# Semi-supervised clustering via Markov Aggregation

*Poster presentation:*

SAC'22, Machine Learning and its Applications Track

Authors: Sophie Steger [1], Bernhard C. Geiger [2], Marek Śmieja [3]

Affiliations: *Graz University of Technology, Austria* [1]
*Know-Center GmbH, Austria* [2]
*Jagiellonian University, Poland* [3]

on April 27, 2022

# » Content

## Introduction

* Unsupervised clustering
* Semi-supervised clustering

## Introduction

* Unsupervised clustering
* Semi-supervised clustering

# » Unsupervised clustering

Given:

# » Unsupervised clustering

Given:

* Set of unlabeled data points

# » Unsupervised clustering

Given:

* Set of unlabeled data points

Aim:

# » Unsupervised clustering

Given:

- Set of unlabeled data points

Aim:

- Group data points based on some "similarity" measure

## Introduction

* Unsupervised clustering
* Semi-supervised clustering

# » Semi-supervised clustering

Prior information:

# » Semi-supervised clustering

Prior information:

* Partition-level side information: labeled data points

# » Semi-supervised clustering

Prior information:

* Partition-level side information: labeled data points
* Instance-level side information: pairwise constraints

# » Semi-supervised clustering

Prior information:

* Partition-level side information: labeled data points
* Instance-level side information: pairwise constraints
  * must-link constraints

# » Semi-supervised clustering

Prior information:

* Partition-level side information: labeled data points
* Instance-level side information: pairwise constraints
    * must-link constraints
    * cannot-link constraints

## Semi-Supervised Clustering via Markov Aggregation

* Markov Aggregation
* Clustering via Markov Aggregation
* Algorithm

## Semi-Supervised Clustering via Markov Aggregation

* Markov Aggregation
* Clustering via Markov Aggregation
* Algorithm

# » Markov Aggregation

## Introduction

Given:

# » Markov Aggregation

Introduction

Given:

* Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

## » Markov Aggregation



**Given:**

* Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

**Aim:**

# » Markov Aggregation



Given:

* Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

Aim:

* Find a Markov chain $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbb{Q})$ on the alphabet $\mathcal{Y}$ that

# » Markov Aggregation

Introduction

Given:

* Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

Aim:

* Find a Markov chain $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbb{Q})$ on the alphabet $\mathcal{Y}$ that
  * best approximates $\mathbf{X}$

# » Markov Aggregation

## Introduction

Given:

* Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

Aim:

* Find a Markov chain $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbb{Q})$ on the alphabet $\mathcal{Y}$ that
  * best approximates $\mathbf{X}$
  * on a smaller alphabet $\mathcal{Y}$

# » Markov Aggregation



**Given:**

- Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

**Aim:**

- Find a Markov chain $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbb{Q})$ on the alphabet $\mathcal{Y}$ that
  - best approximates $\mathbf{X}$
  - on a smaller alphabet $\mathcal{Y}$

**How:**

# » Markov Aggregation



Given:

- Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

Aim:

- Find a Markov chain $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbb{Q})$ on the alphabet $\mathcal{Y}$ that
  - best approximates $\mathbf{X}$
  - on a smaller alphabet $\mathcal{Y}$

How:

- Mapping of $\mathbf{X}$ onto $\mathbf{Y}$ via a

# » Markov Aggregation



Given:

* Markov chain $\mathbf{X} \sim \mathrm{Mar}(\mathcal{X}, \mathbb{P})$ on the alphabet $\mathcal{X}$

Aim:

* Find a Markov chain $\mathbf{Y} \sim \mathrm{Mar}(\mathcal{Y}, \mathbb{Q})$ on the alphabet $\mathcal{Y}$ that
  * best approximates $\mathbf{X}$
  * on a smaller alphabet $\mathcal{Y}$

How:

* Mapping of $\mathbf{X}$ onto $\mathbf{Y}$ via a
  * deterministic mapping $\rightarrow g\colon \mathcal{X} \rightarrow \mathcal{Y}$

# » Cost function

## Rationale

*The mapping $g$ should achieve that . . .*

## » Cost function

Rationale

*The mapping $g$ should achieve that . . .*

1. the hidden Markov process $\mathbf{Y}$ is *Markov* &

# » Cost function

## Rationale

*The mapping $g$ should achieve that . . .*

1. the hidden Markov process $\mathbf{Y}$ is *Markov* &
2. preserves the *temporal dependence structure* of $\mathbf{X}$.

# » Cost function

$$\boxed{\mathcal{C}_\beta(\mathbf{X}, \mathbb{W}) := (1 - 2\beta)\left(H(Y_2|Y_1) - H(Y_2|X_1)\right) - \beta I(Y_1; Y_2)}$$

Rana Ali Amjad, Clemens Blochl, and Bernhard C. Geiger. "A Generalized Framework For Kullback–Leibler Markov Aggregation". In: *IEEE Transactions on Automatic Control* 65.7 (July 2020), pp. 3068–3075. ISSN: 2334-3303. DOI: 10.1109/tac.2019.2945891. URL: http://dx.doi.org/10.1109/TAC.2019.2945891

# » Cost function

$$\boxed{\mathcal{C}_\beta(\mathbf{X}, \mathbb{W}) := (1-2\beta)\left(H(Y_2|Y_1) - H(Y_2|X_1)\right) - \beta I(Y_1;Y_2)}$$

☞ Special cases:

Amjad, Blochl, and Geiger, "A Generalized Framework For Kullback–Leibler Markov Aggregation"

# » Cost function

$$\boxed{\mathcal{C}_\beta(\mathbf{X}, \mathbb{W}) := (1-2\beta)\left(H(Y_2|Y_1) - H(Y_2|X_1)\right) - \beta I(Y_1;Y_2)}$$

☞ Special cases:

- $\beta = 0.5$: *information-theoretic pairwise clustering*

---

Amjad, Blochl, and Geiger, "A Generalized Framework For Kullback–Leibler Markov Aggregation"

# » Cost function

$$\boxed{\mathcal{C}_\beta(\mathbf{X}, \mathbb{W}) := (1-2\beta)\left(H(Y_2|Y_1) - H(Y_2|X_1)\right) - \beta I(Y_1;Y_2)}$$

☞ Special cases:

- $\beta = 0.5$: *information-theoretic pairwise clustering*
- $\beta = 1$: *information bottleneck problem*

