Bigfoot, Sasquatch, and the Yeti: The Missing Links

What we don't know about the AS graph

Matthew Roughan, Jonathon Tuke, Olaf Maennel

<matthew.roughan@adelaide.edu.au>

<simon.tuke@adelaide.edu.au>

<olaf@maennel.net>

Discipline of Applied Mathematics
School of Mathematical Sciences
University of Adelaide

The Yeti

Apparently we have found the Yeti

http://www.canberratimes.com.au/news/local/news/general/yeti-truth-a-hairs-breadth-away/1227921.aspx

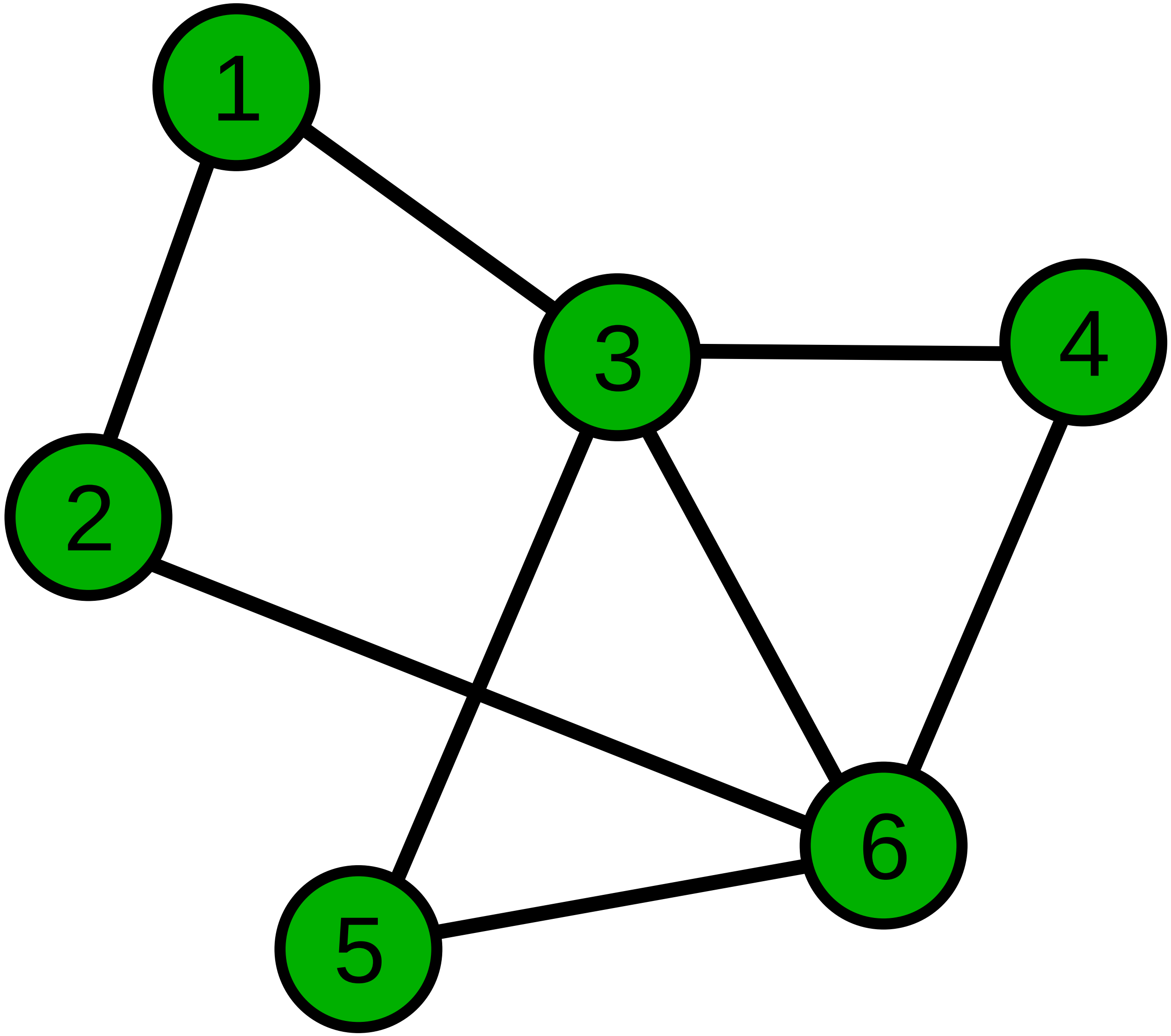
What about the other missing links?

Graph Theory and the Internet

- The Internet is made up of a bunch of connected devices
  - devices = nodes or vertices
  - connections = links or edges
- Represent as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$
  - set of nodes $\mathcal{N}$
  - set of edges $\mathcal{E}$
- e.g. AS-graph
  - nodes are Autonomous Systems (ASs)
  - edges mean two ASs are connected by a "link"
    - a link can actually represent multiple connections

Example

$$\mathcal{N} = \{1,2,3,4,5,6\}$$
$$\begin{aligned}\mathcal{E} = \{ &(1,2),\\ &(1,3),\\ &(2,6),\\ &(3,4),\\ &(3,5),\\ &(3,6),\\ &(4,6),\\ &(5,6)\}\end{aligned}$$



Measuring Graphs

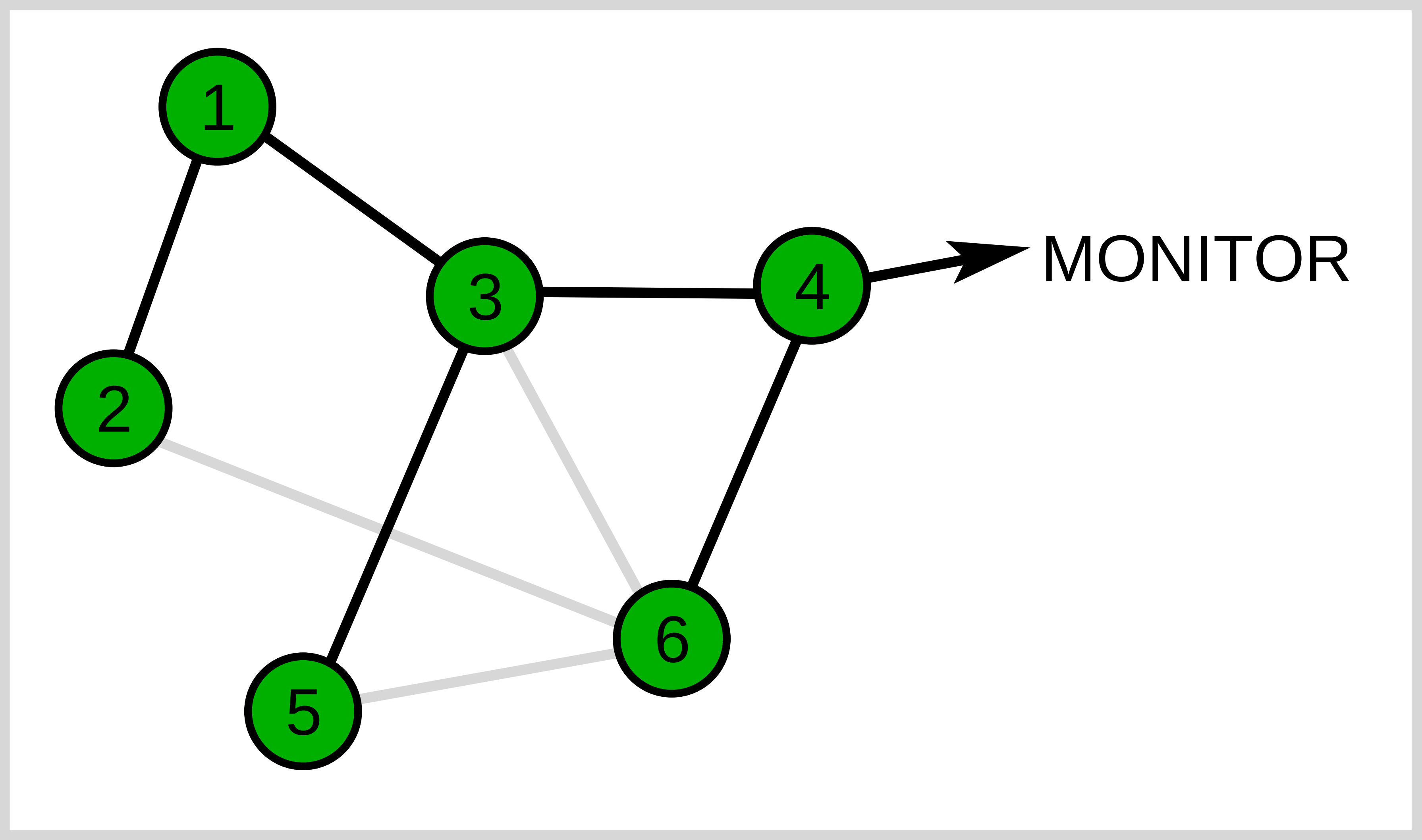
- We often want to measure a graph
  - structure of graph can tell us something
  - graph might be used later (e.g. to predict paths)
- Measurements in the Internet
  - tomography
  - traceroute
  - route monitors
- All measurements have problems
  - we'll focus on route monitors here
  - provide the most up to date information
  - can see dynamics

Route monitors

- Install our own "node"
  - listens for routing messages
  - can infer some of the routes in the network
  - each route tells us about some links
- Problem
  - missing links
  - a single viewpoint only sees a subset of links
  - multiple viewpoints increase coverage
  - how many are enough?
  - how do we know what we are missing?

