EMORY UNIVERSITY | Department of Computer Science

The 28th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)

Source Localization of Graph Diffusion via Variational Autoencoders for Graph Inverse Problems

Presented by: Chen Ling

chen.ling@emory.edu

lingchen0331.github.io

Joint work with Junji Jiang, Junxiang Wang, and Liang Zhao

Aug 14, 2022

Outline

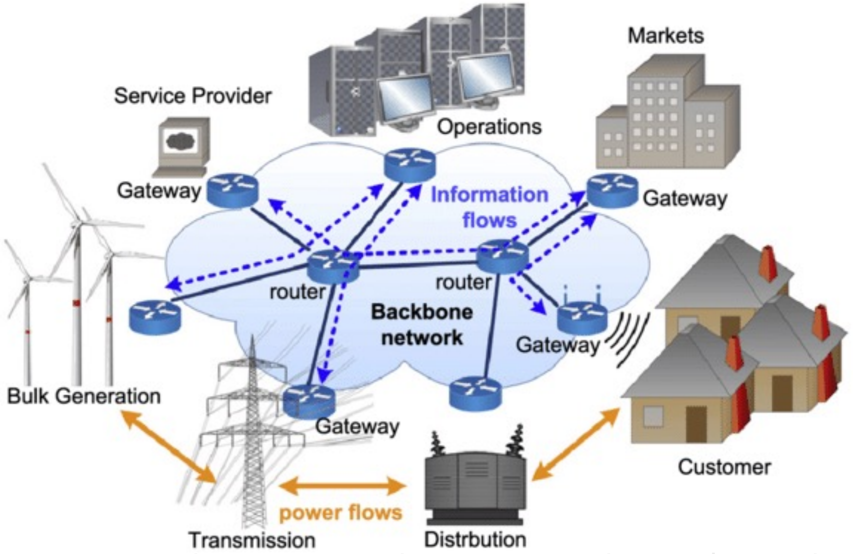
- Background
- Proposed Method – SL-VAE
- Experiment
- Concluding Remarks

Graph Information Diffusion

- As a prevalent data structure, graph can represent various network-structured data.
- The ubiquity of networks has also made us vulnerable to various network risks.

(a) Social Network

(b) Computer Network



(c) Smart Grid Network

Graph Information Diffusion

- As a prevalent data structure, graph can represent various network-structured data.
- The ubiquity of networks has also made us vulnerable to various network risks.

