The ins and outs of snowball sampling: ERGM estimation for very large directed networks

Alex Stivala David Rolls Garry Robins

Melbourne School of Psychological Sciences, The University of Melbourne, Australia

Sunbelt XXXV, June 23–28, 2015, Brighton, UK

Introduction (1)

- Exponential random graph models (ERGMs) are useful for analyzing social networks.
- But estimating parameters is computationally intensive.
- This restricts the size of the networks that can have an ERGM fitted, to a few thousand nodes at most.
- The Markov chain Monte Carlo (MCMC) methods often used are inherently sequential, limiting the use of high performance parallel computing.

Introduction (2)

- Previously, for undirected networks only, we have overcome this problem by taking multiple snowball samples, estimating ERGM parameters for each in parallel, and pooling the estimates.
- But snowball sampling does not necessarily make sense for directed networks.
- So what can we do?

Exponential random graph models (ERGMs)

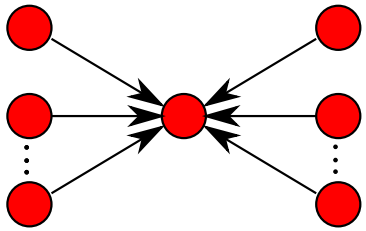
$$\Pr(X = x) = \frac{1}{\kappa} \exp\left(\sum_A \theta_A z_A(x)\right)$$

where

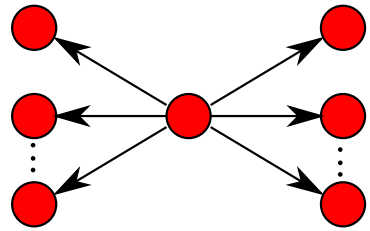
- $X = [X_{ij}]$ is a 0-1 matrix of random tie variables,
- x is a realization of X,
- A is a subgraph configuration,
- $z_A(x)$ is the network statistic for configuration A,
- θ_A is a model parameter corresponding to configuration A,
- κ is a normalizing constant to ensure a proper distribution.

Model configurations — structural

Alternating k-stars: useful for capturing degree distribution

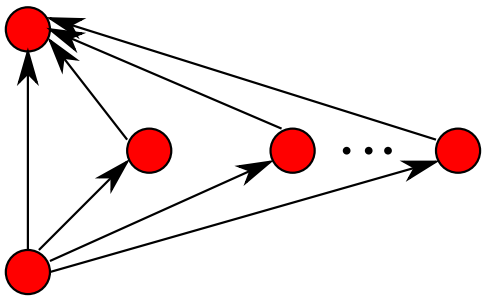


Popularity spread
Alt. in-star
AinS

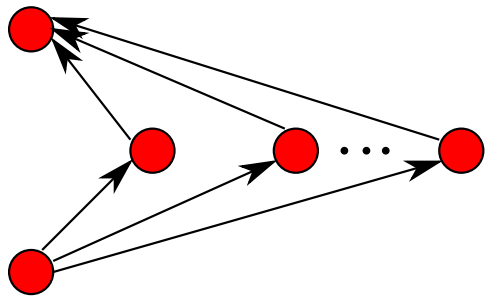


Activity spread
Alt. out-star
AoutS

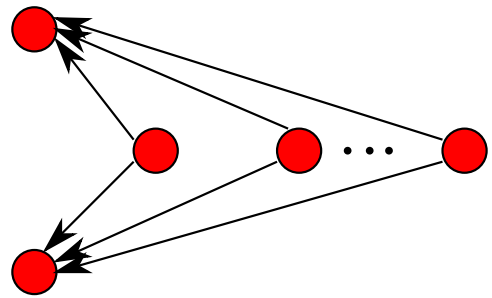
Alternating k-triangles (AT) and k-2-paths (A2P): useful for modelling social circuit dependence



