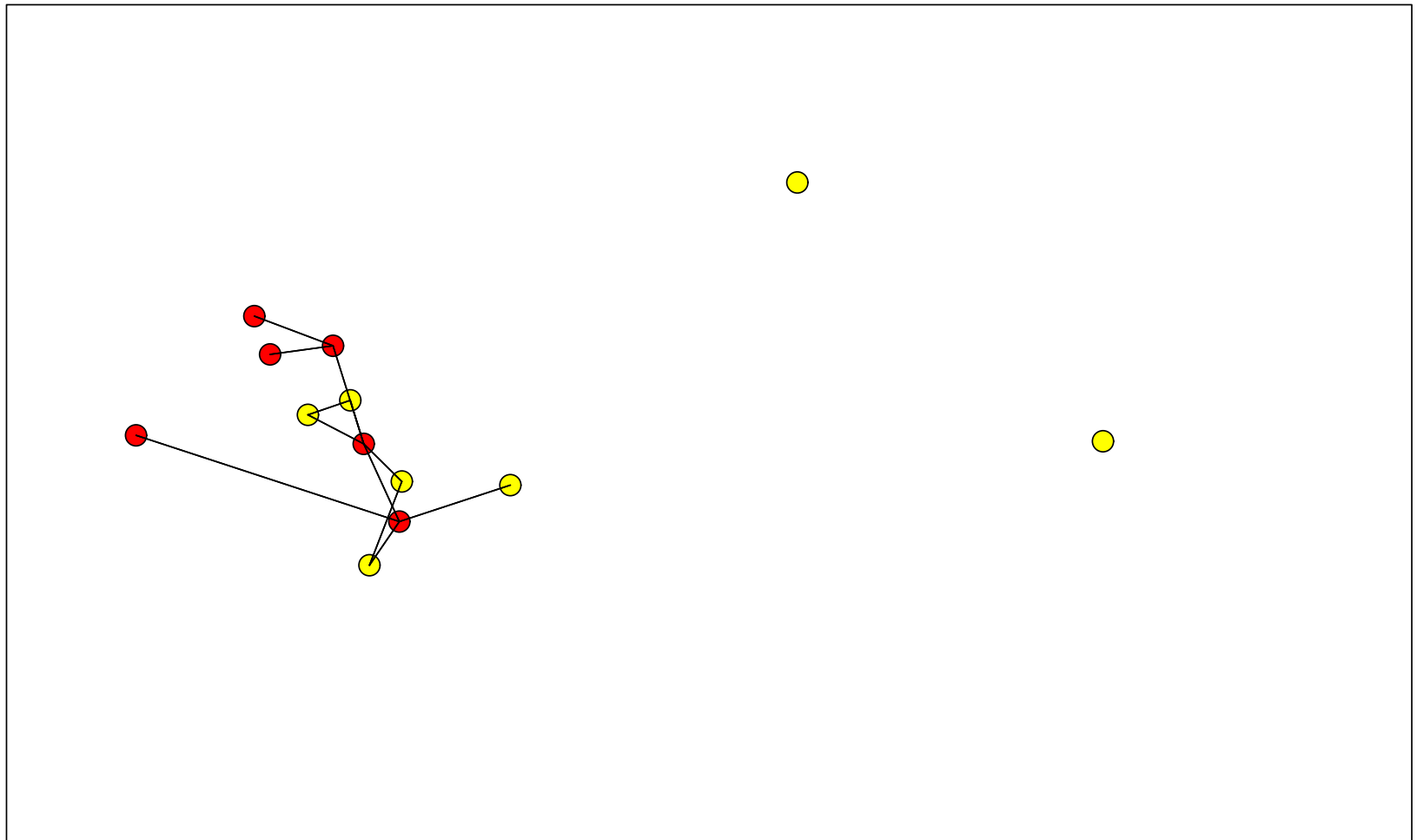
Adaptive web design

weighted links



Adaptive web design

weighted links



Inference

Estimation of a population characteristic such as a population mean, degree distribution, or other quantity, based on the sample data.

- Design-based
 - simple preliminary estimator
 - improve with Rao-Blackwell or resampling

Inference

Estimation of a population characteristic such as a population mean, degree distribution, or other quantity, based on the sample data.

- Design-based
 - simple preliminary estimator
 - improve with Rao-Blackwell or resampling
- Model-based
 - assume stochastic graph model
 - produce realizations from predictive posterior

Design-unbiased estimators

- Start with some preliminary unbiased estimator $\hat{\mu}_0$, such as the **initial sample mean**, an **unequal probability estimator**, or **conditional probability estimator**

Design-unbiased estimators

- Start with some preliminary unbiased estimator $\hat{\mu}_0$, such as the **initial sample mean**, an **unequal probability estimator**, or **conditional probability estimator**
- Improve it using the Rao-Blackwell method:

$$\hat{\mu} = E(\hat{\mu}_0 | d) = \sum_{\text{paths}} \hat{\mu}_0(s) p(s | d)$$

Design-unbiased estimators

- Start with some preliminary unbiased estimator $\hat{\mu}_0$, such as the **initial sample mean**, an **unequal probability estimator**, or **conditional probability estimator**
- Improve it using the Rao-Blackwell method:

$$\hat{\mu} = E(\hat{\mu}_0 | d) = \sum_{\text{paths}} \hat{\mu}_0(s) p(s | d)$$

d is the minimal sufficient statistic

Model based inference

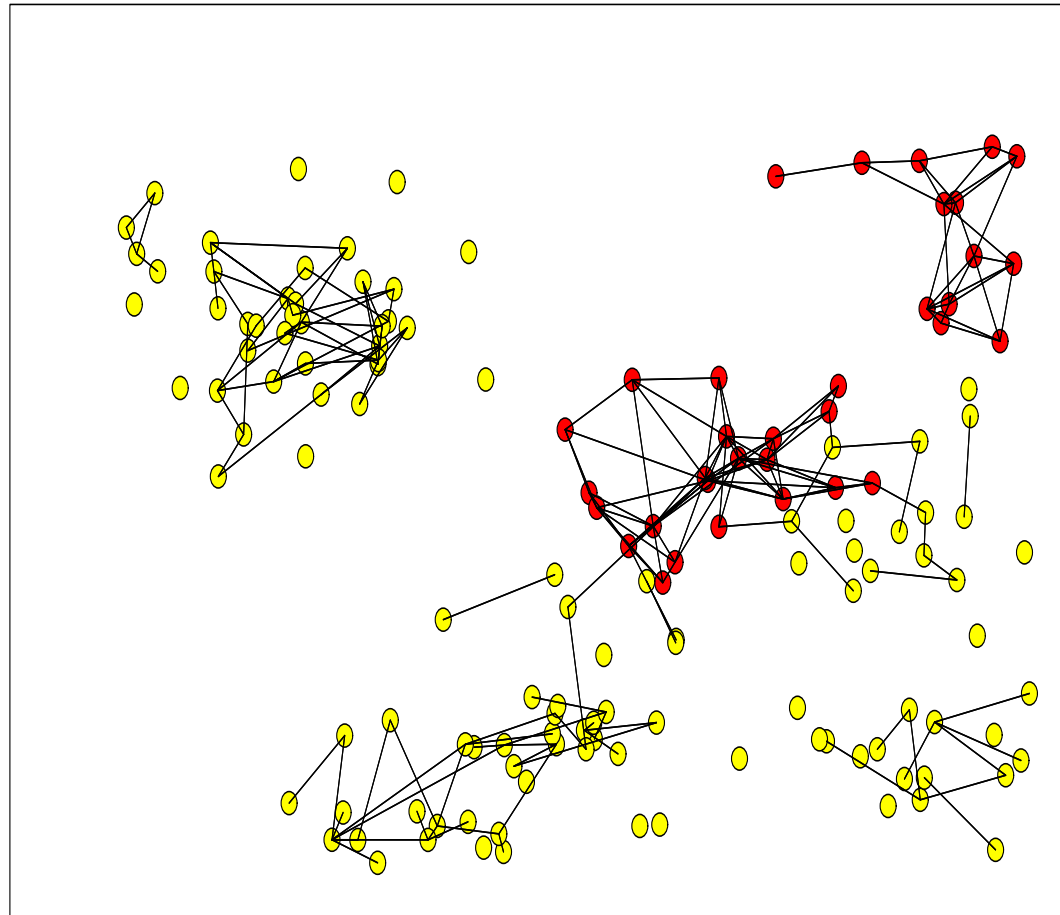
Bayes approach: The object is to produce realizations $(\mathbf{y}_s, \mathbf{y}_{\bar{s}}, \mathbf{w}_s, \mathbf{w}_{\bar{s}})$ of the entire graph from its posterior distribution given the sample data.

1. Using current values of θ and β , select a realization of $(\mathbf{y}_{\bar{s}}, \mathbf{w}_{\bar{s}})$ from $P(\mathbf{y}_{\bar{s}}, \mathbf{w}_{\bar{s}} \mid d)$.
2. Using the values $(\mathbf{y}_{\bar{s}}, \mathbf{w}_{\bar{s}})$ obtained in step (1) to augment the data values $(\mathbf{y}_s, \mathbf{w}_s)$, select new parameter values (θ, β) from the posterior distribution of the parameters given the whole graph realization $\pi(\theta_0, \beta_0, \beta_1, \beta_2 \mid \mathbf{y}_s, \mathbf{y}_{\bar{s}}, \mathbf{w}_s, \mathbf{w}_{\bar{s}})$

Repeat.

