The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

補助回答を利用した意見集約法による化合物の合成可能性判定

Synthetic Accessibility Assessment Using Auxiliary Responses

伊藤 駿 *1 Shun Ito　馬場 雪乃 *2 Yukino Baba　磯村 哲 *3 Tetsu Isomura　鹿島 久嗣 *1*4 Hisashi Kashima

*1京都大学 Kyoto University　*2筑波大学 University of Tsukuba　*3三菱ケミカルホールディングス Mitsubishi Chemical Holdings Corporation

*4理化学研究所 革新知能統合研究センター RIKEN Center for Advanced Intelligence Project

Despite of the recent advances in computational approaches to discovering new compounds, synthetic accessibility assessment of chemical compounds is still an AI-hard task. Therefore human computation approaches are potentially promising. We introduce a new aggregation method of semi-expert judgments that consist of five-grade synthetic accessibility scores along with annotations of obstructive substructures to their synthesis. Our proposed two-stage method treats the obstructive substructures as auxiliary responses. The first stage divides both the semi-experts and the atoms in the substructures into clusters using stochastic block models. The second stage generates five-grade responses based on the clusters. Our experiments show that the use of the auxiliary responses improves the prediction performance.

1. はじめに

創薬における分子設計プロセスは計算機による自動化が進んでおり，新しい候補化合物を短時間で大量に生成することが可能となっている．計算機により生成された化合物は合成可能性が保証されていないため，それを検証する工程が必要であるが，過去に提案された自動化手法 [1, 2] は精度や計算時間の面で実用レベルに達しておらず，専門家による判定作業に依存しており，プロセス全体の高速化のための課題となっている．少数の作業者に依存することによる非効率化に対処するためのアプローチとして，クラウドソーシングの仕組みを活用し多数のワーカで作業を分担する方法が存在する．合成可能性判定においても，専門家が行なっている判定作業を専門家ではない多数のワーカが分担して行うことで，より効率的な判定が可能となる．しかし，合成可能性判定は高度な専門知識を必要とする作業であるため，化学に関する知識を全く持たない人による回答から正しい判定が得られることは期待できない．この問題に対処するため，回答者として化学にある程度知識を持つ準専門家を導入することが提案されており [3]，本研究でも準専門家の回答を集約することによって合成可能性の予測を行う．

本研究では，着眼点の違いから準専門家の能力や化合物の性質などの情報を得ることを目的とし，判定回答とは別の補助回答として，適切でないと考えた部分構造を選択するように依頼する．つまり，準専門家は図 1 に示されるような 2 種類の回答を行う．(1) 5 段階での判定では，判定対象の化合物が合成可能であると判断するほど高い値を回答する．(2) 悪い部分構造の選択では，判定対象の化合物のうち合成するのに不適切な部分構造を構成している原子を 0 個以上選択する．

部分構造の選択で、似た能力の準専門家による原子の選び方・似た性質の原子の選ばれ方は，それぞれ似たような傾向になると仮定し，補助回答に対して確率的ブロックモデルを適用することで，準専門家と原子をいくつかのクラスタに分類する．これを第 1 段階とし，第 2 段階で，各準専門家が独立して持つ潜在パラメータと，準専門家・原子のクラスタを考慮した潜在パラメータの合計 3 種類のパラメータを用いて判定回答の生成をモデル化する，という 2 段階での予測手法を提案する．潜在クラスタを考慮したモデルには，ワーカの能力をクラスタに依存して決定したもの [4] や，ワーカとタスクの両方のクラスタを考慮したもの [5] 等が存在するが，これらの既存手法とは異なり，本研究はクラスタを補助回答から推定し，判定回答の生成モデルを構築する．

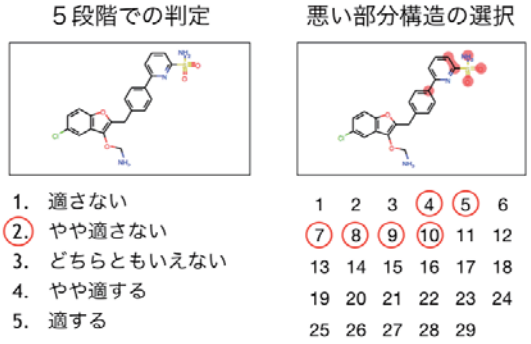


図 1: 準専門家の回答の例.