---

Amjad, Blochl, and Geiger, "A Generalized Framework For Kullback–Leibler Markov Aggregation"

## Semi-Supervised Clustering via Markov Aggregation

* Markov Aggregation
* Clustering via Markov Aggregation
* Algorithm

# » Clustering via Markov Aggregation

Introduction

Idea:

## » Clustering via Markov Aggregation



Idea:

* Minimizing the cost function $\rightarrow$ grouping of states $\mathbf{X}$

# » Clustering via Markov Aggregation



Idea:

* Minimizing the cost function $\rightarrow$ grouping of states $\mathbf{X}$
* Use for grouping of data points

# » Clustering via Markov Aggregation

Introduction

Idea:

* Minimizing the cost function $\rightarrow$ grouping of states $\mathbf{X}$
* Use for grouping of data points

How:

# » Clustering via Markov Aggregation



Idea:

- Minimizing the cost function $\rightarrow$ grouping of states $\mathbf{X}$
- Use for grouping of data points

How:

- Construct the transition probability matrix $\mathbb{P}$

$$P_{i \to j} \propto \mathrm{e}^{-\frac{\|x_i - x_j\|_2^2}{\sigma_k}}$$

# » Clustering via Markov Aggregation



Idea:

* Minimizing the cost function $\to$ grouping of states $\mathbf{X}$
* Use for grouping of data points

How:

* Construct the transition probability matrix $\mathbb{P}$

$$P_{i \to j} \propto \mathrm{e}^{-\frac{\|x_i - x_j\|_2^2}{\sigma_k}}$$

* $x_i$, $x_j$ . . . . . coordinate of $i$-th and $j$-th data point

# » Clustering via Markov Aggregation



Idea:

- Minimizing the cost function $\rightarrow$ grouping of states $\mathbf{X}$
- Use for grouping of data points

How:

- Construct the transition probability matrix $\mathbb{P}$

$$P_{i\to j} \propto \mathrm{e}^{-\frac{\|x_i - x_j\|_2^2}{\sigma_k}}$$

- $x_i$, $x_j$ . . . . . coordinate of $i$-th and $j$-th data point
- $\|\cdot\|_2^2$ . . . . . squared Euclidean distance

# » Clustering via Markov Aggregation



Idea:

* Minimizing the cost function → grouping of states $\mathbf{X}$
* Use for grouping of data points

How:

* Construct the transition probability matrix $\mathbb{P}$

$$P_{i\to j} \propto \mathrm{e}^{-\frac{\|x_i - x_j\|_2^2}{\sigma_k}}$$

* $x_i$, $x_j$ . . . . . coordinate of $i$-th and $j$-th data point
* $\|\cdot\|_2^2$ . . . . . squared Euclidean distance
* $\sigma_k$ . . . . . . scaling parameter

## Semi-Supervised Clustering via Markov Aggregation

* Markov Aggregation
* Clustering via Markov Aggregation
* Algorithm

# » Semi-supervised sequential algorithm

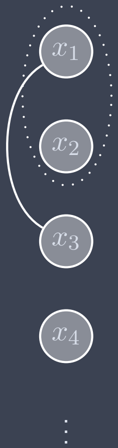
## Visualization

## » Semi-supervised sequential algorithm

Visualization

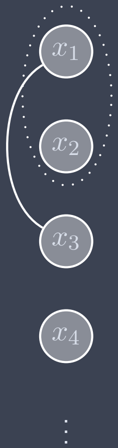
# » Semi-supervised sequential algorithm

## Visualization



# » Semi-supervised sequential algorithm

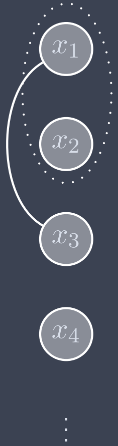
## Visualization




* Clusters:

# » Semi-supervised sequential algorithm

## Visualization

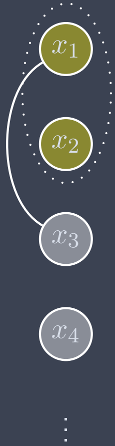



* Clusters:


* Initial aggregation function $g_{init} \longrightarrow$ *greedy coloring algorithm*

# » Semi-supervised sequential algorithm

## Visualization

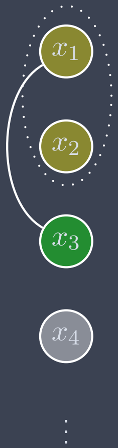



* Clusters:


* Initial aggregation function $g_{init} \longrightarrow$ *greedy coloring algorithm*

# » Semi-supervised sequential algorithm

Visualization

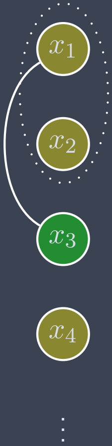



* Clusters:


* Initial aggregation function $g_{init} \longrightarrow$ *greedy coloring algorithm*

# » Semi-supervised sequential algorithm

## Visualization




- Clusters:

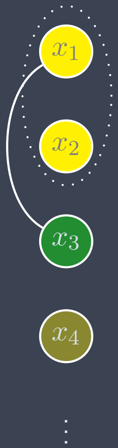

- Initial aggregation function $g_{init} \longrightarrow$ *greedy coloring algorithm*

# » Semi-supervised sequential algorithm

## Visualization

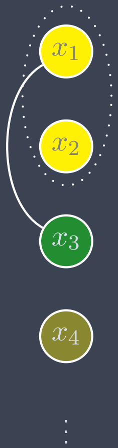
# » Semi-supervised sequential algorithm

## Visualization



# » Semi-supervised sequential algorithm

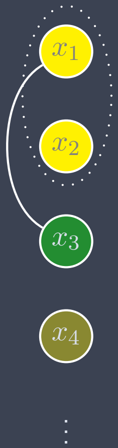
Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