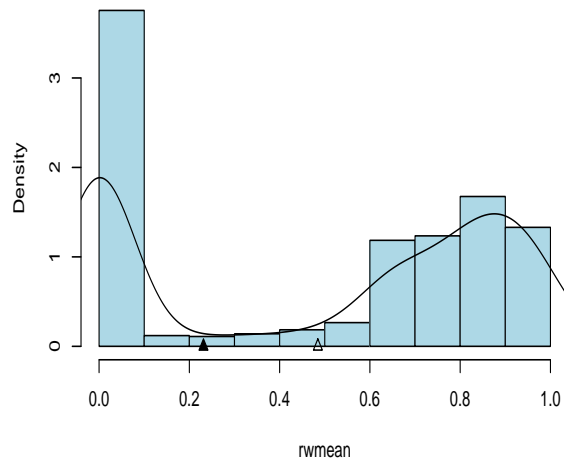
Design and estimation comparisons

population graph

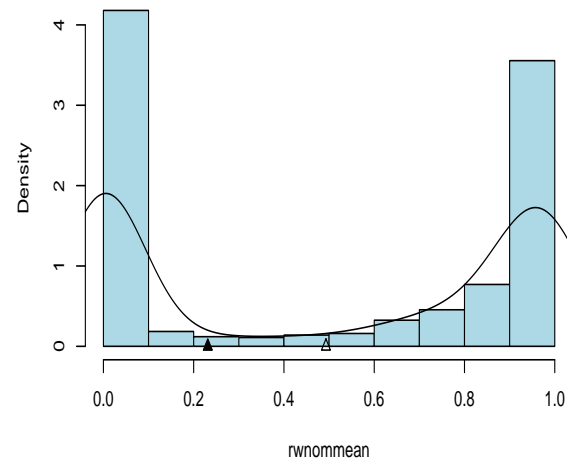


Random walk n=20, initial pp-degree

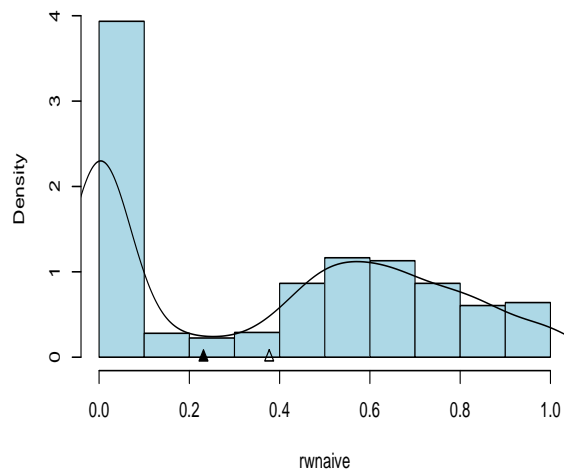
sample mean, random walk



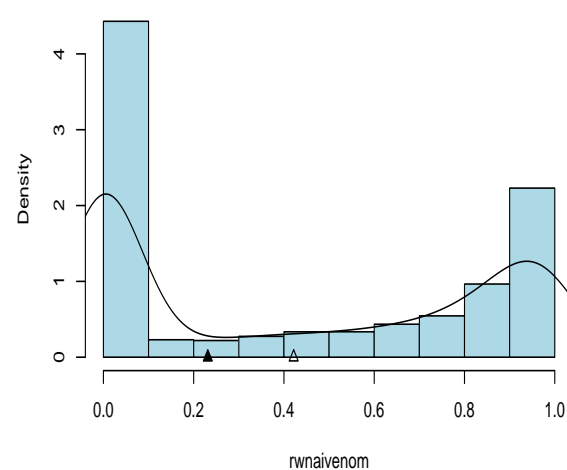
mean of draws, random walk



gen ratio est, random walk

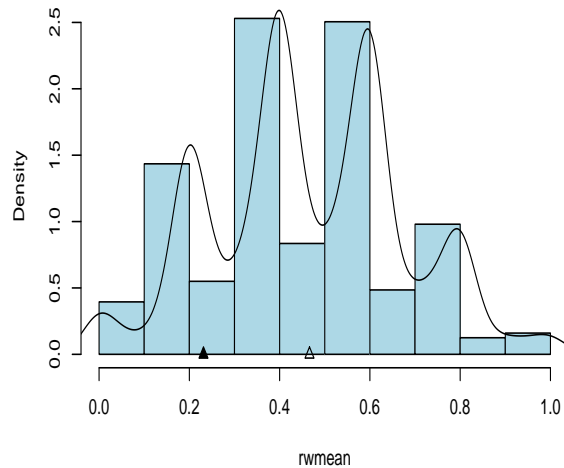


gen ratio of draws, random walk

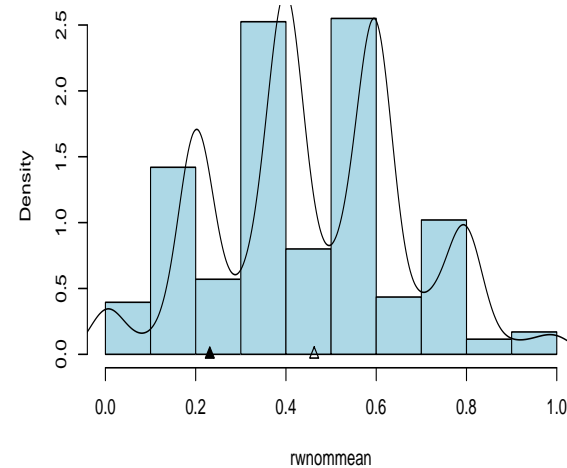


5 random walks, n=4 each, pp-deg starts

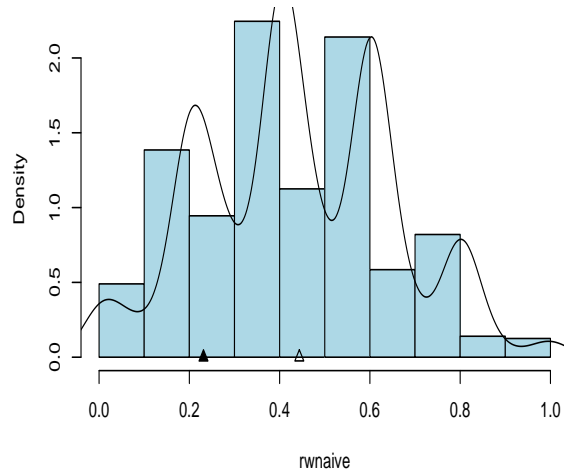
sample mean, random walk



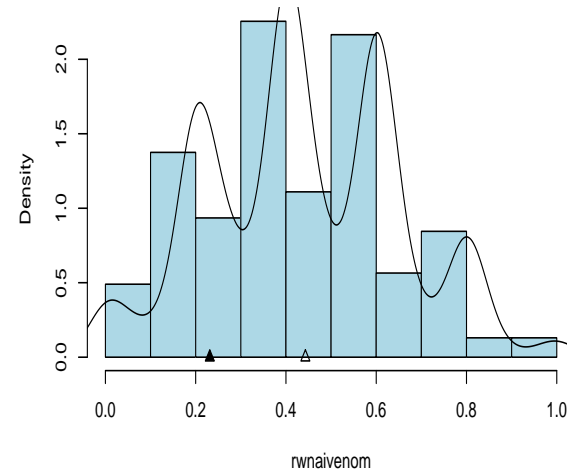
mean of draws, random walk



gen ratio est, random walk

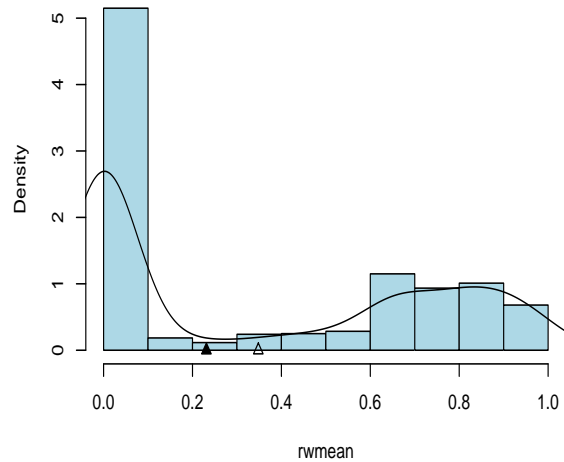


gen ratio of draws, random walk

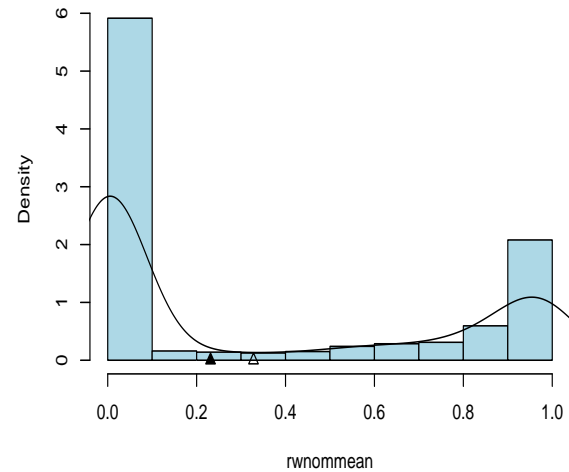


random walk, $n=20$, equal probability start

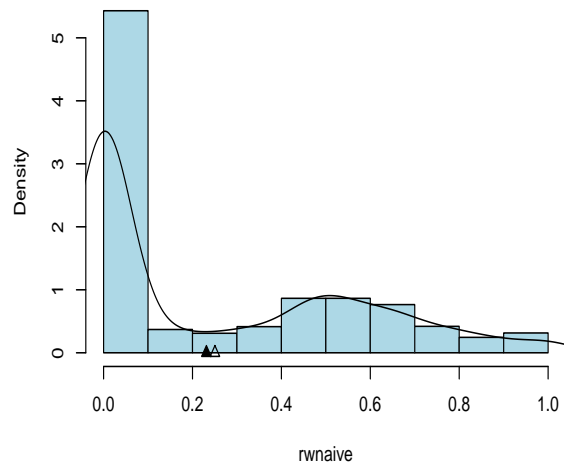
sample mean, random walk



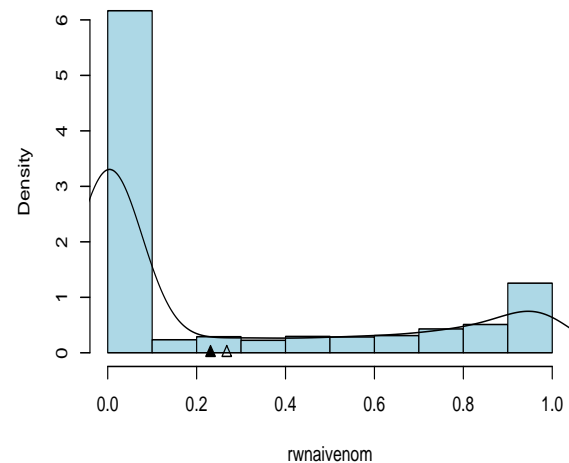
mean of draws, random walk



gen ratio est, random walk

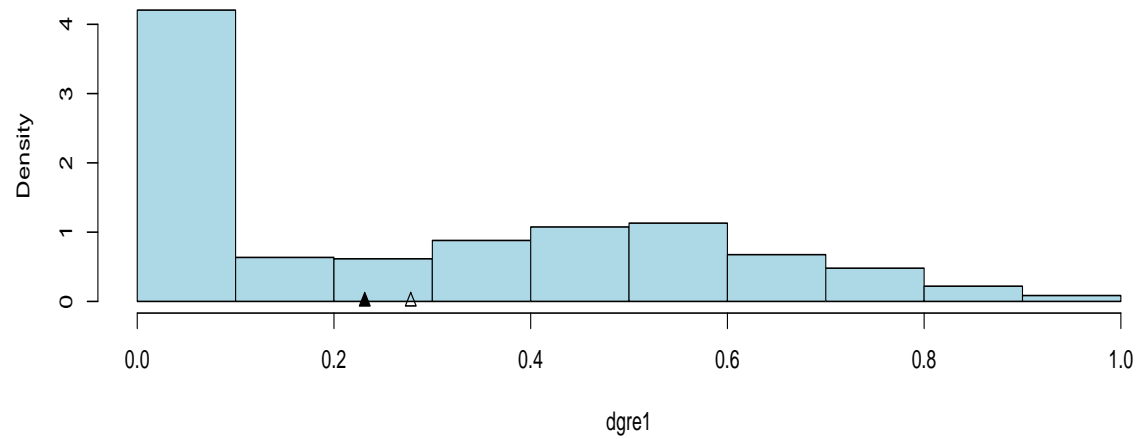


gen ratio of draws, random walk

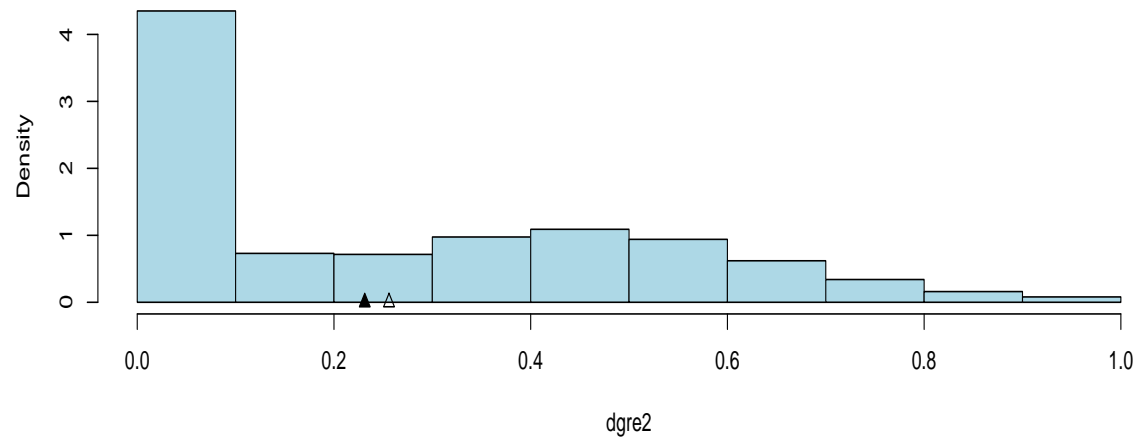


AWS, $n_0=1$, $n=20$, random links, jump=.1

generalized ratio estimate 1

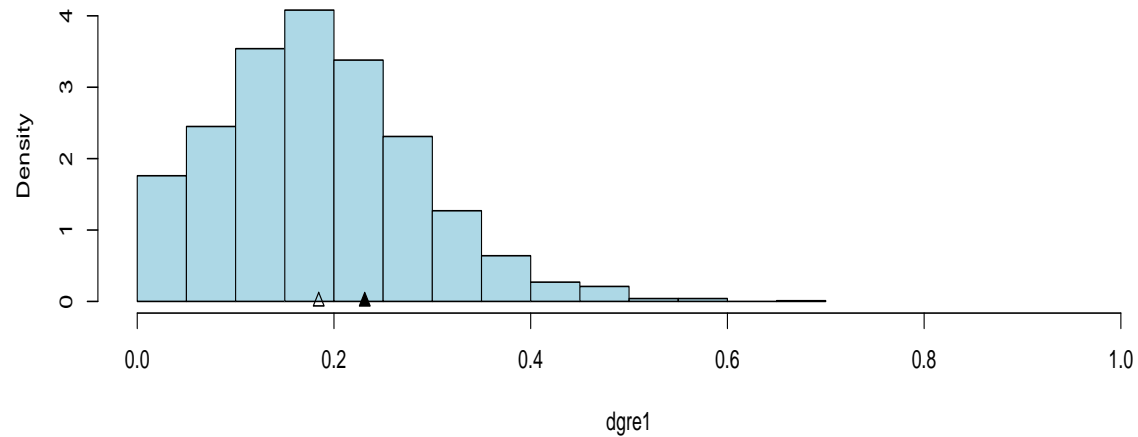


generalized ratio estimate 2

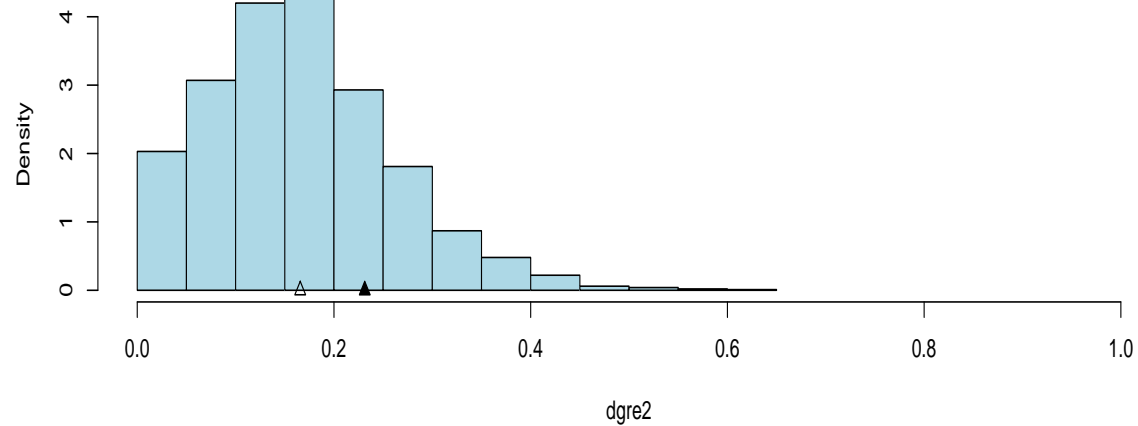


AWS, $n_0=10$, $n=20$, random links, jump=.1

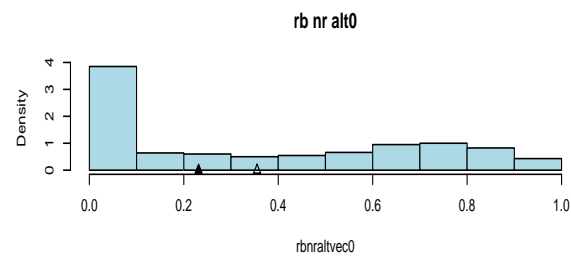
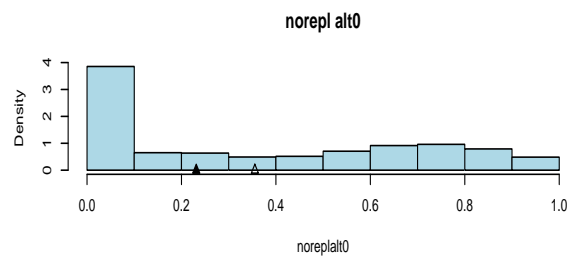
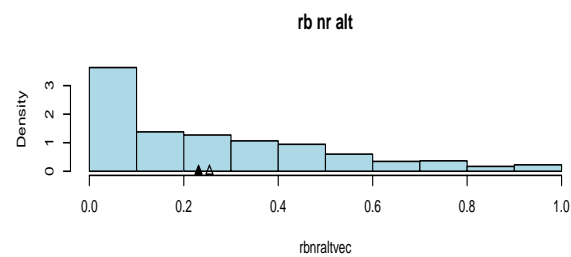
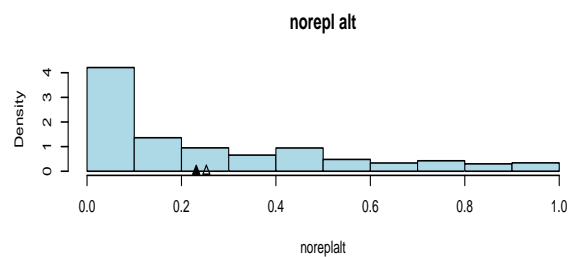
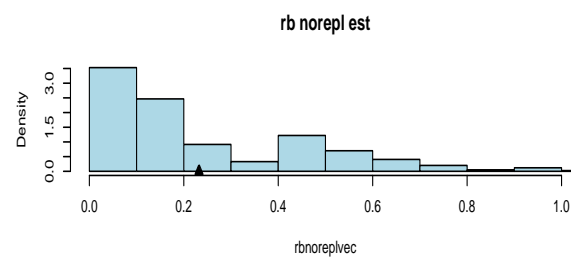
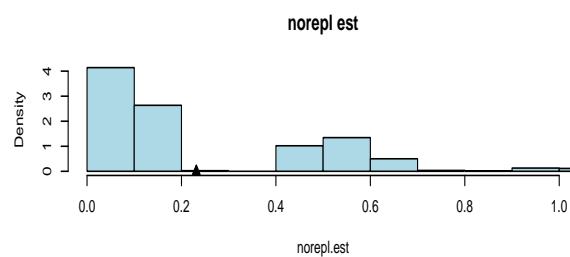
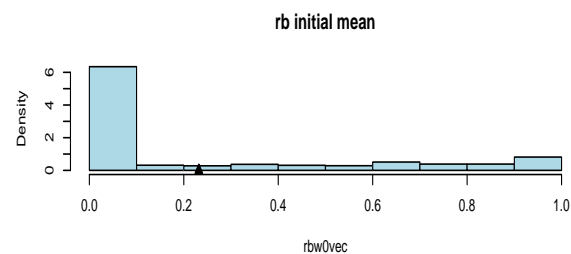
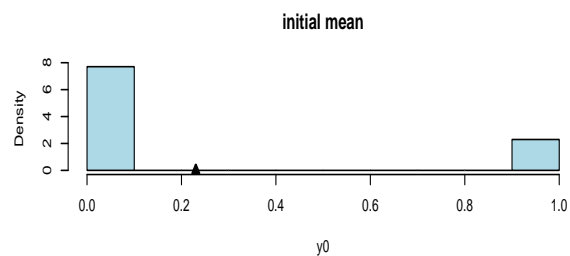
generalized ratio estimate 1



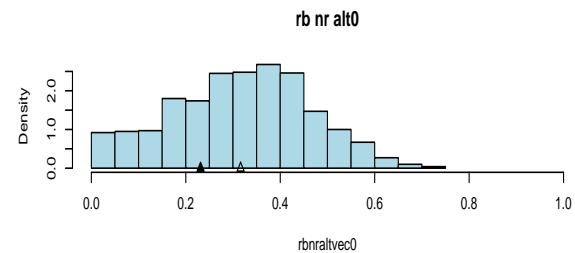
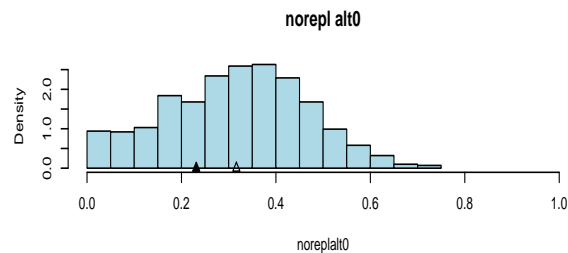
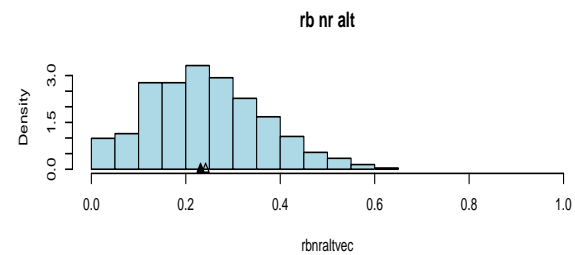
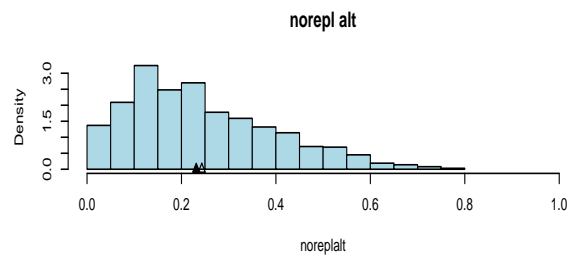
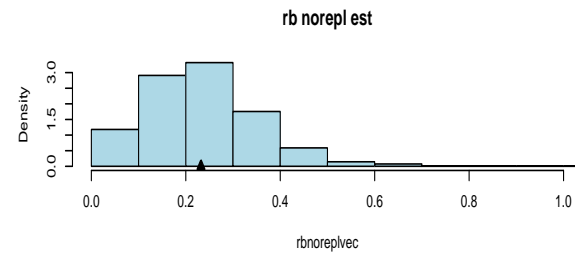
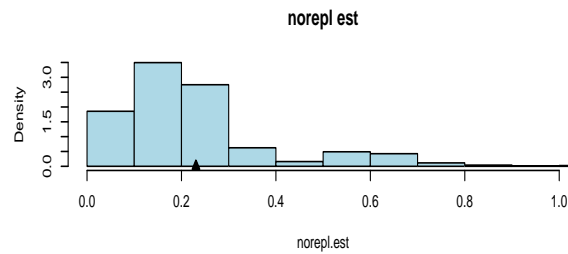
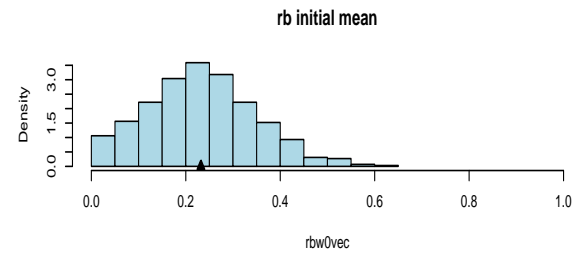
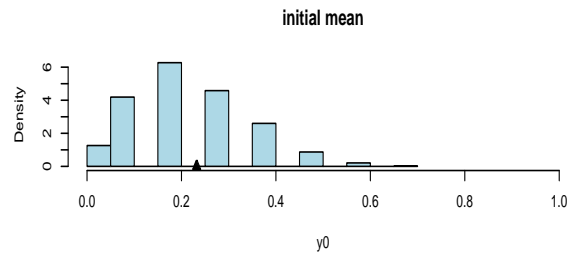
generalized ratio estimate 2



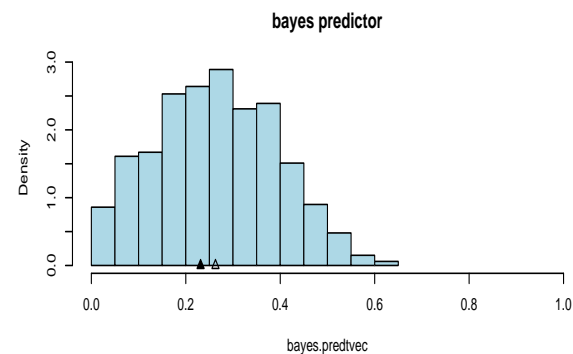
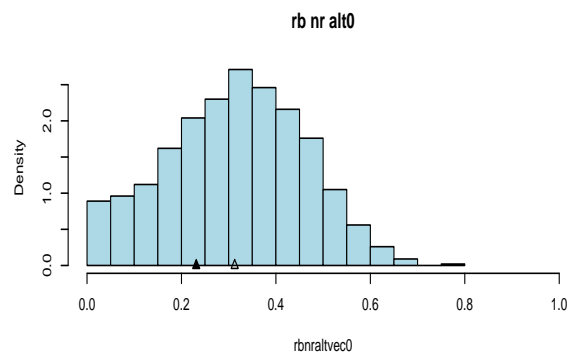
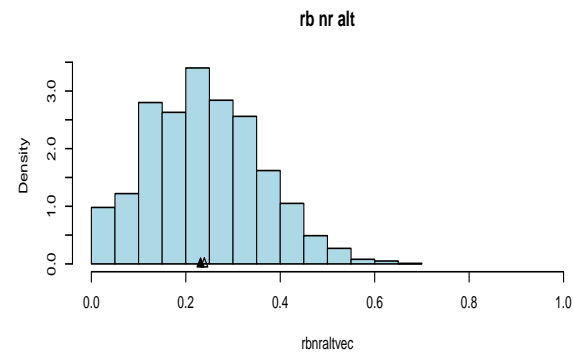
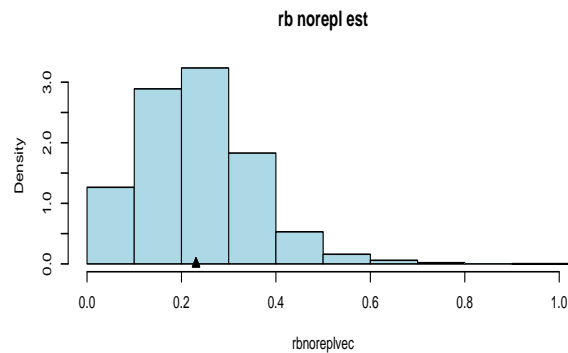
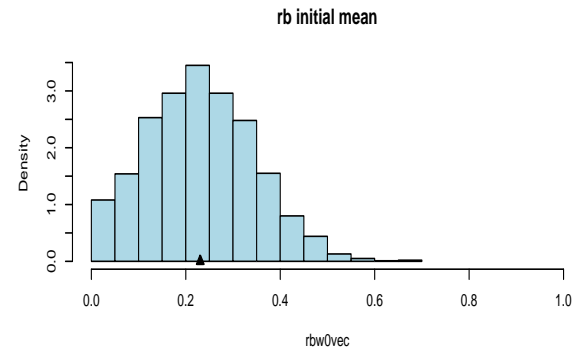
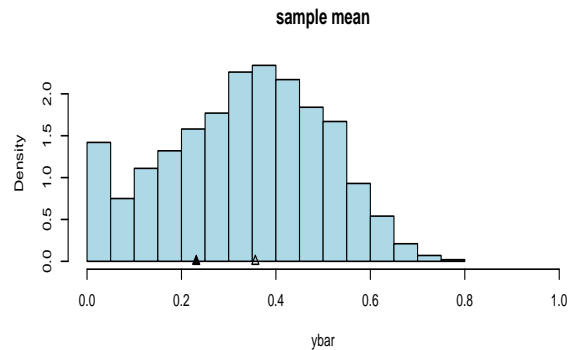
Design based estimators, AWS, n0=1



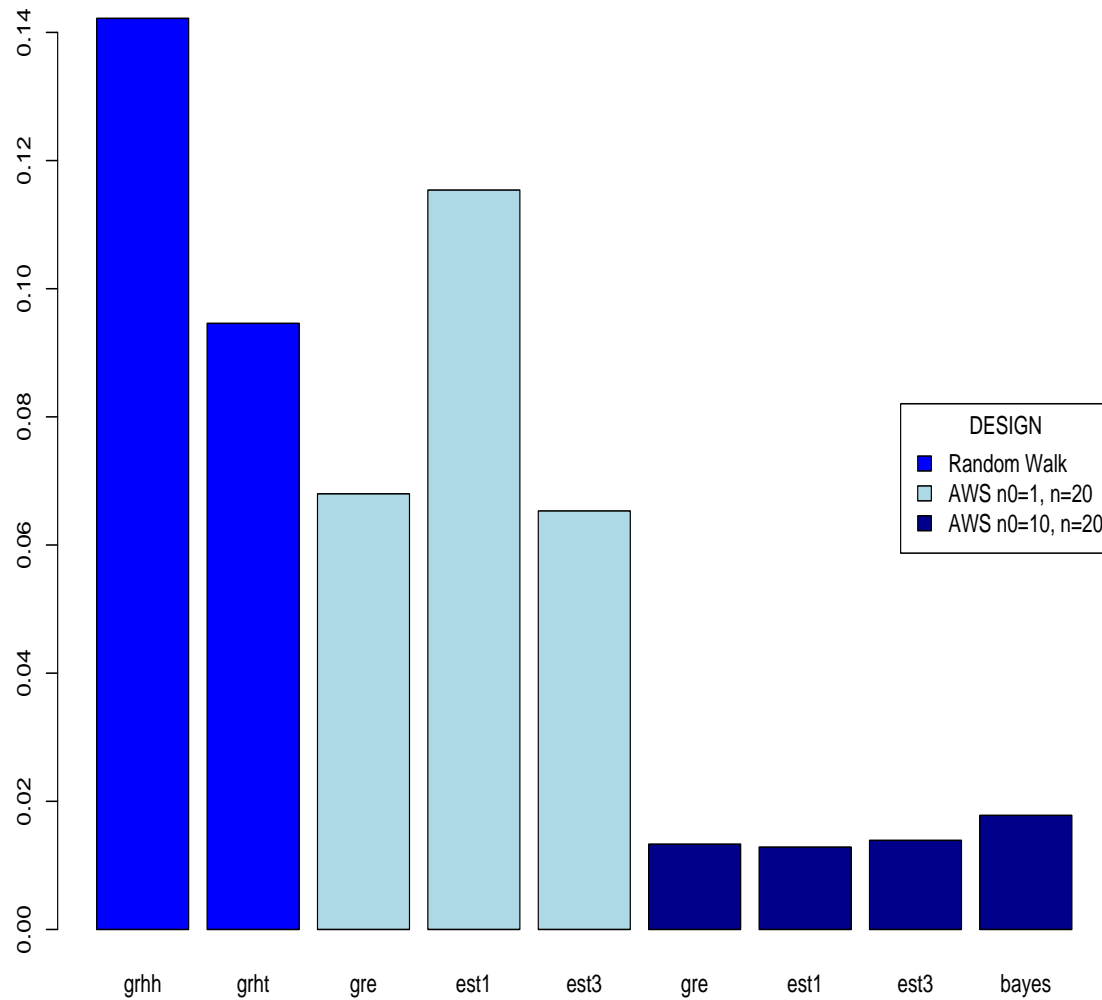
Design based estimators, AWS, n0=10



Design and model based estimators, AWS $n_0=10$, $n=20$



Designs and Estimators

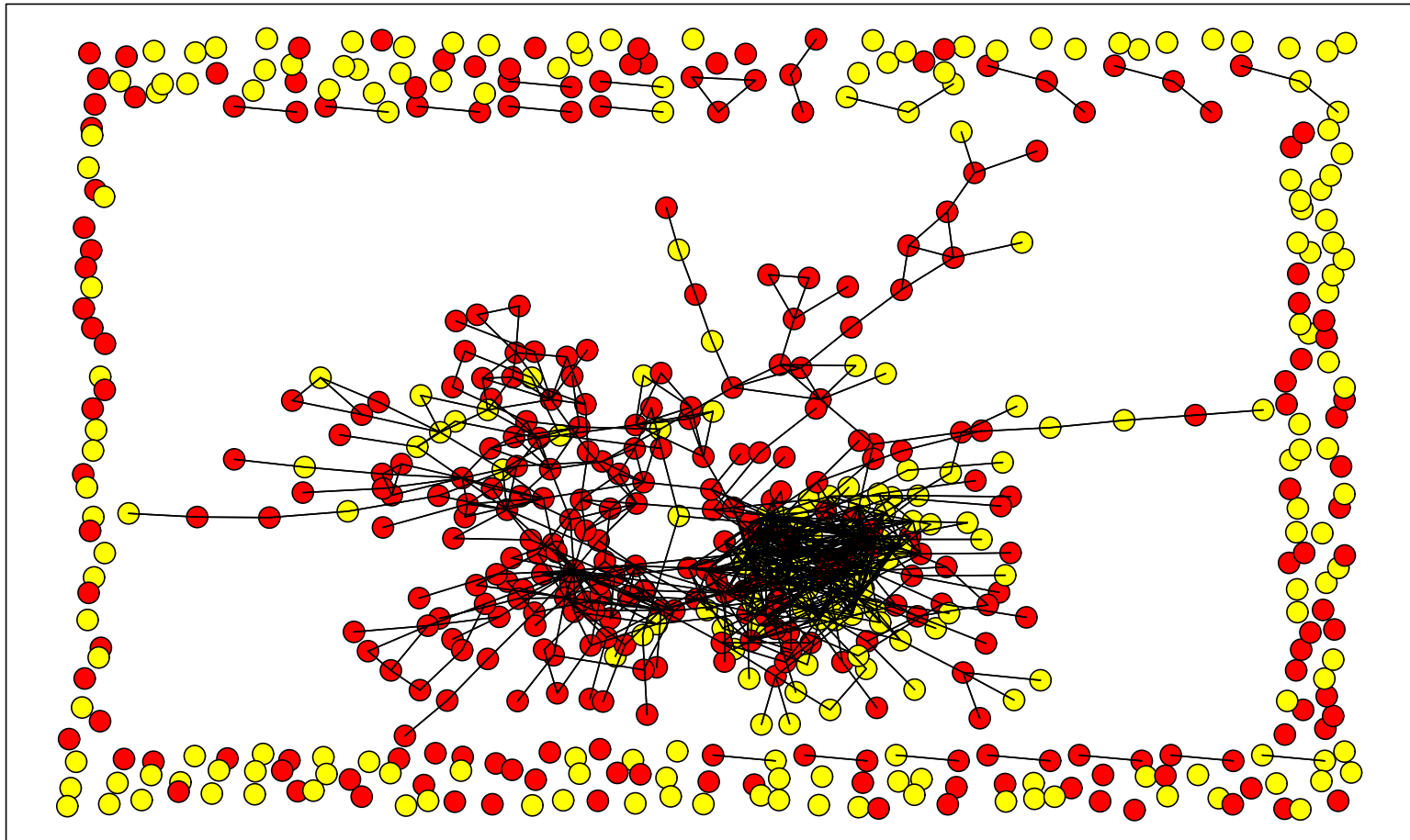


Empirical Example

HIV/AIDS at-risk hidden population: Colorado Springs
Study on the heterosexual transmission of HIV/AIDS
(Potterat et al. 1993, Rothenberg et al. 1995, Darrow et al.
1999)