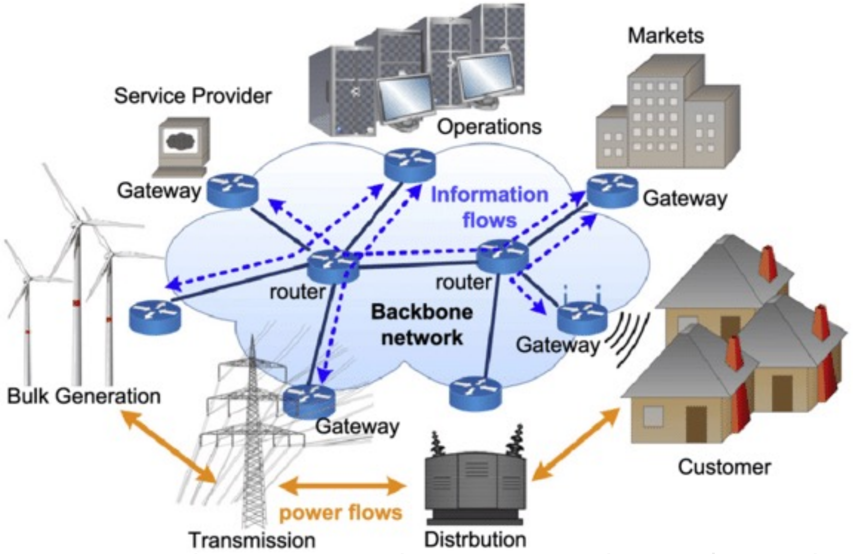
The spread of misinformation

(a) Social Network

The spread of computer virus

(b) Computer Network

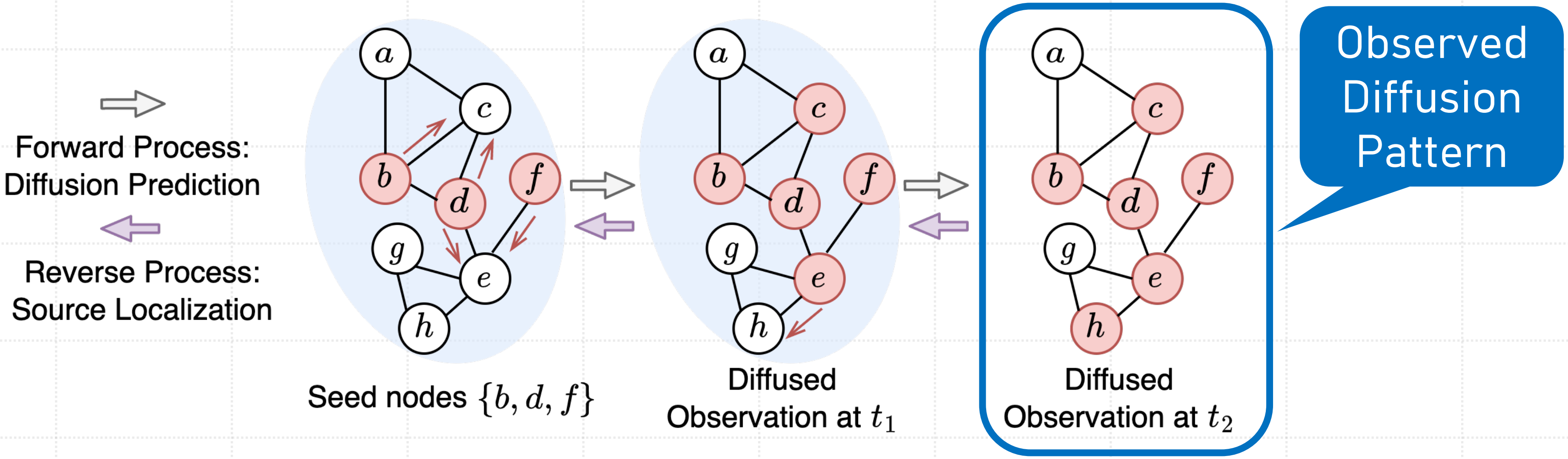
The electric blackout



(c) Smart Grid Network

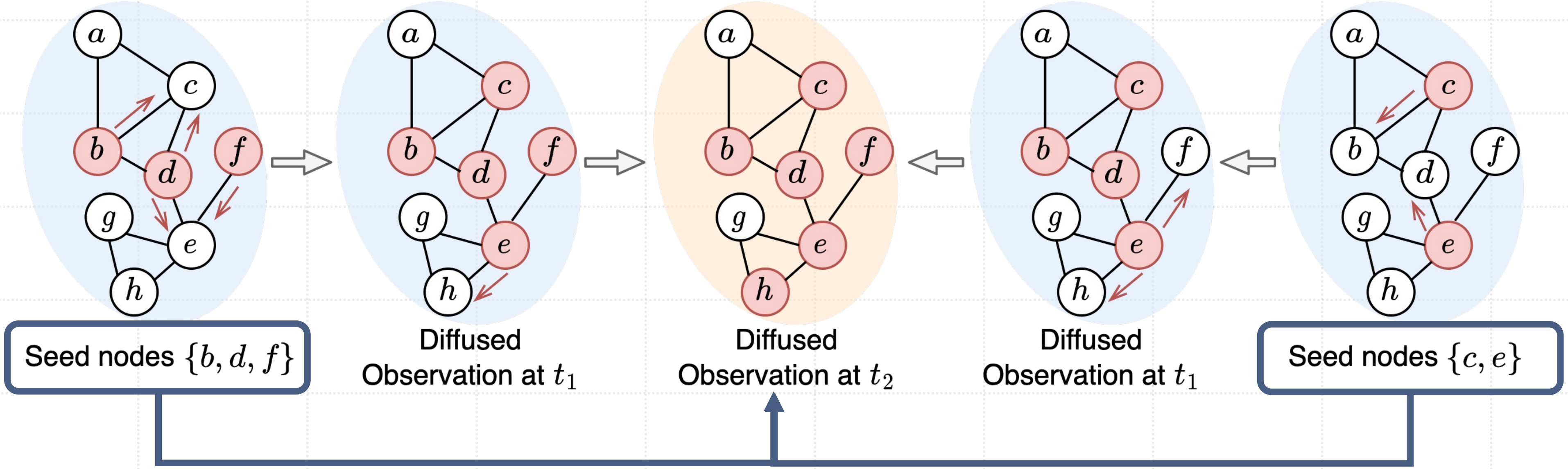
Graph Diffusion and Source Localization

- Information diffusion occurs on graph and has been studies for decades.
- **Information Diffusion Estimation** aims to predict the future graph diffusion patterns given source nodes.
- Its inverse process – **source localization** aims to locate diffusion sources given the observed graph diffusion pattern.



Limitations of Existing Works

- Existing works in graph source localization are proposed as deterministic algorithm to compute sources directly from the diffusion observation.
- However, multiple set of source nodes in graph information diffusion may lead to the same diffusion cascade pattern.



Potential Solution and Challenges

- As an ill-posed inverse problem of diffusion prediction, source localization needs to locate the diffusion source from all feasible sources.

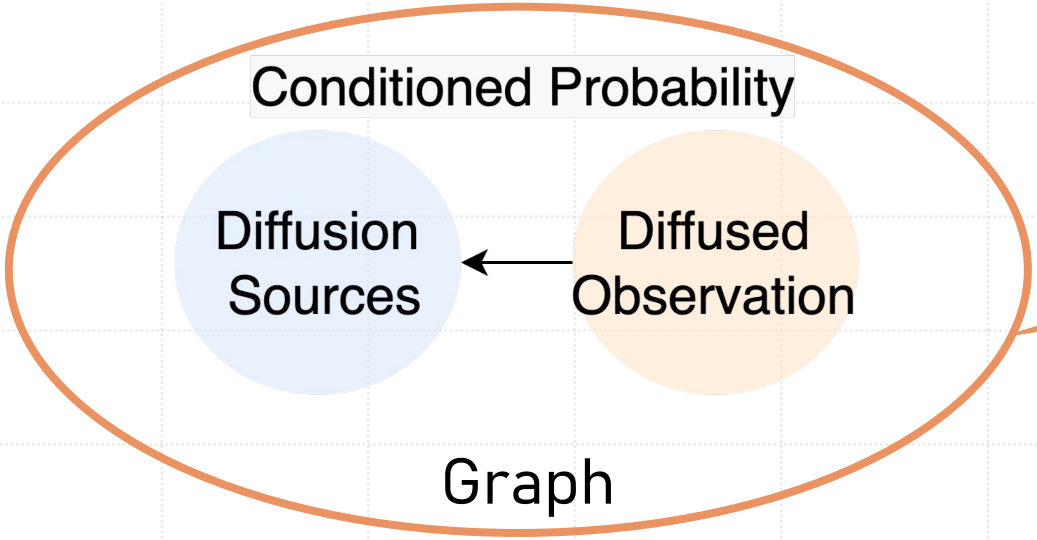
Question: Can we quantify the uncertainty in the ill-posed source localization problem?

