Social influence models with missing data

Alex Stivala[1] H. Colin Gallagher[1] David Rolls[1]
Peng Wang[2] Garry Robins[1]

[1]Melbourne School of Psychological Sciences, The University of Melbourne, Australia
[2]Centre for Transformative Innovation, Faculty of Business and Law, Swinburne University of Technology, Australia

Sunbelt XXXVI, April 5–10, 2016, Newport Beach CA

Social selection and social influence

Social selection model (ERGM) Given the observed network and actor attributes, what processes lead to network tie formation?

- The outcome variables are network ties.
- The model predicts the probability of a tie based on structure (other ties) and actor attributes.

Social influence model (ALAAM) Given the observed network, actor attributes, and (binary) dependent variable, what processes lead to the outcome variable being true?

- The outcome variables are the binary dependent variable.
- The model predicts the probability of the dependent variable being true based on structure (network ties), actor attributes, and dependent variable of other actors.

Motivation (1)

- Logistic regression assumes independence of the outcome variables,
- and random samples can be used when the population is large to make valid inferences.
- But if we are looking for evidence of social influence, our hypothesis specifically contradicts these individualistic assumptions.
- Hence the need for a social influence model such as ALAAM.
- But in real world studies, missing data is inevitable,
- And simple random sampling neglects network structure, which is captured by snowball sampling.

Motivation (2)

- The practical motivation is an epidemiological study in large-scale community samples.
- Hence in this work we examine the validity of inferences made using ALAAMs when nodes are missing at random, or snowball sampling is used.

Autologistic actor attribute model (ALAAM)

Models the probability of vector of binary attributes Y given network (matrix of 0-1 tie variables) X:

$$\Pr(Y = y|X = x) = \frac{1}{\kappa(\theta_I)} \exp\left(\sum_I \theta_I z_I(y, x, w)\right)$$

where

- z_I is a network-attribute statistic,
- θ_I is the parameter corresponding to z_I,
- the "configuration" I is defined by a combination of dependent attribute variables y, network variables x, and actor covariates w,
- $\kappa(\theta_I)$ is a normalizing quantity which ensures a proper probability distribution.

ALAAM dependence assumptions

Assumptions about which attributes Y are independent determines which configurations I are allowed in the model.

- If any two Y_i and Y_j are assumed independent, the only configuration is a single node, and there are no network effects, and the model is just logistic regression.
- If Y_i is conditionally dependent on network tie X_{jk} if and only if $\{i\} \cap \{k, j\} \neq \emptyset$, that is, if and only if the actor i is one end of the tie X_{jk} then configurations include stars and *contagion*.
- Other assumptions with higher order configurations are possible, but we will use this dependence assumption.

ALAAM configurations

Attribute density The number of nodes with attribute Y;

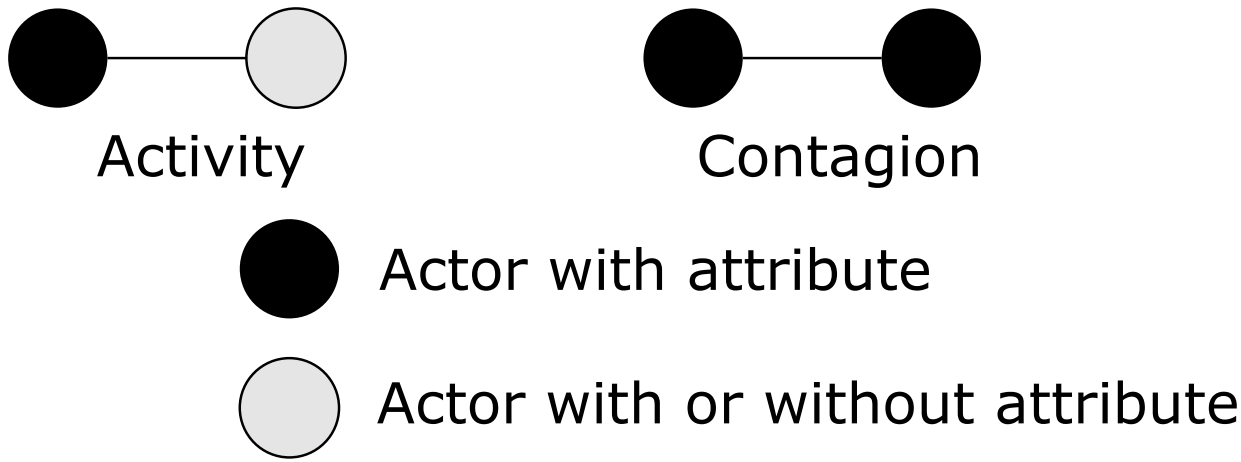
Activity Presence of a tie at a node with attribute Y. That is, whether having attribute Y is associated with having a tie to others;

Contagion The propensity for two nodes with a tie between them to both have attribute Y;

Binary The propensity for a node to have attribute Y based on another binary attribute U;

Continuous The propensity for a node to have attribute Y based on its continuous attribute V.

