# Inferring Concept Prerequisite Relations from Online Educational Resources

## PREREQ

Sudeshna Roy[1] Meghana Madhyastha[3] Sheril Lawrence[3] Vaibhav Rajan[2]

IAAI 2019

[1]VideoKen

[2]National University of Singapore

[3]IIIT-B, India (*authors were intern at VideoKen)

# Table of Contents

1. Introduction

2. PREREQ Algorithm

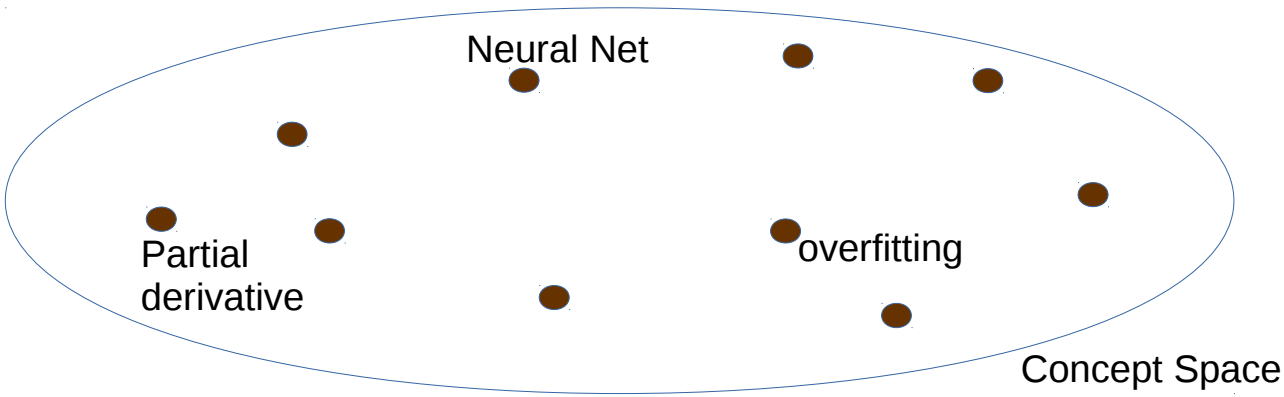
3. Experiments

4. Illustration

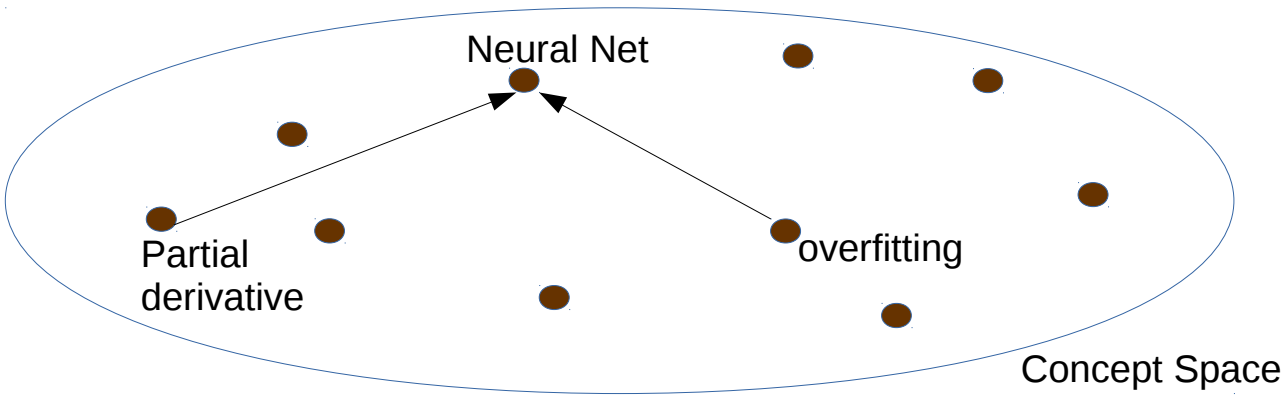
5. Conclusion

# Introduction

# Motivation



# Motivation

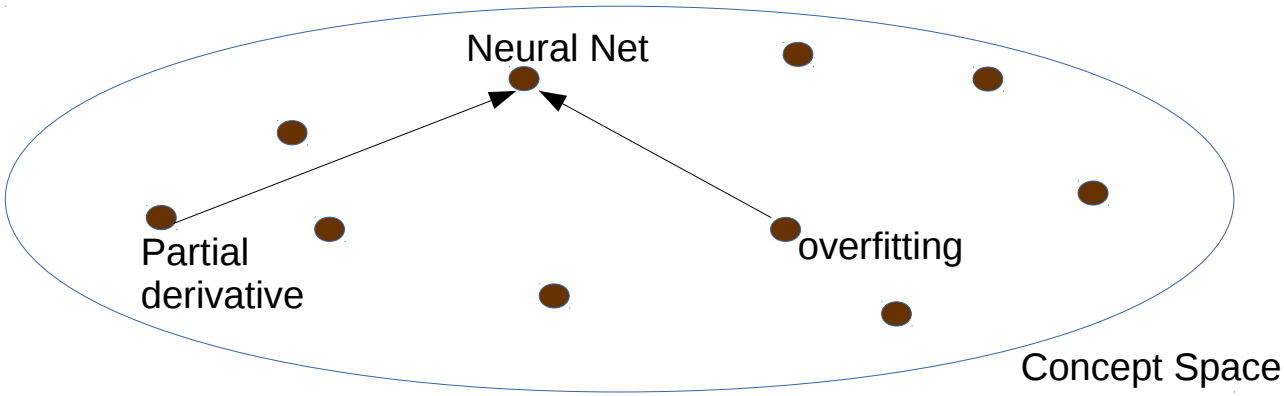


# Motivation

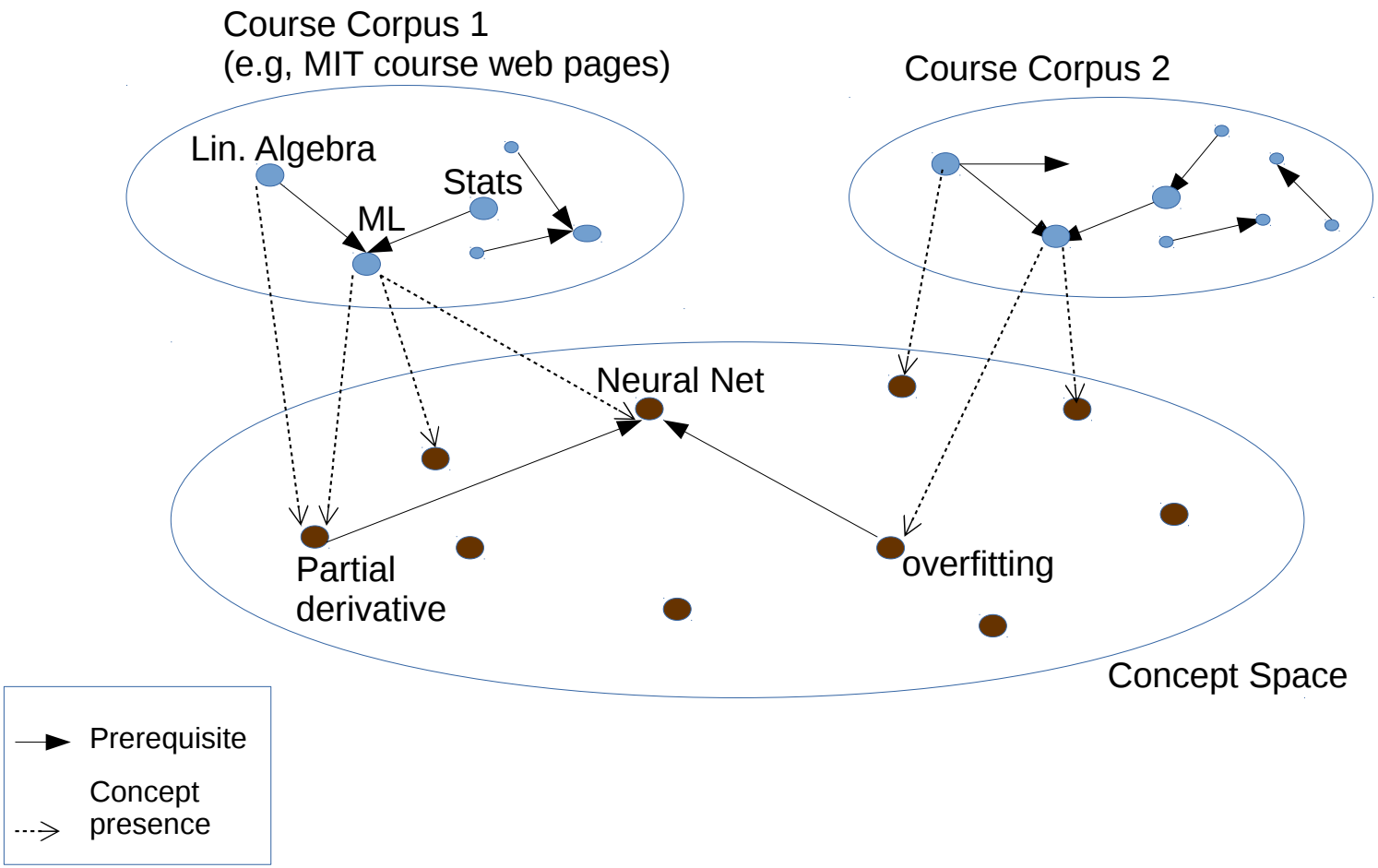
Concept Prerequisite plays a fundamental role in the following applications,