Solution: Deep generative models can characterize the latent distribution of sources so to quantify the uncertainty.

Any Potential Challenges?

Challenge 1: Quantifying Uncertainty in Graph Source Localization

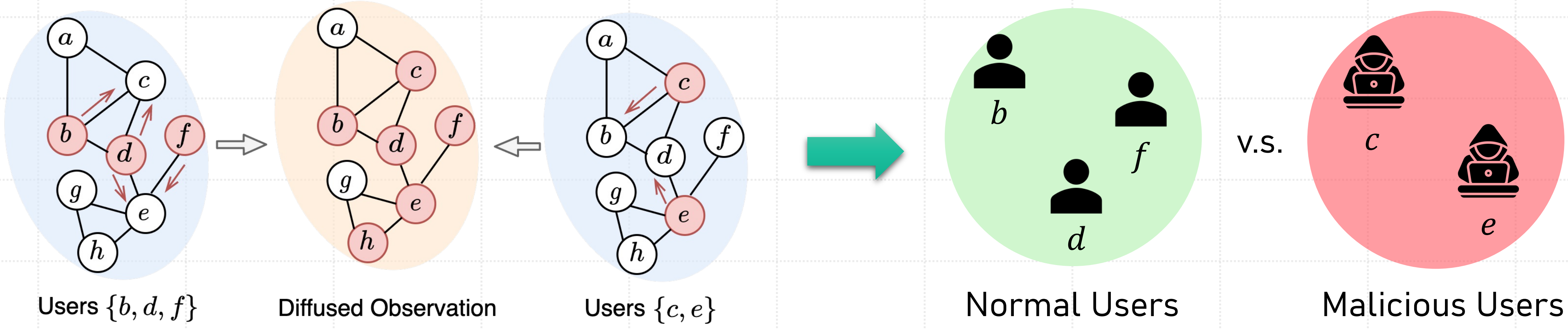
- Quantifying the uncertainty requires building a probabilistic model between the source and observation so that one can estimate the “optimal” source to the observation.
- However, approximating the conditional probability needs to consider the graph topology since graph topology is essential and dominates the diffusion process.



1. Graph topology determines the diffusion process.
2. the probability distribution of graph data is hard to optimize

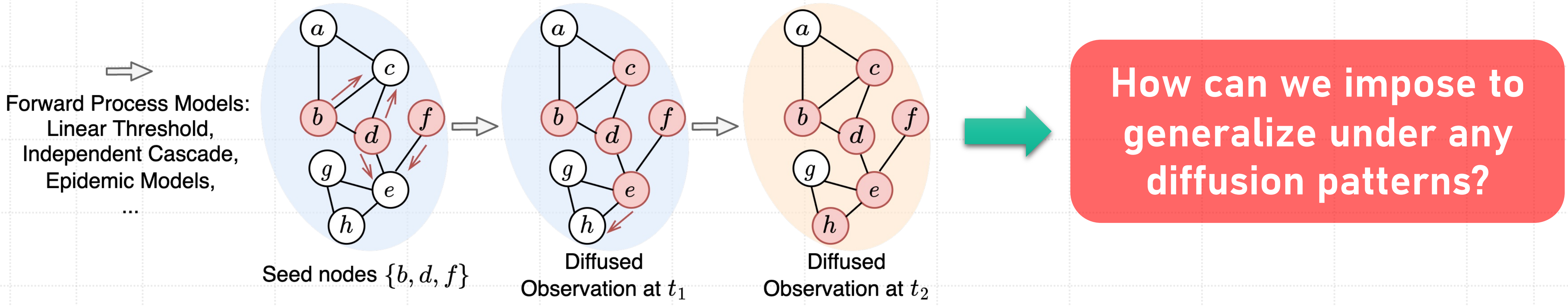
Challenge 2: Characterizing the Intrinsic Patterns of Diffusion Sources

- Characterizing the patterns of diffusion sources also conditioned on the intrinsic nature of the nodes and their connections.
- Such information is apart from the diffused observations but can predominantly help determine the sources.
- Consider a rumor source detection task in a social network:



Challenge 3: Generalizing under any underlying diffusion patterns

- Most existing source localization methods are tailored for specific diffusion processes such as Linear Threshold, Independent Cascade, and Epidemic models.
- Apart from the prescribed diffusion processes, the diffusion principle in real-world scenarios are more complex and cannot be described by mathematical models.



