MATOMC

SDU

Exact and Agnostic Recovery in the Data Block Model

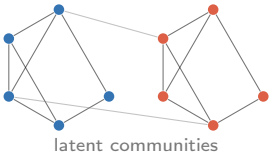
Akbar Davoodi

Department of Mathematics and computer science (IMADA), University of Southern Denmark

41^{st} TBI Winterseminar in Bled

How to generate random graphs with clustered structure?

Clustering goal: infer latent communities (labels) $X^n = (X_1, \ldots, X_n)$ from the observed graph G_n.



Where SBM fits among random graph models:

Erdős–Rényi	edges are i.i.d.; *homogeneous* connectivity, no built-in clusters
Configuration model	preserves degree sequence; captures hubs/heterogeneity, but no explicit community mechanism
SBM	introduces latent labels $X_i \in [k]$ and makes edge probabilities depend on (X_i, X_j) via W

Why it's a good candidate model:

- Explicit clustered structure: *within/between*-community connectivity controlled by W.
- Flexible: assortative or disassortative patterns depending on W_{ab}.
- Analytically tractable.

Stochastic Block Model (SBM)

Model parameters: number of vertices n, number of communities k, prior $P = (p_1, \dots, p_k)$ on $[k]$, and symmetric edge-probability matrix

$$W = [W_{ab}]_{1 \le a,b \le k} \in [0,1]^{k \times k}.$$

Latent labels: draw i.i.d.

$$\mathbb{P}(X_i = a) = p_a, \qquad a \in [k],\ i \in [n],$$

and write $X^n = (X_1, \dots, X_n)$.

Graph: on vertex set $V_n = \{v_1, \dots, v_n\}$, let

$$G_n = (Y_{ij})_{1 \le i < j \le n}, \qquad Y_{ij} \in \{0,1\},$$

and conditionally on X^n, the edges are independent with

$$\mathbb{P}(Y_{ij} = 1 \mid X_i = a, X_j = b) = W_{ab}, \qquad 1 \le i < j \le n.$$

Notation: $(G_n, X^n) \sim \mathrm{SBM}(n, k, P, W)$.

Logarithmic-degree regime: $W^{(n)} = \dfrac{\log n}{n}\, Q$ for a fixed symmetric $Q \in \mathbb{R}_+^{k \times k}$.

Why should we go beyond SBM?

SBM viewpoint: in $(G_n, X^n) \sim \mathrm{SBM}(n, k, P, W)$ we infer latent labels $X^n = (X_1, \ldots, X_n)$ using *only* the connectivity G_n.

But many real networks have vertex-level data:

- Social networks: profile/location/interests alongside friendships.
- Citation graphs: paper text/keywords alongside citations.
- Biological networks: gene/protein attributes alongside interactions.

Key idea: community membership influences *both*

1. the graph structure (edge probabilities via W), and
2. the distribution of node-associated data files $U^{(i)}$.

DBM = SBM + side information. We attach data $U^n := (U^{(1)}, \ldots, U^{(n)})$, and assume the conditional-independence structure $G_n \leftrightarrow X^n \leftrightarrow U^n$, so that G_n and U^n are independent given X^n.

Why it matters: vertex data can enable (or sharpen) recovery when graph structure alone is near/below the SBM exact-recovery threshold.

Data Block Model (DBM)

The **data block model** is the triple (G_n, X^n, U^n).

(1) Underlying SBM. Draw labels i.i.d. with prior $P = (p_1, \dots, p_k)$: $\mathbb{P}(X_i = a) = p_a, \qquad a \in [k],\ i \in [n]$.
Given X^n, generate the graph $G_n = (Y_{ij})_{1 \le i < j \le n}$ on $V_n = \{v_1, \dots, v_n\}$ with conditional independence and

$$\mathbb{P}(Y_{ij} = 1 \mid X_i = a, X_j = b) = W_{ab}, \qquad 1 \le i < j \le n.$$

(2) Vertex data. Attach a data file to each vertex: $U^n := (U^{(1)}, U^{(2)}, \dots, U^{(n)})$,
and assume the Markov chain $G_n \leftrightarrow X^n \leftrightarrow U^n$. Moreover, conditional on X^n, the data files are independent and

$$U^{(i)} \mid \{X_i = x\} \sim P_{U|X}(\,\cdot\,|x), \qquad x \in [k].$$

Here $P_{U|X}$ is the **data–community random transformation**.