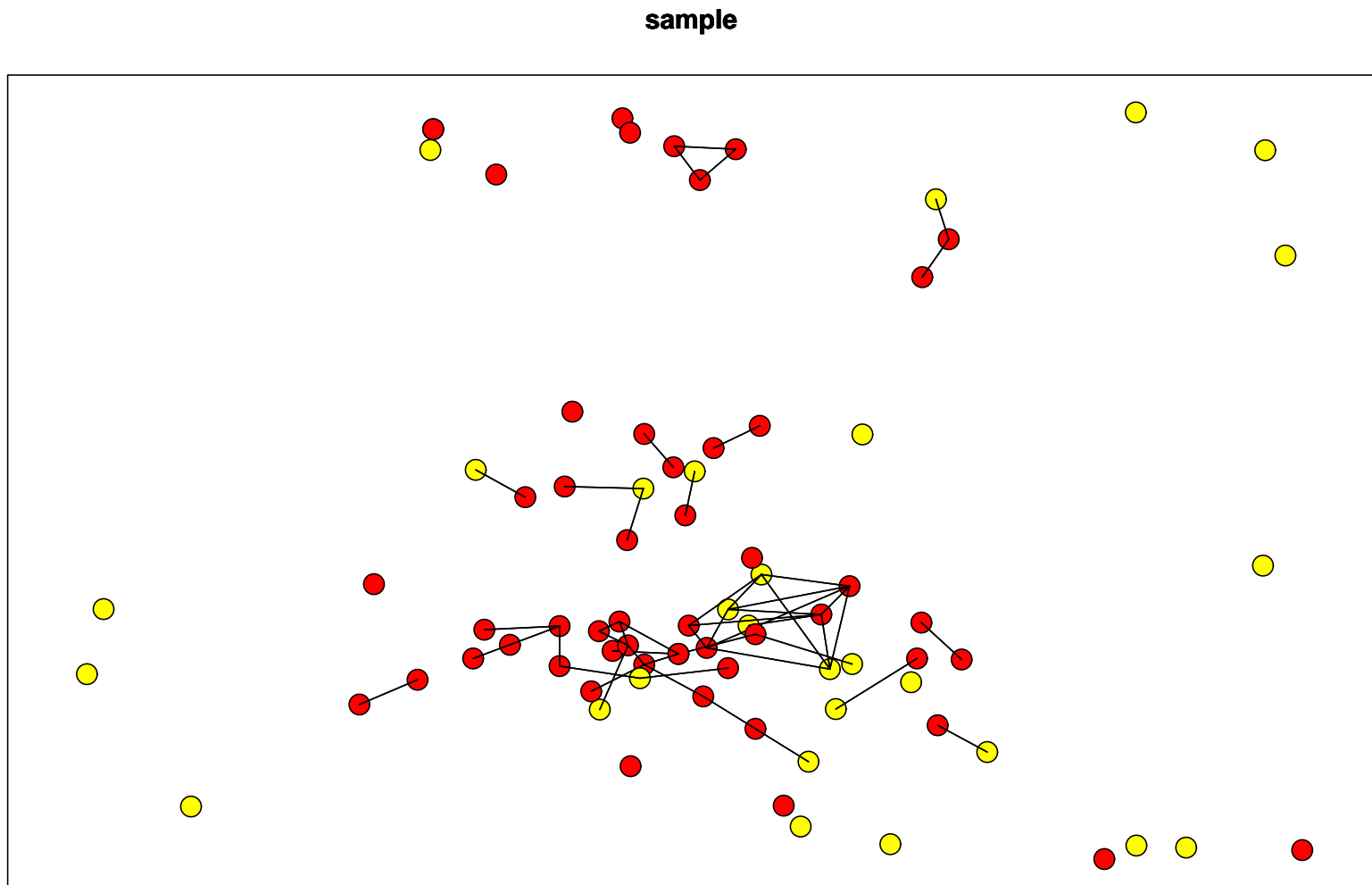
Colorado springs study population

population graph

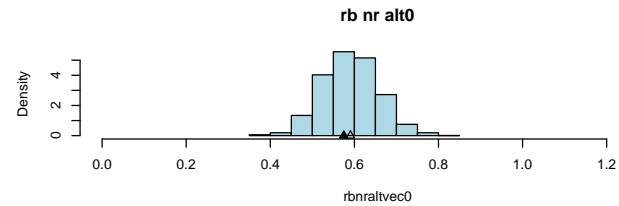
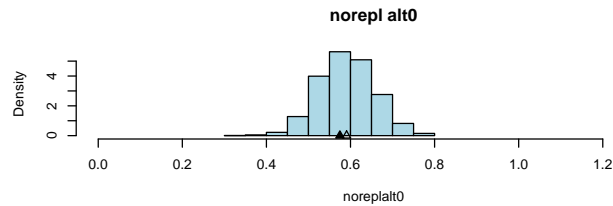
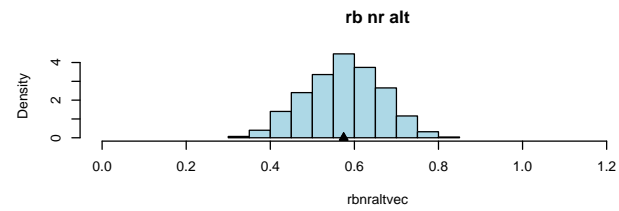
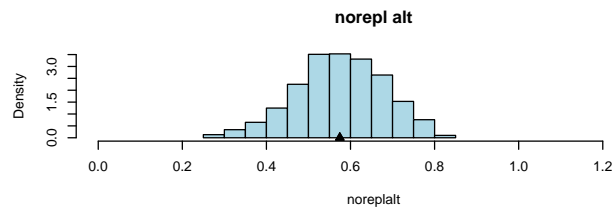
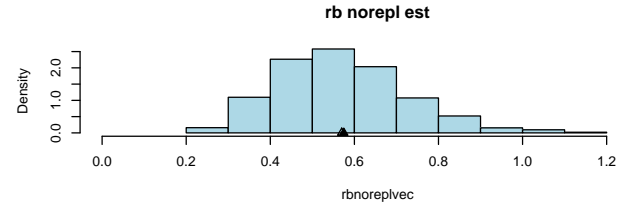
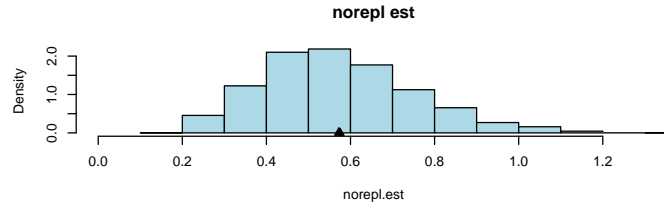
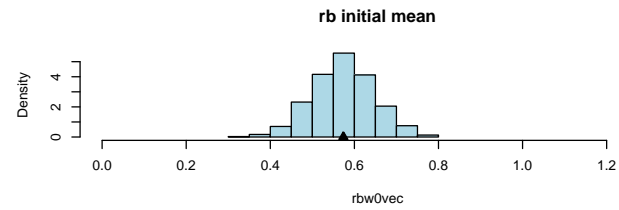
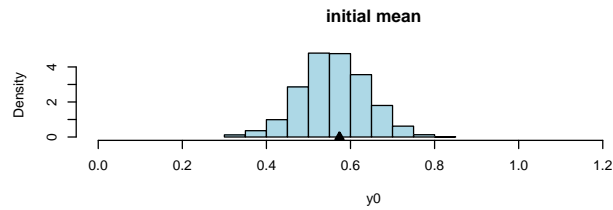


Sample of 80 individuals

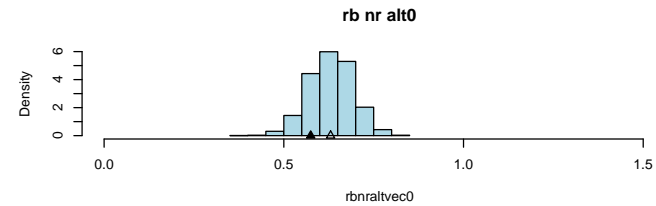
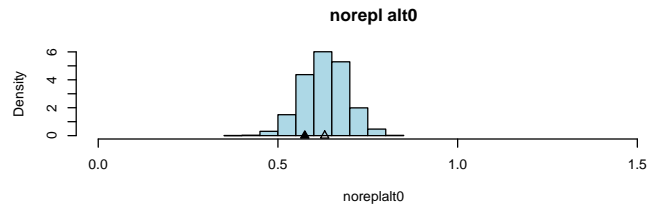
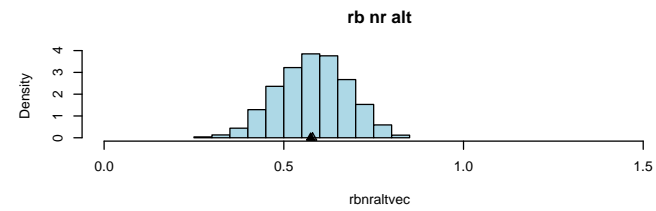
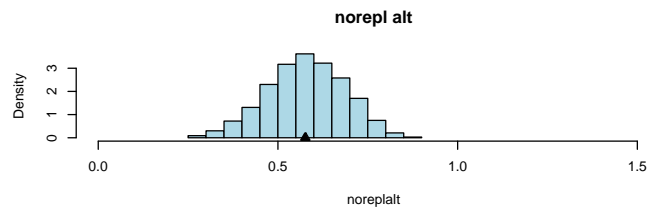
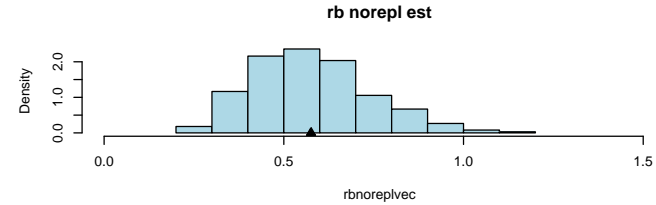
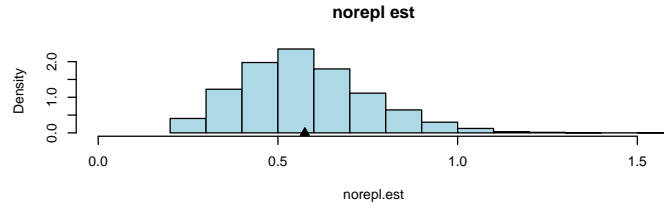
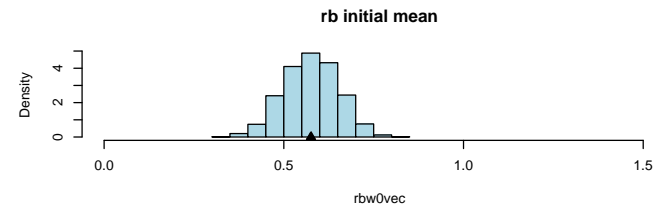
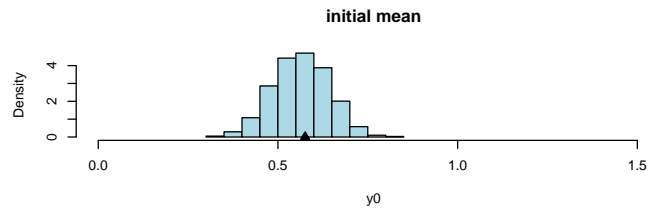
Initial $n_0 = 10$, final $n = 20$, $m = 4$ independent selections.



Estimating idu use, random links design

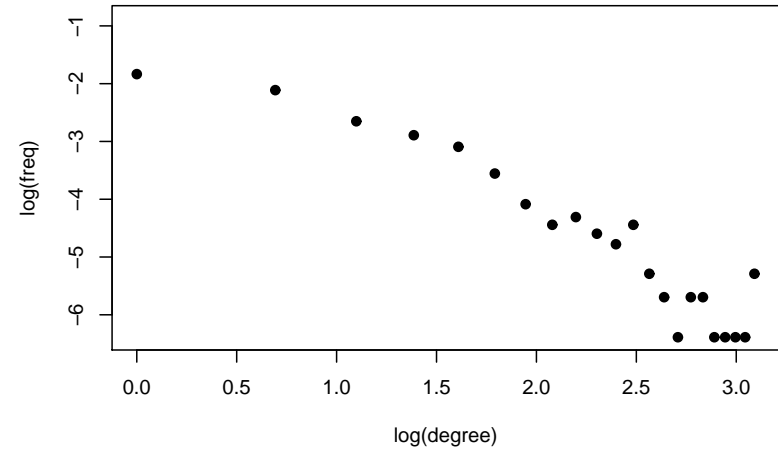
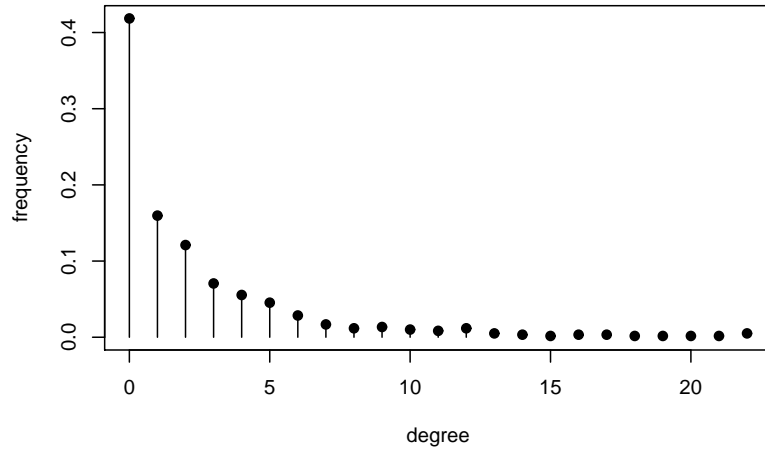


Estimating idu use, weighted links design

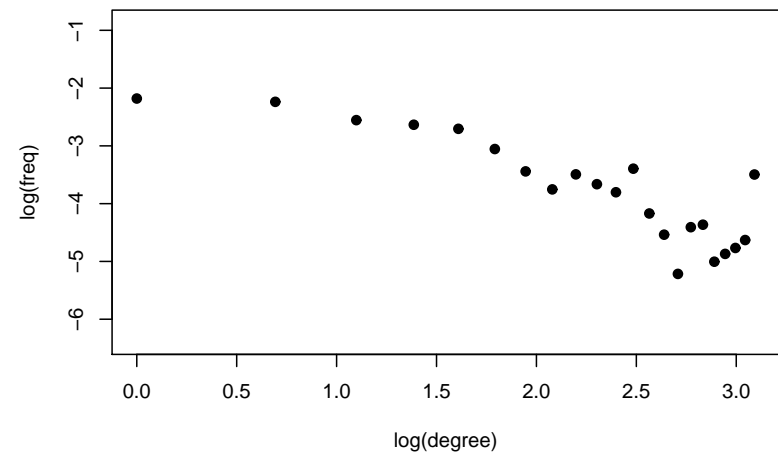
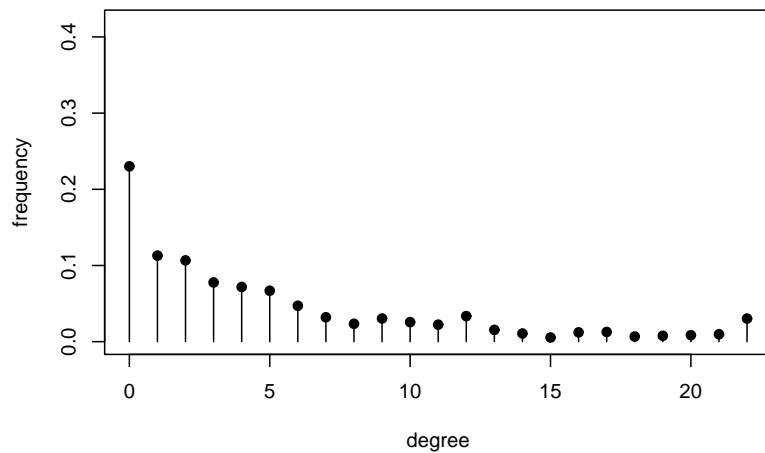


Degree distribution, HIV/AIDS study

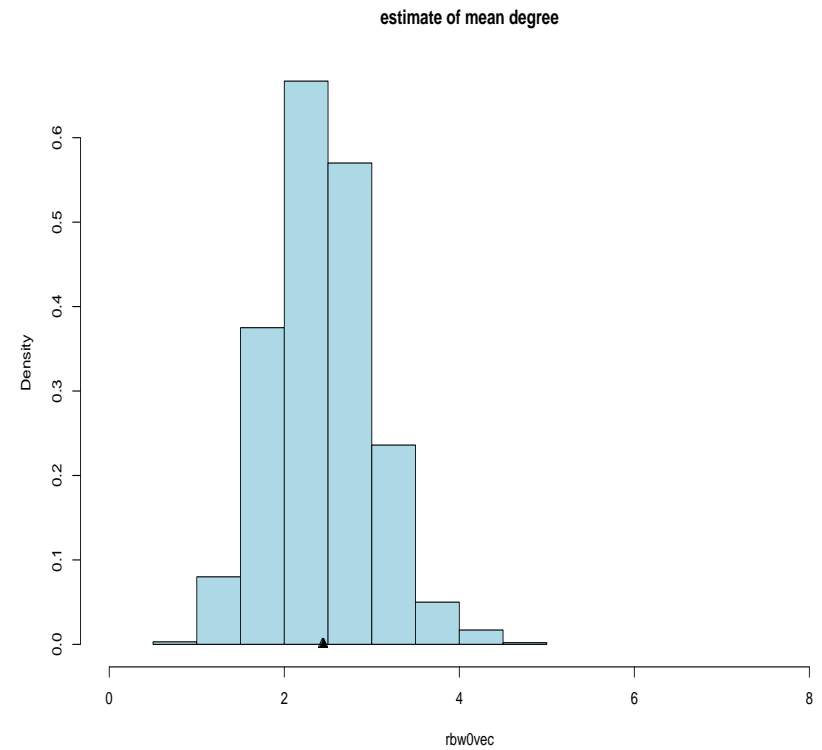
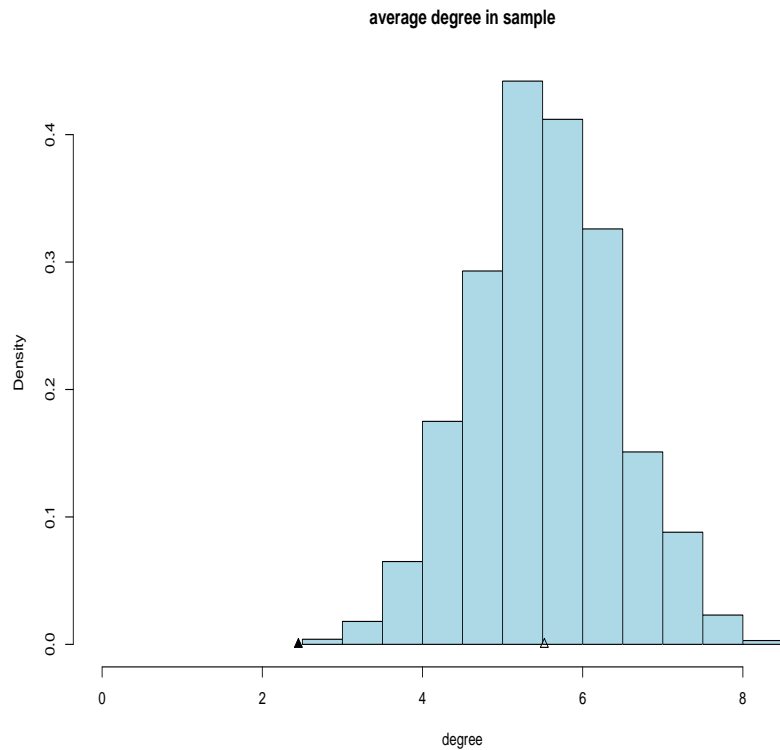
degree distribution



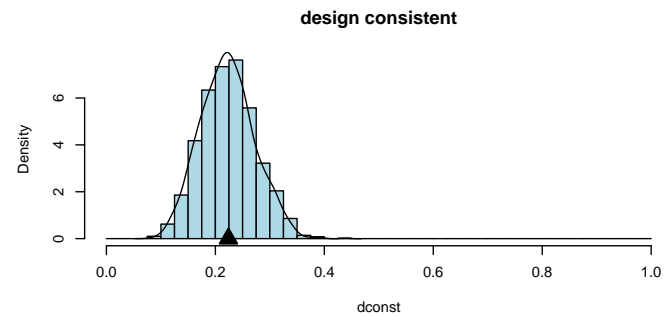
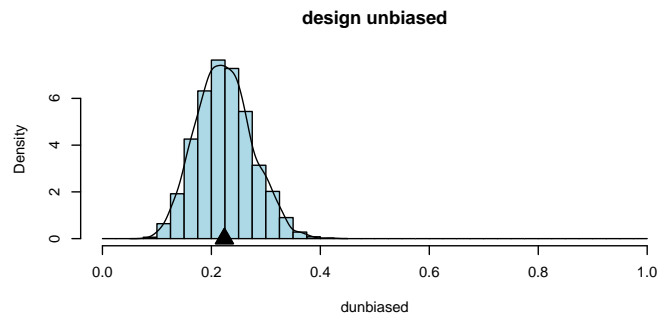
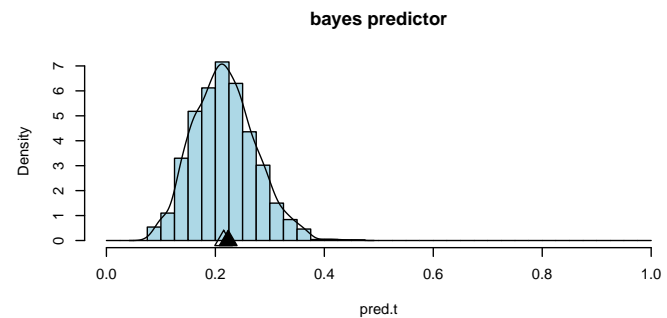
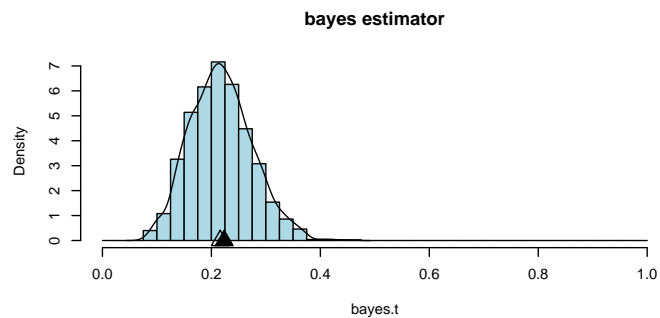
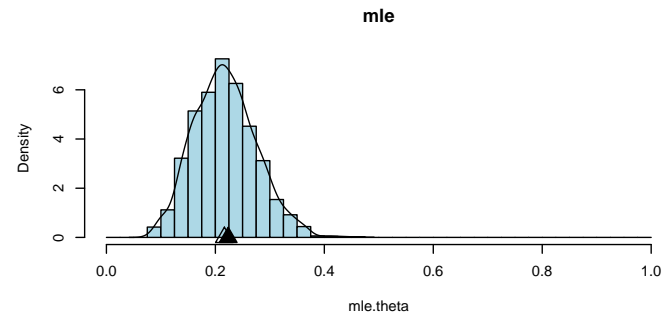
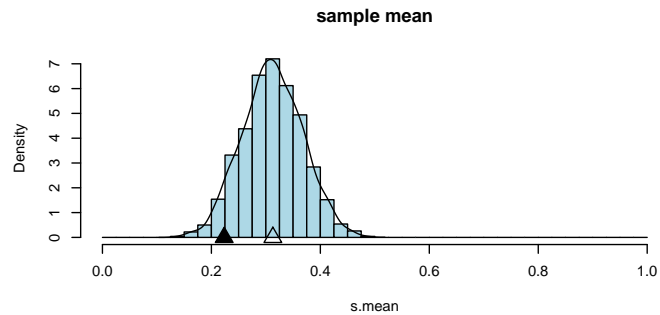
sample degree distribution