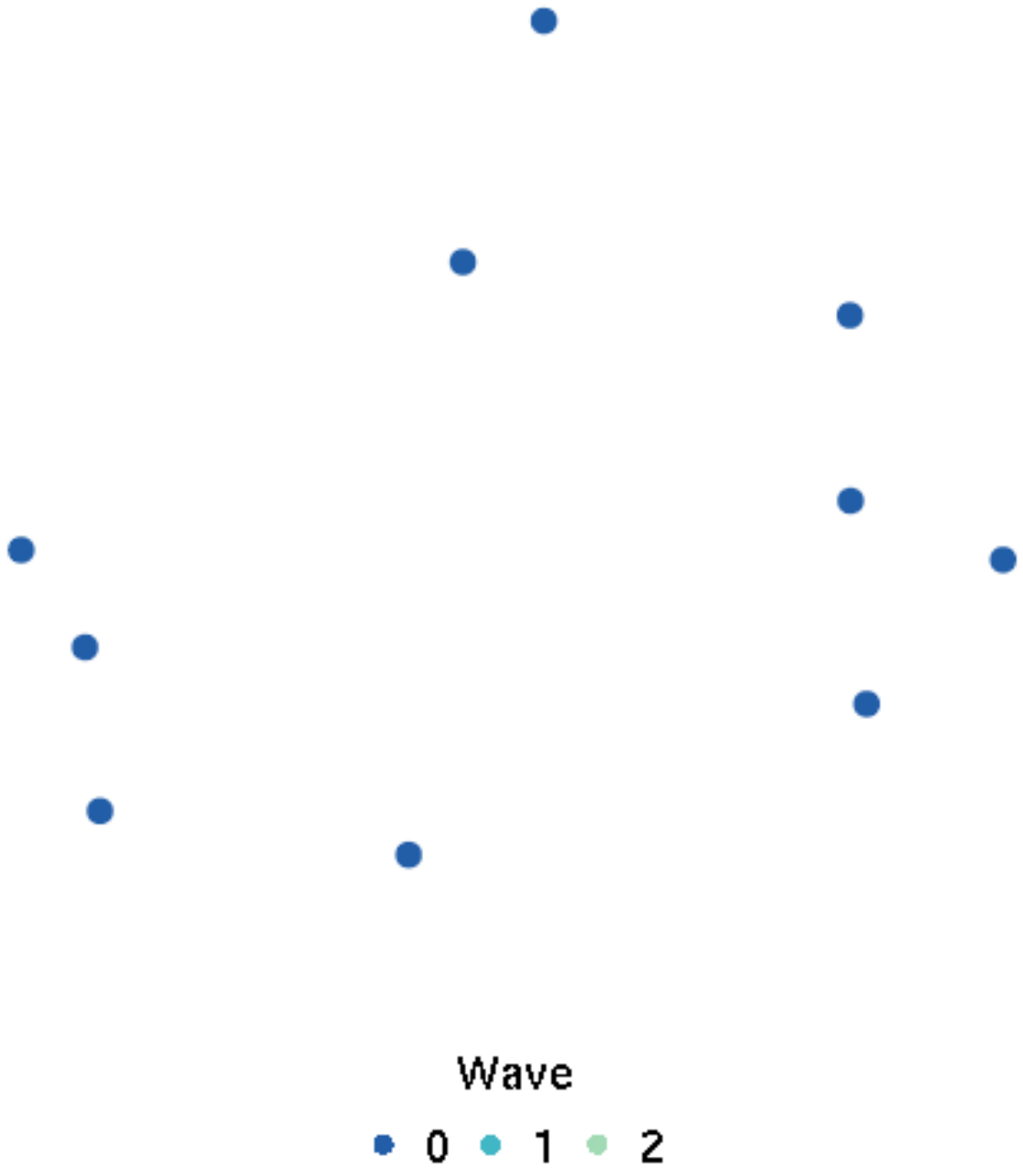
Activity and contagion



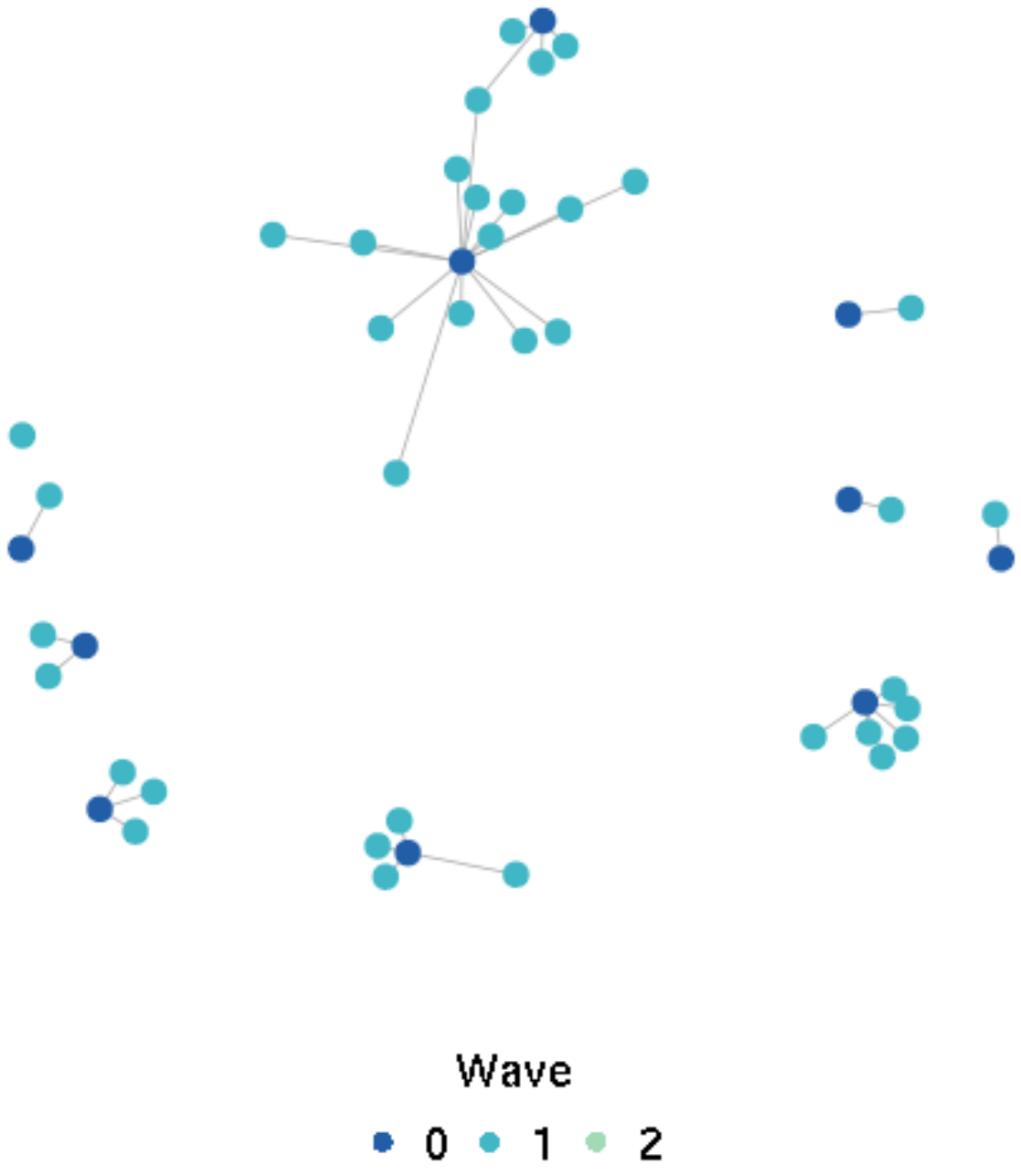
Snowball sampling

- Start with N_0 seed nodes (wave 0).
- Follow all their ties to get a further set of nodes (wave 1).
  - (Following all ties is BFS. If instead at most a fixed number m of ties are followed, we call it "fixed choice" sampling or "degree censoring".)
- 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...

