Recurrent Bayesian Classifier Chains for Exact Multi-Label Classification

NeurIPS 2021

Walter Gerych*, Tom Hartvigsen, Luke Buquicchio, Emmanuel Agu, Elke Rundensteiner

Worcester Polytechnic Institute

Worcester, MA

{wgerych, twhartvigsen, ljbuiquicchio, emmanuel, rundenst}@wpi.edu

Multi-Label Data Is Common

Computer Vision

Predictive Medicine

Text Mining

X =

$\mathbb{C}$ =

$\left\{ \begin{array}{c} \text{Person} \\ \text{Laptop} \\ \text{Soda can} \end{array} \right\}$

$\left\{ \begin{array}{c} \text{Hypertension} \\ \text{Arrhythmia} \end{array} \right\}$

$\left\{ \begin{array}{c} \text{Editorial} \\ \text{Science} \\ \text{Health} \end{array} \right\}$

Multi-Label Classification

$x, c_1, c_2, \ldots, c_L \sim (X, C_1, C_2, \ldots, C_L)$

such that $c_i = 1$ if class i applies to x, and $c_i = 0$ otherwise

Multi-Label Classification

$x, c_1, c_2, \ldots, c_L \sim (X, C_1, C_2, \ldots, C_L)$

such that $c_i = 1$ if class i applies to x, and $c_i = 0$ otherwise

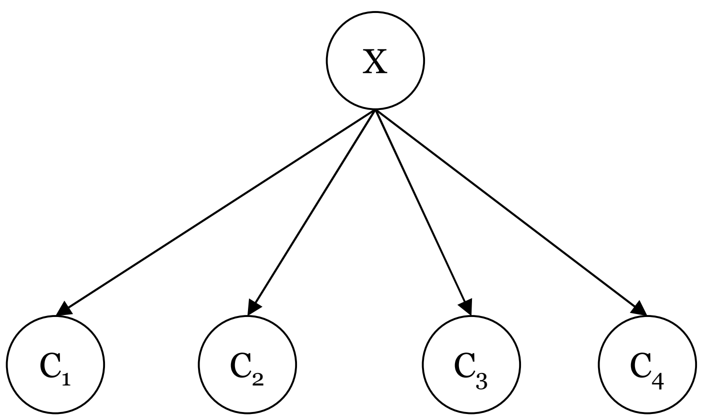
Goal:

Construct $f(x) = c_1, c_2, \ldots, c_L$

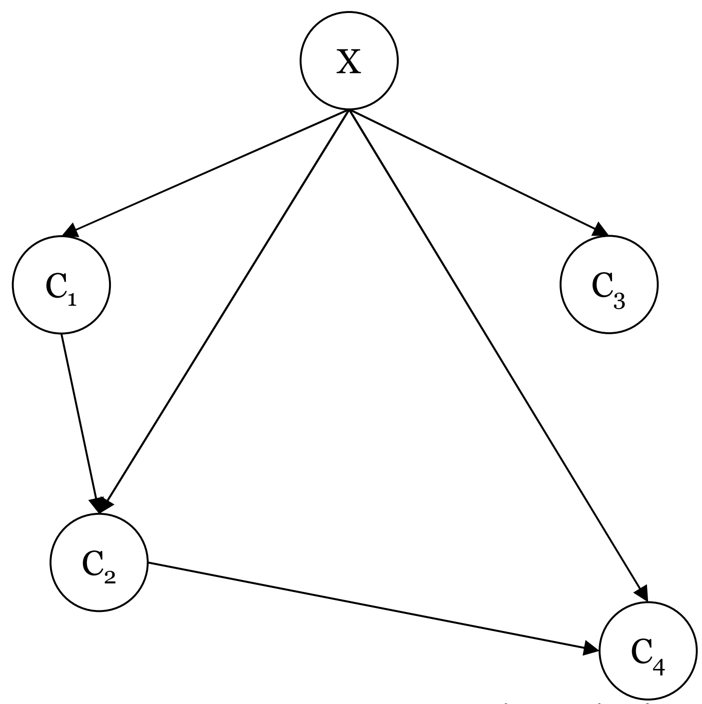
Background

Exploiting Label Relationships

Binary Approach

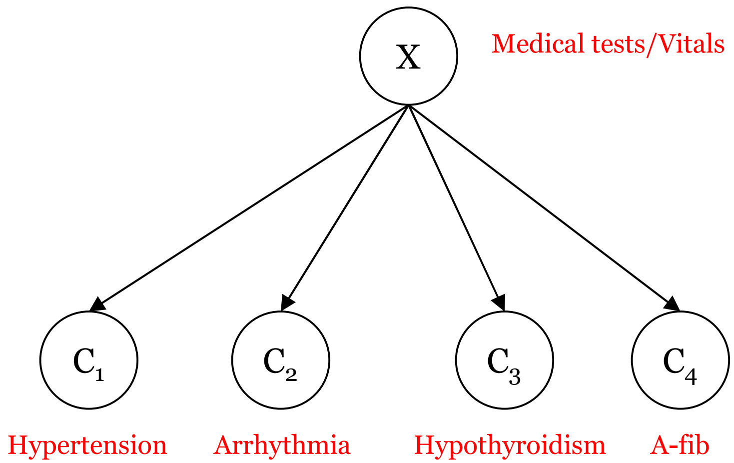


Modeling Label Dependencies

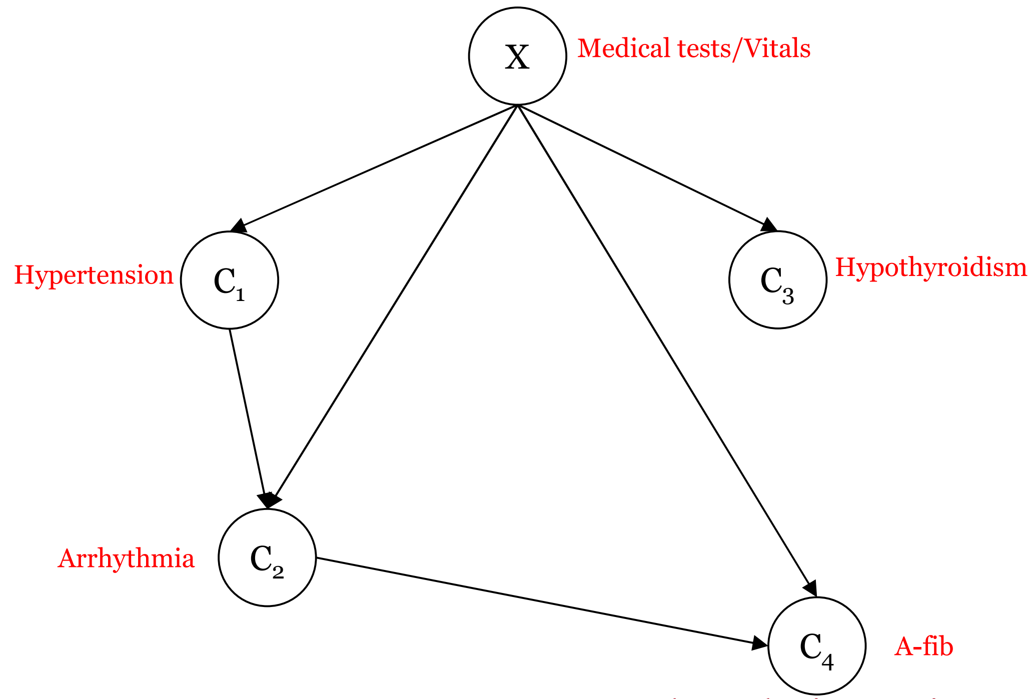


Exploiting Label Relationships

Binary Approach

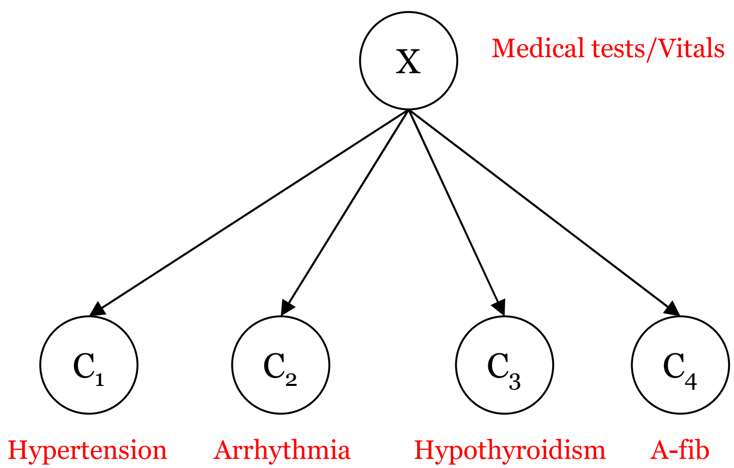


Modeling Label Dependencies

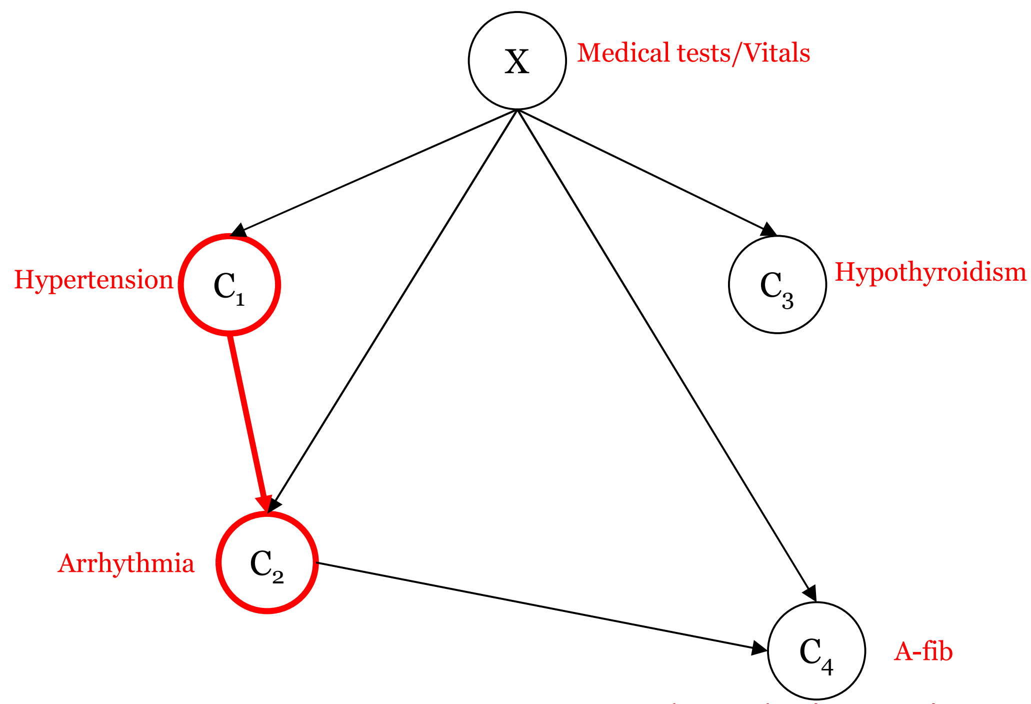


Exploiting Label Relationships

Binary Approach



Modeling Label Dependencies



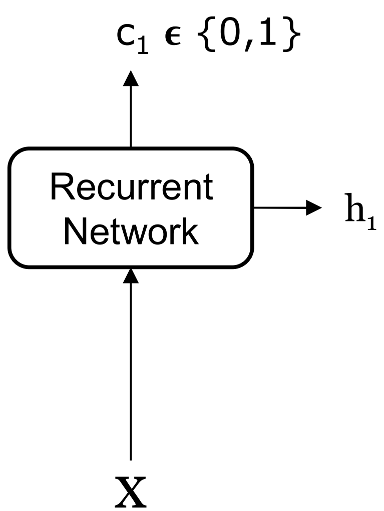
Leading Approach: Recurrent Classifier Chains

$$P(C_1, C_2, \ldots, C_L | X) = P(C_1 | X) \prod_{i=2}^{L} P(C_i | C_{<i}, X)$$

Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.