Path closure
AT-T



Multiple 2-paths
A2P-T



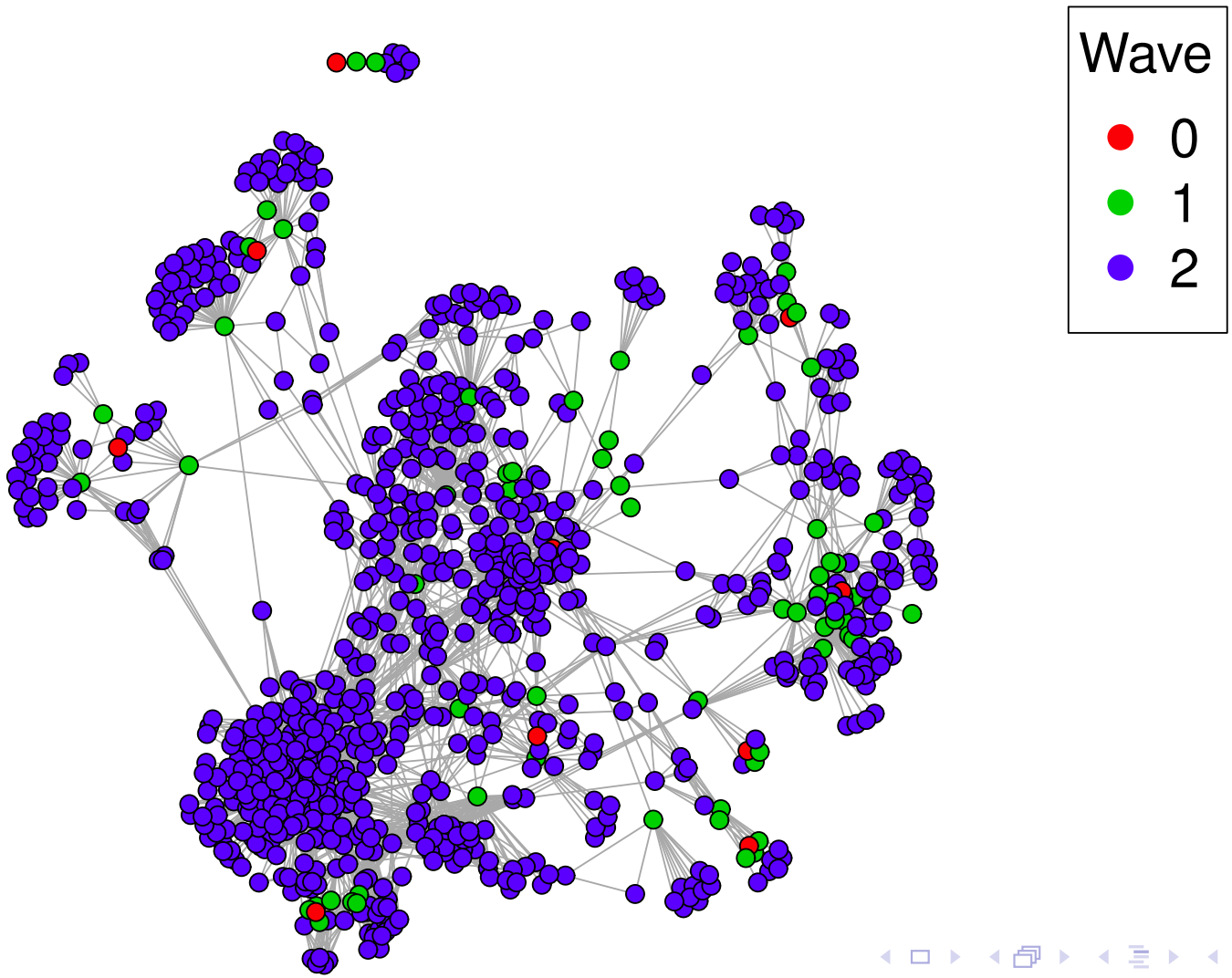
Shared popularity
A2P-D

Snowball sampling

- Start with N_0 seed nodes (wave 0).
- Follow all their ties to get a further set of nodes (wave 1).
- In general, follow the ties from nodes in wave $k-1$ to get the nodes in wave k.
- There is a picture on the next slide...
- We can use conditional estimation procedures (Pattison *et al.* 2013) on each snowball sample.

Snowball sampling example from Nexus condmatcollab2005

Snowball sample ($n = 907$) from condensed matter collaborations network $N = 40421$, 2 waves, 10 seeds.



Pooling the estimates

- The ERGM parameter estimates can be combined using weighted least squares (WLS) or median for the point estimate.
- The error in the combined estimate is estimated using a non-parametric adjusted percentile bootstrap (BCa).
- This uses the estimated standard errors for each snowball sample estimate to adjust for bias and skewness in the bootstrap distribution.

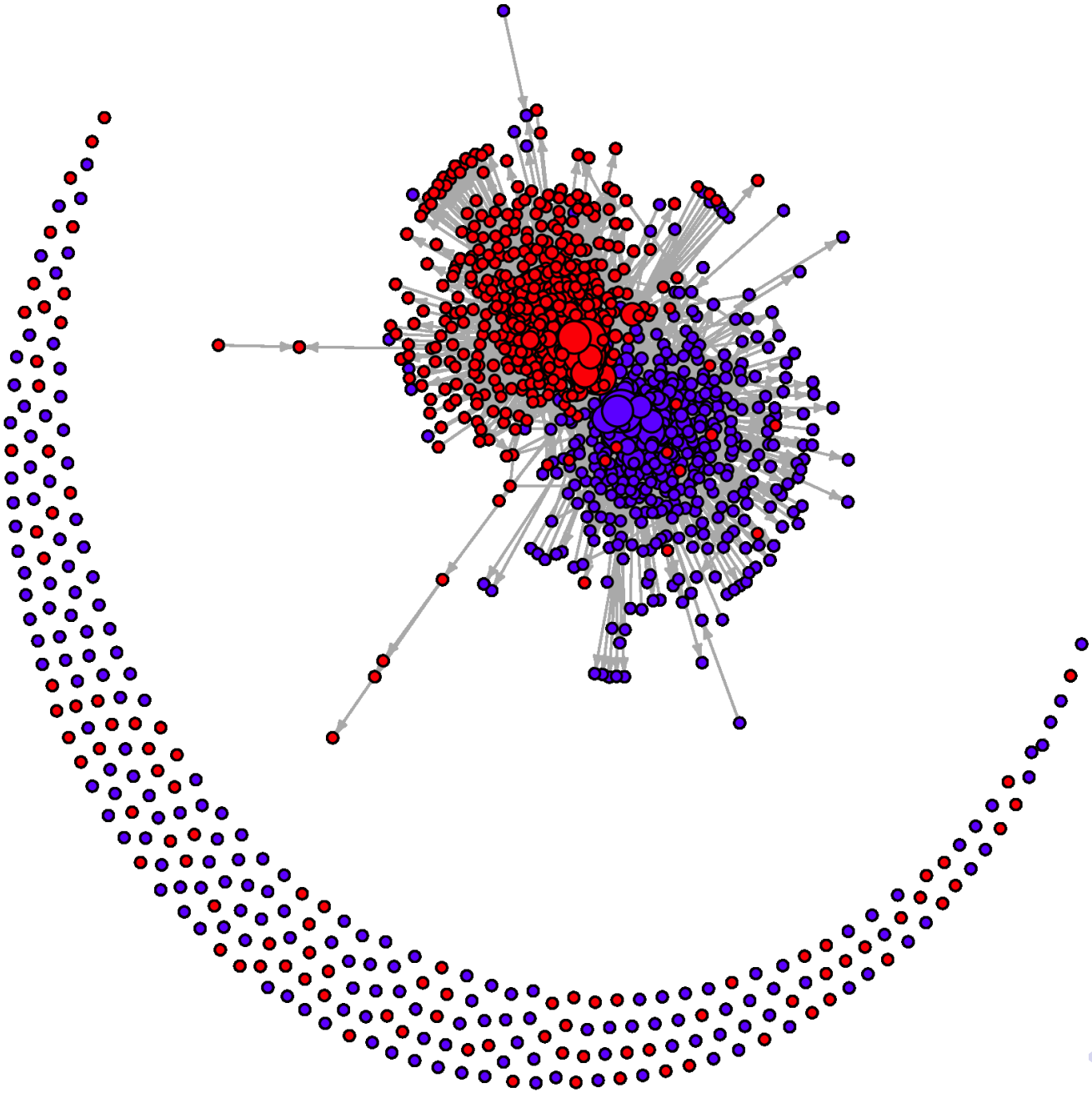
OK, but what about directed networks?

- Snowball sampling works by "following" links starting from the seed nodes.
- But it does not capture inward links, so such links can be missing if unreciprocated.
- If we think of a "usual" case of snowball sampling in an unknown social network, we are only following outgoing links.
- But we are using snowball sampling merely as a computational device to sample in a known network...

