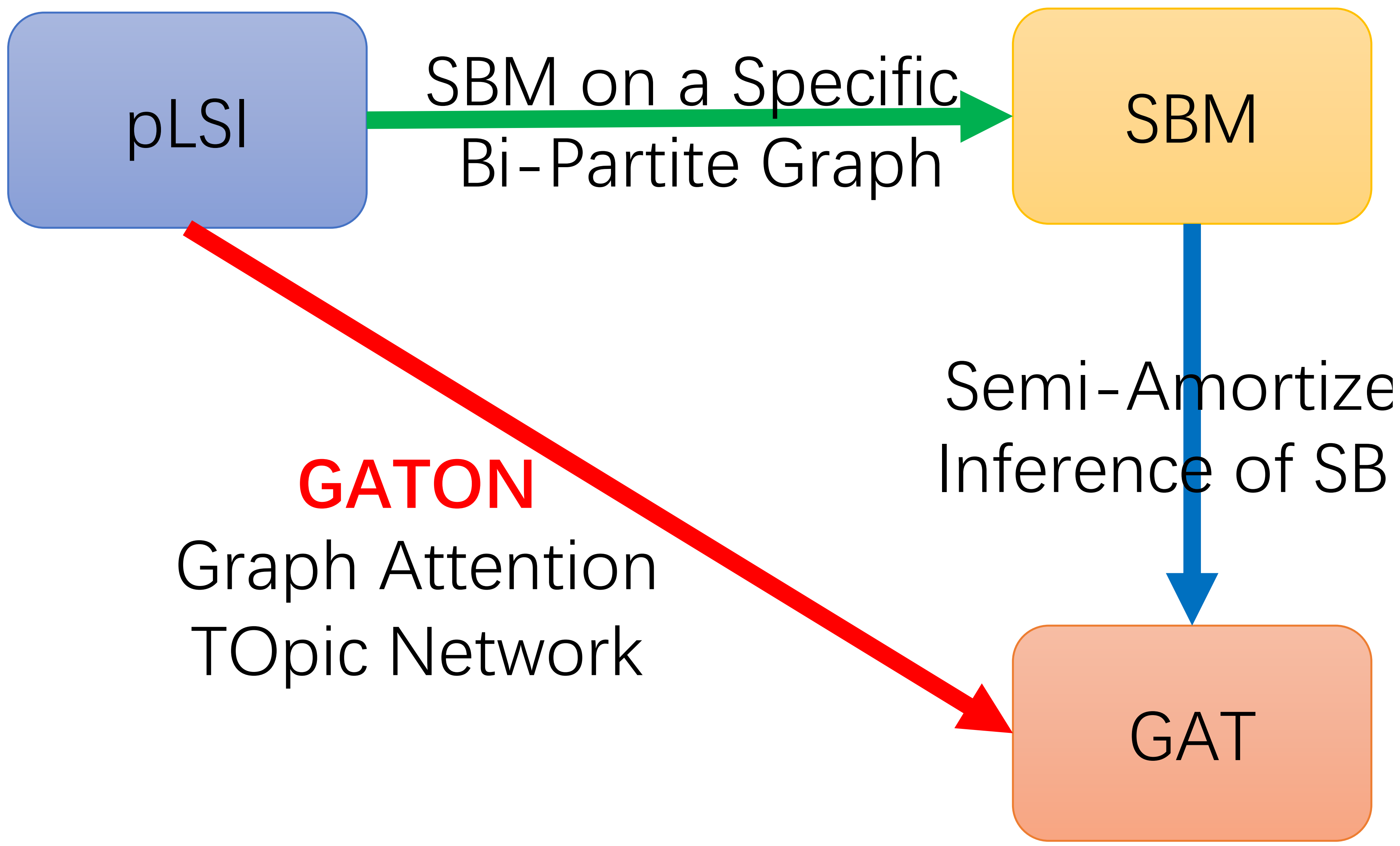
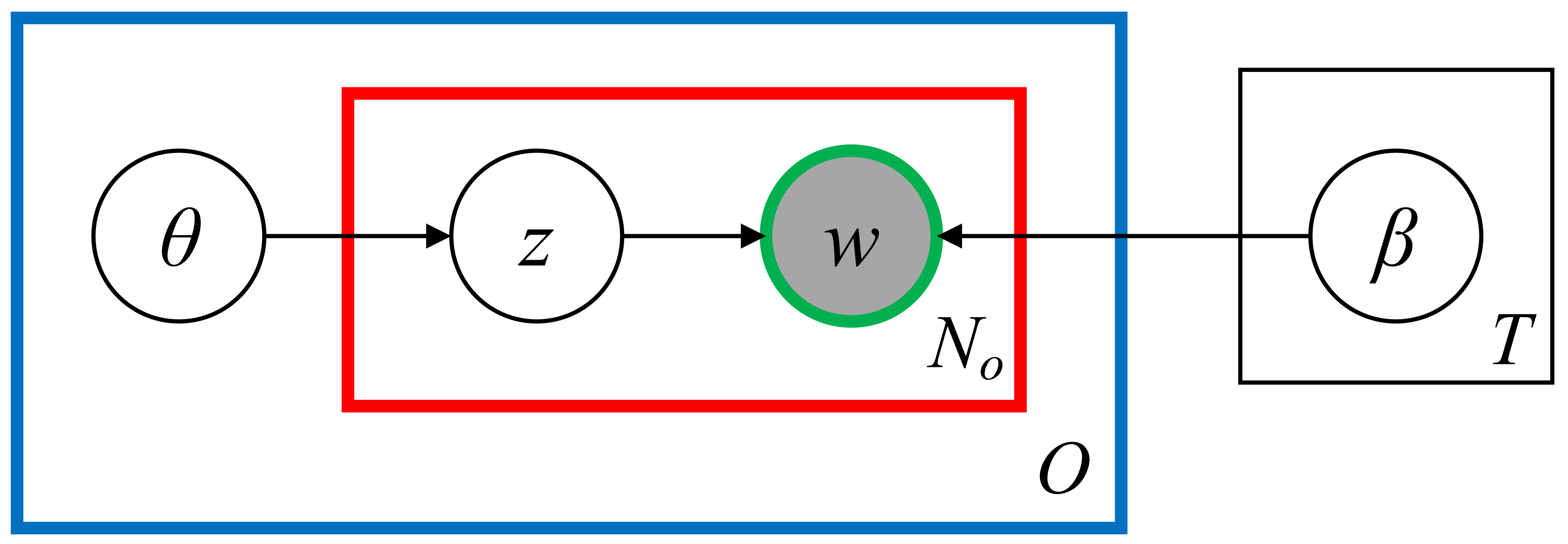


