The Annals of Statistics
2015, Vol. 43, No. 1, 462–466
DOI: 10.1214/14-AOS1271

CORRECTION TO THE PROOF OF CONSISTENCY OF COMMUNITY DETECTION

BY PETER J. BICKEL, AIYOU CHEN, YUNPENG ZHAO, ELIZAVETA LEVINA AND JI ZHU

University of California, Berkeley, Google Inc, George Mason University, University of Michigan and University of Michigan

This note corrects an error in two related proofs of consistency of community detection: under stochastic block models by Bickel and Chen [*Proc. Natl. Acad. Sci. USA* **106** (2009) 21068–21073] and under degree-corrected stochastic block model by Zhao, Levina and Zhu [*Ann. Statist.* **40** (2012) 2266–2292].

This note provides a correction to the proof of consistency of community detection under degree-corrected stochastic block models [2], published in this journal. The same error appeared earlier in the proof of consistency under the stochastic block models [1]. In this note, we provide the correction for the proof of [2], using the notation of that paper, since the case of the degree-corrected stochastic block models is more general and includes the regular stochastic block models as a special case. Very similar arguments can be used to correct the proof of [1] directly.

We start by very briefly restating notation. Let $\mathbf{e}$ be an arbitrary set of label assignments, $\mathbf{c}$ be the true label assignments and $\hat{\mathbf{c}}$ be the maximizer of a community detection criterion. Let $O(\mathbf{e}) \in \mathcal{R}^{K\times K}$, $V(\mathbf{e}) \in \mathcal{R}^{K\times K\times M}$, $\hat{\Pi} \in \mathcal{R}^{K\times M}$, $f(\mathbf{e}) \in \mathcal{R}^{K}$, where

$$O_{kl}(\mathbf{e}) = \sum_{ij} A_{ij} I\{e_i = k, e_j = l\},$$

$$V_{kau}(\mathbf{e}) = \frac{\sum_{i=1}^n I(e_i = k, c_i = a, \theta_i = x_u)}{\sum_{i=1}^n I(c_i = a, \theta_i = x_u)},$$

$$\hat{\Pi}_{au} = \frac{1}{n}\sum_{i=1}^n I(c_i = a, \theta_i = x_u),$$

$$f_k(\mathbf{e}) = \frac{1}{n}\sum_{i=1}^n I(e_i = k) = \sum_{au} V_{kau}(\mathbf{e})\hat{\Pi}_{au}.$$

Received August 2014; revised September 2014.
MSC2010 subject classifications. 62G20.
Key words and phrases. Network communities, stochastic block model, degree-corrected stochastic block model, consistency of community detection.

We considered community detection criteria that can be written in the form

$$Q(\mathbf{e}) = F\left(\frac{O(\mathbf{e})}{\mu_n}, f(\mathbf{e})\right),$$

where $\mu_n = n^2\rho_n$ and $\rho_n \to 0$ is the average probability of an edge in the network. For any matrix B, $\|B\|_\infty = \max_{kl}|B_{kl}|$.

The statement $|\Delta(\mathbf{e},\mathbf{c})| \le M_1(\|X(\mathbf{e}) - X(\mathbf{c})\|_\infty)$ below (A.11) in [2] is incorrect. (We have replaced M' and C' in the original with M_1 and C_1 in this correction since we will need more constants.) For the proof to go through, we need a different way of proving

$$\mathbb{P}\Big(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n}|\Delta(\mathbf{e},\mathbf{c})| - C_1\|V(\mathbf{e}) - \mathbb{I}\|_1/4 \le 0\Big) \to 1, \tag{1.1}$$

where $\delta_n \to 0$. Note that (1.1) is similar to the (A.14) in [2], with an extra constraint $|\mathbf{e}-\mathbf{c}| \le \delta_n n$. Since we have already proved $\mathbb{P}(\frac{1}{n}|\hat{\mathbf{c}} - \mathbf{c}| \le \delta_n) \to 1$ in [2], (1.1) will complete the proof, and the conclusion of Theorem 4.1 in [2] remains valid.

We first need a lemma based on Bernstein's inequality.

LEMMA 1.1. *For $m \in \{1,\dots,n\}$*,

$$\mathbb{P}\Big(\max_{|\mathbf{e}-\mathbf{c}|\le m}\|X(\mathbf{e})\|_\infty \ge \varepsilon\Big) \le 2\binom{n}{m}K^{m+2}\exp\left(-\frac{3\mu_n\varepsilon^2}{4(\varepsilon+3)}\right). \tag{1.2}$$

The proof of Lemma 1.1 closely follows the proof of (A.2) and (A.3) in [2] and hence is omitted here.

