AEA Papers and Proceedings 2023, 113: 156–160
https://doi.org/10.1257/pandp.20231094

BIG DATA: AT-SCALE METHODS AND APPLICATIONS

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

By Juan Nelson Martínez Dahbura, Shota Komatsu, Takanori Nishida, and Angelo Mele*

Professional and business networks are an important determinant of labor market outcomes and efficiency. We study the emergence of such networks, using anonymized data on over 30,000 users of Eight, a contact and career management app with over three million users in Japan.

Our empirical analysis is guided by a structural model of network formation with observed and unobserved individual heterogeneity. In our model, users are first randomly assigned to one of a finite number of unobservable types. The network is then formed sequentially, as users randomly meet and establish business connections based on their benefits and costs of forming, maintaining, and deleting a link.

The Eight data offer a unique view into the mechanisms behind the formation of face-to-face professional networks at scale, since the exchange of business cards is a traditional business practice in Japan. Eight's users have access to a mobile app to scan and manage the business cards they receive and to become contacts within the Eight professional network. We use these links and (some) individual covariates as our network data.

We overcome several computational challenges that plague these structural models by using a highly scalable two-step estimation method. The first step makes efficient use of the sparsity of the network and information on observable characteristics to recover the unobserved types, through computationally convenient approximations of the likelihood function. In the second step, a pseudolikelihood estimator recovers the structural parameters of the utility functions, given the estimated unobserved heterogeneity.

Our results bring light into the role of homophily and shared professional contacts on the emergence of business networks.

I. Model

We model the Eight users' decisions to form, maintain, and delete links. Individuals are characterized by a vector of observables x_i and unobservables z_i, both with finite support. Each user belongs to one of K types, so $z_i = (z_{i1}, \ldots, z_{iK})$ and user i belongs to type k if $z_{ik} = 1$. The network is described by the (symmetric) adjacency matrix g, with entries $g_{ij} = 1$ if users i and j are linked and $g_{ij} = 0$ otherwise. The utility for user i from network g is given by

$$U_i(g) = \sum_{j=1}^{n} g_{ij}\Big(u_{ij} + \sum_{r\neq i,j} g_{jr}\, g_{ri}\, v_{ijr}\Big),$$

where $u_{ij} = u_{ij}^{w} := \alpha_w + \sum_{p=1}^{P} \beta_{wp}\, \mathbf{1}_{(x_{ip}=x_{jp})}$ if $z_i = z_j$ and $u_{ij} = u_{ij}^{b} := \alpha_b + \sum_{p=1}^{P} \beta_{bp} \mathbf{1}_{(x_{ip}=x_{jp})}$ if $z_i \neq z_j$, and where $v_{ijr} = \gamma$ if $z_i = z_j = z_r$ and $v_{ijr} = 0$ otherwise. In this formulation, u_{ij} is the net utility of forming a direct link between i and j, including both costs and benefits of each link, and v_{ijr} is the payoff that i receives because of the common connections with j. In our specification, we allow u_{ij} to depend on observables and unobservables, and v_{ijr} is normalized to 0 when considering users of different unobserved types.

The types z_i's are randomly assigned at time $t = 0$ and do not change over time. Links

*Martínez Dahbura: Sansan, Inc. (email: martinez.dahbura@sansan.com); Komatsu: Sansan, Inc. (email: shota.komatsu@sansan.com); Nishida: Sansan, Inc. (email: nishida@sansan.com); Mele: Johns Hopkins Carey Business School (email: angelo.mele@jhu.edu). We thank Michael Schweinberger for comments and suggestions.

[†] Go to https://doi.org/10.1257/pandp.20231094 to visit the article page for additional materials and author disclosure statement(s).

are formed sequentially, and in each period $t = 1, 2, \ldots$, we have the following:

(i) Users i and j meet with probability $\rho_{ij} > 0$.

(ii) Users i and j observe a logistic matching shock ε_{ij}, iid among players and time.

(iii) Users i and j decide whether to form or delete the link g_{ij} by maximizing their joint surplus.