Leading Approach: RCC

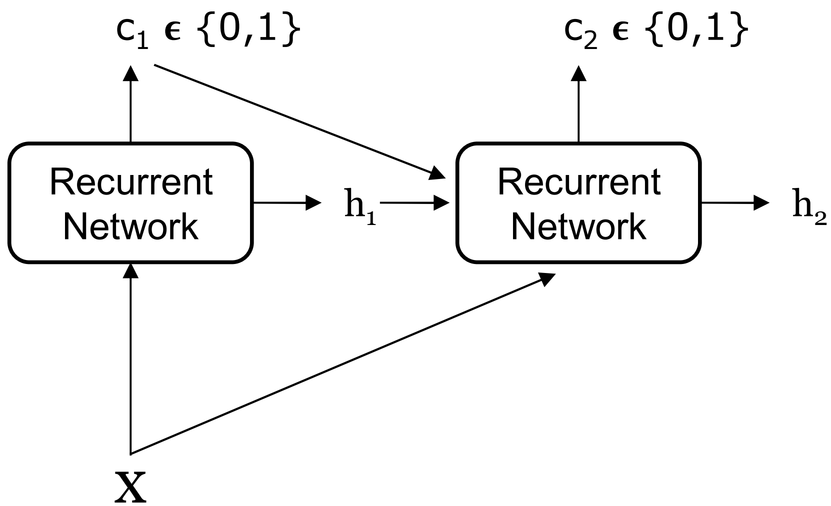
$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$



Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.

Leading Approach: RCC

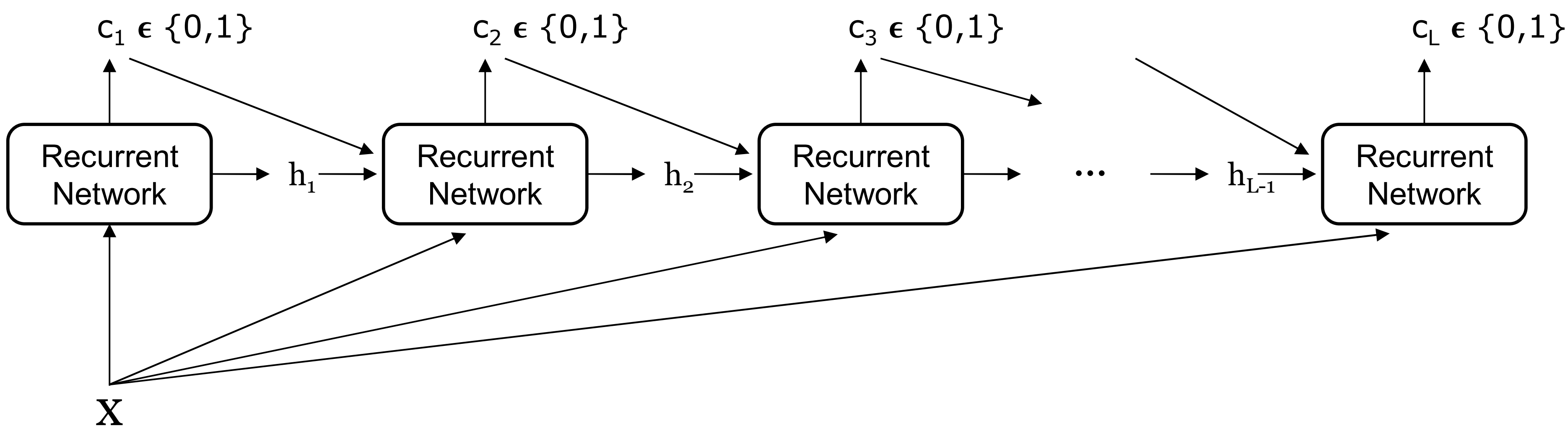
$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$



Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.

Leading Approach: RCC

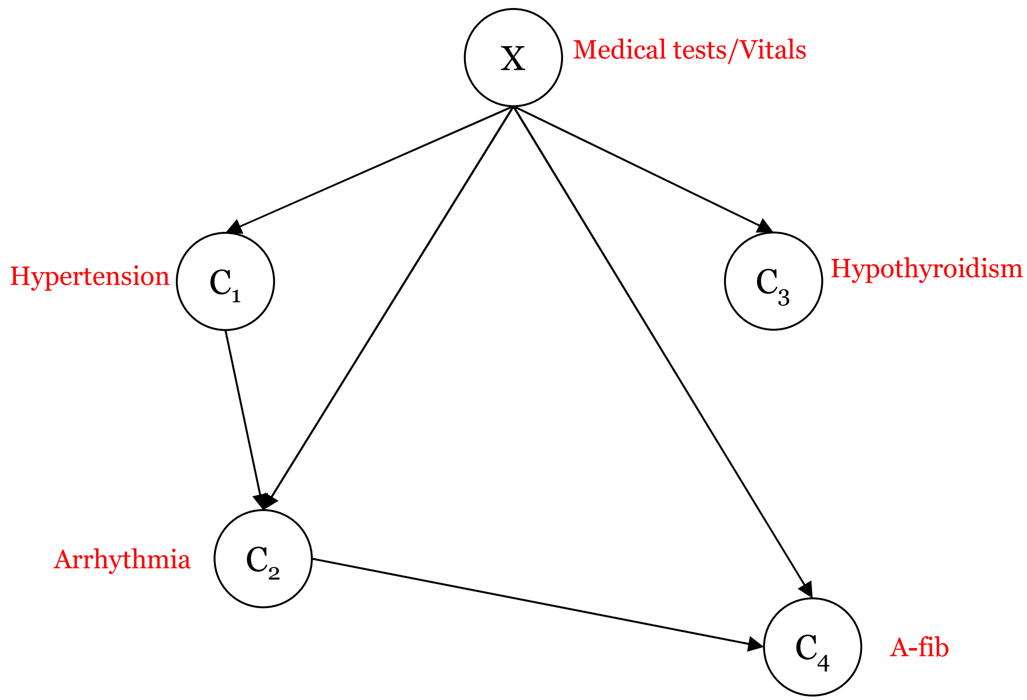
$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$



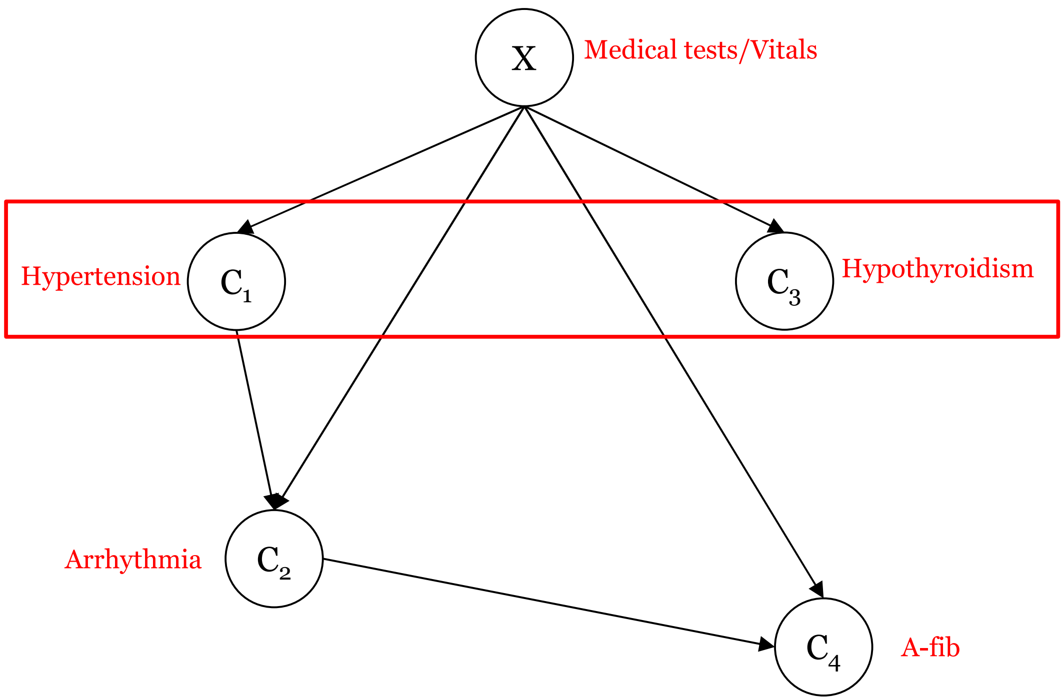
Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.