- Personalized online education
- Reading list generation
- Automatic curriculum planning
- Automatic evaluation of curriculum

# Problem Statement

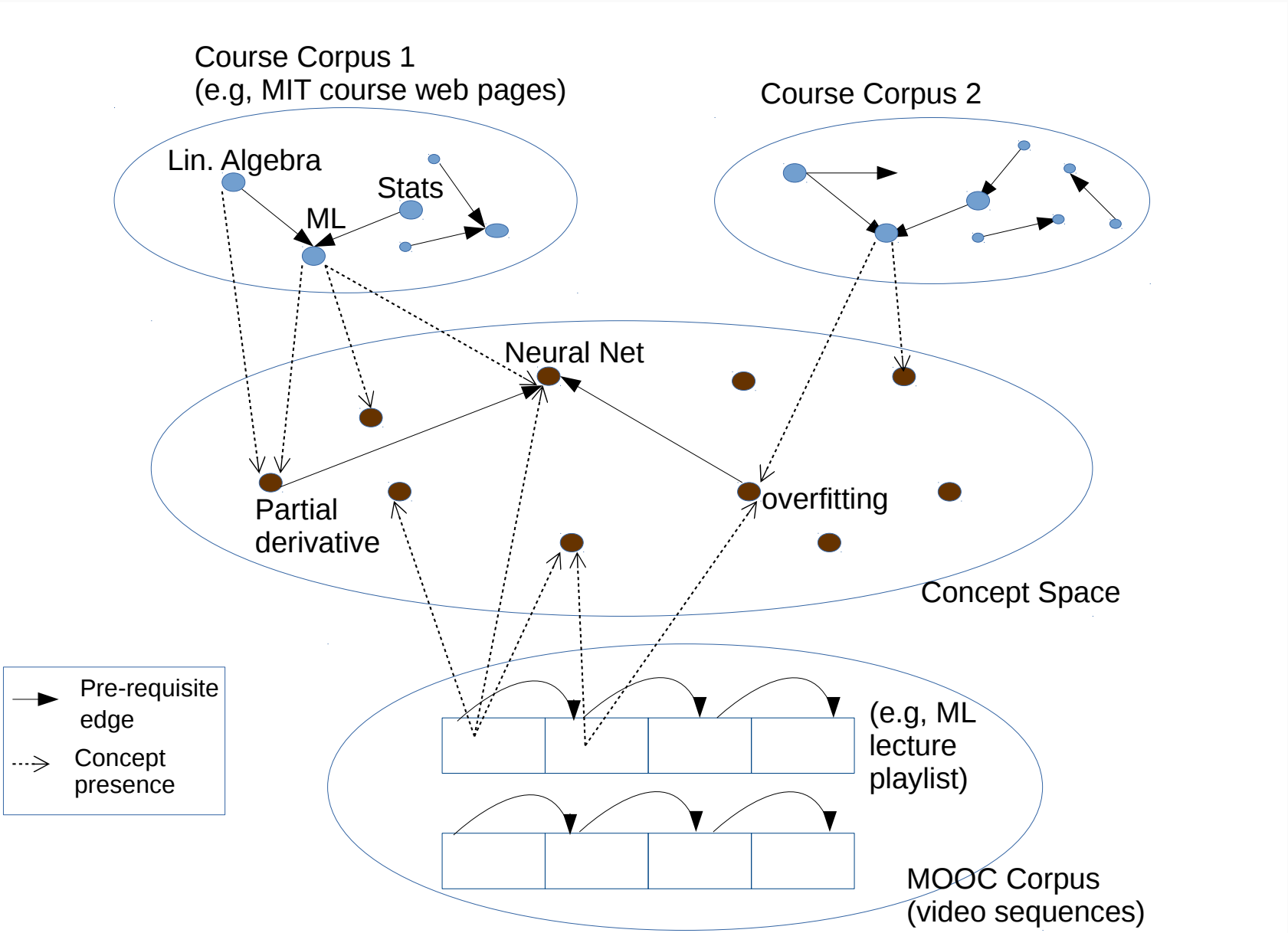


# Problem Statement




[1] CPR-Recover (EAAI '17)

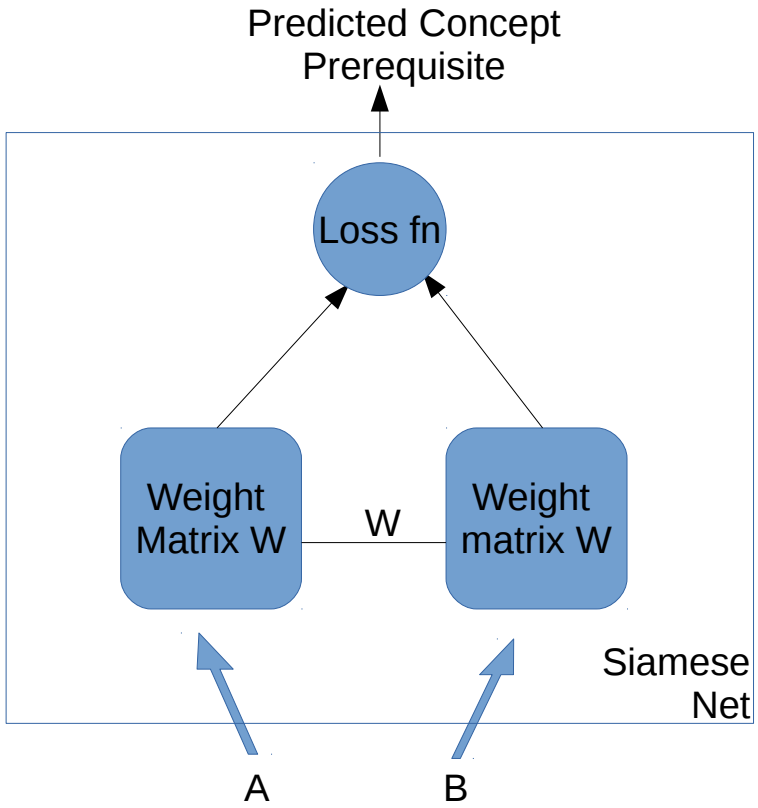
# Problem Statement



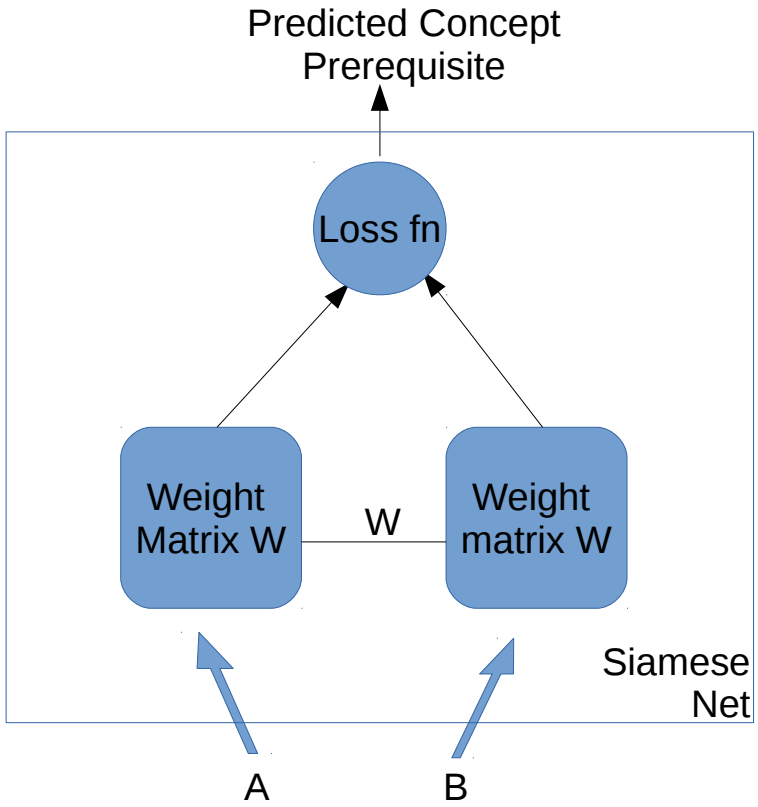

[1] CPR-Recover (EAAI '17) , [3] MOOC-RF (ACL '17)