Outline

Background

Proposed Method – SL-VAE

Experiment

Concluding Remarks

SL-VAE: Variational Inference

- The *diffusion source* set is defined over V as $x = \{0,1\}^{|V|}$, $x_i = 1$ or 0 denotes seed node or not.
- The *probability* of each node being infected is defined as $y = [0,1]^{|V|}$ (diffused observation).
- We Utilize the Maximum A Posteriori (MAP) to estimate the optimal diffusion source:

$$\tilde{x} = \arg\max_x p(y|x,G) \cdot p(x) = \arg\max_x p(x,y|G)$$

- Motivated from *deep generative models*: we map *high-dimensional* and *intractable* $p(x)$ to latent variable z in lower dimension.
- The latent variable z is obtained by the posterior $p(z|x,y,G)$.
- Due to the intractability of $p(x)$, the approximate posterior $q_\phi(z|x,y,G)$ is adopted to infer z:

$$\begin{aligned} q_\phi(z|x,y,G) &= \min_\phi KL\big[q_\phi(z|x,y,G)||p(z|x,y,G)\big] \\ &= \min_\phi \big[\mathbb{E}_{q_\phi}[\log q_\phi(z|x,y,G)] - \mathbb{E}_{q_\phi}[\log p(x,y,G,z)] + \boxed{\log p(x,y,G)}\big] \end{aligned}$$

Intractable Joint Distribution

SL-VAE: Training Objective

- We utilize Evidence Lower BOund (ELBO) to approximate the posterior $q_\phi(z|x,y,G)$:

minimize the KL divergence by minimizing the negative ELBO

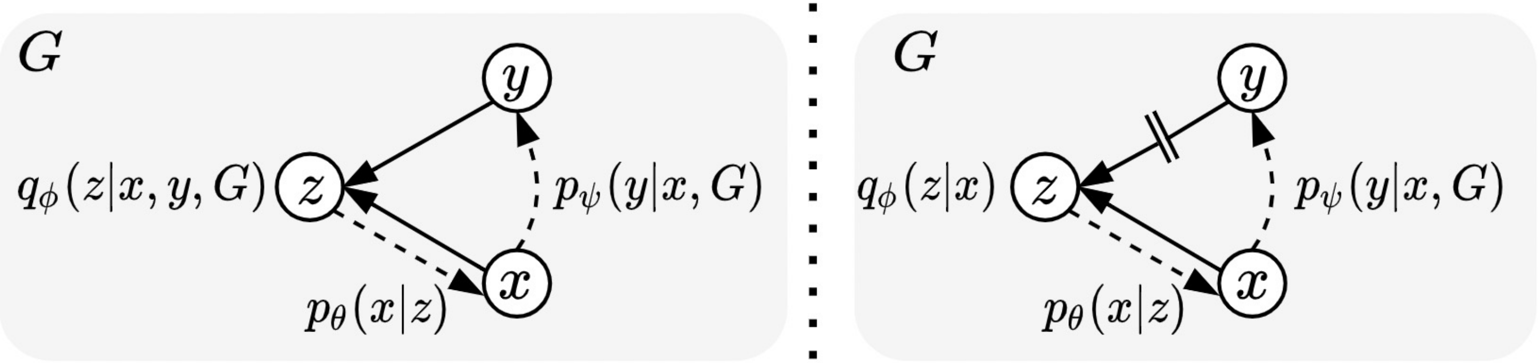
$$ELBO = \mathbb{E}_{q_\phi}[\log\ p(x,y,G,z)] - \mathbb{E}_{q_\phi}[\log\ q_\phi(z|x,y,G)]$$

$$-ELBO = -\mathbb{E}_{q_\phi}[\boxed{\log\ p_\theta(x,y,G|z)}] + KL[q_\phi(z|x,y,G)\|\boxed{p(z)}]$$

Graph Prior Distribution

- In most information diffusion estimation models, the diffused observation y is only determined by the diffusion source x under the graph G:

$$\boxed{\log p_\theta(x,y,G|z)} = \log[p_\psi(y|x,G)] + \log[p_\theta(x|z)]$$



SL-VAE: Training Objective

- We minimize the negative ELBO to jointly train three components:
  - An inference network – encoder $q_\phi(z|x)$ approximates the posterior.
  - A generation network – decoder $p_\theta(x|z)$ decodes information from latent variable $z \sim q_\phi(z|x)$.
  - A *forward information diffusion estimation model* $p_\psi(y|x, G)$ that takes the diffusion source $x \sim p_\theta(x|z)$ and graph G to predict the infecting probability y of each node.

Information Estimation Model

Generation Network (decoder)

Inference Network (encoder)

$$-ELBO = -\mathbb{E}_{q_\phi}\Big[\underbrace{\log\ p_\psi(y|x, G)}_{\text{Forward Loss}} + \underbrace{\log\ p_\theta(x|z)}_{\text{Reconstruction Loss}}\Big] + \underbrace{KL[p(z)||q_\phi(z|x)]}_{\text{Regularization}}$$

Training Objective: Monotonic Constraint on Information Diffusion

- In addition to optimize the variational inference framework (negative ELBO), the information diffusion needs to respect the monotone increasing property:

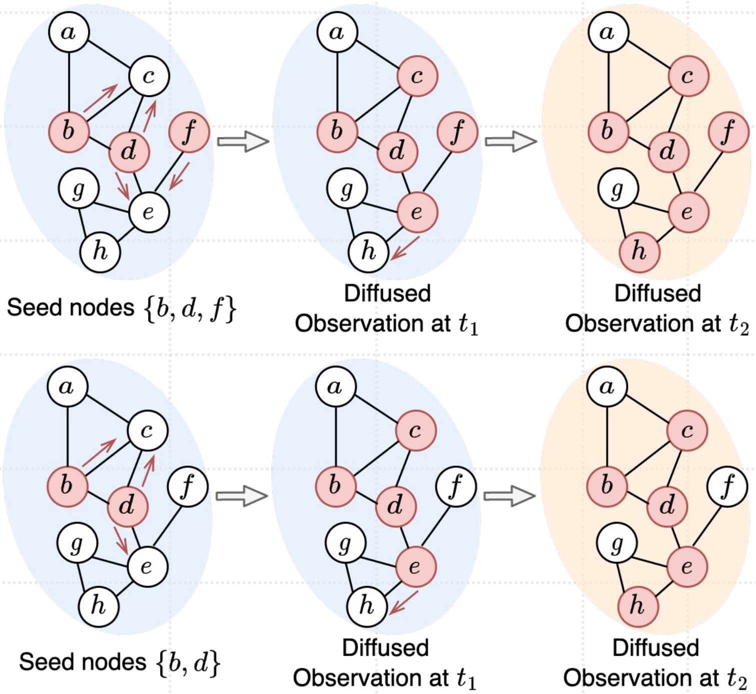
$$y^{(i)} \geq y^{(j)}, \forall\, x^{(i)} \supseteq x^{(j)}$$

