Homophily and Community Structure at Scale: An Application to a Large Professional Network

Juan Nelson Martínez Dahbura martinez.dahbura@sansan.com
Shota Komatsu shota.komatsu@sansan.com
Takanori Nishida nishida@sansan.com
Sansan, Inc.

Angelo Mele angelo.mele@jhu.edu
Johns Hopkins University

ASSA Meeting 2023, New Orleans

web: http://meleangelo.github.io
R package: https://github.com/sansan-inc/lighthergm

Motivation

1. Professional networks:
   - how do they form?
   - how do they affect labor markets?
2. Observable and unobservable heterogeneity; network externalities
3. Structural models of network formation
   - Econometric challenges

     De Paula (2018), Chandrasekhar (2016), Graham (2017), Mele (2017,2021), Menzel (2016), Sheng (2021), De Paula et Al (2018), Graham (2016), Boucher and Mourifie (2017), Gao (2020), Badev (2021), Leung and Moon (2021) and many more...
   - Computational challenges

     Dahbura et al 2021, Gaonkar and Mele (2018), Babkin et al (2021), Mele et al (2021), Vu et al (2013)
4. Potential uses: recommentation system, effect of policy or shocks on professional networks, key player analysis

R package

https://github.com/sansan-inc/lighthergm

- Open Source library for scalable estimation of this class of models
- Solves memory problems of original `hergm` library
- Allows the inclusion of (discrete) covariates in the model
- Improvements in speed by a factor of 14000 for some operations

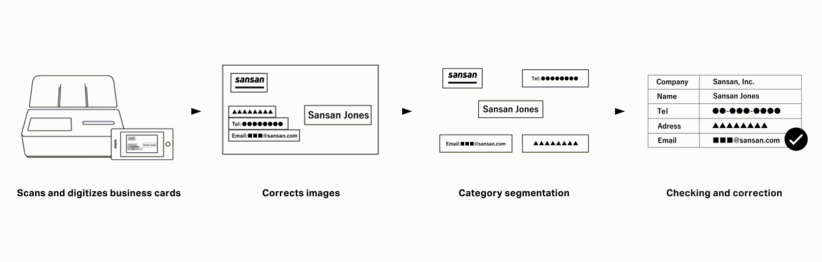
Estimation in the paper uses a Ubuntu Linux machine with 32GB of memory and 8 cores. The computation is performed with about 20 GB of memory for the block recovery step accounting for node covariates. All cores are in use during most of the calculation time

Eight

- In this work we use anonymized data from Eight on connections formed Jan-Dec 2019
- We include only users located in Tokyo who have uploaded a profile card at least once by the end of 2019 and have accepted terms of service
- We keep only nodes for which all covariates used in the analysis have non-missing values and that belong to the largest connected component of the network's 10-core.

Eight

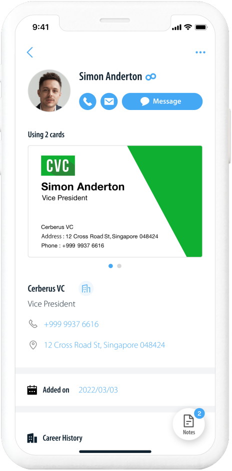
- Contact and career management app[1] with over 3 million users in Japan
- Allows users to scan physical business cards employing the smartphone's camera
- High quality digitization is achieved through the usage of advanced OCR algorithms and the help of human operators[2]

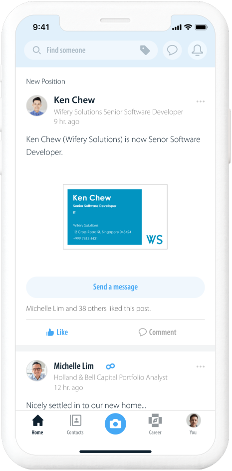


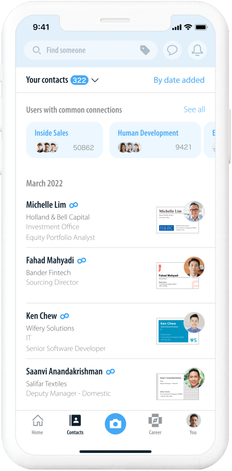
[1] https://8card.net/en/

[2] More about the digitization process at: https://bit.ly/3CqMOxp

Eight


9:41
Simon Anderton
Message
Using 2 cards
CVC
Simon Anderton
Vice President
Cerberus VC
Address : 12 Cross Road St, Singapore 048424
Phone : +999 9937 6616
Cerberus VC
Vice President
+999 9937 6616
12 Cross Road St, Singapore 048424
Added on 2022/03/03
Notes
Career History