Limitations of RCCs

Limitation 1: Noisy Conditioning

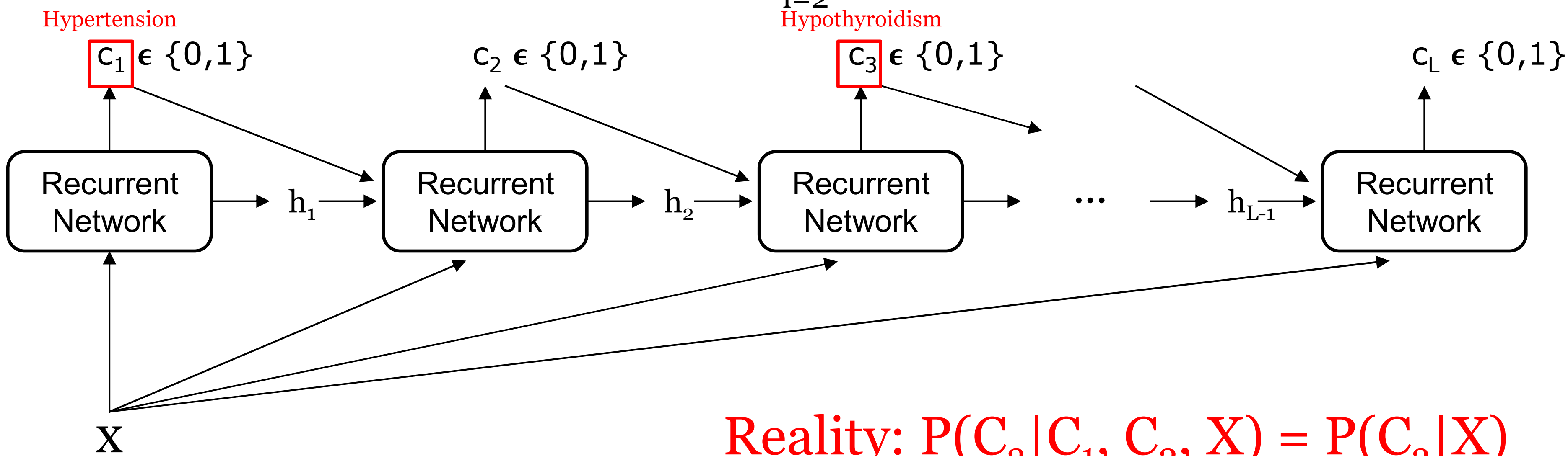


Limitation 1: Noisy Conditioning



Limitation 1: Noisy Conditioning

$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$

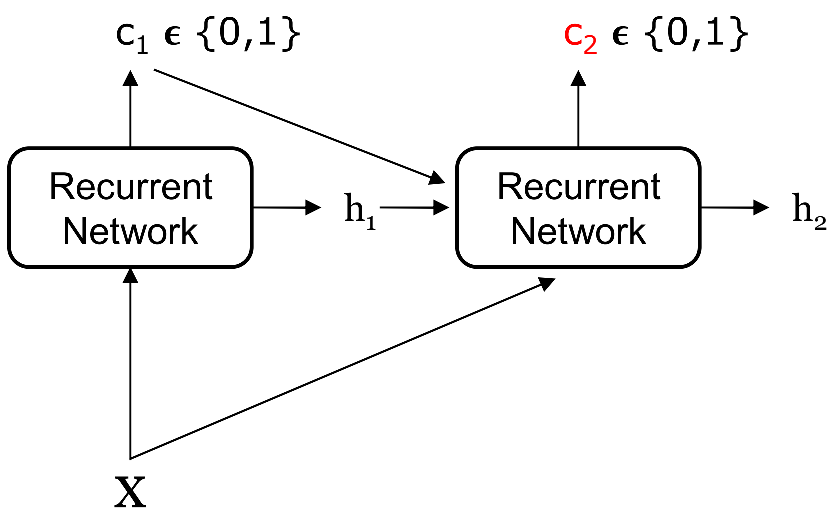


Reality: $P(C_3|C_1, C_2, X) = P(C_3|X)$

RCC model: $f(C_3|C_1, C_2, X) \neq f(C_3|X)$

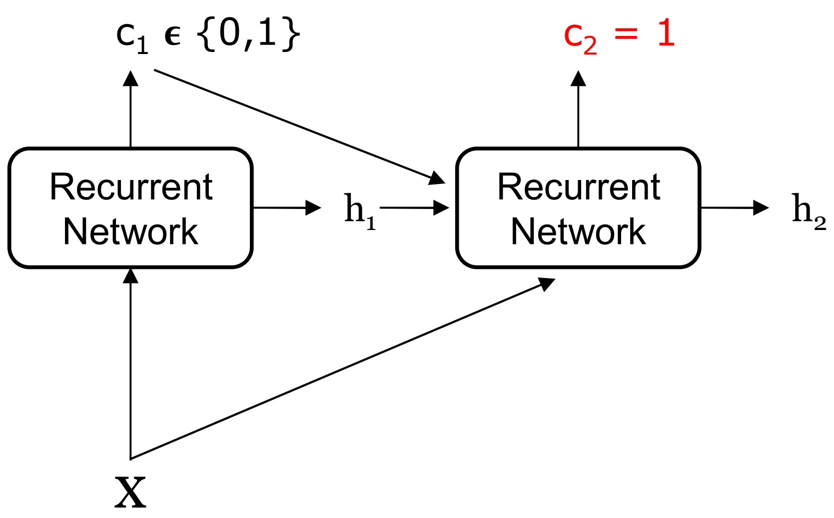
Limitation 2: Error Propagation

$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$



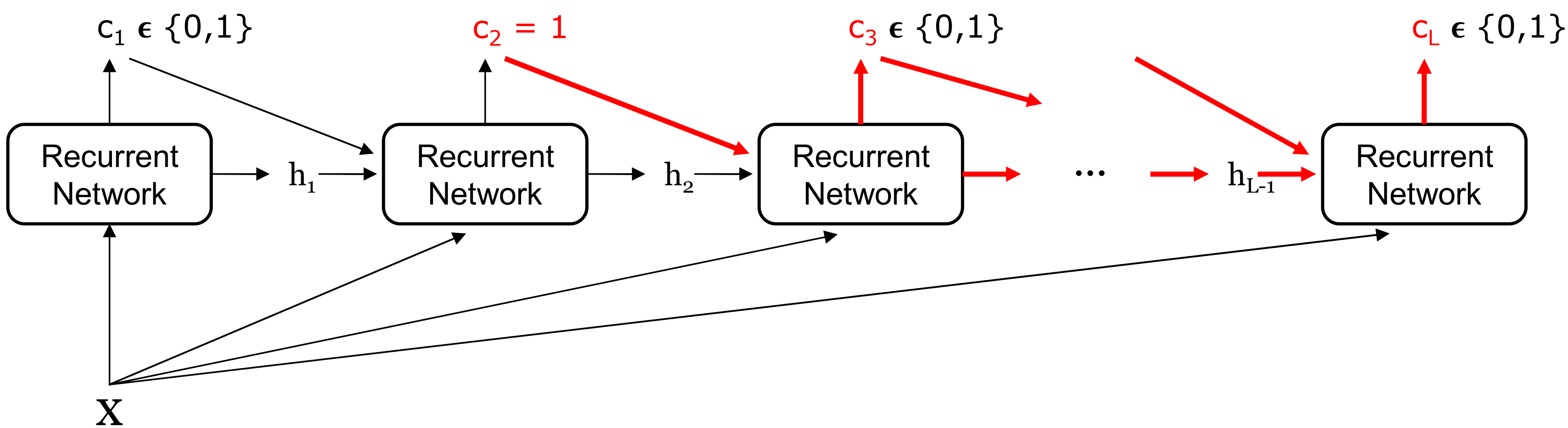
Limitation 2: Error Propagation

$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$



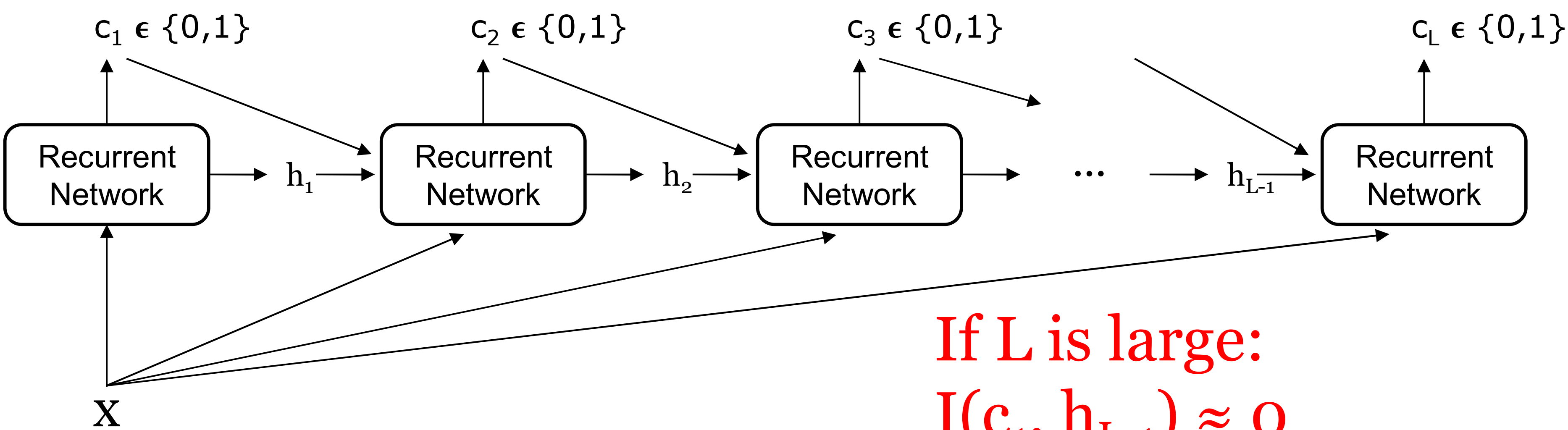
Limitation 2: Error Propagation

$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$



Limitation 3: Large Label Sets

$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$



If L is large:
$I(c_1, h_{L-1}) \approx 0$

Our Approach:
Recurrent Bayesian Classifier Chains

Overview of Recurrent Bayesian Classifier Chains

RBCC key components:

1. Infer Bayesian network of label dependencies
2. Modify RCC architecture to only use parent classes (defined by Bayesian network) for inference

Overview of Recurrent Bayesian Classifier Chains

RBCC key components:

1. Infer Bayesian network of label dependencies

2. Modify RCC architecture to only use parent classes (defined by Bayesian network) for inference

Tackles challenges by:

- Eliminating noisy conditioning

Overview of Recurrent Bayesian Classifier Chains

RBCC key components:

1. Infer Bayesian network of label dependencies
2. Modify RCC architecture to only use parent classes (defined by Bayesian network) for inference

Tackles challenges by:

- Eliminating noisy conditioning
- Minimizing error propagation

Overview of Recurrent Bayesian Classifier Chains

RBCC key components:

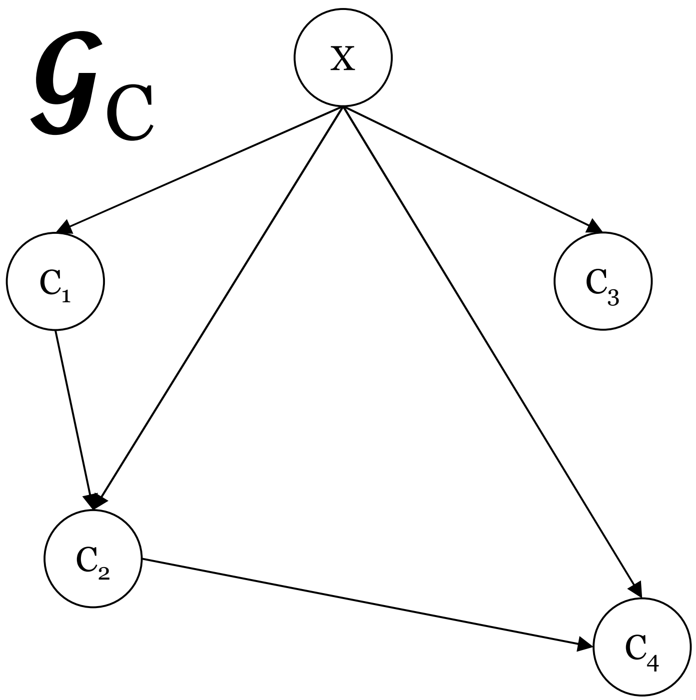
1. Infer Bayesian network of label dependencies

2. Modify RCC architecture to only use parent classes (defined by Bayesian network) for inference

Tackles challenges by:

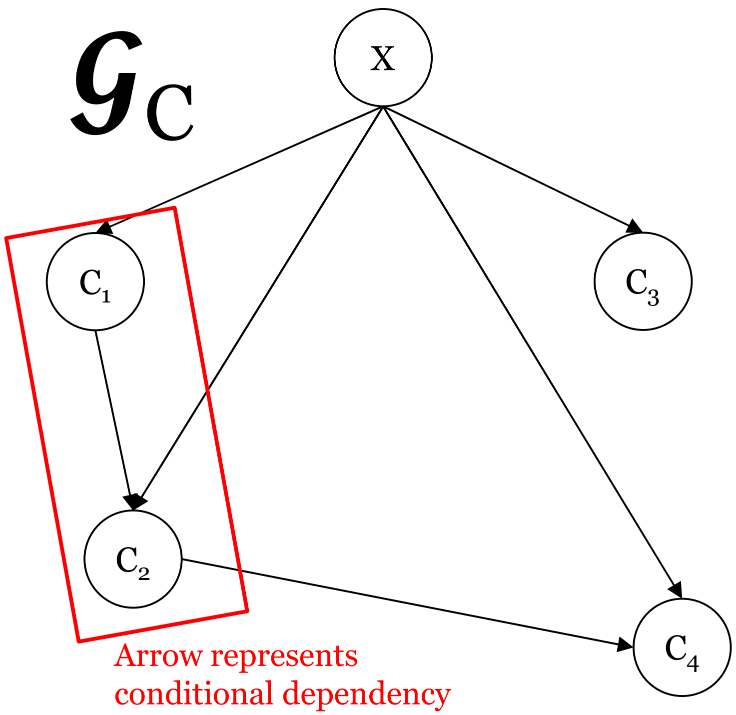
- Eliminating noisy conditioning
- Minimizing error propagation
- Removing need for long-term memory