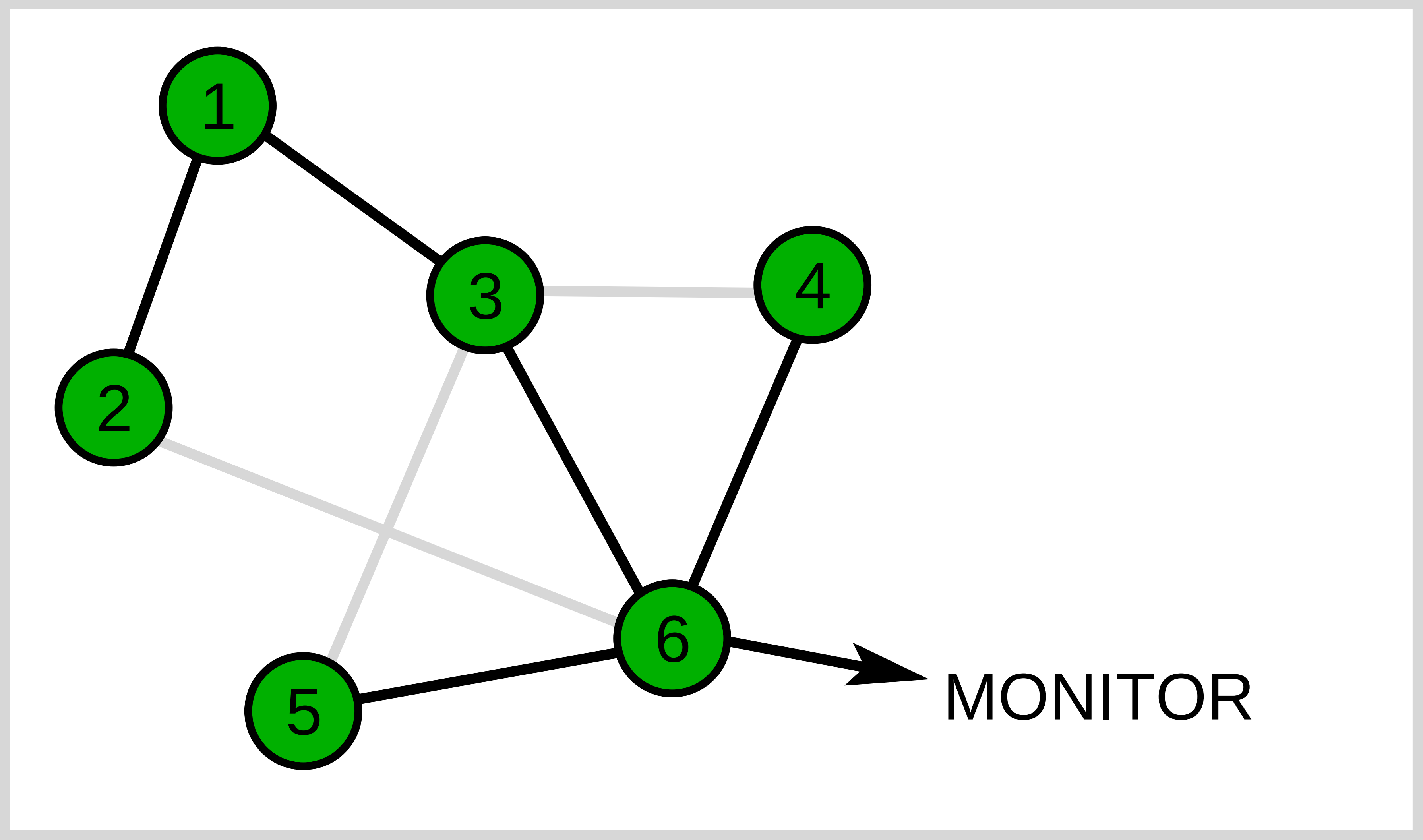
Example

Monitor 1



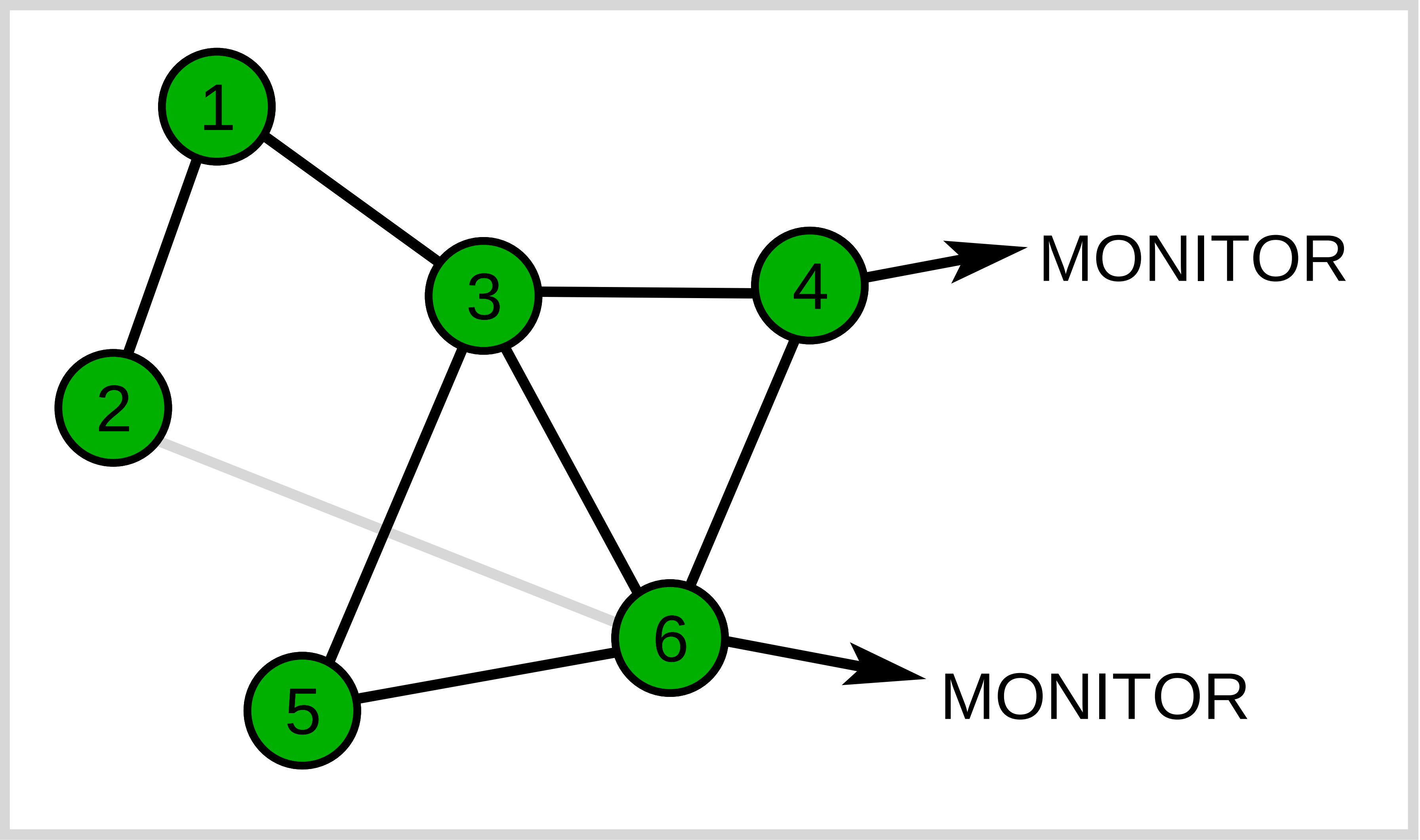
Example

Monitor 2



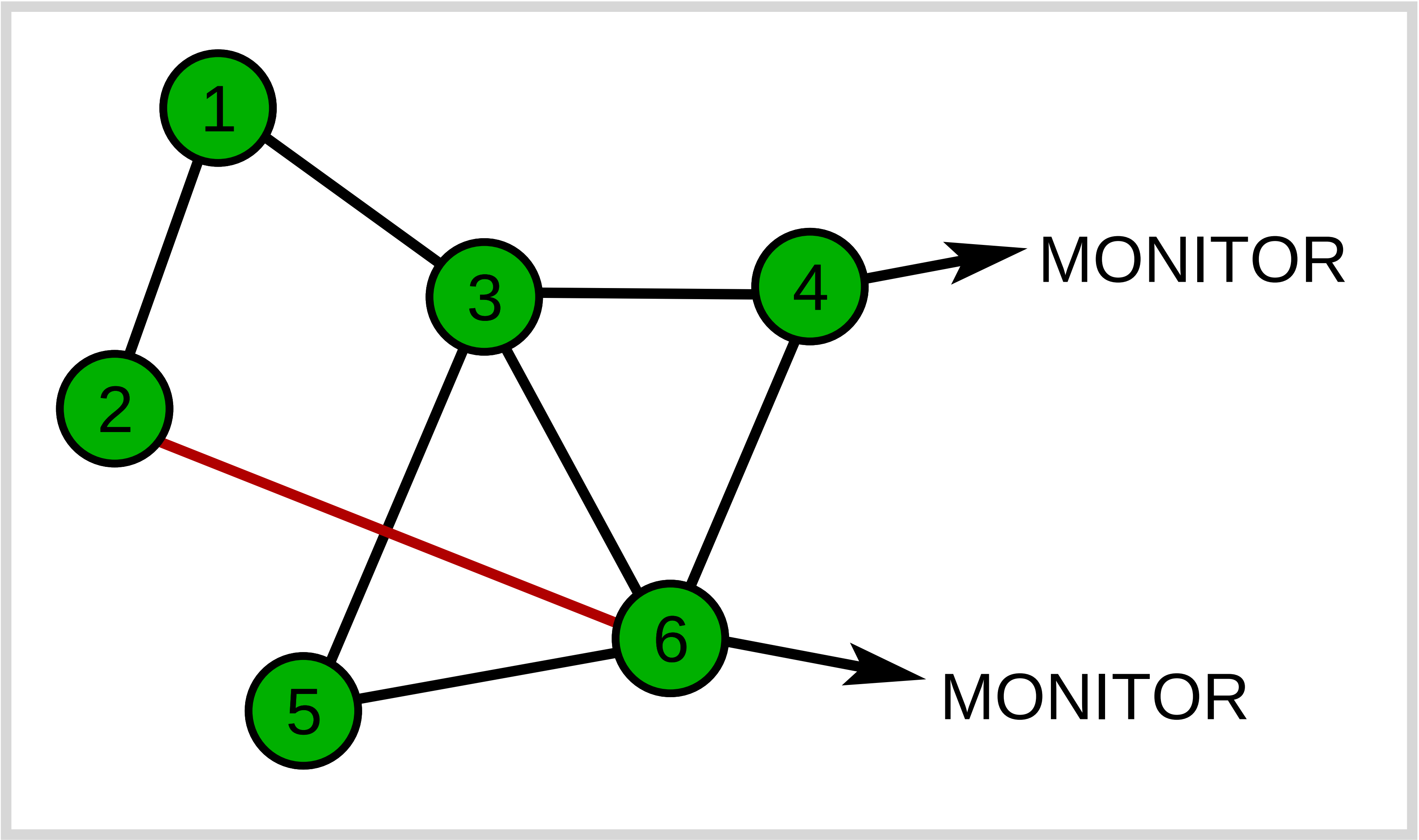
Example

Both monitors

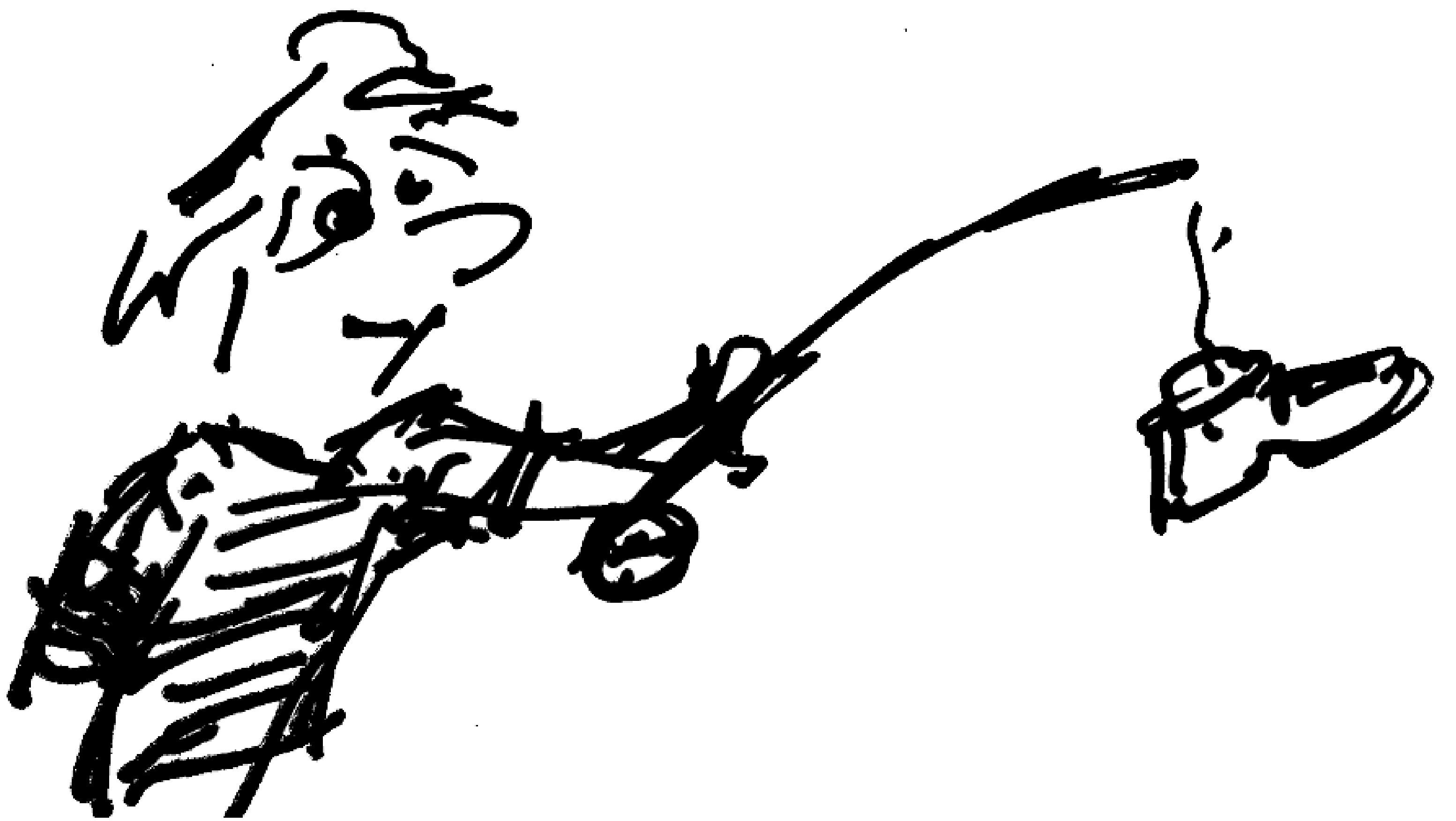


Example

Missing links



Capture-recapture



How many fish are there in the lake?

Standard biological approach

- capture a group of fish, tag them, and release
- some time later
  - capture another group of fish
  - note how many are tagged

Petersen's formula

$$\hat{E} = \frac{E_1 E_2}{E_{12}}$$

where

$$\begin{aligned} E_1 &= \text{the number of "fish" seen in capture 1} \\ E_2 &= \text{the number of "fish" seen in capture 2} \\ E_{12} &= \text{the number of tagged "fish" seen in capture 2} \\ \hat{E} &= \text{the estimated number of "fish" in the pond} \end{aligned}$$

Links = fish

- Capture-recapture Assumptions
  - No change in population over time
  - Tags don't fall off
  - Homogeneity: all fish are the same
  - Independence between experiments
- In our case we want to estimate links
  - number of links = number of fish