# » Semi-supervised sequential algorithm

Visualization

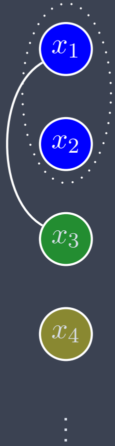



$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$y_2$ *not allowed*

# » Semi-supervised sequential algorithm

Visualization

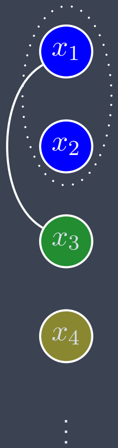



$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$y_2$ *not allowed*

# » Semi-supervised sequential algorithm

## Visualization



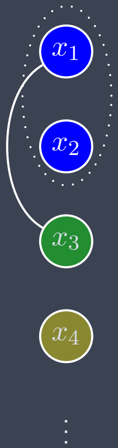

$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$y_2$ *not allowed*

$$C_{g_{y_3}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_3})$$

# » Semi-supervised sequential algorithm

## Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

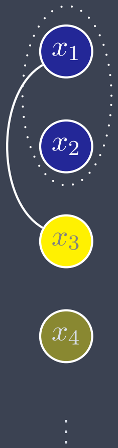
$y_2$ *not allowed*

$$C_{g_{y_3}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_3})$$

$$\Rightarrow g = \arg\min_{g_{y_i}} C_{g_{y_i}}$$

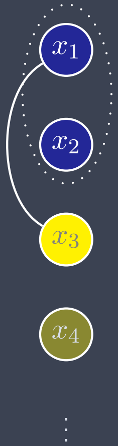
# » Semi-supervised sequential algorithm

Visualization



# » Semi-supervised sequential algorithm

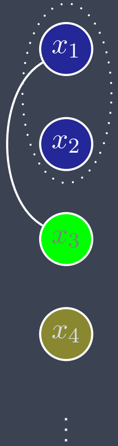
## Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

## » Semi-supervised sequential algorithm

Visualization

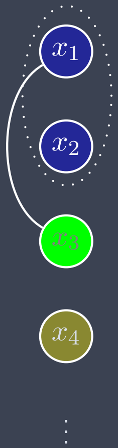



$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

# » Semi-supervised sequential algorithm

Visualization



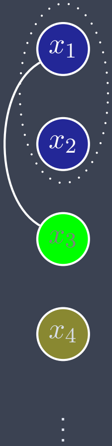

$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

$y_3$ *not allowed*

## » Semi-supervised sequential algorithm

### Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

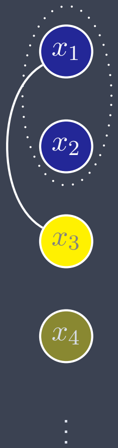
$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

$y_3$ *not allowed*

$$\Rightarrow g = \arg\min_{g_{y_i}} C_{g_{y_i}}$$

# » Semi-supervised sequential algorithm

Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

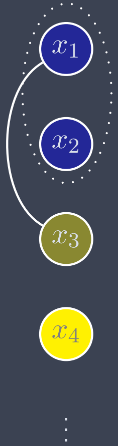
$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

$y_3$ *not allowed*

$$\Rightarrow g = \arg\min_{g_{y_i}} C_{g_{y_i}}$$

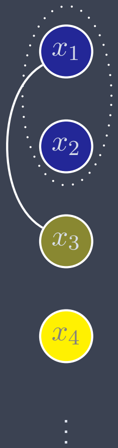
# » Semi-supervised sequential algorithm

## Visualization



# » Semi-supervised sequential algorithm

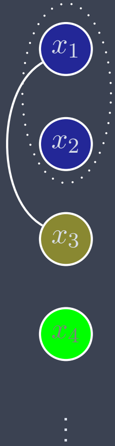
## Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

# » Semi-supervised sequential algorithm

Visualization

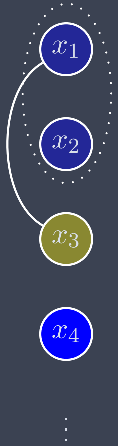



$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

## » Semi-supervised sequential algorithm

### Visualization



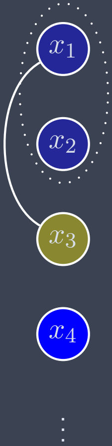

$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

$$C_{g_{y_3}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_3})$$

# » Semi-supervised sequential algorithm

## Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

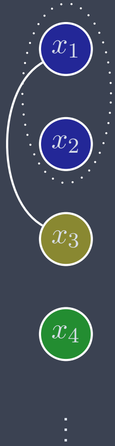
$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

$$C_{g_{y_3}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_3})$$

$$\Rightarrow g = \arg\min_{g_{y_i}} C_{g_{y_i}}$$

# » Semi-supervised sequential algorithm

## Visualization




$$C_{g_{y_1}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_1})$$

$$C_{g_{y_2}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_2})$$

$$C_{g_{y_3}} = \mathcal{C}_\beta(\mathbf{X}, g_{y_3})$$

$$\Rightarrow g = \arg\min_{g_{y_i}} C_{g_{y_i}}$$

## Experiments

* Experimental setup
* Influence of the hyperparameters
* Side-information
* Few labeled classes

# Experiments

* Experimental setup
* Influence of the hyperparameters
* Side-information
* Few labeled classes

# » Clustering accuracy

Normalized Mutual Information:

$$\mathsf{NMI}(\mathcal{X},\mathcal{Y}) = \frac{2I(\mathcal{Y};\mathcal{X})}{H(\mathcal{Y})+H(\mathcal{X})}$$

# » Clustering accuracy

Normalized Mutual Information:

$$\mathsf{NMI}(\mathcal{X},\mathcal{Y}) = \frac{2I(\mathcal{Y};\mathcal{X})}{H(\mathcal{Y}) + H(\mathcal{X})}$$

* $\mathcal{X}$ . . . . . true partition

# » Clustering accuracy

Normalized Mutual Information:

$$\mathrm{NMI}(\mathcal{X}, \mathcal{Y}) = \frac{2I(\mathcal{Y}; \mathcal{X})}{H(\mathcal{Y}) + H(\mathcal{X})}$$

* $\mathcal{X}$ . . . . . true partition
* $\mathcal{Y}$ . . . . . estimated partition

# » Clustering accuracy

Normalized Mutual Information:

$$\mathrm{NMI}(\mathcal{X},\mathcal{Y}) = \frac{2I(\mathcal{Y};\mathcal{X})}{H(\mathcal{Y})+H(\mathcal{X})}$$

* $\mathcal{X}$ . . . . . true partition
* $\mathcal{Y}$ . . . . . estimated partition
* $\mathrm{NMI} = 0$: *no mutual information*