Proof of (1.1):

By Taylor's expansion,

$$\begin{aligned} F&\left(\frac{O(\mathbf{e})}{\mu_n}, f(\mathbf{e})\right) - F\big(\hat{T}(\mathbf{e}), f(\mathbf{e})\big) \\ &= \frac{\partial F}{\partial M}\bigg|_{M=\hat{T}(\mathbf{e}),\mathbf{t}=f(\mathbf{e})}\operatorname{vec}\big(X(\mathbf{e})\big) + O\big(\|X(\mathbf{e})\|_\infty^2\big),\end{aligned}$$

where $\frac{\partial F}{\partial M}$ is the partial derivative over the first component (vectorized) of $F(M,\mathbf{t})$. Similarly,

$$\begin{aligned} F&\left(\frac{O(\mathbf{c})}{\mu_n}, f(\mathbf{c})\right) - F\big(\hat{T}(\mathbf{c}), f(\mathbf{c})\big) \\ &= \frac{\partial F}{\partial M}\bigg|_{M=\hat{T}(\mathbf{c}),\mathbf{t}=f(\mathbf{c})}\operatorname{vec}\big(X(\mathbf{c})\big) + O\big(\|X(\mathbf{c})\|_\infty^2\big).\end{aligned}$$

Since $\frac{\partial F}{\partial M}$ is continuous with respect to M and t, and $\hat{T}(\mathbf{e})$ and $f(\mathbf{e})$ are continuous with respect to $\mathbf{e}$,

$$\frac{\partial F}{\partial M}\bigg|_{M=\hat{T}(\mathbf{e}),\mathbf{t}=f(\mathbf{e})} = \frac{\partial F}{\partial M}\bigg|_{M=\hat{T}(\mathbf{c}),\mathbf{t}=f(\mathbf{c})} + O\big(\|V(\mathbf{e}) - \mathbb{I}\|_1\big). \tag{1.3}$$

Therefore, since

$$\Delta(\mathbf{e},\mathbf{c}) = F\left(\frac{O(\mathbf{e})}{\mu_n}, f(\mathbf{e})\right) - F(\hat{T}(\mathbf{e}), f(\mathbf{e})) - F\left(\frac{O(\mathbf{c})}{\mu_n}, f(\mathbf{c})\right) + F(\hat{T}(\mathbf{c}), f(\mathbf{c}))$$
$$= \frac{\partial F}{\partial M}\bigg|_{M=\hat{T}(\mathbf{c}),\mathbf{t}=f(\mathbf{c})} \operatorname{vec}(X(\mathbf{e}) - X(\mathbf{c})) + O(\|V(\mathbf{e}) - \mathbb{I}\|_1) \operatorname{vec}(X(\mathbf{e}))$$
$$+ O(\|X(\mathbf{e})\|_\infty^2) + O(\|X(\mathbf{c})\|_\infty^2),$$

we have

$$|\Delta(\mathbf{e},\mathbf{c})| \le M_1\|X(\mathbf{e}) - X(\mathbf{c})\|_\infty + M_2\|V(\mathbf{e}) - \mathbb{I}\|_1\|X(\mathbf{e})\|_\infty + M_3\|X(\mathbf{e})\|_\infty^2$$
$$+ M_4\|X(\mathbf{c})\|_\infty^2.$$

Now we prove (1.1), which holds if the following four statements hold:

$$\mathbb{P}\left(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n} M_1\|X(\mathbf{e}) - X(\mathbf{c})\|_\infty - C_1\|V(\mathbf{e}) - \mathbb{I}\|_1/16 \le 0\right) \to 1, \tag{1.4}$$

$$\mathbb{P}\left(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n} M_2\|X(\mathbf{e})\|_\infty - C_1/16 \le 0\right) \to 1, \tag{1.5}$$

$$\mathbb{P}\left(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n} M_3\|X(\mathbf{e})\|_\infty^2 - C_1\|V(\mathbf{e}) - \mathbb{I}\|_1/16 \le 0\right) \to 1, \tag{1.6}$$

$$\mathbb{P}\left(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n} M_4\|X(\mathbf{c})\|_\infty^2 - C_1\|V(\mathbf{e}) - \mathbb{I}\|_1/16 \le 0\right) \to 1. \tag{1.7}$$

The proof of (1.4) is similar to the proof of (A.15) in [2]. Note that $\frac{1}{n}|\mathbf{e}-\mathbf{c}| \le \frac{1}{2}\|V(\mathbf{e}) - \mathbb{I}\|_1$. So for each $m \ge 1$,

$$\mathbb{P}\left(\max_{|\mathbf{e}-\mathbf{c}|=m} M_1\|X(\mathbf{e}) - X(\mathbf{c})\|_\infty - C_1\|V(\mathbf{e}) - \mathbb{I}\|_1/16 > 0\right)$$
$$\le \mathbb{P}\left(\max_{|\mathbf{e}-\mathbf{c}|\le m}\|X(\mathbf{e}) - X(\mathbf{c})\|_\infty > \frac{C_1 m}{8M_1 n}\right) = I_1.$$

Let $\alpha = C_1/8M_1$ if $\alpha \ge 6C$, by (A.2) in [2],

$$I_1 \le 2K^{m+2}n^m \exp\left(-\alpha\frac{3m}{8n}\mu_n\right)$$
$$= 2K^2[K\exp(\log n - \alpha\mu_n/(8/3n))]^m.$$