# PREREQ Algorithm

# Our Approach

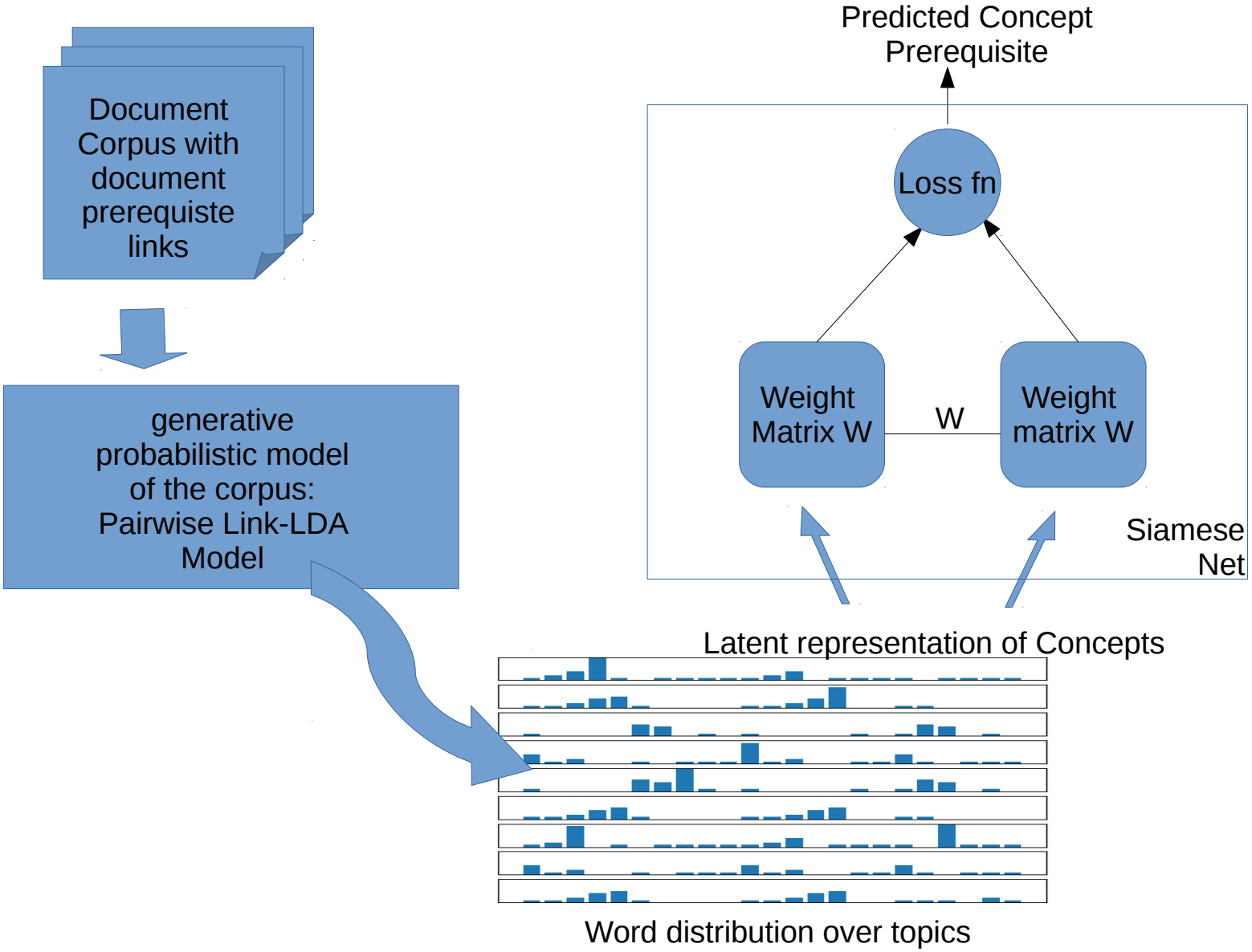


# Our Approach

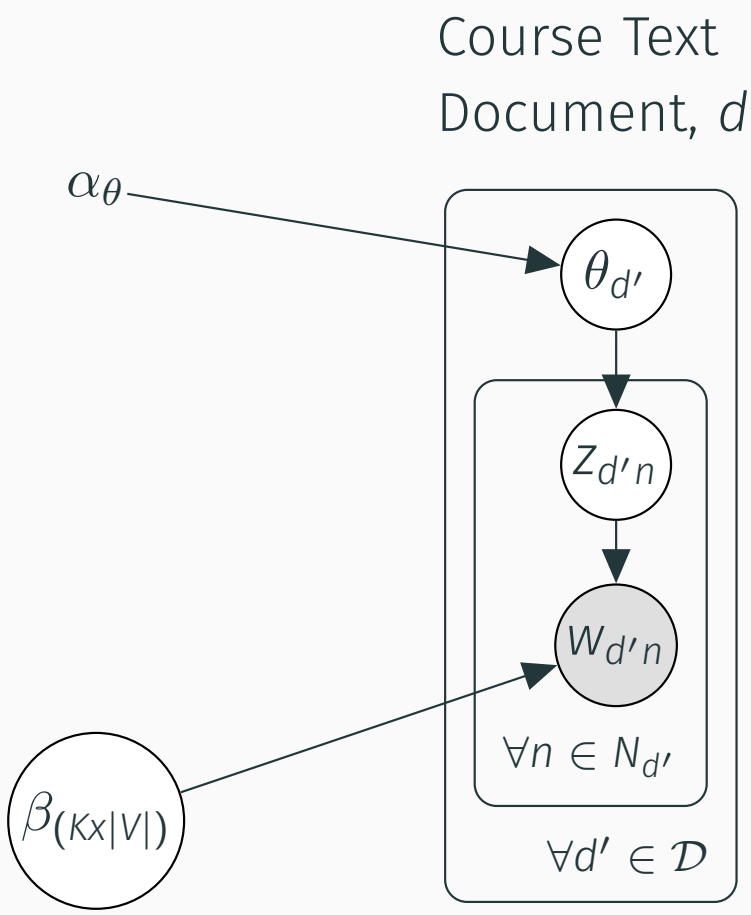



Problem: We do not have enough annotated data

# Our Approach

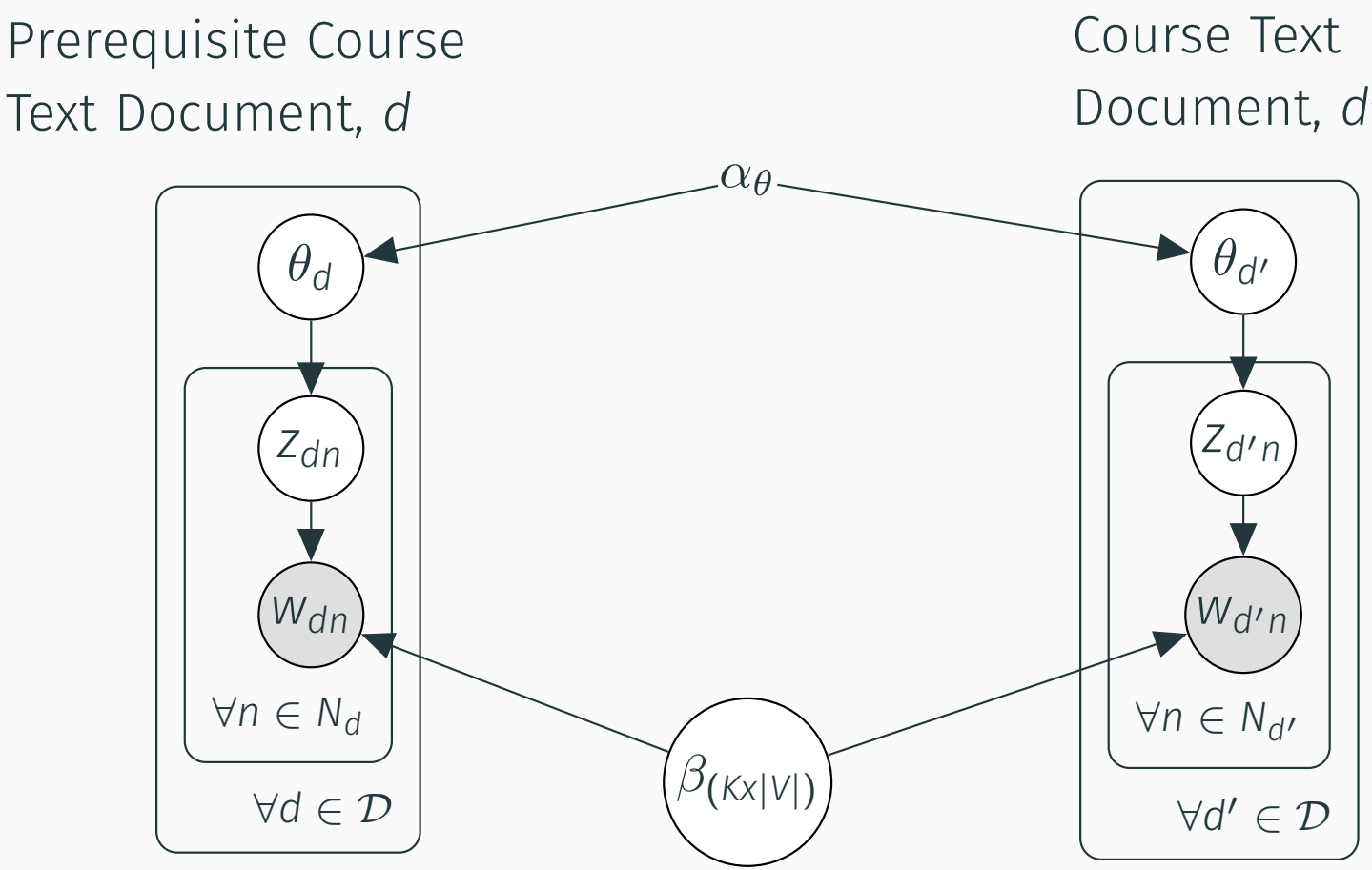


# Concept Representation: Pairwise-link LDA



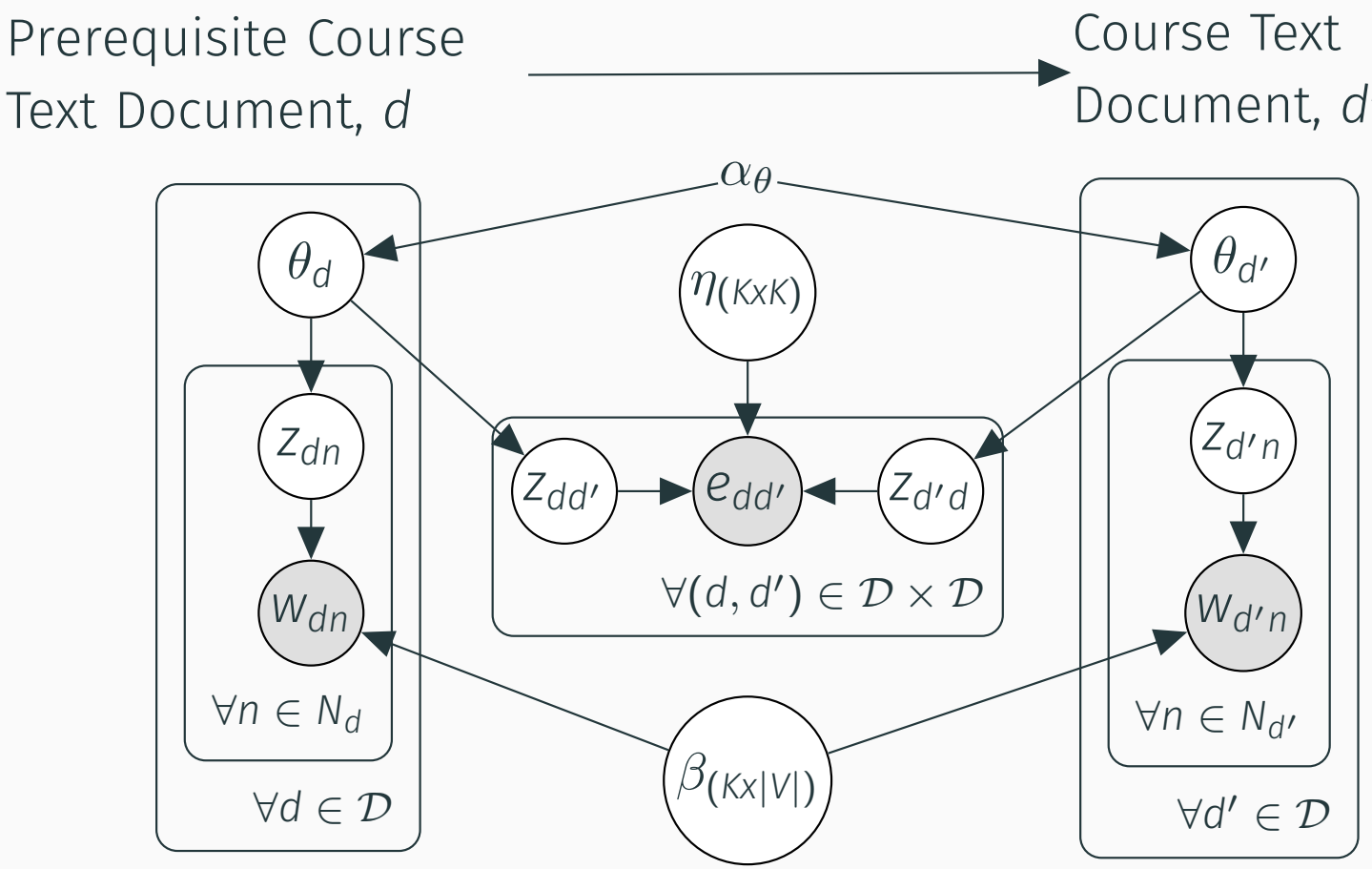

$K$ = number of topics, $V$ = vocabulary

# Concept Representation: Pairwise-link LDA




$K$ = number of topics, $V$ = vocabulary