# » Clustering accuracy

Normalized Mutual Information:

$$\mathrm{NMI}(\mathcal{X},\mathcal{Y}) = \frac{2I(\mathcal{Y};\mathcal{X})}{H(\mathcal{Y}) + H(\mathcal{X})}$$

* $\mathcal{X}$ . . . . . . true partition
* $\mathcal{Y}$ . . . . . . estimated partition
* NMI $= 0$: *no mutual information*
* NMI $= 1$: *identical partitions*

# » Side-information

Pairwise constraints:

# » Side-information

Pairwise constraints:

* Select a fraction of data points:

# » Side-information

Pairwise constraints:

* Select a fraction of data points:
  - ⇒ Assign the true label

# » Side-information

Pairwise constraints:

- Select a fraction of data points:
  - ⇒ Assign the true label
  - ⇒ Generation of pairwise constraints

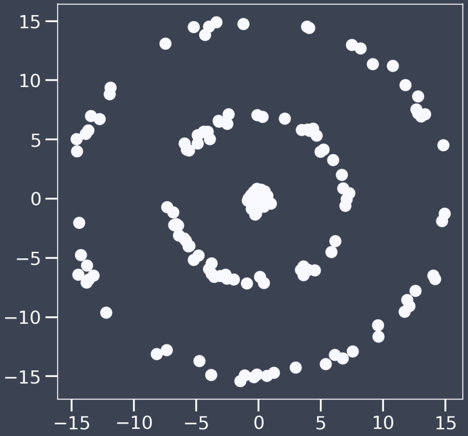
# » Side-information

Pairwise constraints:

* Select a fraction of data points:
  * ⇒ Assign the true label
  * ⇒ Generation of pairwise constraints
* Average resulting NMI over 10 runs

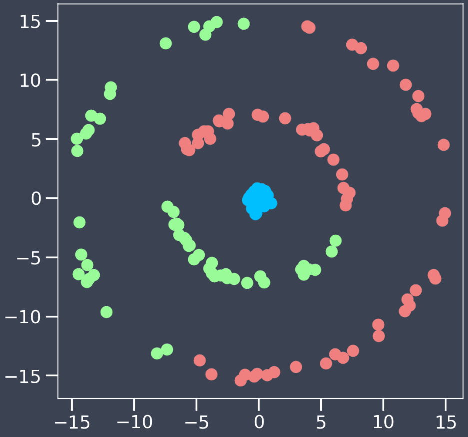
# » Side-information

## Unsupervised Clustering



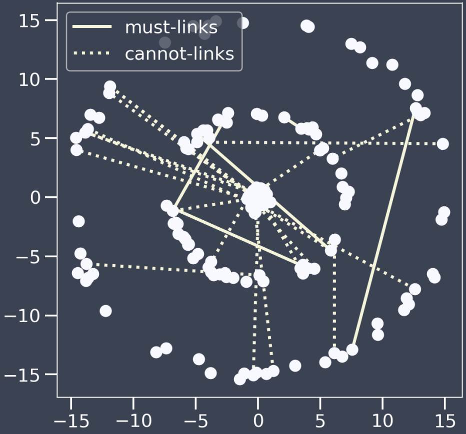
# » Side-information

## Unsupervised Clustering



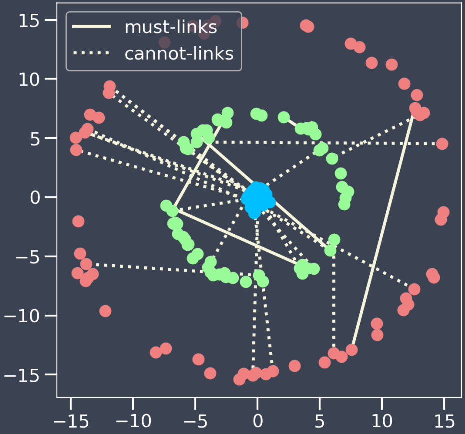
# » Side-information

## Semisupervised Clustering



# » Side-information

Semisupervised Clustering



# Experiments

* Experimental setup
* Influence of the hyperparameters
* Side-information
* Few labeled classes

# » Influence of k

Setup

Setup:

# » Influence of k

Setup

Setup:

* Parameters: $\beta = \{0, 0.5, 1\}$, $k$ varied

# » Influence of k

Setup

Setup:

- Parameters: $\beta = \{0, 0.5, 1\}$, $k$ varied
- Fraction of labeled data: 0, 10, 20, 30%

# » Influence of k

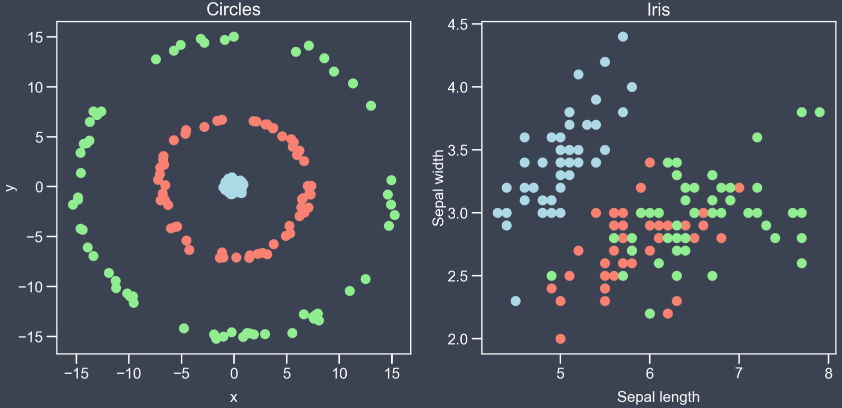
Setup

Setup:

- Parameters: $\beta = \{0, 0.5, 1\}$, $k$ varied
- Fraction of labeled data: 0, 10, 20, 30%
- Datasets: circles, iris

# » Influence of k

## Setup



# » Influence of k

Setup

Setup:

* Parameters: $\beta = \{0, 0.5, 1\}$, $k$ varied
* Fraction of labeled data: 0, 10, 20, 30%
* Datasets: circles, iris

Observations:

# » Influence of k

Setup

Setup:

- Parameters: $\beta = \{0, 0.5, 1\}$, $k$ varied
- Fraction of labeled data: 0, 10, 20, 30%
- Datasets: circles, iris

Observations:

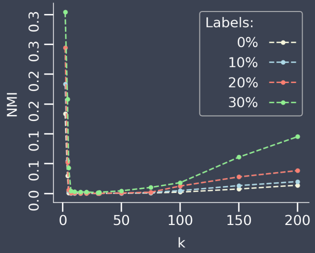
- Influence on accuracy for *unsupervised* and *semi-supervised* case

## » Influence of k

Results: Circles dataset

# » Influence of k

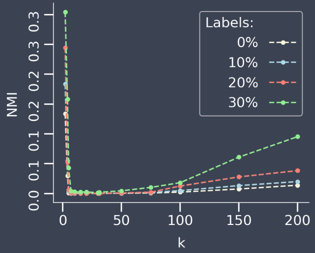
## Results: Circles dataset



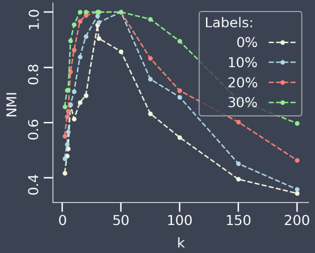

$\beta = 0$

# » Influence of k

## Results: Circles dataset



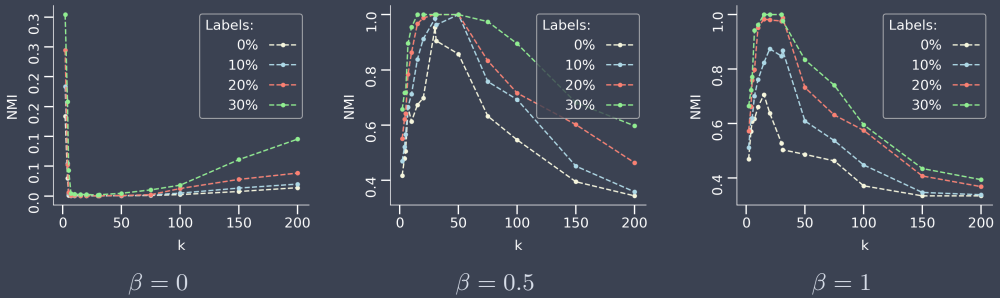

$\beta = 0$




$\beta = 0.5$

# » Influence of k

## Results: Circles dataset

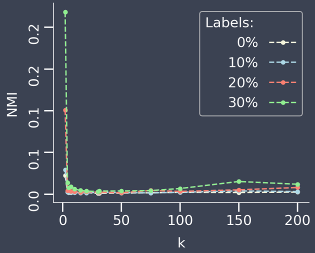


## » Influence of k

Results: Iris dataset

## » Influence of k

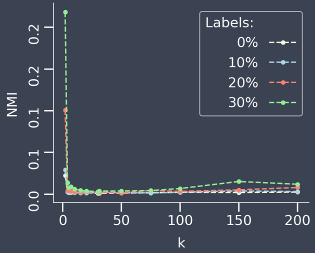
Results: Iris dataset



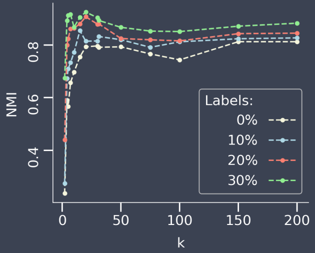

$\beta = 0$

# » Influence of k

## Results: Iris dataset



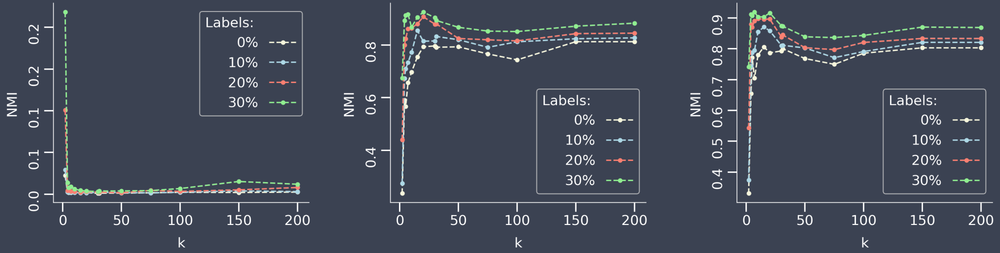

$\beta = 0$




$\beta = 0.5$

# » Influence of k

## Results: Iris dataset




$\beta = 0$ $\beta = 0.5$ $\beta = 1$

# » Influence of beta

Setup

Setup:

# » Influence of beta

Setup

Setup:

* Parameters: $k = 20$, $\beta$ varied

# » Influence of beta



Setup:

* Parameters: $k = 20$, $\beta$ varied
* Datasets: selected UCI datasets

# » Influence of beta

## Setup

Setup:

* Parameters: $k = 20$, $\beta$ varied
* Datasets: selected UCI datasets

# » Influence of beta

## Setup

Setup:

- Parameters: $k = 20$, $\beta$ varied
- Datasets: selected UCI datasets

Observations:

# » Influence of beta

Setup

Setup:

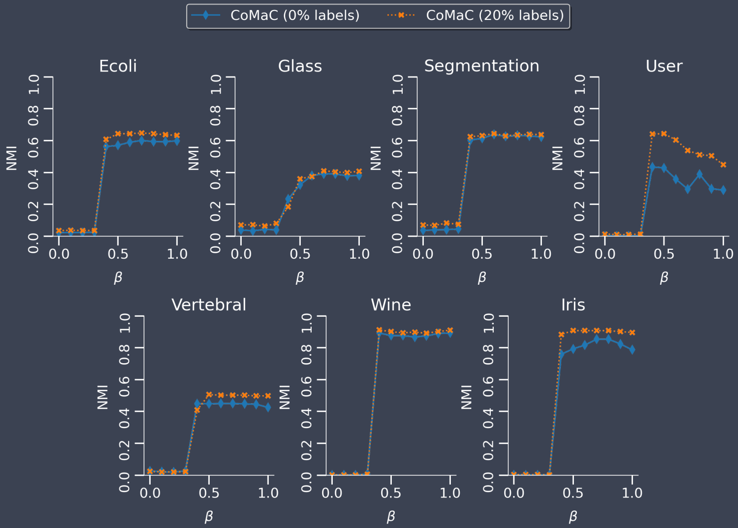
* Parameters: $k = 20$, $\beta$ varied
* Datasets: selected UCI datasets

Observations:

* Influence on accuracy for *unsupervised* and *semi-supervised* case

# » Influence of beta

## Results: semi-supervised (20% labeled data)



## Experiments

* Experimental setup
* Influence of the hyperparameters
* Side-information
* Few labeled classes

# » Side-information



Setup:

# » Side-information

Setup

Setup:

* Fraction of labeled data: 0, 10, 20, 30%

# » Side-information

Setup

Setup:

- Fraction of labeled data: 0, 10, 20, 30%
- $\beta = 0.5$, $k = 20$

[1] Marek Smieja and Bernhard C. Geiger. "Semi-supervised cross-entropy clustering with information bottleneck constraint". In: *Information Sciences* 421 (Dec. 2017), pp. 254–271

# » Side-information

Setup

Setup:

* Fraction of labeled data: 0, 10, 20, 30%
* $\beta = 0.5$, $k = 20$
* Comparison to state-of-the-art-algorithms [1]

[1] Smieja and Geiger, "Semi-supervised cross-entropy clustering with information bottleneck constraint"

# » Side-information

Setup

Setup:

* Fraction of labeled data: 0, 10, 20, 30%
* $\beta = 0.5$, $k = 20$
* Comparison to state-of-the-art-algorithms [1]

Observations:

[1] Smieja and Geiger, "Semi-supervised cross-entropy clustering with information bottleneck constraint"

# » Side-information

Setup

Setup:

- Fraction of labeled data: 0, 10, 20, 30%
- $\beta = 0.5$, $k = 20$
- Comparison to state-of-the-art-algorithms [1]

Observations:

- Influence of labeled data on accuracy

[1] Smieja and Geiger, "Semi-supervised cross-entropy clustering with information bottleneck constraint"

# » Side-information

Setup

Setup:

* Fraction of labeled data: 0, 10, 20, 30%
* $\beta = 0.5$, $k = 20$
* Comparison to state-of-the-art-algorithms [1]

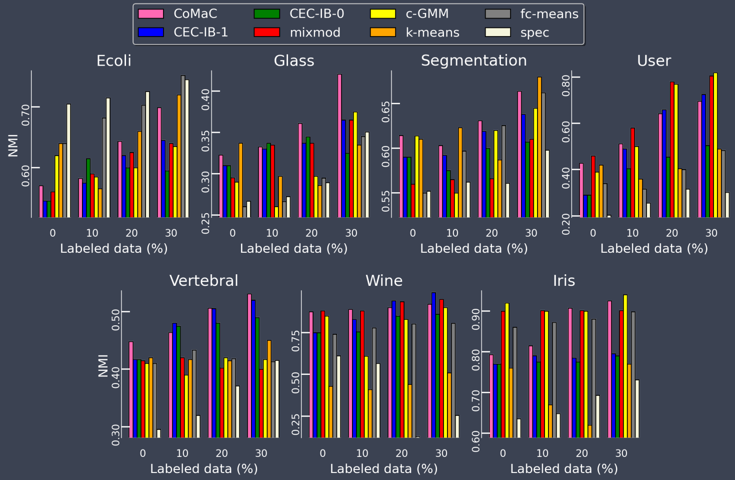
Observations:

* Influence of labeled data on accuracy
* Performance compared to other methods

[1] Smieja and Geiger, "Semi-supervised cross-entropy clustering with information bottleneck constraint"