Graph Attention Topic Modeling Network

Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, Yuanfang Guo

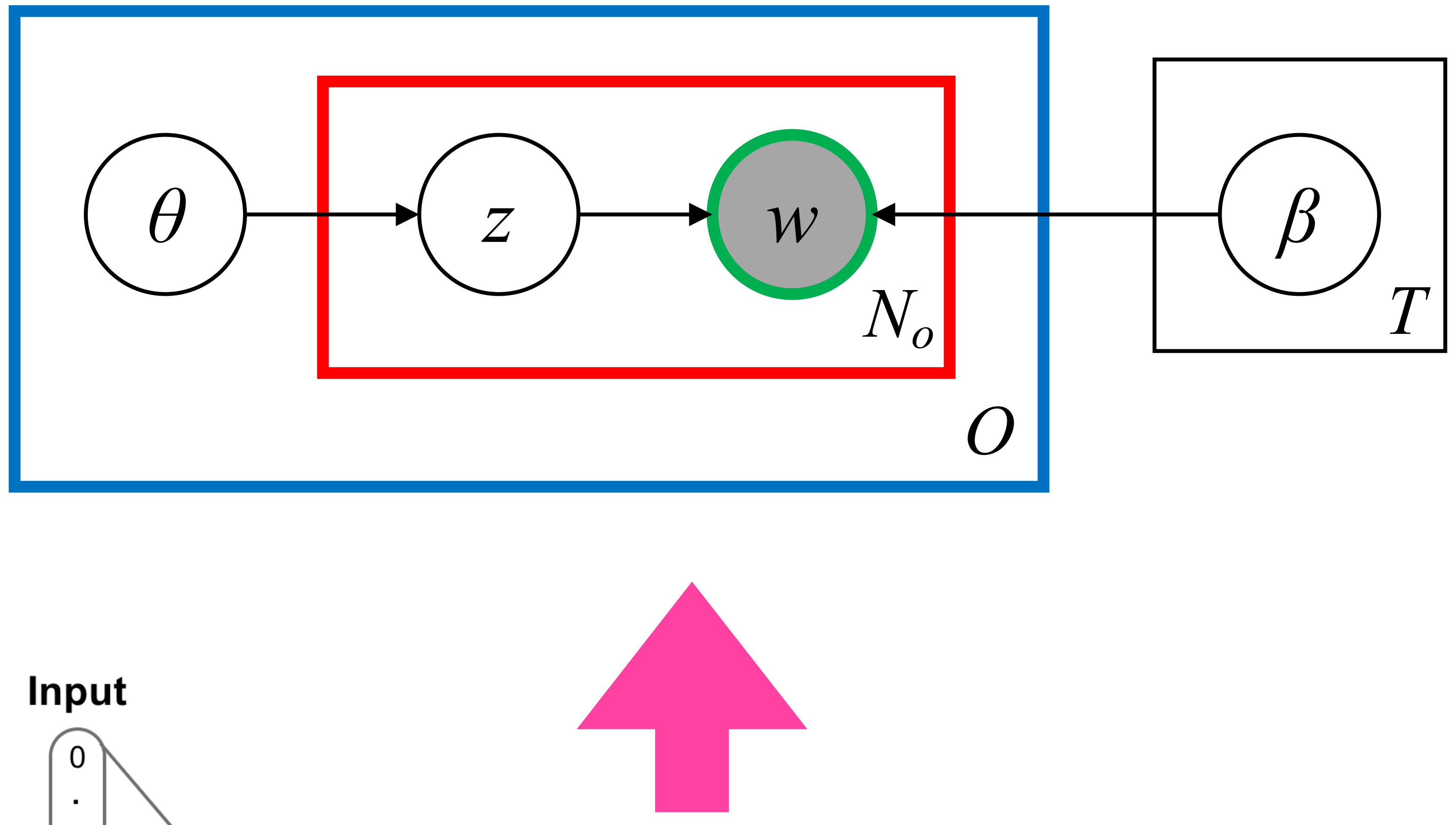
Topic Modeling

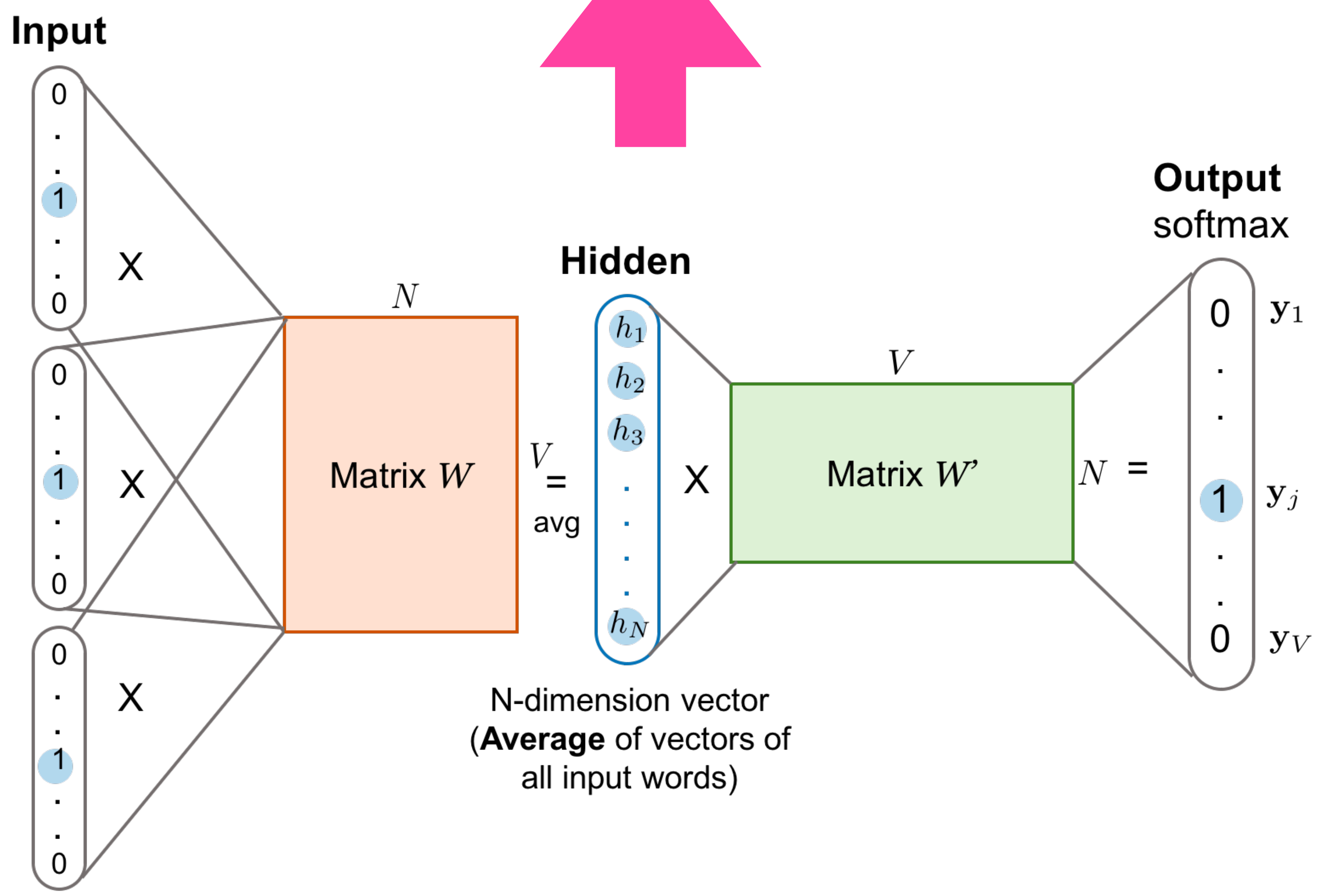


Large number of latent variables makes the inferences inefficient and induces overfitting

Issue: Latent Dirichlet Allocation alleviates the overfitting issue by introducing Dirichlet priors for latent variables, but it fails to capture the rich topical correlations among topics.

Topic Modeling





Large number of latent variables makes the inferences inefficient and induces overfitting

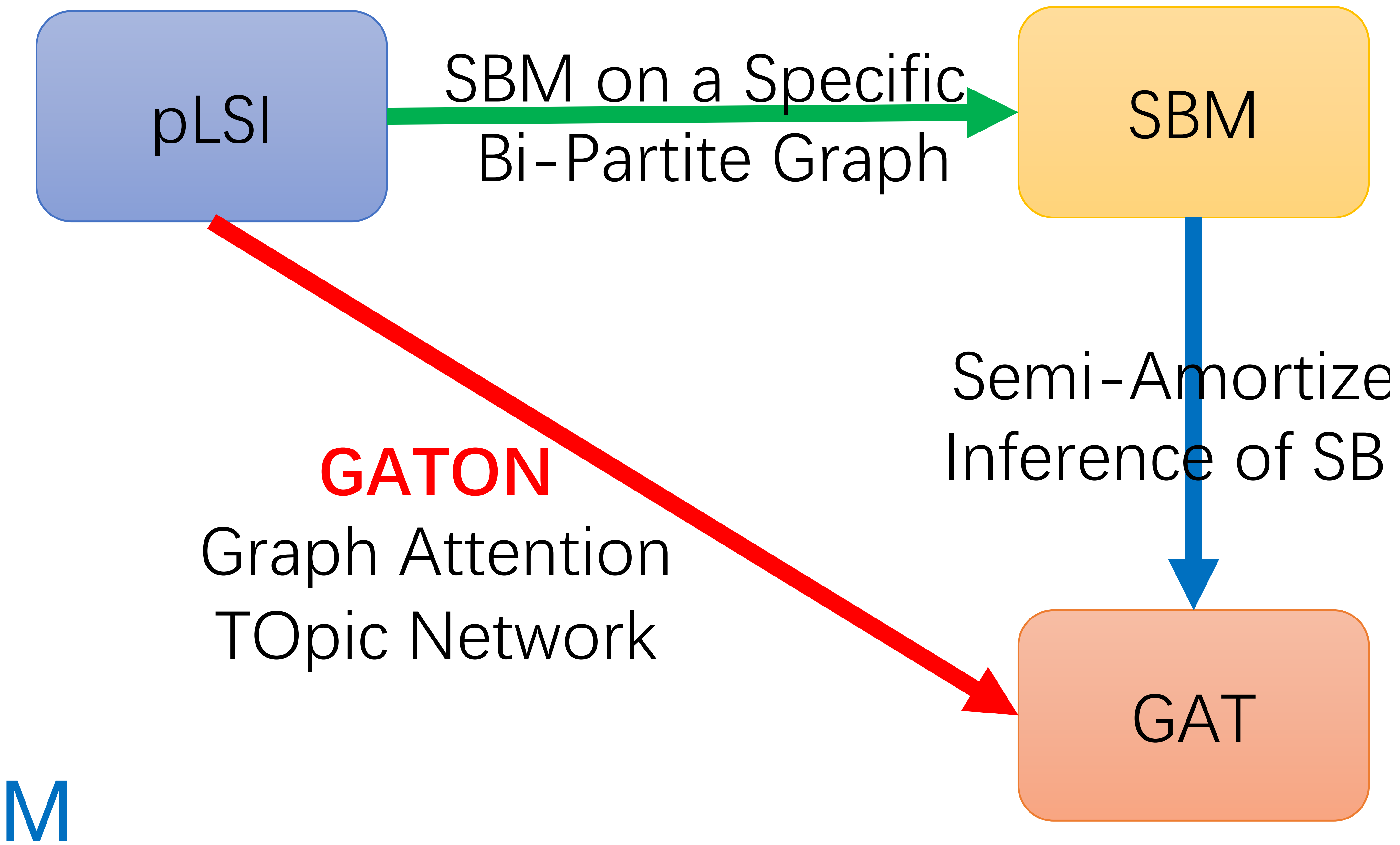
Issue: Latent Dirichlet Allocation alleviates the overfitting issue by introducing Dirichlet priors for latent variables, but it fails to capture the rich topical correlations among topics.

Intent: Overcome the overfitting issue of pLSI by exploiting the word embedding.

Question: How to integrate word embedding into generative topic modeling?

Outline

- Stochastic Block Model (SBM)
- Graph Attention Network (GAT)
- Amortized (Variational) Inference (AVI)
- GAT as Semi-Amortized Inference of SBM
- Probabilistic Latent Semantic Indexing (pLSI)
- Topic Modeling as SBM on Bi-partite Graph
- ***Graph Attention TOpic Network (GATON)***



Stochastic Block Model (SBM)

observed edge between
the nodes vi and vj

expected number of edges
between the nodes vi and vj

propensity of node vi
belonging to community k

$$P(G|\Theta) = \prod_{i<j} \boxed{\frac{\left(\sum_k \theta_{ik}\theta_{jk}\right)^{\boxed{a_{ij}}}}{a_{ij}!}\exp\left(-\boxed{\sum_k \boxed{\theta_{ik}\theta_{jk}}}\right)} \prod_i \boxed{\frac{(\sum_k \theta_{ik}\theta_{ik})^{a_{ii}/2}}{(a_{ii}/2)!}\exp\left(-\frac{1}{2}\sum_k \boxed{\theta_{ik}}\theta_{ik}\right)}.$$

Poisson distribution with
the mean value as the
expected number of edges

expected number of edges
in community k between
the nodes vi and vj

Self-loop

Stochastic Block Model (SBM)

observed edge between the nodes vi and vj

expected number of edges between the nodes vi and vj

propensity of node vi belonging to community k

$$P(G|\Theta) = \prod_{i<j} \boxed{\frac{\left(\sum_k \theta_{ik}\theta_{jk}\right)^{\boxed{a_{ij}}}}{a_{ij}!} \exp\left(-\boxed{\sum_k \boxed{\theta_{ik}\theta_{jk}}}\right)} \prod_i \boxed{\frac{\left(\sum_k \theta_{ik}\theta_{ik}\right)^{a_{ii}/2}}{(a_{ii}/2)!} \exp\left(-\frac{1}{2}\sum_k \boxed{\theta_{ik}}\theta_{ik}\right)}.$$

Poisson distribution with the mean value as the expected number of edges

expected number of edges in community k between the nodes vi and vj

Self-loop

Jensen’s inequality

Variational function

$$\log P(G|\Theta) = \sum_{i<j} a_{ij} \log\left(\sum_k \theta_{ik}\theta_{jk}\right) - \sum_{ijk} \theta_{ik}\theta_{jk} \geq \sum_{ijk} \left[a_{ij}\boxed{q_{ij}(k)}\log \frac{\theta_{ik}\theta_{jk}}{q_{ij}(k)} - \theta_{ik}\theta_{jk}\right],$$

Stochastic Block Model (SBM)

observed edge between
the nodes vi and vj

expected number of edges
between the nodes vi and vj

propensity of node vi
belonging to community k

$$P(G|\Theta) = \prod_{i<j} \boxed{\frac{\left(\sum_k \theta_{ik}\theta_{jk}\right)^{\boxed{a_{ij}}}}{a_{ij}!} \exp\left(-\boxed{\sum_k \boxed{\theta_{ik}\theta_{jk}}}\right)} \prod_i \boxed{\frac{\left(\sum_k \theta_{ik}\theta_{ik}\right)^{a_{ii}/2}}{(a_{ii}/2)!} \exp\left(-\frac{1}{2}\sum_k \boxed{\theta_{ik}}\theta_{ik}\right)}.$$

Poisson distribution with
the mean value as the
expected number of edges

expected number of edges
in community k between
the nodes vi and vj

Self-loop

Jensen's inequality

Variational function

$$\log P(G|\Theta) = \sum_{i<j} a_{ij} \log\left(\sum_k \theta_{ik}\theta_{jk}\right) - \sum_{ijk} \theta_{ik}\theta_{jk} \geq \sum_{ijk} \left[a_{ij}\boxed{q_{ij}(k)}\log\frac{\theta_{ik}\theta_{jk}}{q_{ij}(k)} - \theta_{ik}\theta_{jk}\right],$$