9:41
Find someone
New Position
Ken Chew
Wifery Solutions Senior Software Developer
9 hr. ago
Ken Chew (Wifery Solutions) is now Senor Software Developer.
Ken Chew
WS
Send a message
Michelle Lim and 38 others liked this post.
Like
Comment
Michelle Lim
Holland & Bell Capital Portfolio Analyst
12 hr. ago
Nicely settled in to our new home...
Home
Contacts
Career
You


9:41
Find someone
Your contacts 322
By date added
Users with common connections
See all
Inside Sales
50862
Human Development
9421
March 2022
Michelle Lim
Holland & Bell Capital
Investment Office
Equity Portfolio Analyst
Fahad Mahyadi
Bander Fintech
Sourcing Director
Ken Chew
Wifery Solutions
IT
Senior Software Developer
Saanvi Anandakrishman
Salifar Textiles
Deputy Manager - Domestic
Home
Contacts
Career
You

Data

- The resulting network has $30,323$ nodes and $321,188$ edges.
- The network is very sparse, with a density of roughly 0.0007.
- The data is highly geographically concentrated. About 84% of the nodes are located in just five districts of Tokyo.
- We also observe industrial concentration, especially in the Technology (22%) and Consulting (14%) industries.

Model

Communities and sequential network formation

- Time is discrete: $t = 0, 1, 2, 3, ...$
- At $t = 0$ *Nature* assigns types

$$\boldsymbol{Z}_i \overset{iid}{\sim} Multinomial(1; \eta_1, ..., \eta_K) \tag{1}$$

Remark: types not too large wrt network
Remark: each node belongs to one type only
(extensions to multiple communities possible as in Airoldi et al 2008)

- Conditional on $\boldsymbol{Z} = \boldsymbol{z}$, **network $\boldsymbol{g}$ is formed sequentially.**
- In each period t
  1. Two users i and j meet
  2. Users receive random matching shock ε_{ij}
  3. Users decide whether to form/cut/keep link g_{ij}
     $\rightarrow$ **maximize surplus generated by g_{ij}**

Assumptions

1. Users have positive probability of meeting any user

$$\text{Prob. } i \text{ and } j \text{ meet} := \rho(g_{-ij}, z_i, z_j, x_i, x_j, n) > 0 \tag{2}$$

2. Payoff of user i

$$U_i\left(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}\right) = \sum_{j=1}^{n} g_{ij} \left(u_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \sum_{r \neq i,j}^{n} g_{jr} g_{ri} v_{ijr}(\boldsymbol{\gamma}) \right)$$

- $u_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} \alpha_w + \sum_{p=1}^{P} \beta_{wp} \mathbf{1}\{\boldsymbol{x}_{ip} = \boldsymbol{x}_{jp}\} & \text{if } i, j \text{ belong to same } k \\ \alpha_b + \sum_{p=1}^{P} \beta_{bp} \mathbf{1}\{\boldsymbol{x}_{ip} = \boldsymbol{x}_{jp}\} & \text{otherwise} \end{cases}$
- **Local transitivity**: $v_{ijr}(\gamma) = \begin{cases} \gamma & \text{if } i, j, r \text{ belong to same } k \\ 0 & \text{otherwise} \end{cases}$

3. Matching shock ε_{ij} is logistic iid

Equilibrium

PROPOSITION. Conditional on $\boldsymbol{z}$, the long-run network distribution factorizes into WITHIN- and BETWEEN-types components

$$\begin{aligned}
\pi(\boldsymbol{g},\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) &= \prod_{k=1}^{K} \frac{\exp\left[Q_{k,k}(\boldsymbol{g}_{k,k},\boldsymbol{x}^{(k)},\boldsymbol{z};\boldsymbol{\alpha}_w,\boldsymbol{\beta}_w,\gamma)\right]}{c_{k,k}(\mathcal{G}_{k,k},\boldsymbol{x}^{(k)};\boldsymbol{\theta})} \\
&\times \prod_{l>k}^{K}\prod_{i\in\mathcal{C}_k}\prod_{j\in\mathcal{C}_l} \frac{\exp\left[g_{ij}\left(u_{ij}(\alpha_b,\boldsymbol{\beta}_b)+u_{ji}(\alpha_b,\boldsymbol{\beta}_b)\right)\right]}{1+\exp\left[(u_{ij}(\alpha_b,\boldsymbol{\beta}_b)+u_{ji}(\alpha_b,\boldsymbol{\beta}_b)\right]}
\end{aligned}$$