# Concept Representation: Pairwise-link LDA



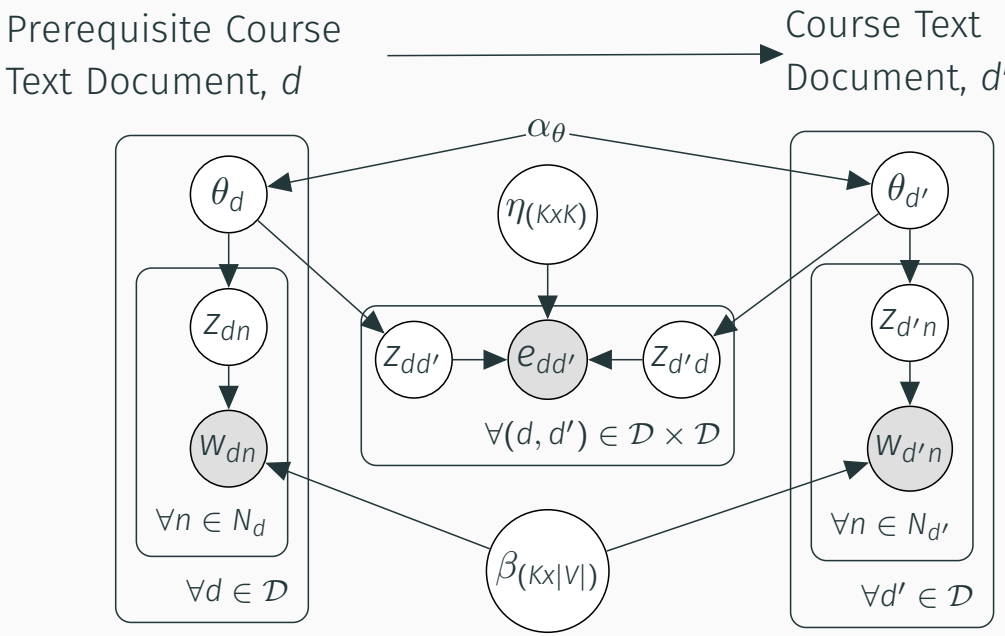

Observed prerequisite relation, $e_{dd'} \sim Bernoulli(.|\eta_{z_{dd'}, z_{d'd}})$

$K$ = number of topics, $V$ = vocabulary

# Pairwise-link LDA: Discriminatory Signal

Given, $V$ is the vocabulary of $n$-grams concepts and $K$ be number of topics,

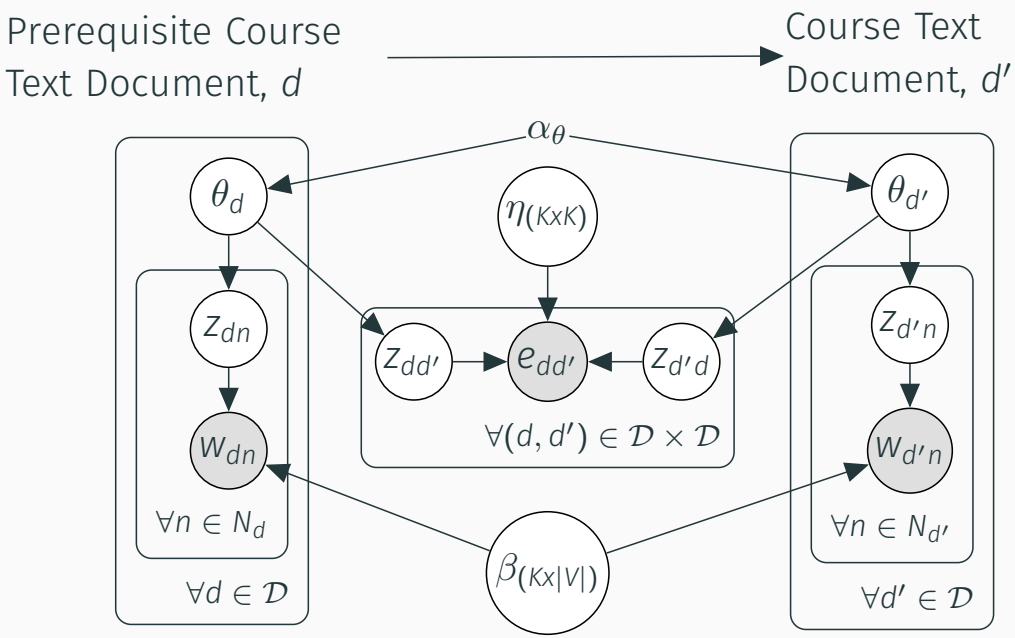
- $\beta_{K\times|V|}$, the word distribution over topics
- $\eta_{K\times K}$, the asymmetric relationship between each pair of topics.