Estimating mean degree



Design and Model based inferences



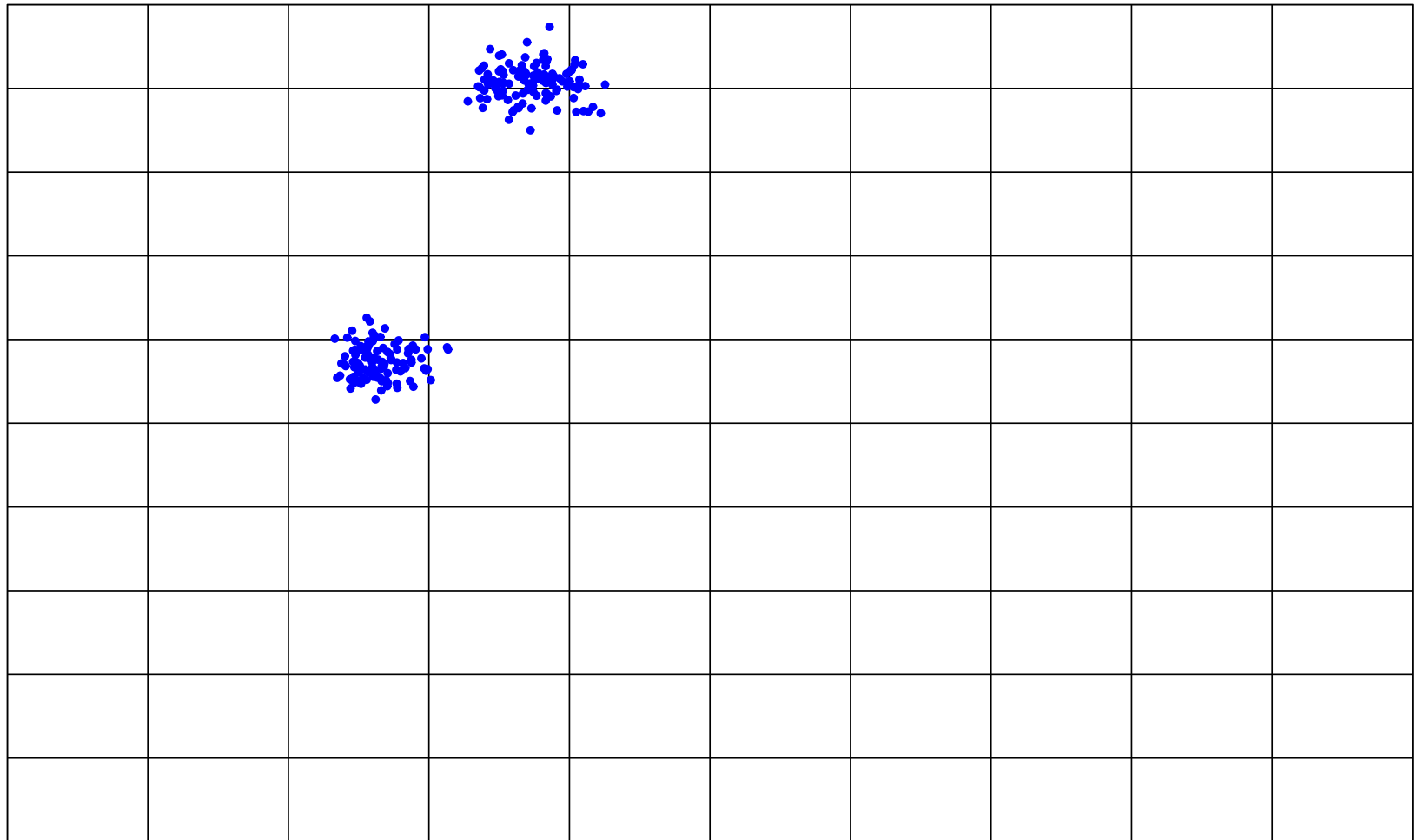
(Sex worker data)