  - don't perform successive experiments
  - each monitor is a separate measurement
    - don't need tags because links have unique ID
    - we have $K \simeq 40$ monitors

But it doesn't work!

- Produces inaccurate estimates
  - below a lower bound
- Assumptions of Petersen aren't valid:
  - **links in AS-graph aren't homogeneous**
  - **P2P and C-P links have different visibility**
- propose a stratified model
  - C different classes of links
  - probability of class j is w_j
  - observation probability of class j is p_j

New model

- New model is called a **Binomial Mixture Model**
- We actually observe a truncated version of this model.
- We have a new EM algorithm for estimating the parameters w_j and p_j for a given number of classes.

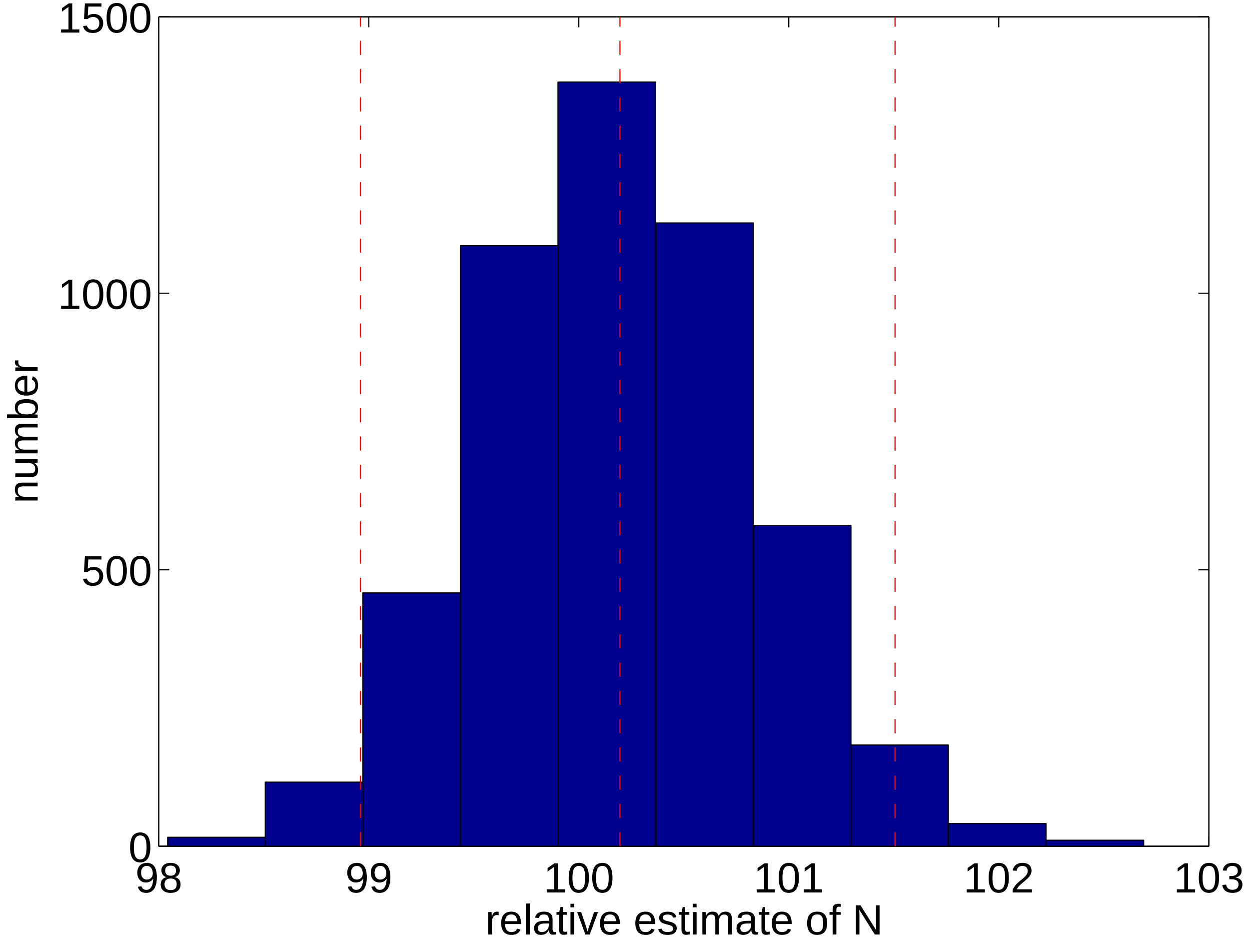
Simulations

Parameters, $C = 7$

Class	Parameter	
	p_j	w_j
1	0.010906	0.248714
2	0.140579	0.052389
3	0.345960	0.036864
4	0.557597	0.049963
5	0.758552	0.060776
6	0.917098	0.068741
7	0.998352	0.482553

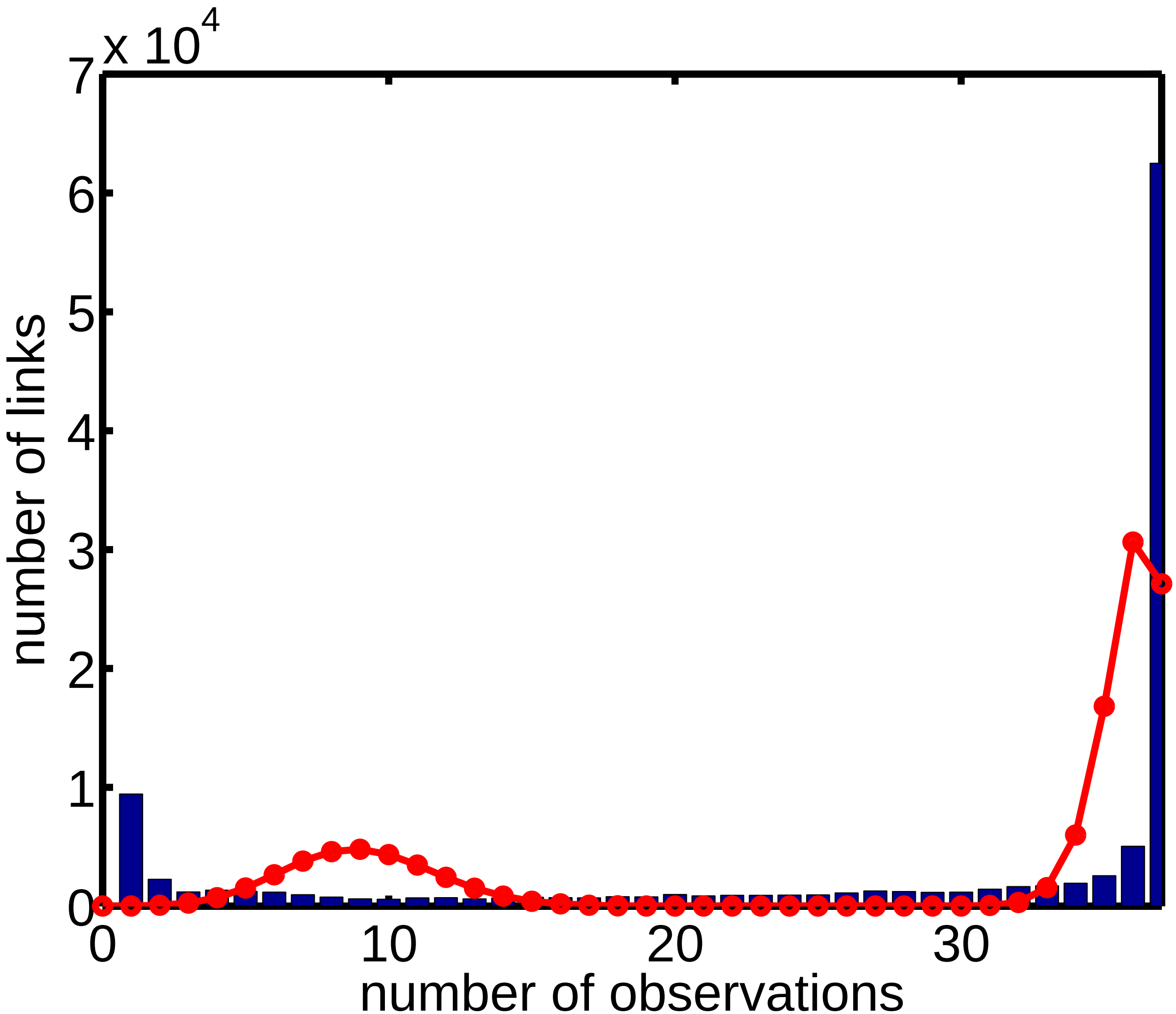
Performance of EM Algorithm

Simulated performance:

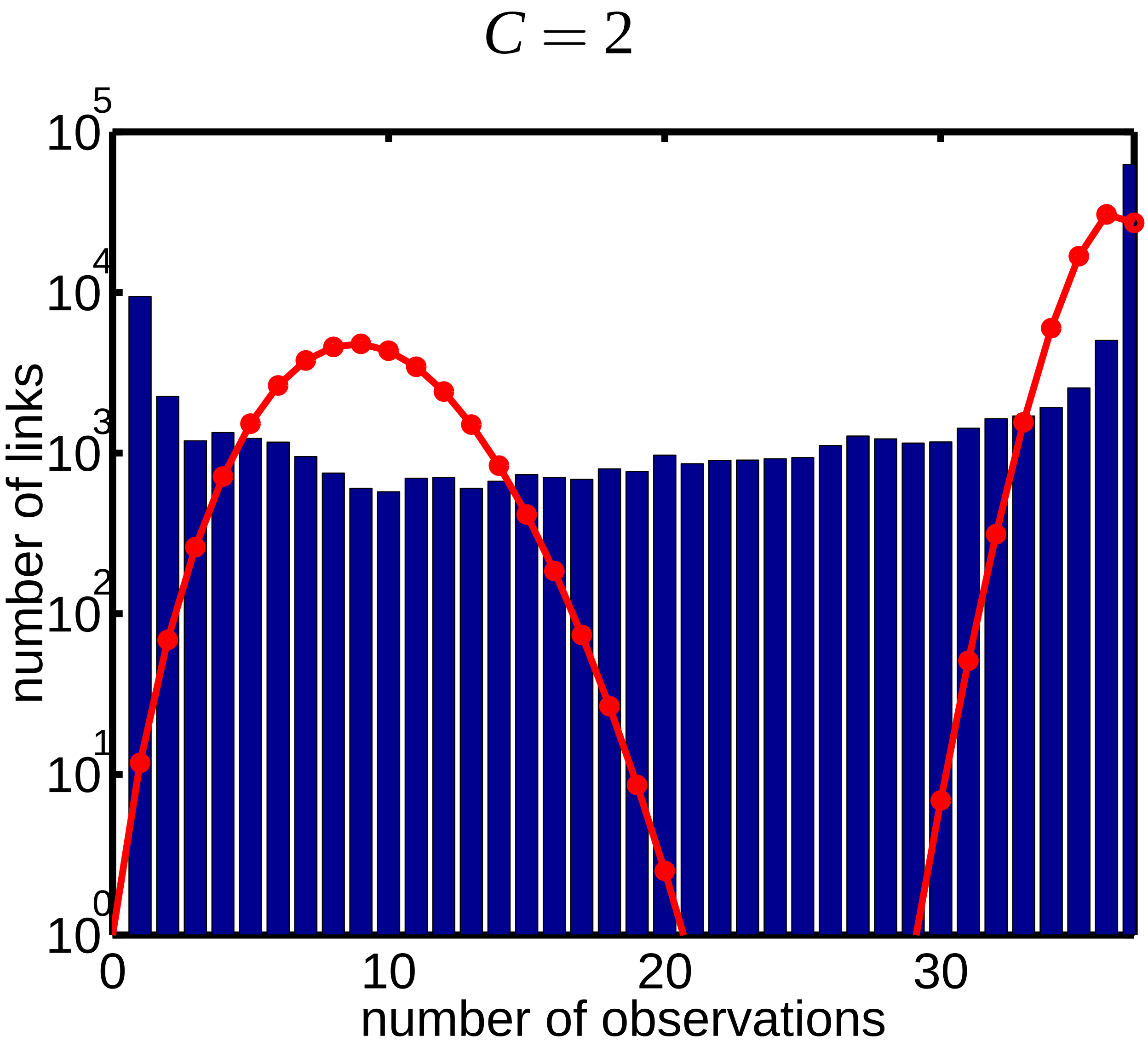


Choice of C

Need to choose C for real data

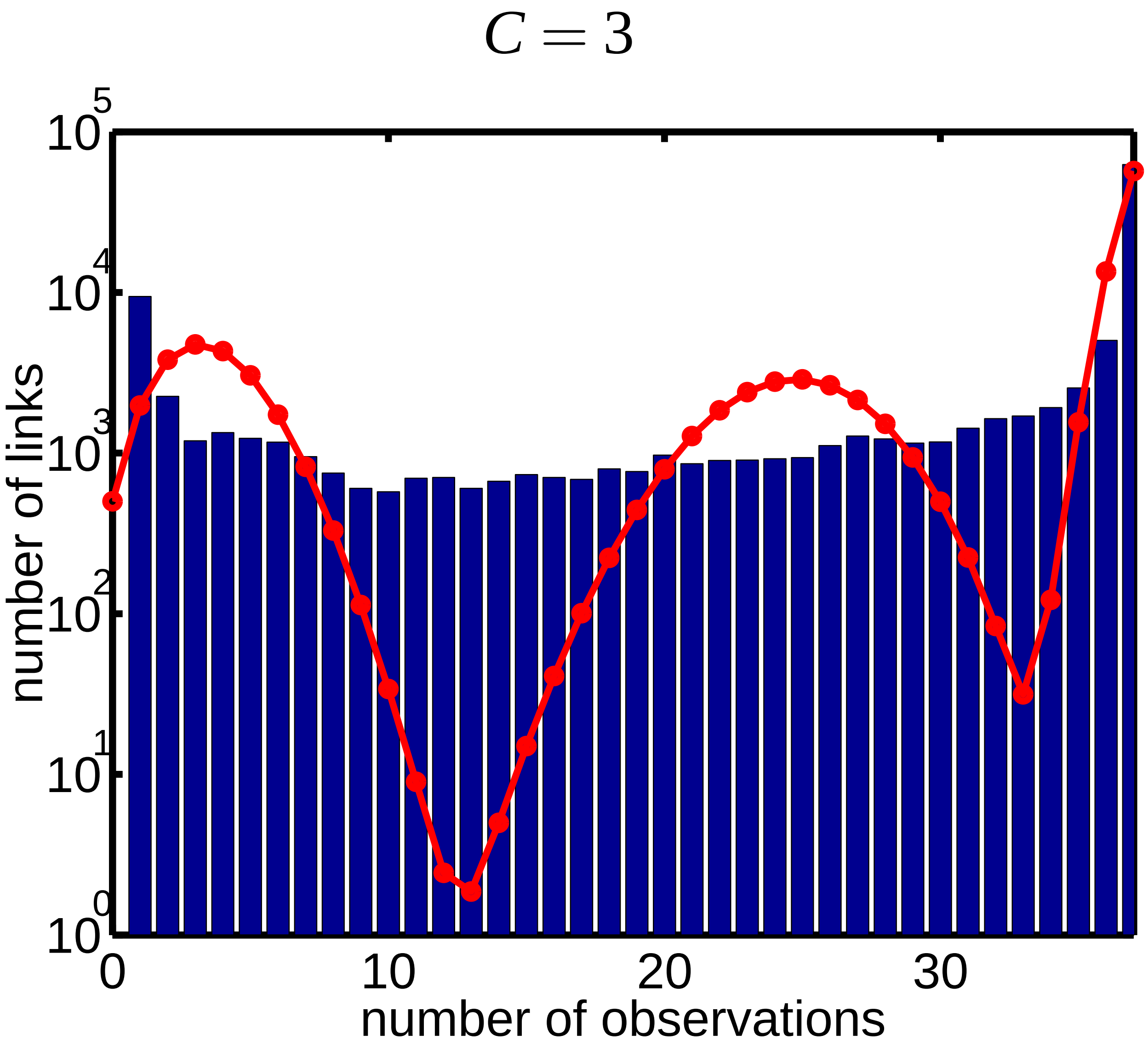


Choice of C



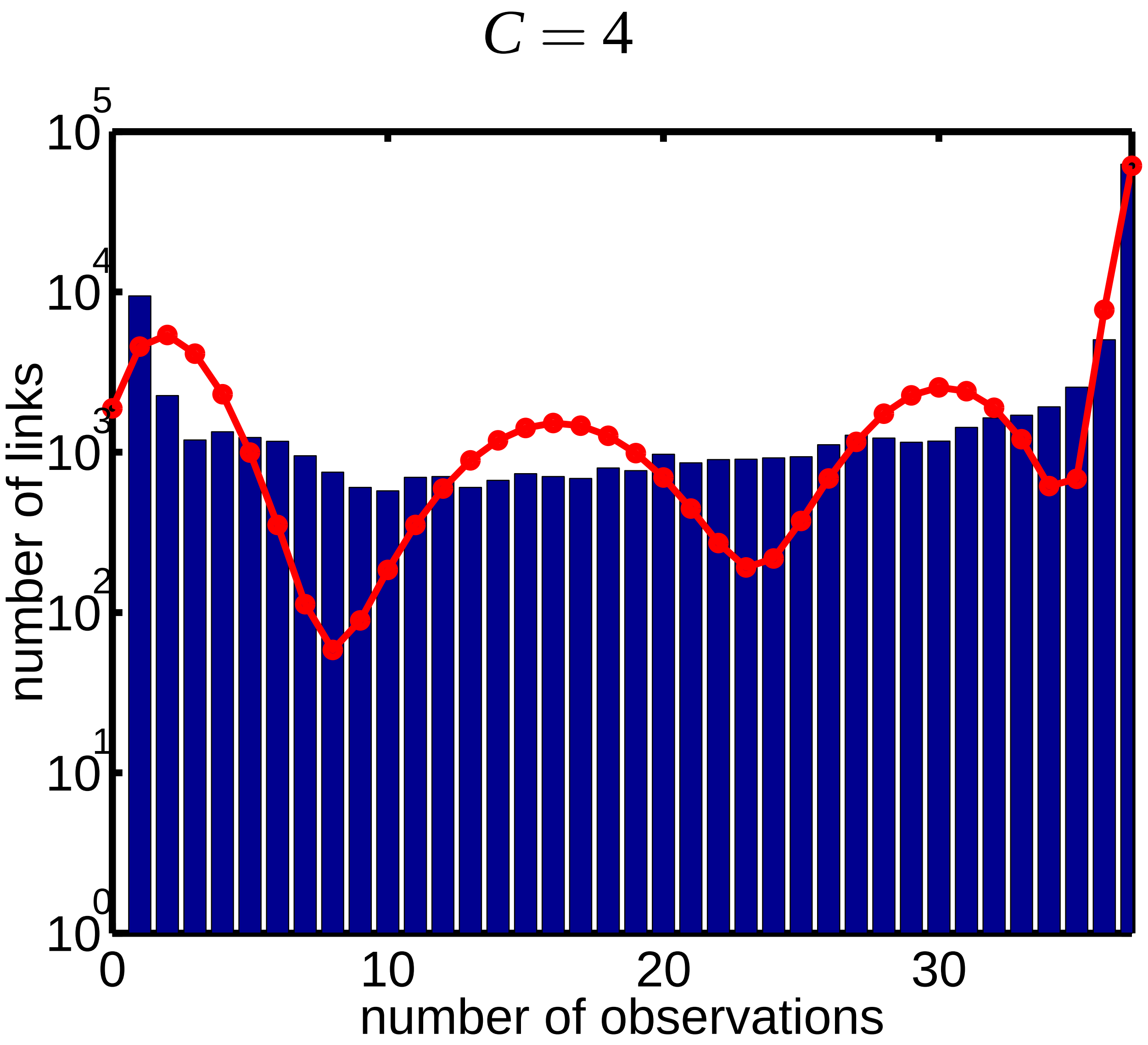
$C = 2$
10^5
10^4
10^3
10^2
10^1
10^0
number of links
0
10
20
30
number of observations