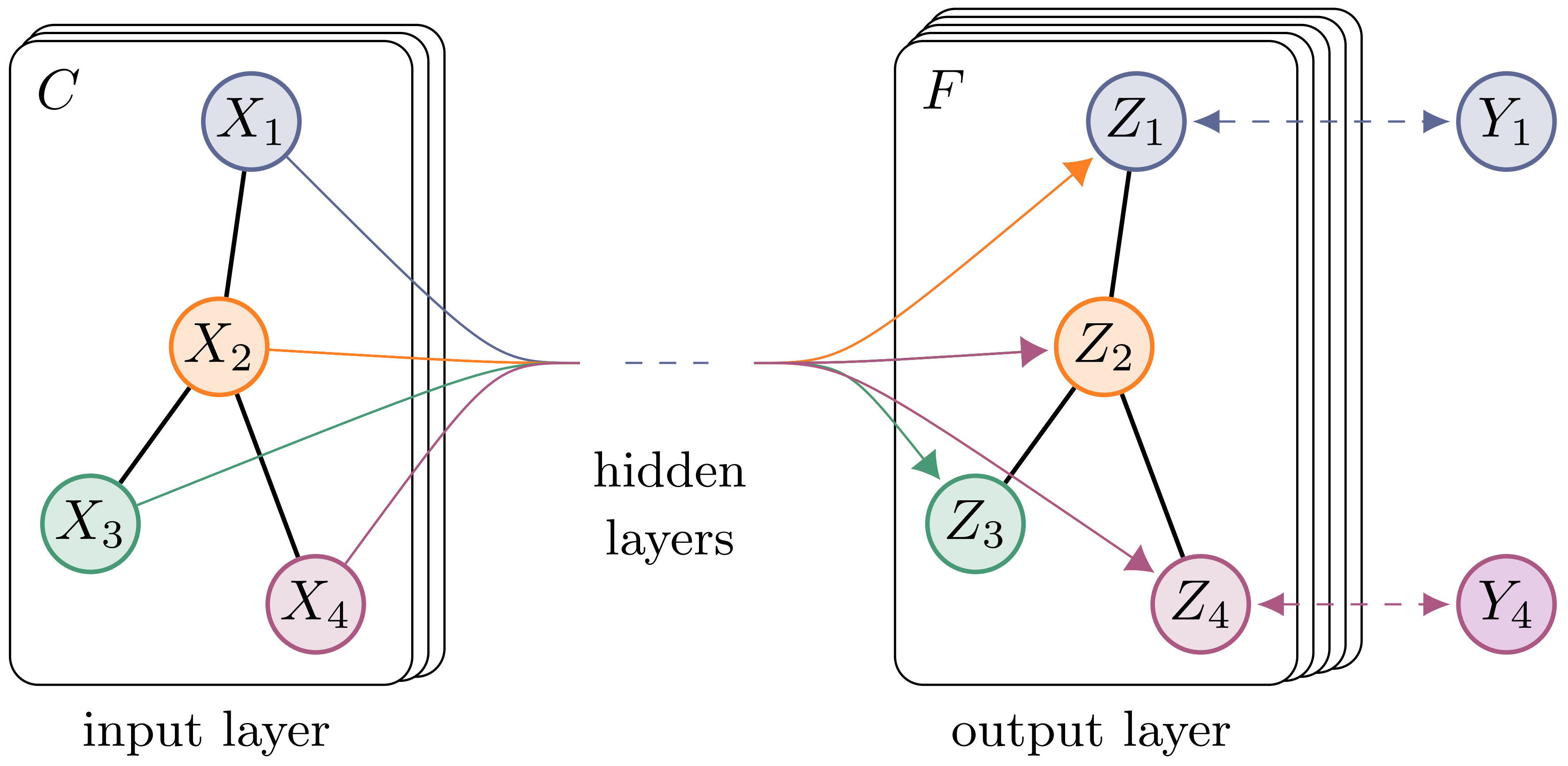
Expectation Maximization

$$q_{ij} = \frac{\theta_i \odot \theta_j}{\theta_i^T \theta_j} = \left(\frac{\theta_i}{\theta_i^T \theta_j}\right) \odot \theta_j, \qquad \theta_{ik} = \frac{\sum_j a_{ij}q_{ij}(k)}{\sum_i \theta_{ik}} = \frac{\sum_j a_{ij}q_{ij}(k)}{\sqrt{\sum_{ij} a_{ij}q_{ij}(k)}} = g_i\left(\sum_j a_{ij}q_{ij}(k)\right)$$

Graph Attention Network (GAT)

Graph Convolutional Network

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i) \cup i} \frac{1}{\sqrt{(d_i+1)(d_j+1)}} W^{(l)} h_j^{(l)}\right),$$



Graph Attention Network (GAT)

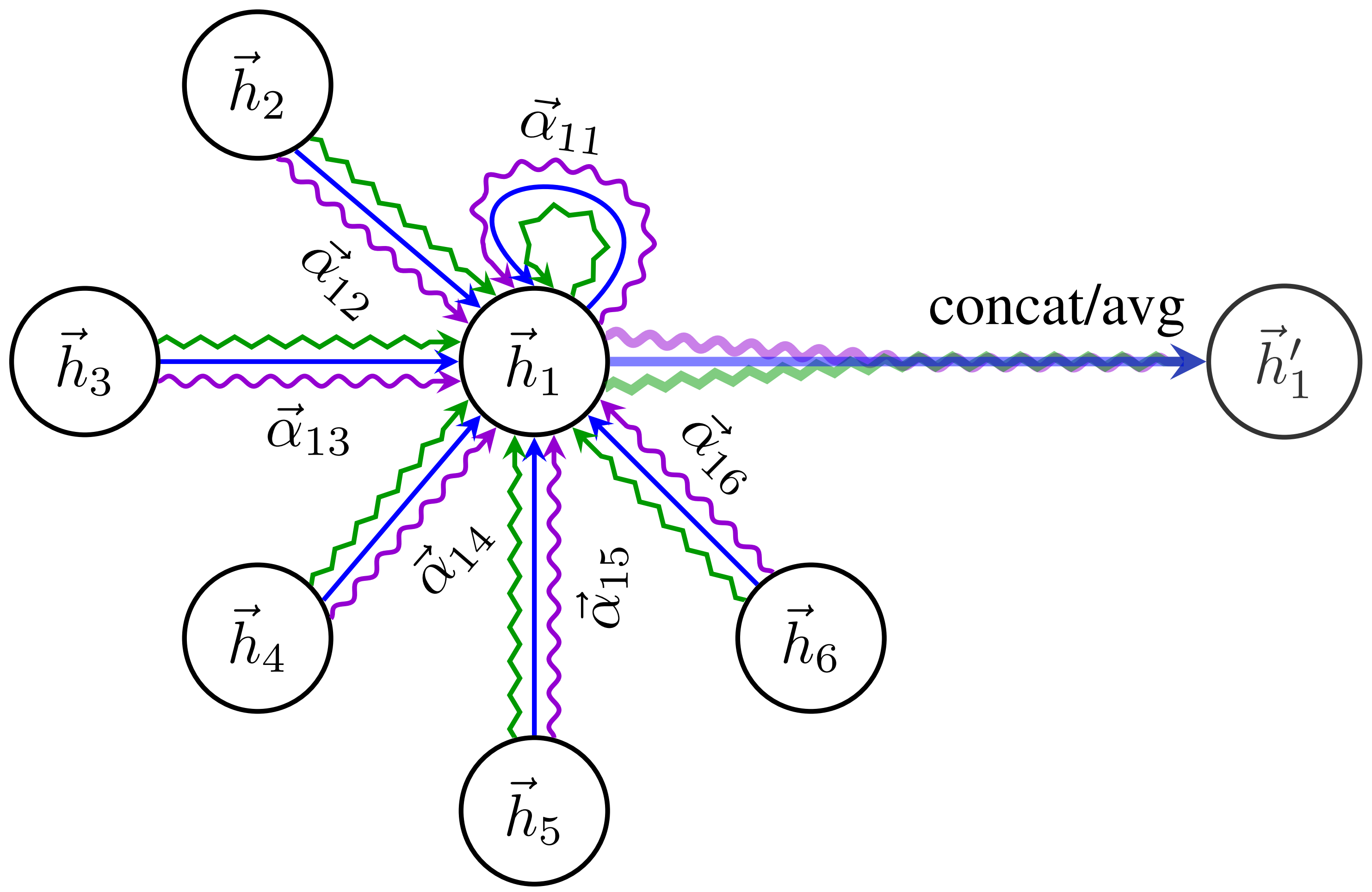
Graph Convolutional Network

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i) \cup i} \frac{1}{\sqrt{(d_i+1)(d_j+1)}} W^{(l)} h_j^{(l)}\right),$$

Graph Attention Network

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W h_j^{(l)}\right),$$

$$\boxed{\alpha_{ij} = \mathrm{softmax}_j(a(Wh_i^{(l)}, Wh_j^{(l)})) = \frac{\exp\left(\mathrm{LeakyReLU}(b^T[Wh_i || Wh_j])\right)}{\sum_{k \in N(i)} \exp\left(\mathrm{LeakyReLU}(b^T[Wh_i || Wh_k])\right)},}$$



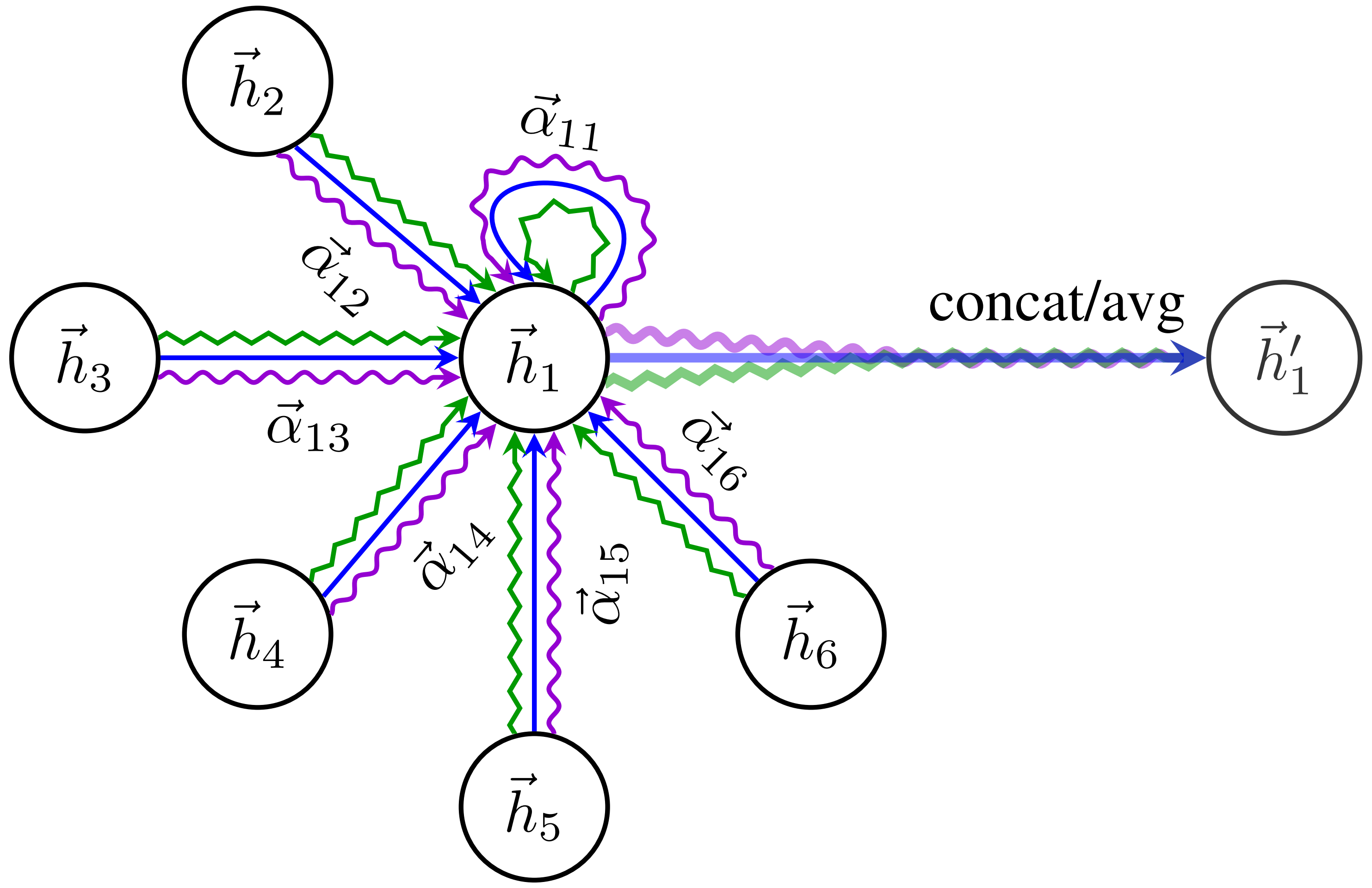
Graph Attention Network (GAT)