Mele (2022) and Dahbura et al. (2021) show that under mild assumptions and conditioning on the unobserved types z, this network formation process leads to a long-run equilibrium distribution of networks $\pi(g, x, z; \theta)$,

$$(1)\quad \pi(g, x, z; \theta) \;=\; \prod_{k=1}^{K} \frac{e^{Q_{kk}}}{c_{kk}} \left[\prod_{l>k}^{K} \prod_{ij} \frac{e^{\Delta_{ij}}}{1 + e^{\Delta_{ij}}} \right],$$

where $\Delta_{ij} \;=\; z_{ik} z_{jl} g_{ij} u_{ij}^{b}$,

$$Q_{kk} \;=\; \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ik} z_{jk} g_{ij} \left(u_{ij}^{w} + \frac{2\gamma}{3} \sum_{r \neq i,j} z_{rk} g_{jr} g_{ri} \right),$$

and $c_{kk} = \sum_{\omega \in \mathcal{G}} \exp(Q_{kk})$. This shows that the likelihood of observing a network g can be decomposed in within- and between-type likelihood contributions. The between-type likelihood consists of independent links because the externality v_{ijr} is normalized to zero, while the within-type likelihood consists of K independent exponential random graphs.

II. Scalable Two-Step Estimation

We take a random effects approach and assume that z_i's are independent from observables and the network,

$$(2)\quad z_i \overset{iid}{\sim} p_\eta(z) = Multinomial(1; \eta_1, \ldots, \eta_K).$$

Let $L(g, x, z, \theta, \eta) := p_\eta(z) \pi(g, x, z; \theta)$, so complete log-likelihood is

$$(3)\quad \mathcal{L}(g, x; \theta, \eta) \;=\; \log \sum_{z \in \mathcal{Z}} L(g, x, z, \theta, \eta).$$

Estimation of this model is challenging for two reasons. First, the likelihood $\pi(g, x, z; \theta)$ is proportional to a normalizing constant that is impractical or infeasible to compute. Second, the complete likelihood involves integrating over all possible unobservable block structures z, which is also impractical.

We therefore perform approximate inference based on an approximation of the likelihood (3). We exploit the fact that our model corresponds to a stochastic block model when $\gamma = 0$. Indeed, most of the links are across blocks, while the contribution to the likelihood of the within-type links is small. Therefore, we estimate the types assignments using the approximate stochastic block model likelihood, setting $\gamma = 0$ for the community discovery step (Babkin et al. 2020). Let $L_0(g, x, z, \alpha, \beta, \eta) := p_\eta(z) \pi(g, x, z; \alpha, \beta, \gamma = 0)$, so we have $L(g, x, z, \theta, \eta) \approx L_0(g, x, z, \alpha, \beta, \eta)$. To estimate the stochastic block model, we use a variational mean-field approximation approach (Bickel et al. 2013; Babkin et al. 2020). This amounts to finding the approximate multinomial distribution $q_\xi(z) = \prod_{i=1}^{n} q_{\xi_i}(z_i)$ that minimizes the Kullback-Leibler divergence from the true likelihood; we thus obtain a lower bound $\ell_B(g, x, \alpha, \beta, \eta; \xi)$ to the log-likelihood

$$\begin{aligned} \mathcal{L}(g, x; \theta, \eta) \;&\approx\; \log \sum_{z \in \mathcal{Z}} L_0(g, x, z, \alpha, \beta, \eta) \\ &\geq\; \ell_B(g, x, \alpha, \beta, \eta; \xi) \\ &=\; \sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \xi_{ik} \xi_{jl} \log \pi_{ij,kl}(x) \\ &\quad +\; \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik} (\log \eta_k - \log \xi_{ik}), \end{aligned}$$

where $\pi_{ij,kl}(x)$ is the probability that i and j of type k and l are connected. Maximizing the lower bound $\ell_B(g, x, \alpha, \beta, \eta; \xi)$ is still computationally intensive for a large network; thus, we resort to a minorization approach first proposed in Vu, Hunter, and Schweinberger (2013). At iteration $s + 1$ of the algorithm, we find the ξ that maximizes the minorizer $M(\xi, \xi^{(s)})$ of the lower bound