where

$$\begin{aligned}
Q_{kk} &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_{ik}z_{jk}g_{ij}\left(u_{ij}(\alpha_w,\boldsymbol{\beta}_w)+\frac{2\gamma}{3}\sum_{r\neq i,j} z_{rk}g_{jr}g_{ri}\right) \\
c_{kk} &= \sum_{\omega\in\mathcal{G}} e^{Q_{kk}}
\end{aligned}$$

REMARK. In long-run → HERGM (Schweinberger-Handcock 2015)

Approximate Estimation

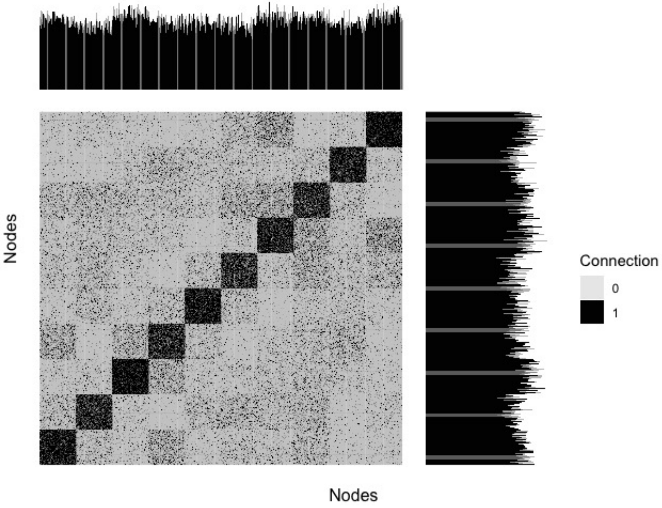
Approximate Maximum Likelihood

- **State-of-the-art**: Bayesian estimation (Mele JBES, forthcoming)
- For large networks, the Bayesian approach is *impractical or infeasible*

- On the other hand, if some conditions are satisfied:
  1. communities small enough and
  2. network large

     $\Longrightarrow$ most probability mass is **across** blocks

     $\Longrightarrow$ conditionally independent links
- Network resembles stochastic blockmodel *except within blocks*


Nodes
Nodes
Connection
0
1

Approximate Maximum Likelihood

Step 1: Compute approximate $\hat{z}$ using SBM ($\gamma = 0$)

- **Conditions for good approximation**: Schweinberger-Stewart 2021, Babkin et al 2020
- **Variational approximations**:
  Jordan and Wainwright 2003, Mele and Zhu 2023, Bickel and Chen 2013
- **Minorization-maximization**:
  Vu et al 2013
- **Spectral methods**:
  Athreya et al 2018, Mele et al 2022, Cong et al 2022, Hao et al 2022

Step 2: Approximate likelihood, given $\hat{z}$

- **Monte Carlo Maximum Likelihood (MCMC-MLE)**
  Geyer and Thompson 1992, Snijders 2002, Mele 2011
- **Maximum Pseudolikelihood (MPLE)**
  Snijders 2002, Boucher and Mourifie 2017, Babkin et al 2020

STEP 1: Variational Approximation

The full log-likelihood of our model can be written as follows

$$\begin{aligned}
\mathcal{L}(\boldsymbol{g},\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\eta}) &= \log\sum_{\boldsymbol{z}\in\mathcal{Z}} P_{\boldsymbol{\eta}}\left(\boldsymbol{Z}=\boldsymbol{z}\right)\pi(\boldsymbol{g},\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) = \log\sum_{\boldsymbol{z}\in\mathcal{Z}} L\left(\boldsymbol{g},\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\eta}\right) \\
&\approx \log\sum_{\boldsymbol{z}\in\mathcal{Z}} L\left(\boldsymbol{g},\boldsymbol{x},\boldsymbol{z};\boldsymbol{\alpha},\boldsymbol{\beta},\gamma=0,\boldsymbol{\eta}\right) && (3) \\
&= \log\sum_{\boldsymbol{z}\in\mathcal{Z}} \frac{q_{\xi}(\boldsymbol{z})}{q_{\xi}(\boldsymbol{z})} L_0\left(\boldsymbol{g},\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta},\boldsymbol{\eta}\right) \\
&\geq \ell_B(\boldsymbol{g},\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta};\boldsymbol{\xi}) && (4) \\
&= \sum_{i<j}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} \xi_{ik}\xi_{jl}\log\pi_{ij,kl}(g_{ij},\boldsymbol{x},\boldsymbol{z}) && (5) \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K} \xi_{ik}\left(\log\eta_k - \log\xi_{ik}\right)
\end{aligned}$$