(3) Joint distribution factorization. For (g_n, x^n, u^n),

$$\mathbb{P}(G_n = g_n, X^n = x^n, U^n = u^n) = \prod_{i=1}^{n} P(x_i) \prod_{1 \le i < j \le n} \left((W_{x_i x_j})^{y_{ij}} (1 - W_{x_i x_j})^{1 - y_{ij}} \right) \prod_{i=1}^{n} P_{U|X}\big(u^{(i)} \mid x_i\big).$$

Notation: $(G_n, X^n, U^n) \sim \mathrm{DBM}(n, k, P, W, P_{U|X})$.

Logarithmic-degree regime: $W^{(n)} = \dfrac{\log n}{n} Q$ for fixed symmetric $Q \in \mathbb{R}_+^{k \times k}$.

Recovery notions: almost exact vs exact

Let an algorithm output $\widehat{X}^n = \widehat{X}^n(G_n)$ (SBM) or $\widehat{X}^n = \widehat{X}^n(G_n, U^n)$ (DBM). Because community labels are only identifiable up to permutation, we compare up to $\pi \in \mathrm{Sym}(k)$.

Almost exact recovery (vanishing fraction of mistakes):

$$\mathbb{P}\left[\max_{\pi \in \mathrm{Sym}(k)} \frac{1}{n} \sum_{i=1}^{n} 1\{\widehat{X}_i = \pi(X_i)\} \geq 1 - o(1)\right] = 1 - o(1).$$

Exact recovery (no mistakes):

$$\mathbb{P}\left[\exists\, \pi \in \mathrm{Sym}(k) :\ \widehat{X}^n = \pi(X^n)\right] = 1 - o(1).$$

Intuition: exact recovery is the event that *every* vertex is correctly labeled (up to a global relabeling). In the log-degree regime, SBMs/DBMs exhibit sharp phase transitions for this event.

Exact recovery for SBM (known sharp threshold)

For positive vectors $a, b \in \mathbb{R}_+^k$ define

$$D_+(a\|b) \;:=\; \max_{t\in[0,1]} \sum_{r=1}^{k} \Big((1-t)b_r + ta_r - a_r^t b_r^{1-t}\Big) \qquad \text{(Chernoff–Hellinger divergence).}$$

Assume the **logarithmic-degree regime**: $W^{(n)} = \frac{\log n}{n} Q, \qquad Q \in \mathbb{R}_+^{k\times k}$ symmetric.
Define community-wise mean vectors

$$\mu_r := (\mathrm{diag}(P)Q)_r \in \mathbb{R}_+^k, \qquad r \in [k],$$

where $(\mathrm{diag}(P)Q)_r$ is the r-th column.

Theorem (Abbe–Sandon, 2015):

Let $(G_n, X^n) \sim \mathrm{SBM}\big(n, k, P, \frac{\log n}{n} Q\big)$.

- If for all $i \neq j$, $D_+(\mu_i\|\mu_j) > 1$, then exact recovery is achievable (efficiently).
- If there exist $i \neq j$ with $D_+(\mu_i\|\mu_j) < 1$, then exact recovery is impossible.

Takeaway: a sharp phase transition at $\min_{i\neq j} D_+(\mu_i\|\mu_j) = 1$.

Key quantity for DBM: Chernoff–TV divergence D_{CT}

DBM combines two sources of evidence:

- graph-based statistics (degree-profile / neighborhood information), and
- node data $U^{(i)}$ through $P_{U|X}(\cdot|s)$.

For discrete pmfs $P_X^{(1)}, P_X^{(2)}$ on $\mathcal{X}$ and $Q_U^{(1)}, Q_U^{(2)}$ on $\mathcal{U}$, define

$$D_{\mathrm{CT}}\Big(P_X^{(1)}, Q_U^{(1)} \,\|\, P_X^{(2)}, Q_U^{(2)}\Big) := -\log\left(\sum_{u\in\mathcal{U}} \min_{\lambda_u\in[0,1]} \sum_{x\in\mathcal{X}} \big(P_X^{(1)}(x)Q_U^{(1)}(u)\big)^{\lambda_u}\big(P_X^{(2)}(x)Q_U^{(2)}(u)\big)^{1-\lambda_u}\right).$$

Two points:

- If $Q_U^{(1)} = Q_U^{(2)}$, then D_{CT} reduces to $D_+(P_X^{(1)}\|P_X^{(2)})$.
- If $P_X^{(1)} = P_X^{(2)}$, then $D_{\mathrm{CT}} = -\log\big(1 - \mathrm{TV}(Q_U^{(1)}, Q_U^{(2)})\big)$.