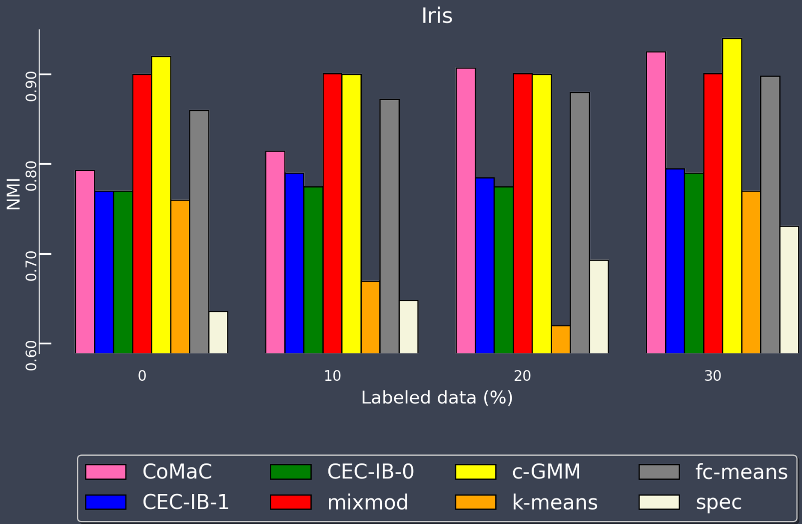
# » Side-information

## Results: Overview



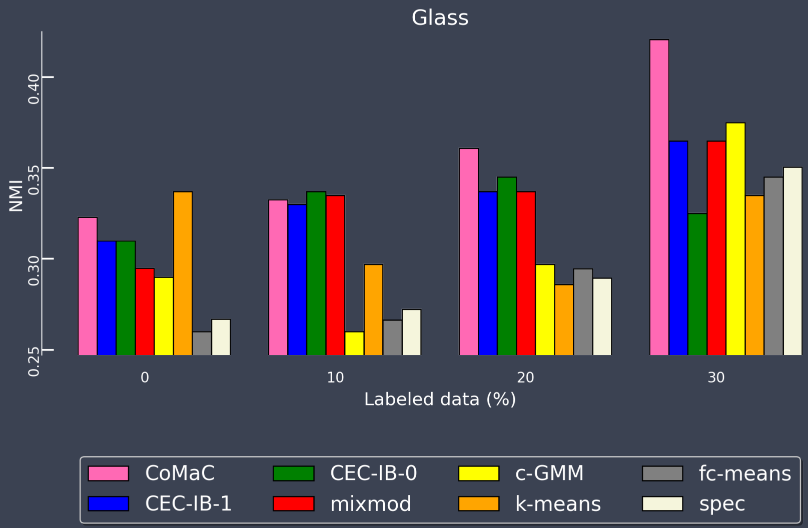
# » Side-information

## Results: Details



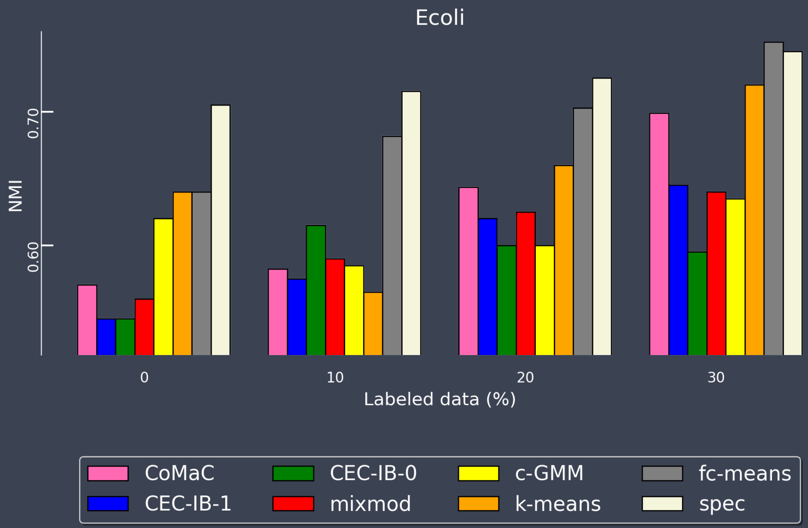
# » Side-information

## Results: Details



# » Side-information

## Results: Details



## Experiments

* Experimental setup
* Influence of the hyperparameters
* Side-information
* Few labeled classes

# » Few labeled classes

Setup

Setup:

## » Few labeled classes

Setup

Setup:

- Fraction of labeled data: 0, 10, 20, 30%

# » Few labeled classes

Setup

Setup:

* Fraction of labeled data: 0, 10, 20, 30%
* Samples taken from 2 classes

# » Few labeled classes

Setup

Setup:

* Fraction of labeled data: 0, 10, 20, 30%
* Samples taken from 2 classes
  * Classes are selected randomly

# » Few labeled classes

Setup

Setup:

* Fraction of labeled data: 0, 10, 20, 30%
* Samples taken from 2 classes
    * Classes are selected randomly
    * Must cover $> 30\%$ of total data

# » Few labeled classes

Setup

Setup:

- Fraction of labeled data: 0, 10, 20, 30%
- Samples taken from 2 classes
  - Classes are selected randomly
  - Must cover $> 30\%$ of total data

Observations:

# » Few labeled classes

Setup

Setup:

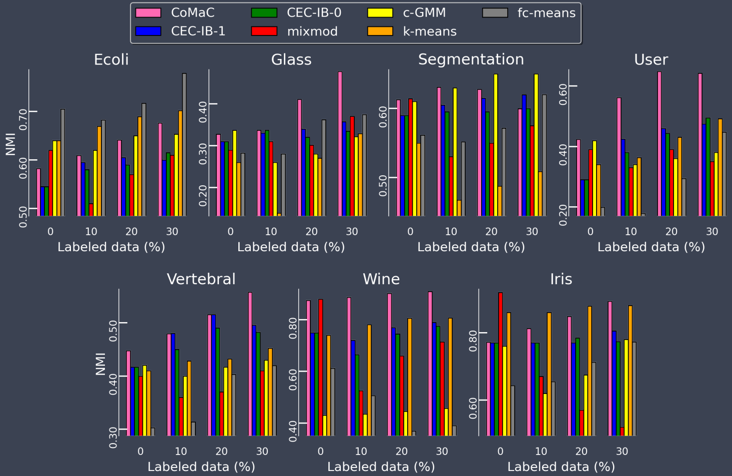
- Fraction of labeled data: 0, 10, 20, 30%
- Samples taken from 2 classes
  - Classes are selected randomly
  - Must cover $> 30\%$ of total data

Observations:

- Influence of labeled data on accuracy

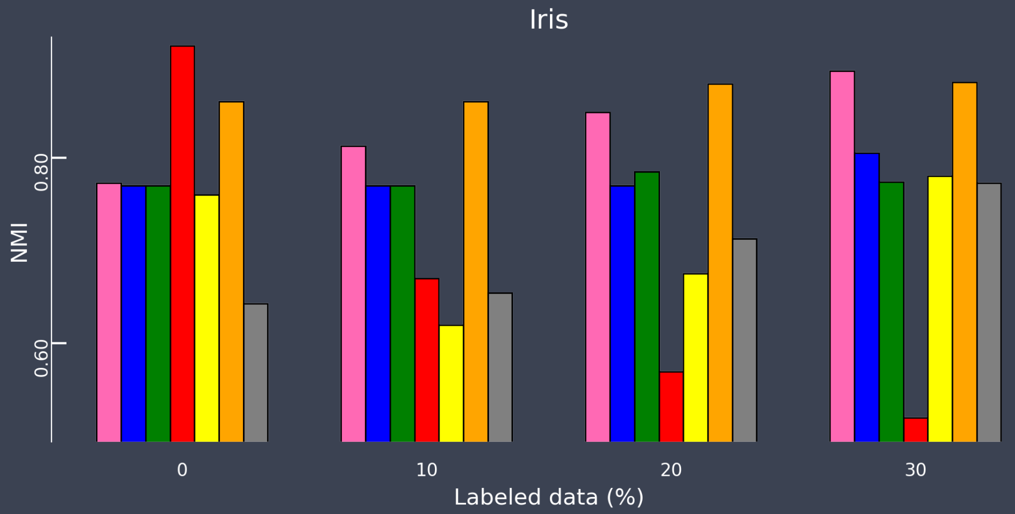
# » Few labeled classes

## Results: Few vs all classes



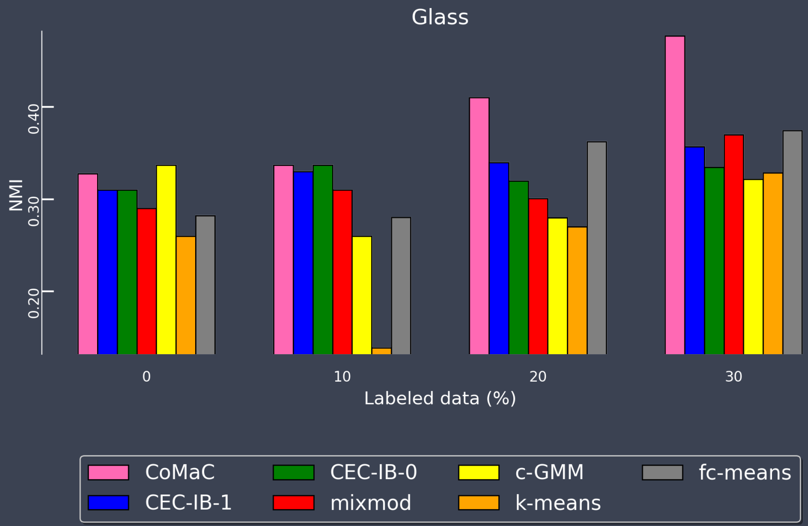
## » Few labeled classes

Results: Details



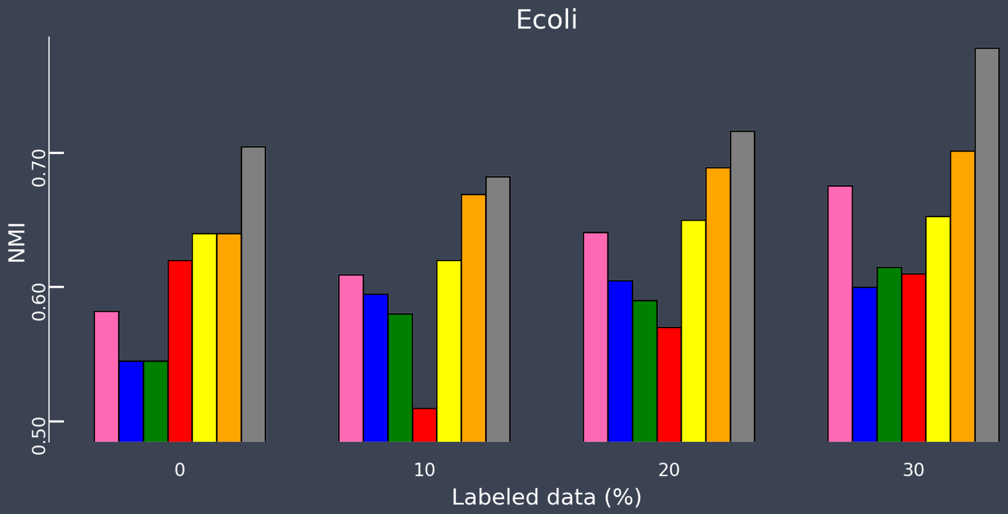
# » Few labeled classes

## Results: Details



# » Few labeled classes

## Results: Details



# Limitations

# » Limitations

⚠ Erroneous partition-level side information

- Sensitive to noisy constraints

# » Limitations

- ⚠ Erroneous partition-level side information
  - ∗ Sensitive to noisy constraints
- ⚠ Non exhaustive pairwise constraints

# » Limitations

- Erroneous partition-level side information
  - Sensitive to noisy constraints
- Non exhaustive pairwise constraints
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{N}$, then also $x'$ and $x''$ should not link

# » Limitations

- ⚠ Erroneous partition-level side information
  - Sensitive to noisy constraints
- ⚠ Non exhaustive pairwise constraints
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{N}$, then also $x'$ and $x''$ should not link
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{M}$, then also $x'$ and $x''$ should link

# » Limitations

- Erroneous partition-level side information
  - Sensitive to noisy constraints
- Non exhaustive pairwise constraints
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{N}$, then also $x'$ and $x''$ should not link
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{M}$, then also $x'$ and $x''$ should link
- Negative influence of cannot-link constraints.

# » Limitations

- Erroneous partition-level side information
  - Sensitive to noisy constraints
- Non exhaustive pairwise constraints
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{N}$, then also $x'$ and $x''$ should not link
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{M}$, then also $x'$ and $x''$ should link
- Negative influence of cannot-link constraints.
  - Moderate numbers of cannot-link constraints

# » Limitations

- Erroneous partition-level side information
  - Sensitive to noisy constraints
- Non exhaustive pairwise constraints
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{N}$, then also $x'$ and $x''$ should not link
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{M}$, then also $x'$ and $x''$ should link
- Negative influence of cannot-link constraints.
  - Moderate numbers of cannot-link constraints
  - Initial coloring: cannot-link constraints $\rightarrow$ assign to the first cluster available

# » Limitations

- Erroneous partition-level side information
  - Sensitive to noisy constraints
- Non exhaustive pairwise constraints
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{N}$, then also $x'$ and $x''$ should not link
  - if $(x, x') \in \mathcal{M}$ and $(x, x'') \in \mathcal{M}$, then also $x'$ and $x''$ should link
- Negative influence of cannot-link constraints.
  - Moderate numbers of cannot-link constraints
  - Initial coloring: cannot-link constraints $\rightarrow$ assign to the first cluster available
  - Problem for few classes: constrained data pairs are forced to stay in place

# Conclusion

# » Conclusion

💡 Unsupervised clustering via Markov aggregation $\longrightarrow$ accept pairwise constraints

# » Conclusion

- Unsupervised clustering via Markov aggregation $\longrightarrow$ accept pairwise constraints
  - Lowers sensitivity to hyperparameters

# » Conclusion

- Unsupervised clustering via Markov aggregation $\longrightarrow$ accept pairwise constraints
  - Lowers sensitivity to hyperparameters
  - Learns non-linear decision boundaries

# » Conclusion

- Unsupervised clustering via Markov aggregation $\longrightarrow$ accept pairwise constraints
  - Lowers sensitivity to hyperparameters
  - Learns non-linear decision boundaries
  - Competes with state-of-the-art semi-supervised clustering techniques

# » Conclusion

- Unsupervised clustering via Markov aggregation $\longrightarrow$ accept pairwise constraints
  - Lowers sensitivity to hyperparameters
  - Learns non-linear decision boundaries
  - Competes with state-of-the-art semi-supervised clustering techniques

Full-text: `https://arxiv.org/abs/2112.09397`

# » Conclusion

- Unsupervised clustering via Markov aggregation $\longrightarrow$ accept pairwise constraints
  - Lowers sensitivity to hyperparameters
  - Learns non-linear decision boundaries
  - Competes with state-of-the-art semi-supervised clustering techniques

Full-text: `https://arxiv.org/abs/2112.09397`

</> `https://github.com/stegsoph/Constrained-Markov-Clustering.git`

Thank you for your attention!