- The assumption is applied in tasks, including influence maximization and information diffusion estimation.
- We update the objective function and optimize the following *constrained objective*:

$$-ELBO = -\,\mathbb{E}_{q_\phi}\left[\log\ p_\psi(y|x,G) + \log\ p_\theta(x|z)\right] + KL[q_\phi(z|x)||p(z)]$$

$$\boxed{\text{s.t. } y^{(i)} \geq y^{(j)}, \forall\, x^{(i)} \supseteq x^{(j)}}$$

$$-ELBO = -\,\mathbb{E}_{q_\phi}\left[\log\ p_\psi(y|x,G) + \log\ p_\theta(x|z)\right]$$
$$+\,KL[q_\phi(z|x)||p(z)] + \lambda\|\max(0, y^{(j)} - y^{(i)})\|_2^2, \forall\, x^{(i)} \supseteq x^{(j)}$$



Predicting the Diffusion Sources

- The training phase aims to learning the generator $p_\theta(x|z)$, and a (arbitrary) diffusion estimater $p_\psi(y|x,G)$ to predict the optimal diffusion source $\tilde{x}$ given the observation y.
- Since $p(x)$ is determined by $p(z)$ such that $p(x) = \sum_z p_\theta(x|z)p(z)$, we may marginalize $p(z)$ and optimize the following MAP problem to find the optimal diffusion source.

$$\max_x \left[p_\psi(y|x,G) \cdot \sum_z p_\theta(x|z) \cdot p(z)\right]$$

- However, optimizing the above objective function could be problematic:

1. Marginalizing $p(z)$ would require lots of sampling to match the desired distribution.
2. The objective function does not contain information of the observed diffusion sources.

Diffusion Source Prediction: Objective Function

- During the VAE training, the latent random variable z is sampled from the encoder $q_\phi(z|x)$, where all the parameters are obtained through stable functions of $\hat{x}$ in the training set.
- we could sample z from the posterior distribution $q_\phi(z|x)$ instead of $p(z)$ if the VAE can approximate the posterior $q_\phi(z|x)$ to match the prior $p(z)$.
- By replacing the prior $p(z)$ by $q_\phi(z|x)$. The objective function for the diffusion source prediction is:

$\hat{x}$ denotes diffusion sources from the training set

$$
\begin{aligned}
\mathcal{L}_{pred} &= \max_x \left[p_\psi(y|x,G) \cdot \sum_z \sum_{\hat{x}} p_\theta(x|z) \cdot q_\phi(z|\hat{x})\right] \\
&= \min_x \Big[-\log p_\psi(y|x,G) - \log \big[\sum_z \sum_{\hat{x}} p_\theta(x|z) \cdot q_\phi(z|\hat{x}) \big] \Big]
\end{aligned}
$$

Outline

- Background
- Proposed Method – SL-VAE
- Experiment
- Concluding Remarks

Experiment Setup

	Nodes	Edges	Average Degree	Clustering Coefficient
Karate	34	78	2.294	0.255
Jazz	198	2,742	27.69	0.52
Cora-ML	2,810	7,981	5.68	0.246
Network Science	1,565	13,532	17.29	0.741
Power Grid	4,941	6,594	2.669	0.081
MemeTracker	12,529	70,466	11.248	0.092
Digg	15,912	78,649	9.885	0.083

Training Data:

- Randomly sample 10% nodes as seeds and simulate the information diffusion based on epidemic models: Susceptible-Infection (SI) and Susceptible-Infection-Recovery (SIR).
- *Digg* and *MemeTracker* are social networks with real-world diffusion cascades.

Evaluation Metrics: Classification Scores between the predicted sources and the ground truth ones.

Comparison Methods:

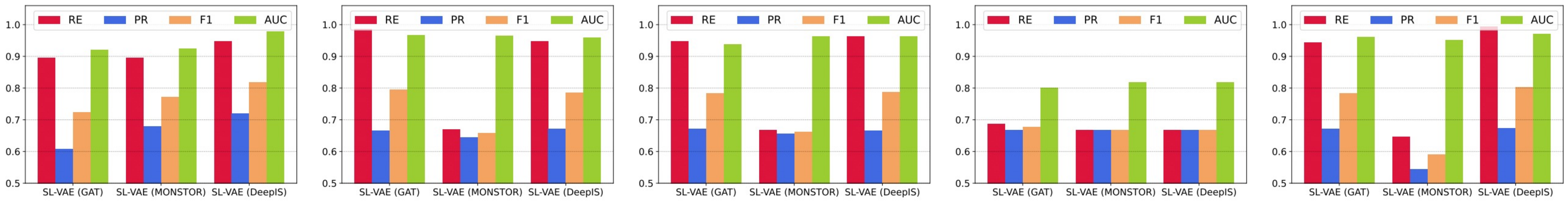
- *Diffusion Estimation Models*: SL-VAE can be coupled with different diffusion estimation models, we select STOAs: GAT [1], MONSTOR [2], and DeepIS [3].
- *Source Localization Methods*. SL-VAE is compared with NetSleuth [4], LPSI [5], OJC [6], and GCNSI [7].