6. Network methods in spatial sampling

1. Relation of spatial and network sampling
2. Adaptive web sampling in a spatial setting

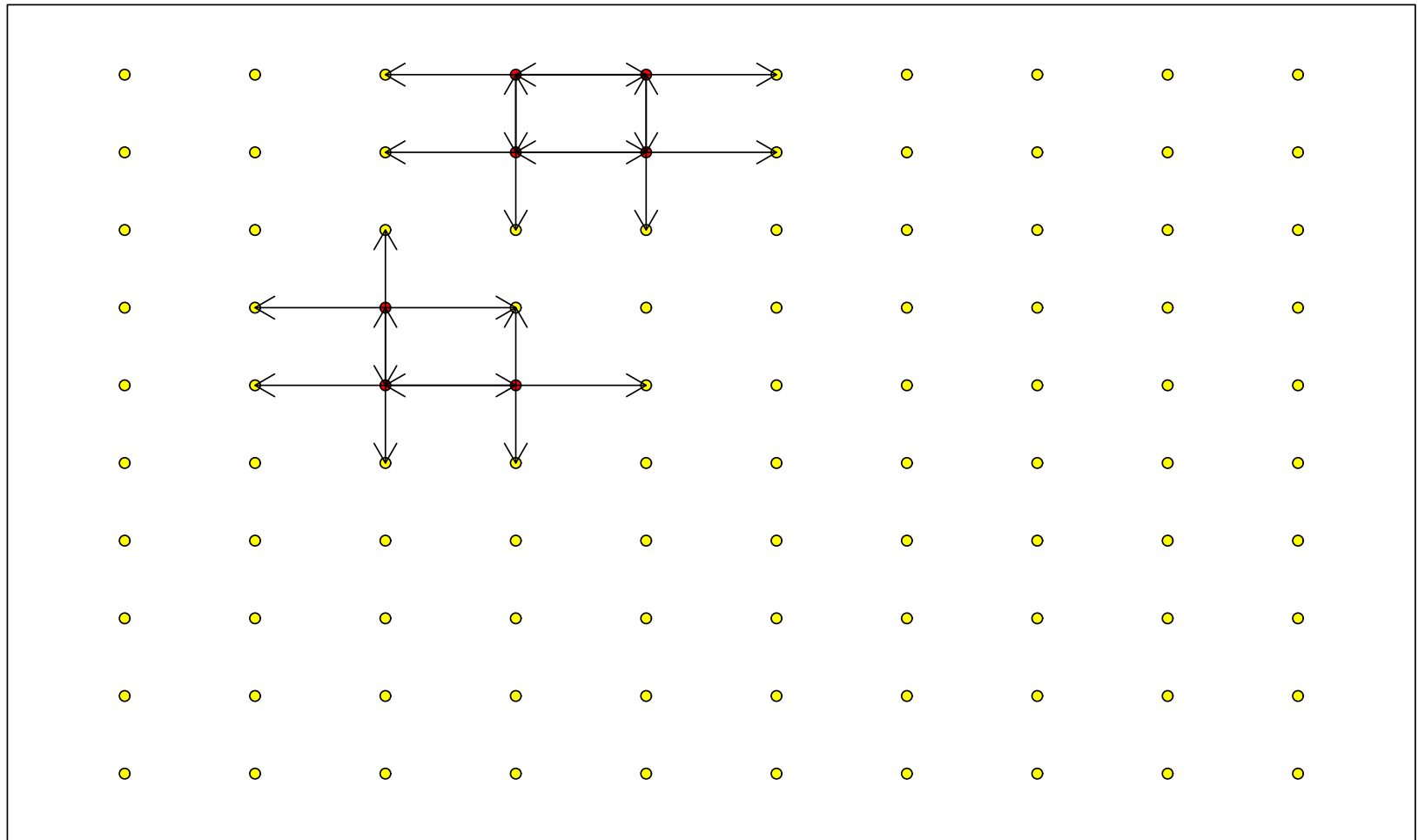
Spatial adaptive web sampling

spatial population



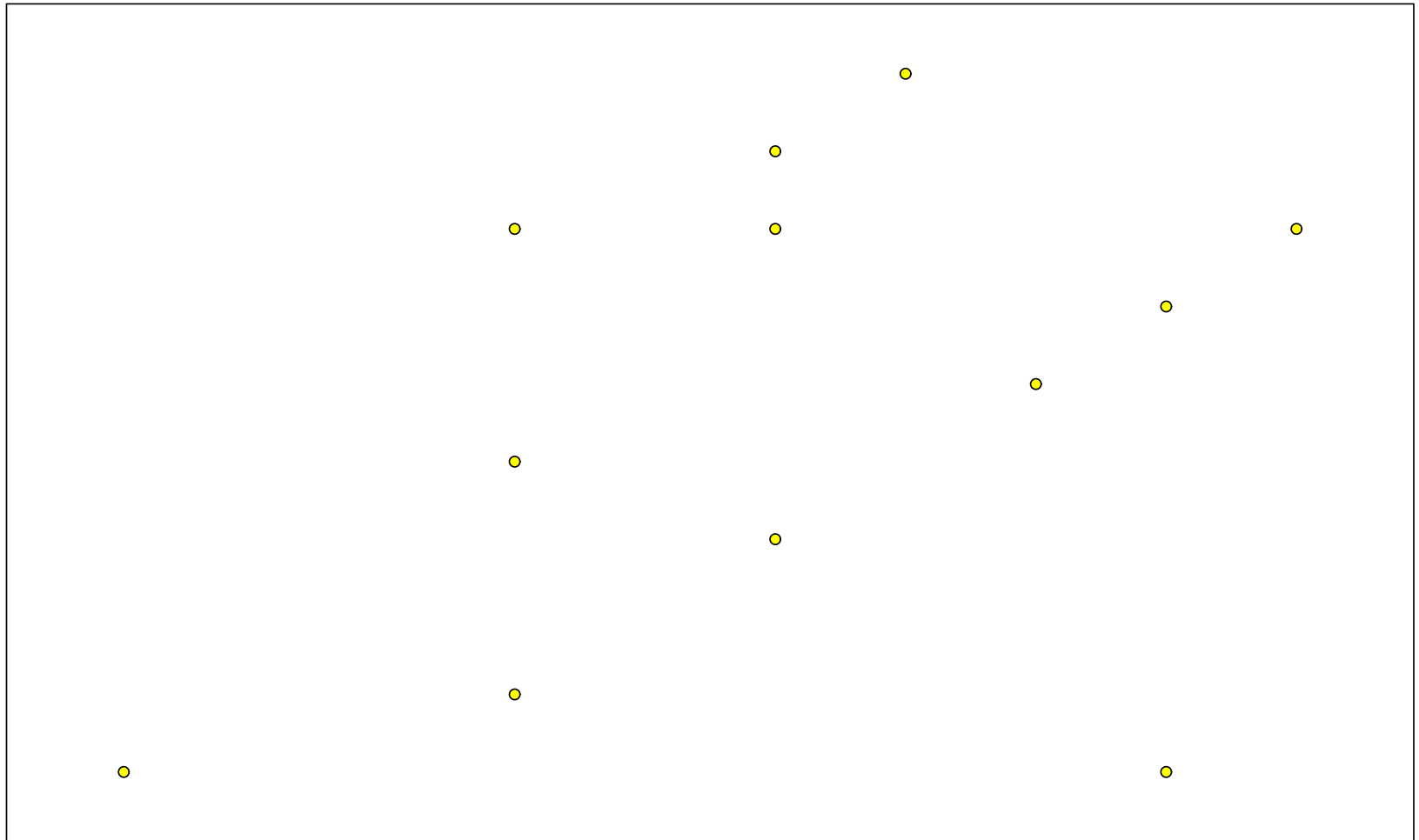
Network structure of spatial population

population graph



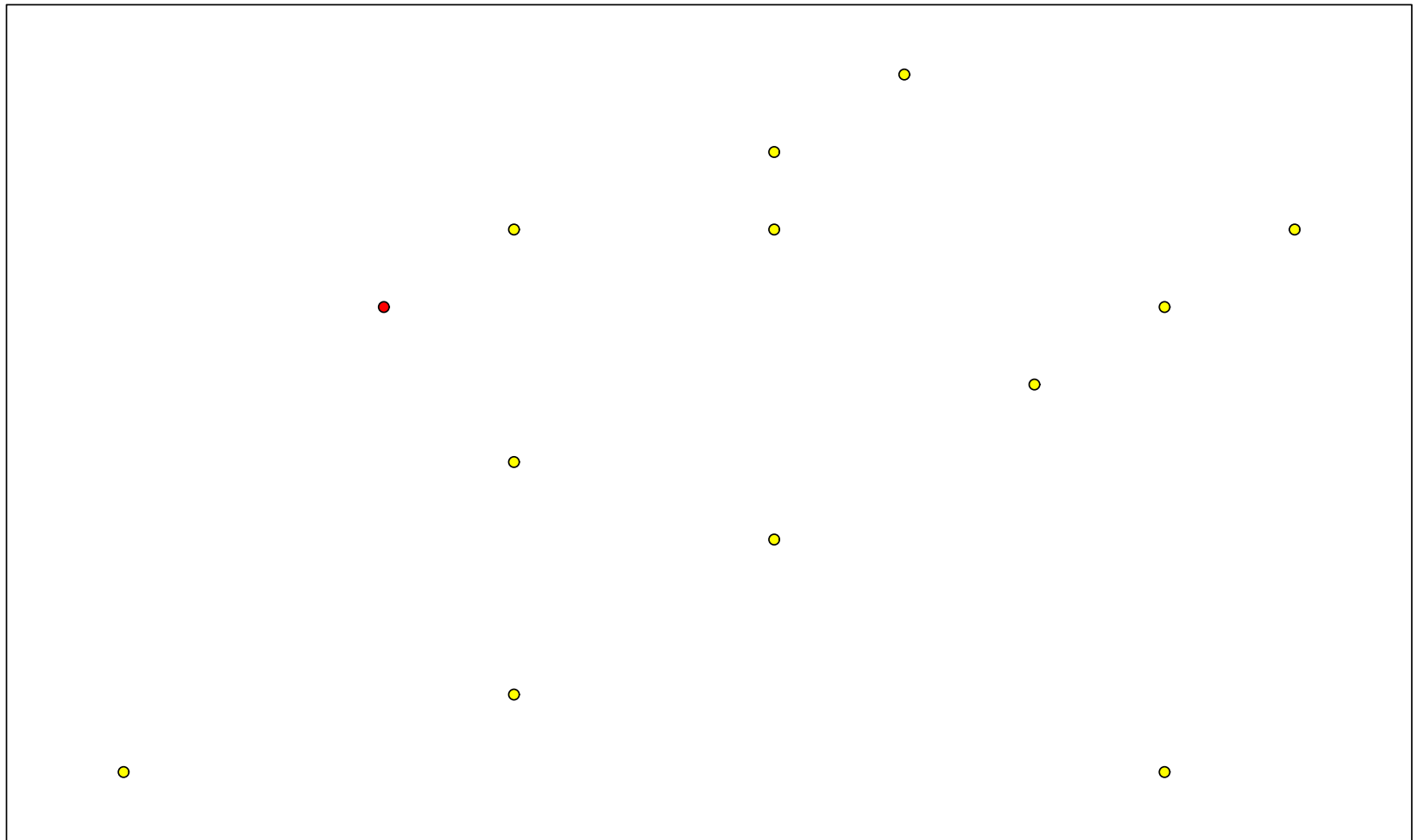
Adaptive web sample

sample



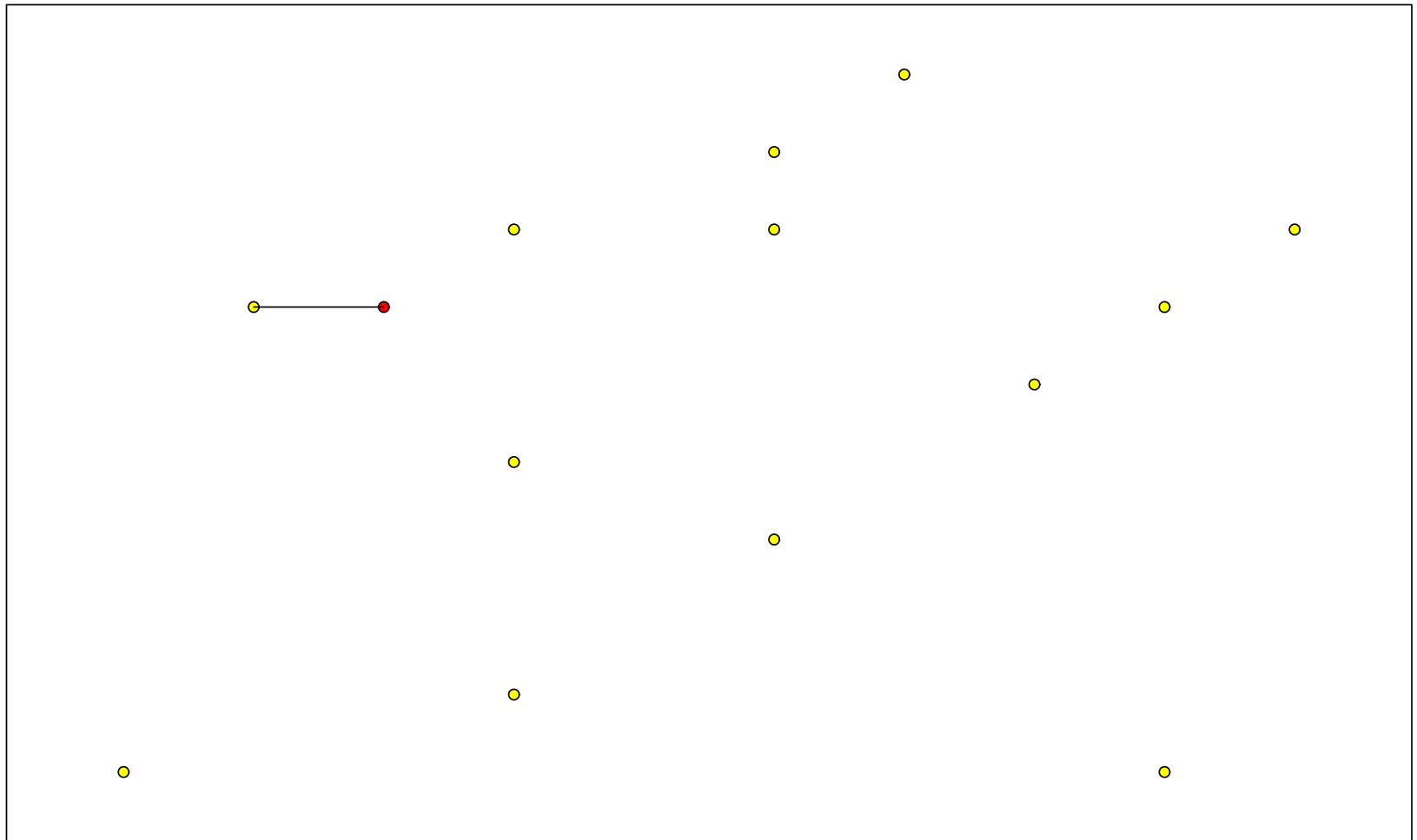
Adaptive web sample

sample



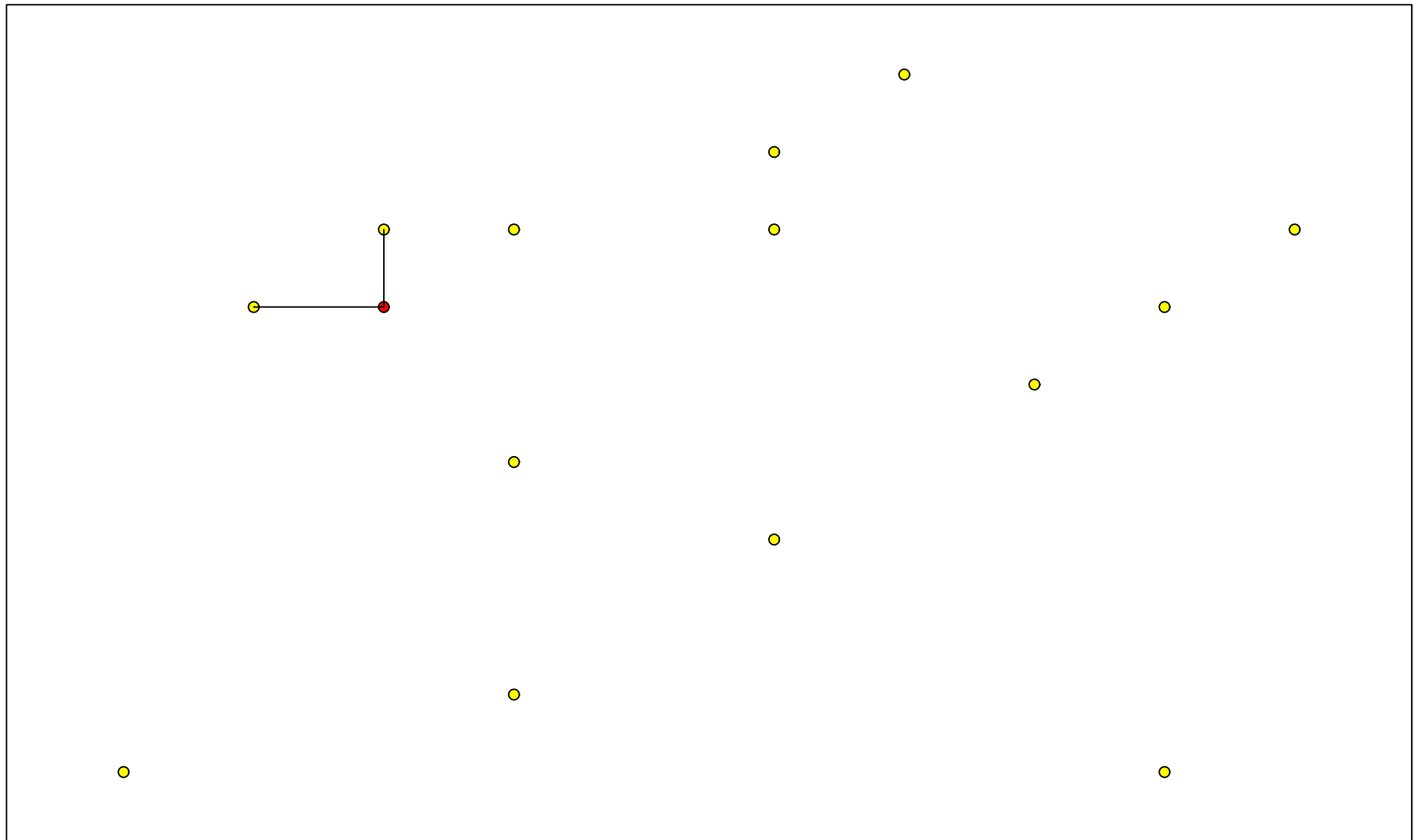
Adaptive web sample

sample



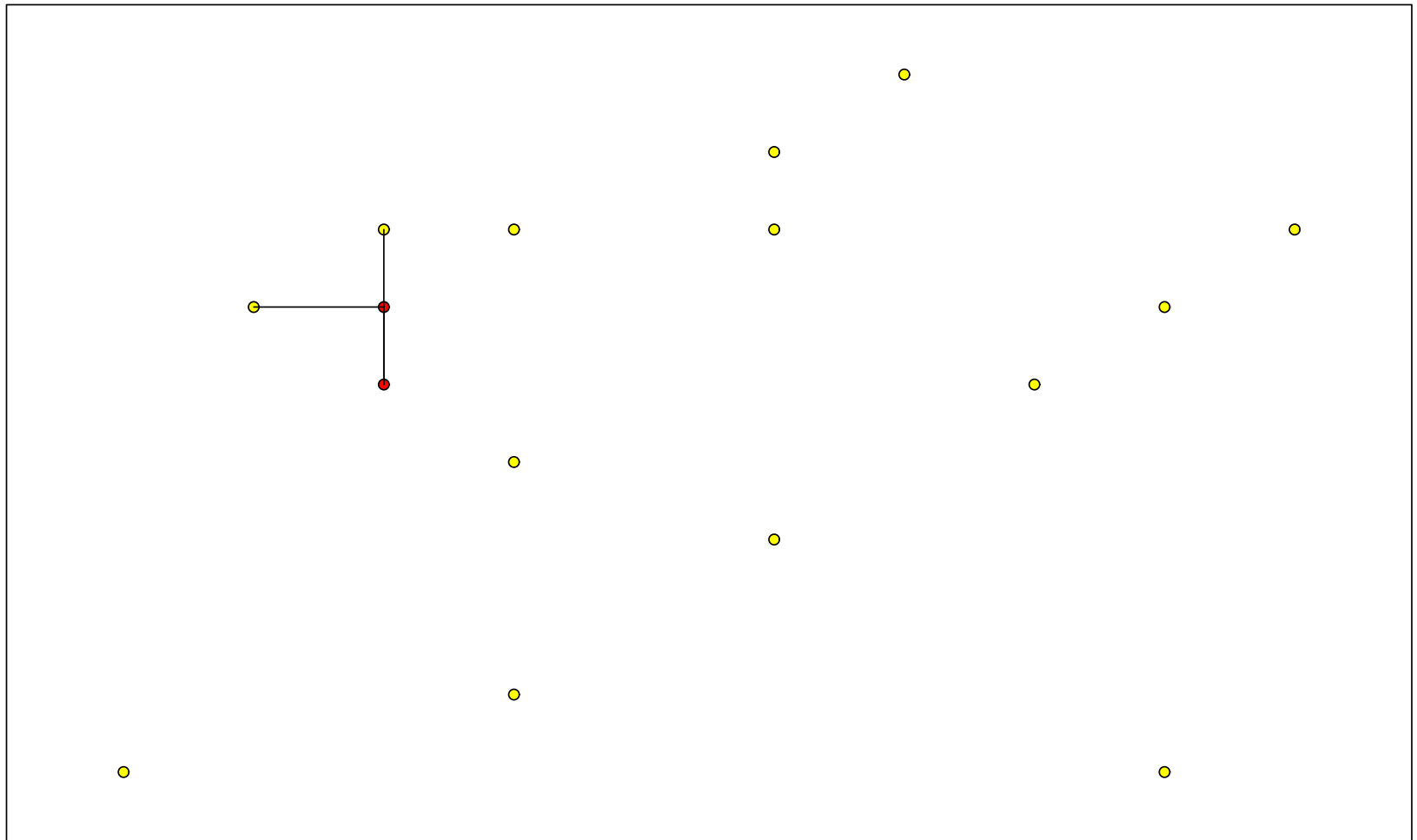
Adaptive web sample

sample



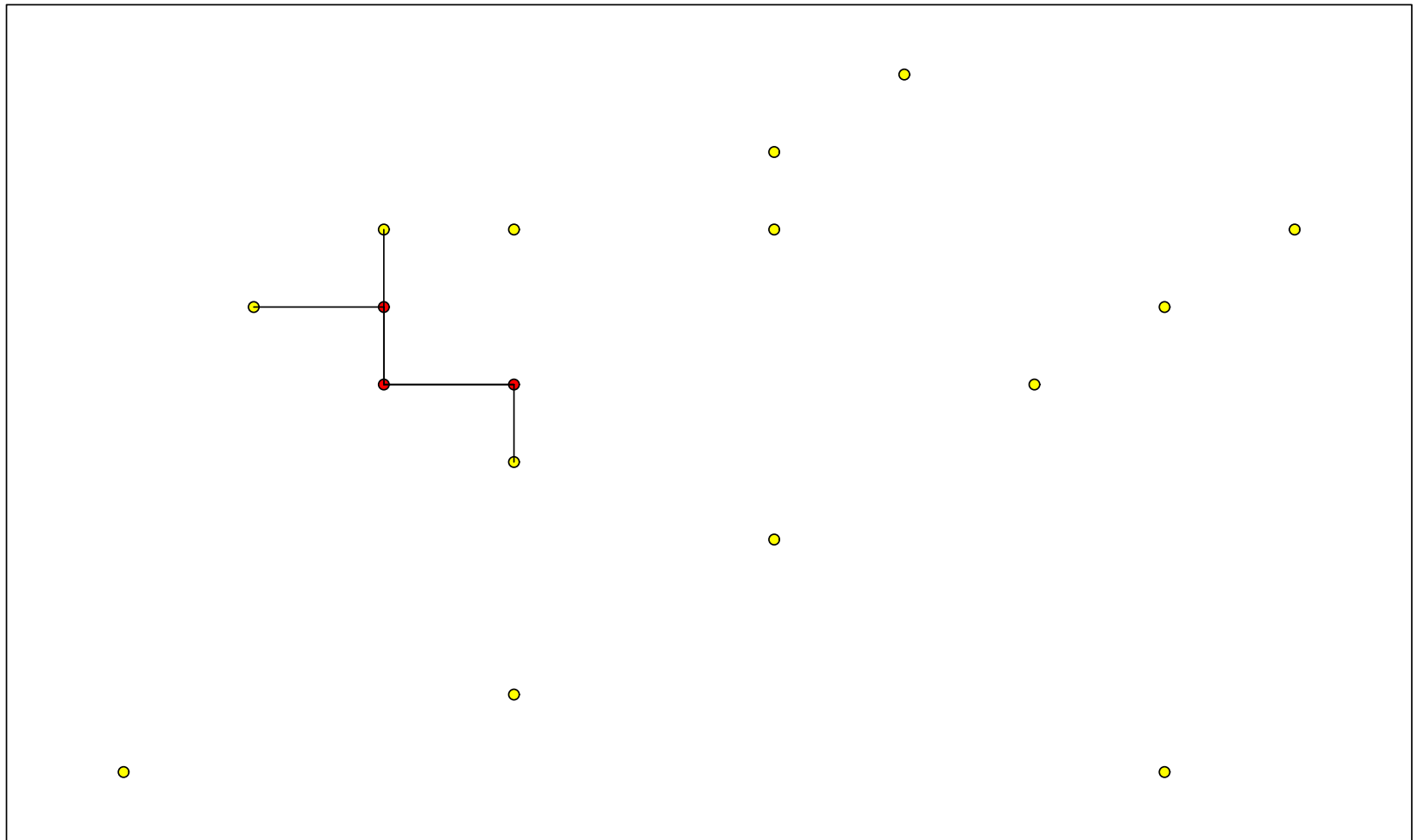
Adaptive web sample

sample



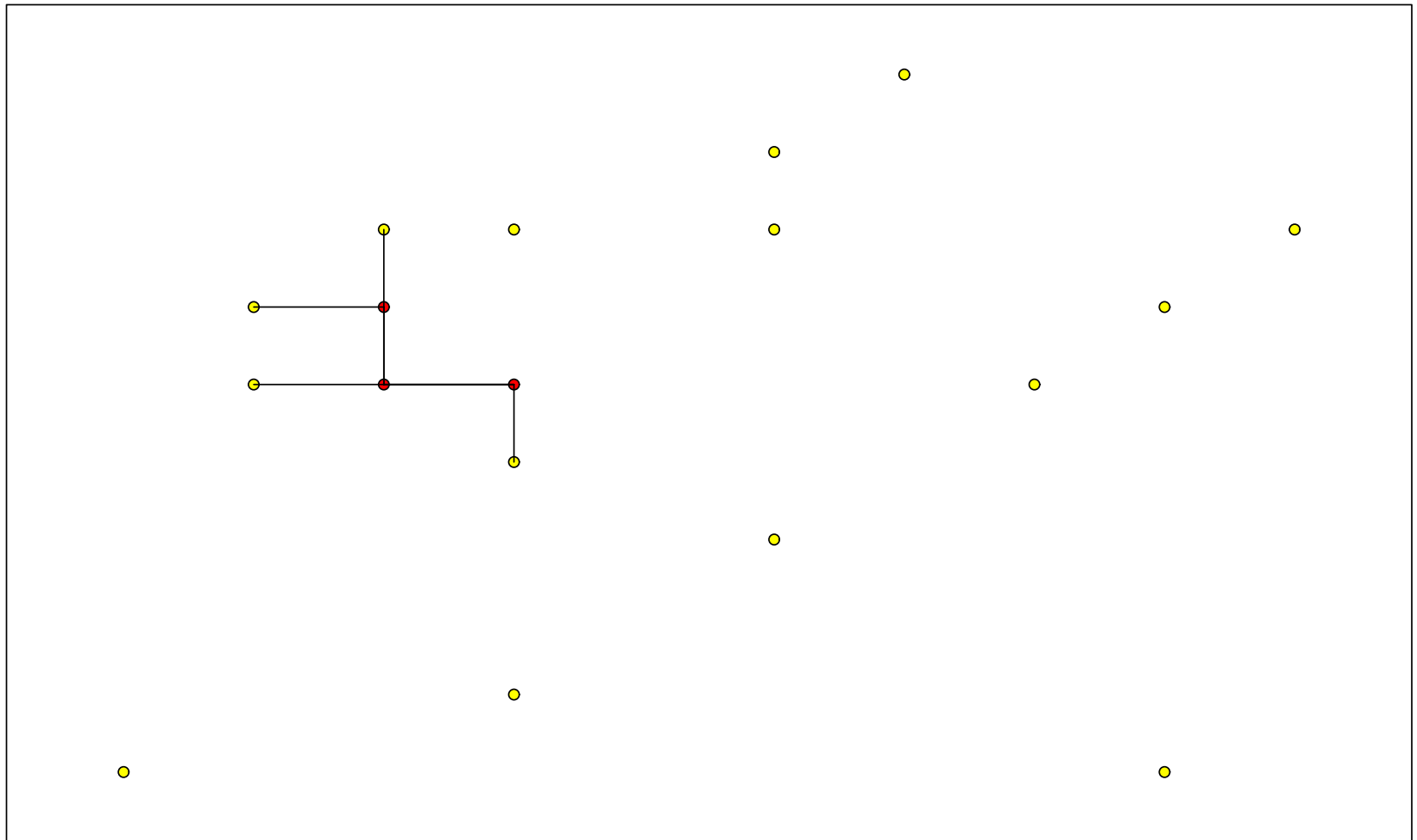
Adaptive web sample

sample



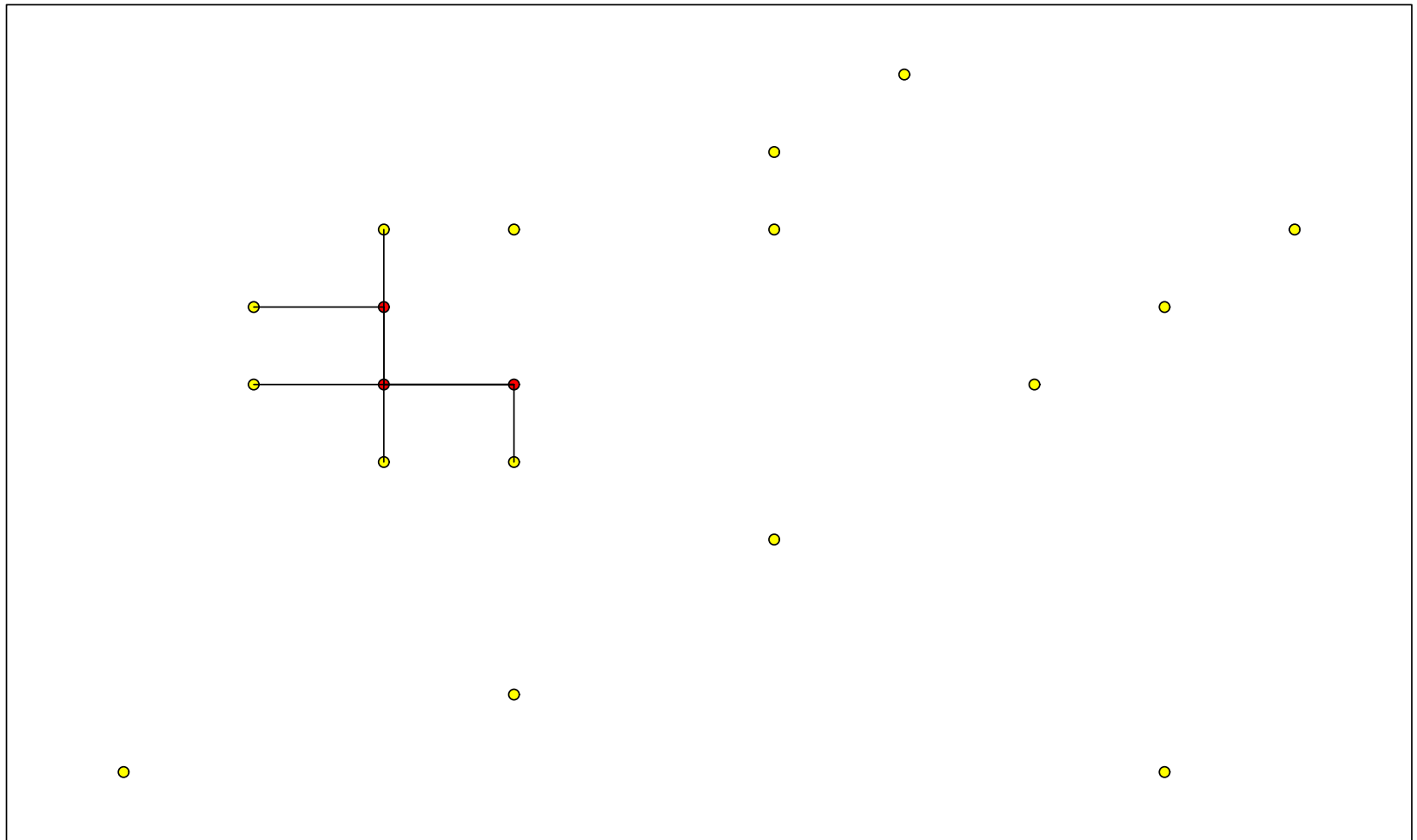
Adaptive web sample

sample



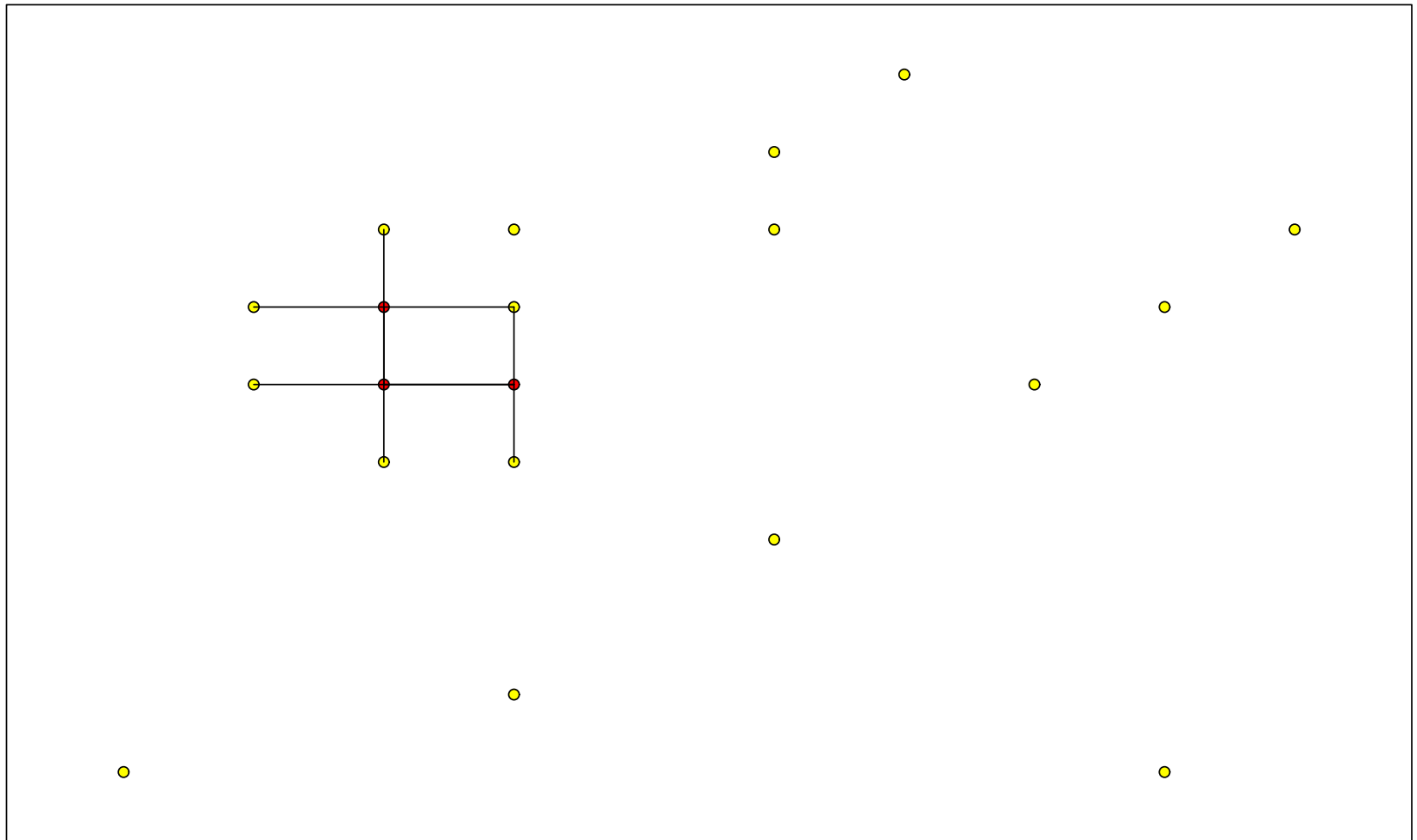
Adaptive web sample

sample



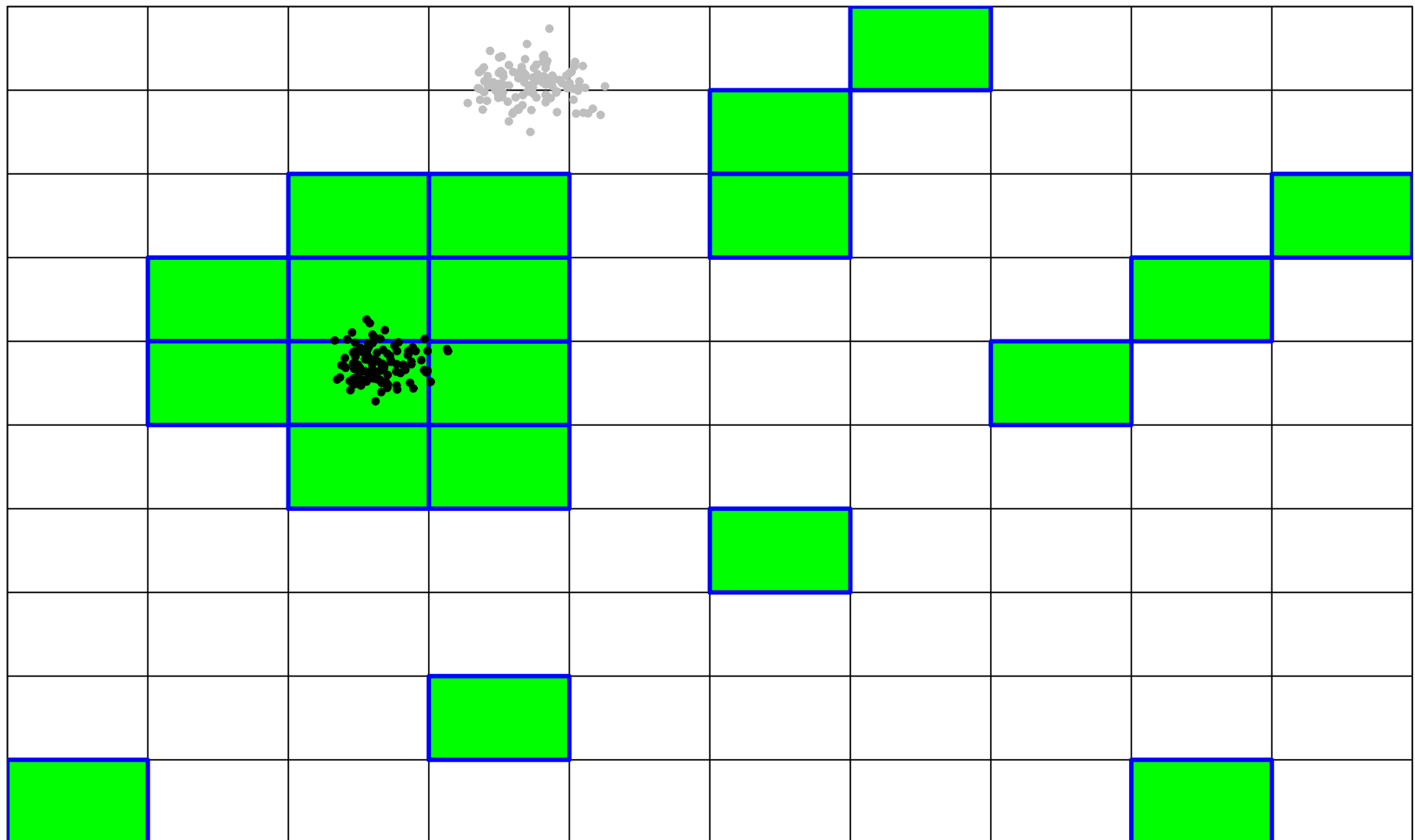
Adaptive web sample

sample



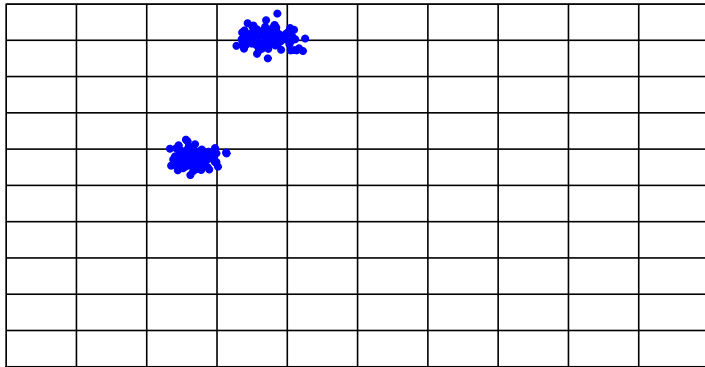