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

The maximization of $M\left(\xi,\xi^{(s)}\right)$ amounts to the solution of n independent maximization problems, amenable to massive parallelization. The update rules for ξ, η, and $\pi_{ij;kl}(x)$ follow

$$\xi^{(s+1)} := \underset{\xi}{\arg\max}\, M\left(\xi,\xi^{(s)}\right),$$

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

$$\pi_{ij;kl}^{(s+1)}(x) := \frac{\sum_{i=1}^{n}\sum_{j\neq i}\xi_{ik}^{(s+1)}\,\xi_{jl}^{(s+1)}\,\mathbf{1}\{g_{ij},\chi_{ij}\}}{\sum_{i=1}^{n}\sum_{j\neq i}\xi_{ik}^{(s+1)}\,\xi_{jl}^{(s+1)}\,\mathbf{1}\{\chi_{ij}\}},$$

for $k,l = 1,\ldots,K$, where $\chi_{ij} = \left\{\chi_{1,ij},\ldots,\chi_{P,ij}\right\}$ and $\chi_{p,ij} = \mathbf{1}\left\{x_{ip} = x_{jp}\right\}$. Once this variational expectation-maximization (EM) algorithm has converged, we assign $\hat{z}$ according to the modal type of each node. Given the estimated types $\hat{z}$, we then estimate the structural parameters $\theta = (\alpha,\beta,\gamma)$ using a pseudolikelihood estimator. Let p_{ij}^{w} and p_{ij}^{b} be the conditional probabilities of links within- and between-type, respectively:

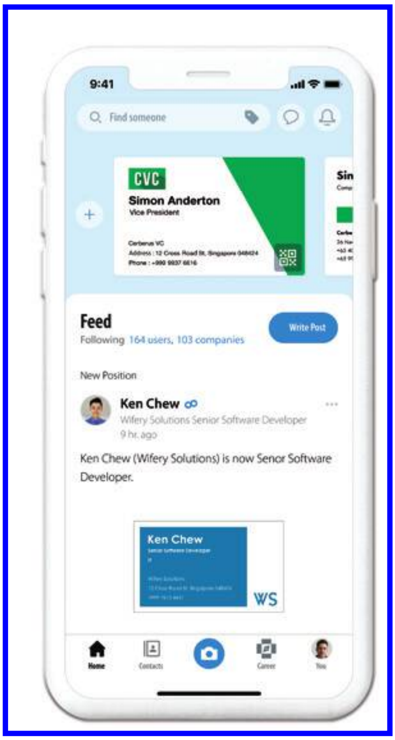
$$p_{ij}^{w} = \Lambda\left(u_{ij}^{w} + u_{ji}^{w} + 4\gamma\sum_{r\neq i,j} I_{ijr}\, g_{jr}\, g_{ir}\right)$$

$$p_{ij}^{b} = \Lambda\left(u_{ij}^{b} + u_{ji}^{b}\right),$$

where $\Lambda(u) = e^{u}/(1+e^{u})$ is the logistic function and $I_{ijr} = 1$ if $z_i = z_j = z_r$. The pseudolikelihood estimator solves

$$\hat{\theta}_{PL} = \underset{\theta}{\arg\max}\sum_{i=1}^{n}\sum_{j>i}^{n}\Big[g_{ij}\log p_{ij} + (1-g_{ij})\log(1-p_{ij})\Big]$$

with $p_{ij} = p_{ij}^{w}$ if $z_i = z_j$ and $p_{ij} = p_{ij}^{b}$ otherwise. Asymptotic theory for this estimator is contained in Boucher and Mourifie (2017).



CVC
Simon Anderton
Vice President
Cerberus VC
Address : 12 Cross Road St, Singapore 048424
Phone : +990 9937 6616

Ken Chew
Wifery Solutions Senior Software Developer
9 hr. ago

Ken Chew (Wifery Solutions) is now Senor Software Developer.

FIGURE 1. THE EIGHT MOBILE APP

Notes: The Eight mobile app allows users to scan physical business cards employing the smartphone's camera. High-quality digitization is achieved through the usage of advanced optical character recognition algorithms and the help of human operators.