Exact recovery for DBM (this work: sharp threshold)

Consider $(G_n, X^n, U^n) \sim \mathrm{DBM}\Big(n, k, P, \frac{\log n}{n} Q, P^{(n)}_{U|X}\Big)$ and set

$$\mu_r := (\mathrm{diag}(P)Q)_r, \qquad \mu_r^{(n)} := (\log n)\mu_r.$$

Let $P_{\mu_r^{(n)}}$ denote the product Poisson distribution on $\mathbb{Z}_+^k$ with mean vector $\mu_r^{(n)}$.

Define the **pairwise divergence rate** for $s \neq t$:

$$D_{s,t} \;:=\; \liminf_{n\to\infty} \frac{1}{\log n} D_{\mathrm{CT}}\Big(P_{\mu_s^{(n)}},\, P^{(n)}_{U|X}(\cdot|s) \;\Big\|\; P_{\mu_t^{(n)}},\, P^{(n)}_{U|X}(\cdot|t)\Big).$$

Theorem (Davoodi et al., 2026+):

- if $D_{s,t} > 1$ for all $s \neq t$, then **exact recovery is achievable** by a polynomial-time (near-linear-time) algorithm.
- if there exist $s \neq t$ with $D_{s,t} < 1$, then **exact recovery is impossible** (information-theoretically).

Consistency checks: recovers the SBM threshold when data are uninformative, and reduces to a data-only limit when the graph is uninformative.

How the DBM threshold is achieved (high-level algorithm)

Two-stage idea:

1. **Coarse clustering from edges:** split the graph, run a fast SBM routine (Sphere-comparison) on the first subgraph to get σ' with $o(n)$ errors.
2. **Local refinement using edges + data:** on the second subgraph, for each node v form its degree profile

$$d(v) = (d_1(v), \ldots, d_k(v)), \quad d_r(v) = \#\{\text{edges from } v \text{ to nodes labeled } r \text{ by } \sigma'\},$$

and perform a per-node MAP decision combining $d(v)$ and the observed data file $U^{(v)}$.

Why it hits the information limit: the MAP step matches the exponent captured by D_{CT}, so it succeeds exactly when $D_{s,t} > 1$ for all $s \neq t$.

DBM exact-recovery algorithm: input / output

Observed (input)

- Graph $G_n = (Y_{ij})_{1 \le i < j \le n}$ on $V_n = \{v_1, \dots, v_n\}$.
- Vertex data $U^n = (U^{(1)}, \dots, U^{(n)})$.

Goal (output)

Produce $\widehat{X}^n \in [k]^n$ such that

$$\mathbb{P}\Big[\exists\, \pi \in \mathrm{Sym}(k) :\ \widehat{X}^n = \pi(X^n)\Big] = 1 - o(1).$$

Model known to the algorithm

- k and prior $P = (p_1, \dots, p_k)$ on $[k]$.
- Log-degree scaling: $W^{(n)} = \dfrac{\log n}{n} Q$ with symmetric $Q \in \mathbb{R}_+^{k \times k}$.
- Data channel: $P_{U|X}^{(n)}(\cdot|s)$ for each $s \in [k]$ (alphabet $\mathcal{U}^{(n)}$).

Implementation knobs

- edge split parameter $\gamma \in (0, 1)$
- coarse-clustering tolerance $\delta \in (0, 1)$

Simulations: setup (two-community DBM, erased labels)

Model. Balanced two-community DBM with $P = (0.5, 0.5)$ and

$$Q = \begin{pmatrix} a & b \\ b & a \end{pmatrix} \ (a > b), \qquad W^{(n)} = \frac{\log n}{n} Q.$$

Side information (erased-label channel). Alphabet $\mathcal{U} = \{1, 2, \varepsilon\}$ and

$$P_{U|X}^{(n)}(x|x) = 1 - n^{-\alpha}, \qquad P_{U|X}^{(n)}(\varepsilon|x) = n^{-\alpha}.$$