Choice of C



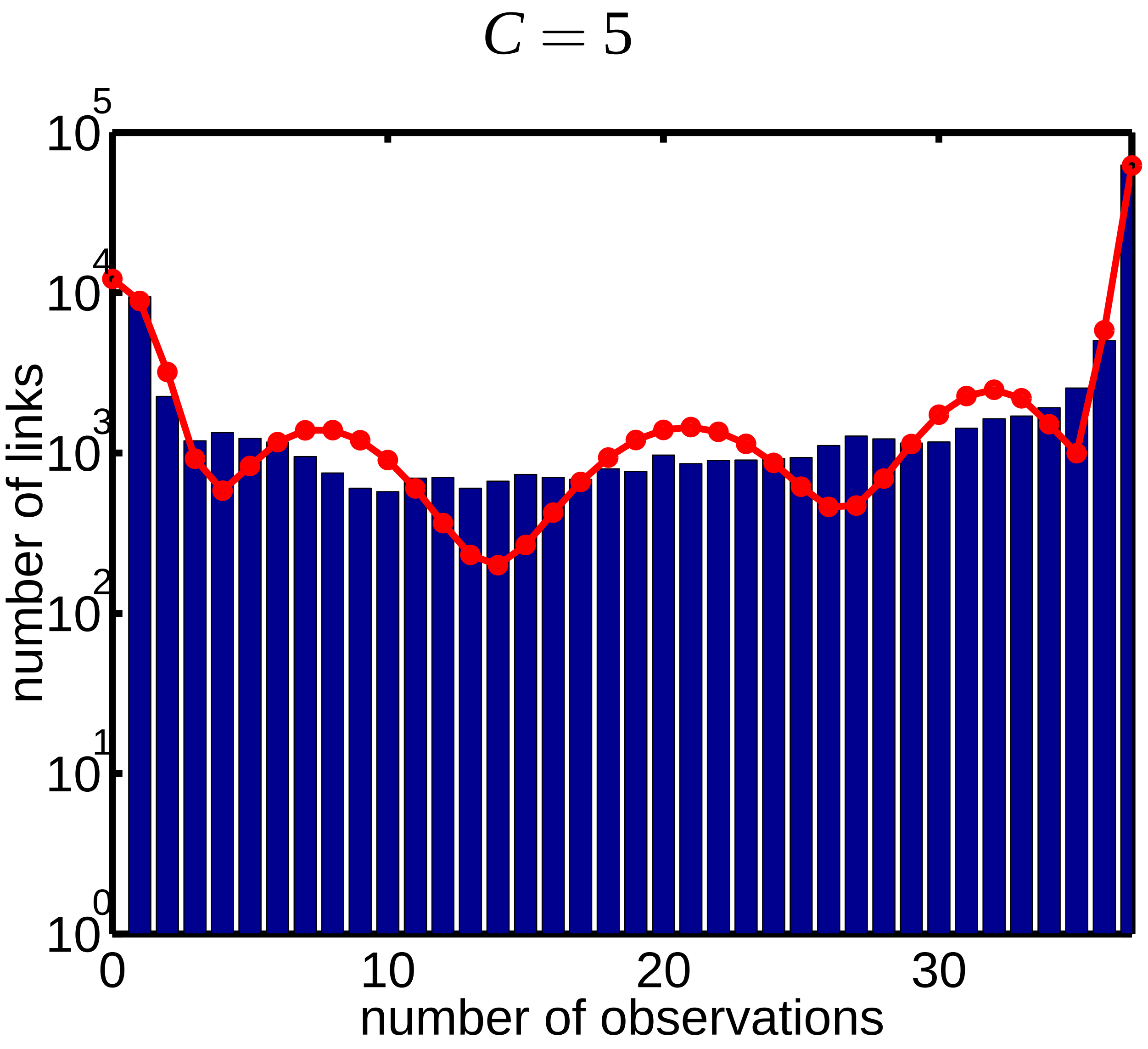
$C = 3$
number of links
number of observations
10^5
10^4
10^3
10^2
10^1
10^0
0
10
20
30

Choice of C



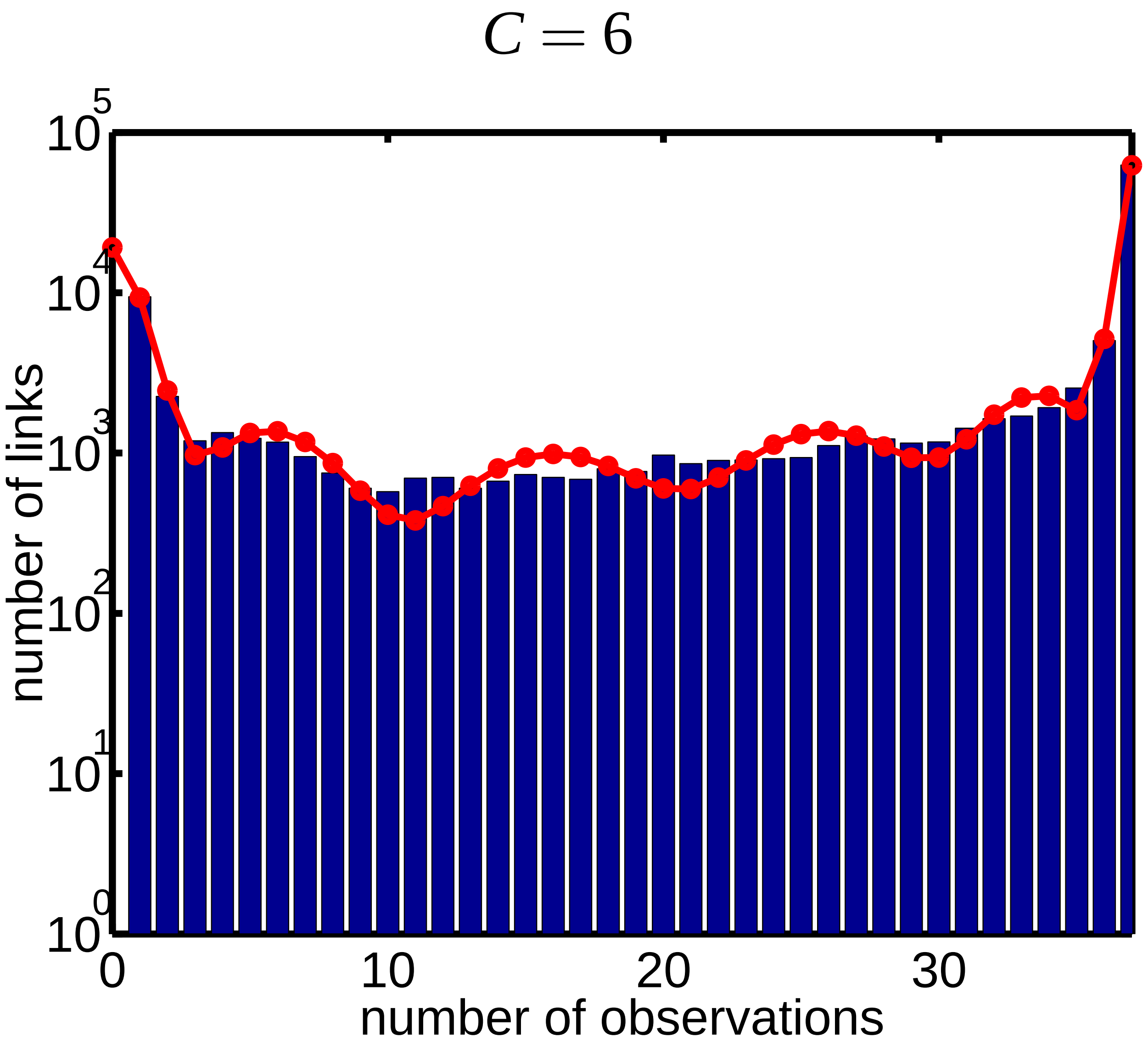
C = 4
10^5
10^4
10^3
10^2
10^1
10^0
number of links
0
10
20
30
number of observations