# Pairwise-link LDA: Discriminatory Signal

Given, $V$ is the vocabulary of $n$-grams concepts and $K$ be number of topics,

- $\beta_{K\times|V|}$, the word distribution over topics
- $\eta_{K\times K}$, the asymmetric relationship between each pair of topics.
- Using ($\beta$ and $\eta$), we may predict $c_s$ –> $c_t$ as $\beta_{c_s}^T \eta \beta_{c_t}$. Later mentioned as **Pairwise LDA**.



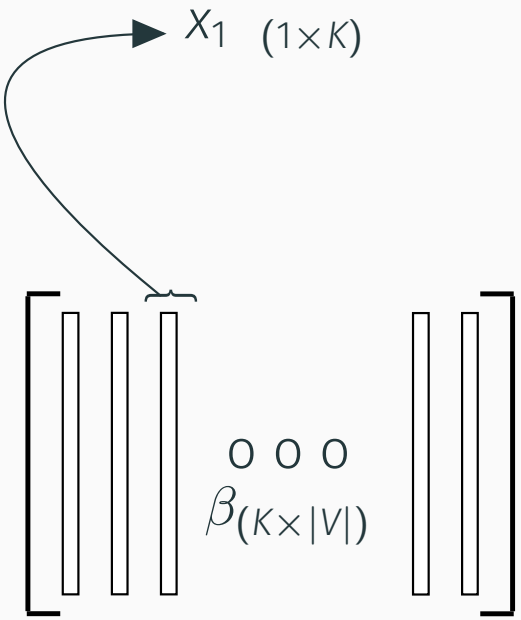
# Pairwise-link LDA: Discriminatory Signal

Given, $V$ is the vocabulary of $n$-grams concepts and $K$ be number of topics,

- $\beta_{K \times |V|}$, the word distribution over topics
- $\eta_{K \times K}$, the asymmetric relationship between each pair of topics.
- Using ($\beta$ and $\eta$), we may predict $c_s$ –> $c_t$ as $\beta_{c_s}^T \eta \beta_{c_t}$. Later mentioned as **Pairwise LDA**.
- Using two different measures that use topics of the learned model, statistical hypothesis testings show that, **topics have enough signal to discriminate between related and unrelated documents**.

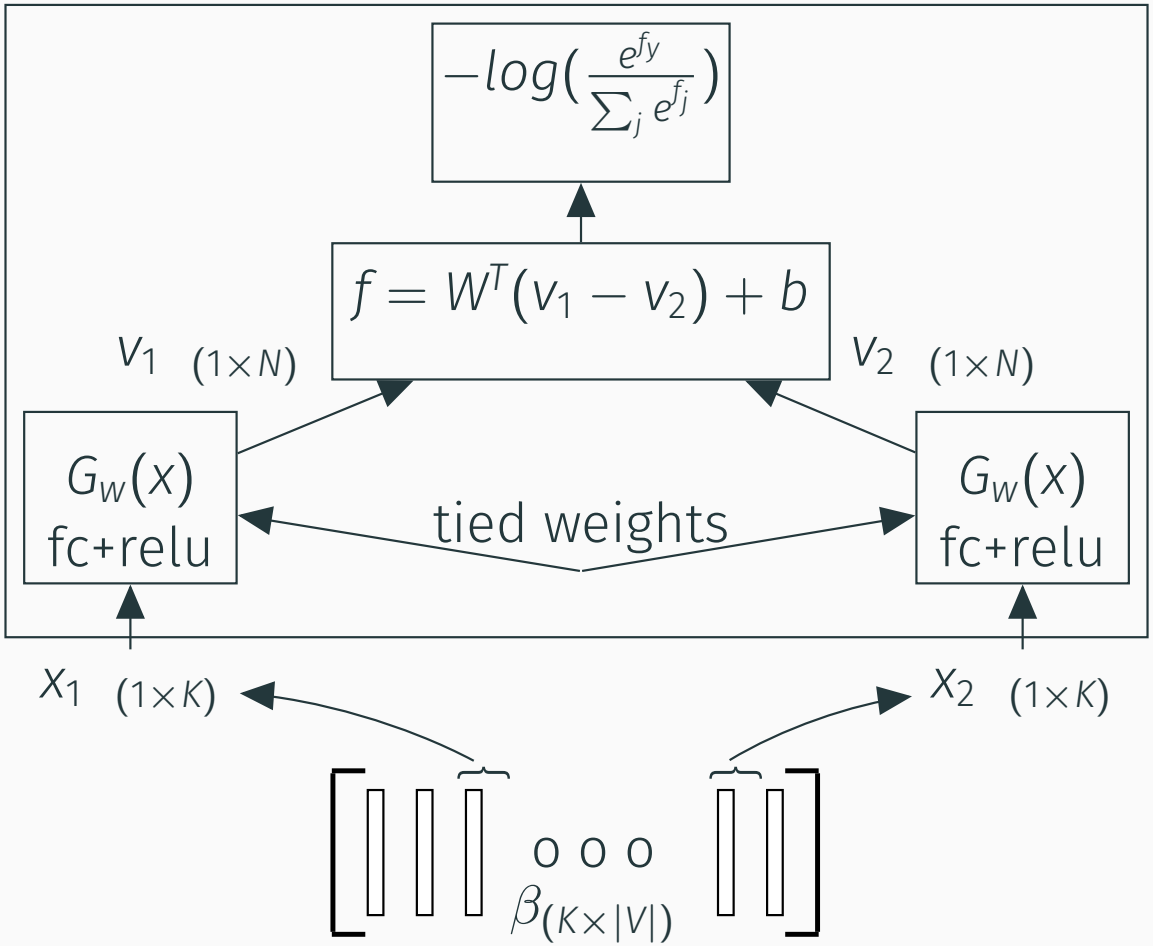
# Concept Representation

Represent each concept as distribution over topics, from the learned Pairwise-link LDA model.




$K$ = number of topics, $V$ = vocabulary

# Relationship Prediction: Siamese Network




$K$ = number of topics, $V$ = vocabulary

$y \in \{0, 1\}$ is the label of the corresponding ordered pair

# Experiments

# Datasets for Evaluation

| Dataset | Number of Course/ Lecture Document | Number of Course Prerequisite Edges | Number of Concepts | Number of Concept Prerequisite Pairs |
|---|---|---|---|---|
| University Course Dataset [1] | 654 | 861 | 365 | 1008 |
| NPTEL MOOC Dataset [1] | 382 | 1445 | 345 | 1008 |

[1] http://nptel.ac.in/

[1] Recovering Concept Prerequisite Relations from University Course Dependencies, EAAI 2017

# Experimental Settings

- Results are presented over 5-fold cross validation
- Evaluation metrics,
  Precision = $\frac{True\ positive}{True\ positive+False\ positive}$, Recall = $\frac{True\ positive}{True\ positive+False\ negative}$
  and,
  F-measure = $2 \times \frac{Precision.Recall}{Precision+Recall}$
  Also,
  Precision@K = $\frac{\sum_{i=1}^{K} rel(i)}{K}$, where rel(.) is a binary indicator of presence of the concept pair $(c_s, c_t)$ in the ground truth

Code and Datasets, https://github.com/suderoy/PREREQ-IAAI-19

# Experiments

We did three sets of experiments,

- Performance evaluation against the baseline methods
- Effect of training data size
- Effectiveness of the learned representation