Theory-predicted exponent (symmetric case):

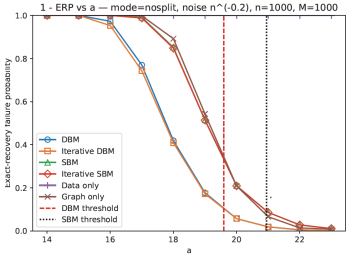
$$D_{1,2} = \alpha + \frac{(\sqrt{a} - \sqrt{b})^2}{2}, \qquad \text{exact recovery iff } D_{1,2} > 1.$$

Equivalently,

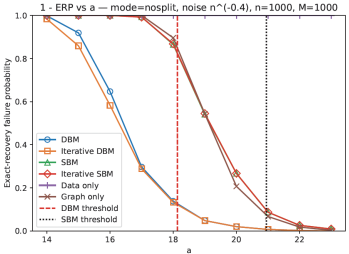
$$a^\star_{\mathrm{DBM}}(b, \alpha) = \left(\sqrt{b} + \sqrt{2(1-\alpha)}\right)^2, \qquad a^\star_{\mathrm{SBM}}(b) = \left(\sqrt{b} + \sqrt{2}\right)^2.$$

Methods compared. DBM / Iterative DBM (graph+data MAP), SBM / Iterative SBM (graph-only MAP), Spectral (graph-only), and Data-only.

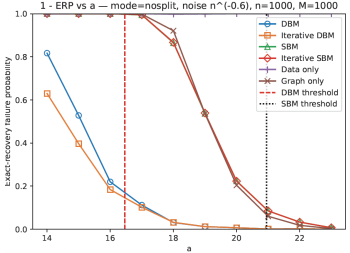
Experiment 1: phase transition across (a, α) (ERP)

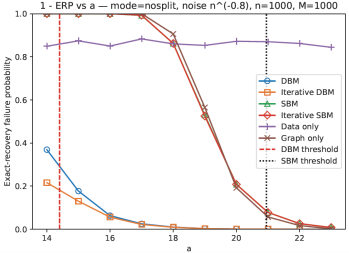


(a) $\alpha = 0.2$



(b) $\alpha = 0.4$





- Fix $n = 1000$, $b = 10$, $M = 1000$ trials.
- Sweep $a \in \{14, \ldots, 23\}$ and $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$.

Key takeaways.

- The empirical transition tracks $a^{\star}_{\mathrm{DBM}}(b, \alpha)$ closely.
- As α increases (more side info), DBM succeeds at smaller a.
- Iteration gives modest ERP gains near the transition.

Punchline. Side information shifts the exact-recovery boundary as expected.

How much side information helps

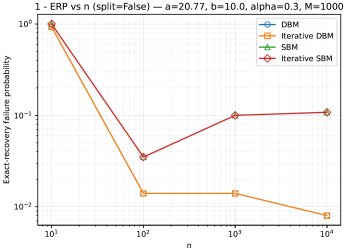
Setting: $n = 1000$, $b = 10$, $M = 1000$ trials; erased-label side channel.

Smallest integer a such that $\mathrm{ERP} \geq 95\%$ (lower is better):

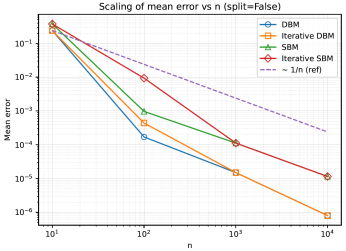
Method	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
DBM	21	19	18	17
SBM	22	22	22	22

Read-off: increasing α shifts the DBM boundary left (easier exact recovery), while graph-only methods remain near the SBM threshold.

Experiment 2: finite-size scaling in n (DBM above threshold, SBM below)



(a) $1 - \mathrm{ERP}$ versus n at $(a, b, \alpha) = (20.77, 10, 0.3)$.



(b) Mean error versus n (log-log) at the same parameters; the thin line marks slope -1.

Figure: $1 - \mathrm{ERP}$ vs n and mean error vs n (log scales).

$$(b, \alpha) = (10, 0.3), \;\; a = 1.10\, a^{\star}_{\mathrm{DBM}}(b, \alpha)$$

so a is *supercritical* for DBM but *subcritical* for SBM.