If $\alpha < 6C$, by (A.3) in [2],

$$I_1 \le 2K^{m+2}n^m \exp\left(-\alpha^2\frac{m}{16Cn}\mu_n\right)$$
$$= 2K^2[K\exp(\log n - \alpha^2\mu_n/(16Cn))]^m.$$

In both cases, since $\lambda_n/\log n \to \infty$ ($\lambda_n = n\rho_n$),

$$\mathbb{P}\Big(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n} M_1\|X(\mathbf{e})-X(\mathbf{c})\|_\infty - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16>0\Big)$$

$$\le \sum_{m=1}^{\infty}\mathbb{P}\Big(\max_{|\mathbf{e}-\mathbf{c}|=m} M_1\|X(\mathbf{e})-X(\mathbf{c})\|_\infty - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16>0\Big)\to 0,$$

as $n\to\infty$, which completes the proof of (1.4).

Equation (1.5) simply follows (A.1) in [2].

We next prove (1.6). For each $1\le m\le \delta_n n$,

$$\mathbb{P}\Big(\max_{|\mathbf{e}-\mathbf{c}|=m} M_3\|X(\mathbf{e})\|_\infty^2 - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16>0\Big)$$

$$\le \mathbb{P}\Big(\max_{|\mathbf{e}-\mathbf{c}|\le m}\|X(\mathbf{e})\|_\infty^2 > \frac{C_1 m}{8M_3 n}\Big) = I_2.$$

Let $\varepsilon=\sqrt{\frac{C_1 m}{8M_3 n}}$, $\alpha = C_1/64M_3$. Then from Lemma 1.1,

$$\begin{aligned} I_2 &\le 2K^{m+2}n^m\exp\Big(-\frac{3\mu_n\varepsilon^2}{4(\varepsilon+3)}\Big)\\ &\le 2K^{m+2}n^m\exp\Big(-\frac{\mu_n\varepsilon^2}{8}\Big)\\ &= 2K^{m+2}n^m\exp\Big(-\alpha\frac{\mu_n}{n}m\Big)\\ &= 2K^2\Big[K\exp\Big(\log n-\alpha\frac{\mu_n}{n}\Big)\Big]^m. \end{aligned}$$

Since $\lambda_n/\log n\to\infty$,

$$\mathbb{P}\Big(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n} M_3\|X(\mathbf{e})\|_\infty^2 - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16>0\Big)$$

$$\le \sum_{m=1}^{\infty}\mathbb{P}\Big(\max_{|\mathbf{e}-\mathbf{c}|=m} M_3\|X(\mathbf{e})\|_\infty^2 - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16>0\Big)\to 0,$$

as $n\to\infty$, which completes the proof of (1.6).

We now complete the proof by showing (1.7). For each $1\le m\le\delta_n n$,

$$\mathbb{P}\Big(\max_{|\mathbf{e}-\mathbf{c}|=m} M_4\|X(\mathbf{c})\|_\infty^2 - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16>0\Big)$$

$$= \mathbb{P}\Big(\|X(\mathbf{c})\|_\infty^2 > \frac{C_1 m}{8M_4 n}\Big) = I_3.$$

Let $\varepsilon=\sqrt{\frac{C_1 m}{8M_4 n}}$, $\alpha = C_1/64M_4$. Then from Bernstein's inequality,

$$I_3 \le 2K^2\exp\Big(-\frac{3\mu_n\varepsilon^2}{4(\varepsilon+3)}\Big)\le 2K^2\exp\Big(-\alpha\frac{\mu_n}{n}m\Big). \tag{1.8}$$

Therefore,

$$\mathbb{P}\Big(\max_{1\le|\mathbf{e}-\mathbf{c}|\le\delta_n n} M_4\|X(\mathbf{c})\|_\infty^2 - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16 > 0\Big)$$
$$\le \sum_{m=1}^{\infty} \mathbb{P}\big(M_4\|X(\mathbf{e})\|_\infty^2 - C_1\|V(\mathbf{e})-\mathbb{I}\|_1/16 > 0\big) \to 0 \qquad \text{as } n\to\infty.$$

Acknowledgements. We are very grateful to Emma Jingfei Zhang, a former Ph.D. student at University of Illinois at Urbana-Champaign now at University of Miami, who discovered the error and persisted in tracking down its root cause.

REFERENCES

[1] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[2] ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083

P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: bickel@stat.berkeley.edu

A. CHEN
GOOGLE INC
1600 AMPHITHEATRE PKWY
MOUNTAIN VIEW, CALIFORNIA 94043
USA
E-MAIL: aiyouchen@google.com

Y. ZHAO
DEPARTMENT OF STATISTICS
GEORGE MASON UNIVERSITY
1714 ENGINEERING BUILDING
4400 UNIVERSITY DRIVE
FAIRFAX, VIRGINIA 22030-4444
USA
E-MAIL: yzhao15@gmu.edu

E. LEVINA
J. ZHU
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
311 WEST HALL
1085 S. UNIVERSITY AVE.
ANN ARBOR, MICHIGAN 48109-1107
USA
E-MAIL: elevina@umich.edu
jizhu@umich.edu