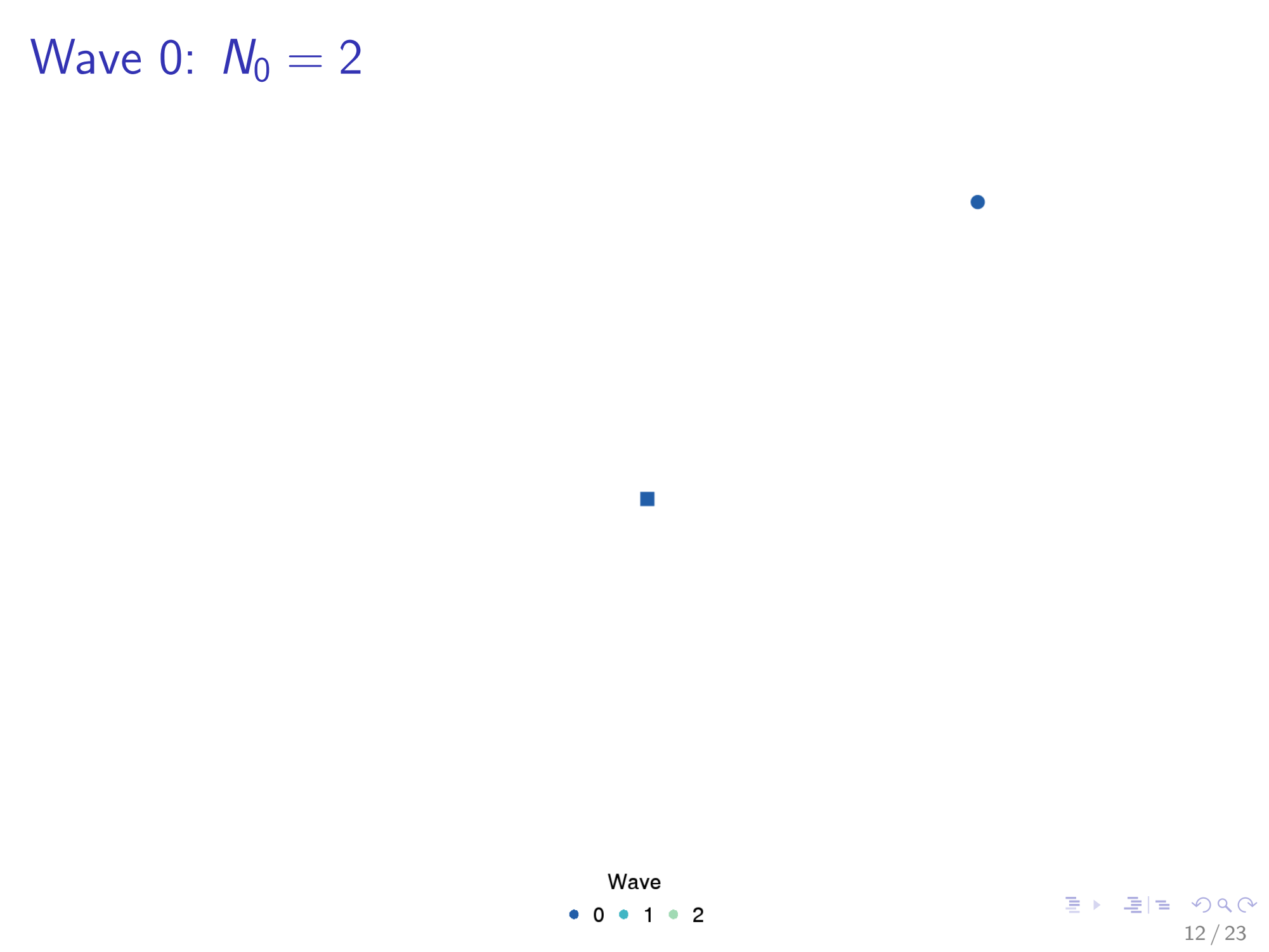
Directed snowball sampling

- So we can pretend all the links are undirected, and snowball sample in the directed network just as if it were undirected.
- This gives us a set of nodes for our sample, from which an induced subgraph can be found.
- The next slides show an example sampling the political bloggers network (Adamic & Glance, 2005), $N = 1490$.
- On first slide: blue is liberal, red is conservative. Node size is proportional to in-degree.
- On later slides: circle is liberal, square is conservative. Colour shows sampling wave.

