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
   - Computational challenges

4. Potential uses: recommendation system, effect of policy or shocks on professional networks, key player analysis
Introduction and motivation

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.
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.
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/3CqM0xp
Introduction and motivation
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: $t = 0, 1, 2, 3, \ldots$
- At $t = 0$ Nature assigns types

$$Z_i \sim iid \sim Multinomial(1; \eta_1, \ldots, \eta_K)$$  \hspace{1cm} (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 $Z = z$, network $g$ is formed sequentially.
- In each period $t$
  1. Two users $i$ and $j$ meet  
  2. Users receive random matching shock $\varepsilon_{ij}$  
  3. Users decide whether to form/cut/keep link $g_{ij}$  
     \hspace{1cm} $\rightarrow$ **maximize surplus generated by** $g_{ij}$
Assumptions

1. Users have positive probability of meeting any user

\[ \text{Prob. } i \text{ and } j \text{ meet} := \rho(g_{ij}, z_i, z_j, x_i, x_j, n) > 0 \]  \hspace{1cm} (2)

2. Payoff of user \( i \)

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

- \( u_{ij}(\alpha, \beta) = \begin{cases} \alpha_w + \sum_{p=1}^{P} \beta_{wp}1\{x_{ip} = x_{jp}\} & \text{if } i, j \text{ belong to same } k \\ \alpha_b + \sum_{p=1}^{P} \beta_{bp}1\{x_{ip} = x_{jp}\} & \text{otherwise} \end{cases} \)

- **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 \( \varepsilon_{ij} \) is logistic iid
**Equilibrium**

**PROPOSITION.** Conditional on \( z \), the long-run network distribution factorizes into **WITHIN-** and **BETWEEN-types** components

\[
\pi(g, x, z; \theta) = \prod_{k=1}^{K} \exp \left[ Q_{kk}(g_{k,k}, x^{(k)}, z; \alpha_w, \beta_w, \gamma) \right] \frac{c_{k,k}(G_{k,k}, x^{(k)}; \theta)}{1 + \exp \left[ u_{ij}(\alpha_b, \beta_b) + u_{ji}(\alpha_b, \beta_b) \right]}
\]

where

\[
Q_{kk} = \sum_{i=1}^{n} \sum_{j=1}^{n} z_i \cdot z_j \cdot g_{ij} \left( u_{ij}(\alpha_w, \beta_w) + \frac{2\gamma}{3} \sum_{r \neq i,j} z_{rk} g_{jr} g_{ri} \right)
\]

\[
c_{kk} = \sum_{\omega \in G} e^{Q_{kk}}
\]

**REMARK.** In long-run \( \rightarrow \) 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
     \[\implies\] most probability mass is **across** blocks
     \[\implies\] conditionally independent links

- Network resembles stochastic blockmodel _except within blocks_
Empirical Strategy

Approximate Maximum Likelihood

Homophily and Community Structure at Scale
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)**

- **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

\[ L(g, x; \theta, \eta) = \log \sum_{z \in Z} P_{\eta}(Z = z) \pi(g, x, z; \theta) = \log \sum_{z \in Z} L(g, x, z; \theta, \eta) \]

\[ \approx \log \sum_{z \in Z} L(g, x, z; \alpha, \beta, \gamma = 0, \eta) \]

\[ = \log \sum_{z \in Z} \frac{q_{\xi}(z)}{q_{\xi}(z)} L_0(g, x, z; \theta, \eta) \]

\[ \geq \ell_B(g, x, \alpha, \beta, \eta; \xi) \]

\[ = \sum_{i<j} \sum_{k=1}^{K} \sum_{l=1}^{K} \xi_{ik} \xi_{jl} \log \pi_{ij,kl}(g_{ij}, x, z) \]

\[ + \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik} (\log \eta_k - \log \xi_{ik}) \]

where

\[ \pi_{ij,kl}(g_{ij}, x, z) = \text{Prob } i \text{ and } j \text{ of type } k \text{ and } l \text{ are connected} \]
**STEP 1: Minorization-Maximization**

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

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

$$M \left( \xi; g, x, \alpha, \beta, \eta, \xi^{(s)} \right) \leq \ell_B(g, x, \alpha, \beta, \eta; \xi) \text{ for all } \xi$$  \hspace{1cm} (7)

$$M \left( \xi^{(s)}; g, x, \alpha, \beta, \eta, \xi^{(s)} \right) = \ell_B(g, x, \alpha, \beta, \eta; \xi^{(s)})$$  \hspace{1cm} (8)