RBCC Step 1: Label Dependency Graph

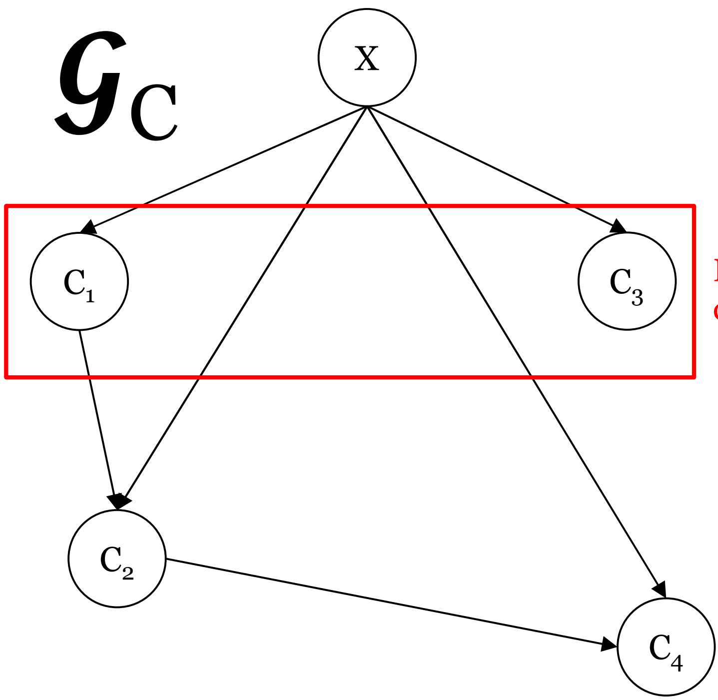


RBCC Step 1: Label Dependency Graph

$\mathcal{G}_C$

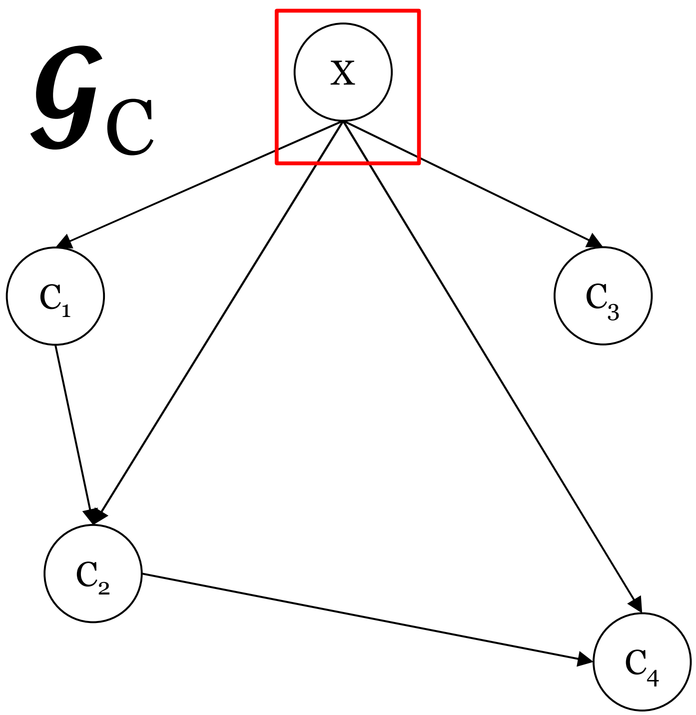


RBCC Step 1: Label Dependency Graph



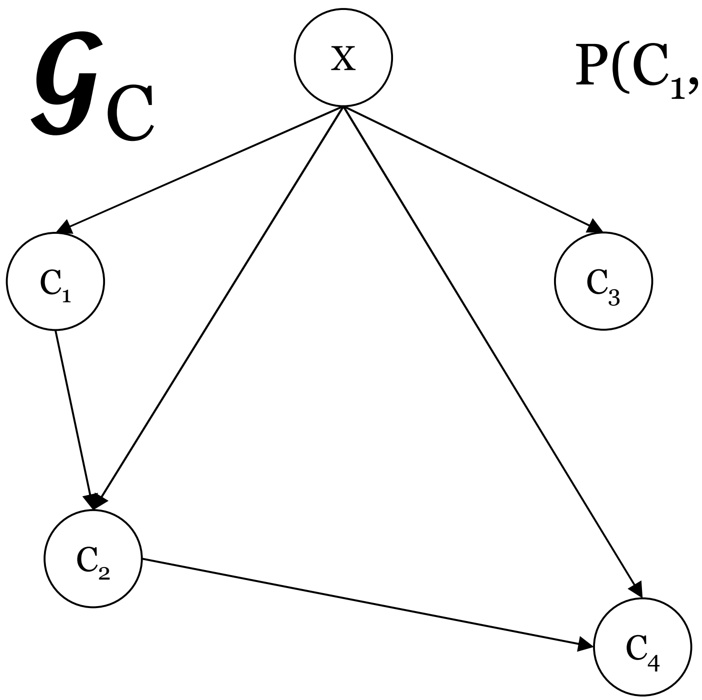
Lack of arrow indicates
conditional independence

RBCC Step 1: Label Dependency Graph



RBCC Step 1: Label Dependency Graph

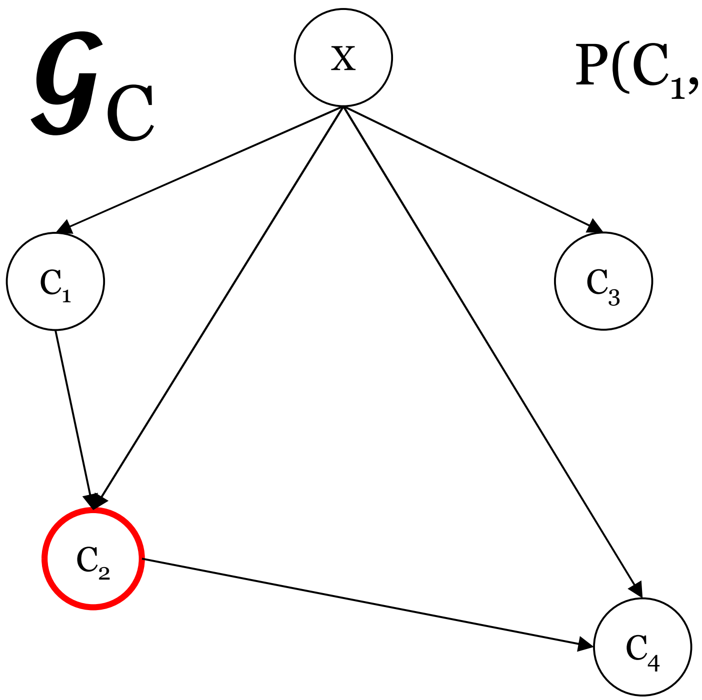
$\mathcal{G}_C$



$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$

$$= \boxed{P(C_1|X)\prod_{i=2}^{L} P(C_i|Pa_{\mathcal{G}C}(C_i))}$$

RBCC Step 1: Label Dependency Graph



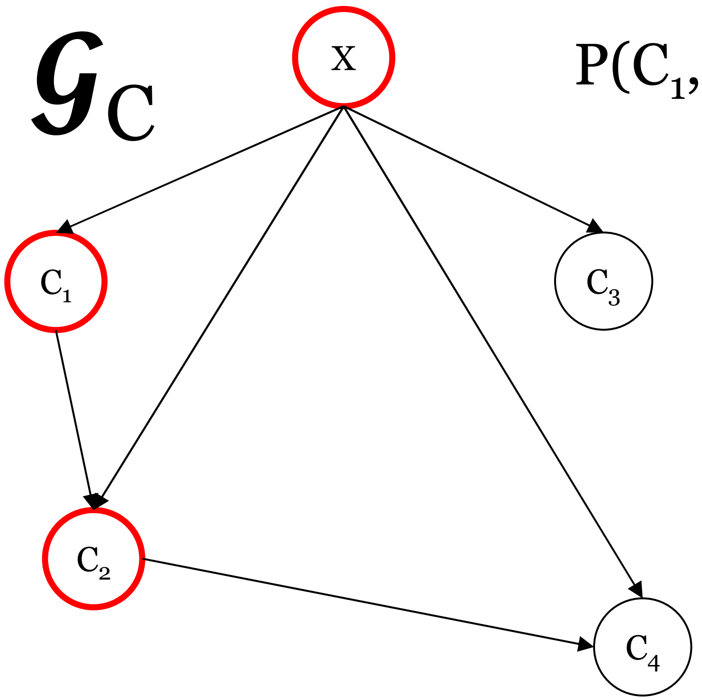
$$P(C_1, C_2, \dots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$

$$= P(C_1|X)\prod_{i=2}^{L} P(C_i|Pa_{\mathcal{G}C}(C_i))$$

ex: $Pa_{\mathcal{G}C}(C_2) = C_1, X$

RBCC Step 1: Label Dependency Graph

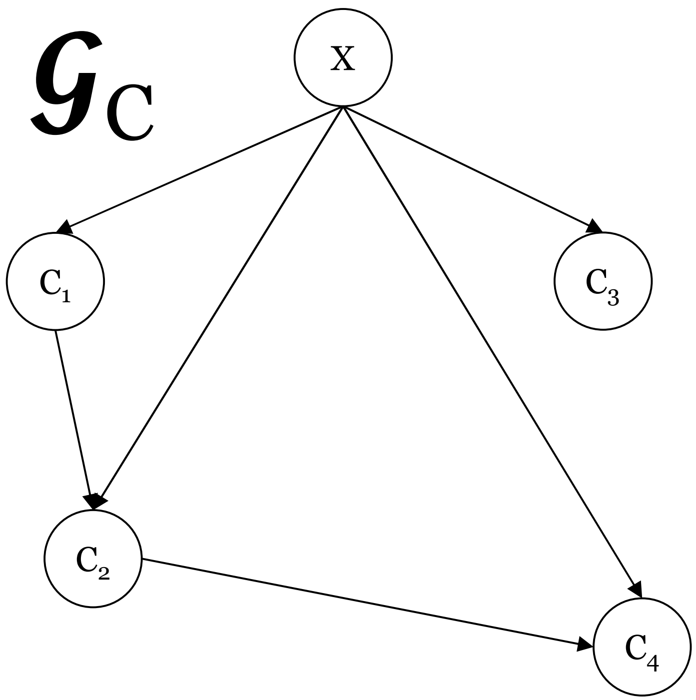
$\mathcal{G}_C$



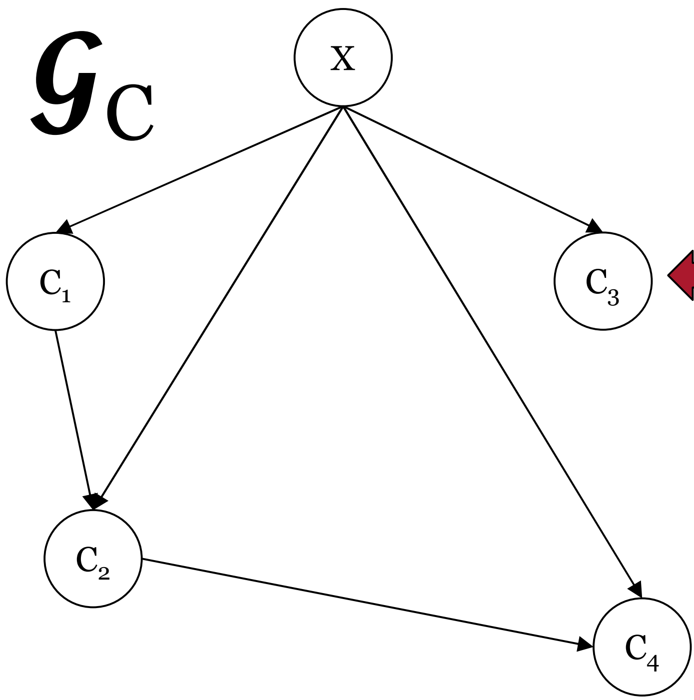
$$P(C_1, C_2, \ldots, C_L|X) = P(C_1|X)\prod_{i=2}^{L} P(C_i|C_{<i},X)$$
$$= P(C_1|X)\prod_{i=2}^{L} P(C_i|Pa_{\mathcal{G}C}(C_i))$$