What happens as n grows.

- DBM: $1 - \mathrm{ERP}$ decays with n (toward exact recovery).
- SBM: $1 - \mathrm{ERP}$ increases with n (cannot reach exact recovery).
- Mean error can be tiny even when ERP isn't 1 (exact recovery is stricter).

"We chose a in the wedge: above the DBM threshold but below the SBM threshold; DBM improves with n while SBM does not."

Agnostic DBM on real data: what must be estimated

Observed

We observe (G_n, U^n): the graph $G_n = (Y_{ij})_{1 \le i < j \le n}$ and node data $U^n = (U^{(1)}, \dots, U^{(n)})$.

Parameters to learn (plug-in)

- k (number of communities).
- $P = (p_1, \dots, p_k)$ (class proportions).
- $Q \in \mathbb{R}_+^{k \times k}$ in the log-degree scaling

$$W^{(n)} = \frac{\log n}{n} Q.$$

- **Data model** $P_{U|X}^{(n)}(\cdot|s)$ (or a surrogate producing $\log P_{U|X}^{(n)}(U^{(i)}|s)$).

Edge split schedule

Split edges into G' (for learning/initialization) and G'' (for final decisions):

$$\mathbb{P}\{(i,j) \in E' \mid (i,j) \in E\} = \gamma_n, \qquad E'' = E \setminus E'.$$

A good asymptotic regime is

$$\gamma_n \to 0 \quad \text{and} \quad \gamma_n \log n \to \infty \quad (\text{e.g. } \gamma_n = 1/\log\log n).$$

Intuition: G' is still strong enough for coarse clustering/estimation, while G'' preserves almost all edges for the final MAP refinement.

A practical plug-in pipeline (agnostic DBM clustering)

1. **Choose / estimate k:** eigengap + stability, or likelihood-based selection (e.g. held-out edges / BIC-style criteria). Denote the selected value by $\widehat{k}$.
2. **Coarse labels (graph-only, fast):** run a scalable method on G_n (e.g. spectral clustering) to get $\sigma^{(0)} \in [\widehat{k}]^n$ with a small fraction of errors.
3. **Estimate the prior P:** $\widehat{p}_s = \frac{|\{i:\sigma^{(0)}(i)=s\}|}{n}, \qquad s \in [\widehat{k}].$
4. **Estimate Q (equivalently $W^{(n)}$):** let $n_s = |\{i : \sigma^{(0)}(i) = s\}|$ and let E_{st} be the number of edges between clusters s, t under $\sigma^{(0)}$. Define

$$\widehat{W}_{st} = \begin{cases} \dfrac{E_{st}}{n_s n_t}, & s \neq t, \\ \dfrac{2E_{ss}}{n_s(n_s-1)}, & s = t, \end{cases} \qquad \widehat{Q}_{st} = \frac{n}{\log n}\widehat{W}_{st}.$$

   (Optionally symmetrize and regularize: $\widehat{Q} \leftarrow (\widehat{Q} + \widehat{Q}^\top)/2$.)
5. **Estimate the data model $P^{(n)}_{U|X}$:**
   - If U is discrete: $\widehat{P}^{(n)}_{U|X}(u|s) = \frac{\#\{i:\sigma^{(0)}(i)=s,\ U^{(i)}=u\}+\eta}{n_s+\eta|\mathcal{U}^{(n)}|}$.
   - If U is continuous/high-dim: fit a per-class generative model $u \mapsto \log \widehat{P}^{(n)}_{U|X}(u|s)$ (e.g. Gaussian/naive Bayes/topic model), or a calibrated discriminative surrogate producing log-likelihood scores.

Real-data case study: chemical network + structure fingerprints

Objects (DBM view).

- Nodes $V_n = \{v_1, \ldots, v_n\}$ are **chemicals**.
- Graph $G_n = (Y_{ij})_{1 \le i < j \le n}$ is an **undirected** chemical–chemical network.
- Node data $U^{(i)}$ is a **binary chemical fingerprint** (structure feature vector).

Clustering target. Latent label $X_i \in [k]$ represents the **disease category** that chemical i treats (*used only for evaluation*).