For stochastic blockmodels, Vu et al 2013 suggest

$$M \left( \xi; g, x, \alpha, \beta, \eta, \xi^{(s)} \right) := \sum_{i<j} \sum_k \sum_{l=1}^K \left( \xi_{ik}^2 \frac{\xi_{jl}^{(s)}}{2\xi_{ik}^{(s)}} + \xi_{jl}^2 \frac{\xi_{ik}^{(s)}}{2\xi_{jl}^{(s)}} \right) \log \pi_{ij;kl}^{(s)}(g_{ij}, x, z)$$

$$+ \sum_{i=1}^n \sum_{k=1}^K \xi_{ik} \left( \log \eta_k^{(s)} - \log \xi_{ik}^{(s)} - \frac{\xi_{ik}^{(s)}}{\xi_{ik}} + 1 \right).$$  \hspace{1cm} (9)

Parallelizes to $n$ independent maximization problems
Variational Updates with discrete covariates

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

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

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

and

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

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

$\chi_{p,ij} = 1\{x_{ip} = x_{jp}\}$. Generalizations of this specification are allowed.
STEP 2: Maximum Pseudolikelihood Estimator

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

**WITHIN BLOCKS**

$$p_{ij}(g, x, \theta; \hat{z}) = \Lambda \left( u_{ij}(\alpha_w, \beta_w) + u_{ji}(\alpha_w, \beta_w) + 4\gamma \sum_{r \neq i, j} g_{jr}g_{ir} \right)$$

**BETWEEN BLOCKS**

$$p_{ij}(g, x, \theta; \hat{z}) = \Lambda \left( u_{ij}(\alpha_b, \beta_b) + u_{ji}(\alpha_b, \beta_b) \right)$$

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

The **pseudolikelihood estimator** solves

$$\hat{\theta}_{PL} = \arg \max_{\theta} \ell_{PL}(g, x, \theta; \hat{z})$$

$$\quad = \arg \max_{\theta} \sum_{i=1}^{n} \sum_{j>i}^{n} \left[ g_{ij} \log p_{ij}(g, x, \theta) + (1 - g_{ij}) \log(1 - p_{ij}(g, x, \theta)) \right]$$
Empirical results: block size
<table>
<thead>
<tr>
<th></th>
<th>Between (1)</th>
<th>Within (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\alpha$)</td>
<td>$-7.709^{***}$</td>
<td>$-4.754^{***}$</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Shared Contacts ($\gamma$)</td>
<td></td>
<td>$0.736^{***}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.004)</td>
</tr>
<tr>
<td>Same Location ($\beta_1$)</td>
<td>$0.333^{***}$</td>
<td>0.006</td>
</tr>
<tr>
<td>(H3 Tile)</td>
<td>(0.007)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>Same Industry ($\beta_2$)</td>
<td>$0.694^{***}$</td>
<td>0.034^{***}</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Same Occupation ($\beta_3$)</td>
<td>$0.409^{***}$</td>
<td>0.041^{***}</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Bayesian Inf. Crit.</td>
<td>4,171,768</td>
<td>808,597</td>
</tr>
</tbody>
</table>

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

\[ U_i(g, x, z; \theta) = \theta_1 \sum_{k=1}^{K} \sum_{j=1}^{n} z_{ik} z_{jk} g_{ij} + \theta_2 \sum_{j=1}^{n} z_{ik} z_{jk} z_{rk} g_{ij} 1_{ij} \]

\[ + \theta_B \sum_{k=1}^{K} \sum_{\ell > k} \sum_{j=1}^{n} z_{ik} z_{j\ell} g_{ij} \]

where \( 1_{ij} = 1 \) 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

Two-step approach: balanced case, large network
Spectral clustering: balanced case, large network
Two-step approach: unbalanced case, medium network
Spectral clustering: unbalanced case, medium network
Results: Approximate ML

Parameters estimates, point estimates and 95% ellipses

- **n = 30, K = 3, balanced case**
- **n = 150, K = 3, balanced case**
- **n = 2,500, K = 100, balanced case**
Results: Approximate ML

Parameters estimates as a function of misclassification rate for $z$

$n = 1000, K = 50, \text{true } z$

$n = 1000, K = 50, 1\% \text{ misclassified}$

$n = 1000, K = 50, 5\% \text{ misclassified}$

$n = 1000, K = 50, 10\% \text{ misclassified}$