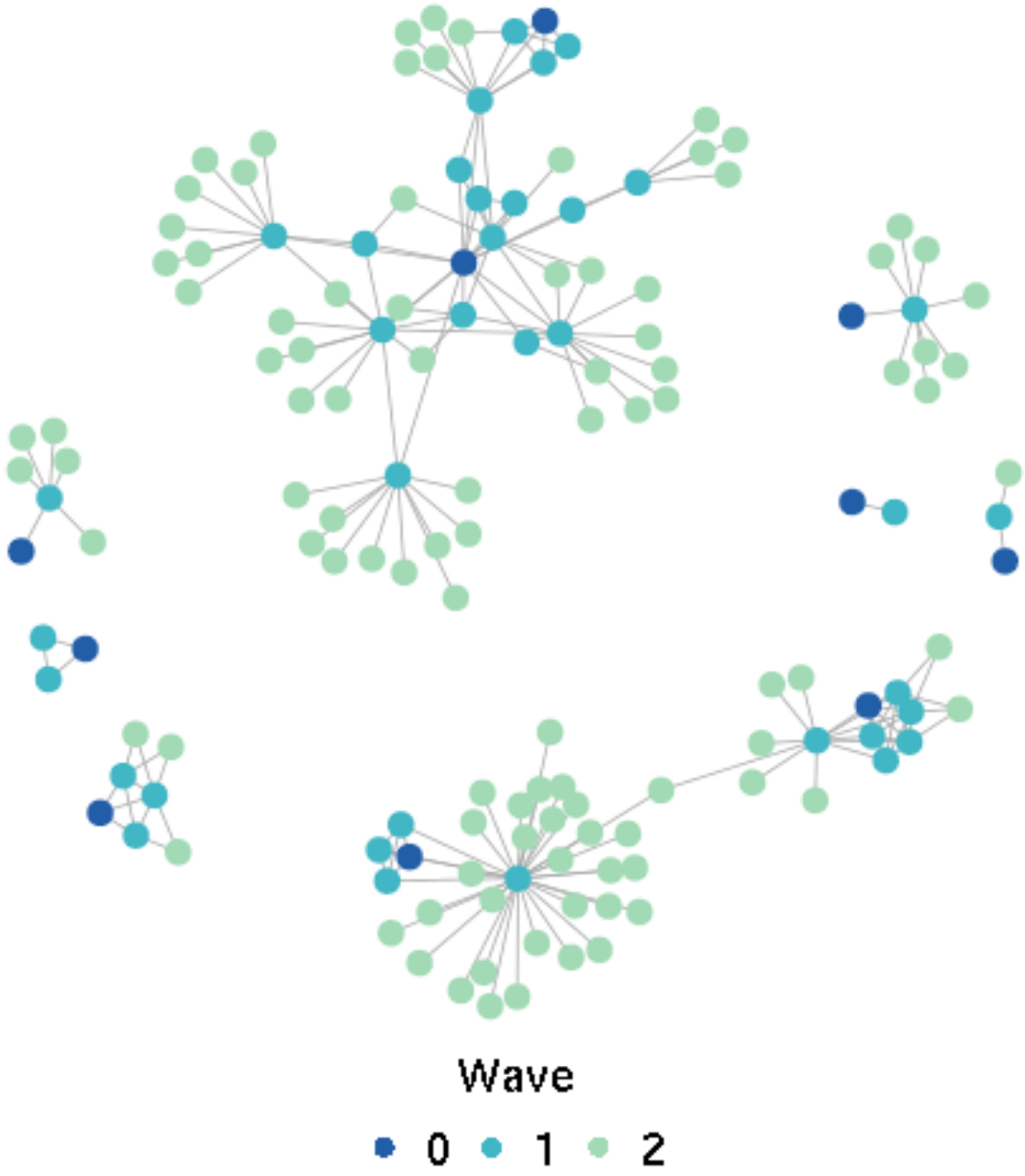
Snowball sampling example, wave 0



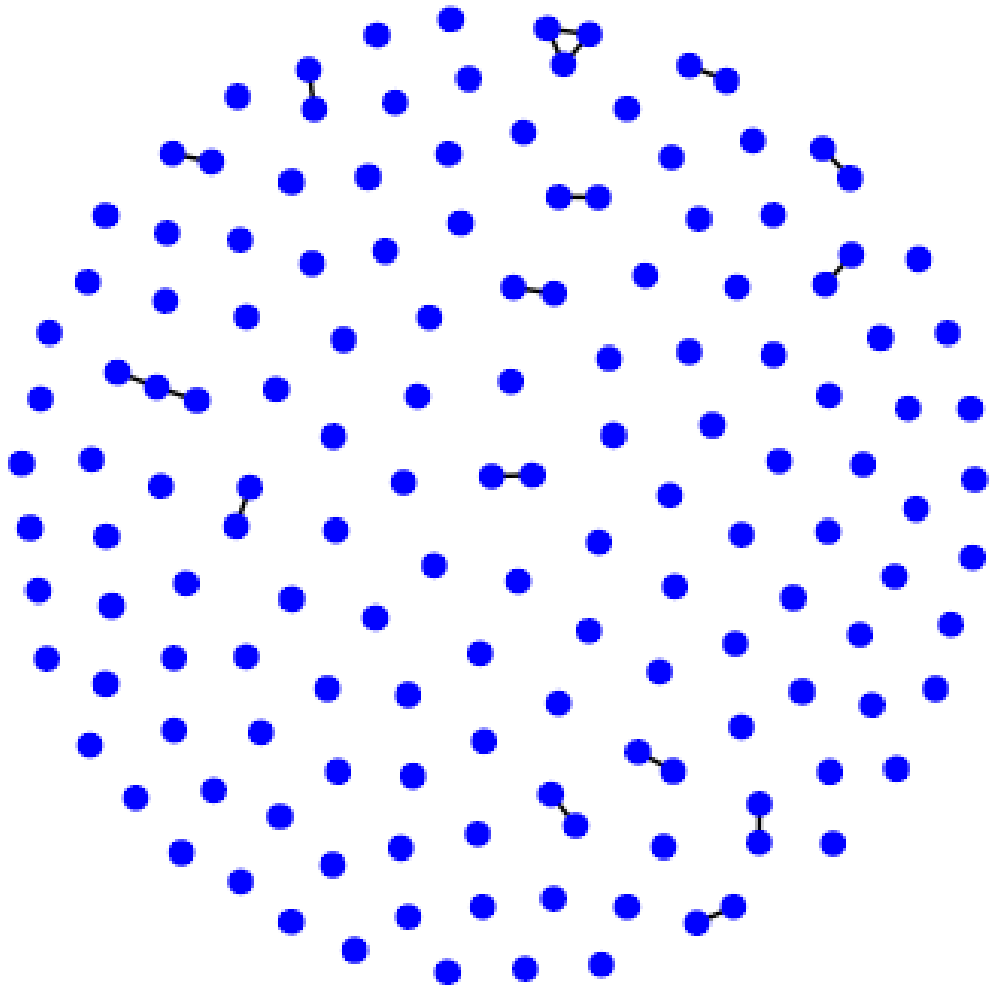
Snowball sampling example, wave 1



Snowball sampling example, wave 2



Random sampling example, same number of nodes (146)



Simulated networks and ALAAMs — methods

- By using simulated ALAAMs (on simulated networks) we can measure errors in the ALAAM parameter estimation (including bias, RMSE and Type II errors in inference).
- By simulating ALAAMs, each with a single parameter set to zero, we can measure Type I error rates in inference.
- We study two kinds of sampling:
  - Simple random sampling (viewed another way, nodes are missing at random)
  - Snowball sampling, with both fixed choice (degree censoring) and without fixed choice (BFS sampling)

Statistics of simulated networks

N	Components	Mean degree	Max. degree	Density	Clustering coefficient	Positive outcome % mean	s.d.
500	5	4.90	11	0.00983	0.10347	15	2.19
1000	3	6.00	13	0.00601	0.07097	15	1.59

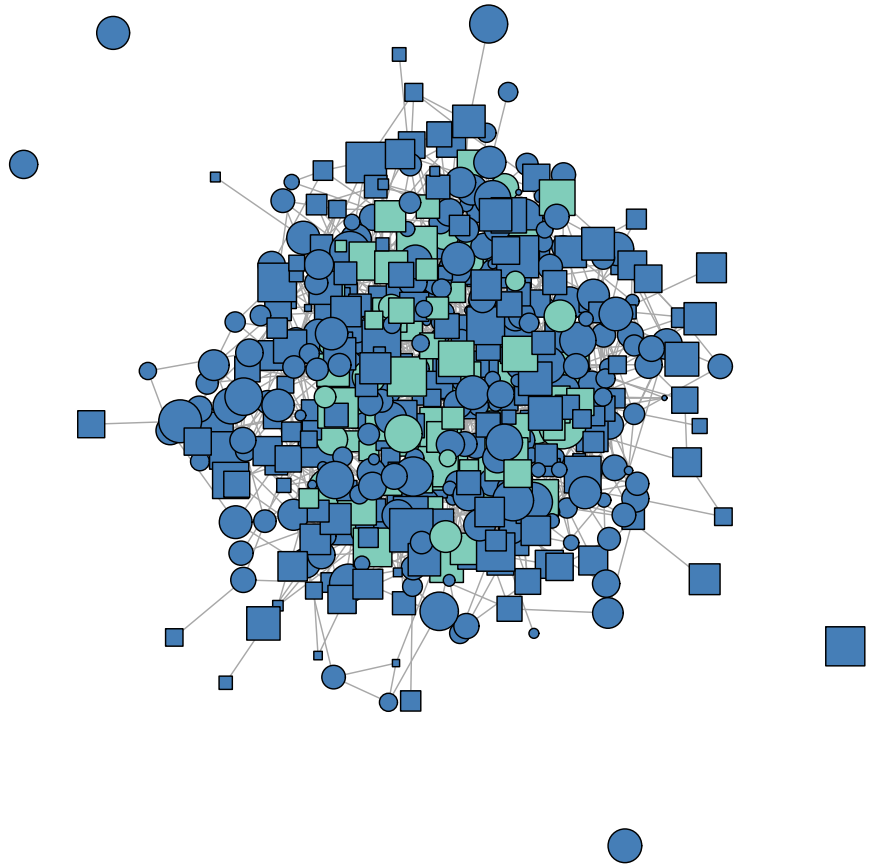
- Each node (actor) has a binary and a continuous attribute.
- The binary attribute is assigned the positive value for 50% of the nodes, chosen at random.
- For the continuous attribute, the attribute value v_i at each node i is $v_i \overset{iid}{\sim} N(0,1)$.

Parameters of the simulated ALAAMs

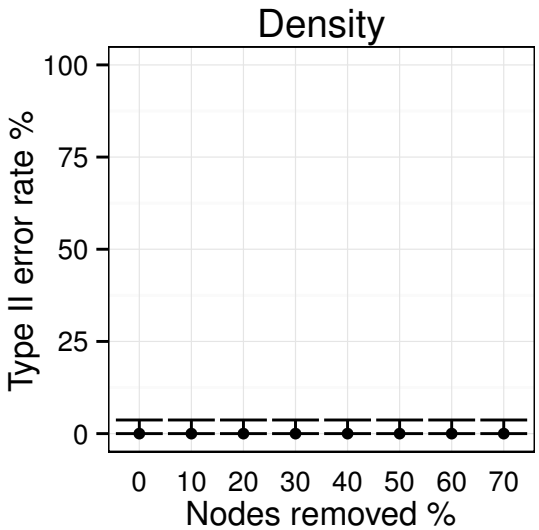
N	Density	Activity	Contagion	Binary	Continuous
500	-7.20	0.55	1.00	1.20	1.15
1000	-8.05	0.55	1.00	1.20	1.15

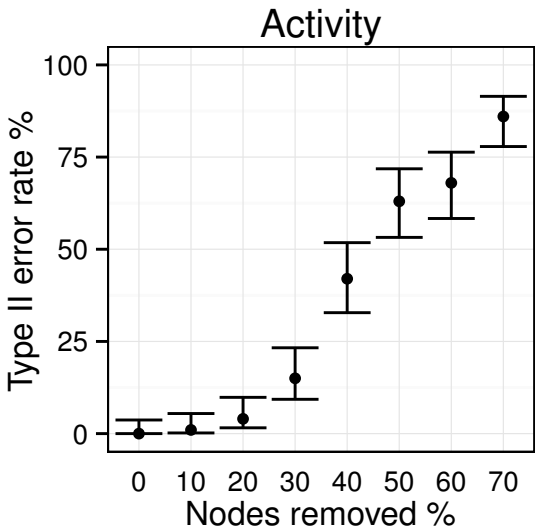
- The parameters were chosen so that the proportion of nodes with a positive outcome was approx. 15%
- This was chosen to correspond roughly to rates of mental health conditions in a disaster-affected population in an actual study (Bryant et al., 2014)