Network summary

#nodes n	1329
#edges m	5534
density	0.0063
avg degree	8.33
median degree	6
max degree	425
#components	1

Node attributes summary

fingerprint dimension d	1024 bits
avg #ones per node	106.96
median #ones	105
min / max #ones	56 / 215

DBM interpretation

DBM mapping

We view the dataset as a realization of (G_n, X^n, U^n) where

- G_n is the observed chemical–chemical graph,
- $U^n = (U^{(1)}, \ldots, U^{(n)})$ are 1024-bit fingerprints,
- $X_i \in [k]$ is the (unknown) disease label of chemical i.

The agnostic method takes **only** (G_n, U^n) **as input**; X^n is used **only for evaluation**.

Ground truth summary

$k = 4$ disease groups (0: bacterial infectious disease, 1: hypertensive disorder, 2: visual epilepsy, 3: asthma) with sizes:

$$(417,\ 308,\ 287,\ 317).$$

Empirical edge densities by disease (evidence of block structure)

	d0	d1	d2	d3
d0	0.01357	0.00299	0.00209	0.00173
d1	0.00299	0.01388	0.01027	0.00425
d2	0.00209	0.01027	0.01601	0.00264
d3	0.00173	0.00425	0.00264	0.01236

Within-disease densities (diag): 0.012–0.016; between-disease: min 0.0017, max 0.0103.

Agnostic DBM on this dataset: model choices + learned parameters

Instantiation choices (practical)

- $k = 4$.
- **Attribute model:** binary fingerprints $\Rightarrow$ per-cluster *product Bernoulli* model

$$U^{(i)} \mid \{X_i = s\} \;\approx\; \prod_{b=1}^{1024} \mathrm{Bern}(\Theta_{s,b}).$$

- **Agnostic learning:** estimate $(\widehat{P}, \widehat{W}, \widehat{\Theta})$ from a coarse partition, then run local MAP refinement combining degree profiles + attribute likelihoods.

Learned parameters (summary)

Estimated cluster proportions $\widehat{\pi}$:

$$(0.253,\ 0.305,\ 0.188,\ 0.254).$$

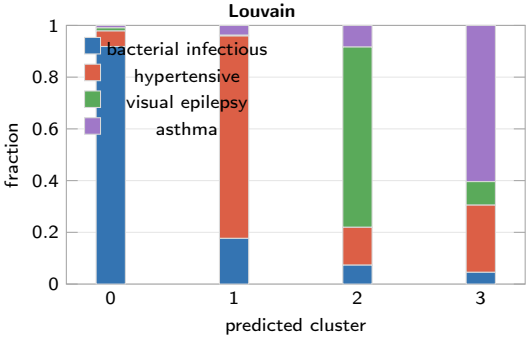
Estimated edge-probability matrix $\widehat{W}$ between inferred clusters:

$$\begin{pmatrix} 0.02054 & 0.00185 & 0.00130 & 0.00774 \\ 0.00185 & 0.01417 & 0.00042 & 0.00183 \\ 0.00130 & 0.00042 & 0.01626 & 0.00124 \\ 0.00774 & 0.00183 & 0.00124 & 0.01907 \end{pmatrix}$$

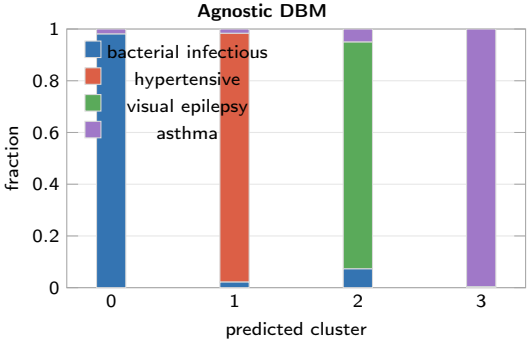
Fingerprint model $\widehat{\Theta}$ (4×1024): range [0.0025, 0.9444]; expected #ones per cluster:

$$(104.8,\ 121.2,\ 111.3,\ 98.3).$$

Results: Louvain vs agnostic DBM



Macro purity (approx.): 0.75



Macro purity (approx.): 0.95

Quantitative evaluation (against disease labels; after best permutation)

Initialization: acc 0.8766, ARI 0.7199, NMI 0.6884 $\longrightarrow$ Refinement: acc 0.9533, ARI 0.8849, NMI 0.8588.

Thank you!

Questions or comments?

akb@sdu.dk