III. Data and Results

We use anonymized data on business card exchanges among users of Eight who have agreed to the terms of usage of the service. Users upload their own business card when creating their profile, which allows us to observe information about them as well as their business connections with other users. We employ information on the user's geocoded location based on the address in their business card. We map the coordinates to an index of the H3 indexing system,[1] which represents a tile of roughly 5.17 squared kilometers. We also employ data on their occupation type and the industrial classification of the company they work for. The rest of the analysis is performed on a subnetwork containing only users located in Tokyo. We employ k-core decomposition to extract the connected component that exhibits a minimum degree of 10. The resulting graph is formed by 30,323 nodes and 321,188 edges. Most users in the data are employed in companies in the technology (22 percent) and consulting (14 percent) industries. The most common occupational categories are sales-related positions (17 percent) followed by company directors (15 percent). We also observe a large degree of geographic concentration: about 84 percent of the nodes are located in

[1] https://eng.uber.com/h3/

just five districts of Tokyo. The network is quite sparse, with a density of 0.0007. The median degree is 16 and the maximum degree is 343, reflecting the highly skewed degree distribution of the network.

Our implementation of the variational EM algorithm is an improved version of the hierarchical exponential-family random graph model (hergm) R package (Schweinberger and Luna 2018). The "lighthergm" package makes it possible to perform the block recovery step on networks with hundreds of thousands of nodes with model specifications including discrete covariates. Scalability is achieved through the usage of sparse matrices and by favoring matrix algebra over nested loops, which allows us to massively parallelize most computations. The code is publicly available on GitHub.[2] Further implementation details can be found in Dahbura et al. (2021). We estimate the model with $K = 100$ types. We initialize the types affiliations with the Infomap algorithm by Rosvall, Axelsson, and Bergstrom (2009). We run 20,000 iterations of the variational EM algorithm without employing the information on node covariates to achieve a better starting partition at a relatively low computational cost. Finally, we apply 100 EM iterations starting from the partition obtained in the previous step, using the version of the algorithm that uses the full specification with node covariates. This is the most complex version of the algorithm in terms of memory consumption and computation. For our network, one iteration takes roughly 1.3 minutes, and the whole computation can be performed with under 20 gigabytes of memory. This is a considerable improvement in terms of processing time and memory usage over previous implementations in Schweinberger and Luna (2018), which did not allow for node covariates. Figure 2, panel A shows the level of the lower bound at each iteration when applying the EM algorithm with covariates.

Figure 2, panel B shows the distribution of block size. The distribution is skewed, with a median size of 230 nodes and a maximum size of 1,155 nodes, although the largest block only accounts for roughly 3.8 percent of the total number of nodes. The resulting partition is an improvement over the initial value obtained from Infomap, which groups roughly half of the nodes in the four largest blocks.

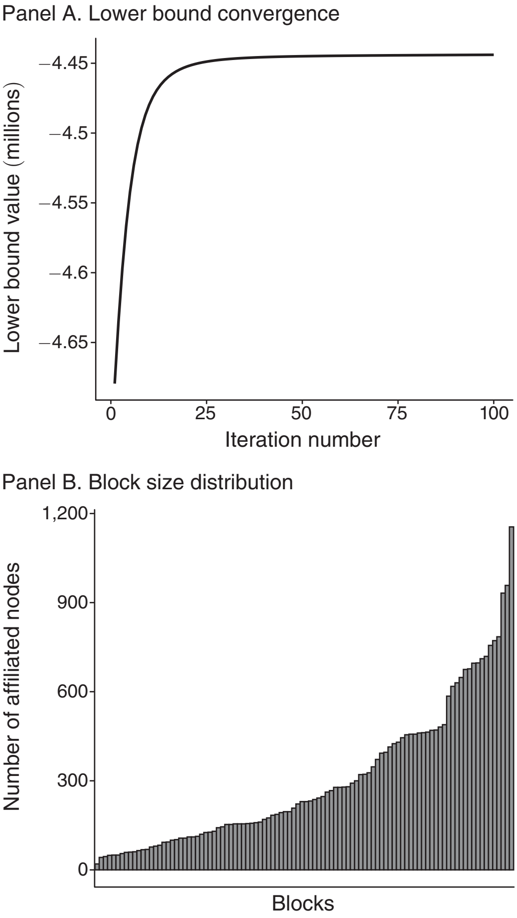


Figure 2. Results of the Block Recovery Step

Notes: The clustering step is first initialized with the resulting partition after applying 20,000 iterations of the EM algorithm without employing covariates. The values shown are obtained by employing for clustering the information on node location (H3 index tile), large industrial category, and occupation type.