Graph Convolutional Network

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i) \cup i} \frac{1}{\sqrt{(d_i+1)(d_j+1)}} W^{(l)} h_j^{(l)}\right),$$

Graph Attention Network

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W h_j^{(l)}\right),$$



$$\alpha_{ij} = \mathrm{softmax}_j(a(Wh_i^{(l)}, Wh_j^{(l)})) = \frac{\exp\left(\mathrm{LeakyReLU}(b^T[Wh_i||Wh_j])\right)}{\sum_{k \in N(i)} \exp\left(\mathrm{LeakyReLU}(b^T[Wh_i||Wh_k])\right)},$$

decompose

Three steps

1. $$h'_i = Wh_i^{(l)}$$
2. $$h''_{ij} = \alpha_{ij} h'_j = \mathrm{softmax}_j(\mathrm{LeakyReLU}(b^T[h'_i||h'_j])h'_j$$
3. $$h_i^{(l+1)} = \sigma\left(\sum_j a_{ij} h''_{ij}\right),$$

Amortized (Variational) Inference (AVI)

- Variational inference analytical approximates to the posterior distribution of latent variables by making some assumptions about the form of posterior distribution. It is challenging for large datasets and non-conjugate models, because it separately updates each latent variable with a conjugate posterior distribution

$$\lambda_i = \lambda_i + \epsilon \nabla \mathrm{ELBO}(\lambda_i, x),$$

- To alleviate this issue, amortized variational inference (AVI) is developed to reformulate the variational inference as a prediction neural network which is shared (amortized) across all the data in the dataset

$$\lambda_i = f(x_i, \boxed{\phi}), \longleftarrow \text{Learnable parameter}$$

GAT as Semi-Amortized Inference of SBM

Stochastic Block Model (SBM)

$$q_{ij} = \frac{\theta_i \odot \theta_j}{\theta_i^T \theta_j} = \boxed{\left(\frac{\theta_i}{\theta_i^T \theta_j} \right)} \odot \boxed{\theta_j,}$$

traditional
inference

$$\theta_{ik} = \frac{\sum_j a_{ij} q_{ij}(k)}{\sum_i \theta_{ik}} = g_i \left(\sum_j a_{ij} q_{ij}(k) \right)$$

GAT as Semi-Amortized Inference of SBM

Stochastic Block Model (SBM)

$$q_{ij} = \frac{\theta_i \odot \theta_j}{\theta_i^T \theta_j} = \boxed{\left(\frac{\theta_i}{\theta_i^T \theta_j}\right)} \odot \boxed{\theta_j,}$$

traditional inference

$$\theta_{ik} = \frac{\sum_j a_{ij} q_{ij}(k)}{\sum_i \theta_{ik}} = g_i\left(\sum_j a_{ij} q_{ij}(k)\right)$$

Graph Attention Network (GAT)

$$\begin{aligned} h_i' &= \boxed{W} \boxed{h_i^{(l)}} \\ h_{ij}'' &= \boxed{\alpha_{ij}} h_j' \\ \boxed{\alpha_{ij}} &= \mathrm{softmax}_j(a(Wh_i^{(l)}, Wh_j^{(l)})) \\ &= \frac{\exp\left(\mathrm{LeakyReLU}(\boxed{b^T}[Wh_i || Wh_j])\right)}{\sum_{k \in N(i)} \exp\left(\mathrm{LeakyReLU}(b^T[Wh_i || Wh_k])\right)} \end{aligned}$$

amortized inference

$$h_i^{(l+1)} = \sigma\left(\sum_j a_{ij} h_{ij}''\right),$$

traditional inference

GAT as Semi-Amortized Inference of SBM

Stochastic Block Model (SBM)

$$q_{ij} = \frac{\theta_i \odot \theta_j}{\theta_i^T \theta_j} = \boxed{\left(\frac{\theta_i}{\theta_i^T \theta_j}\right)} \odot \boxed{\theta_j},$$

traditional inference

$$\theta_{ik} = \frac{\sum_j a_{ij} q_{ij}(k)}{\sum_i \theta_{ik}} = g_i\left(\sum_j a_{ij} q_{ij}(k)\right)$$

Latent Variable

Propagation Weight

Propagation

Graph Attention Network (GAT)

$$h_i' = \boxed{W}\boxed{h_i^{(l)}}$$

$$h_{ij}'' = \boxed{\alpha_{ij}} h_j'$$

amortized inference

$$\boxed{\alpha_{ij}} = \mathrm{softmax}_j(a(Wh_i^{(l)}, Wh_j^{(l)}))$$

$$= \frac{\exp\left(\mathrm{LeakyReLU}(\boxed{b^T}[Wh_i||Wh_j])\right)}{\sum_{k \in N(i)} \exp\left(\mathrm{LeakyReLU}(b^T[Wh_i||Wh_k]\right.}$$

$$h_i^{(l+1)} = \sigma\left(\sum_j a_{ij} h_{ij}''\right),$$

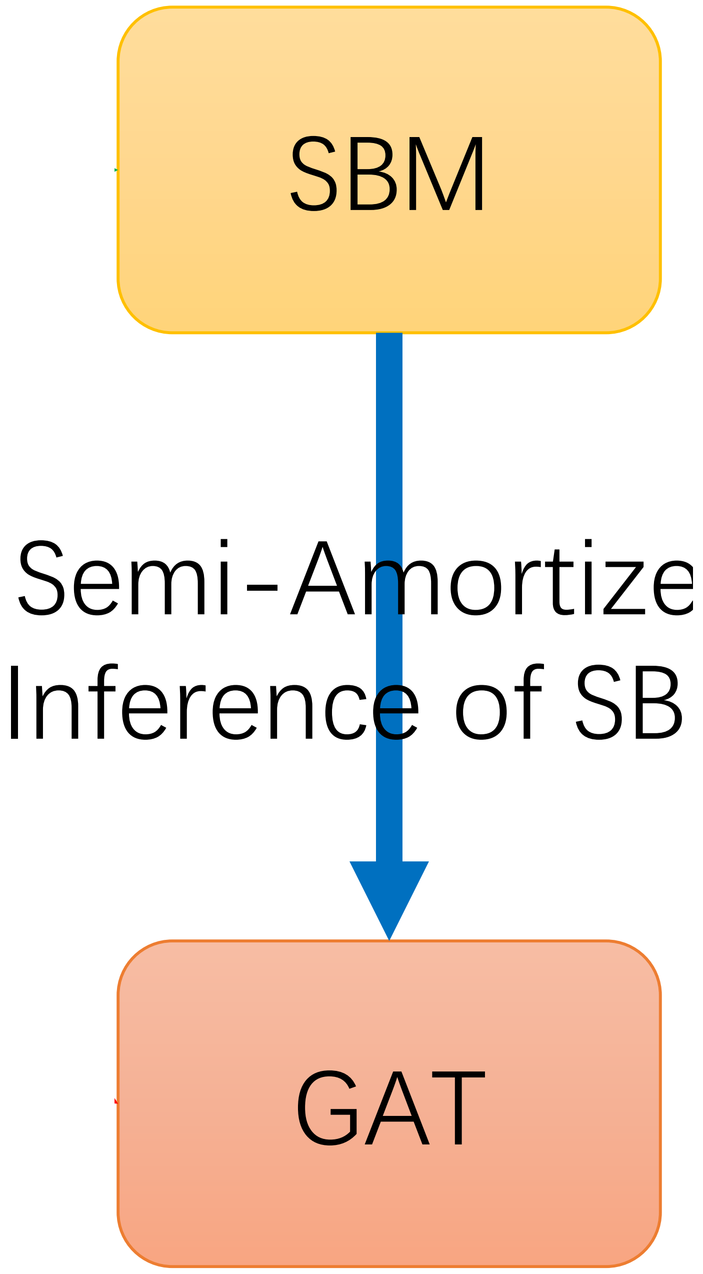
traditional inference

GAT can be regarded as the Semi-Amortized Inference (SAI) of SBM, which alternately performs the amortized inference and traditional inference.

GAT as Semi-Amortized Inference of SBM

Table 1: Comparisons between Stochastic Block Model and Graph Attention Network.

	Stochastic Block Model	Graph Attention Network
Latent Variable	θ_i (community membership)	h_i (node representation)
Initialization	random initialization	x_i (node attributes)
Amortized Mapping	without mapping	$h'_i = W h_i^{(l)}$ with learnable parameter W
Propagation Weight	$\frac{\theta_i}{\theta_i^T \theta_j}$	$\text{softmax}_j(\text{LeakyReLU}(b^T [h'_i \| \| h'_j])$
Propagation Weight Granularity	element-wise	edge-wise
Propagation Weight Learnability	without learnable parameters	with learnable parameter b
Propagated Information	θ_i (original latent variable)	h'_i (latent variable after mapping)
Weighted Information	$q_{ij} = \left(\frac{\theta_i}{\theta_i^T \theta_j}\right) \odot \theta_j$	$h''_{ij} = \text{softmax}_j(\text{LeakyReLU}(b^T [h'_i \| \| h'_j]) h'_j$
Propagation Rule	$\theta_i = g_i\left(\sum_j a_{ij} q_{ij}\right)$	$h_i^{(l+1)} = \sigma\left(\sum_j a_{ij} h''_{ij}\right)$