[1] Veličković, Petar, et al. "Graph attention networks." *arXiv preprint arXiv:1710.10903* (2017).
[2] Ko, Jihoon, et al. "Monstor: an inductive approach for estimating and maximizing influence over unseen networks." ASONAM, 2020.
[3] Xia, Wenwen, et al. "Deepis: Susceptibility estimation on social networks." WSDM. 2021.
[4] Prakash, B. Aditya, et al. "Spotting culprits in epidemics: How many and which ones?", ICDM 2012.
[5] Wang, Zheng, et al. "Multiple source detection without knowing the underlying propagation model." AAAI. 2017.
[6] Zhu, Kai, et al. "Catch'em all: Locating multiple diffusion sources in networks with partial observations." AAAI. 2017.
[7] Dong, Ming, et al. "Multiple rumor source detection with graph convolutional networks." CIKM. 2019.

Experiment: Flexibility

How accurate does SL-VAE perform in source localization task when equipped with various forward diffusion estimation models $p_{\psi}(y|x,G)$?

- AUCs for each dataset are above **90%** with only one exception.
- We cannot find **noticeable difference** in performance between each variant of SL-VAE.



(a) Jazz (b) Cora_ML (c) Power Grid (d) Karate (e) Network Science

Figure 3: The performance of SL-VAE with different forward models under the SI diffusion pattern.

Experiment: Accuracy and Adaptation

- SL-VAE excels others in terms of both F1 and AUC on average **15%** in predicting the diffusion sources given the diffused observation under SI diffusion pattern.
- While others experience a performance decline in recovering sources under real-world diffusion pattern, SL-VAE still achieves the best by leading the second best **20%**.

The performance comparison on source localization under SI diffusion Pattern

	Jazz				Cora-ML				Power Grid				Karate				Network Science			
Methods	RE	PR	F1	AUC	RE	PR	F1	AUC	RE	PR	F1	AUC	RE	PR	F1	AUC	RE	PR	F1	AUC
LPSI	0.4789	0.1054	0.1716	0.4841	0.5954	0.1556	0.2466	0.6675	0.4953	0.4546	0.4737	0.9337	0.4667	0.1861	0.2855	0.6344	0.6044	0.4231	0.4978	0.8378
GCNSI	0.4368	0.1589	0.2329	0.6428	0.3619	0.1182	0.1780	0.5383	0.3477	0.1413	0.2099	0.5044	0.4333	0.1999	0.2613	0.6022	0.2247	0.1375	0.1706	0.4759
OJC	0.1798	0.1005	0.1289	0.5045	0.2239	0.2036	0.2133	0.5633	0.2871	0.1044	0.1531	0.5011	0.3611	0.2708	0.3095	0.6335	0.1233	0.3708	0.1851	0.5331
Netsleuth	0.1315	0.1087	0.1191	0.5432	0.2647	0.2647	0.2647	0.4688	0.5972	0.4975	0.5428	0.7651	0.3333	0.3333	0.3333	0.4355	0.3948	0.3283	0.3585	0.6528
SL-VAE	**0.9474**	**0.7193**	**0.8182**	**0.9777**	**0.9466**	**0.6717**	**0.7858**	**0.9582**	**0.9636**	**0.6648**	**0.7868**	**0.9636**	**0.6667**	**0.6667**	**0.6667**	**0.8172**	**0.9937**	**0.6738**	**0.8031**	**0.9705**

The performance comparison on source localization under Real-world diffusion Pattern

	Digg-7556				Digg				Memetracker-7884				Memetracker			
Methods	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC
LPSI	0.0026	0.0123	0.0043	0.4432	0.0079	0.2727	0.0155	0.5618	0.0132	0.3184	0.0253	0.5112	0.0087	0.2913	0.0169	0.5377
GCNSI	0.0114	0.3700	0.0221	0.4450	0.0123	0.2100	0.0232	0.4129	0.0211	0.3219	0.0396	0.4357	0.0197	0.2342	0.0363	0.4103
OJC	0.0118	0.0107	0.0112	0.5023	0.0635	0.0696	0.0664	0.5142	0.0542	0.0433	0.0481	0.4812	0.0331	0.0207	0.0255	0.5077
SL-VAE	**0.4131**	**0.6217**	**0.4655**	**0.5541**	**0.4297**	**0.5421**	**0.4792**	**0.6213**	**0.5113**	**0.6214**	**0.5610**	**0.5954**	**0.4612**	**0.5181**	**0.4880**	**0.6245**

Experiment: Accuracy and Adaptation

- SL-VAE excels others in terms of both F1 and AUC on average **15%** in predicting the diffusion sources given the diffused observation under SI diffusion pattern.
- While others experience a performance decline in recovering sources under real-world diffusion pattern, SL-VAE still achieves the best by leading the second best **20%**.