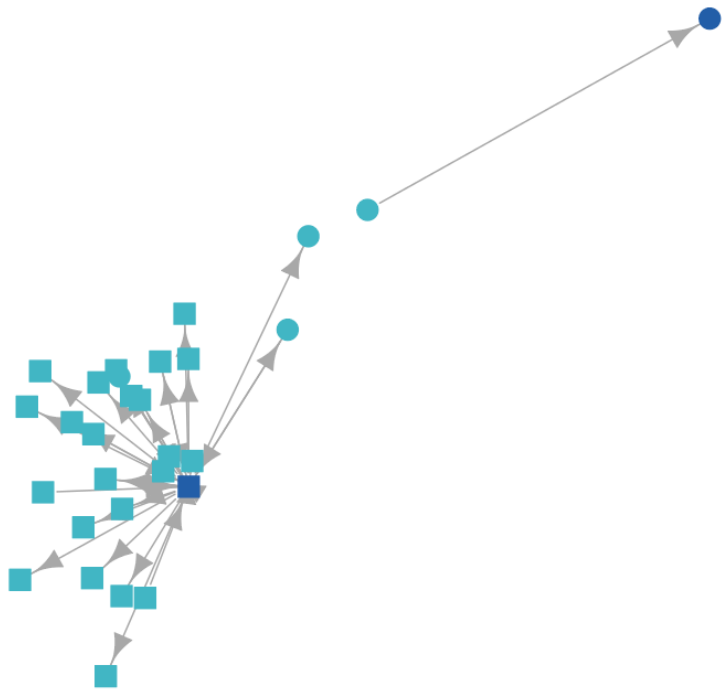
Political bloggers network $N = 1490$



Wave 0: $N_0 = 2$

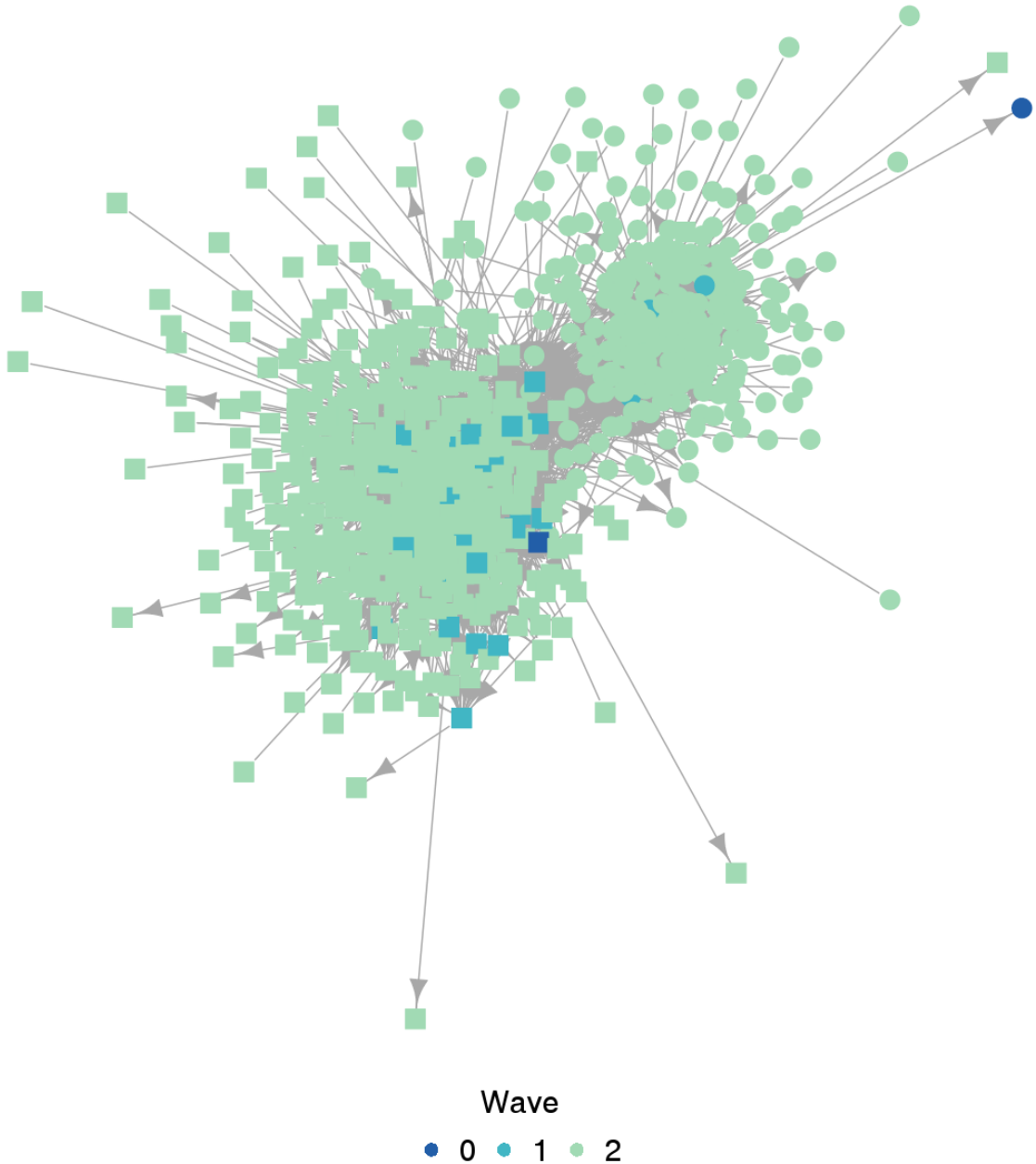


Wave 1: $N_1 = 2 + 27 = 29$



Wave
0 1 2

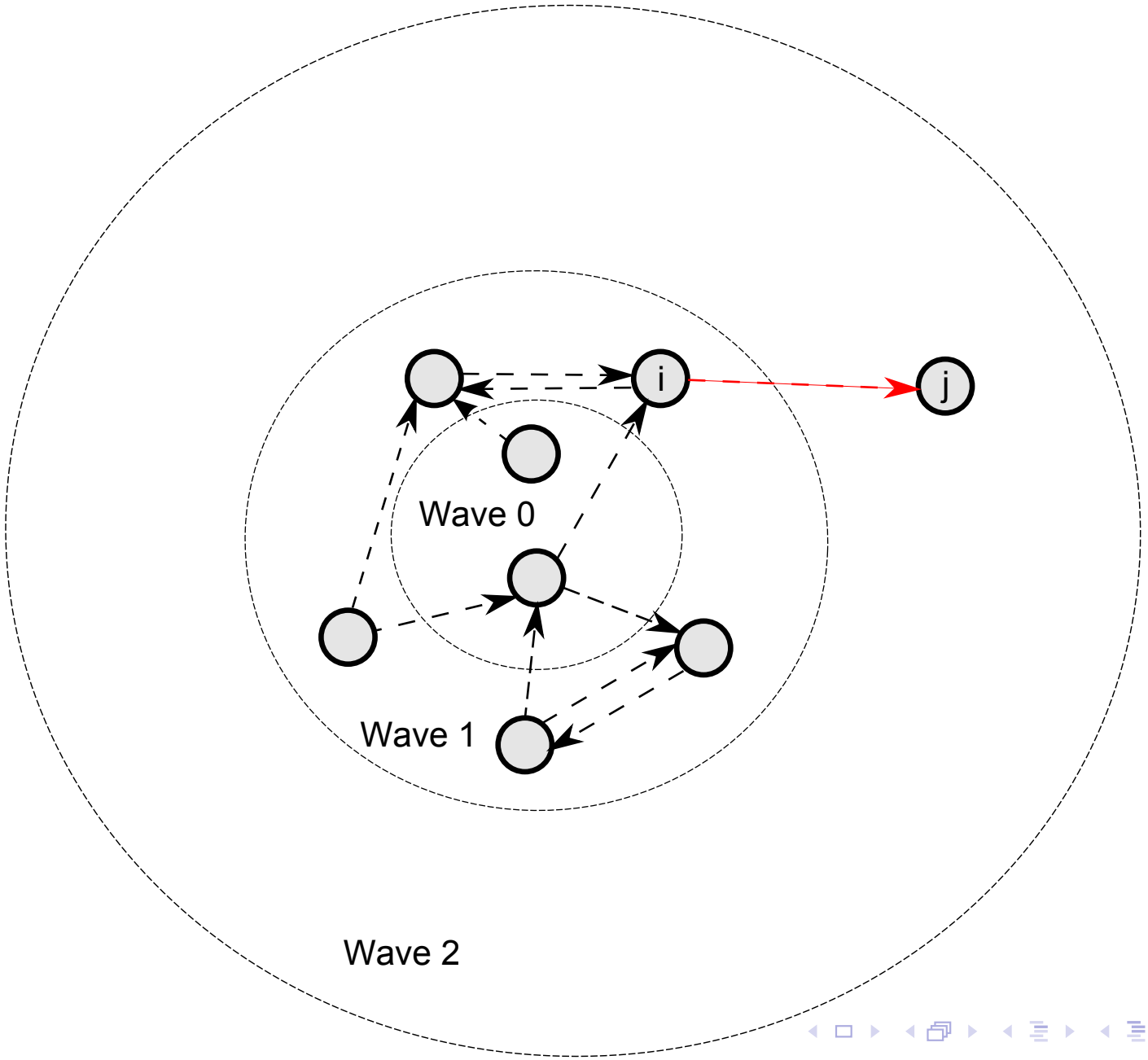
Wave 2: $N_2 = 2 + 27 + 605 = 634$



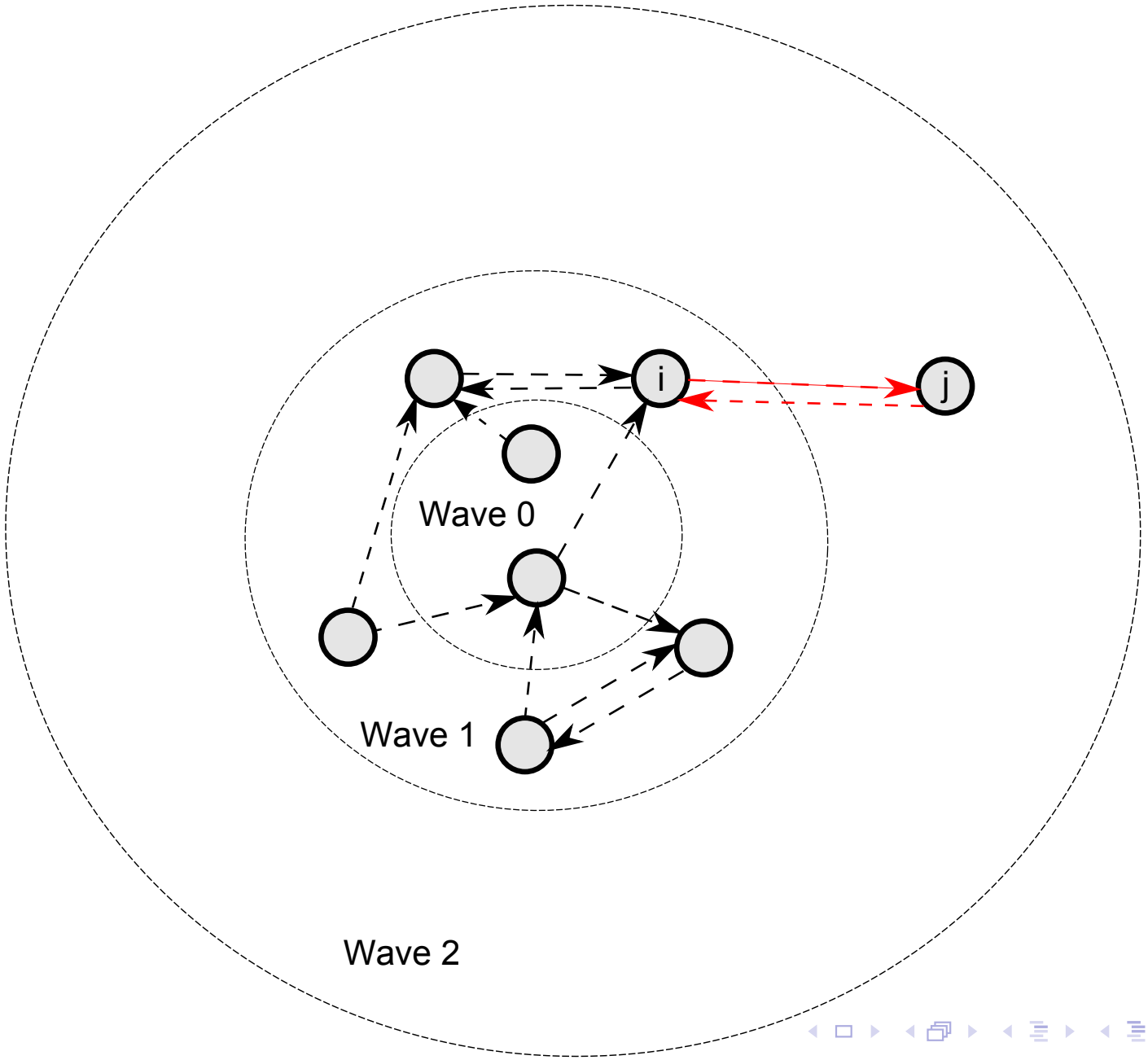
Conditional estimation for directed snowball samples

- The MCMC procedure for conditional estimation simulates new networks by altering edges in the inner waves while keeping ties within the outermost wave fixed.
- In adding or deleting arcs in the inner waves, the snowball sampling structure must be respected.
- I.e. an arc can't be added that skips over a wave(s),
- And can't be deleted if it is the last remaining connection to a node from an earlier wave.
- The arcs are directed, but in obeying the rule above, we take account of the fact that the snowball sampling treated them as undirected.

Arc $i \to j$ cannot be deleted



Arc $i \to j$ or $j \to i$ (but not both) can be deleted



Simulated networks

N	Arc	Reciprocity	Popularity spread [AinS]	Activity spread [AoutS]	Path closure [AT-T]
5000	-4.00	4.25	-1.00	-0.50	1.50

100 digraphs are simulated with the above parameters, resulting in digraphs with the following statistics:

N	Mean components	Mean out-degree	Mean density	Mean global clustering coefficient
5000	1.00	6.52	0.00131	0.01529