2. 提案手法

補助回答を観測データとする第 1 段階と，判定回答を観測データとする第 2 段階とで構成される，2 段階での合成可能性の予測手法を提案する．第 1 段階では確率的ブロックモデル (Stochastic Block Model; SBM) [6] または混合メンバシップ確率的ブロックモデル (Mixed Membership Stochastic Blockmodels; MMSB) [7] によるクラスタリングを行い，第 2 段階では 3 種類の潜在パラメータによる判定回答の生成モデルを構築する．観測データを元にこれらのモデルの推定を行い，化合物の合成可能性の予測値を得る．以下では，準専

連絡先: 伊藤 駿，京都大学大学院情報学研究科知能情報学専攻，s.itoh248@ml.ist.i.kyoto-u.ac.jp

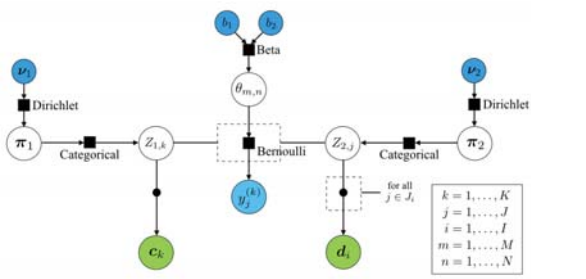


図 2: SBM によるクラスタリング.

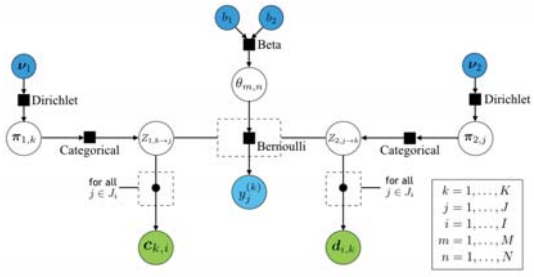


図 3: MMSB によるクラスタリング.

門家を $k = 1, \ldots, K$，化合物を $i = 1, \ldots, I$, 化合物 i の真の合成可能性を $t_i \in \{0, 1\}$，化合物 i の構成原子の集合を $\mathcal{J}_i$，全化合物に対する原子を $j = 1, \ldots, J$ $(J = \Sigma_i |\mathcal{J}_i|)$，とする．ただし，$t_i = 1$ が化合物 i が合成可能であることに対応する．観測データは，準専門家 k の化合物 i に対する判定回答 $x_i^{(k)} \in \{1, \ldots, 5\}$ と原子 j に対する回答 $y_j^{(k)} \in \{0, 1\}$ で，$y_j^{(k)} = 1$ の時に準専門家 k が原子 j を選択したことを表す．

2.1 クラスタリング

第 1 段階では，補助回答データに SBM または MMSB を適用することで，準専門家と原子をいくつかのクラスタに分類する．SBM は，関係ネットワークにおいて，あるノードとノードとの間にエッジが存在する確率は，それぞれのノードが所属するクラスタに依存するという仮定をおくモデルである．SBM のもとで補助回答 $y_j^{(k)}$ は次のベルヌーイ分布に従って生成される．

$$y_j^{(k)} | \theta_{m,n}, Z_{1,k}, Z_{2,j} \sim \text{Bernoulli}(y_j^{(k)} | \theta_{Z_{1,k}, Z_{2,j}})$$

ここで $Z_{1,k} = 1, \ldots, M$, $Z_{2,j} = 1, \ldots, N$ はそれぞれ準専門家・原子の所属クラスタで，$\theta_{m,n}$ はクラスタ m に属する準専門家がクラスタ n に属する原子を選択する確率を表す．これらのパラメータは次のような分布によって生成される．

$$Z_{1,k} | \boldsymbol{\pi}_1 \sim \text{Categorical}(Z_{1,k} | \boldsymbol{\pi}_1)$$
$$Z_{2,j} | \boldsymbol{\pi}_2 \sim \text{Categorical}(Z_{2,j} | \boldsymbol{\pi}_2)$$
$$\theta_{m,n} \sim \text{Beta}(\theta_{m,n} | b_1, b_2)$$