Choice of C



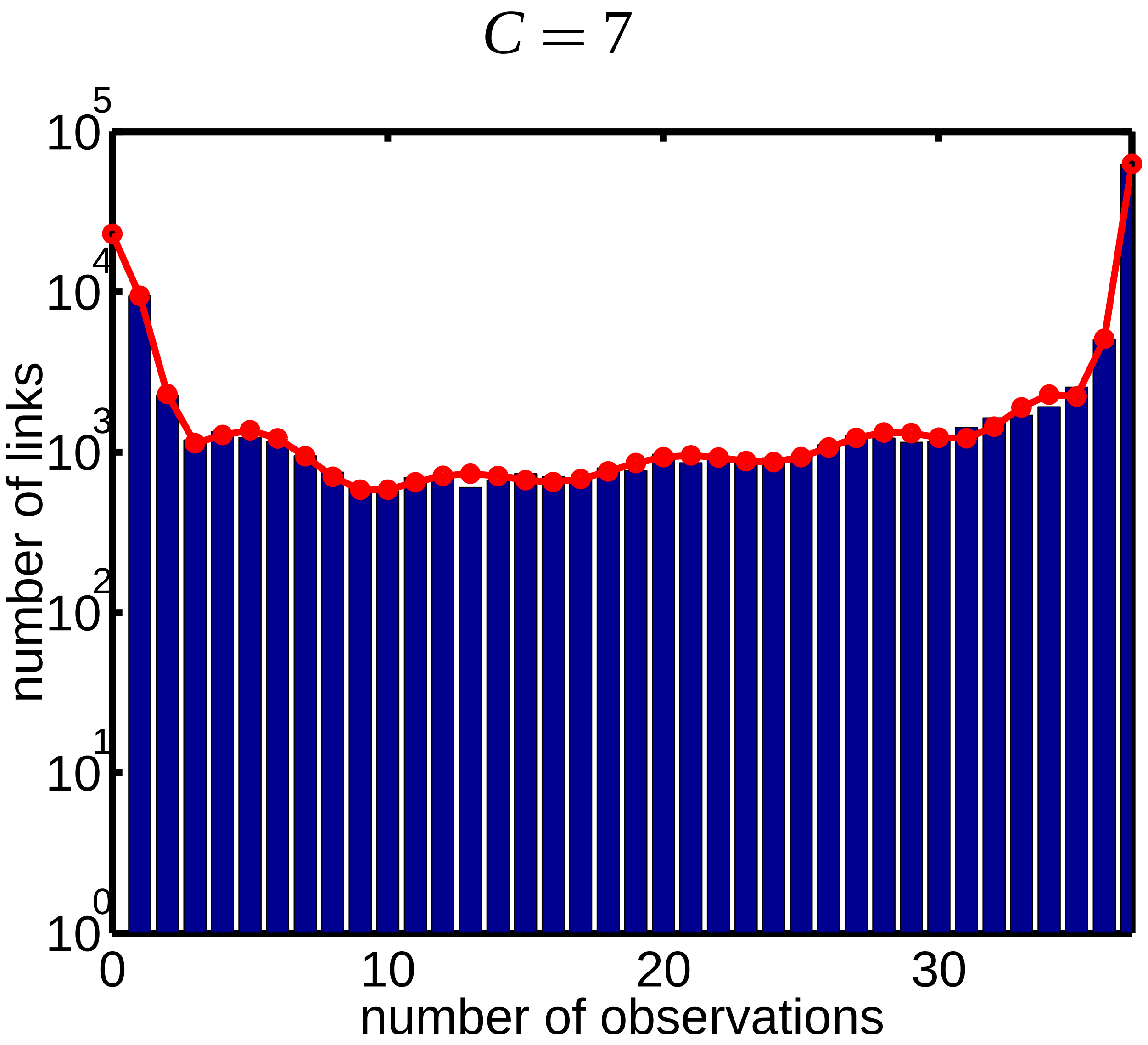
$C = 5$
10^5
10^4
10^3
10^2
10^1
10^0
number of links
0
10
20
30
number of observations

Choice of C



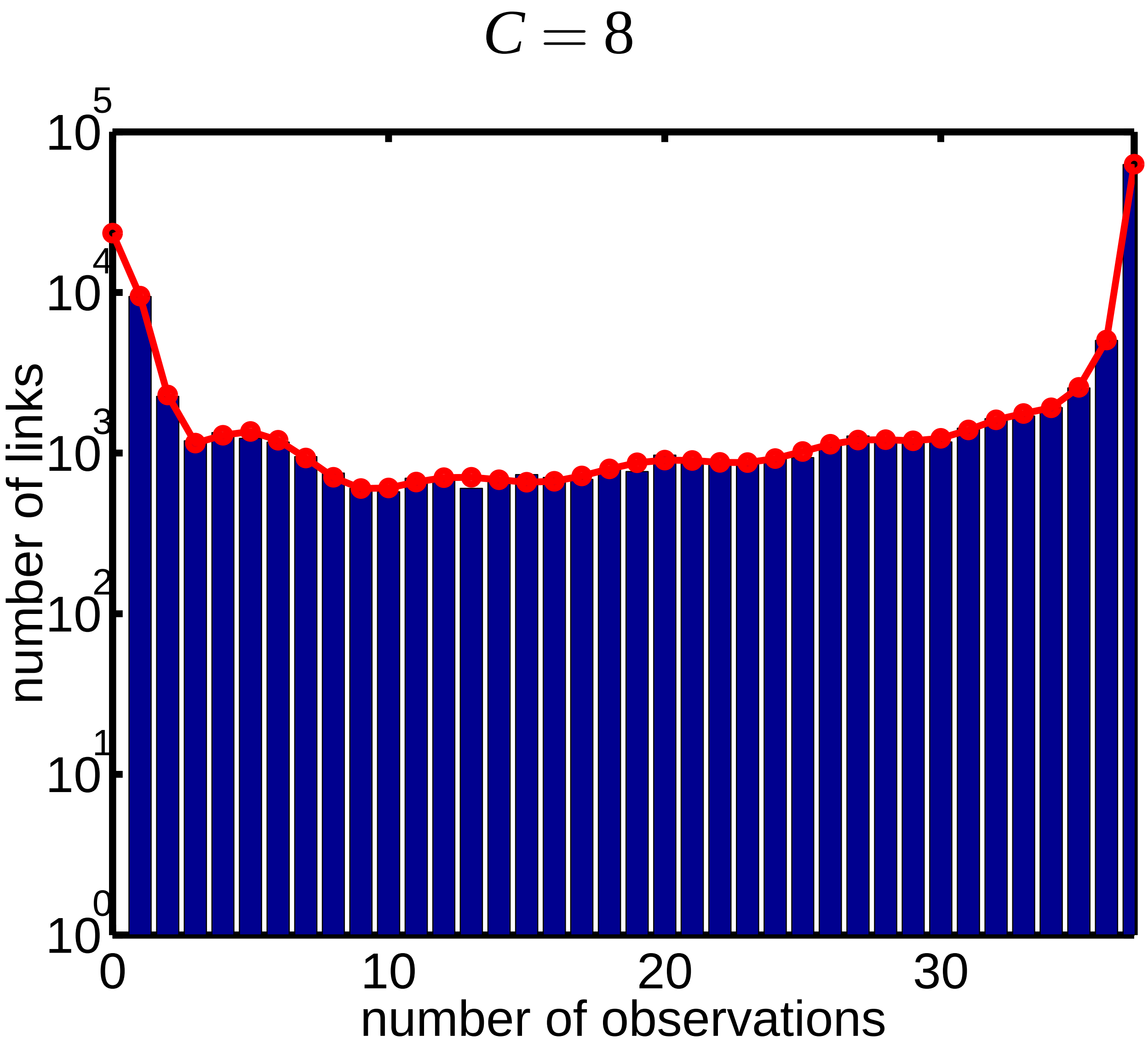
$C = 6$
10^5
10^4
10^3
10^2
10^1
10^0
number of links
0
10
20
30
number of observations

Choice of C



$C = 7$
number of links
number of observations
10^5
10^4
10^3
10^2
10^1
10^0
0
10
20
30