Directed snowball conditional estimation — methods

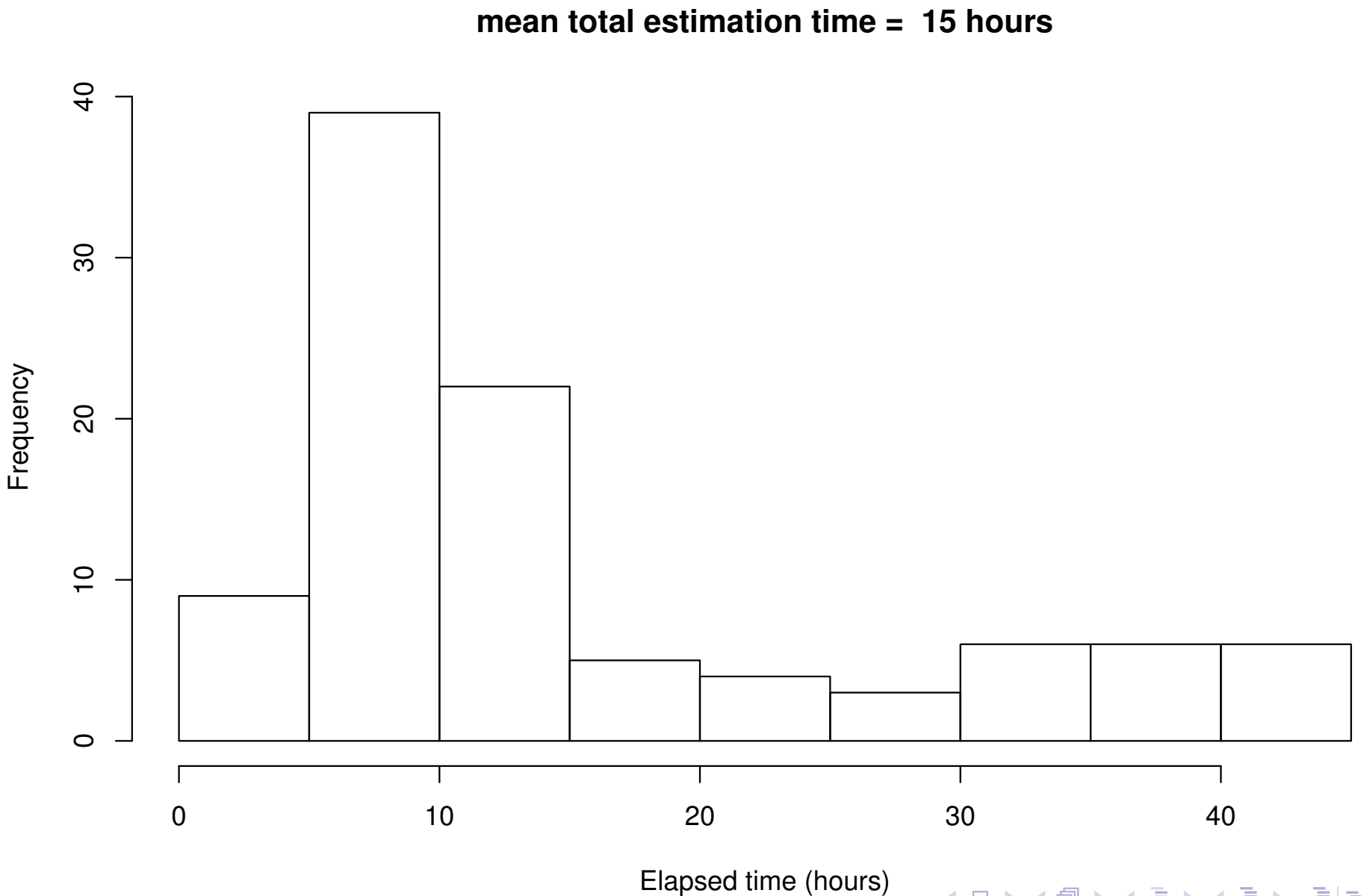
For each of the 100 simulated networks, 20 directed snowball samples are taken, each with the following parameters:

- 2 waves
- 5 seeds

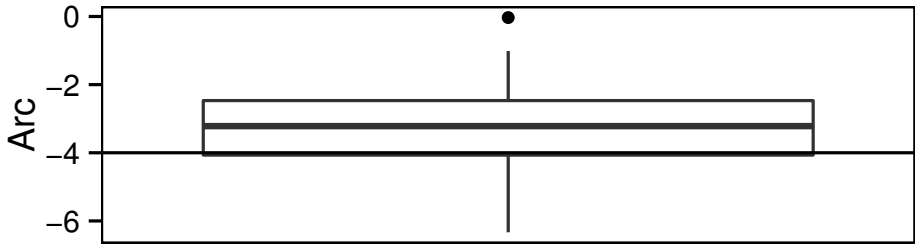
and conditional estimation is done in parallel with 20 MPI tasks (one per snowball sample).

The cluster system used is a 48 node compute cluster, each node has 16 cores (AMD Opteron 6128, 2.0 GHz) and 32 GB memory, running Linux and OpenMPI.