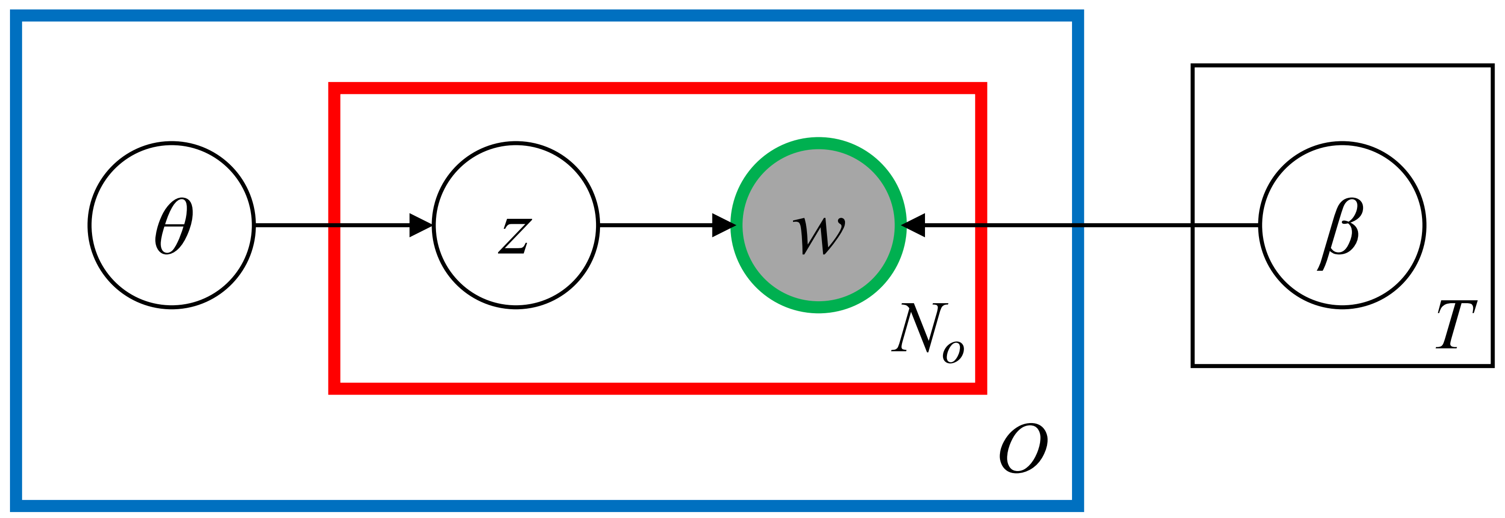
Probabilistic Latent Semantic Indexing (pLSI)

(1) Choose the number of word $N_o \sim \text{Poisson}(\eta_o)$ for document o;

(2) For each of the N_o words w_{on} in document o;

  (a) Choose a topic $z_{on} \sim \text{Multinomial}(\theta_o)$;

  (b) Choose a word $w_{on} \sim \text{Multinomial}(\beta_{z_{on}})$.



(a) The probabilistic graphical model of pLSI

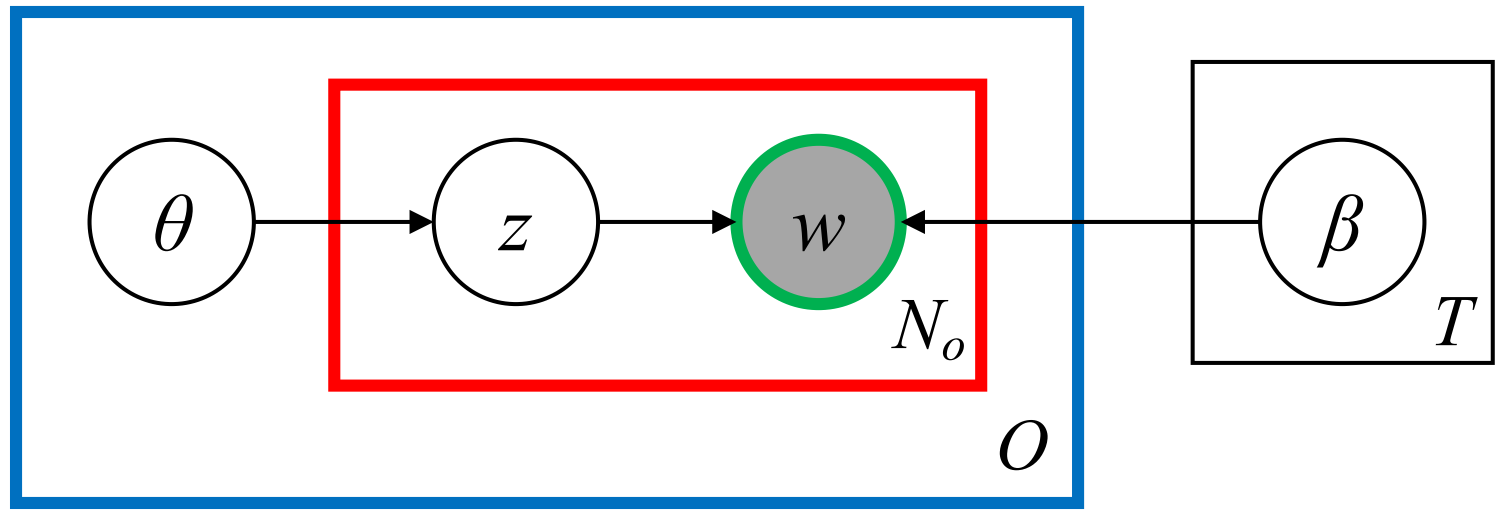
Probabilistic Latent Semantic Indexing (pLSI)

(1) Choose the number of word $N_o \sim \text{Poisson}(\eta_o)$ for document o;

(2) For each of the N_o words w_{on} in document o;

(a) Choose a topic $z_{on} \sim \text{Multinomial}(\theta_o)$;

(b) Choose a word $w_{on} \sim \text{Multinomial}(\beta_{z_{on}})$.



(a) The probabilistic graphical model of pLSI

$$P(O|\eta, \Theta, B) = \prod_{o=1}^{M} p(N_o|\eta_o) \prod_{n=1}^{N_o} \sum_{z_{on}=1}^{T} p(z_{on}|\theta_o) p(w_{on}|z_{on}, B)$$

$$\propto \prod_{o=1}^{M} \eta_o^{N_o} \exp(-\eta_o) \prod_{n=1}^{N_o} \sum_{z=1}^{T} \prod_{u=1}^{U} (\theta_{oz}\beta_{zu})^{w_{on}^{u}}. \quad (19)$$

$$= \prod_{o=1}^{M} \eta_o^{N_o} \exp(-\eta_o) \prod_{u=1}^{U} \frac{(\sum_{z=1}^{T} \theta_{oz}\beta_{zu})^{n_{ou}}}{n_{ou}!}.$$

$$= \prod_{o=1}^{M} \prod_{u=1}^{U} \exp\left(-\sum_{z=1}^{T} \theta'_{oz}\beta_{zu}\right) \frac{(\sum_{z=1}^{T} \theta'_{oz}\beta_{zu})^{n_{ou}}}{n_{ou}!}.$$

$$n_{ou} = \sum_{n=1}^{N_o} w_{on}^{u}$$

the frequency of word u appearing in document o

$$\boxed{\eta_o^{N_o} \sum_{u=1}^{U} \sum_{z=1}^{T} \theta_{oz}\beta_{zu} = \eta_o^{N_o} \sum_{z=1}^{T} \theta_{oz} = \eta_o^{N_o},}$$

$$\theta'_{oz} = \eta_o \theta_{oz},$$

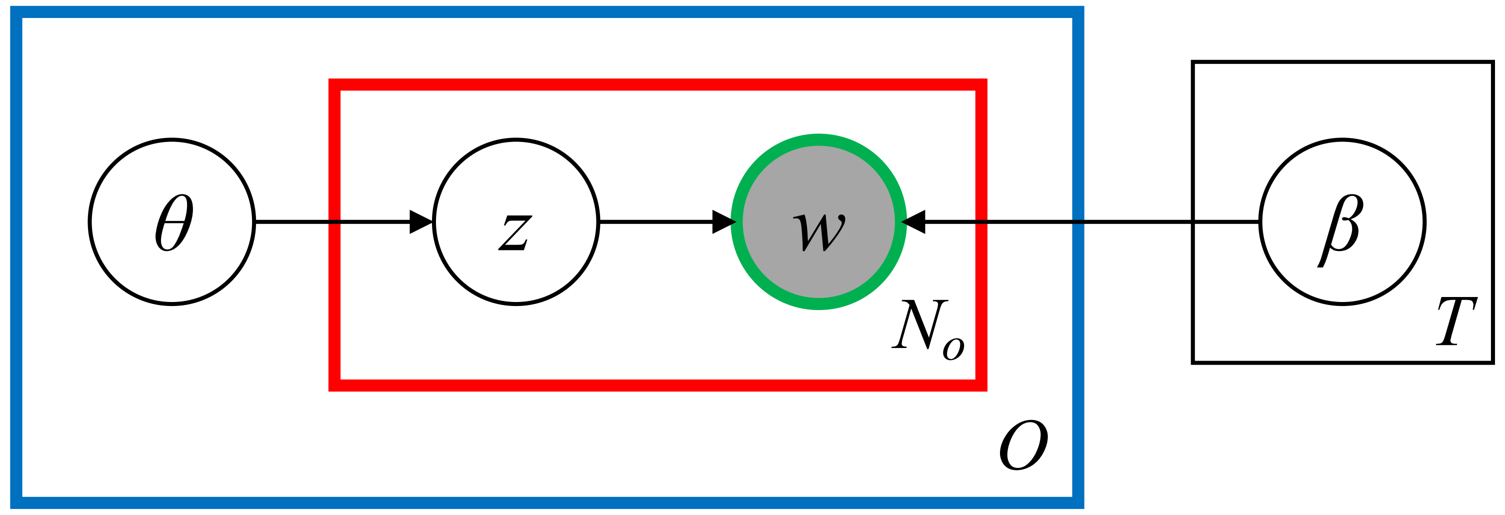
Topic Modeling as SBM on Bi-partite Graph

Probabilistic Latent Semantic Indexing (pLSI)

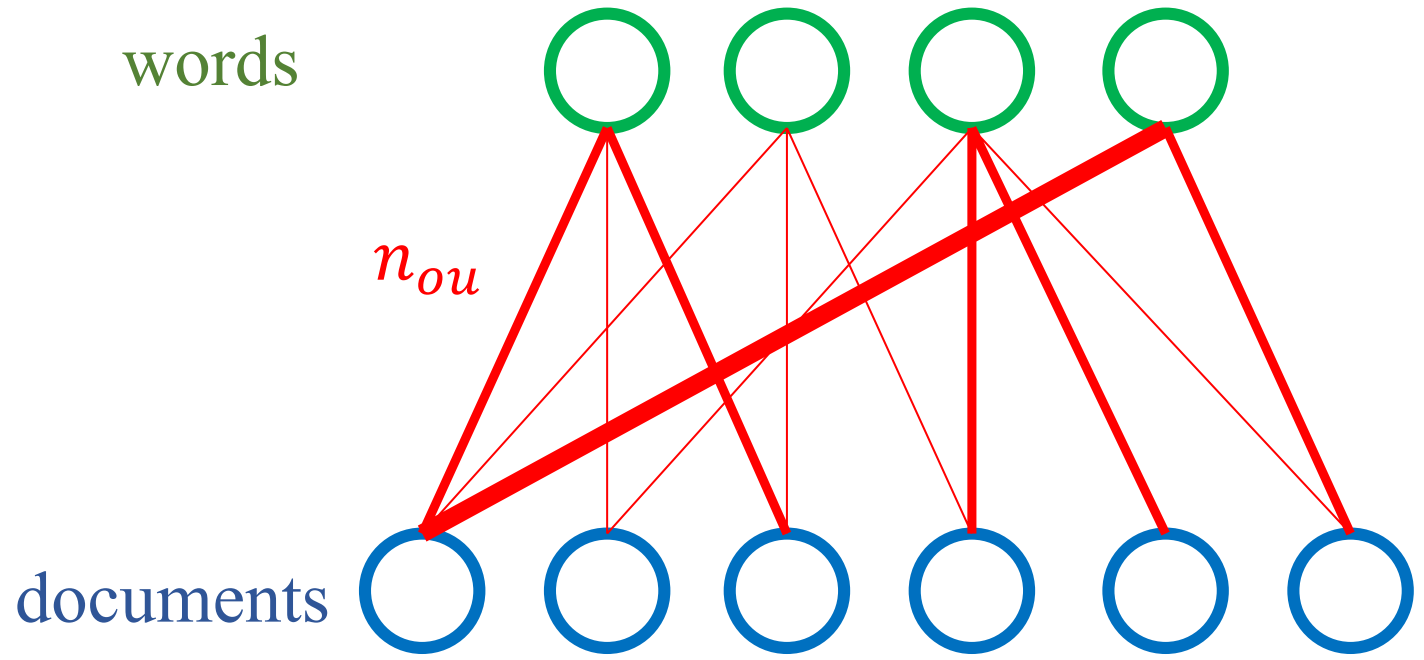
$$P(G|\Theta) = \prod_{i<j} \frac{\left(\sum_k \theta_{ik}\theta_{jk}\right)^{\boxed{a_{ij}}}}{a_{ij}!} \exp\left(-\sum_k \theta_{ik}\theta_{jk}\right) \prod_i \frac{(\sum_k \theta_{ik}\theta_{ik})^{a_{ii}/2}}{(a_{ii}/2)!} \exp\left(-\frac{1}{2}\sum_k \theta_{ik}\theta_{ik}\right).$$