ex: $Pa_{\mathcal{G}C}(C_2) = C_1, X$

RBCC Step 1: Label Dependency Graph

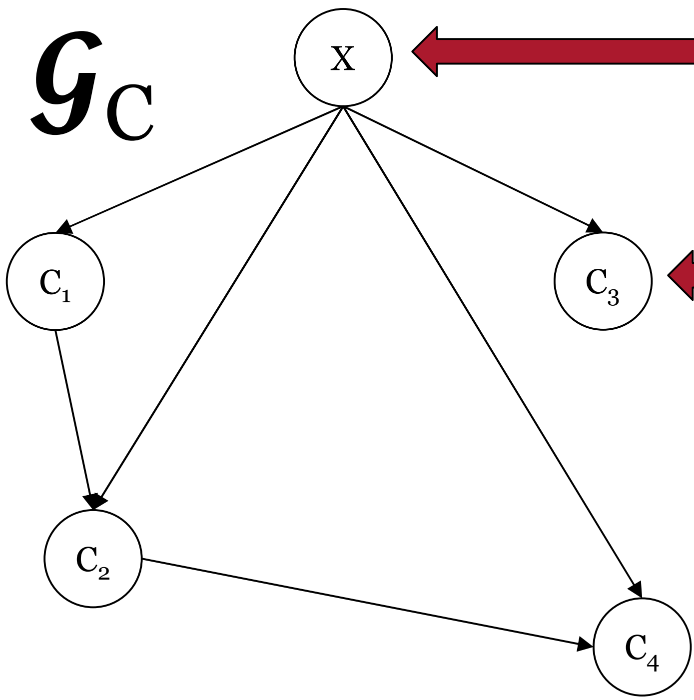


RBCC Step 1: Label Dependency Graph



Binary, univariate

RBCC Step 1: Label Dependency Graph

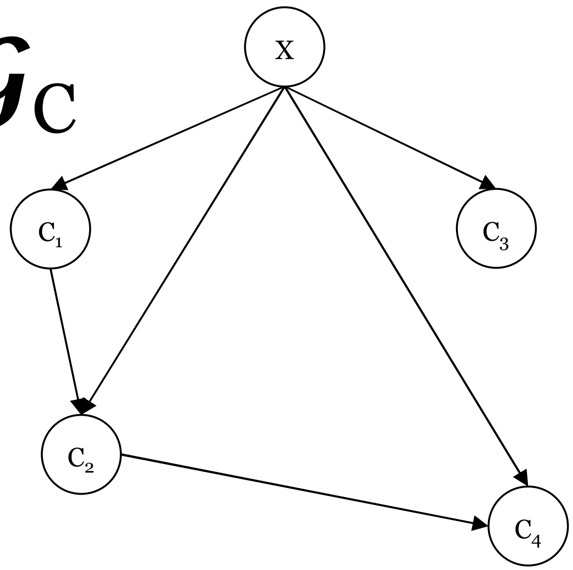


Continuous, multivariate

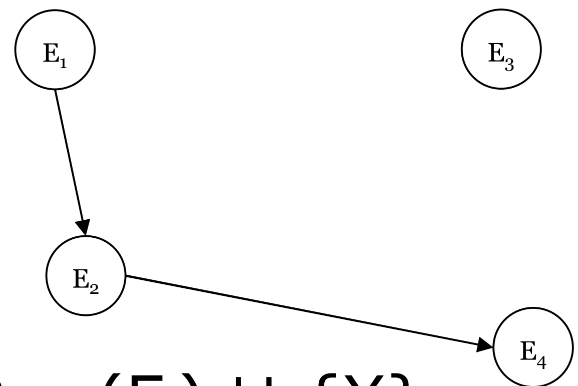
Binary, univariate

RBCC Step 1: Label Dependency Graph

$\mathcal{G}_C$



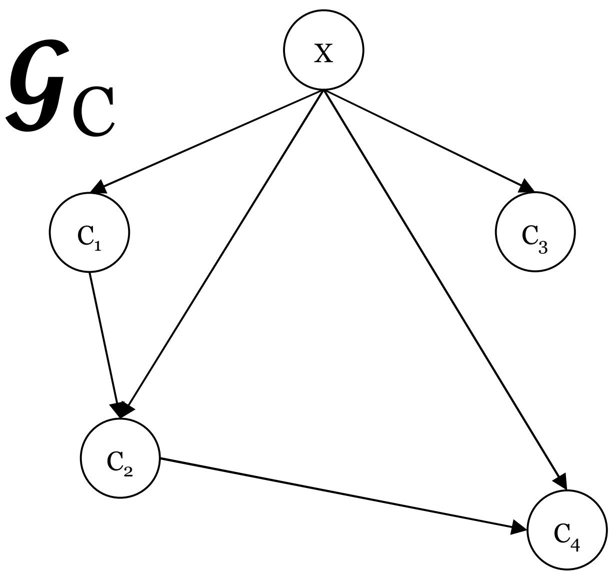
$\mathcal{G}_E$

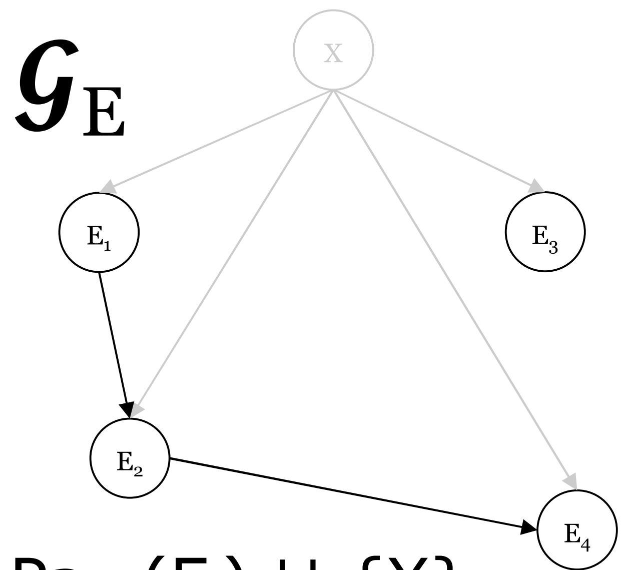


$$Pa_{\mathcal{G}C}(C_i) = Pa_{\mathcal{G}E}(E_i) \cup \{X\}$$

Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.

RBCC Step 1: Label Dependency Graph

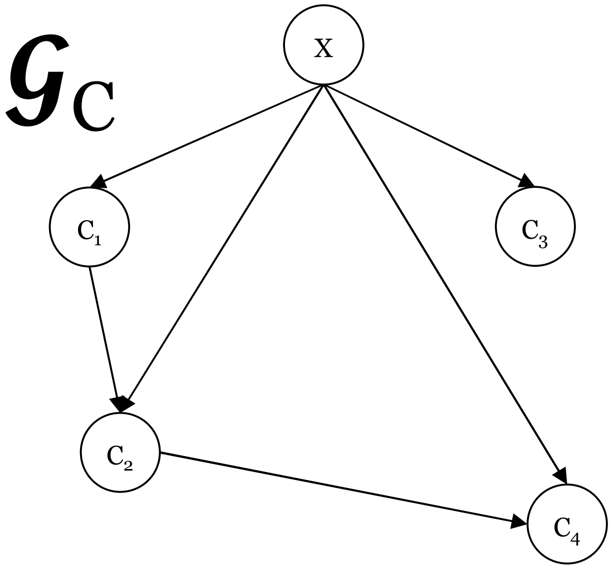


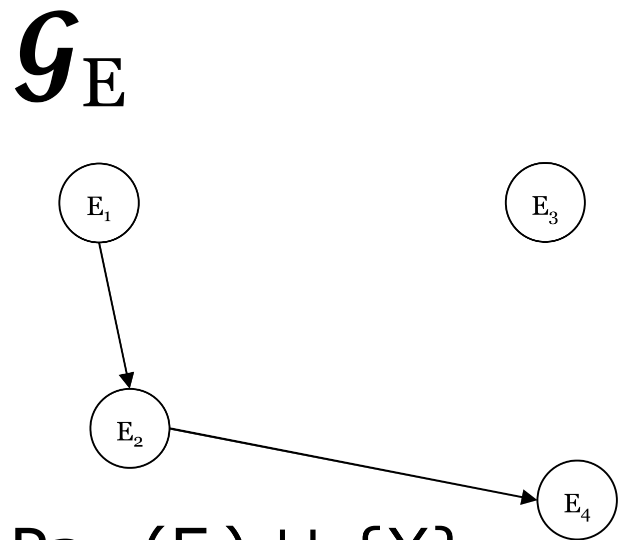


$$Pa_{\mathcal{G}C}(C_i) = Pa_{\mathcal{G}E}(E_i) \cup \{X\}$$

Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.

RBCC Step 1: Label Dependency Graph

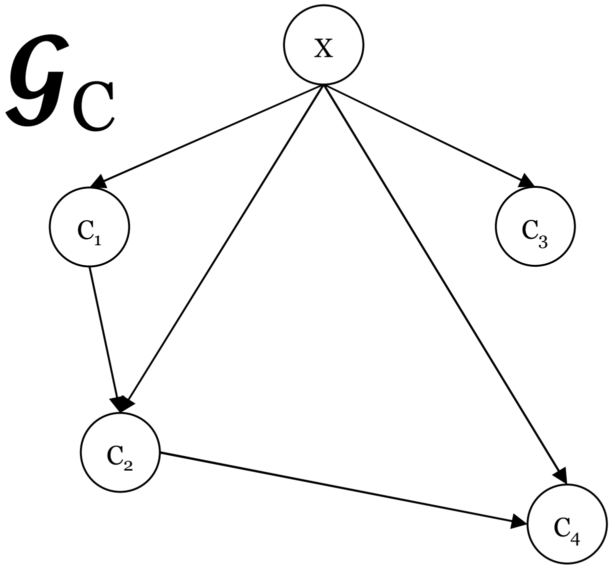


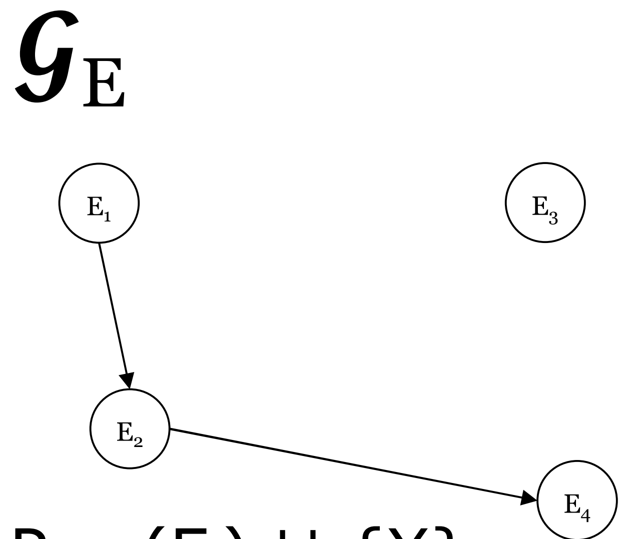


$$Pa_{\mathcal{G}C}(C_i) = Pa_{\mathcal{G}E}(E_i) \cup \{X\}$$

Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.

RBCC Step 1: Label Dependency Graph

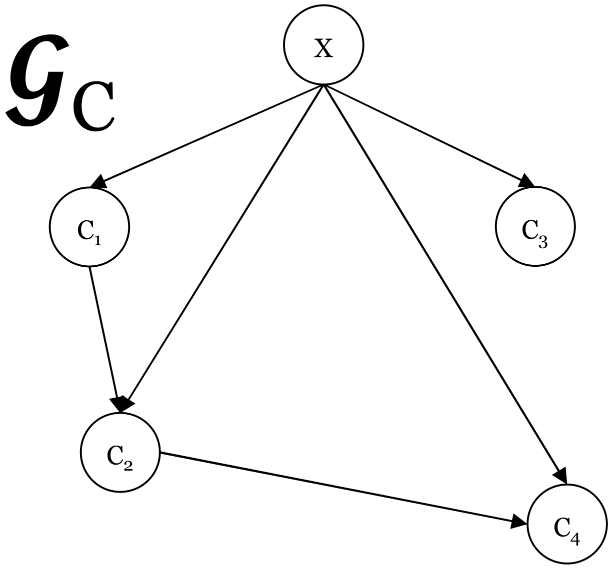


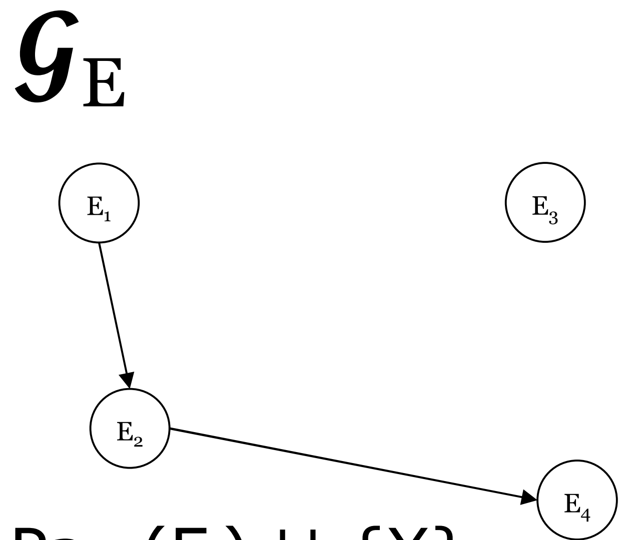


$$Pa_{\mathcal{G}C}(C_i) = Pa_{\mathcal{G}E}(E_i) \cup \{X\}$$

$$Pa_{\mathcal{G}C}(C_2) = \{C_1, X\}$$

RBCC Step 1: Label Dependency Graph





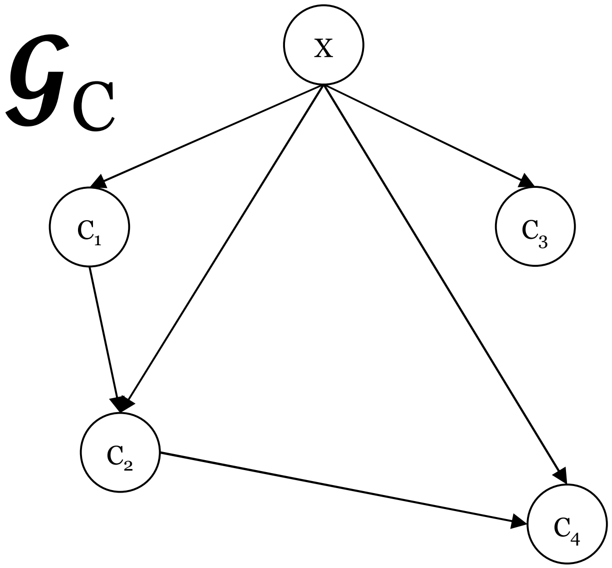
$$Pa_{\mathcal{G}C}(C_i) = Pa_{\mathcal{G}E}(E_i) \cup \{X\}$$