$\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ はディリクレ分布により生成されるパラメータで，それぞれ各次元が各クラスタ m, n に所属する確率を表す．以上の過程により準専門家と原子のクラスタ $\{Z_{1,k}\}_k, \{Z_{2,j}\}_j$ が得られるが，第 2 段階のためにこれらを準専門家と化合物のクラスタ $\{\boldsymbol{c}_k\}_k, \{\boldsymbol{d}_i\}_i$ に変換する．化合物 i に対するクラスタは，原子 $j \in \mathcal{J}_i$ が所属しているクラスタを複合したものとして決定されると仮定し，次のように変換を行う．

$$\boldsymbol{c}_k = (c_{k,1}, \ldots, c_{k,M})^\top$$
$$c_{k,m} = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} \mathbb{I}(Z_{1,k} = m)$$
$$\boldsymbol{d}_i = (d_{i,1}, \ldots, d_{i,N})^\top$$
$$d_{i,n} = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} \mathbb{I}(Z_{2,j} = n)$$

ここで $\mathbb{I}(\cdot)$ は指示関数を表す．以上の SBM により準専門家と原子，化合物のクラスタが得られる過程を図 2 に示す．

一方，MMSB は SBM の拡張として提案されたモデルで，SBM がインデックス毎にクラスタリングを行うのに対し，MMSB はインデックスのペア毎によってクラスタリングを行うため，一般に一つのインデックスは複数のブロックに割り当てられる．MMSB のもとで補助回答 $y_j^{(k)}$ は次のベルヌーイ分布に従って生成される．

$$y_j^{(k)} | \theta_{m,n}, Z_{1,k \to j}, Z_{2,j \to k} \sim \text{Bernoulli}(y_j^{(k)} | \theta_{Z_{1,k \to j}, Z_{2,j \to k}})$$

ここで $Z_{1,k \to j} = 1, \ldots, M$, $Z_{2,j \to k} = 1, \ldots, N$ はそれぞれ準専門家・原子の所属クラスタであるが，SBM の場合とは異なりどちらも k と j のペア毎にクラスタが決定しており，次のカテゴリカル分布から生成される．

$$Z_{1,k \to j} | \boldsymbol{\pi}_{1,k} \sim \text{Categorical}(Z_{1,k \to j} | \boldsymbol{\pi}_{1,k})$$
$$Z_{2,j \to k} | \boldsymbol{\pi}_{2,j} \sim \text{Categorical}(Z_{2,j \to k} | \boldsymbol{\pi}_{2,j})$$

$\boldsymbol{\pi}_{1,k}, \boldsymbol{\pi}_{2,j}$ は SBM と同様にディリクレ分布により生成されるパラメータであるが，MMSB では各準専門家 k，原子 j 毎に異なる値をとる．また，$\theta_{m,n}$ は SBM と同様にベータ分布から生成される．以上の過程により準専門家と原子のクラスタ $\{Z_{1,k \to j}\}_{k,j}, \{Z_{2,j \to k}\}_{j,k}$ が得られるが，これらも第 2 段階のために準専門家と化合物のクラスタ $\{\boldsymbol{c}_{k,i}\}_{k,i}, \{\boldsymbol{d}_{i,k}\}_{i,k}$ に変換する．化合物 i に対するクラスタは，原子 $j \in \mathcal{J}_i$ が所属しているクラスタの複合として決定される．

$$\boldsymbol{c}_{k,i} = (c_{k,i,1}, \ldots, c_{k,i,M})^\top$$
$$c_{k,i,m} = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} \mathbb{I}(Z_{1,k \to j} = m)$$
$$\boldsymbol{d}_{i,k} = (d_{i,k,1}, \ldots, d_{i,k,N})^\top$$
$$d_{i,k,n} = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} \mathbb{I}(Z_{2,j \to k} = n)$$