Simulated networks — conditional estimation elapsed time

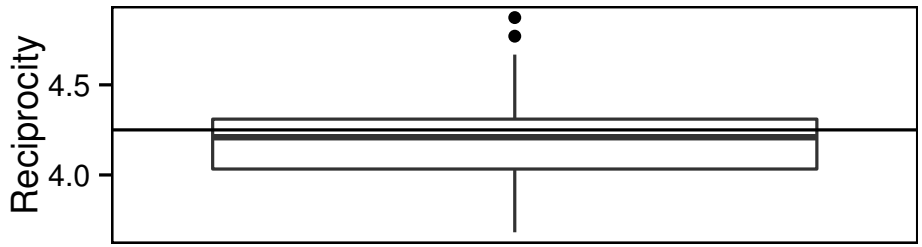


Simulated networks — results



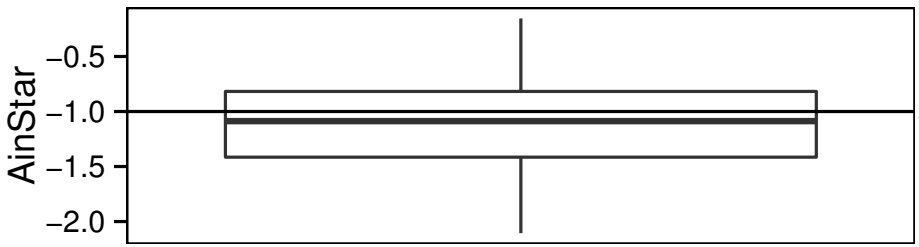
bias = 0.7977, RMSE = 1.365

% In CI = 100, FNR% = 45



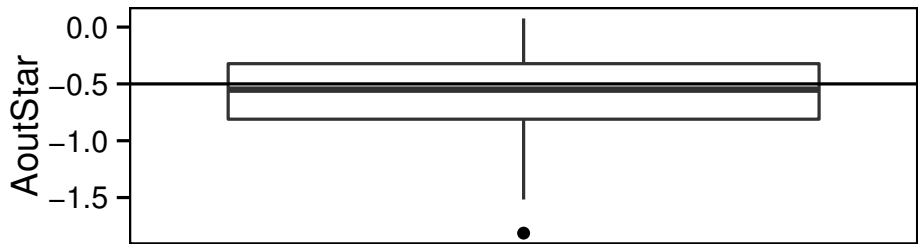
bias = −0.06254, RMSE = 0.2334

% In CI = 96, FNR% = 0



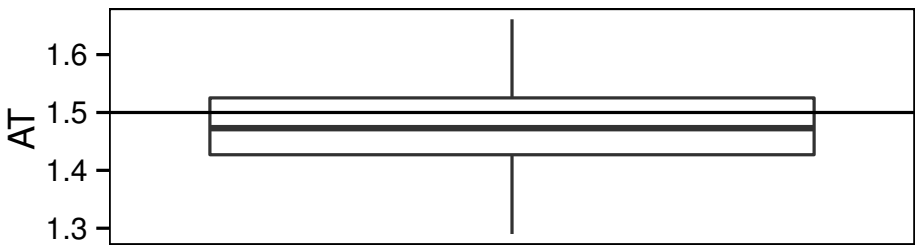
bias = −0.1182, RMSE = 0.4364

% In CI = 100, FNR% = 60



bias = −0.0874, RMSE = 0.3777

% In CI = 99, FNR% = 86



bias = −0.02373, RMSE = 0.07687

% In CI = 100, FNR% = 0

Results for political bloggers network

Effect	N_s	Estimate	C.I. lower	C.I. upper	
Shared pop.,2-paths [A2P-TD]	20	-0.0188	-0.0454	0.0077	
Arc	20	-19.5606	-29.5288	-9.5923	*
Popularity spread [AinS]	20	3.2070	1.0355	5.3785	*
Activity spread [AoutS]	20	3.4345	2.0965	4.7725	*
Reciprocity	20	1.3081	0.1185	2.4976	*
Matching	20	1.6025	0.6410	2.5639	*
Matching reciprocity	20	-0.1334	-1.5643	1.2975	

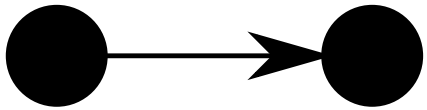
2 waves, 2 seeds, 20 samples (parallel), 133 hours (SGI Altix XE Cluster, 1088 Intel Nehalem cores (8 per node) 2.66 GHz, CentOS 5, OpenMPI). Could not get PNet whole network estimate in 800 hour limit.

Acknowledgments

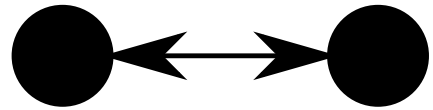
- Co-authors: Garry Robins, David Rolls
- Johan Koskinen, Peng Wang
- University of Melbourne ITS High Performance Computing
- This research was supported by Victorian Life Sciences Computation Initiative (VLSCI) grant numbers VR0261 and VR0297 on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government, Australia.
- And the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575, with Prof. Noshir Contractor and Dr Yun Huang at Northwestern U.

Hidden bonus slides

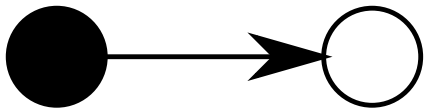
Model configurations — categorical attributes



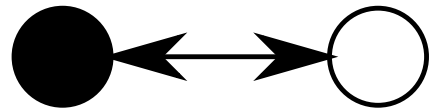
Matching



Matching reciprocity



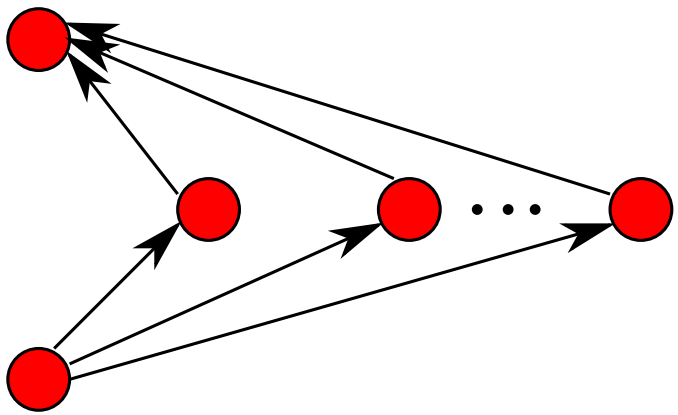
Mismatching



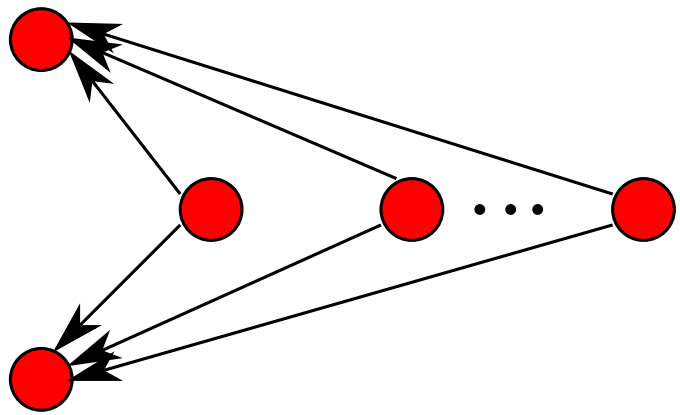
Mismatching reciprocity

Model configurations: A2P-TD

A2P-TD counts multiple 2-paths (A2P-T) and shared popularity (A2P-D) in a single configuration, adjusting for double counting