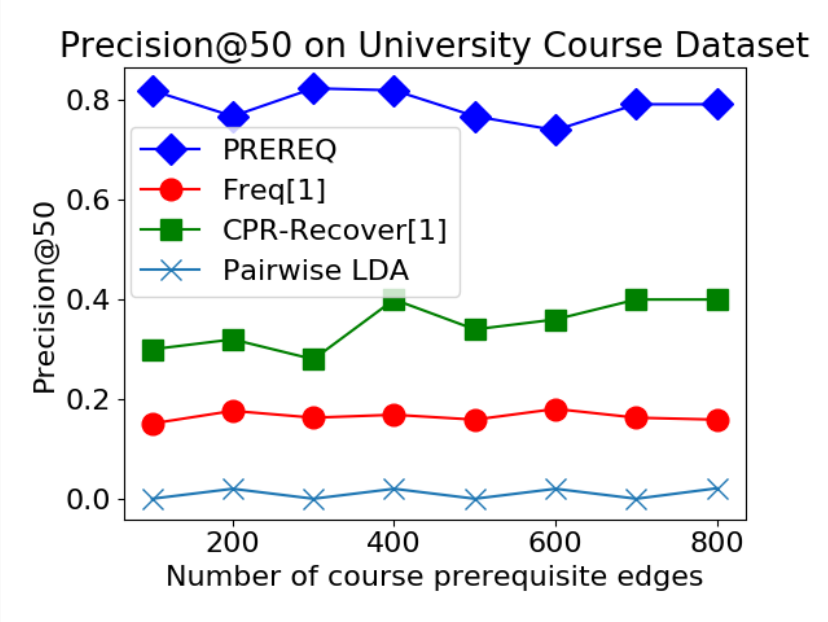
# Performance Evaluation

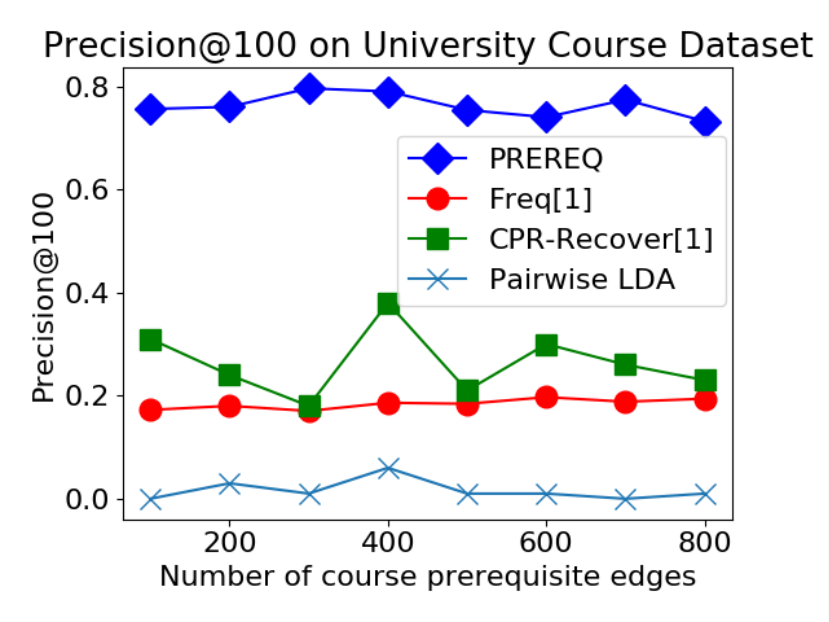
| Dataset | University Course Dataset[1] | | | NPTEL MOOC Dataset | | |
|---|---|---|---|---|---|---|
| Method | Precision | Recall | F-score | Precision | Recall | F-score |
| PREREQ | 46.76 | **91.64** | **59.68** | 55.60 | **75.74** | **60.73** |
| Pairwise LDA | **98.27** | 16.42 | 28.14 | 48.43 | 10.47 | 17.22 |
| CPR-Recover[1] | 16.66 | 46.51 | 24.54 | 17.18 | 52.97 | 25.94 |
| MOOC-RF[3] | 43.70 | 53.43 | 50.95 | **59.74** | 56.48 | 58.07 |

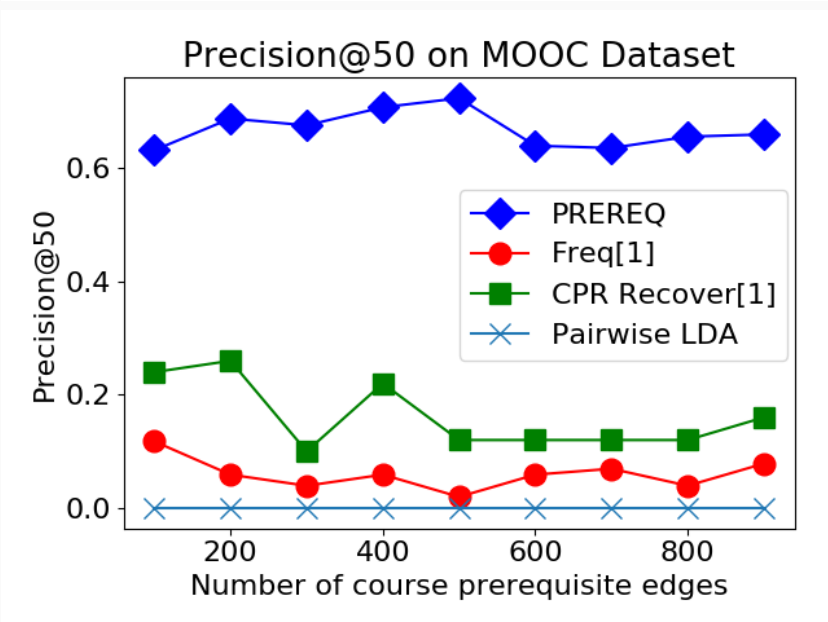
[1] Recovering Concept Prerequisite Relations from University Course Dependencies, EAAI 2017

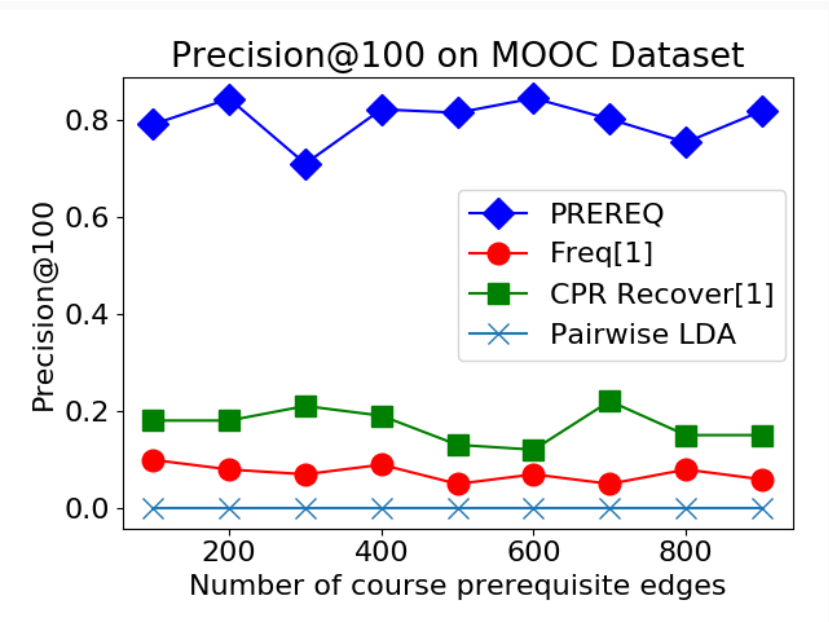
[3] Prerequisite Relation Learning for Concepts in MOOCs, ACL 2017