Stochastic Block Model (SBM)

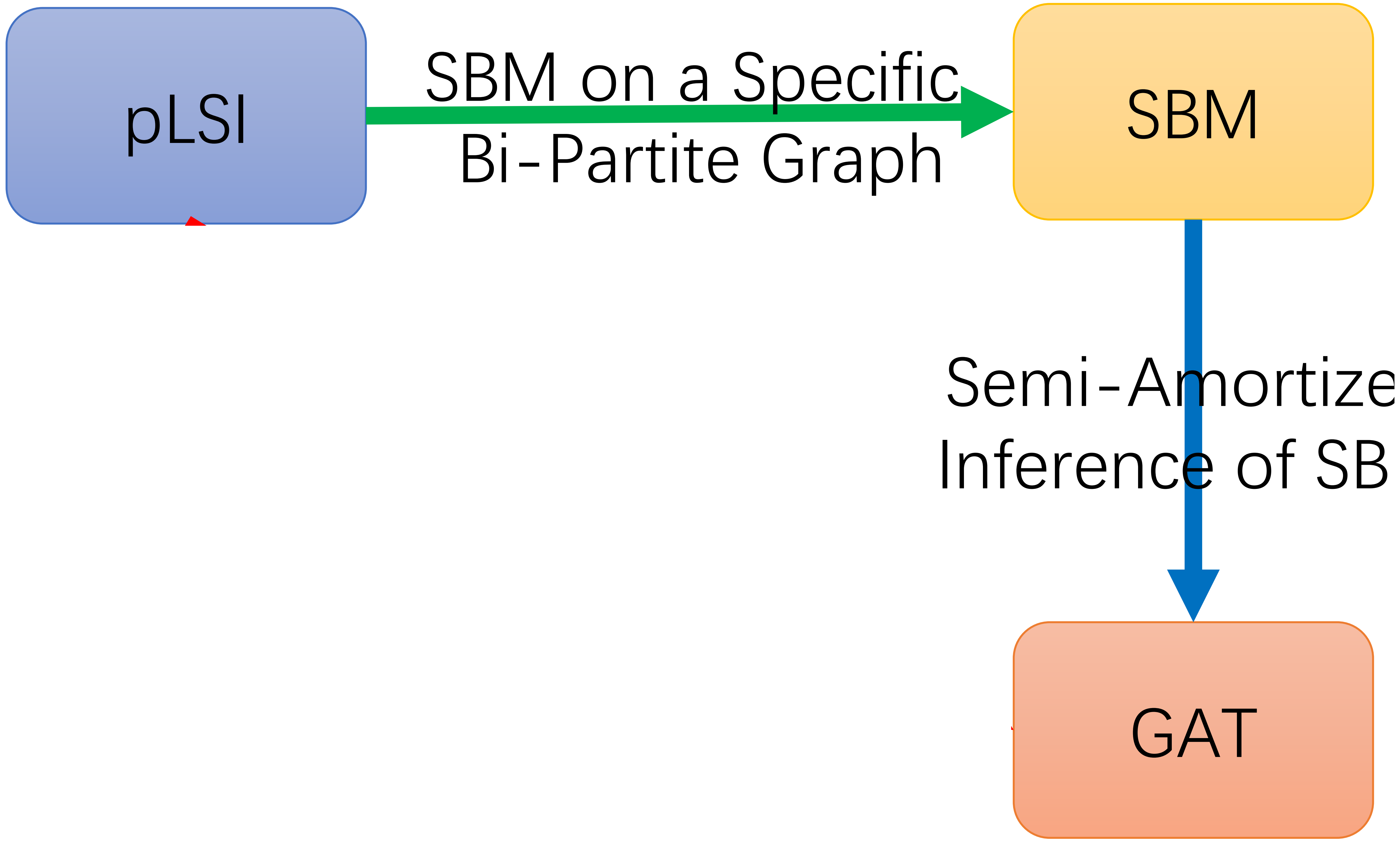
$$P(O|\Theta, B) = \prod_{o=1}^{M}\prod_{u=1}^{U} \exp\left(-\sum_{z=1}^{T} \theta'_{oz}\beta_{zu}\right) \frac{(\sum_{z=1}^{T} \theta'_{oz}\beta_{zu})^{\boxed{n_{ou}}}}{n_{ou}!}.$$

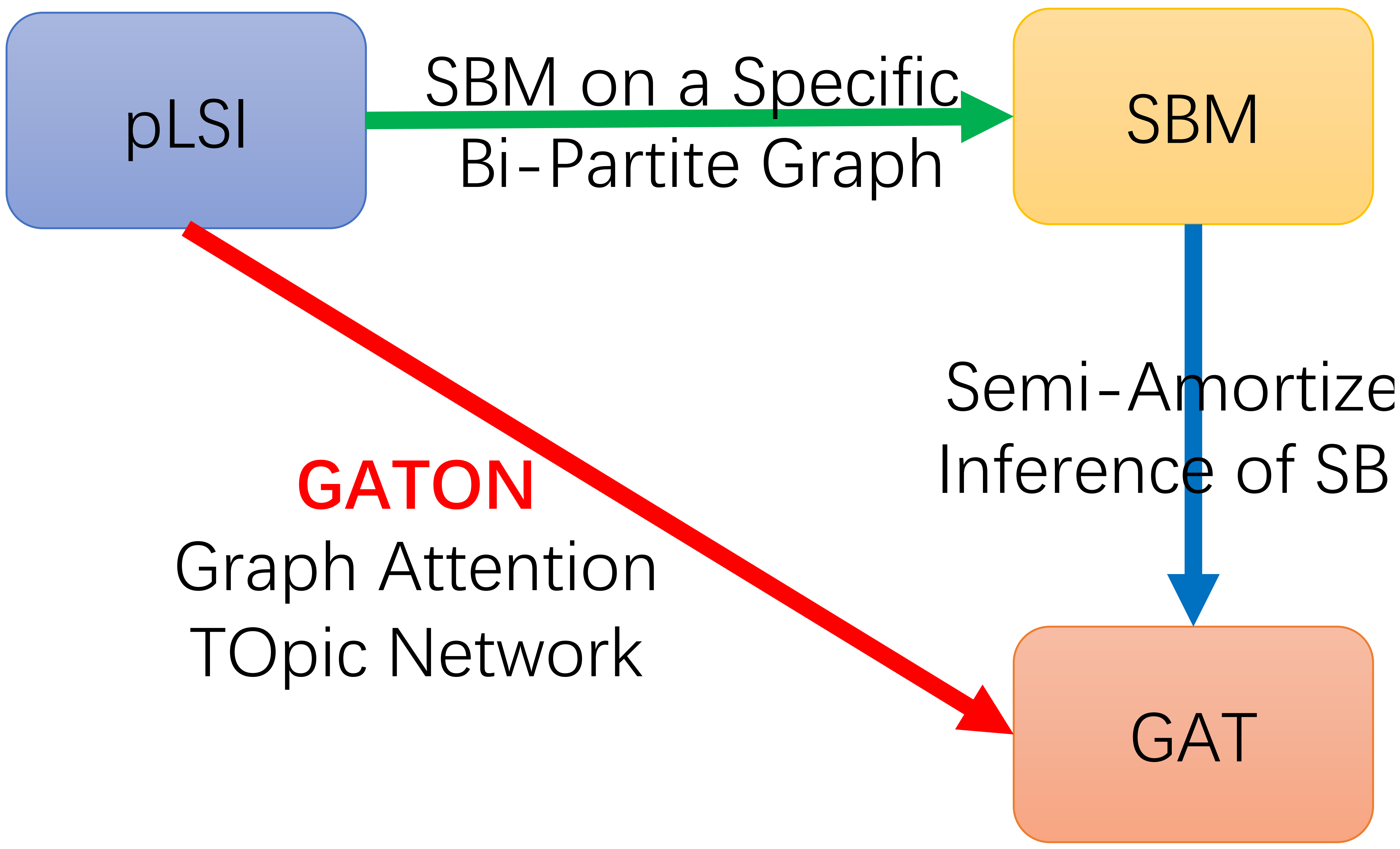


(a) The probabilistic graphical model of pLSI



(b) The bi-partite graph of pLSI


pLSI
SBM on a Specific
Bi-Partite Graph
SBM
Semi-Amortize
Inference of SB
GAT

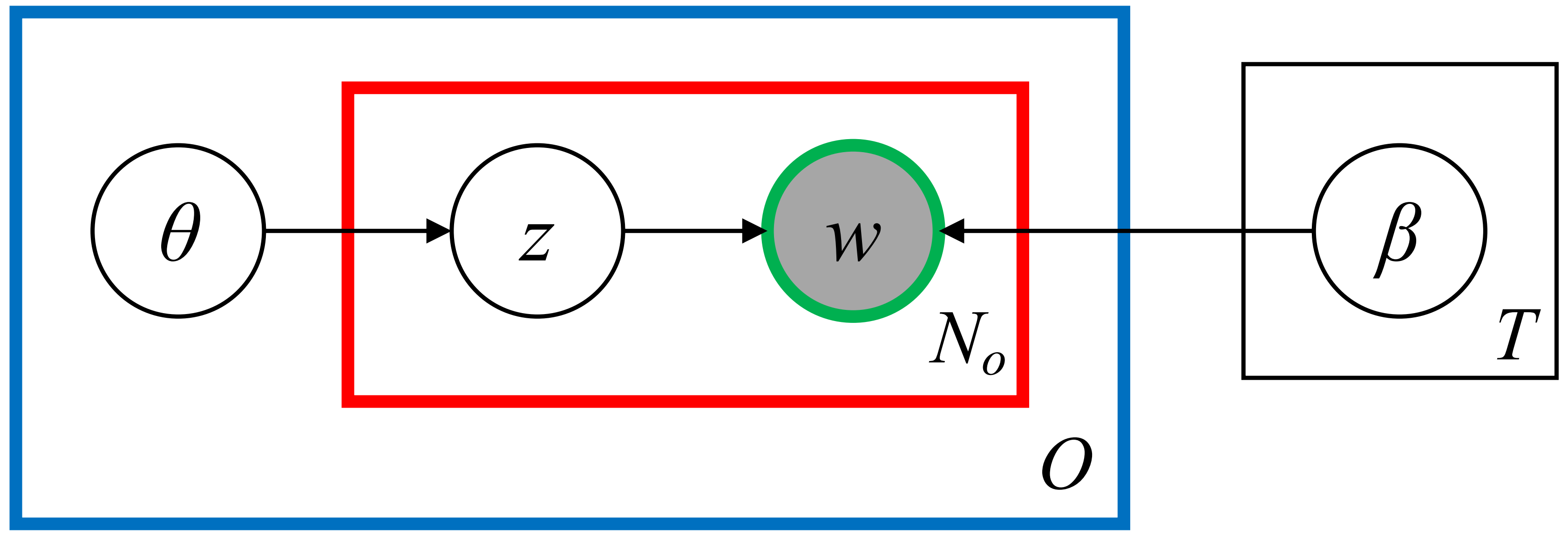

pLSI
SBM on a Specific
Bi-Partite Graph
SBM
Semi-Amortize
Inference of SB
GATON
Graph Attention
TOpic Network
GAT