以上の MMSB により準専門家と原子，化合物のクラスタが得られる過程を図 3 に示す．

2.2 判定回答生成モデル

第 2 段階では，第 1 段階で得られた準専門家と化合物に対するクラスタリング結果を考慮し，5 段階の判定回答の生成を確率モデルで表現する．本手法では，判定回答を生成する潜在パラメータとして (1) 各準専門家 k が独立に持つパラメータ $\boldsymbol{\alpha}_t^{(k)}$，(2) 準専門家のクラスタ m 毎に共通するパラメータ $\boldsymbol{\beta}_{t,m}$，(3) 化合物のクラスタ n 毎に共通するパラメータ $\boldsymbol{\gamma}_n$ の 3 種類を仮定する．ここで $t \in \{0, 1\}$ は真の合成可能性を表し，$t = 1$ が合成可能な場合に対応する．これら 3 種類のパラメー

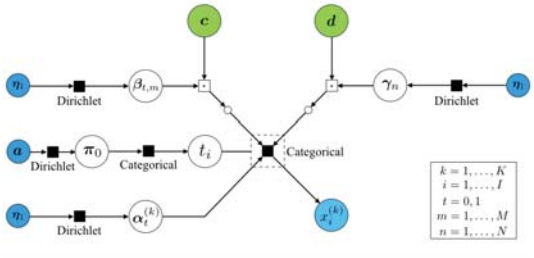


図 4: 判定回答生成モデル.

タは全て 5 次元ディリクレ分布から生成され，各次元が 5 段階での各段階の生成確率を表す．つまり，$\boldsymbol{\alpha}_t^{(k)}$ は真の合成可能性が t の時の準専門家 k の回答傾向，$\boldsymbol{\beta}_{t,m}$ は真の合成可能性が t の時のクラスタ m に属する準専門家の回答傾向，$\boldsymbol{\gamma}_n$ はクラスタ n に属する化合物の被回答傾向となる．$\boldsymbol{\gamma}_n$ は化合物自体の性質を表すため，真の合成可能性には依存しないと仮定する．これらのパラメータの生成は次のようになる．

$$\boldsymbol{\alpha}_t^{(k)} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_t^{(k)}|\boldsymbol{\eta}_1)$$
$$\boldsymbol{\beta}_{t,m} \sim \mathrm{Dirichlet}(\boldsymbol{\beta}_{t,m}|\boldsymbol{\eta}_2)$$
$$\boldsymbol{\gamma}_n \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}_n|\boldsymbol{\eta}_3)$$

一方，化合物 i の真の合成可能性 t_i は 2 次元パラメータ $\boldsymbol{\pi}_0$ を事前分布として，次のカテゴリカル分布から生成される．

$$t_i|\boldsymbol{\pi}_0 \sim \mathrm{Categorical}(t_i|\boldsymbol{\pi}_0)$$

以上のパラメータを考慮することで，判定回答 $x_i^{(k)}$ が生成される．

$$x_i^{(k)}|t_i, \boldsymbol{\alpha}_{t,k}, \boldsymbol{B}_t, \boldsymbol{\Gamma}, \boldsymbol{c}_k, \boldsymbol{d}_i \sim \mathrm{Categorical}\left(x_i^{(k)} \middle| \frac{\boldsymbol{\alpha}_{t_i}^{(k)} + \boldsymbol{c}_k^\top \boldsymbol{B}_{t_i} + \boldsymbol{d}_i^\top \boldsymbol{\Gamma}}{\omega}\right)$$

ここで ω は正規化定数，$\boldsymbol{B}_t \in [0,1]^{M\times 5}, \boldsymbol{\Gamma} \in [0,1]^{N\times 5}$ はそれぞれ $\{\boldsymbol{\beta}_{t,m}\}_m, \{\boldsymbol{\gamma}_n\}_n$ を並べた行列であるとする．

$$\boldsymbol{B}_t = \begin{pmatrix} \boldsymbol{\beta}_{t,1} \\ \vdots \\ \boldsymbol{\beta}_{t,M} \end{pmatrix}, \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_N \end{pmatrix}$$

判定回答を観測データとして以上のモデルを推定することにより，合成可能性の予測値 $\{\hat{t}_i\}_i$ を得る．第 2 段階の過程を図 4 に示す．

3. 実験

提案手法を実際に準専門家から得られた回答データに対して適用し，性能を検証する．今回用いるデータセットは，分子設計研究に携わる 9 名を準専門家とし，全員が 5 段階の判定回答と，部分構造の選択を行なったものである．対象となる新規化合物は，公開データベース上の化合物に対してランダムな部分構造置換を行うことによって生成されたものである．新規化合物の真の合成可能性は，創薬の専門家 5 名が 95 個の新規化合物に対して 2 度の回答を行い，化合物 i に対して 5 名全員が 3 以下の判定をした場合には $t_i = 0$，全員が 3 以上の判定をした場合には $t_i = 1$ となるように決定した．今回はこの方法で正解が得られた 59 個の新規化合物を対象に実験を行う．

表 1: 提案手法と比較手法の実データに対する予測性能.