The performance comparison on source localization under SI diffusion Pattern

	Jazz				Cora-ML				Power Grid				Karate				Network Science			
Methods	RE	PR	F1	AUC	RE	PR	F1	AUC	RE	PR	F1	AUC	RE	PR	F1	AUC	RE	PR	F1	AUC
LPSI	0.4789	0.1054	0.1716	0.4841	0.5954	0.1556	0.2466	0.6675	0.4953	0.4546	0.4737	0.9337	0.4667	0.1861	0.2855	0.6344	0.6044	0.4231	0.4978	0.8378
GCNSI	0.4368	0.1589	0.2329	0.6428	0.3619	0.1182	0.1780	0.5383	0.3477	0.1413	0.2099	0.5044	0.4333	0.1999	0.2613	0.6022	0.2247	0.1375	0.1706	0.4759
OJC	0.1798	0.1005	0.1289	0.5045	0.2239	0.2036	0.2133	0.5633	0.2871	0.1044	0.1531	0.5011	0.3611	0.2708	0.3095	0.6335	0.1233	0.3708	0.1851	0.5331
Netsleuth	0.1315	0.1087	0.1191	0.5432	0.2647	0.2647	0.2647	0.4688	0.5972	0.4975	0.5428	0.7651	0.3333	0.3333	0.3333	0.4355	0.3948	0.3283	0.3585	0.6528
SL-VAE	**0.9474**	**0.7193**	**0.8182**	**0.9777**	**0.9466**	**0.6717**	**0.7858**	**0.9582**	**0.9636**	**0.6648**	**0.7868**	**0.9636**	**0.6667**	**0.6667**	**0.6667**	**0.8172**	**0.9937**	**0.6738**	**0.8031**	**0.9705**

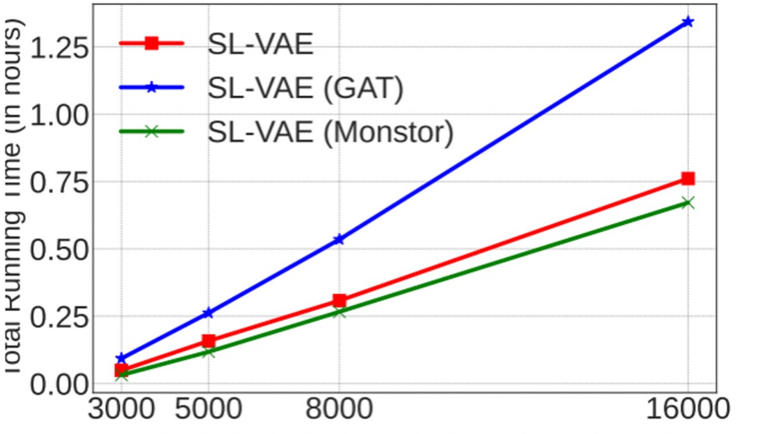
The performance comparison on source localization under Real-world diffusion Pattern

	Digg-7556				Digg				Memetracker-7884				Memetracker			
Methods	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC
LPSI	0.0026	0.0123	0.0043	0.4432	0.0079	0.2727	0.0155	0.5618	0.0132	0.3184	0.0253	0.5112	0.0087	0.2913	0.0169	0.5377
GCNSI	0.0114	0.3700	0.0221	0.4450	0.0123	0.2100	0.0232	0.4129	0.0211	0.3219	0.0396	0.4357	0.0197	0.2342	0.0363	0.4103
OJC	0.0118	0.0107	0.0112	0.5023	0.0635	0.0696	0.0664	0.5142	0.0542	0.0433	0.0481	0.4812	0.0331	0.0207	0.0255	0.5077
SL-VAE	**0.4131**	**0.6217**	**0.4655**	**0.5541**	**0.4297**	**0.5421**	**0.4792**	**0.6213**	**0.5113**	**0.6214**	**0.5610**	**0.5954**	**0.4612**	**0.5181**	**0.4880**	**0.6245**

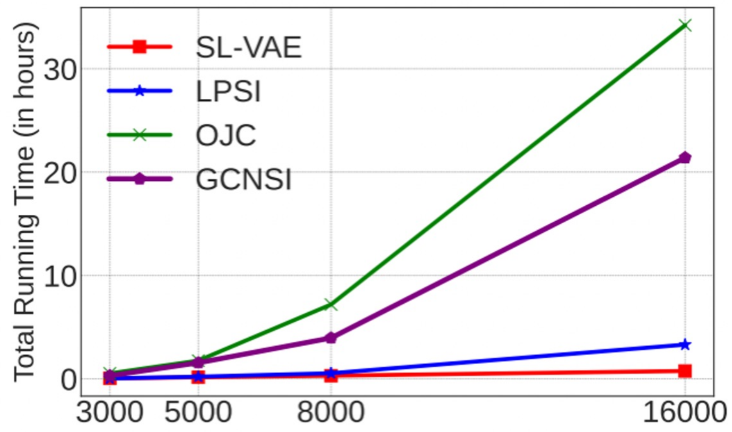
Experiment: Scalability

We record the runtime regarding the number of nodes in Digg dataset. Fig. a is the runtime of SL-VAE with different forward models; Fig. b is the runtime of SL-VAE against other source localization algorithms.

1. All variants of SL-VAE demonstrate the **linear** runtime with the growth of graph size.
2. Only LPSI has comparable runtime with SLVAE in node size (≤ 5,000), other models are slower than SL-VAE in operating on large graphs (≥ 8,000).