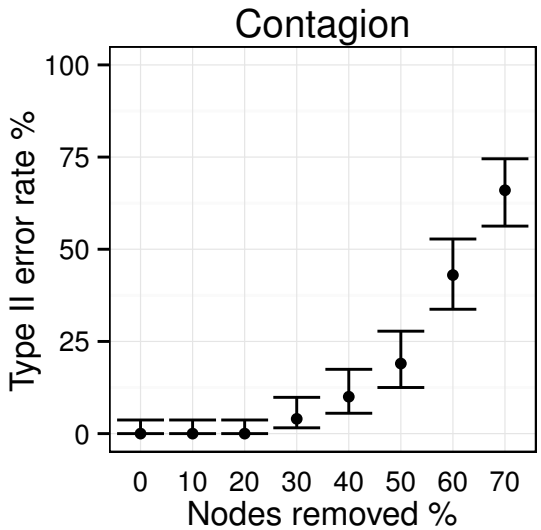
Simulated network $N = 500$

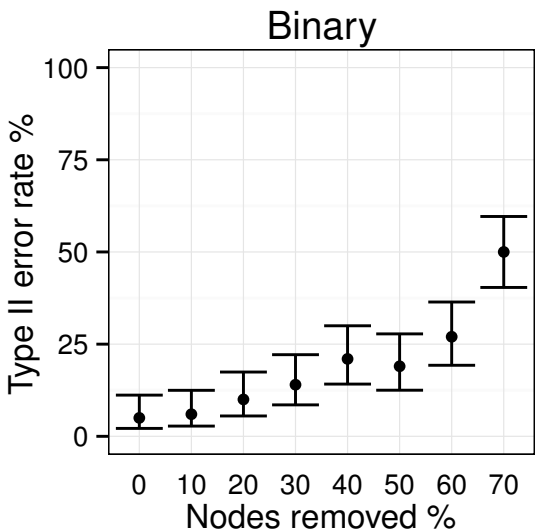


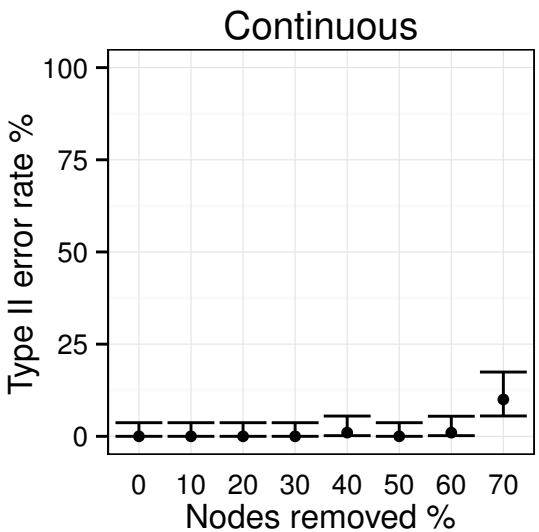
Effect on type II error of random removal of nodes





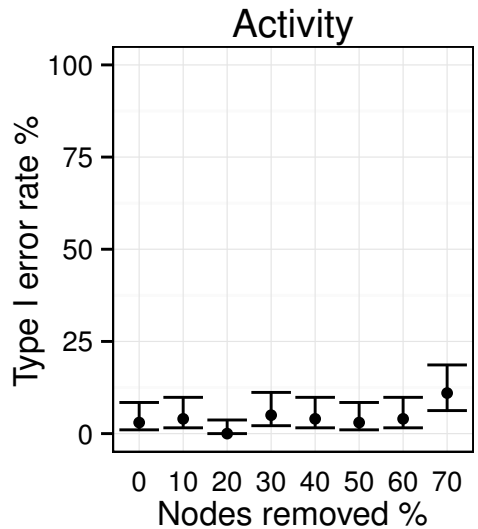


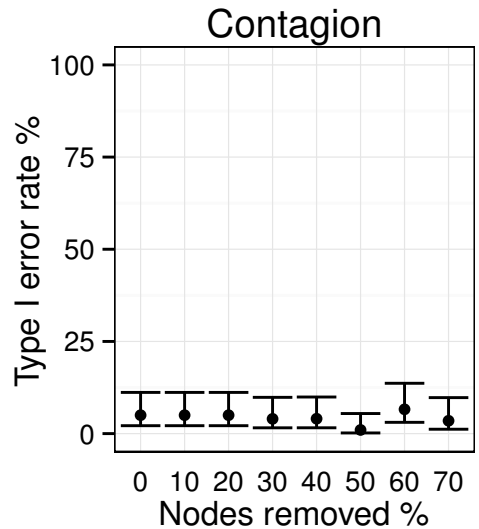


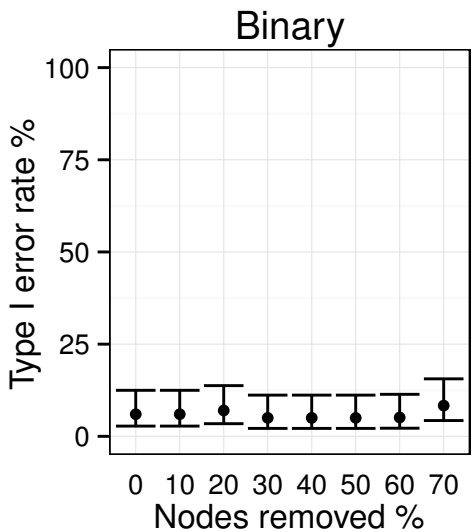


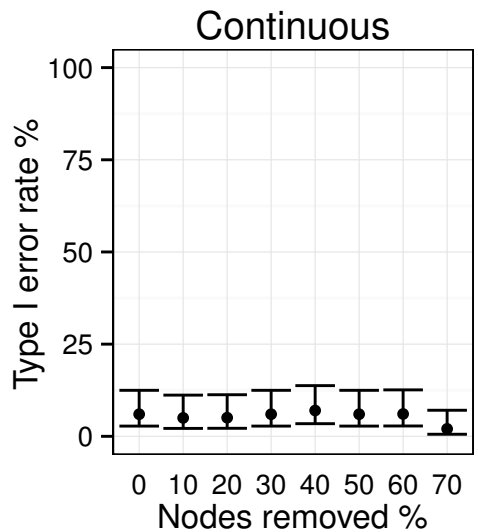
Effect on type I error of random removal of nodes