The resulting spatial sample

spatial population

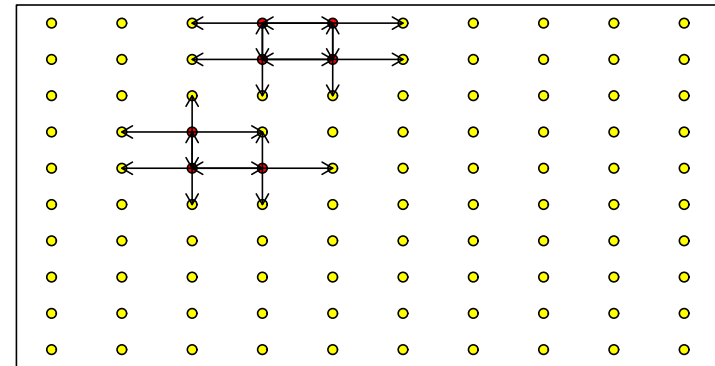


Active set design variations

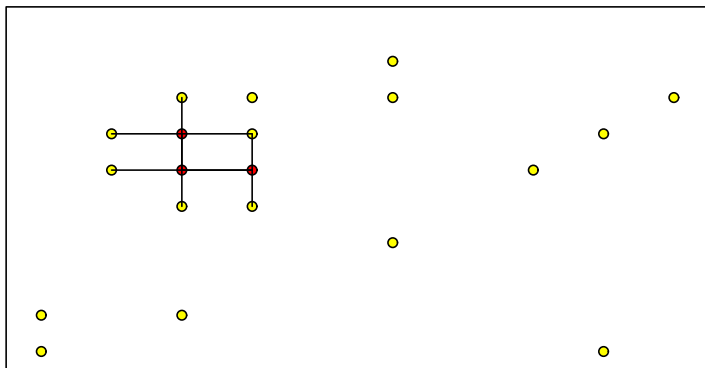
spatial population



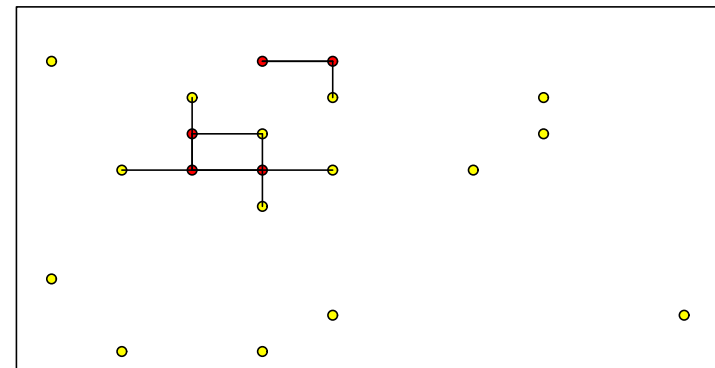
population graph



active set sample



active set sample



Migratory waterfowl survey



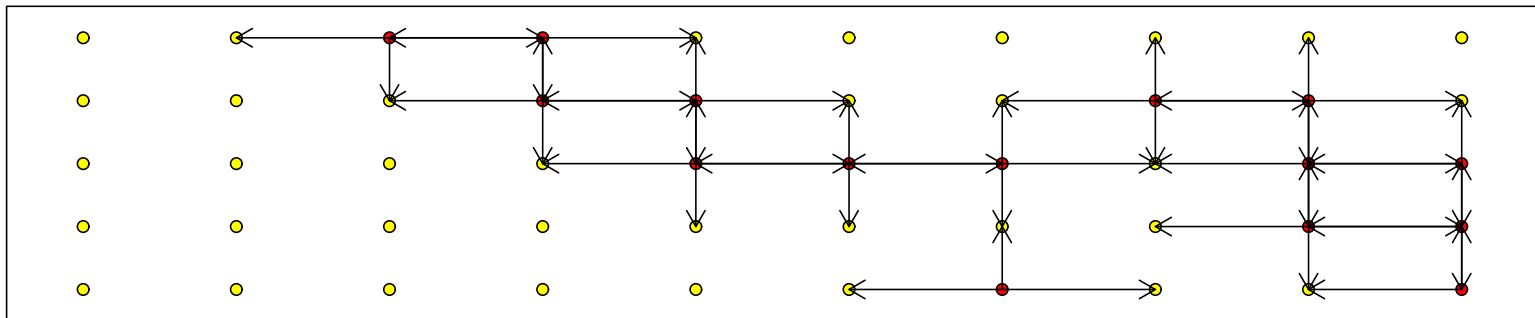
J.I. Hodges

Blue-winged teal population

spatial population

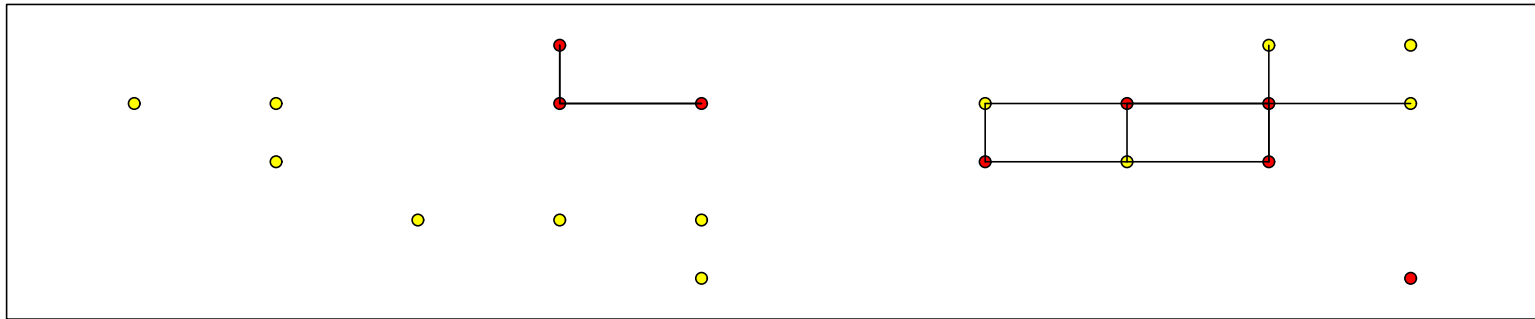
0	0	3	5	0	0	0	0	0	0
0	0	0	24	14	0	0	10	103	0
0	0	0	0	2	3	2	0	13639	1
0	0	0	0	0	0	0	0	14	122
0	0	0	0	0	0	2	0	0	177

population graph

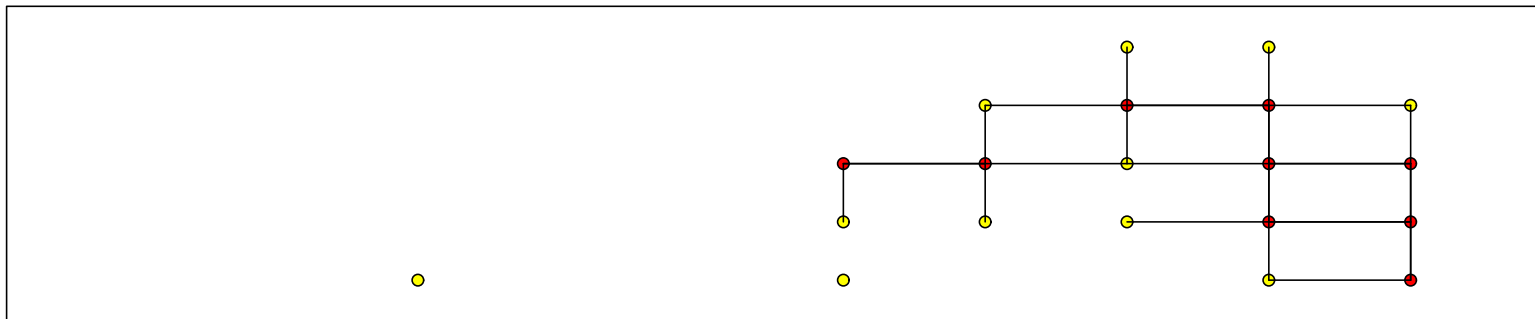


Two samples, n=20. Top: $n_0 = 13$. Bottom: $n_0 = 1$.

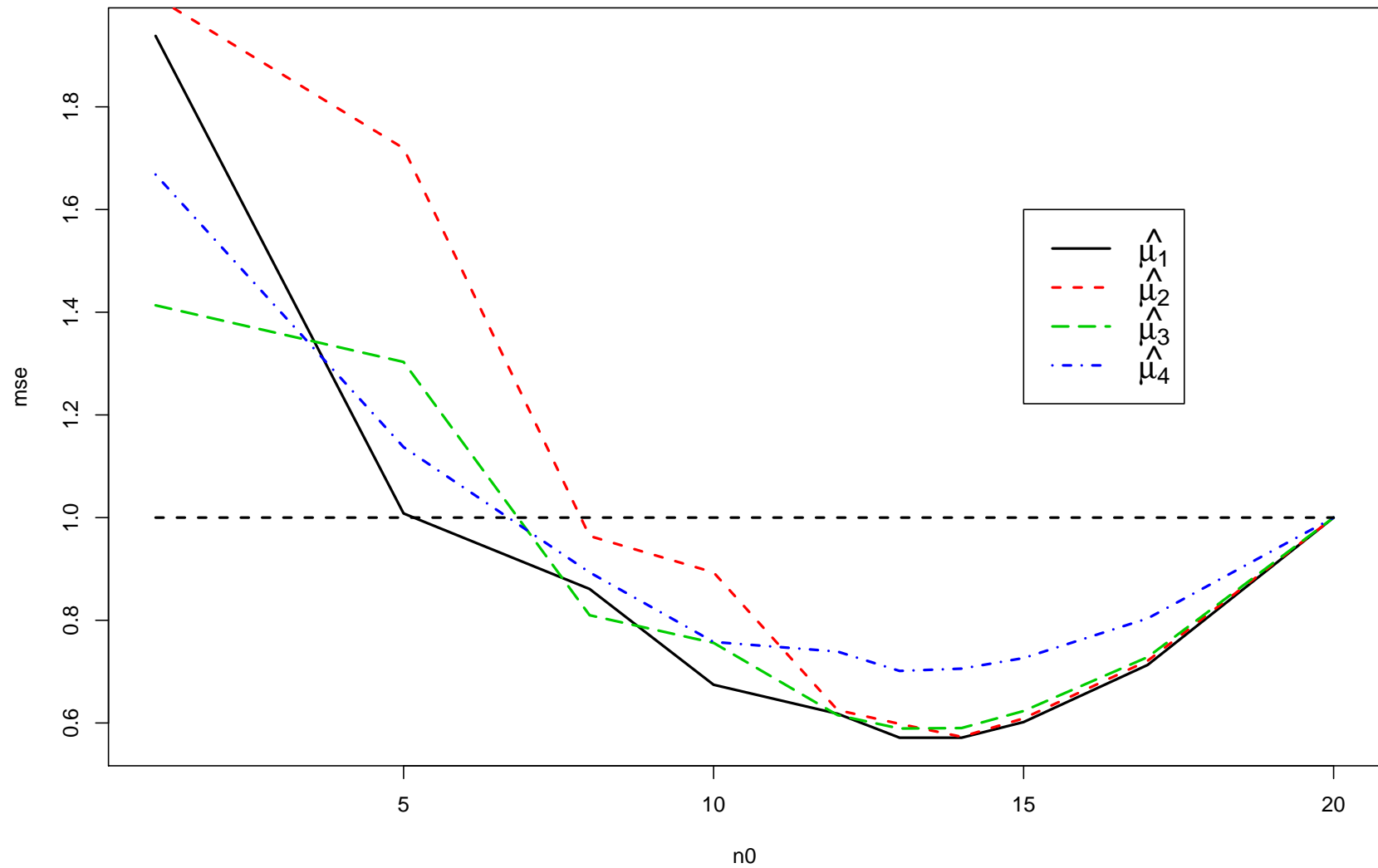
sample



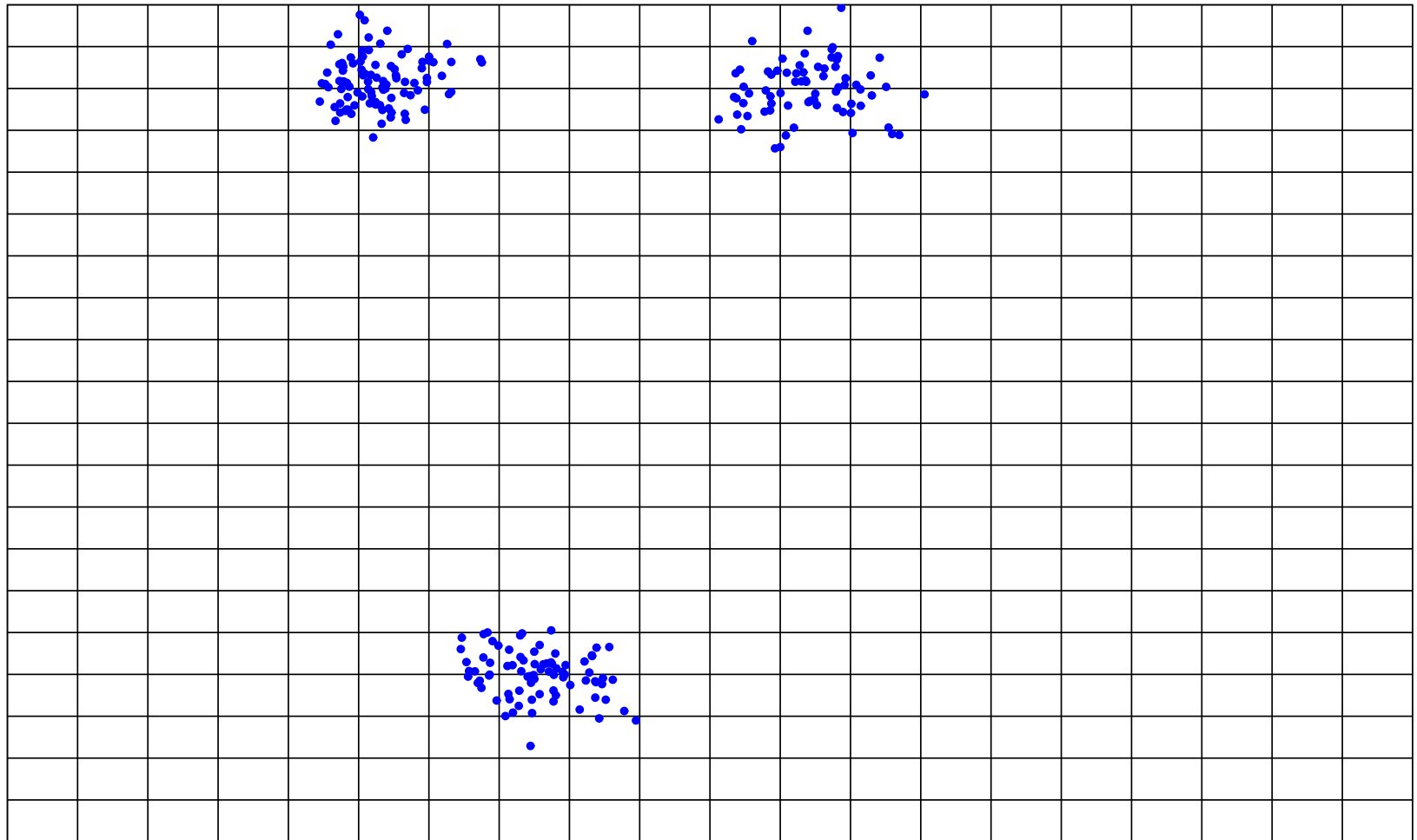
sample



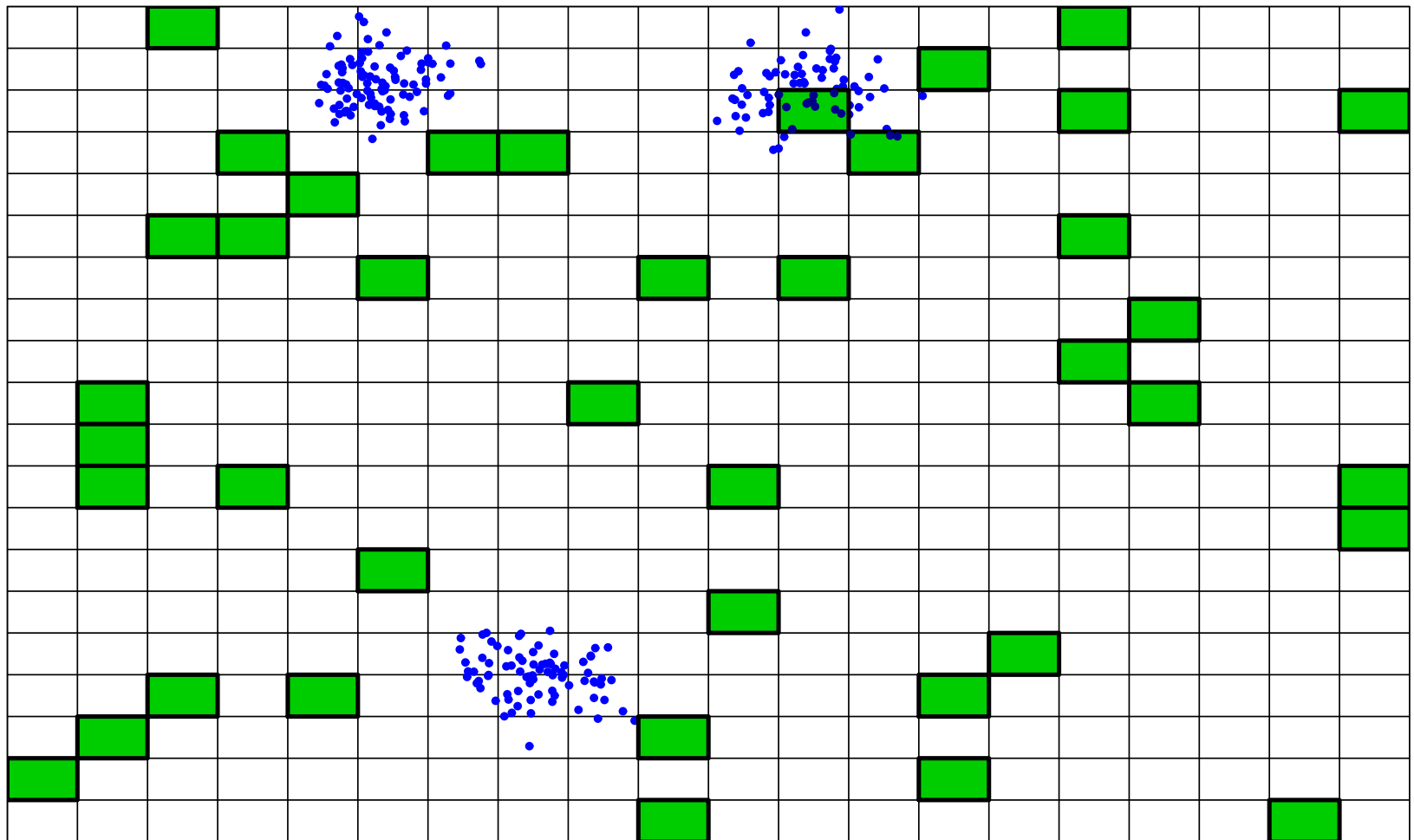
MSE of estimators depending on n_0 , with $n = 20$



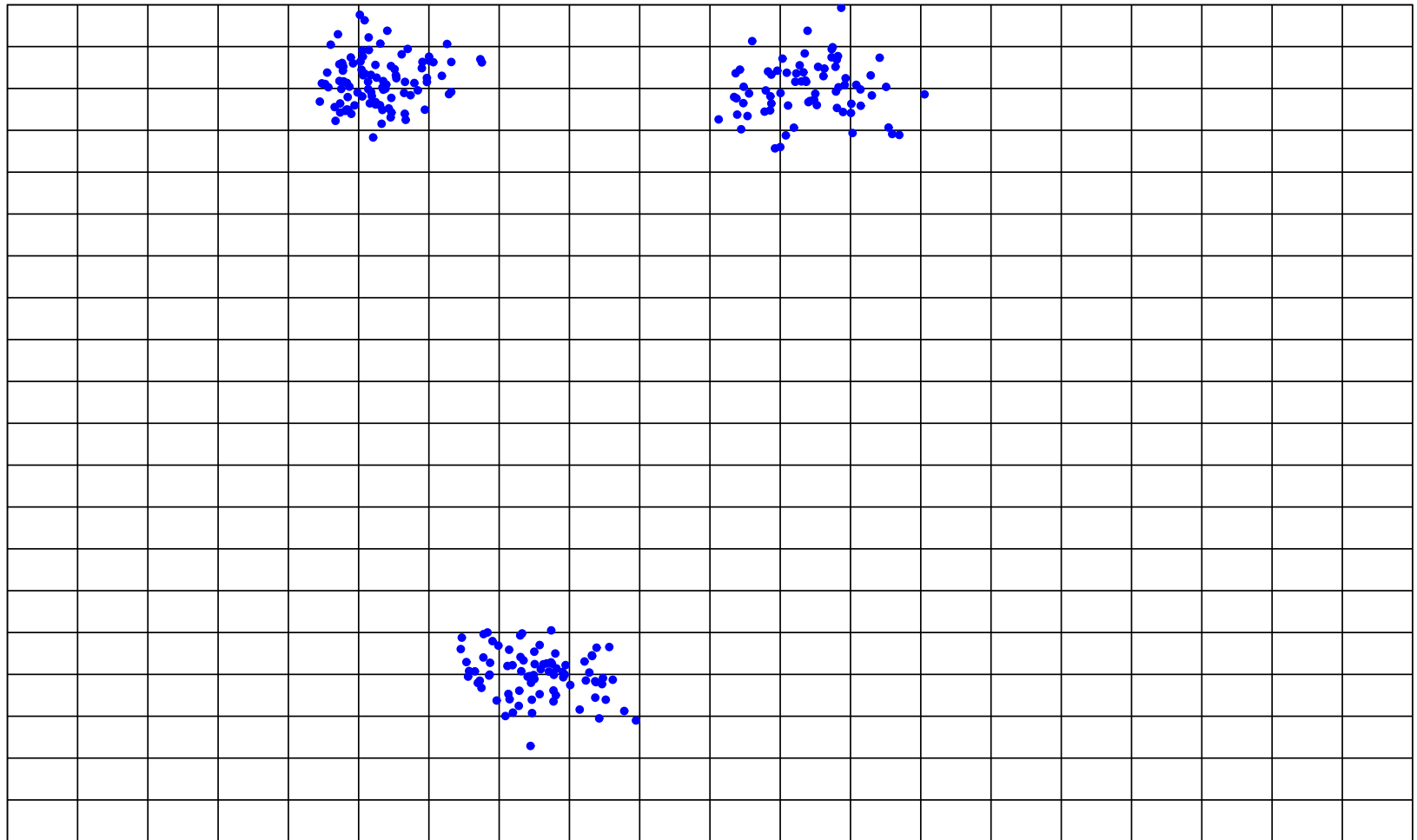
Rare, clustered population (Adaptive cluster sampling)