# Evaluation: Precision@K over 5-fold cross validation

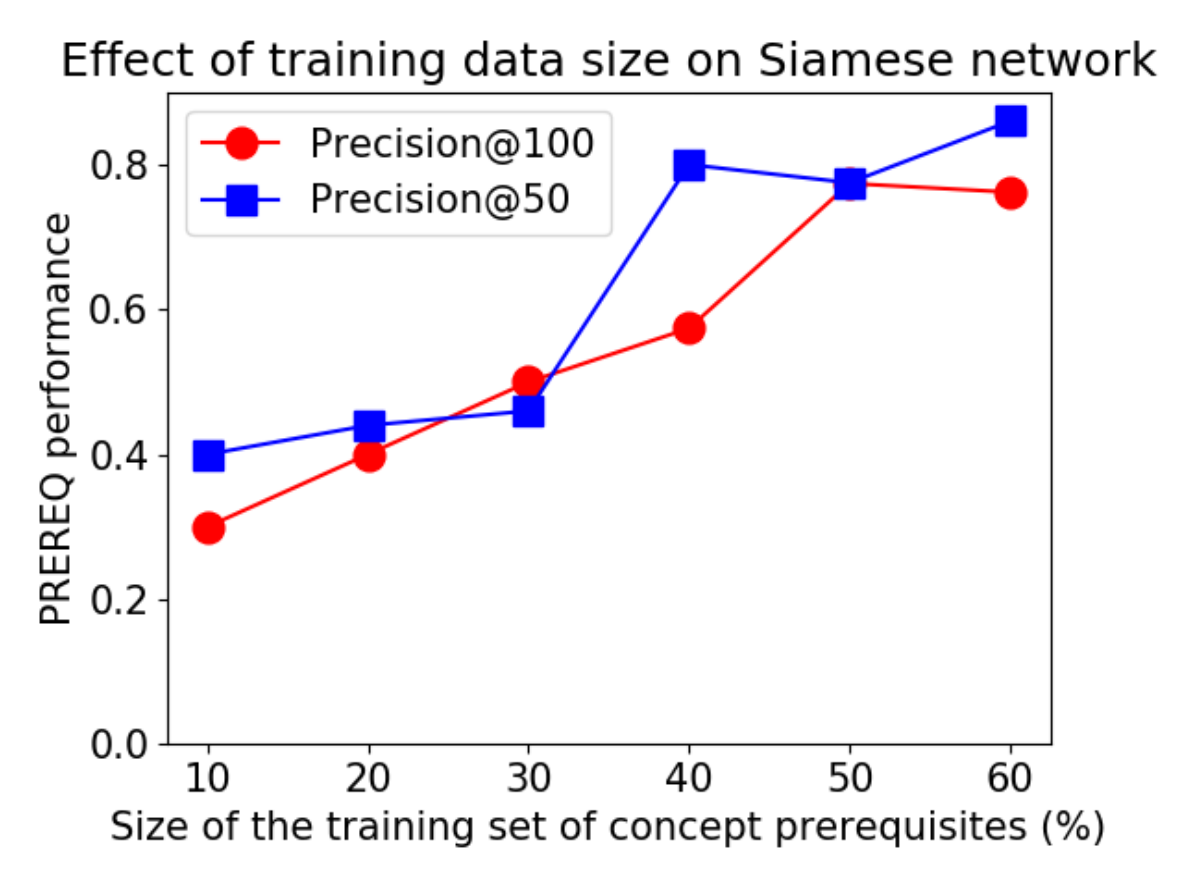


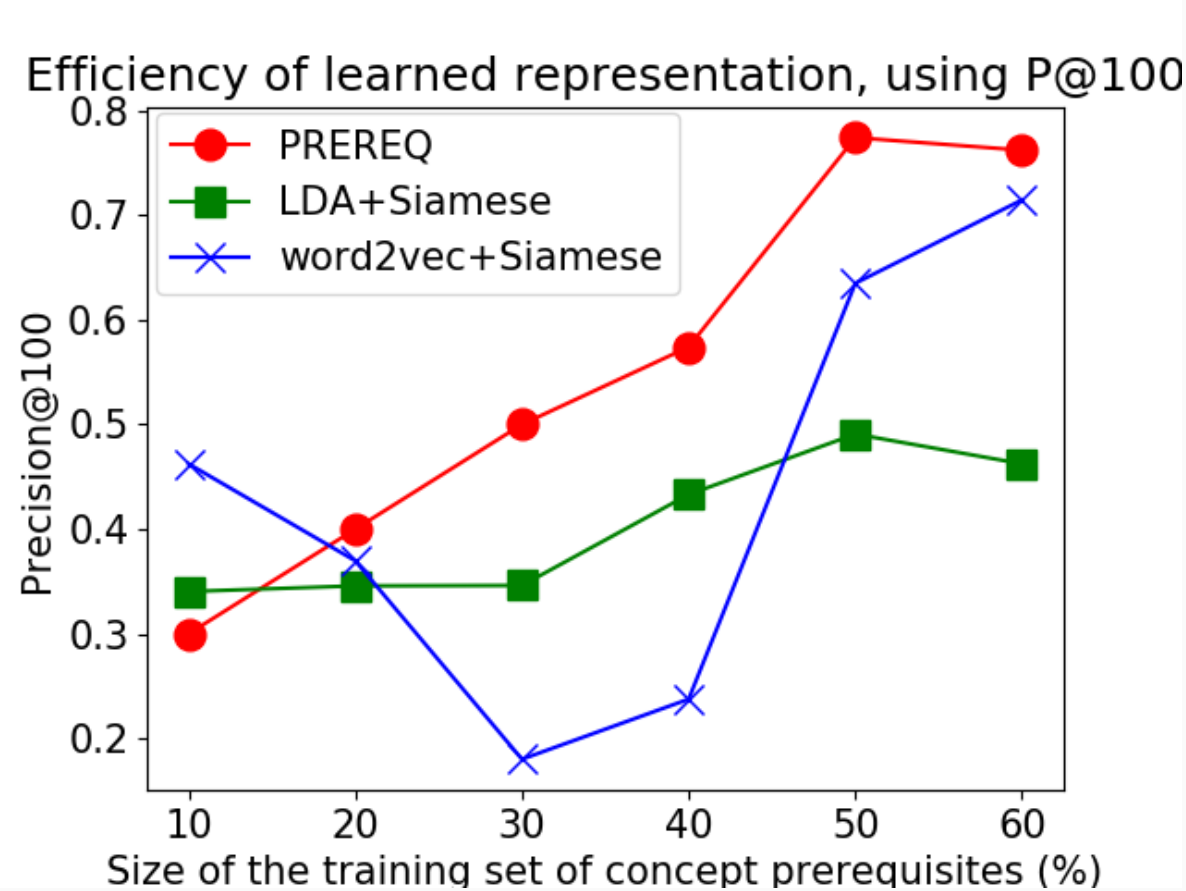



# Effect of Training Data Size




Effect of training data size, averaged over multiple random train-test splits.

# Efficacy of the Learned Representation

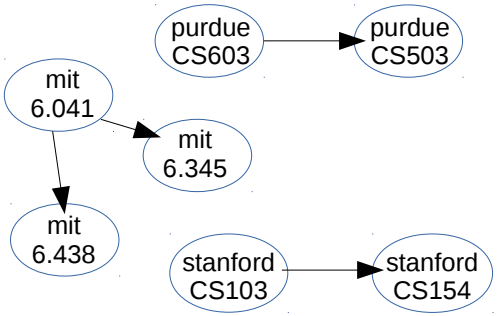



Comparison between different Concept Representations, averaged over multiple random train-test splits.

*word2vec embeddings are trained on latest wikipedia corpus, with best threshold to get all the ngram concept phrases.

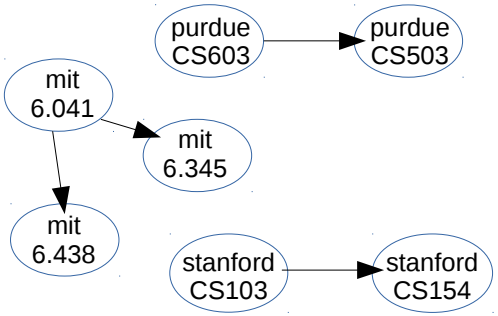
# Illustration

# Illustration



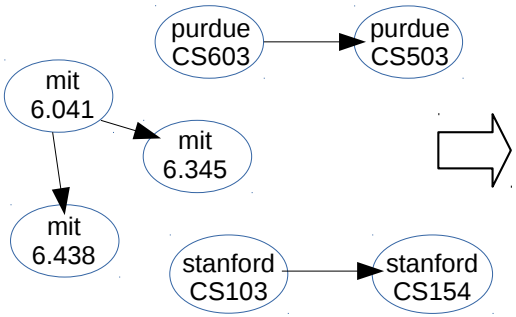

| | |
|---|---|
| mit 6.041 | An introduction to probability theory, and the modeling and analysis of probabilistic systems. Probabilistic models, conditional probability. Discrete and continuous random variables. Expectation and conditional expectation. Limit Theorems. Bernoulli and Poisson processes. Markov chains. Bayesian estimation and hypothesis testing. Elements of statistical inference. Meets with graduate subject 6.431, but assignments differ. |
| mit 6.438 | Introduction to statistical inference with probabilistic graphical models. Directed and undirected graphical models, and factor graphs, over discrete and Gaussian distributions; hidden Markov models, ... junction tree algorithms; forward-backward ... algorithms. Variational methods, mean-field theory, and loopy belief propagation. ... and Chow-Liu algorithms. |
| mit 6.345 | Automatic Speech Recognition. Introduces the rapidly developing fields of automatic speech recognition and spoken language processing. Topics include acoustic theory of speech production and perception, acoustic-phonetics, signal representation, acoustic and language modeling, search, hidden Markov modeling, robustness, ... |
| purdue_CS 603 | Design and control of distributed computing systems (operating systems and database systems). Topics include principles of naming and location, atomicity, resource sharing, concurrency control and other synchronization, deadlock detection and avoidance, security, distributed data access and control, integration of operating systems and computer networks, distributed systems design, consistency control, and fault tolerance |
| purdue_CS 503 | Basic principles of operating systems: addressing modes, indexing, relative addressing, indirect addressing, stack maintenance; implementation of multitask systems; control and coordination of tasks, deadlocks, synchronization, mutual exclusion; storage management, segmentation, paging, virtual memory; protection, sharing, access control; file systems; resource management; ... using and modifying a small operating system |
| stanford_CS 103 | Mathematical foundations required for computer science, including propositional predicate logic, induction, sets, functions, and relations. Formal language theory, including regular expressions, grammars, finite automata, Turing machines, and NP-completeness. Mathematical rigor, proof techniques, and applications |
| stanford_CS 154 | Introduction to Automata and Complexity Theory. This course provides a mathematical introduction to the following questions: What is computation? Given a computational model, ... problems can we hope to efficiently solve? In many cases we can give completely rigorous answers; .. able to classify computational problems in terms of their computational complexity (Is the problem regular? Not regular? Decidable? Recognizable? Neither? Solvable in P? NP-complete? PSPACE-complete?, etc.). ... technology, such as the Church-Turing Thesis and the P versus NP problem |