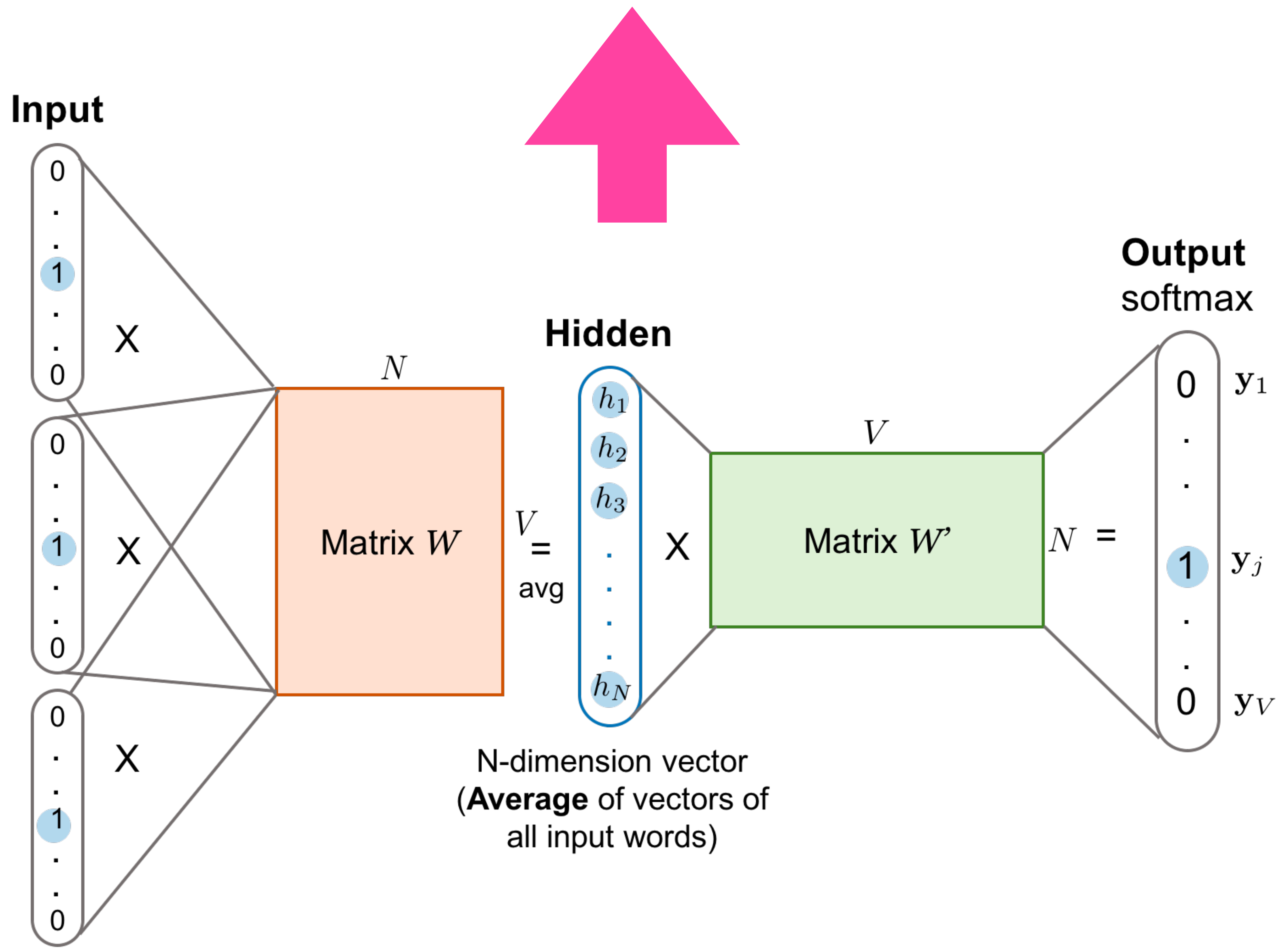
Graph Attention TOpic Network (GATON)

Issue: Latent Dirichlet Allocation alleviates the overfitting issue by introducing Dirichlet priors for latent variables, but it fails to capture the rich topical correlations among topics.

Intent: Overcome the overfitting issue of pLSI by exploiting the word embedding.

Question: How to integrate word embedding into generative topic modeling?





Graph Attention TOpic Network (GATON)

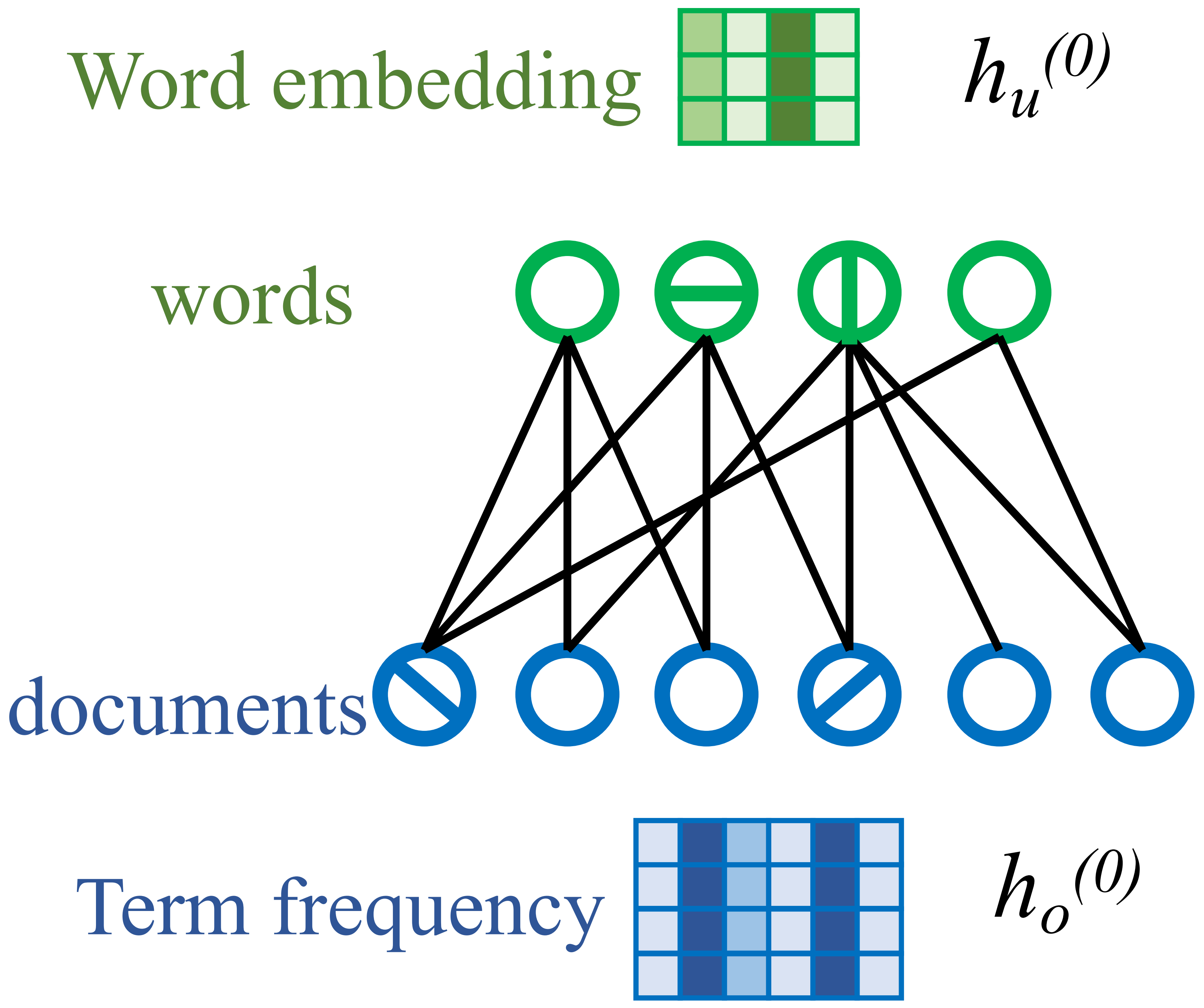
Issue: Latent Dirichlet Allocation alleviates the overfitting issue by introducing Dirichlet priors for latent variables, but it fails to capture the rich topical correlations among topics.

Intent: Overcome the overfitting issue of pLSI by exploiting the word embedding.

Question: How to integrate word embedding into generative topic modeling?

Answer: Graph Convolutional Networks!!!