no Density parameter

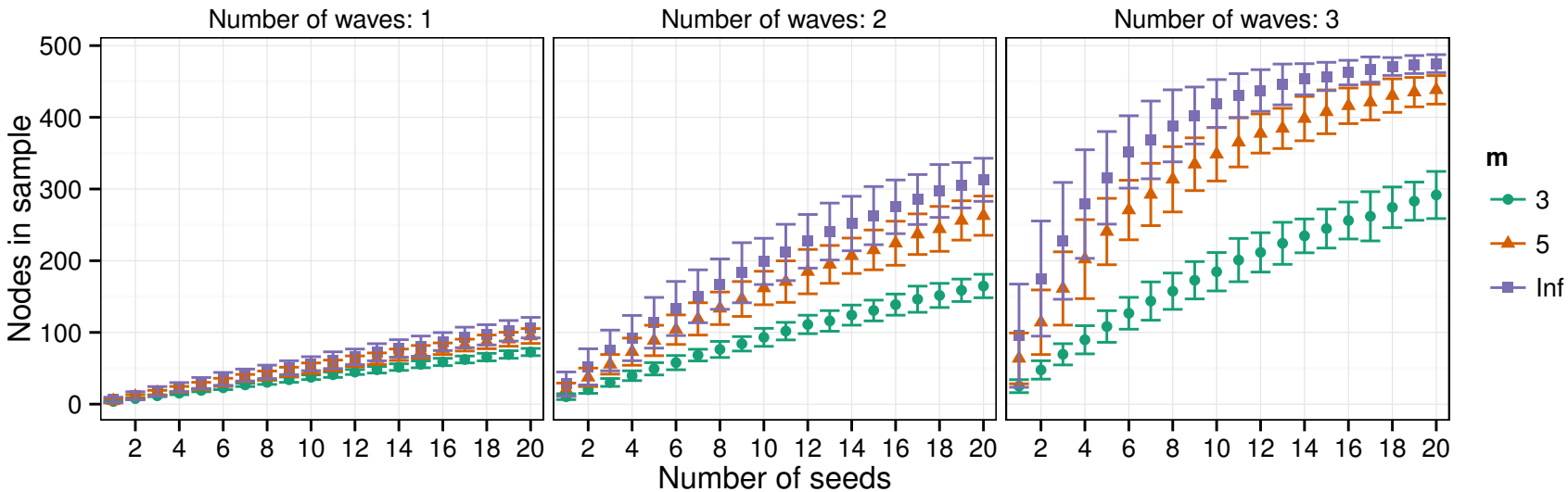




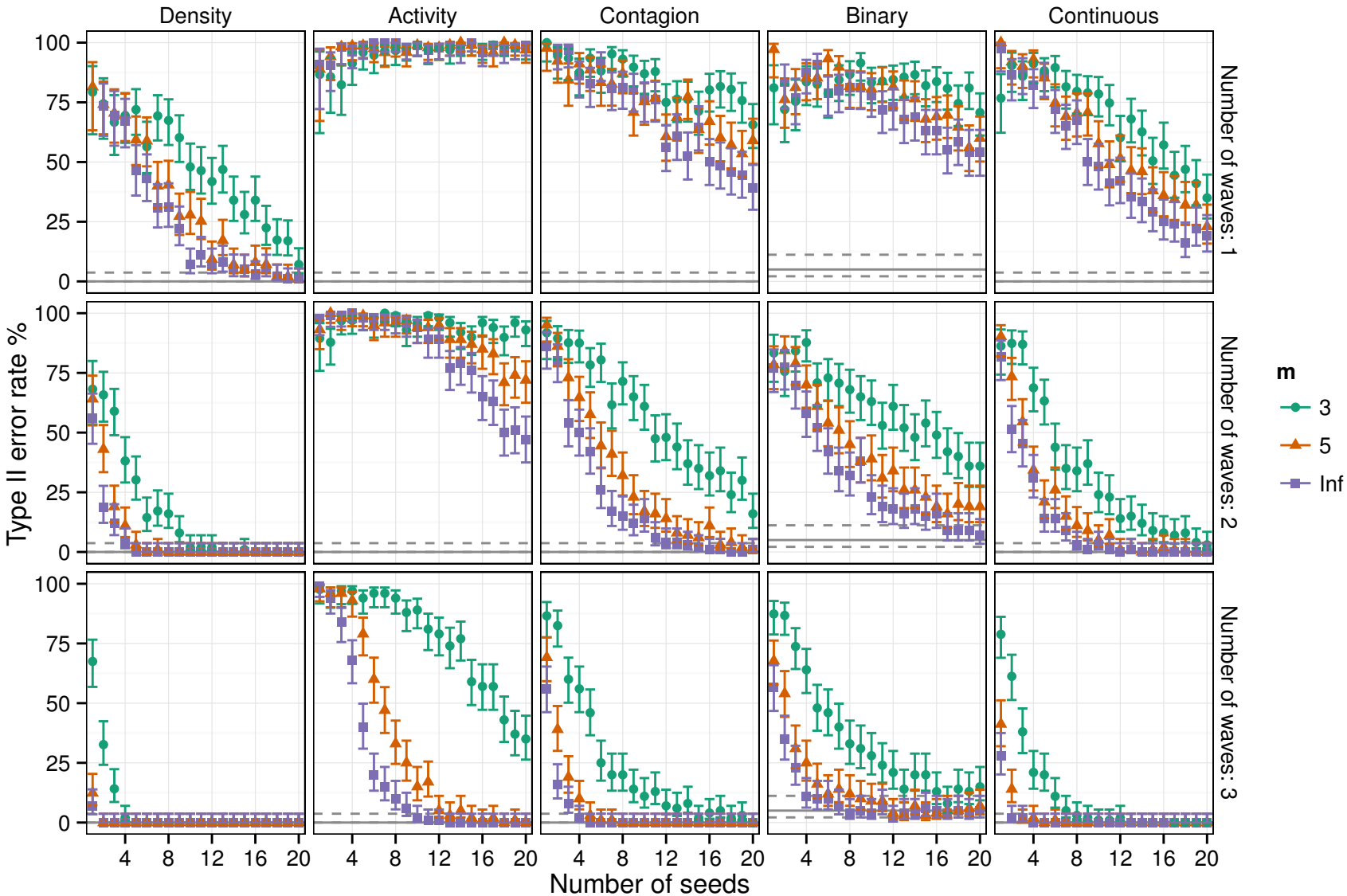




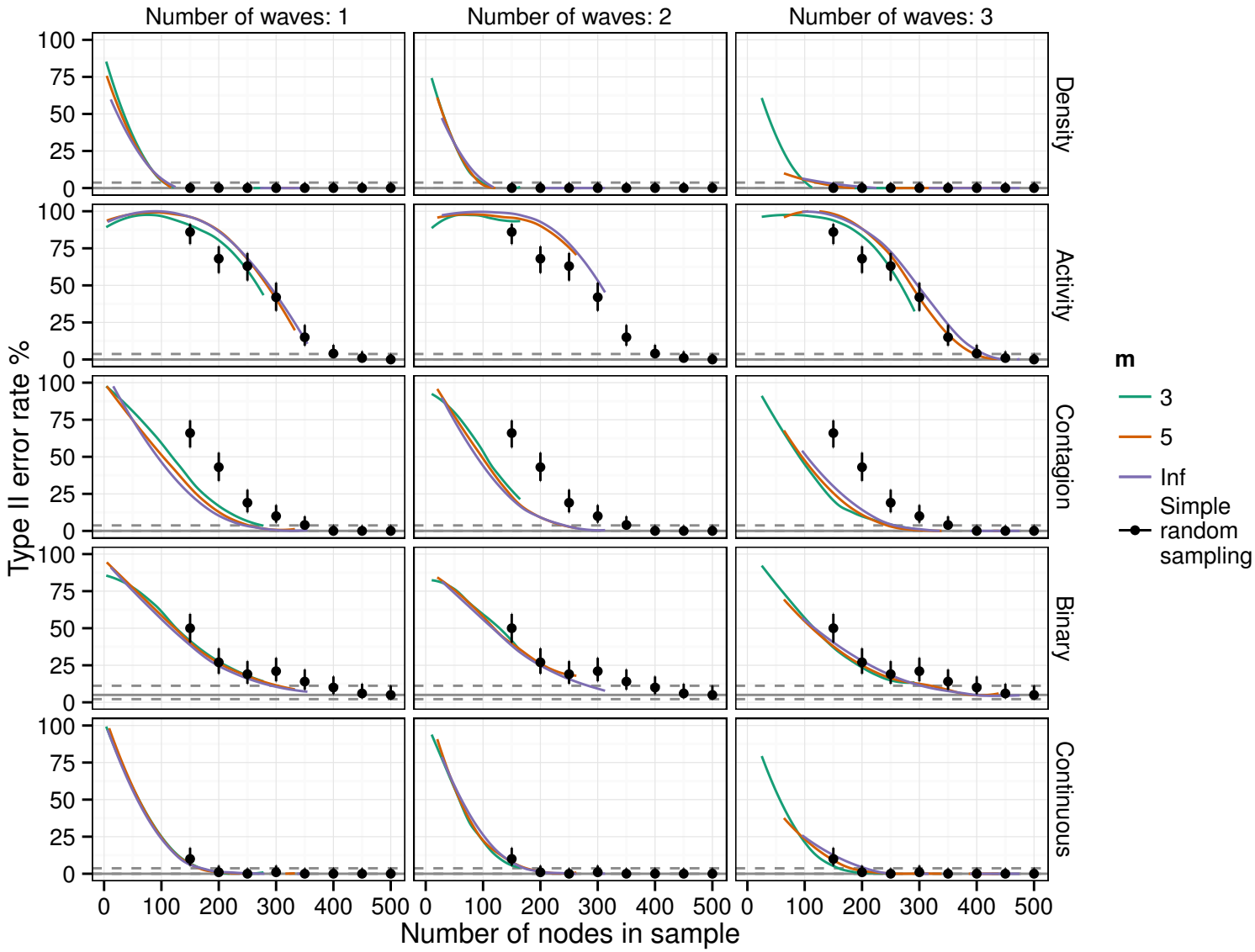
Snowball sample size by number of waves and degree censoring



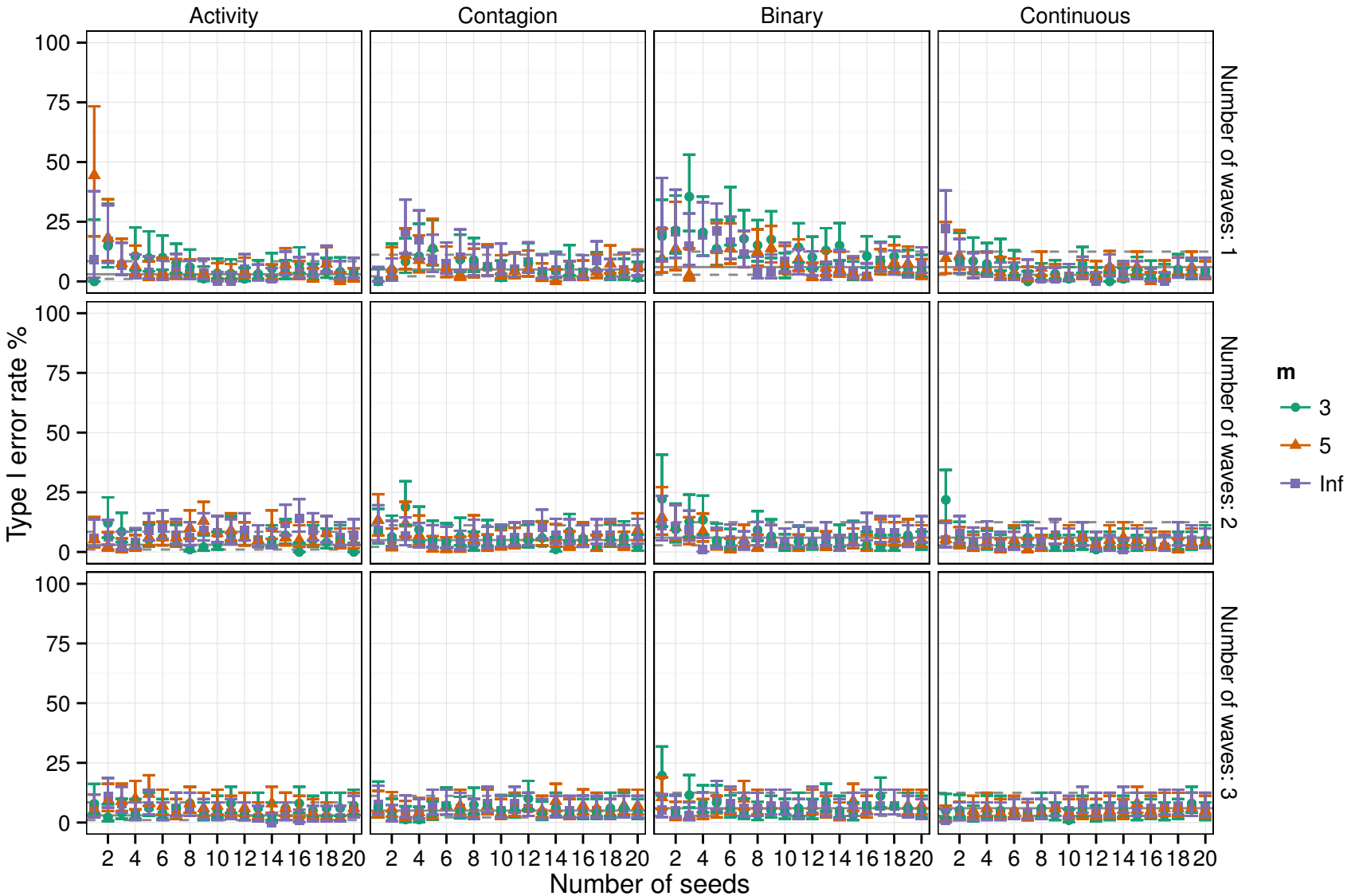
Snowball sampling effect on type II error rate



Snowball sample size effect on type II error rate



Snowball sampling effect on type I error rate



Conclusion

- ALAAM parameter inference can work with non-trivial amounts of missing network data.
- The Type I error rate does not significantly increase for any amount of data missing at random, and for any snowball sampling scheme tested (except for unreasonably small network samples).
- The Type II error rate remains reasonable even for high rates of missing data, except for the Activity parameter.
- For a given network sample size, snowball sampling gives significantly higher power on the Contagion parameter than simple random sampling.