手法	AUC
MV	0.886
Raykar	0.939
Only-α	0.925
SBM-$(\alpha+\beta)$	0.924
SBM-$(\alpha+\gamma)$	0.944
SBM-$(\beta+\gamma)$	0.887
SBM-$(\alpha+\beta+\gamma)$	**0.955**
MMSB-$(\alpha+\beta)$	0.927
MMSB-$(\alpha+\gamma)$	0.947
MMSB-$(\beta+\gamma)$	0.875
MMSB-$(\alpha+\beta+\gamma)$	**0.959**

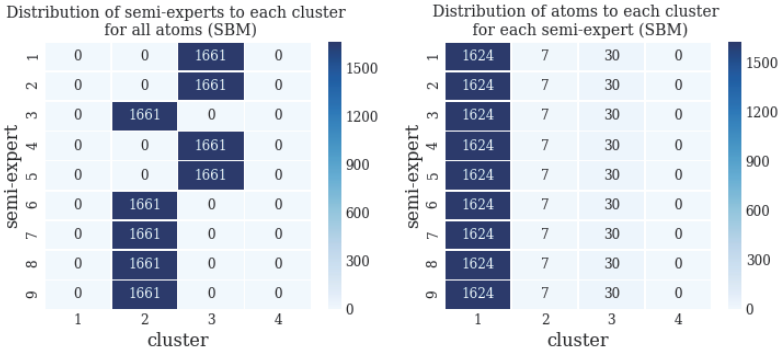
比較手法として，補助回答を用いない手法を 3 つ，補助回答を用いる手法を 6 つ用意した．補助回答を用いない（5 段階の判定回答のみを用いる）手法は，9 人の判定回答の平均値を予測値として用いる手法 (MV)，回答が段階ラベルで得られる場合に感度と特異度によってワーカの能力をモデル化した手法 [8] (Raykar)，本手法での $x_i^{(k)}$ の生成を準専門家依存のパラメータ $\boldsymbol{\alpha}_t^{(k)}$ のみを用いて行う手法 (Only-α) の 3 つを用意する．補助回答を用いる手法は Only-α と同様に提案手法の第 2 段階で用いるパラメータを変化させた手法であり，$\boldsymbol{\alpha}_t^{(k)}$ と準専門家クラスタ依存のパラメータ $\boldsymbol{\beta}_{t,m}$ を用いる手法 (SBM-$(\alpha+\beta)$, MMSB-$(\alpha+\beta)$)，$\boldsymbol{\alpha}_t^{(k)}$ と化合物クラスタ依存のパラメータ $\boldsymbol{\gamma}_n$ を用いる手法 (SBM-$(\alpha+\gamma)$, MMSB-$(\alpha+\gamma)$)，準専門家・原子クラスタ依存のパラメータ $\boldsymbol{\beta}_{t,m}, \boldsymbol{\gamma}_n$ を用いる手法 (SBM-$(\beta+\gamma)$, MMSB-$(\beta+\gamma)$) の計 6 つを用意する．

推論は第 1 段階のクラスタリング部分と第 2 段階の判定回答生成モデル部分のそれぞれで，最大事後確率推定を独立して行う．準専門家と原子のクラスタ数は $M = 4, N = 4$ に設定した．