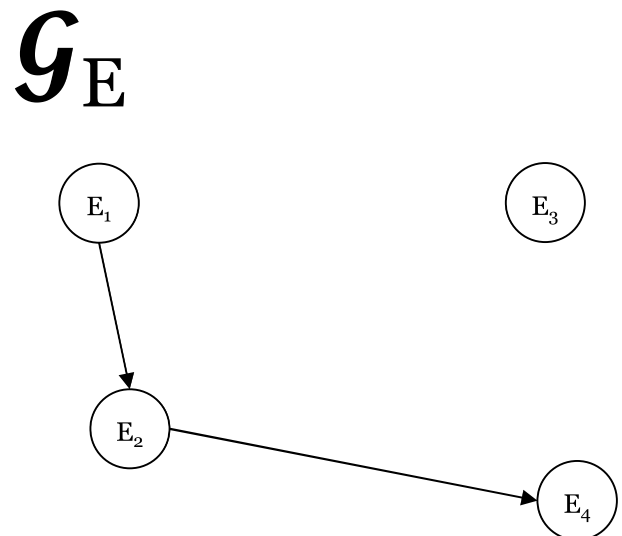
$$Pa_{\mathcal{G}C}(C_2) = \{C_1, X\}$$

$$Pa_{\mathcal{G}E}(E_2) = \{E_1\}$$

Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.

RBCC Step 1: Label Dependency Graph



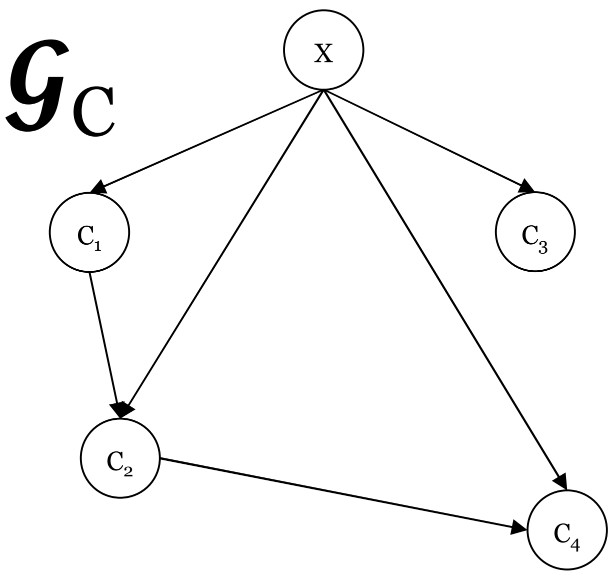


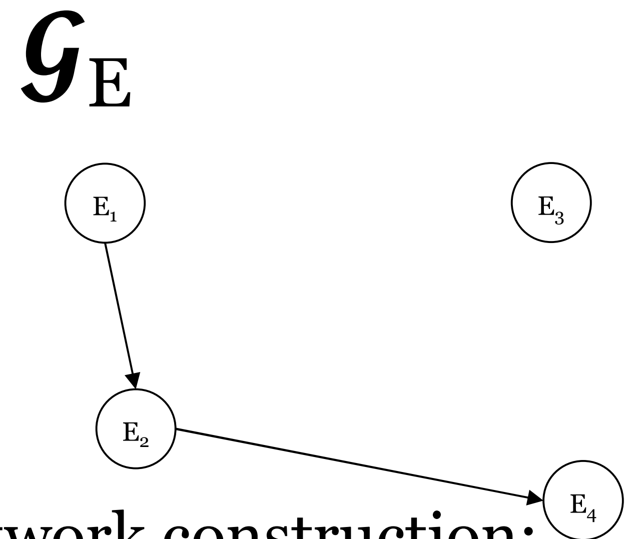
$$C_i = k_i(X) + E_i \Rightarrow E_i = C_i - f(x)$$

Where k_i is found by maximizing data likelihood

Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.

RBCC Step 1: Label Dependency Graph





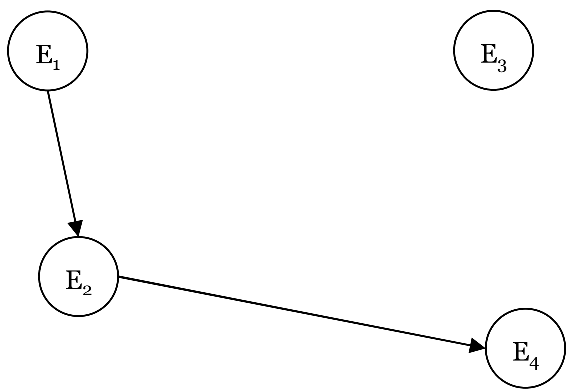
Network construction:

- Hill climbing [1]
- Constraint based [2]
- Chow Liu algorithm [3]

[1] Daly , Rónán, et al. “Methods to accelerate the learning of bayesian network structures.” UKCI 2007.
[2] Verma , Thomasand, et al. “Equivalence and synthesis of causal models.” 1991.
[3] Chow, C., et al. “Approximating discrete probability distributions with dependence trees.”. IEEE Transactions on Information Theory 1968.

RBCC Step 1: Label Dependency Graph

$\mathcal{G}_{\mathrm{E}}$



$$P(C_1, C_2, \ldots, C_L | X) = P(C_1|X)\prod_{i=2}^{L} P(C_i | C_{<i}, X)$$

$$= P(C_1|X)\prod_{i=2}^{L} P(C_i | Pa_{\mathcal{G}C}(C_i))$$

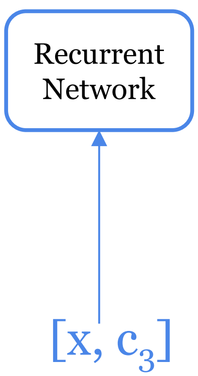
$$= P(C_1|X)\prod_{i=2}^{L} P(C_i | Pa_{\mathcal{G}E}(E_i), X)$$

RBCC Step 2: Model Training

$$\left\{\begin{array}{l} C_1: [C_3, C_5, X] \\ C_2: [C_3, C_8, C_{10}\ X] \\ C_3: [X] \\ \vdots \\ C_L: [C_1, C_5, X] \end{array}\right\}$$

RBCC Step 2: Model Training

$$\left\{ \begin{array}{l} C_1: [C_3, C_5, X] \\ C_2: [C_3, C_8, C_{10}\ X] \\ C_3: [X] \\ \vdots \\ C_L: [C_1, C_5, X] \end{array} \right\}$$



RBCC Step 2: Model Training

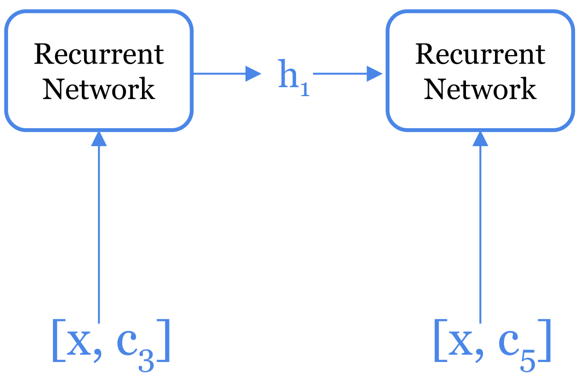
C_1: [C_3, C_5, X]

C_2: [C_3, C_8, C_{10} X]

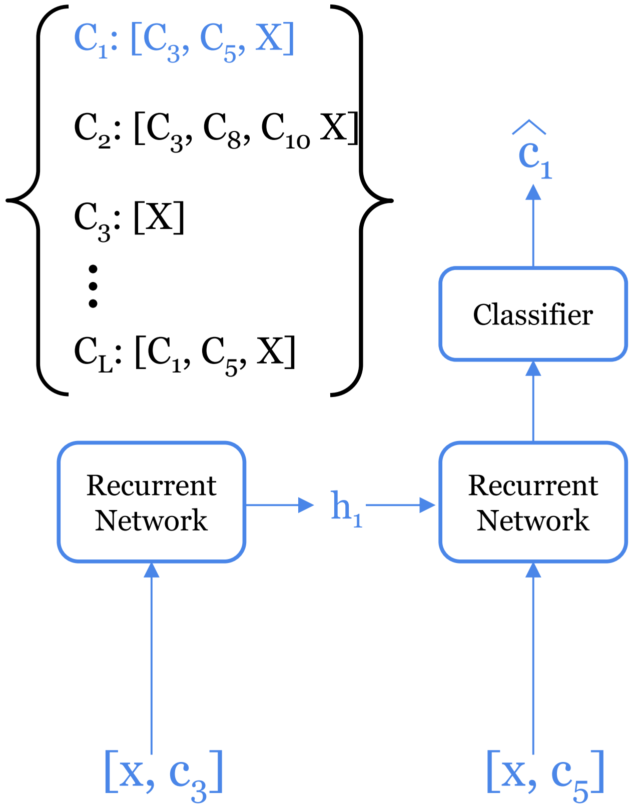
C_3: [X]

⋮

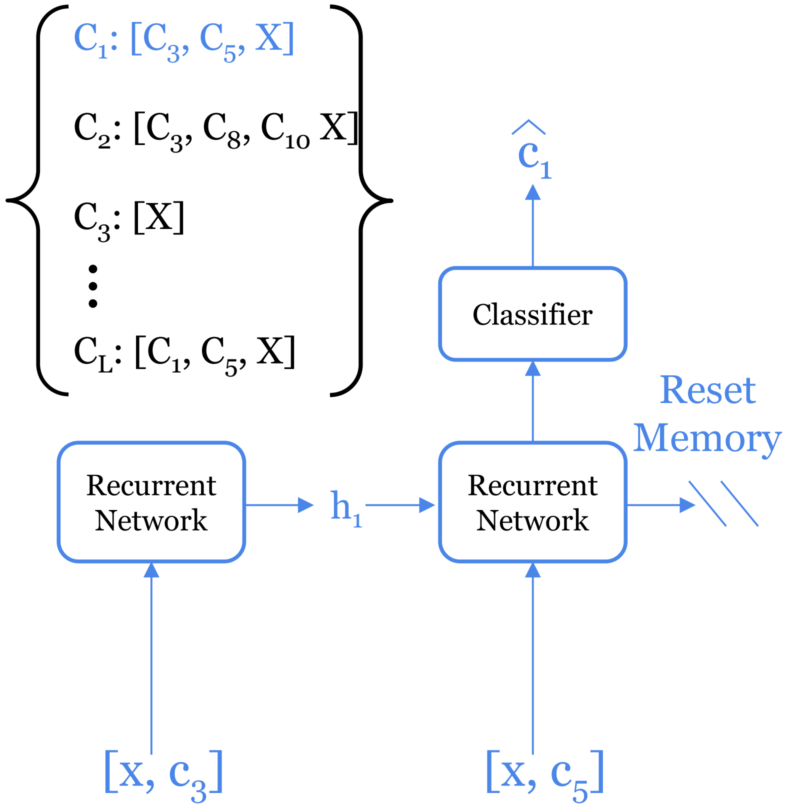
C_L: [C_1, C_5, X]