Acknowledgments

- Co-authors: Colin Gallagher, David Rolls, Peng Wang, Garry Robins
- Johan Koskinen
- University of Melbourne ITS High Performance Computing
- This research was supported by Victorian Life Sciences Computation Initiative (VLSCI) grant number VR0261 on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government, Australia.

You might also like...

- H. Colin Gallagher *et al.* "Social Influence Models in Community Setting: The Case of Post-Disaster Mental Health" Thursday 11:40 Salon 5
- Peng Wang *et al.* "Combined Analysis of Social Structure and Individual Outcomes Using ERGMs" Friday 16:20 Baycliff

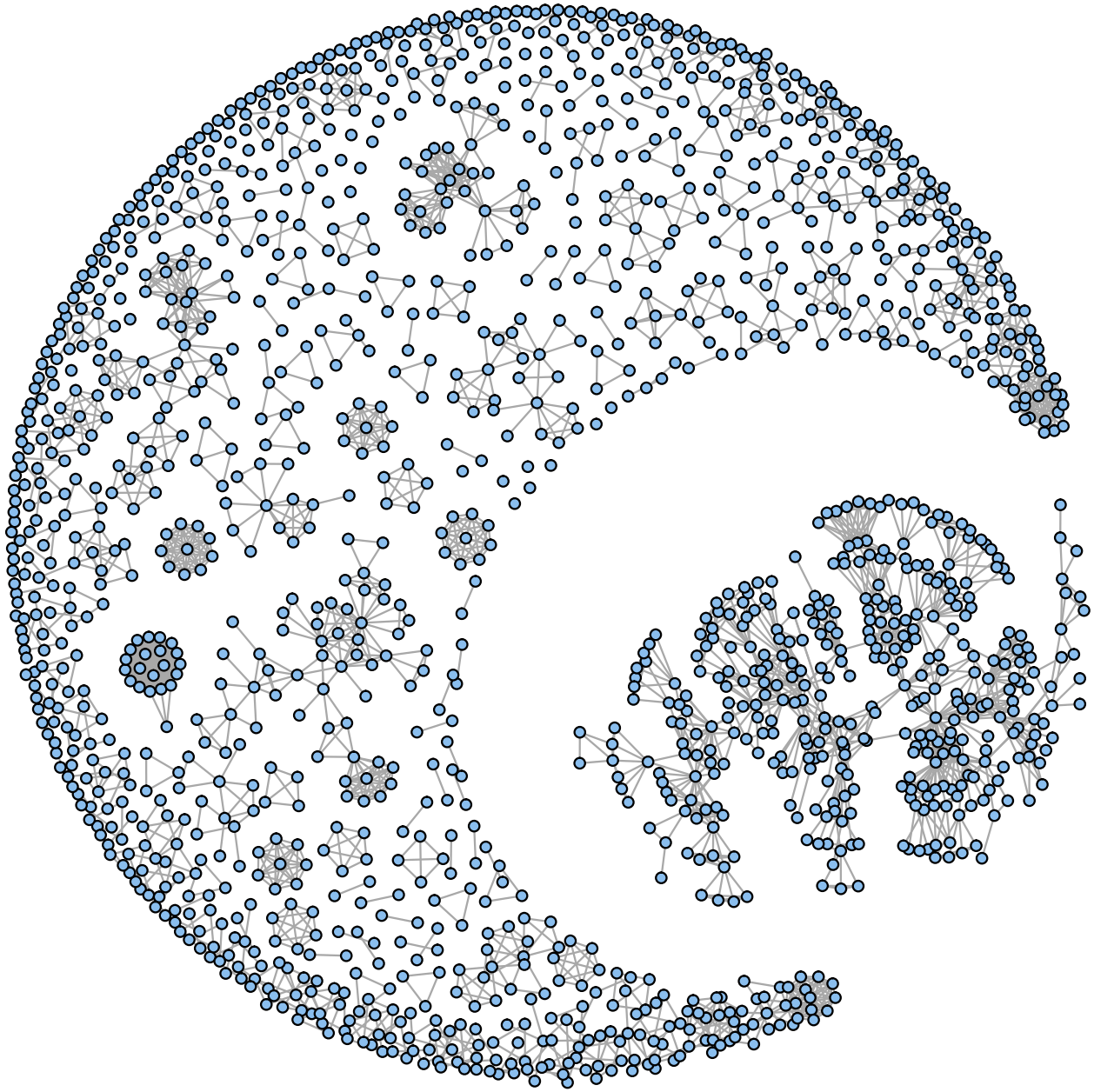
These slides

Slides:
`http://munk.cis.unimelb.edu.au/~stivalaa/alaam_extended_slides.pdf`

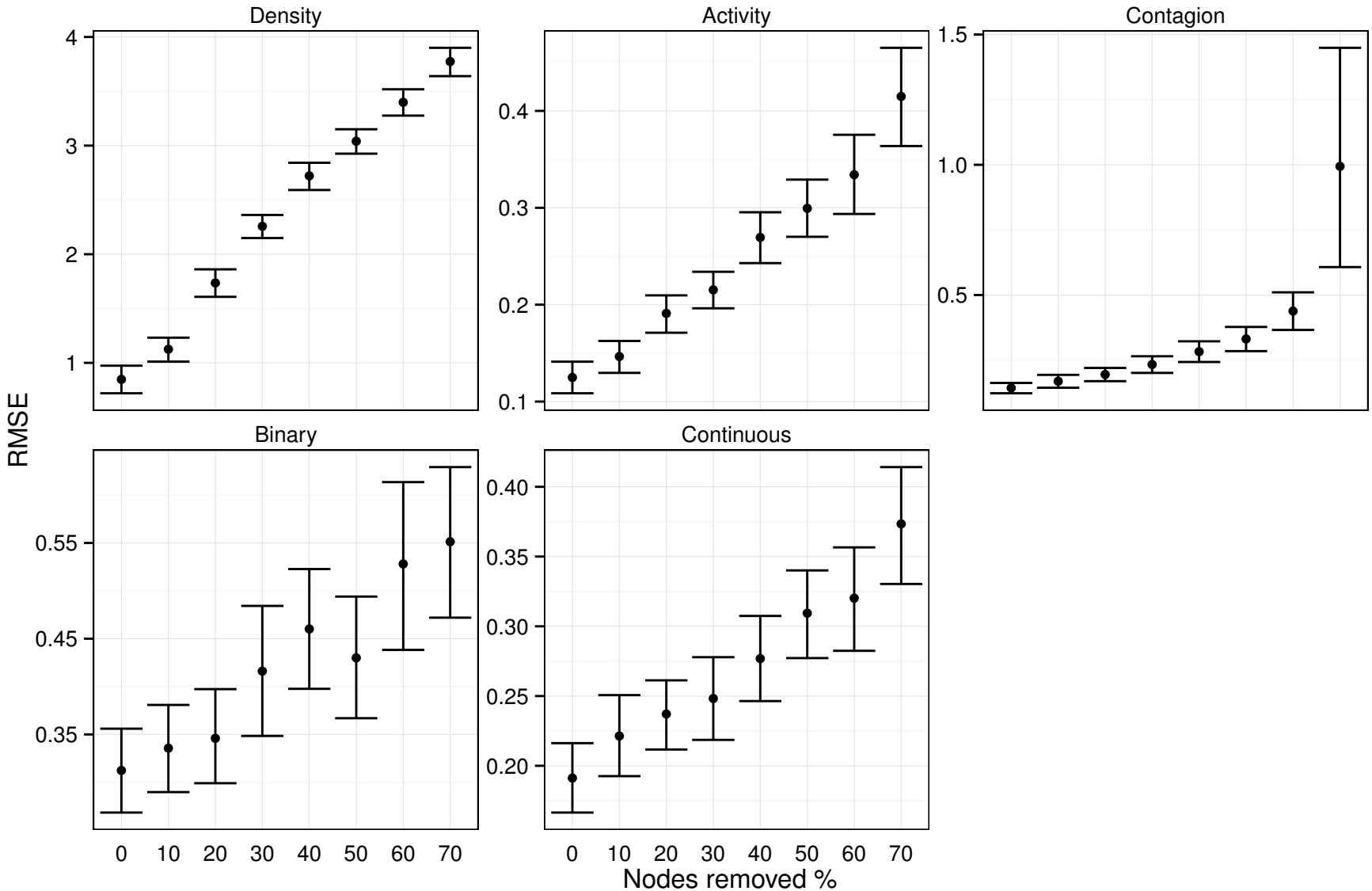
Printable version:
`http://munk.cis.unimelb.edu.au/~stivalaa/alaam_extended_handout.pdf`