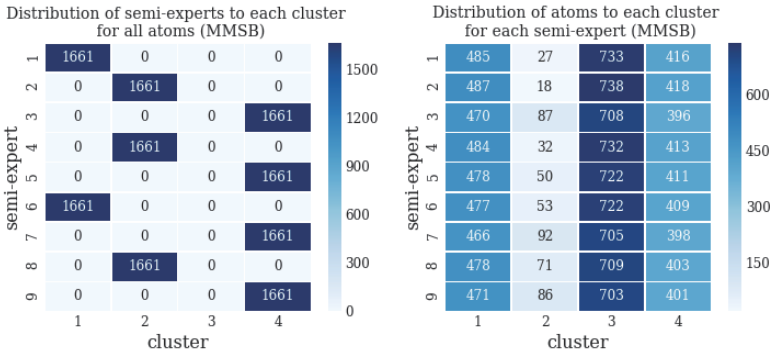
4. 実験結果

提案手法と比較手法を実際のデータセットに適用した結果を AUC により評価したものを表 1 に示す．補助回答に対するクラスタリングに SBM と MMSB のどちらを用いた場合にも，提案手法の SBM-$(\alpha+\beta+\gamma)$, MMSB-$(\alpha+\beta+\gamma)$ が最も高い AUC の値を得られた．また，判定回答の生成パラメータとして α, β を使っている手法と α, γ を使っている手法を比較すると，γ を考慮した手法の方が高い AUC が得られている．これは，α と β が準専門家に関する潜在パラメータであるのに対して γ が化合物に関する潜在パラメータであるため，準専門家と化合物の両方の性質を同時に考慮することでより正確なモデルとなったことが理由として考えられる．さらに，3 つのパラメータ全てを考慮した提案手法が最も高い AUC の値を得られていることから，β が α からは得られない準専門家同士に共通する能力を抽出していると考えられる．

続いて SBM と MMSB によって得られたクラスタリングの結果について図 5 を用いて検証する．まず準専門家のクラスタリング結果を見ると，SBM による結果の方が MMSB によ



(a) SBM



(b) MMSB

図 5: SBM と MMSB によるクラスタリング結果．左側が準専門家 9 人の全原子 1661 個に対するクラスタ，右側が原子 1661 個の準専門家 9 人に対するクラスタを表す．準専門家は個人の判定回答の AUC が低い順に上から並べられている．

る結果よりも明確であるが，いずれの場合でも AUC の比較的高い準専門家のクラスタと低い準専門家のクラスタが生成されており，能力による分類がうまく行われていることがわかる．続いて原子のクラスタリング結果を比較する．SBM による結果ではほとんど全ての原子がクラスタ 1 に割り当てられているのに対し，MMSB による結果ではより詳細な分類結果が得られていることがわかる．また，MMSB でのクラスタ 2 に注目すると，ペアとなる準専門家の AUC が高くなるにつれてクラスタ 2 に属する原子の個数が増加する傾向にあるため，悪い部分構造を構成している可能性が高い原子が MMSB によるクラスタリングによりうまく分類されていると考えられる．

5. むすび

補助回答から得たクラスタリングの結果を考慮し，準専門家の判定回答を統合することによって化合物の合成可能性判定を行うモデルを提案し，提案手法の有効性を実験により示した．今回は，回答者が適切でない部分構造を発見することにより化合物が合成不可能であると判定するような一種の減点方式であったが，適切である，もしくは特に良いと思った部分構造があれば判定対象に良い判定を与えるというような加点方式の設定も考慮していくことが，今後の展望として挙げられる．

参考文献

[1] P. Ertl and A. Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, Vol. 1, pp. 1–11, 2009.

[2] K. Boda, T. Seidel, and J. Gasteiger. Structure and reaction based evaluation of synthetic accessibility. *Journal of Computer-Aided Molecular Design*, Vol. 21, pp. 311–325, 2007.

[3] Y. Baba, T. Isomura, and H. Kashima. Wisdom of crowds for synthetic accessibility evaluation. *Journal of Molecular Graphics and Modelling*, 2018. in press.

[4] M. Venanzi, J. Guiver, Kohli Kazai, G., P., and M. Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the the 23rd World Wide Web Conference*, pp. 155–164, 2014.

[5] H. Lakkaraju, J. Leskovec, J. Kleinberg, and S. Mullainathan. A bayesian framework for modeling human evaluations. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 181–189, 2015.

[6] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, Vol. 96, pp. 1077–1087, 2001.

[7] E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, Vol. 9, pp. 1981–2014, 2008.

[8] V. C. Raykar and S. Yu. Ranking annotators for crowdsourced labeling tasks. In *Advances in Neural Information Processing Systems 24*, pp. 1809–1817. 2011.