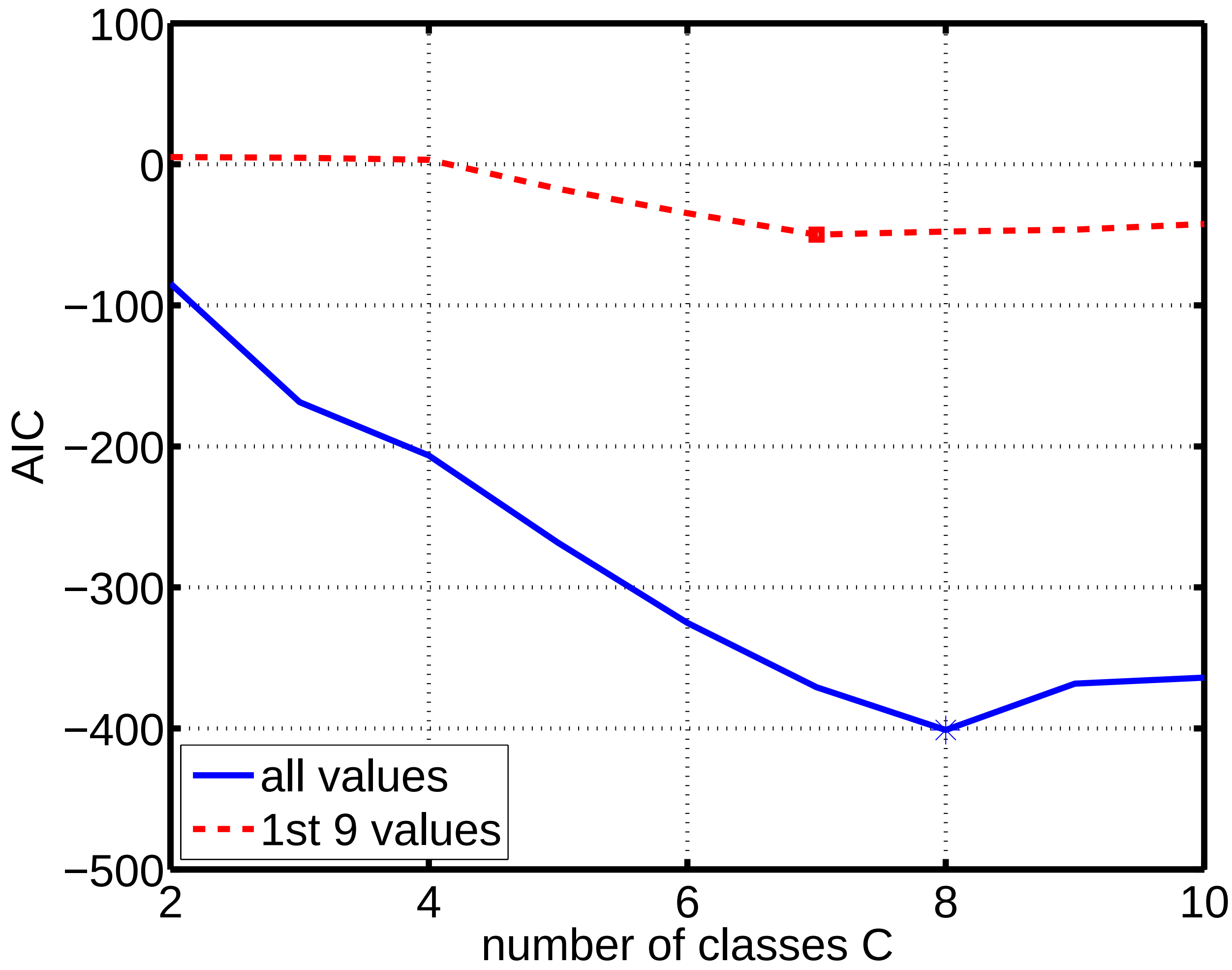
Choice of C



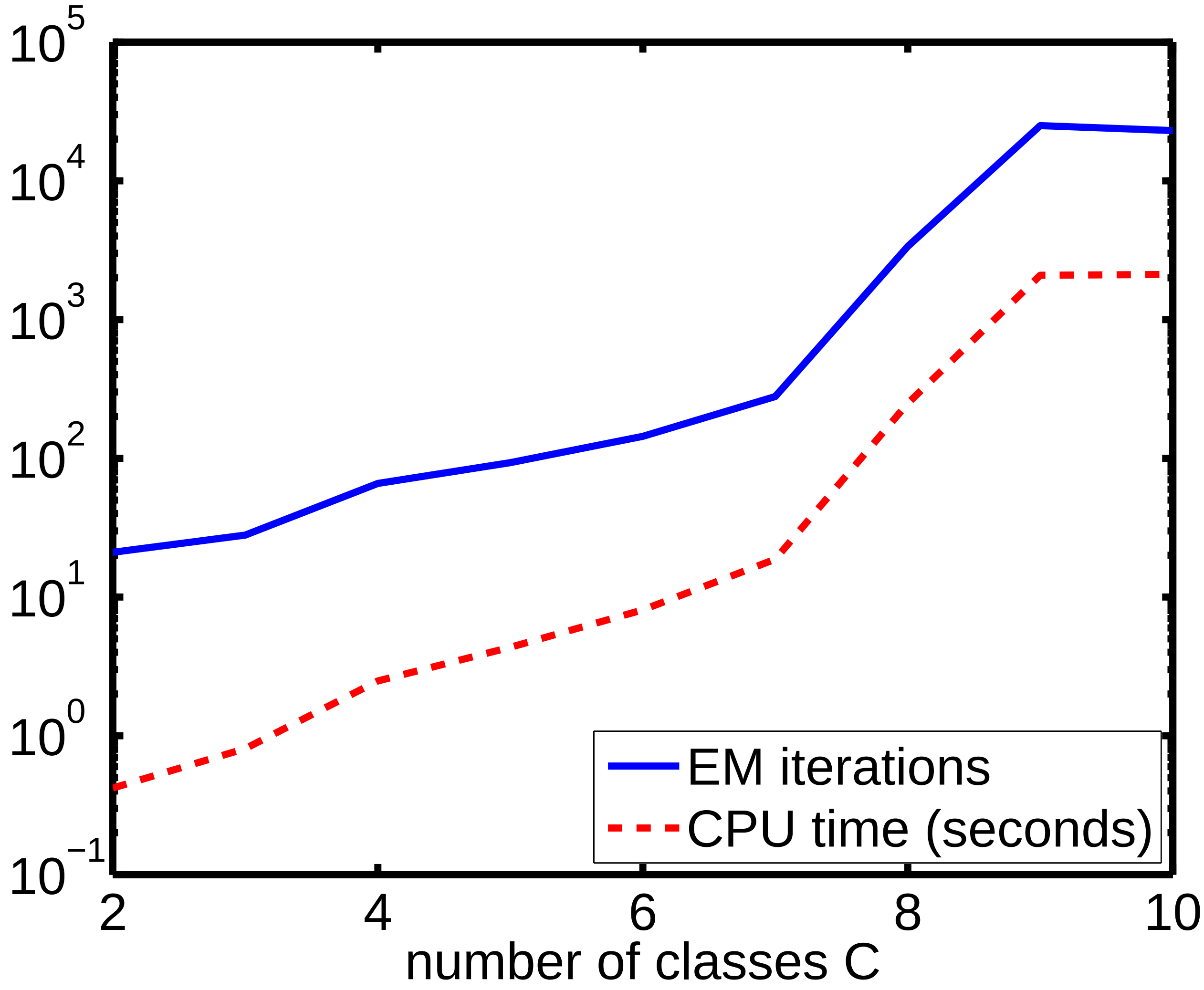
$C = 8$
10^5
10^4
10^3
10^2
10^1
10^0
number of links
0
10
20
30
number of observations

Systematic choice of C

Akaike's Information Criteria $= n[\ln(2\pi \mathrm{RSS}/n) + 1] + 2C,$



Workload



10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
2
4
6
8
10
number of classes C
EM iterations
CPU time (seconds)

Previous studies

Paper	label	date	$\hat{E}$
Zhang et al. [1]	Updates(1M)	2004-10-24	55,388
He et al. [2]	All	2005-05-12	59,500
Mühlbauer et al. [3]	N/A	2005-11-13	58,903

References

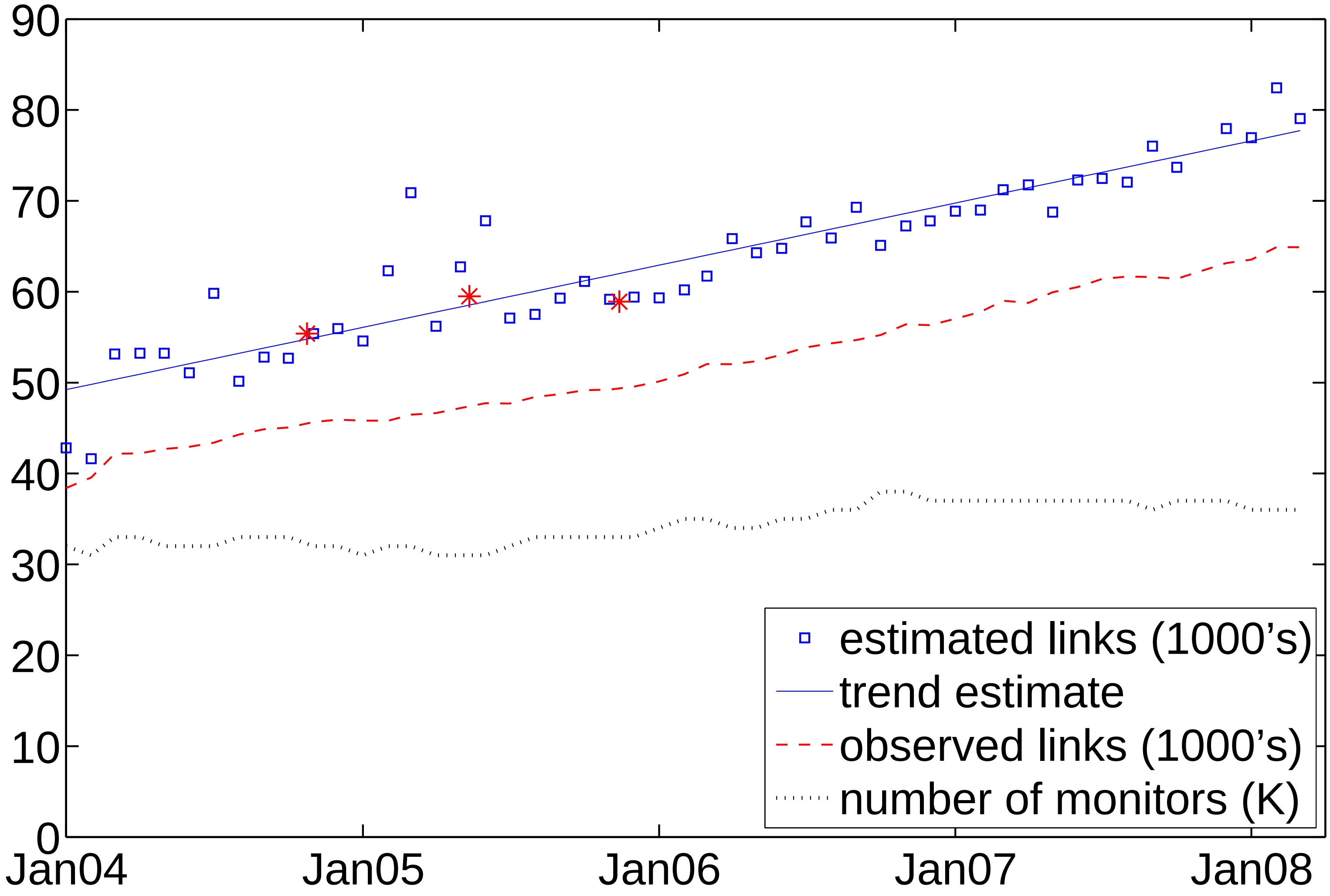
[1] B. Zhang, R. Liu, D. Massey, and L. Zhang, "Collecting the Internet AS-level topology," *ACM SIGCOMM Computer Communication Review (CCR) special issue on Internet Vital Statistics*, January 2005.

[2] Y. He, G. Siganos, M. Faloutsos, and S. V. Krishnamurthy, "A systematic framework for unearthing the missing links: Measurements and impact," in *USENIX/SIGCOMM NSDI*, (Cambridge, MA, USA), April 2007.

[3] W. Mühlbauer, A. Feldmann, M. R. O. Maennel, and S. Uhlig, "Building an AS-topology model that captures route diversity," in *ACM SIGCOMM*, (Pisa, Italy), 2006.

Results: $C = 7$

Monthly data since January 2004.



Conclusion

- Method for estimating how much we don't know
- Used it to study the AS graph
  - Potential improvements
    - account for monitor dependencies
    - account for heterogeneity amongst monitors
- There still might be something missing – what about a class of links that we **never** observe?
- Much wider applicability
  - Social networks?
  - Network Dynamics

Truncated binomial

Using same assumptions as Petersen's the number of observations k of a link will follow a Binomial distribution

$$\text{prob}\{k\} = \binom{K}{k} p^k (1-p)^{(K-k)}$$

However, we only observe a link if $k > 0$, so we observe the conditional distribution

$$\text{prob}\{k | k > 0\} = \binom{K}{k} \frac{p^k (1-p)^{(K-k)}}{1-(1-p)^K}$$

which is a truncated Binomial distribution.

Estimator

MLE (Maximum Likelihood Estimator) $\hat{p}$ has to satisfy

$$E_{\text{obs}} K \hat{p} = [1 - (1-\hat{p})^K] \sum_{i=1}^{E_{\text{obs}}} k_i$$

where

$$\begin{aligned} K &= \text{the number of monitors} \\ E_{\text{obs}} &= \text{the number of observed links (via all monitors)} \\ k_i &= \text{the number of observations of the } i\text{th link} \\ \hat{p} &= \text{the MLE estimator of the observation probability } p \end{aligned}$$