Hidden bonus slides

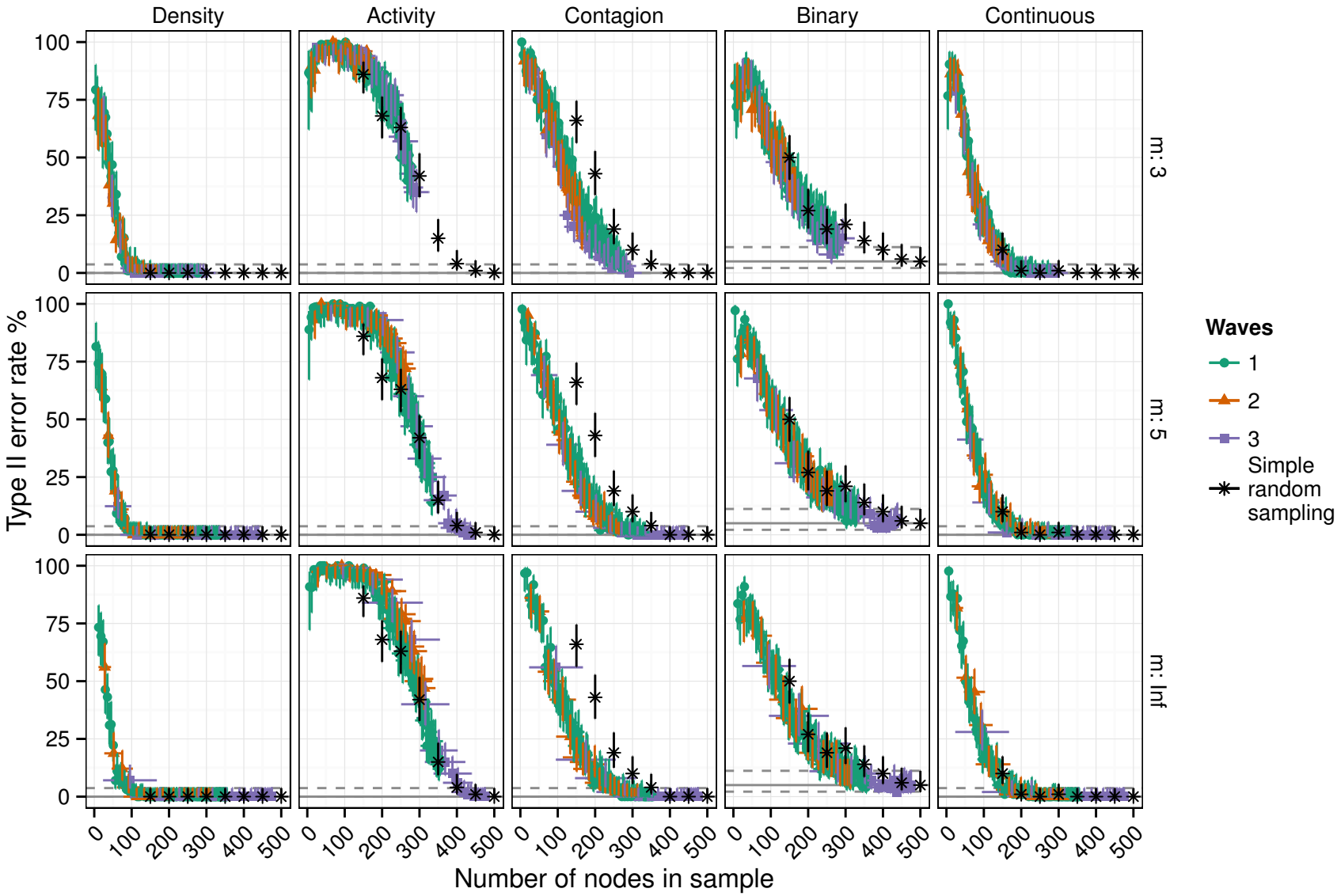
Network science collaboration network $N = 1589$



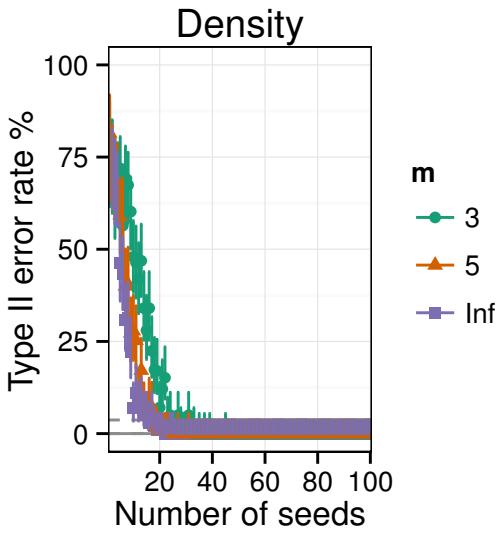
Effect on RMSE of random removal of nodes

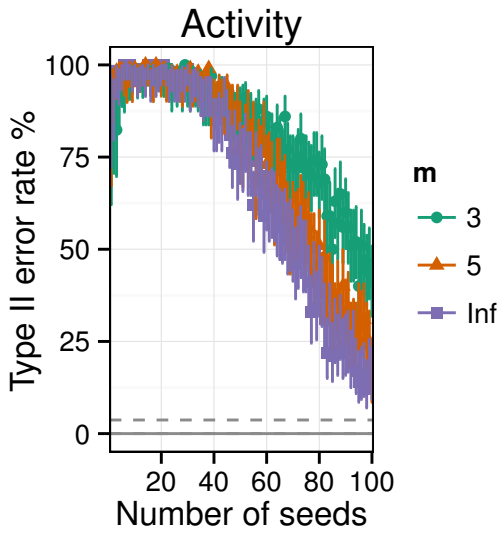


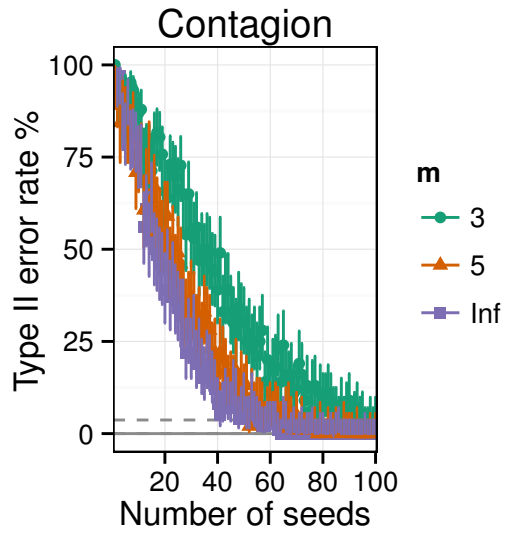
Snowball sample size effect on type II error rate

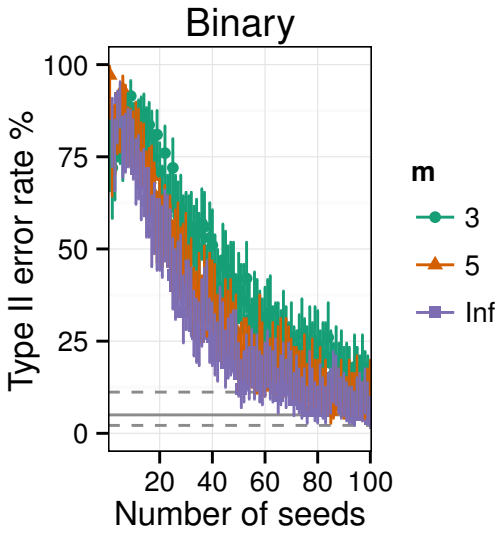


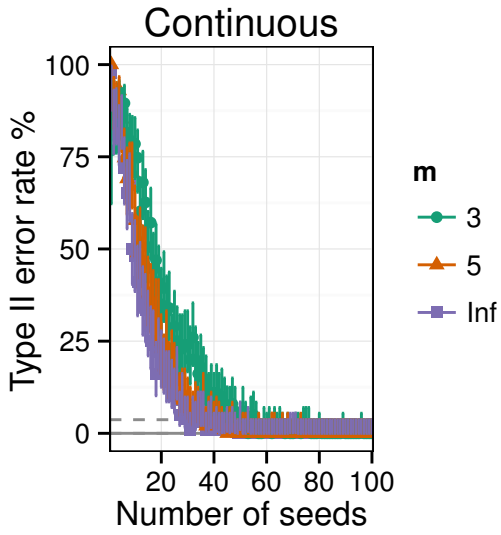
1 wave snowball sampling effect on type II error rate