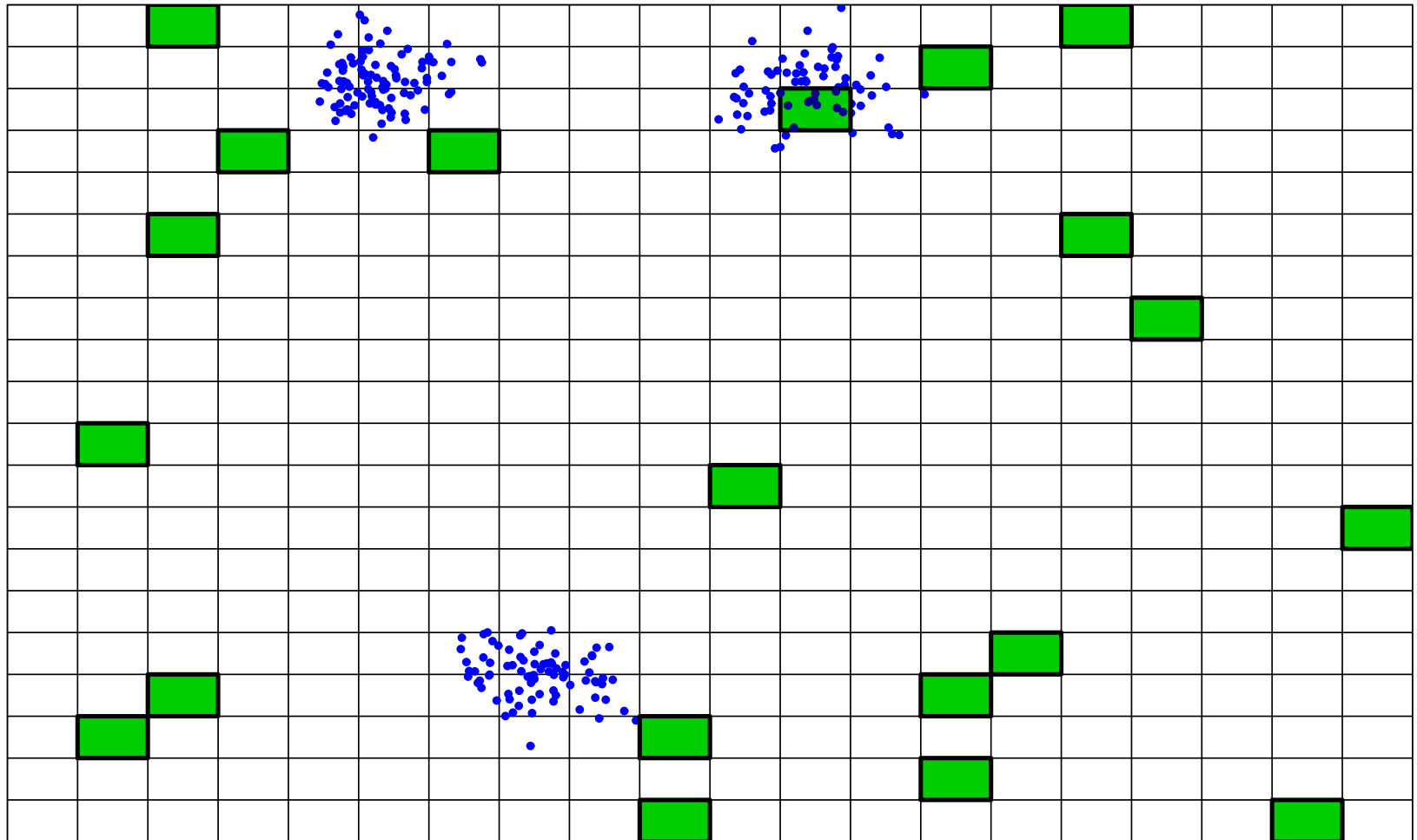
Random sample of 40 units



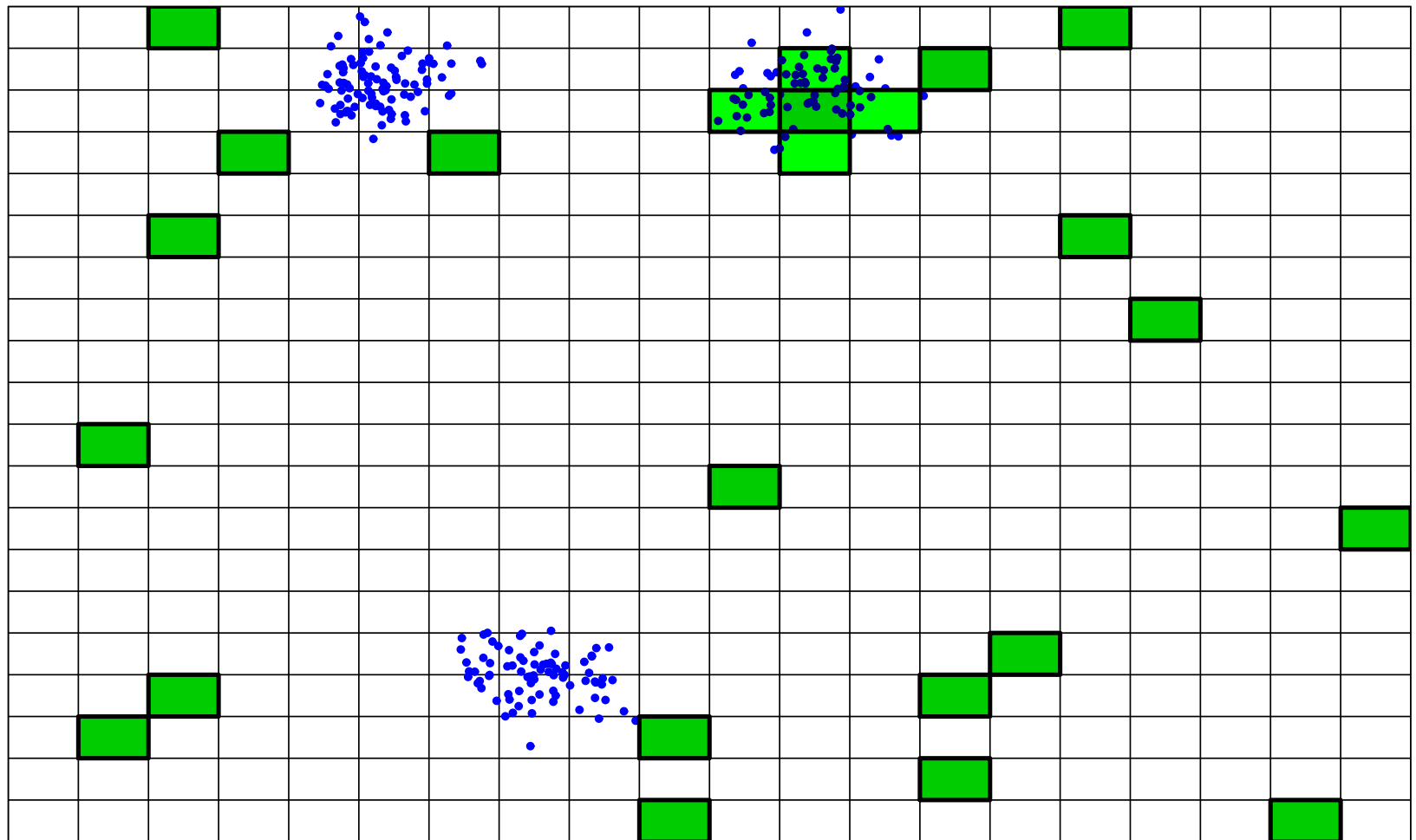
Same population



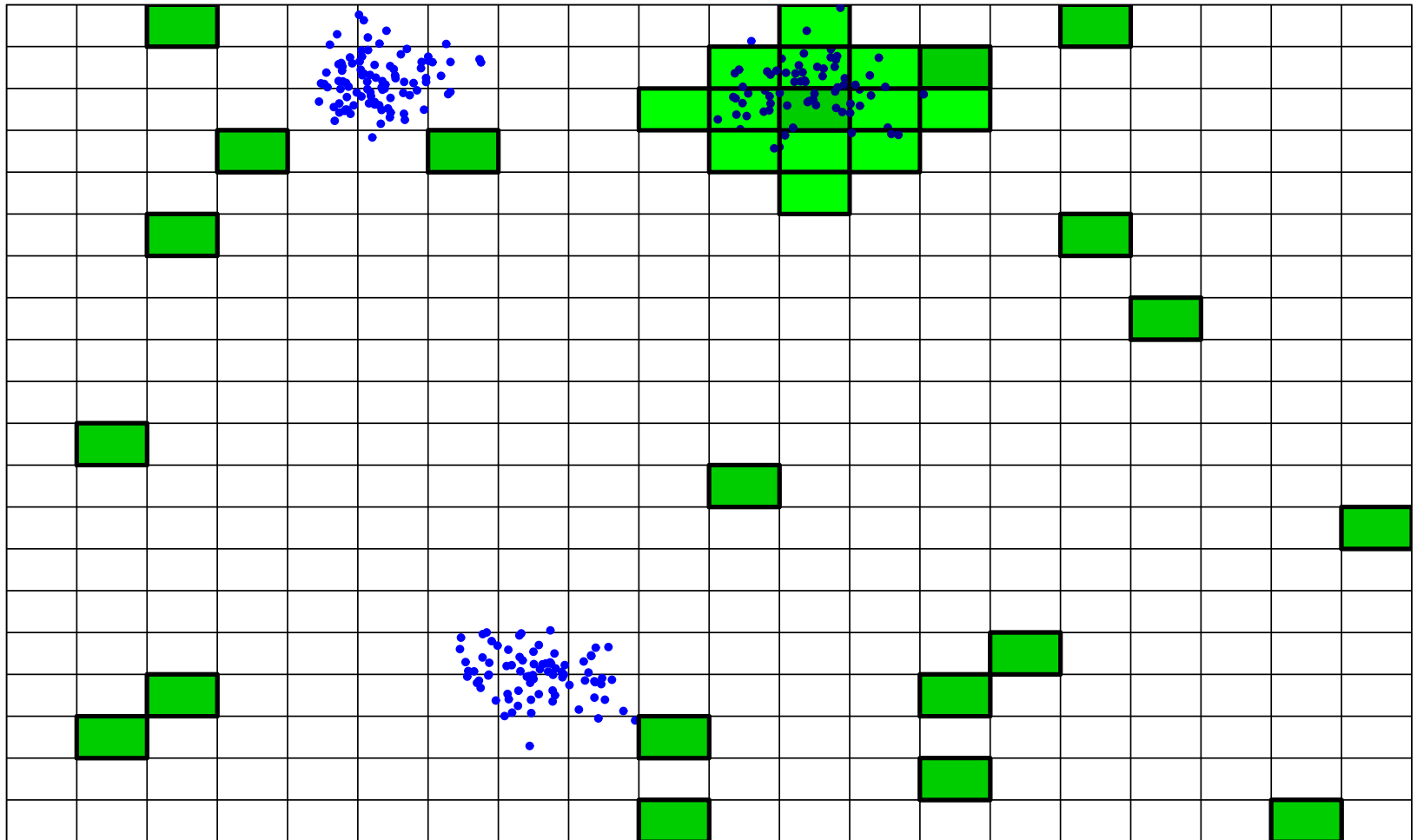
Initial sample of 20 units



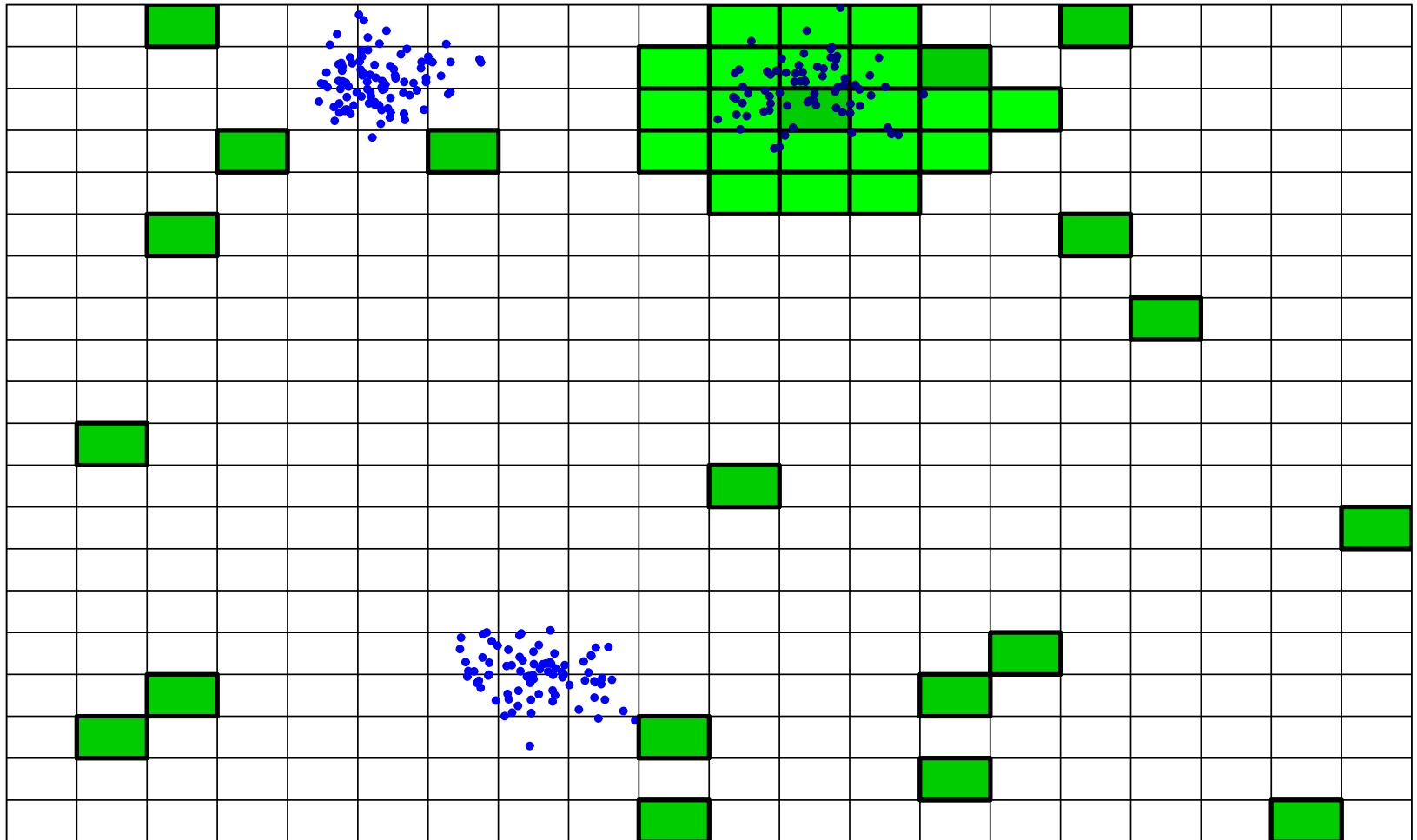
Adaptive cluster sample



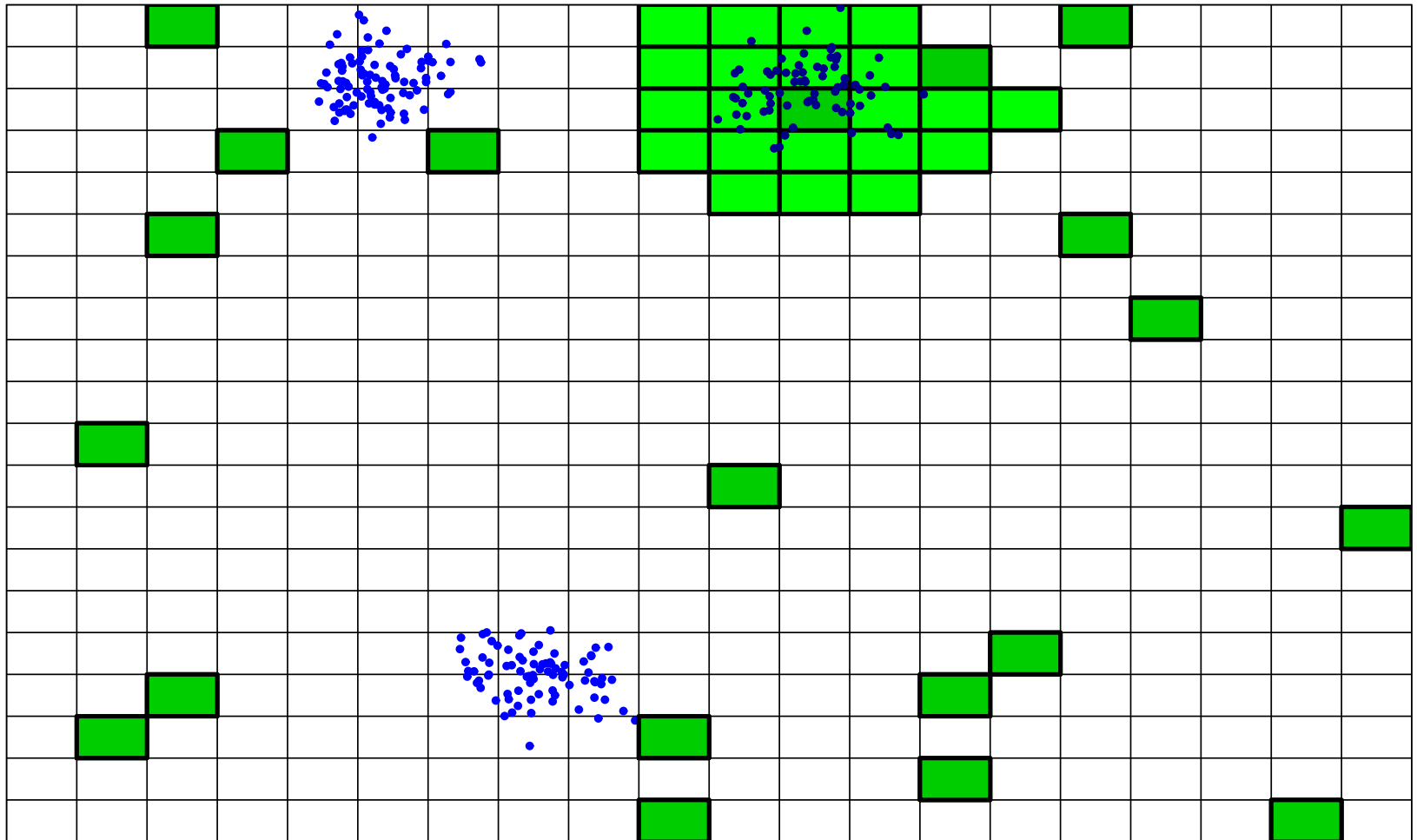
Adaptive cluster sample



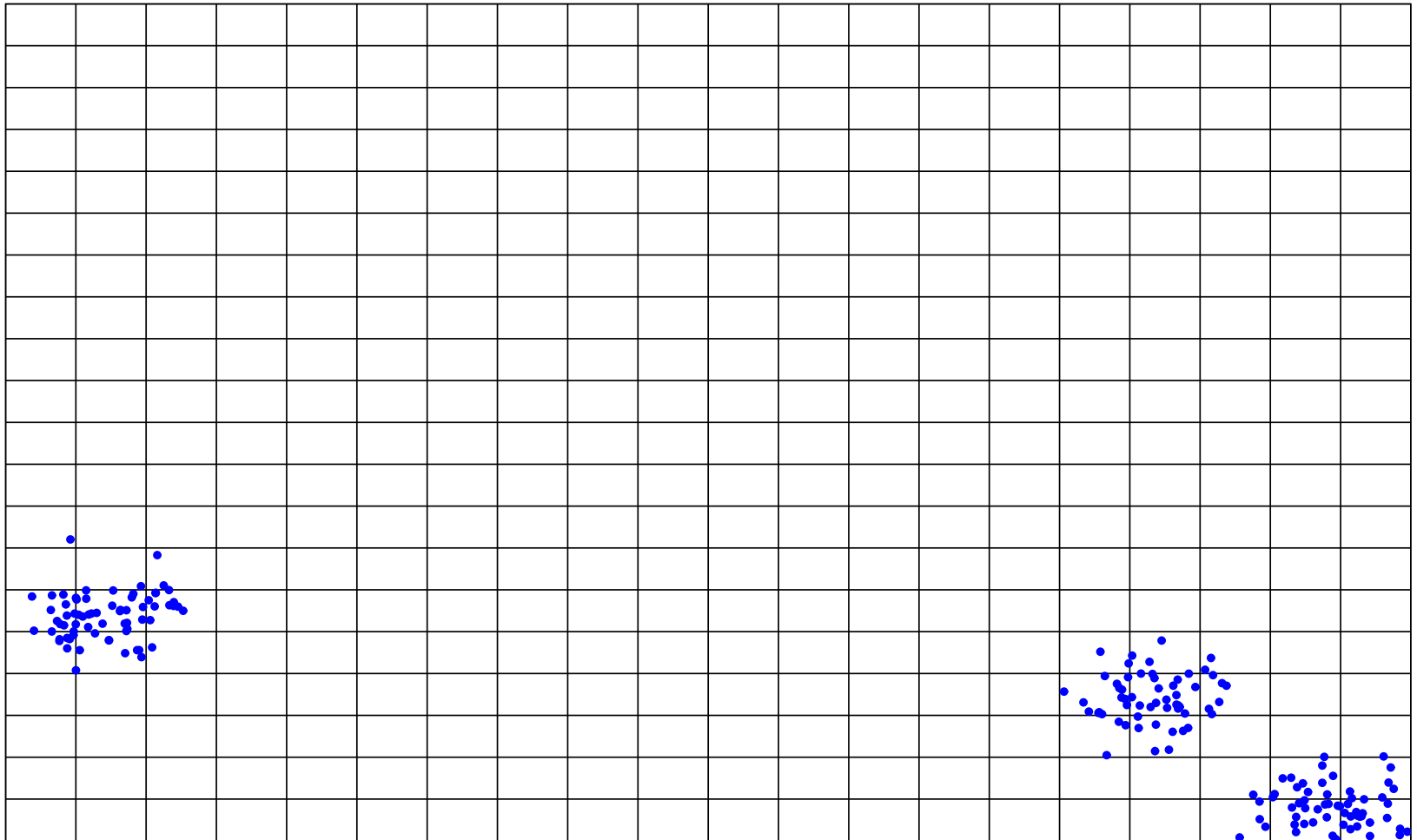
Adaptive cluster sample



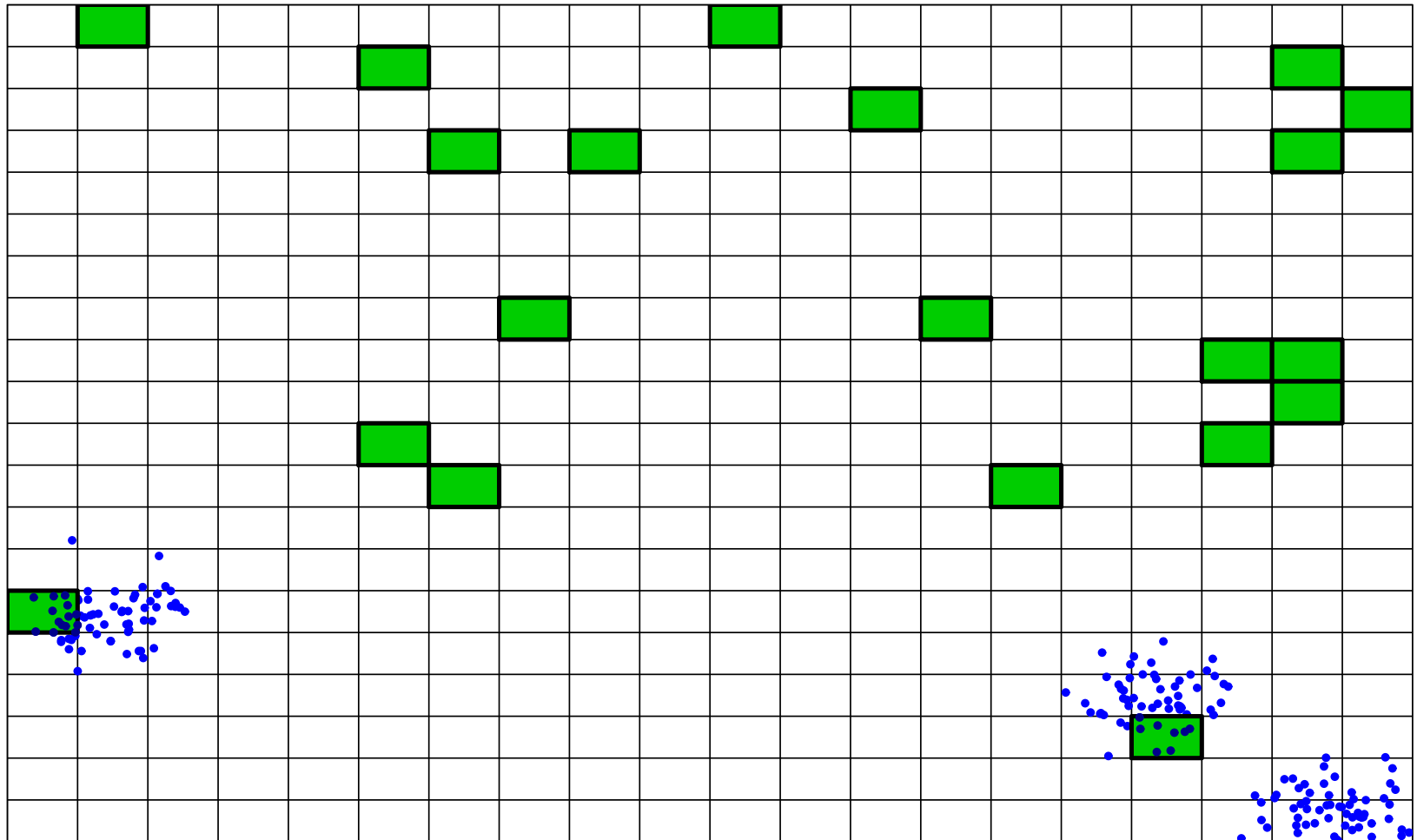
Adaptive cluster sample



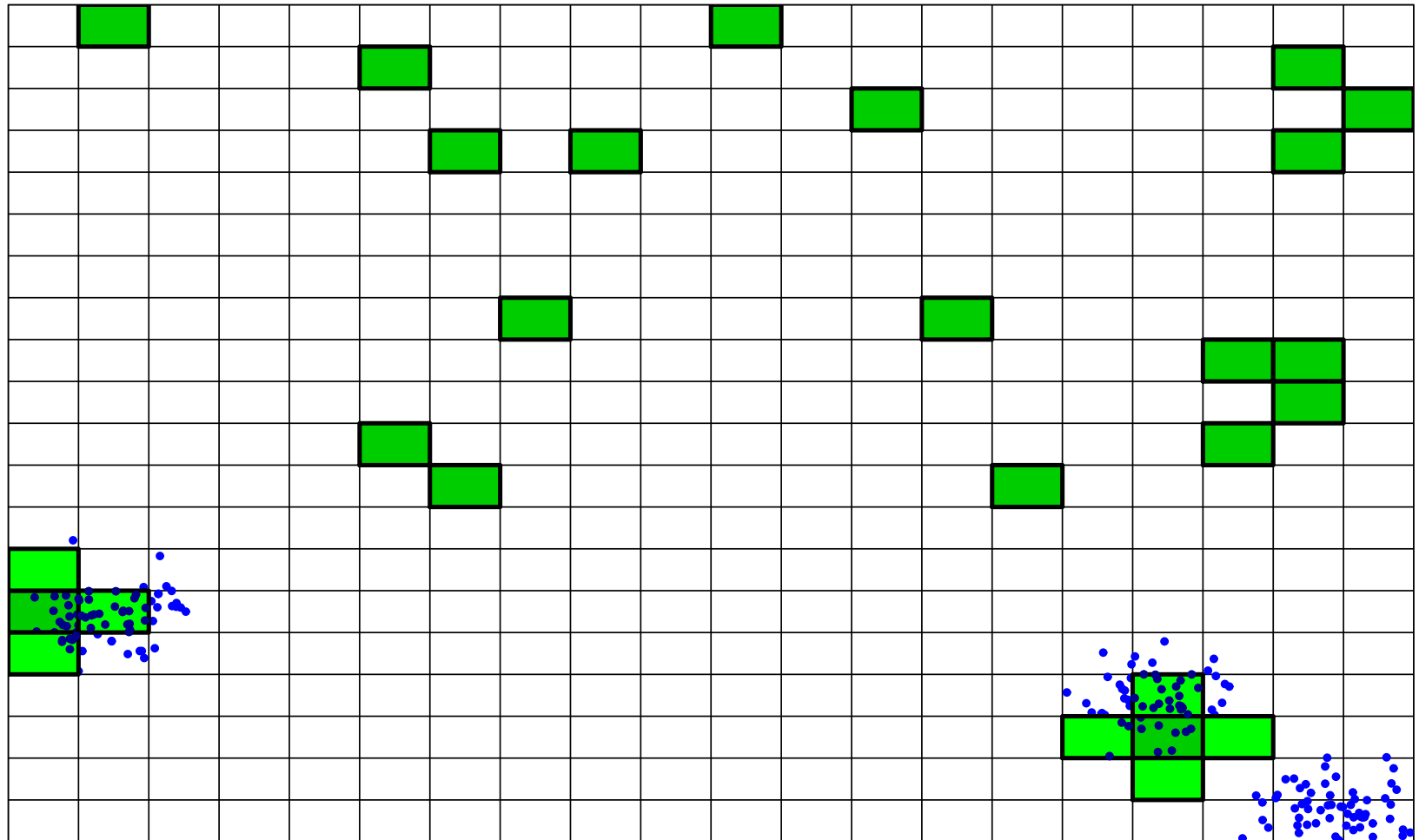
Changed population!



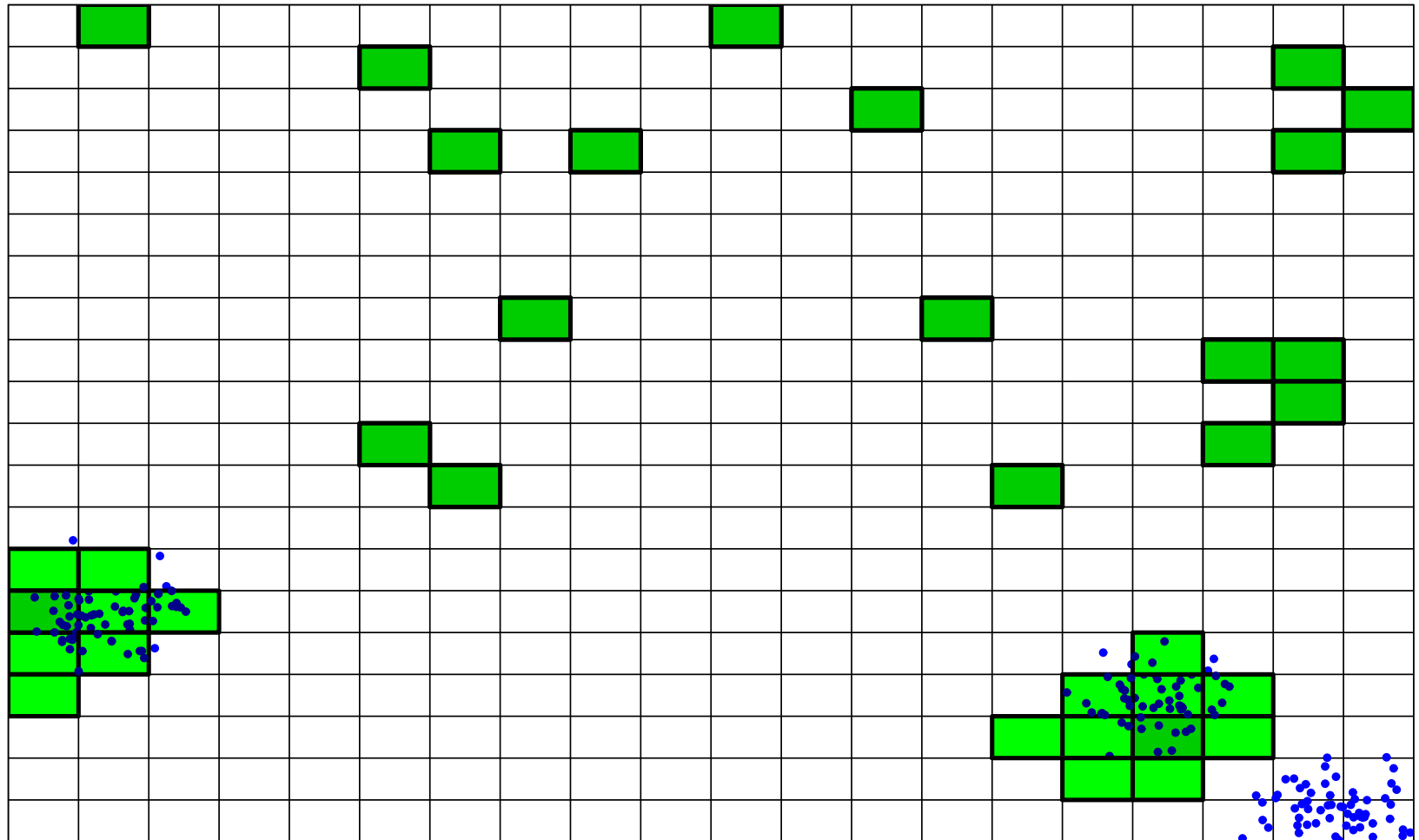
Initial sample of 20 units



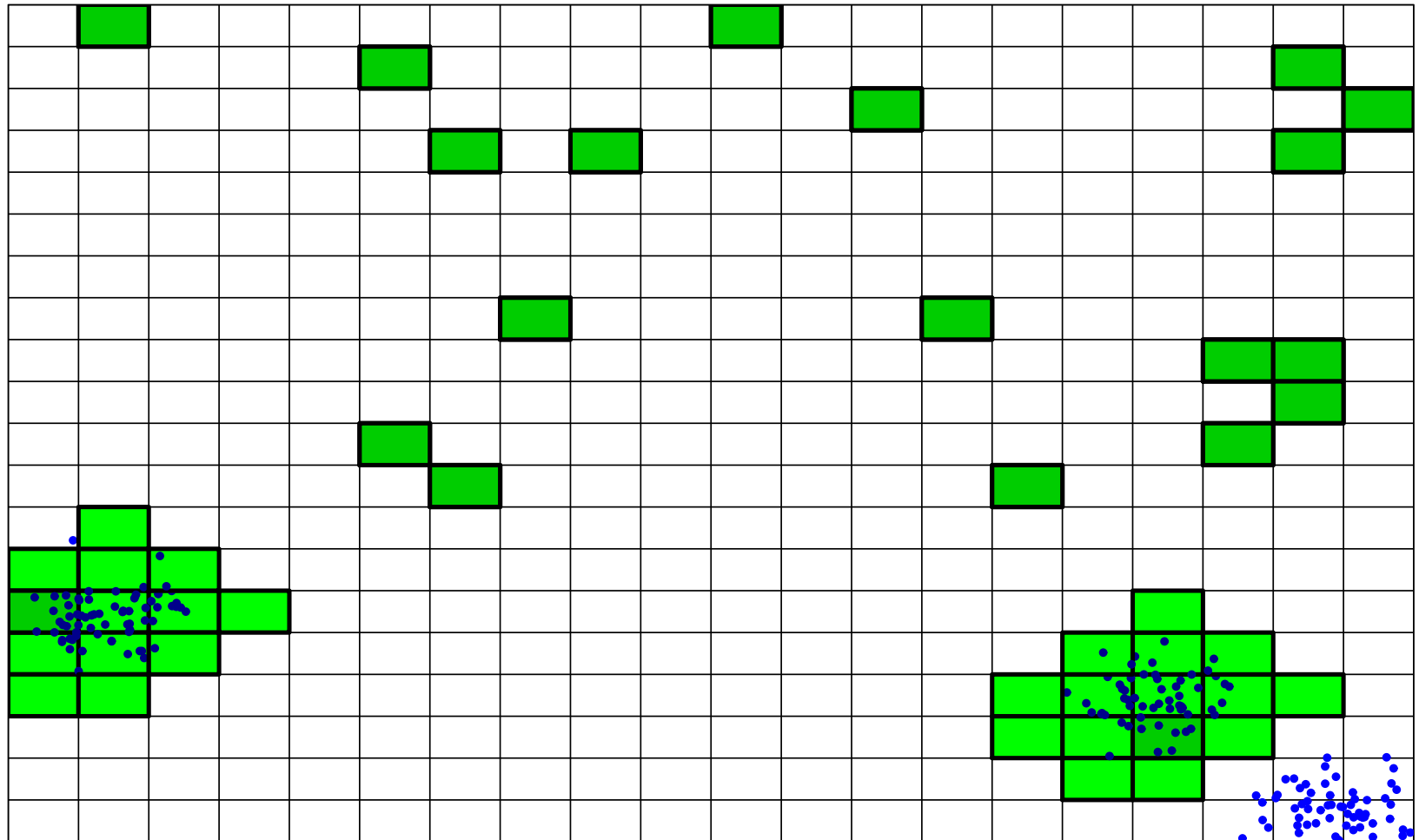
Adaptive cluster sample



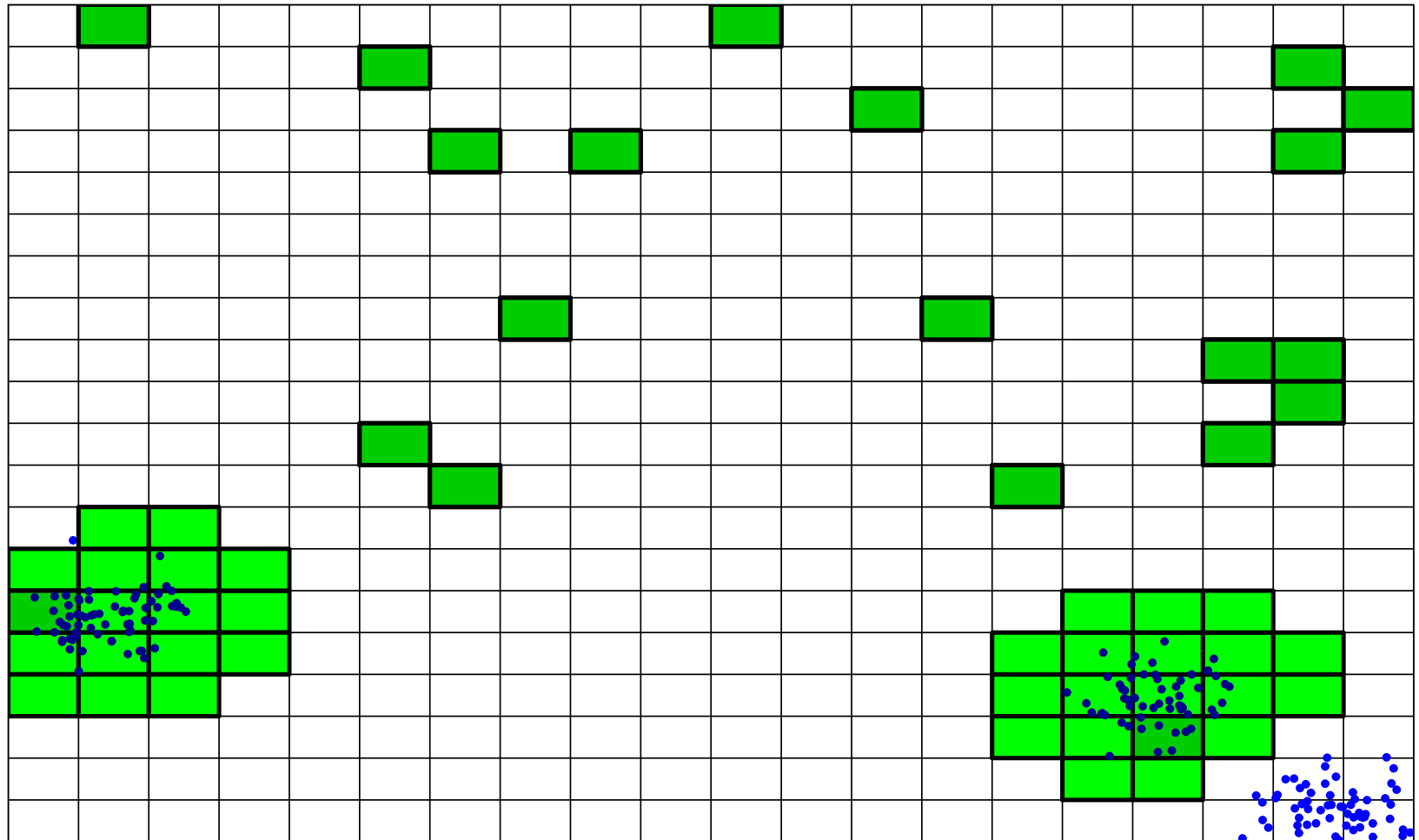
Adaptive cluster sample