The parameters of the utility function are obtained by maximum pseudolikelihood estimation (MPLE), conditioning on the estimated node partition. The corresponding estimates are shown in Table 1. We observe evidence of homophily in professional networking. Business persons in our data exhibit a significant preference for connections with other users who are similar in terms of geospatial location, occupational category, and industrial classification, especially

[2] https://github.com/sansan-inc/lighthergm

Table 1—hergm Parameter Estimates

	Between (1)	Within (2)
Intercept (α)	-7.709 (0.002)	-4.754 (0.005)
Shared contacts (γ)		0.736 (0.004)
Same location (β_1) (H3 tile)	0.333 (0.007)	0.006 (0.012)
Same industry (β_2)	0.694 (0.005)	0.034 (0.009)
Same occupation (β_3)	0.409 (0.006)	0.041 (0.010)
Bayesian inf. crit.	4,171,768	808,597

Note: Coefficient estimates are obtained by the method of MPLE.

for connections across the communities recovered by our model. This is consistent with other studies on industrial and spatial agglomeration (although geographic homophily within types is not significant). We also observe that pairs of users with contacts in common are significantly more likely to form new connections. Finding evidence of externalities even after controlling for homophily is an important result for such a sparse network. It may suggest that the network as a whole can benefit from lower costs to triadic closure. This can potentially be achieved through improvements in the way that users connect with each other through the web app and in the information they observe in their feed.

IV. Conclusion

In this work, we study the decision process behind the formation of a large professional network, guided by an empirical network formation model with unobserved types. We use data of mostly face-to-face encounters from a popular professional networking service in Japan, and make use of a variational EM algorithm to recover unobservable types after controlling for node covariates. Our implementation is scalable and can be employed to analyze networks with hundreds of thousands of nodes at a low computational cost. A scalable block recovery algorithm can be useful for other downstream tasks such as node classification, link prediction, and the improvement of search engines by employing information on the similarity of nodes on the unobserved types recovered by our model. Although our two-steps method significantly enhances researchers' ability to estimate models with large scale networks, additional improvements are possible, especially in the handling of a large number of covariates as well as in model selection. We leave these developments to future research.

REFERENCES

Babkin, Sergeii, Jonathan Stewart, Xiaochen Long, and Michael Schweinberger. 2020. "Large-Scale Estimation of Random Graph Models with Local Dependence." arXiv: 1703.09301.

Bickel, Peter, David Choi, Xiangyu Chang, and Hai Zhang. 2013. "Asymptotic Normality of Maximum Likelihood and Its Variational Approximation for Stochastic Blockmodels." *Annals of Statistics* 41 (4): 1922–43.

Boucher, Vincent, and Ismael Mourifie. 2017. "My Friend Far, Far Away: A Random Field Approach to Exponential Random Graph Models." *Econometrics Journal* 20 (3): S14–46.

Dahbura, Juan Nelson Martínez, Shota Komatsu, Takanori Nishida, and Angelo Mele. 2021. "A Structural Model of Business Cards Exchange Networks." arXiv: 2105.12704.

Mele, Angelo. 2022. "A Structural Model of Homophily and Clustering in Social Networks." *Journal of Business and Economic Statistics* 40 (3): 1377–89.

Rosvall, M., D. Axelsson, and C.T. Bergstrom. 2009. "The Map Equation." *European Physical Journal Special Topics* 178: 13–23.

Schweinberger, Michael, and Pamela Luna. 2018. "hergm: Hierarchical Exponential-Family Random Graph Models." *Journal of Statistical Software* 85 (1): 1–39.

Vu, Duy Q., David R. Hunter, and Michael Schweinberger. 2013. "Model-Based Clustering of Large Networks." *Annals of Applied Statistics* 7 (2): 1010–39.