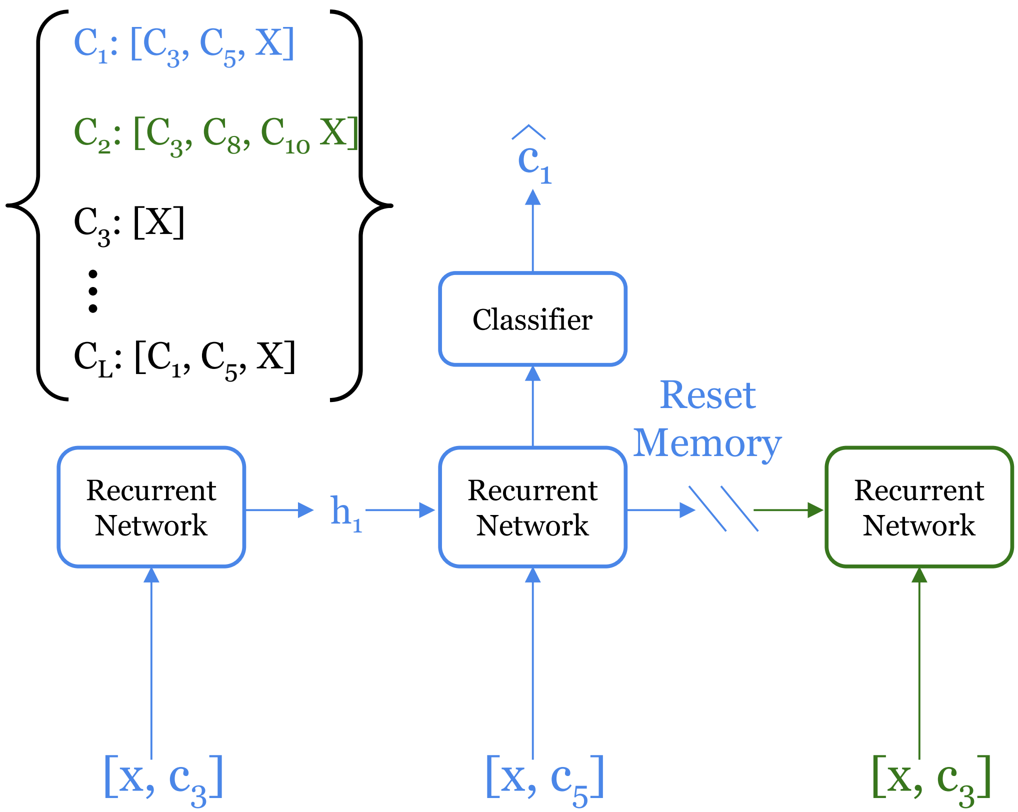
RBCC Step 2: Model Training



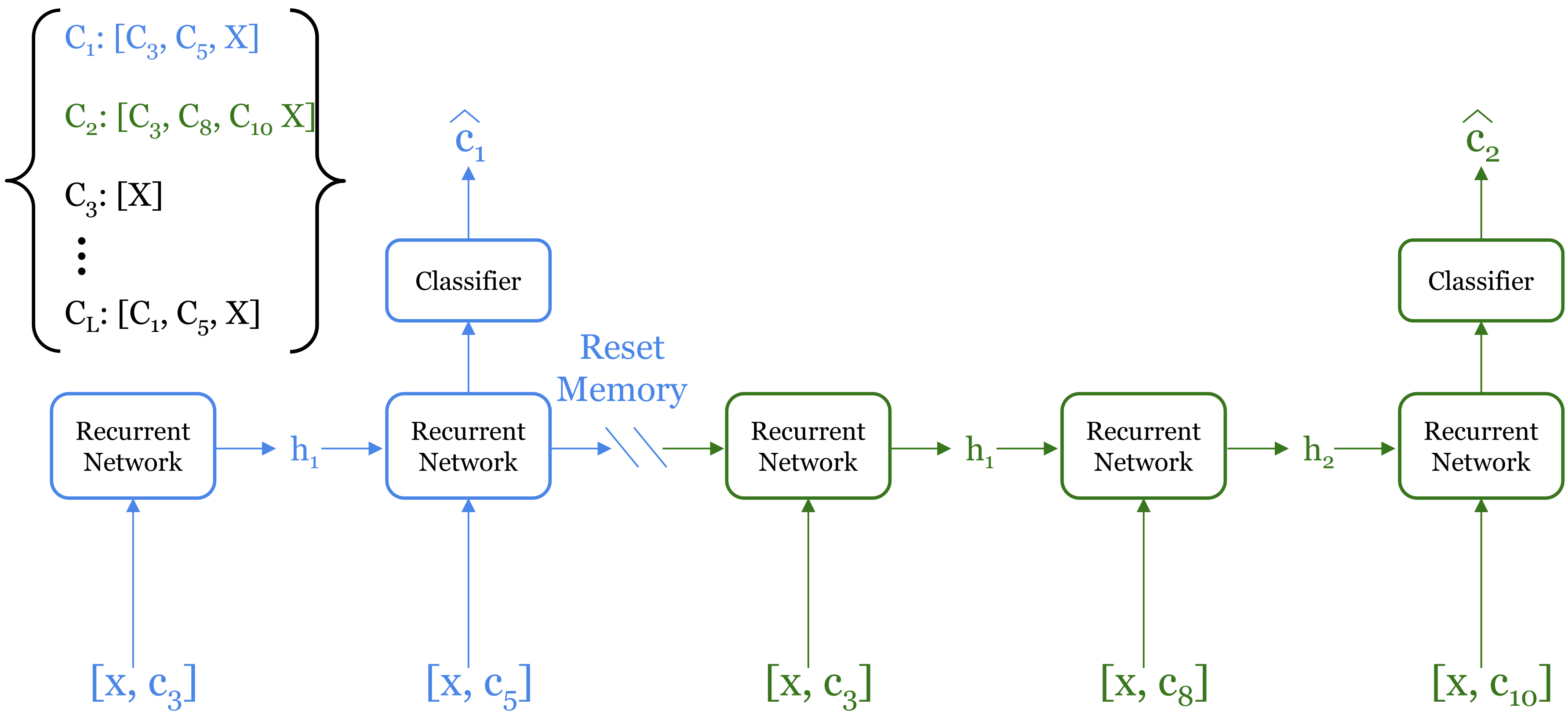
RBCC Step 2: Model Training



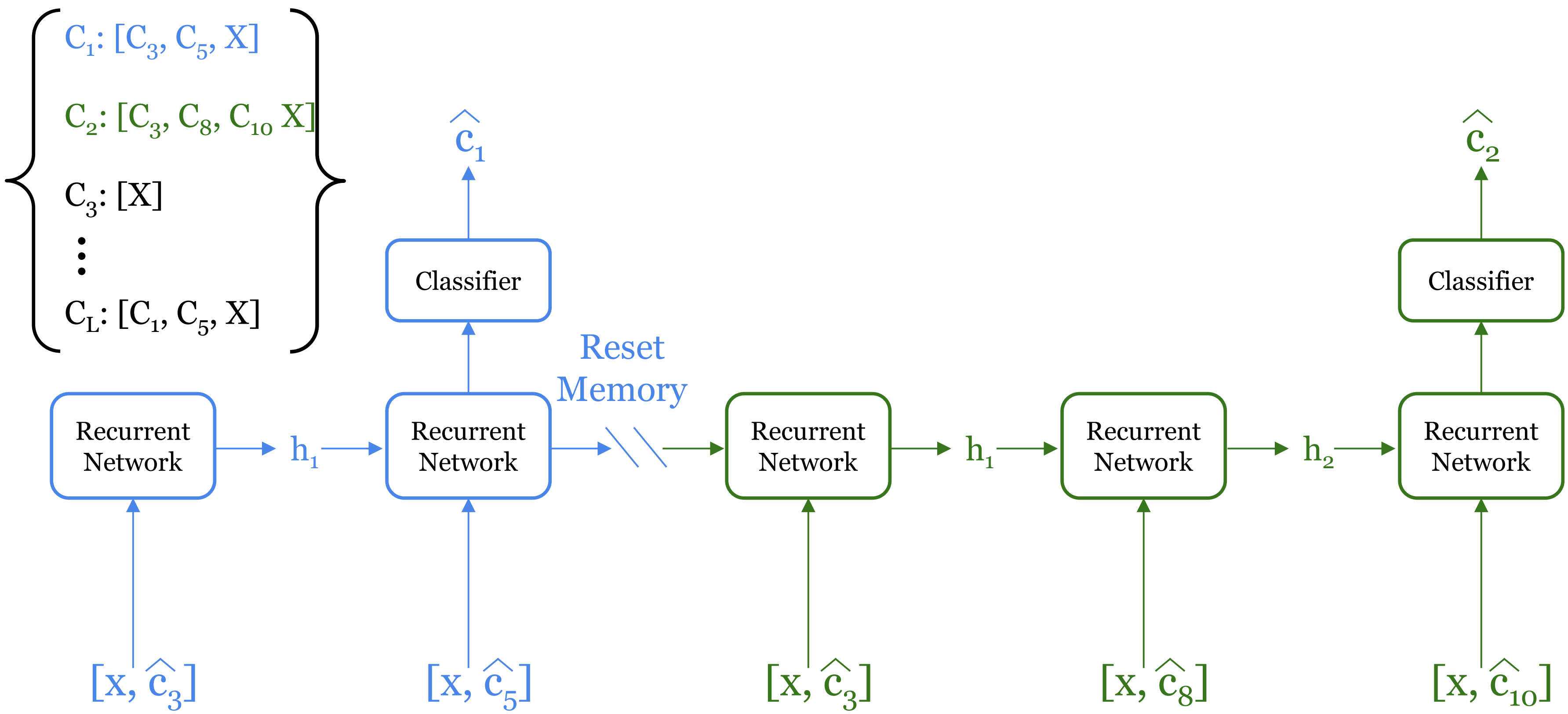
RBCC Step 2: Model Training



RBCC Step 2: Model Training



RBCC Step 3: Inference

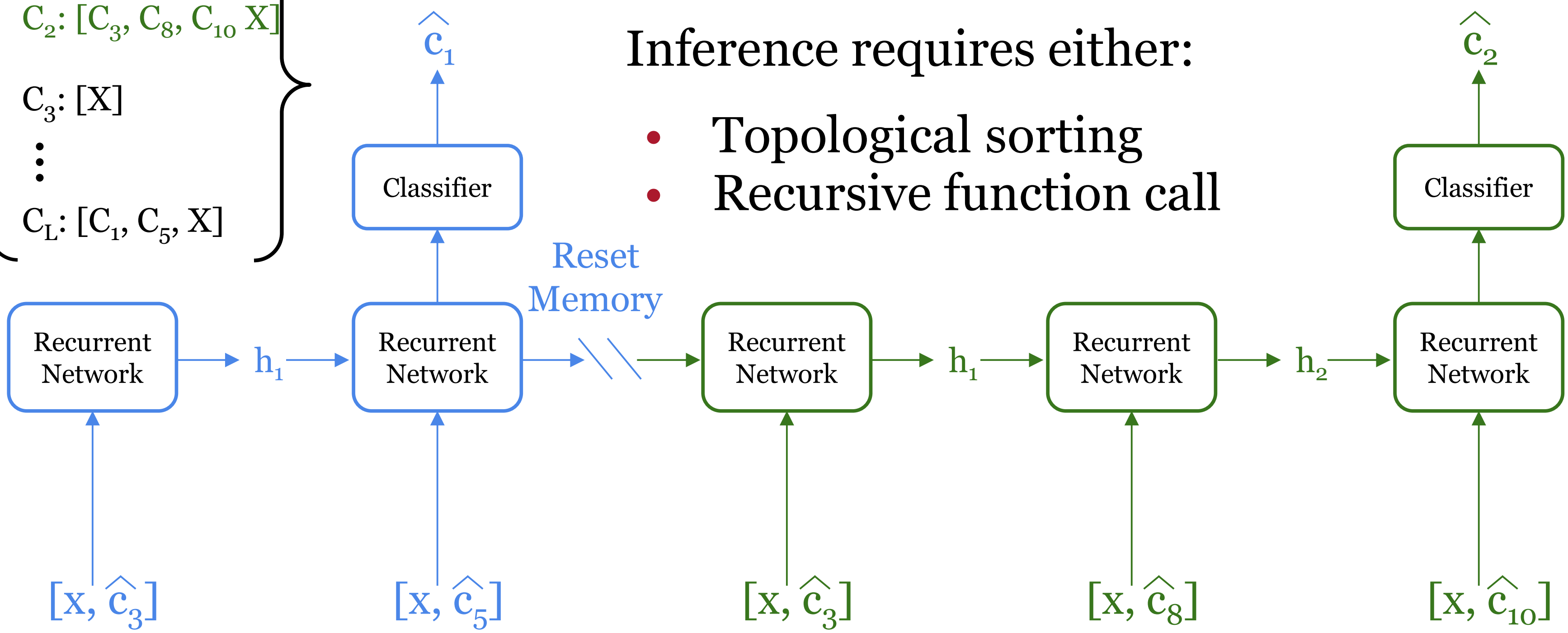


RBCC Step 3: Inference

$$\left\{ \begin{array}{l} C_1: [C_3, C_5, X] \\ C_2: [C_3, C_8, C_{10}\ X] \\ C_3: [X] \\ \vdots \\ C_L: [C_1, C_5, X] \end{array} \right\}$$

Inference requires either:

- Topological sorting
- Recursive function call



Evaluation

Compared Methods

- Recurrent Classier Chains (RCC) [1]
- Topological-Sort RCC (TS-RCC) [1]
- Order-Free RCC (OF-RCC) [2]
- Bayesian Classifier Chains (BCC) [3]
- Binary Decomposition (BD) [4]

[1] Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.
[2] Shang-Fu Chen, et al. “Order-free RNN with visual attention for multi-label classification.” AAAI 2018.
[3] Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.
[4] Tsoumakas , Grigorios Tsoumakas aet al. “Multi label classification: An overview.” IJDWM 2007.