where

$$\pi_{ij,kl}(g_{ij},\boldsymbol{x},\boldsymbol{z}) = \text{Prob } i \text{ and } j \text{ of type } k \text{ and } l \text{ are connected} \quad (6)$$

STEP 1: Minorization-Maximization

Find function approximating $\ell_B(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{\xi})$, but simpler to maximize.

$M\left(\boldsymbol{\xi}; \boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\xi}^{(s)}\right)$ minorizes $\ell_B(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{\xi})$ at $\boldsymbol{\xi}^{(s)}$ and iteration s if

$$M\left(\boldsymbol{\xi}; \boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\xi}^{(s)}\right) \leq \ell_B(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{\xi}) \quad \text{for all } \boldsymbol{\xi} \tag{7}$$

$$M\left(\boldsymbol{\xi}^{(s)}; \boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\xi}^{(s)}\right) = \ell_B(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{\xi}^{(s)}) \tag{8}$$

For stochastic blockmodels, Vu et al 2013 suggest

$$\begin{aligned} M\left(\boldsymbol{\xi}; \boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\xi}^{(s)}\right) &:= \sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(\xi_{ik}^2 \frac{\xi_{jl}^{(s)}}{2\xi_{ik}^{(s)}} + \xi_{jl}^2 \frac{\xi_{ik}^{(s)}}{2\xi_{jl}^{(s)}} \right) \log \pi_{ij;kl}^{(s)}(g_{ij}, \boldsymbol{x}, \boldsymbol{z}) \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik} \left(\log \eta_k^{(s)} - \log \xi_{ik}^{(s)} - \frac{\xi_{ik}}{\xi_{ik}^{(s)}} + 1 \right). \end{aligned} \tag{9}$$

Parallelizes to n independent maximization problems

Variational Updates with discrete covariates

The update rules for $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and $\pi_{ij;kl}(g_{ij}, \boldsymbol{x}, \boldsymbol{z})$ follow

$$\boldsymbol{\xi}^{(s+1)} := \arg\max_{\boldsymbol{\xi}} M\left(\boldsymbol{\xi}; \boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(s)}, \boldsymbol{\eta}^{(s)}, \boldsymbol{\xi}^{(s)}\right),$$

$$\eta_k^{(s+1)} := \frac{1}{n}\sum_{i=1}^{n} \xi_{ik}^{(s+1)}, \quad k = 1, \ldots, K,$$

and

$$\pi_{ij;kl}^{(s+1)}(d, \chi_1, \ldots, \chi_p, \boldsymbol{z}) := \frac{\sum_{i=1}^{n}\sum_{j\neq i} \xi_{ik}^{(s+1)}\xi_{jl}^{(s+1)} \mathbf{1}\{g_{ij} = d, \chi_{1,ij} = \chi_1, \ldots, \chi_{p,ij} = \chi_p\}}{\sum_{i=1}^{n}\sum_{j\neq i} \xi_{ik}^{(s+1)}\xi_{jl}^{(s+1)} \mathbf{1}\{\chi_{1,ij} = \chi_1, \ldots, \chi_{p,ij} = \chi_p\}},$$

for $k, l = 1, \ldots, K$ and $d, \chi_1, \ldots, \chi_p \in \{0, 1\}$, respectively.

$\chi_{p,ij} = \mathbf{1}\{x_{ip} = x_{jp}\}$. Generalizations of this specification are allowed.