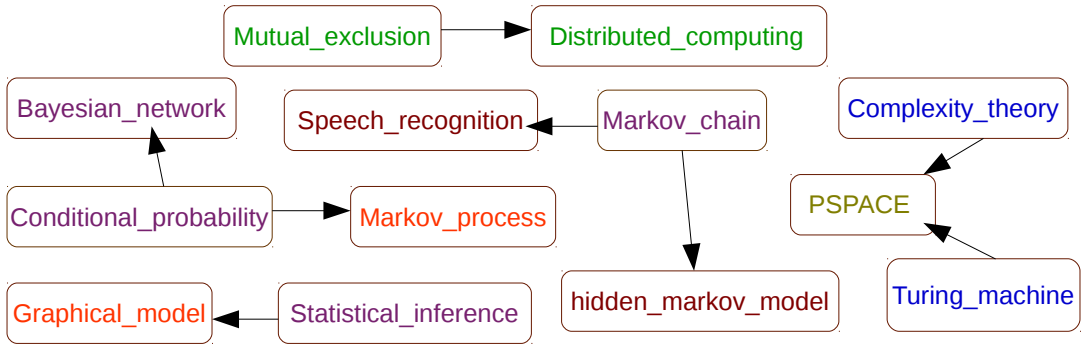
# Illustration




| | |
|---|---|
| mit 6.041 | An introduction to probability theory, and the modeling and analysis of probabilistic systems. Probabilistic models, conditional probability. Discrete and continuous random variables. Expectation and conditional expectation. Limit Theorems. Bernoulli and Poisson processes. Markov chains. Bayesian estimation and hypothesis testing. Elements of statistical inference. Meets with graduate subject 6.431, but assignments differ. |
| mit 6.438 | Introduction to statistical inference with probabilistic graphical models. Directed and undirected graphical models, and factor graphs, over discrete and Gaussian distributions; hidden Markov models, ... junction tree algorithms; forward-backward ... algorithms. Variational methods, mean-field theory, and loopy belief propagation. ... and Chow-Liu algorithms. |
| mit 6.345 | Automatic Speech Recognition. Introduces the rapidly developing fields of automatic speech recognition and spoken language processing. Topics include acoustic theory of speech production and perception, acoustic-phonetics, signal representation, acoustic and language modeling, search, hidden Markov modeling, robustness, ... |
| purdue_CS 603 | Design and control of distributed computing systems (operating systems and database systems). Topics include principles of naming and location, atomicity, resource sharing, concurrency control and other synchronization, deadlock detection and avoidance, security, distributed data access and control, integration of operating systems and computer networks, distributed systems design, consistency control, and fault tolerance |
| purdue_CS 503 | Basic principles of operating systems: addressing modes, indexing, relative addressing, indirect addressing, stack maintenance; implementation of multitask systems; control and coordination of tasks, deadlocks, synchronization, mutual exclusion; storage management, segmentation, paging, virtual memory; protection, sharing, access control; file systems; resource management; ... using and modifying a small operating system |
| stanford_CS 103 | Mathematical foundations required for computer science, including propositional predicate logic, induction, sets, functions, and relations. Formal language theory, including regular expressions, grammars, finite automata, Turing machines, and NP-completeness. Mathematical rigor, proof techniques, and applications |
| stanford_CS 154 | Introduction to Automata and Complexity Theory. This course provides a mathematical introduction to the following questions: What is computation? Given a computational model, ... problems can we hope to efficiently solve? In many cases we can give completely rigorous answers; .. able to classify computational problems in terms of their computational complexity (Is the problem regular? Not regular? Decidable? Recognizable? Neither? Solvable in P? NP-complete? PSPACE-complete?, etc.). ... technology, such as the Church-Turing Thesis and the P versus NP problem |

# Illustration




| | |
|---|---|
| mit 6.041 | An introduction to probability theory, and the modeling and analysis of probabilistic systems. Probabilistic models, conditional probability. Discrete and continuous random variables. Expectation and conditional expectation. Limit Theorems. Bernoulli and Poisson processes. Markov chains. Bayesian estimation and hypothesis testing. Elements of statistical inference. Meets with graduate subject 6.431, but assignments differ. |
| mit 6.438 | Introduction to statistical inference with probabilistic graphical models. Directed and undirected graphical models, and factor graphs, over discrete and Gaussian distributions; hidden Markov models, ... junction tree algorithms; forward-backward ... algorithms. Variational methods, mean-field theory, and loopy belief propagation. ... and Chow-Liu algorithms. |
| mit 6.345 | Automatic Speech Recognition. Introduces the rapidly developing fields of automatic speech recognition and spoken language processing. Topics include acoustic theory of speech production and perception, acoustic-phonetics, signal representation, acoustic and language modeling, search, hidden Markov modeling, robustness, ... |
| purdue_CS 603 | Design and control of distributed computing systems (operating systems and database systems). Topics include principles of naming and location, atomicity, resource sharing, concurrency control and other synchronization, deadlock detection and avoidance, security, distributed data access and control, integration of operating systems and computer networks, distributed systems design, consistency control, and fault tolerance |
| purdue_CS 503 | Basic principles of operating systems: addressing modes, indexing, relative addressing, indirect addressing, stack maintenance; implementation of multitask systems; control and coordination of tasks, deadlocks, synchronization, mutual exclusion; storage management, segmentation, paging, virtual memory; protection, sharing, access control; file systems; resource management; ... using and modifying a small operating system |
| stanford_CS 103 | Mathematical foundations required for computer science, including propositional predicate logic, induction, sets, functions, and relations. Formal language theory, including regular expressions, grammars, finite automata, Turing machines, and NP-completeness. Mathematical rigor, proof techniques, and applications |
| stanford_CS 154 | Introduction to Automata and Complexity Theory. This course provides a mathematical introduction to the following questions: What is computation? Given a computational model, ... problems can we hope to efficiently solve? In many cases we can give completely rigorous answers; .. able to classify computational problems in terms of their computational complexity (Is the problem regular? Not regular? Decidable? Recognizable? Neither? Solvable in P? NP-complete? PSPACE-complete?, etc.). ... technology, such as the Church-Turing Thesis and the P versus NP problem |

# Demo

A Demo

# Conclusion

# Summary

**Model** PREREQ obtains concept representations through the Pairwise-Link LDA model, followed by, Siamese net based classifier to identify prerequisite relations.

**Generic** PREREQ can learn effectively from course webpages as well as unlabeled video playlists, using minimal training data.

**Utility** PREREQ can effectively utlize the large course corpora and MOOCs to solve a fundamental problem, essential for online educational technology applications.

# References


C. Liang, J. Ye, Z. Wu, B. Pursel, and C. L. Giles.
**Recovering concept prerequisite relations from university course dependencies.**
In *AAAI*, 2017.

R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen.
**Joint latent topic models for text and citations.**
In *ACM SIGKDD*, 2008.

L. Pan, C. Li, J. Li, and J. Tang.
**Prerequisite relation learning for concepts in moocs.**
In *ACL*, volume 1, pages 1447–1456, 2017.

# Pairwise Link-LDA: Generative Process

For each document $d \in \mathcal{D}$

$\theta_d \sim Dirichlet(.|\alpha)$

For each $w_n \in d$

$z_n \sim Multinomial(.|\theta_d)$

$w_n \sim Multinomial(.|\beta_{z_n})$

For each document pair $(d, d')$

$z_{dd'} \sim Multinomial(.|\theta_d)$

$z_{d'd} \sim Multinomial(.|\theta_{d'})$

$e_{dd'} \sim Bernoulli(.|\eta_{z_{dd'}, z_{d'd}})$

# Pairwise Link-LDA: Likelihood

let $d$ be a prerequisite of $d'$ and $z_{dd'}$ and $z_{d'd}$ be the latent topics sampled from $d$ and $d'$ respectively for this interaction. Then the parameter used to generate the Bernoulli random variable $e_{d,d'}$ will be $\eta_{z_{dd'},z_{d'd}}$ which is different from $\eta_{z_{d'd},z_{dd'}}$ thus modeling the directionality in the relationship.

The likelihood of the observed data with respect to the model parameters is,

$$p(c, w, \theta, z|\alpha, \beta, \eta) = (\prod_{d=1}^{\mathcal{D}} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)) \\ (\prod_{d,d'} p(z_{dd'}, z_{d'd}|\theta_d, \theta_{d'}) p(c_{dd'}|z_{dd'}, zd'd, \eta) \tag{1}$$

We refer the readers to [2] for more details of the model.

# Pairwise Link-LDA: Evidence Lower BOund (ELBO)

ELBO = The expected log-likelihood of the data +
the KL divergence between the prior and approximate conditionals.
Say, variational params for the RV $\theta_d$, $z_d$ and $z_{dd'}$ are $\gamma_d$, $\phi_d$ and $\lambda_{dd'}$.

$$\begin{aligned}\mathcal{L}(\phi, \lambda, \gamma; \alpha, \beta, \eta) = \sum_{d=1}^{D}(E_q[logp(\theta|\alpha)]+ \\ \sum_{n=1}^{N} E_q[logp(z_{dn}|\theta_d)]E_q[p(w_{dn}|z_{dn}, \beta))])+ \\ \sum_{d,d'}(E_q[logp(z_{dd'}, z_{d'd}|\theta_d, \theta_{d'})] + E_q[logp(c_{dd'}|z_{dd'}, zd'd, \eta)] \\ - \sum_{d=1}^{D}(E_q[logq(\theta_d)] - \sum_{n=1}^{N} E_q[logq(z_{dn})]) \\ - \sum_{d,d'} E_q[logq(z_{dd'}, z_{d'd})]\end{aligned} \tag{2}$$

Mean-field variational approximation is used to infer the values of $\beta$ and $\eta$ by alternately maximizing the ELBO on log-likelihood and estimating the variational parameters.

# Topics in Related and Unrelated Documents

$E_D$ be the set of document pairs $(d_i, d_j)$ such that there is a prerequisite relation between $d_i$ and $d_j$ in either direction.
$\overline{E_D}$ be the set of document pairs $(d_i, d_j)$ such that there is no prerequisite relation between $d_i$ and $d_j$ in either direction.

- Jaccard Index $\frac{\theta_i \cap \theta_j}{\theta_i \cup \theta_j}$ is significantly different between document pairs in the sets $E_D$ and $\overline{E_D}$, using the University Course dataset[1], p-value: 1.54901e-51.
- Kullback-Leibler divergence $D_{KL}(\theta_i || \theta_j)$ between topics in the document pairs are significantly different between document sets $E_D$ and $\overline{E_D}$, using the University Course dataset, p-value: 5.14549e-31.
- The tests suggest that inferred topics have discriminatory signal at the document level.