Compared Methods

- Recurrent Classier Chains (RCC) [1]
- Topological-Sort RCC (TS-RCC) [1]
- Order-Free RCC (OF-RCC) [2]
- Bayesian Classifier Chains (BCC) [3]
- Binary Decomposition (BD) [4]

[1] Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.
[2] Shang-Fu Chen, et al. "Order-free RNN with visual attention for multi-label classification." AAAI 2018.
[3] Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.
[4] Tsoumakas , Grigorios Tsoumakas aet al. "Multi label classification: An overview." IJDWM 2007.

Compared Methods

- Recurrent Classier Chains (RCC) [1]
- Topological-Sort RCC (TS-RCC) [1]
- Order-Free RCC (OF-RCC) [2]
- Bayesian Classifier Chains (BCC) [3]
- Binary Decomposition (BD) [4]

[1] Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.
[2] Shang-Fu Chen, et al. "Order-free RNN with visual attention for multi-label classification." AAAI 2018.
[3] Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.
[4] Tsoumakas , Grigorios Tsoumakas aet al. "Multi label classification: An overview." IJDWM 2007.

Compared Methods

- Recurrent Classier Chains (RCC) [1]
- Topological-Sort RCC (TS-RCC) [1]
- Order-Free RCC (OF-RCC) [2]
- Bayesian Classifier Chains (BCC) [3]
- Binary Decomposition (BD) [4]

[1] Nam, Jinseok, et al. "Maximizing subset accuracy with recurrent neural networks in multi-label classification." NeurIPS 2017.
[2] Shang-Fu Chen, et al. "Order-free RNN with visual attention for multi-label classification." AAAI 2018.
[3] Zhang, Min-Ling, et al. "Multi-label learning by exploiting label dependency." KDD 2010.
[4] Tsoumakas , Grigorios Tsoumakas aet al. "Multi label classification: An overview." IJDWM 2007.

Datasets

We compare on 6 benchmark multi-label datasets:

- PASCAL VOC 2007
- Scene
- Yeast
- Enron
- EukaryoteGO
- Yeast

M. Everingham, et al. “The “PASCAL Visual Object Classes Challenge” 2007
Boutell, Matthew, et al. “Learning multi-label scene classification.” Pattern Recognition 2004.
Sajnani, Hitesh et al. “Classifying yelp reviews into relevant categories”. 2012.
Klimt, B., et. al. “The Enron Corpus: A New Dataset for Email Classification Research.” ECML 2004.
Xu, Jianhua et al. “A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously”. Knowledge-Based Systems 2016.
Elisseeff, A., et al. “A Kernel Method for Multi-Labelled Classification.” NeurIPS 2001.

Results

Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	0.212 ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	0.189 ± 0.000
Macro-F1 ↑	0.556 ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	0.662 ± 0.004	0.653 ± 0.003	0.638 ± 0.003

Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

Results

Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	<u>0.212</u> ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	<u>0.189</u> ± 0.000
Macro-F1 ↑	<u>0.556</u> ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	<u>0.662</u> ± 0.004	0.653 ± 0.003	0.638 ± 0.003

Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

Results

Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	<u>0.212</u> ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	<u>0.189</u> ± 0.000
Macro-F1 ↑	<u>0.556</u> ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	<u>0.662</u> ± 0.004	0.653 ± 0.003	0.638 ± 0.003

Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

Results

Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	<u>0.212</u> ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	<u>0.189</u> ± 0.000
Macro-F1 ↑	<u>0.556</u> ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	<u>0.662</u> ± 0.004	0.653 ± 0.003	0.638 ± 0.003

Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

Results

Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	0.212 ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	0.189 ± 0.000
Macro-F1 ↑	0.556 ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	0.662 ± 0.004	0.653 ± 0.003	0.638 ± 0.003

Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

Results

Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	$\boxed{0.212}$ ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	0.189 ± 0.000
Macro-F1 ↑	0.556 ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	0.662 ± 0.004	0.653 ± 0.003	0.638 ± 0.003

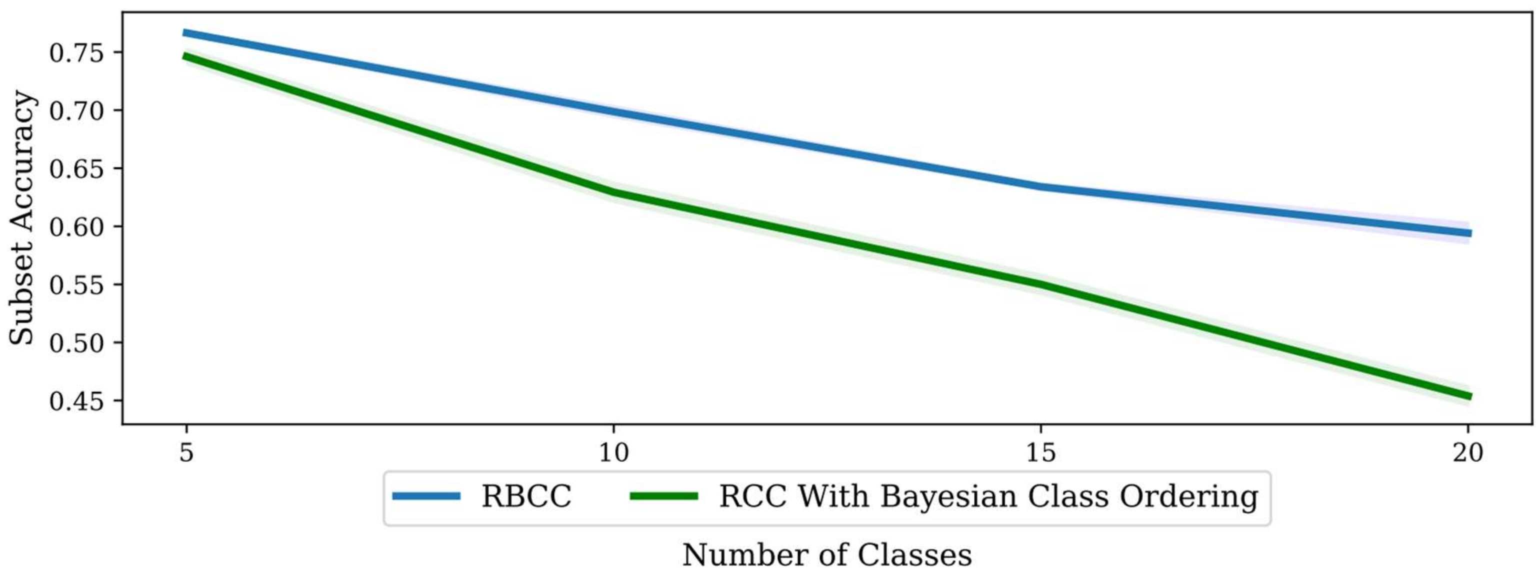
Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

Results

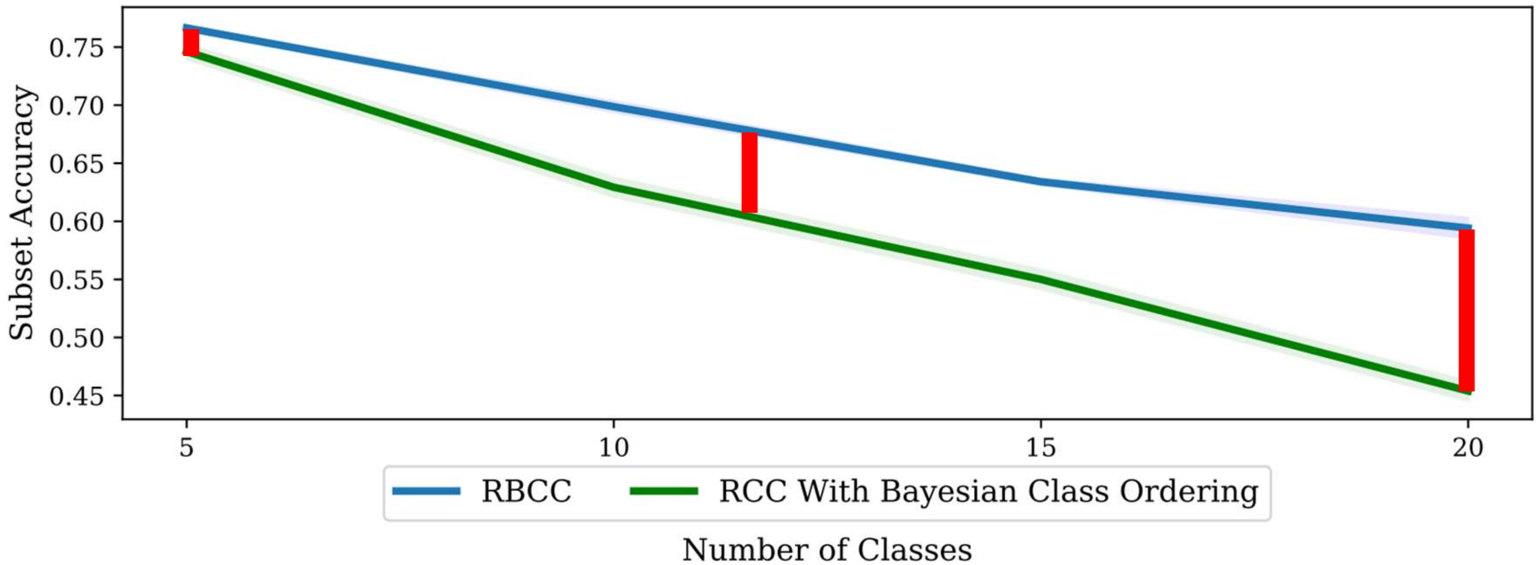
Evaluation Metrics	Methods					
	RBCC (Ours)	RCC	TS-RCC	OF-RCC	BCC	BD
Subset Accuracy ↑	**0.240** ± 0.008	<u>0.212</u> ± 0.002	0.192 ± 0.010	0.169 ± 0.009	0.210 ± 0.000	0.202 ± 0.002
Hamming Loss ↓	**0.186** ± 0.003	0.204 ± 0.001	0.209 ± 0.004	0.218 ± 0.004	0.199 ± 0.001	<u>0.189</u> ± 0.000
Macro-F1 ↑	<u>0.556</u> ± 0.008	0.526 ± 0.004	0.506 ± 0.004	**0.569** ± 0.004	0.551 ± 0.005	0.517 ± 0.008
Micro-F1 ↑	**0.670** ± 0.006	0.639 ± 0.002	0.628 ± 0.004	<u>0.662</u> ± 0.004	0.653 ± 0.003	0.638 ± 0.003

Table 2: Classification results for the Yelp dataset. Bolded is best performer, underlined is second best.

Performing Better on Large Label Sets



Performing Better on Large Label Sets



Conclusions

In this work we:

- Identified flaws with state-of-the-art multi-label approach (RCC)
- Proposed new multi-label approach that leverages label dependence and independence to improve RCC training and inference
- Performed experimental study illustrating the practical improvement of our approach

Acknowledgements

- WPI WASH Research group
- WPI DAISY Lab
- DARPA WASH Grant #FA8750-18-2-0077