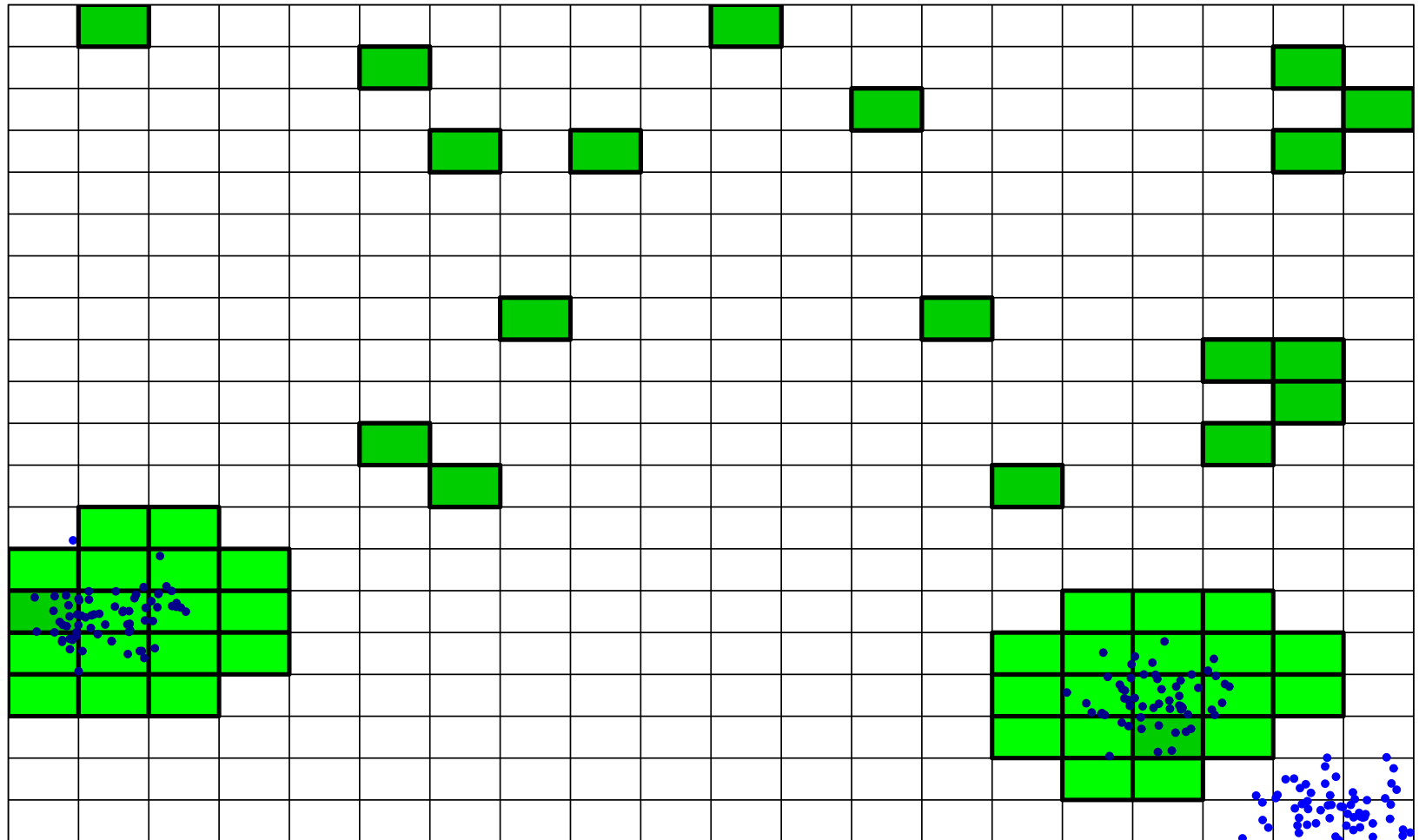
Adaptive cluster sample



Adaptive cluster sample



Adaptive cluster sample



7. Computational methods

1. Markov chain resampling in design-based estimation
2. Markov chain Monte Carlo in Bayes estimation
3. Computing and software

Aside: On computing

R 9-10 times faster than commercial counterpart

C 20 times faster than R

For exploratory research, graphics needed at each iteration.

Limits of computing

Exact Rao-Blackwell estimate with adaptive web sampling involves selection and estimation computations for all reorderings of the sample

Can do up to sample size 10

With a computer 1000 times faster I could handle a sample size of 13.

Markov chain resampling a more practical answer.