STEP 2: Maximum Pseudolikelihood Estimator

Given estimated $\widehat{\boldsymbol{z}}$, compute conditional prob of link

WITHIN BLOCKS

$$p_{ij}(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\theta}; \widehat{\boldsymbol{z}}) = \Lambda\left(u_{ij}(\boldsymbol{\alpha}_w, \boldsymbol{\beta}_w) + u_{ji}(\boldsymbol{\alpha}_w, \boldsymbol{\beta}_w) + 4\gamma \sum_{r \neq i,j} g_{jr} g_{ir} \right)$$

BETWEEN BLOCKS

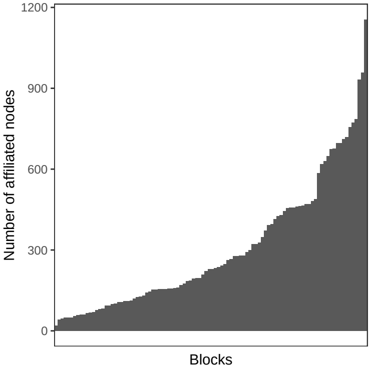
$$p_{ij}(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\theta}; \widehat{\boldsymbol{z}}) = \Lambda\left(u_{ij}(\boldsymbol{\alpha}_b, \boldsymbol{\beta}_b) + u_{ji}(\boldsymbol{\alpha}_b, \boldsymbol{\beta}_b) \right)$$

where $\Lambda(u) = e^u/(1+e^u)$ is the logistic function.

The **pseudolikelihood estimator** solves

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{PL} &= \arg\max_{\boldsymbol{\theta}} \ell_{PL}(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\theta}; \widehat{\boldsymbol{z}}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{j>i}^{n} \left[g_{ij} \log p_{ij}(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\theta}) + (1 - g_{ij}) \log(1 - p_{ij}(\boldsymbol{g}, \boldsymbol{x}, \boldsymbol{\theta})) \right] \end{aligned}$$

Empirical results: block size



	Between	Within
	(1)	(2)
Intercept (α)	-7.709^{***}	-4.754^{***}
	(0.002)	(0.005)
Shared Contacts (γ)		0.736^{***}
		(0.004)
Same Location (β_1)	0.333^{***}	0.006
(H3 Tile)	(0.007)	(0.012)
Same Industry (β_2)	0.694^{***}	0.034^{***}
	(0.005)	(0.009)
Same Occupation (β_3)	0.409^{***}	0.041^{***}
	(0.006)	(0.010)
Bayesian Inf. Crit.	4,171,768	808,597

Notes: *p<0.1; **p<0.05; ***p<0.01.
All estimates are obtained using a maximum pseudolikelihood estimator, conditioning on the estimated block structure. Block recovery was performed for a total of 100 blocks. We employ 20,000 EM iterations without employing node covariates, and a final run of 100 iterations employing covariates.

Summary

Summary

- Equilibrium model with community structure
- Approximate maximum likelihood
  - Use SBM likelihood
  - Variational Approximations for SBM

    Bickel et al 2013; Jordan and Wainwright 2003; Mele and Zhu 2020
  - Use Minorization algorithm to speed up computation

    Vu et al 2013; Babkin et al 2021
- Find evidence of homophily and transitivity (see also Dahbura et al 2021)

In progress

- Improve package speed: MPLE vs. MC-MLE; Spectral Methods instead of Variational approximations; initialization with InfoMap; Other clustering methods: Louvain, etc.
- Estimation with more covariates (discrete)
- Goodness of fit, counterfactual simulations
- Effect of networks on outcomes: key player simulations

THANK YOU!

More of this at:
arxiv: **https://arxiv.org/abs/2105.12704**

Contact:
Juan Nelson Martínez Dahbura **martinez.dahbura@sansan.com**
Shota Komatsu **shota.komatsu@sansan.com**
Takanori Nishida **nishida@sansan.com**
Angelo Mele **angelo.mele@jhu.edu**