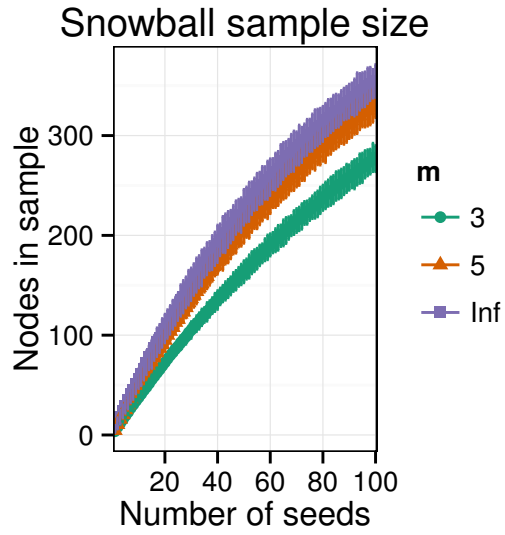




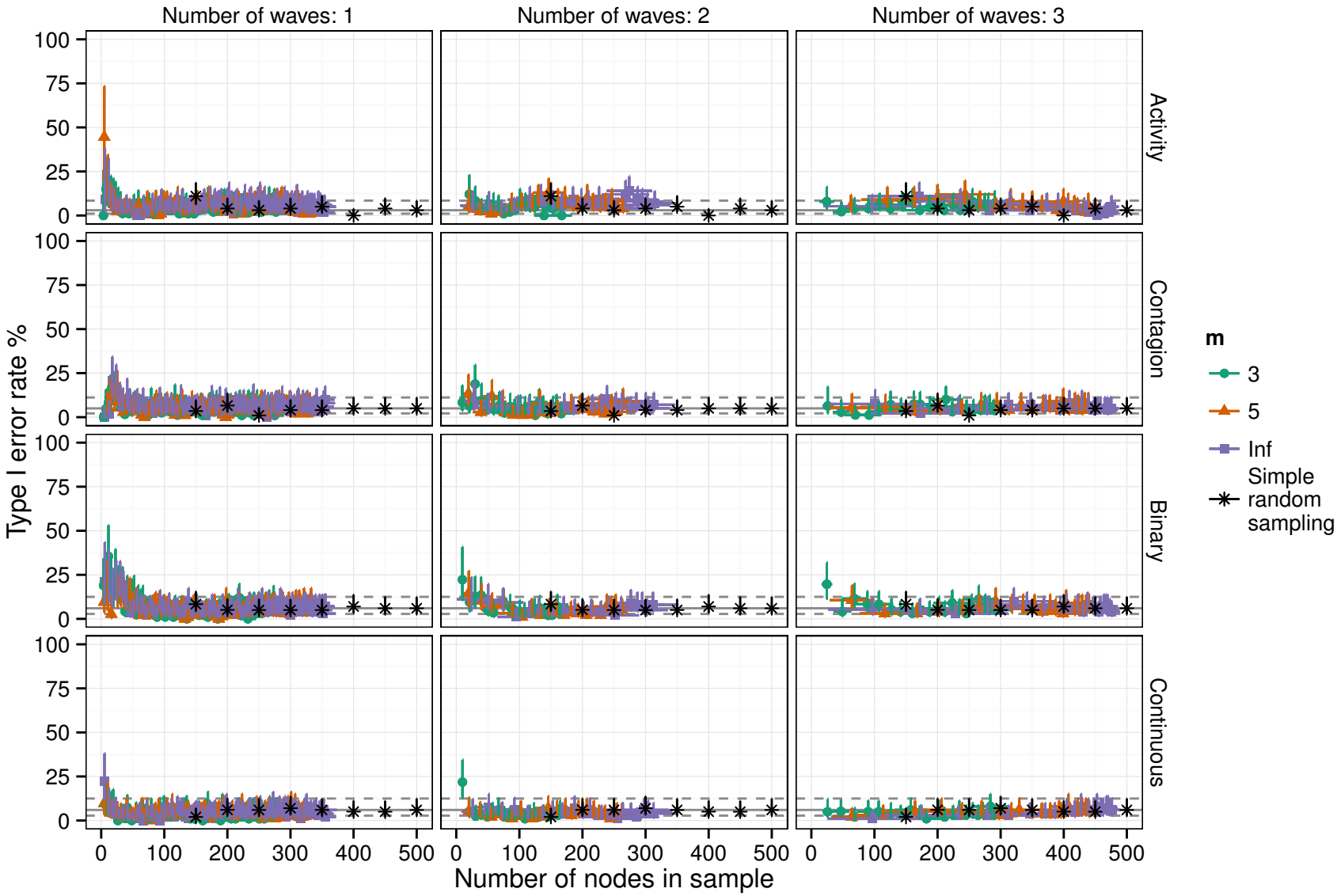




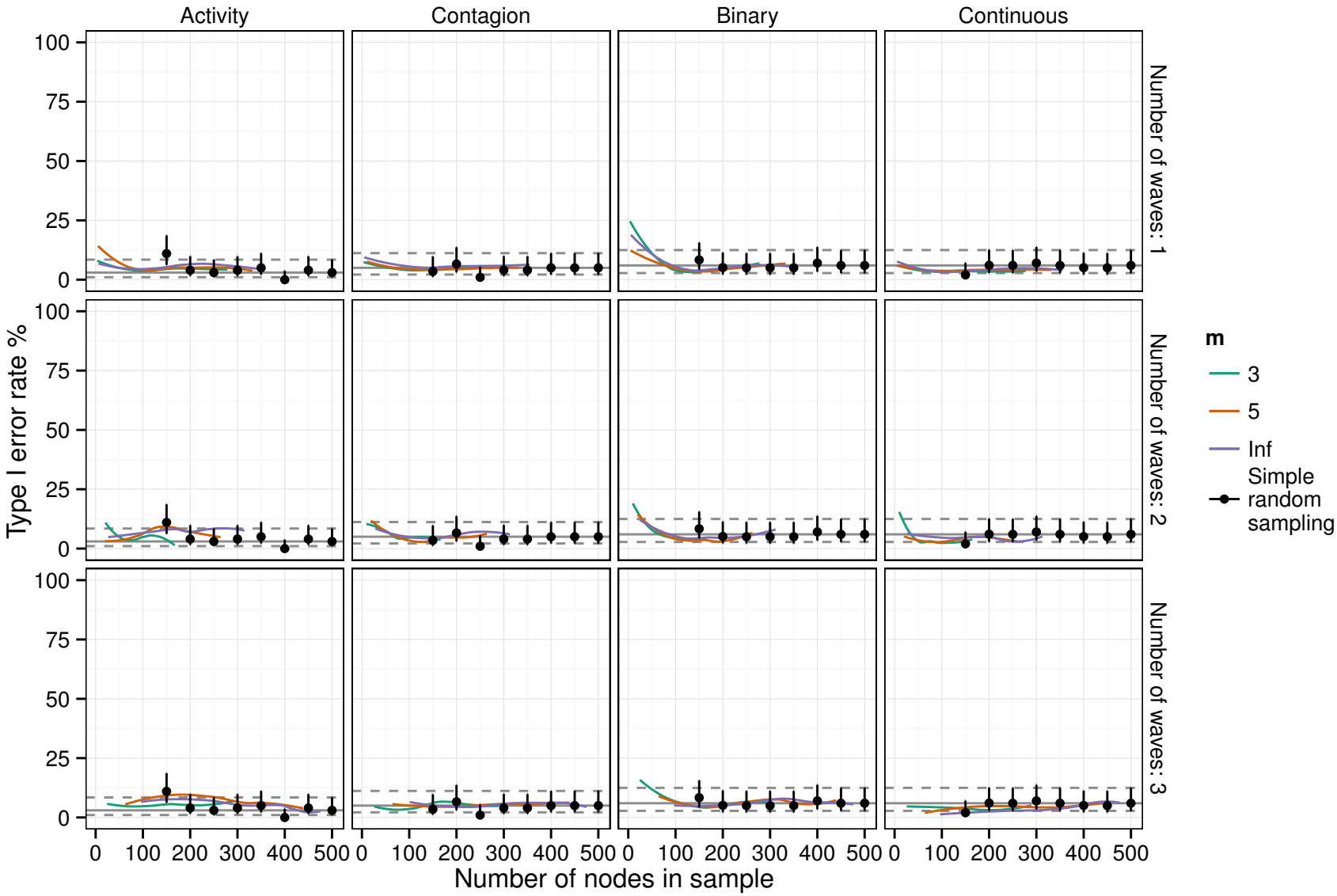




Snowball sample size effect on type I error rate



Snowball sample size effect on type I error rate (smoothed)



Positive outcome statistics of simulated ALAAMs with zero effect

N	Zero effect	Positive outcome %	
		mean	s.d.
500	-	15	2.19
500	Activity	14	2.08
500	Contagion	6	0.926
500	Binary	5	1.16
500	Continuous	9	1.87