Configuration of GATON



Graph Attention TOpic Network (GATON)

1 Mapping

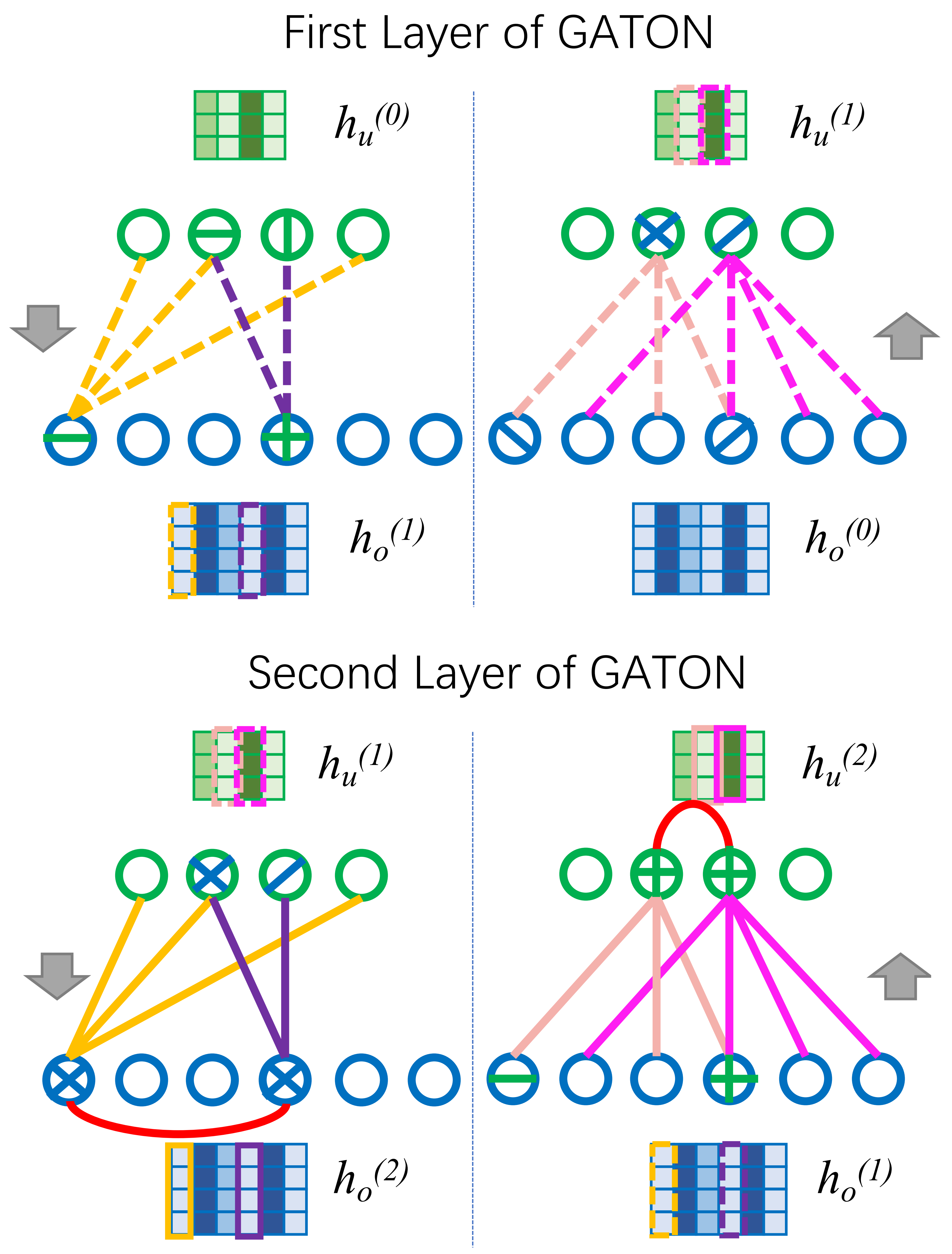
$$\hat{h}_u^{word} = W^{word} x_u^{word},$$
$$\hat{h}_o^{document} = W^{document} x_o^{document}.$$

2 Propagation weights

$$\alpha_{o\to u} = \frac{\exp\left(\text{LeakyReLU}(b_{o\to u}^T[\hat{h}_o || \hat{h}_u])\right)}{\sum_{t\in N(o)} \exp\left(\text{LeakyReLU}(b_{o\to u}^T[\hat{h}_o || \hat{h}_t])\right)},$$
$$\alpha_{u\to o} = \frac{\exp\left(\text{LeakyReLU}(b_{u\to o}^T[\hat{h}_u || \hat{h}_o])\right)}{\sum_{z\in N(u)} \exp\left(\text{LeakyReLU}(b_{u\to o}^T[\hat{h}_u || \hat{h}_z])\right)},$$

3 Propagation

$$h_o = \sigma\left(\sum_{t\in N(o)} \alpha_{u\to o} \hat{h}_t\right),$$
$$h_u = \sigma\left(\sum_{z\in N(u)} \alpha_{o\to u} \hat{h}_z\right),$$



Evaluations

Table 3: Document classification performances on datasets.

Dataset	20News			Reuters		
Metrics	Prec.	Recall	F1	Prec.	Recall	F1
NMF	0.704	0.701	0.697	0.911	0.877	0.891
pLSI	0.722	0.712	0.709	0.919	0.896	0.906
LDA	0.727	0.722	0.719	0.888	0.870	0.879
Gauss-LDA	0.309	0.265	0.227	0.462	0.315	0.353
LF-LDA	0.716	0.714	0.709	0.893	0.591	0.661
CLM	0.825	0.818	0.816	0.944	0.916	0.929
TWE	0.525	0.466	0.437	0.794	0.512	0.626
PV-DBOW	0.510	0.491	0.459	0.755	0.505	0.549
PV-DM	0.428	0.386	0.361	0.681	0.434	0.507
TopicVec	0.713	0.713	0.712	0.925	0.921	0.922
MeanWV	0.704	0.703	0.701	0.920	0.896	0.905
TV+Mean	0.718	0.715	0.716	0.922	0.916	0.916
GATON-C	0.822	0.803	0.812	**0.975**	**0.979**	**0.977**
GATON-S	**0.859**	**0.842**	**0.850**	0.944	0.937	0.940
GATON-G	0.716	0.767	0.741	0.914	0.896	0.905

Table 2: Topic coherence performances on both datasets.

Dataset	20News			Reuters		
#Top-words	5	10	20	5	10	20
NMF	-18.05	-85.53	-417.19	-11.28	-66.41	-335.61
pLSI	-15.15	-78.59	-365.69	-13.22	-70.07	-333.57
LDA	-15.30	-80.48	-368.82	-12.09	-69.80	-352.29
Gauss-LDA	-19.45	-94.52	-435.90	-24.22	-108.45	-478.43
LF-LDA	-16.58	-78.54	-385.73	-13.26	-71.35	-369.00
CLM	-11.62	-60.30	-282.79	-11.48	-63.08	-313.45
GATON-C	**-10.17**	**-55.82**	-245.29	**-10.06**	-57.46	-285.90
GATON-S	-10.92	-55.98	**-244.73**	-10.35	**-56.75**	**-277.34**
GATON-G	-11.55	-58.13	-285.91	-11.66	-61.03	-299.35

Table 4: Word embedding performances on 20News dataset.

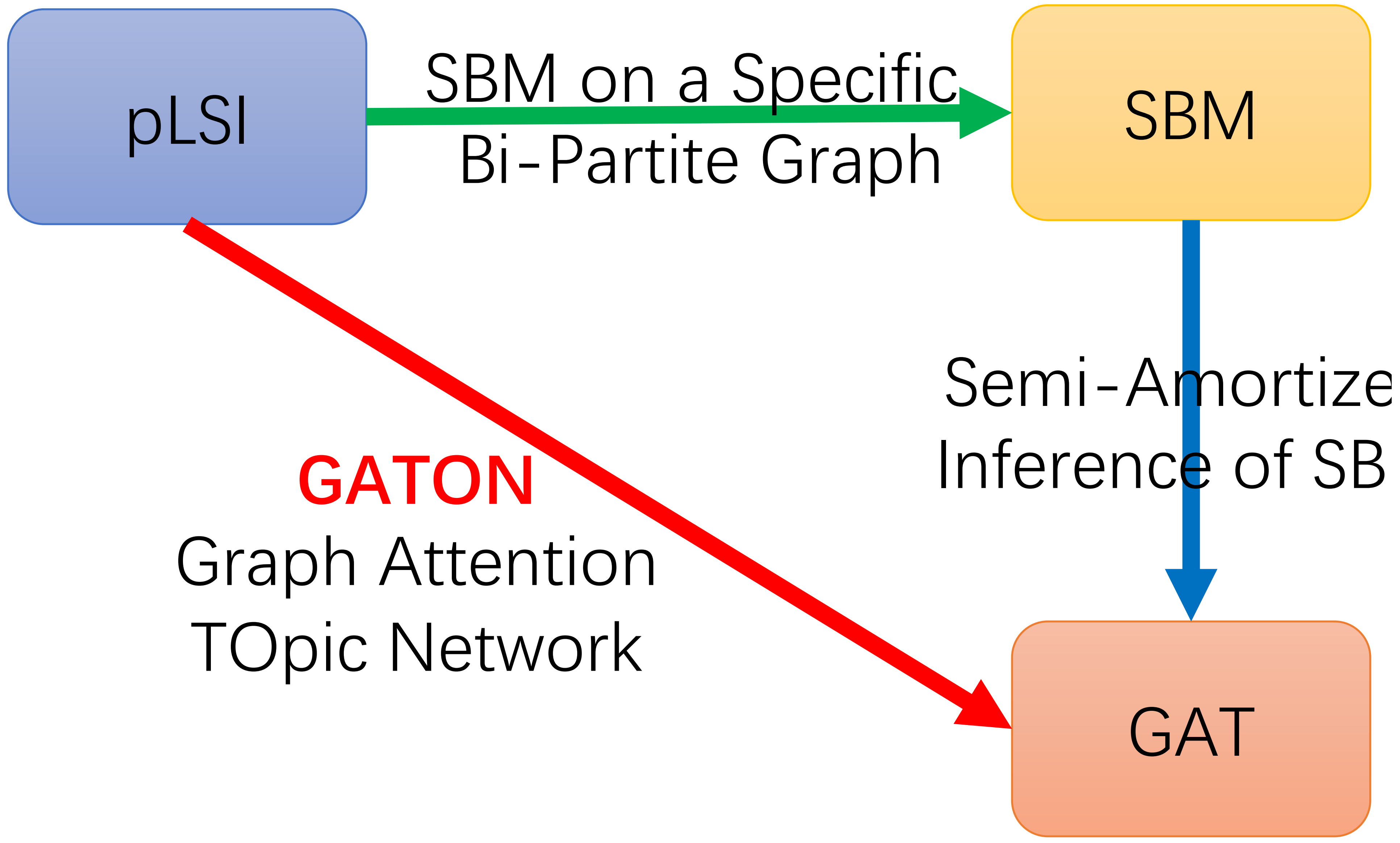
	W353	WRel	WSim	Men	Turk	SimL	Rare
SPPMI	0.461	0.444	0.465	0.444	0.551	0.131	0.245
SPPMI+SVD	0.451	0.435	0.449	0.426	0.489	0.166	0.349
PV-DBOW	0.477	0.442	0.486	0.449	0.488	0.139	0.285
TWE	0.317	0.231	0.407	0.190	0.260	0.084	0.184
CLM	0.526	0.486	0.550	0.477	0.525	0.189	0.411
CBOW	0.488	0.451	0.494	0.432	0.529	0.151	0.407
Skip-Gram	0.492	0.479	0.473	0.456	0.512	0.155	0.407
GloVe	0.300	0.279	0.320	0.192	0.268	0.049	0.230
GATON-C	**0.563**	**0.531**	**0.579**	0.505	**0.569**	0.232	0.470
GATON-S	0.552	0.527	0.573	**0.516**	0.560	**0.242**	**0.473**
GATON-G	0.461	0.405	0.460	0.352	0.435	0.154	0.358

Conclusions

- We propose a novel approach to overcome the overfitting issue in topic modeling by adopting amortized inference, with the word embedding as input, to significantly reduce the number of to-be-estimated parameters.
- We reveal the connections between the generative stochastic block model (SBM) and graph neural networks (GNNs), especially graph attention network (GAT). GAT is equivalent to the Semi- Amortized inference algorithm of SBM.
- We observe that the probabilistic latent semantic indexing (pLSI) can be seen as SBM on a specific bi-partite graph, where the documents and the words are the two kinds of the nodes, respectively.
- To relax the i.i.d. data assumption of vanilla amortized inference, we pioneer to propose a novel graph neural network model, named Graph Attention TOpic Network (GATON), for correlated topic modeling. GATON, which constructs the graph topology with the bi-partite graph of documents and words, explores the topic structure by convolving the node attributes over the graph with an attention mechanism.

thank you

THE WEB CONFERENCE | TAIPEI 2020



Graph Attention Topic Modeling Network