Estimator

MLE (Maximum Likelihood Estimator) $\hat{p}$ has to satisfy

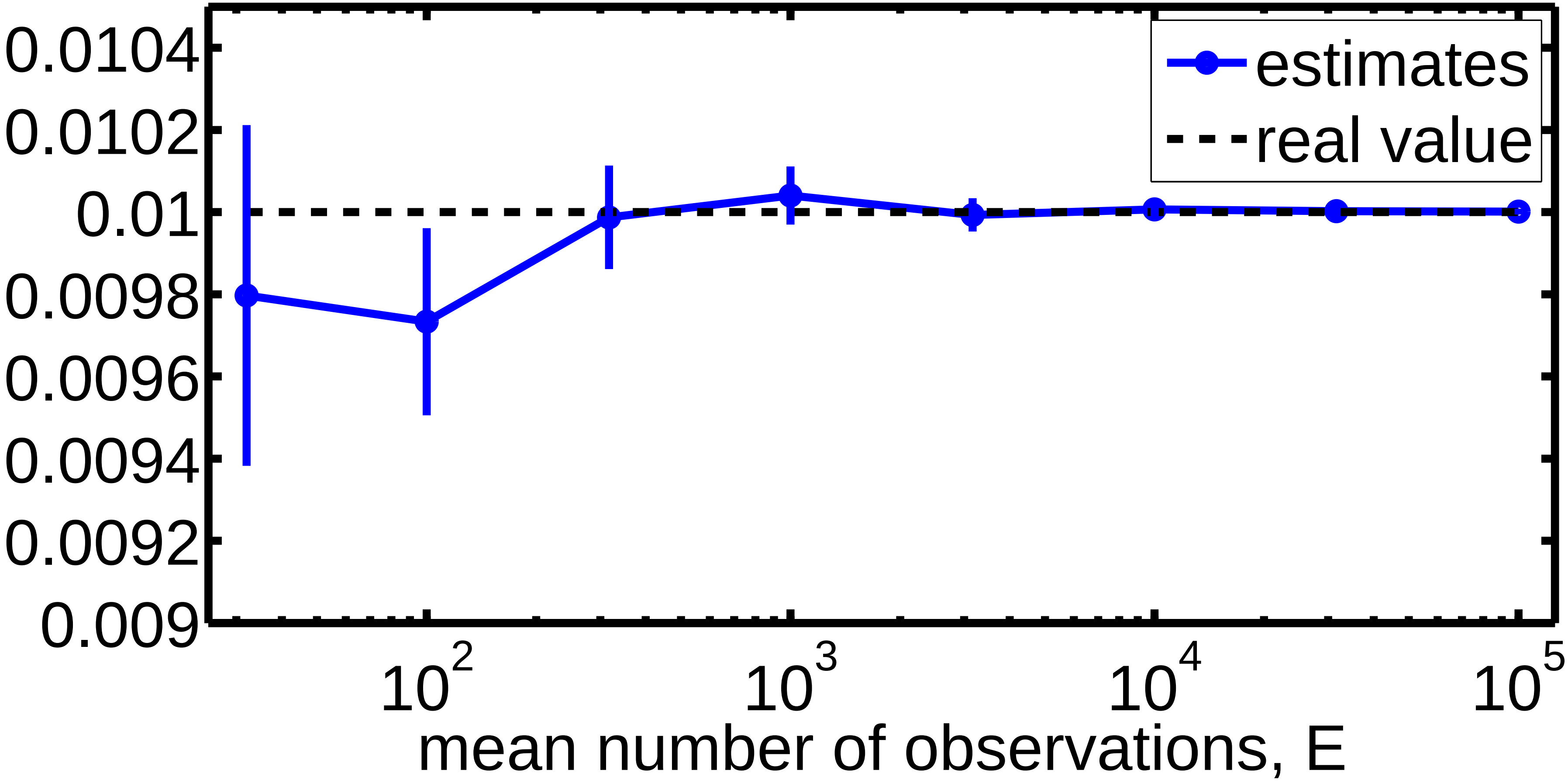
$$E_{\text{obs}} K p = \left[1 - (1-p)^K\right] \sum_{i=1}^{E_{\text{obs}}} k_i$$

Solution by repeated substitution

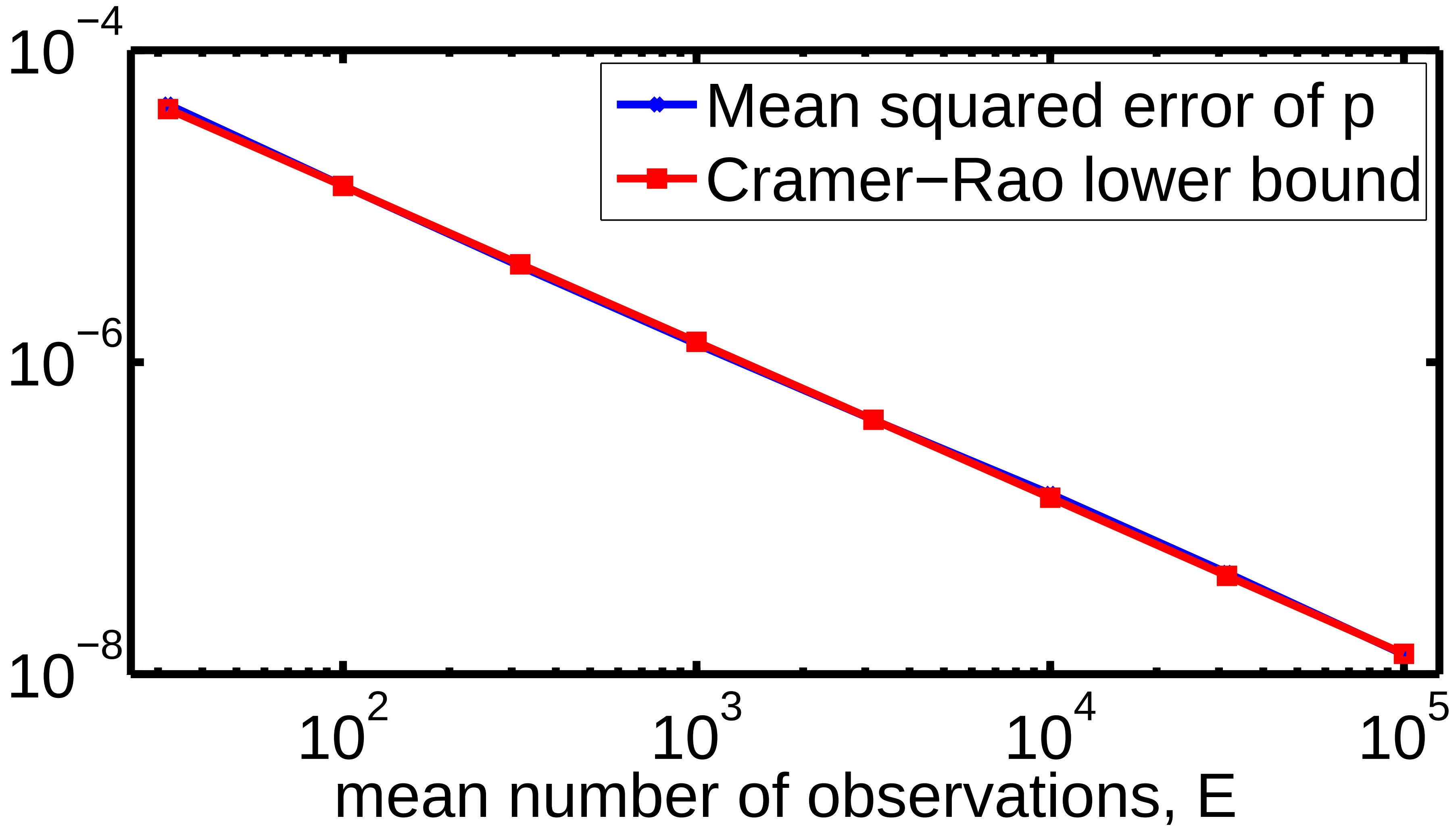
$$\begin{aligned} \hat{p}_0 &= \frac{\sum_{i=1}^{E_{\text{obs}}} k_i}{E_{\text{obs}} K} \\ \hat{p}_{i+1} &= \frac{\sum_{i=1}^{E_{\text{obs}}} k_i}{E_{\text{obs}} K} \left[1 - (1-\hat{p}_i)^K\right] \end{aligned}$$

Can prove that this converges to a fixed point of the above equation.

Simulated estimates $\hat{p}$



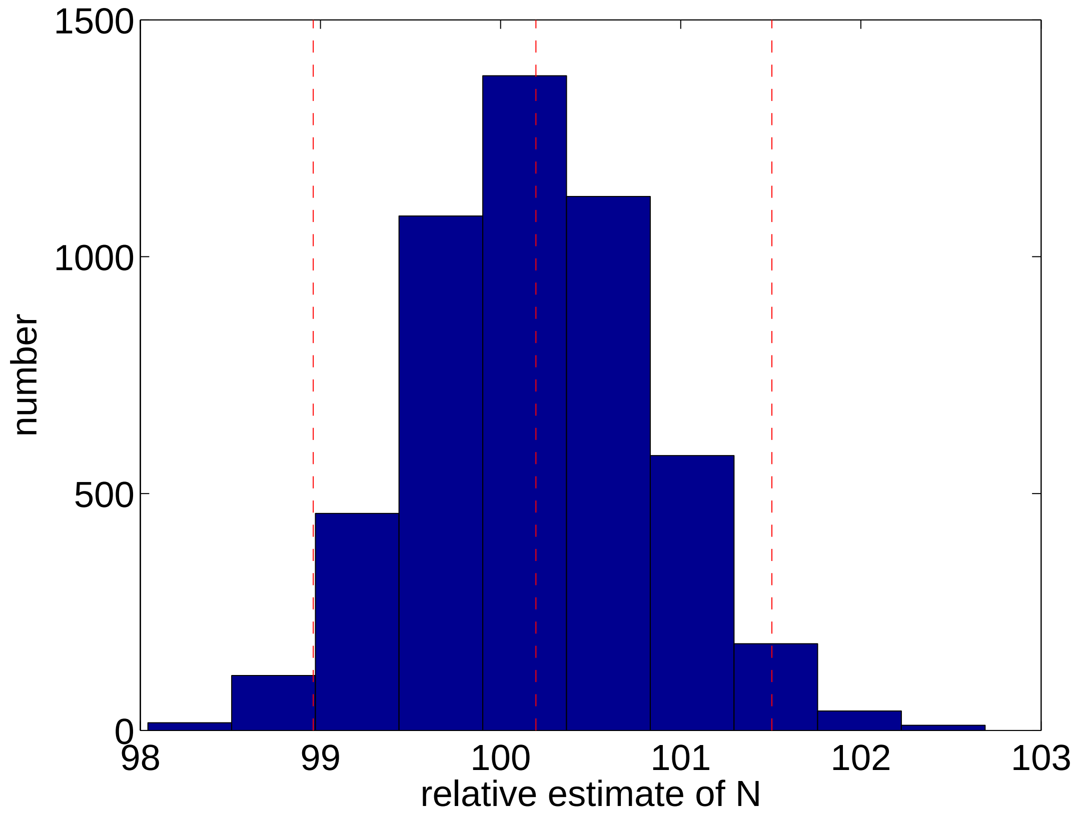
Variance of $\hat{p}$



Estimator $\hat{E}$

Once we know p, then MLE for E is

$$\hat{E} = \frac{E_{\text{obs}}}{1-(1-\hat{p})^K}$$



New model

Binomial mixture model

- probability of class j is w_j
- Binomial distribution $B(K, p_j)$ for each class

Distribution function

$$\text{prob}\{k\} = \sum_{j=1}^{C} w_j \binom{K}{k} p_j^k (1 - p_j)^{(K-k)}$$

Of course, we observe a truncated version of this.

EM Algorithm

```
While (not converged 1) do
 E step:
  estimate c_j^(i)
```

$$c_j^{(i)} \leftarrow \hat{w}_j P\{k_i | K, \hat{p}_j\}$$

```
 M step:
  for j=1 to C
   While (not converged 2) do
```

$$\hat{p}_j \leftarrow \frac{\sum_i k_i c_j^{(i)}}{K \sum_i c_j^{(i)}} [1 - (1 - \hat{p}_j)^K]$$

```
   end while 2
```

$$\hat{w}_j \leftarrow \sum_i c_j^{(i)} / (E(1 - (1 - \hat{p}_j)^K))$$

```
  end for
end while 1
```

Systematic choice of C