R package: **https://github.com/sansan-inc/lighthergm**

Does it work?
Monte Carlo experiments

Monte Carlo simulation to check approximate ML

Specification of payoff - no covariates

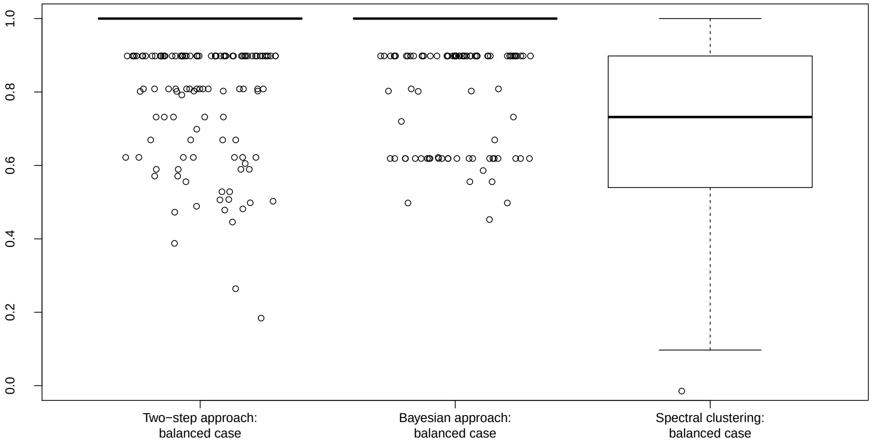
$$\begin{aligned} U_i(g,x,z;\theta) &= \theta_1 \sum_{k=1}^{K}\sum_{j=1}^{n} z_{ik}z_{jk}g_{ij} + \theta_2 \sum_{j=1}^{n} z_{ik}z_{jk}z_{rk}g_{ij}\mathbf{1}_{ij} \\ &+ \theta_B \sum_{k=1}^{K}\sum_{\ell>k}\sum_{j=1}^{n} z_{ik}z_{j\ell}g_{ij} \end{aligned}$$

where $\mathbf{1}_{ij} = 1$ if i and j have at least 1 partner in common

- $n = \{30, 150, 2500\}$
- $K = \{3, 100\}$
- Model with no covariates, only edges and transitive triples
- Simulate 500 networks and estimate model
- Parameters $(\theta_1, \theta_2) = (-1, .5)/\log(n_k)$ for $n = \{30, 150\}$
- Parameters $(\theta_1, \theta_2) = (-2, 1)/\log(n_k)$ for $n = 2500$
- Parameter $\theta_B = -3/\log(n)$

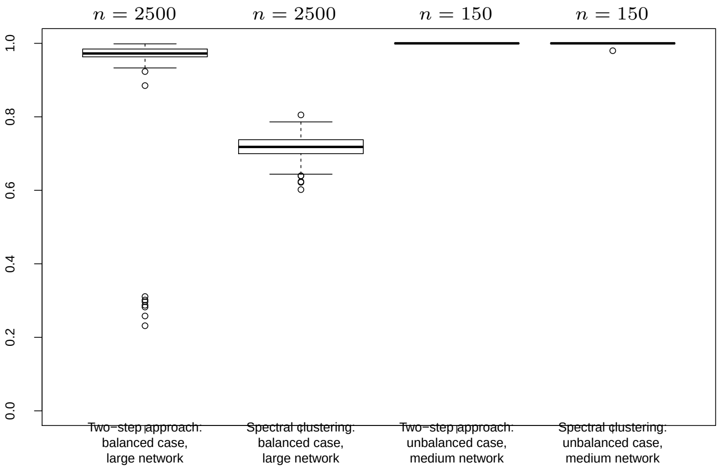
Results: Approximate ML vs Bayesian vs Spectral

Estimation of $\hat{z}$, $n = 30, K = 3$



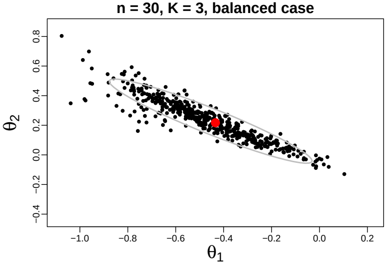
Results: Approximate ML

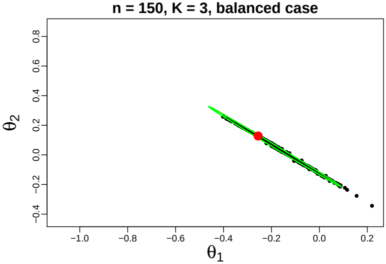
Estimation of $\hat{z}$, large and medium size networks

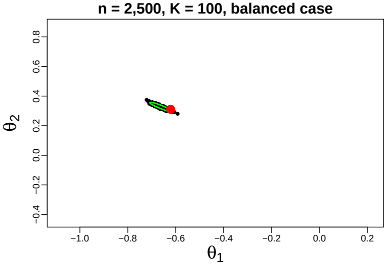


Results: Approximate ML

Parameters estimates, point estimates and 95% ellipses







Results: Approximate ML

Parameters estimates as a function of misclassification rate for z