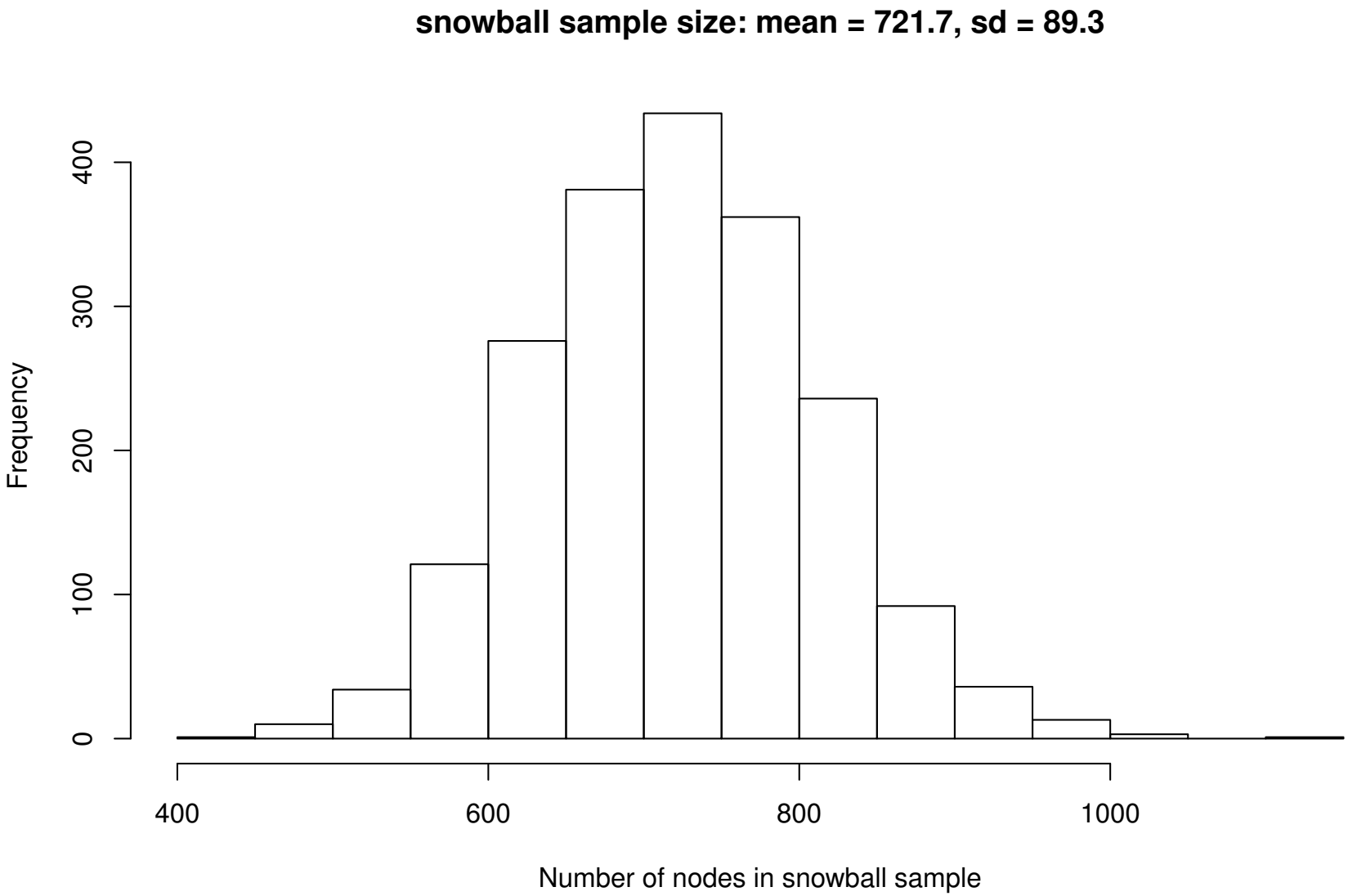


Multiple 2-paths
A2P-T

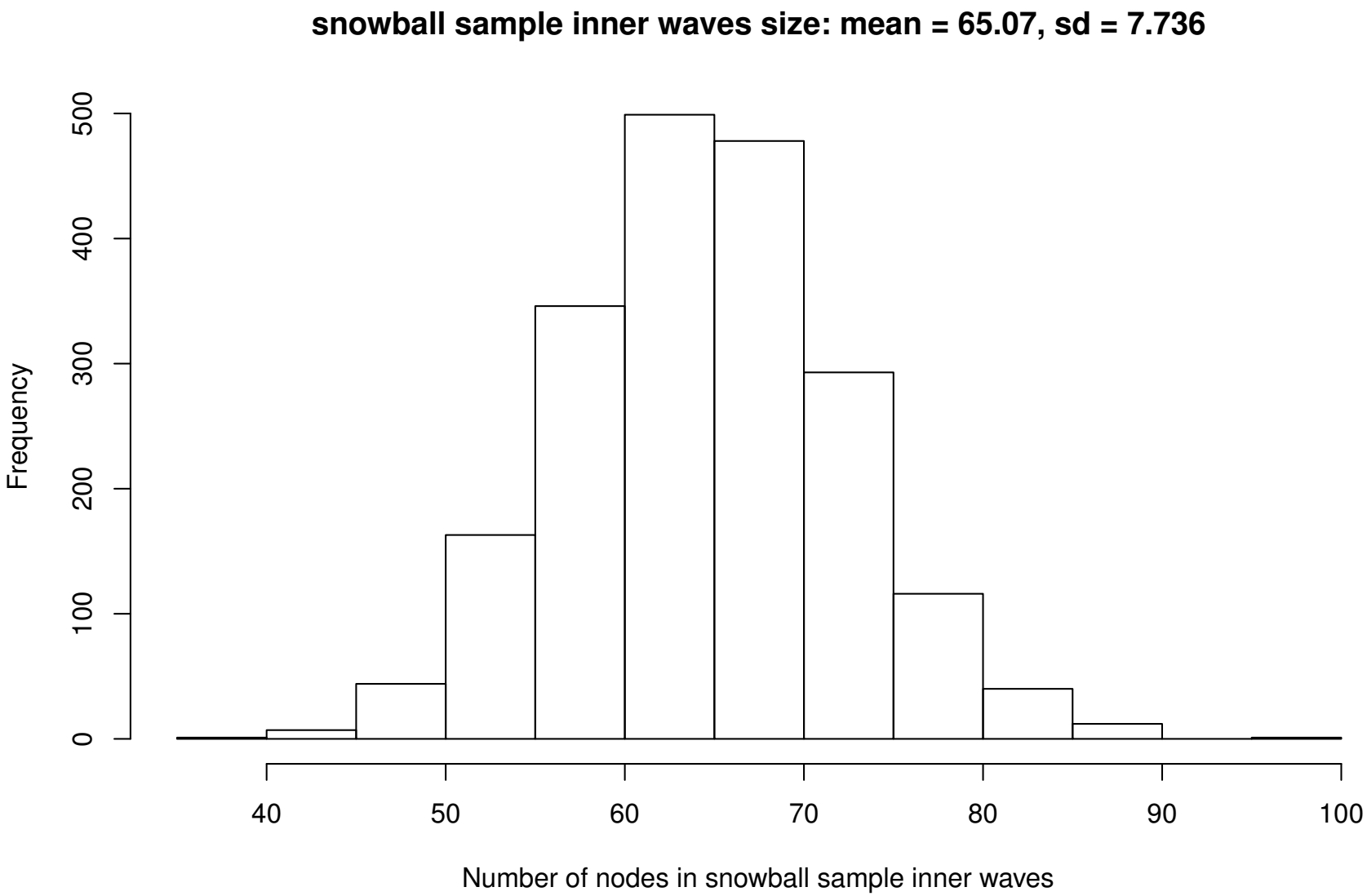


Shared popularity
A2P-D

Simulated networks — snowball sample sizes (1)



Simulated networks — snowball sample sizes (2)



Political bloggers network snowball sample sizes