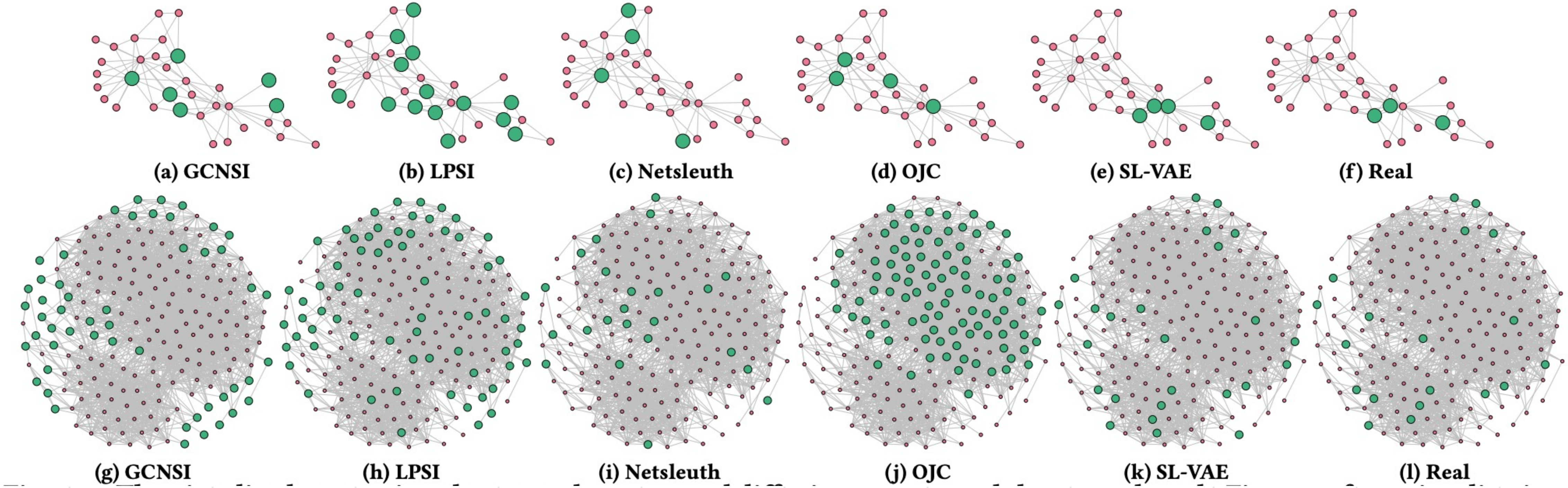
(a) Different forward models



(b) Different comparison methods

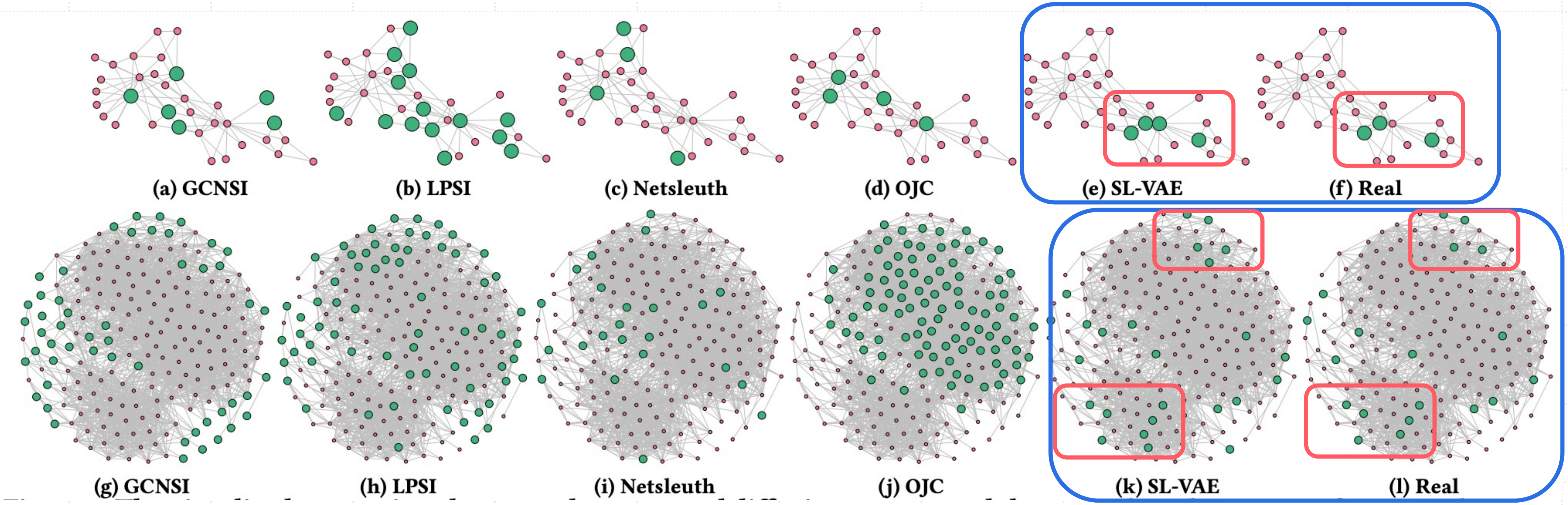
Experiment: Visualization

- We also present a *case study* to visually demonstrate the performance comparison in recovering the true seed nodes. (*Karate* on the top, *Jazz* on the bottom.)
- SL-VAE generates the overall most similar seed nodes to the ground truth compared with others.



Experiment: Visualization

- We also present a *case study* to visually demonstrate the performance comparison in recovering the true seed nodes. (*Karate* on the top, *Jazz* on the bottom.)
- SL-VAE generates the overall most similar seed nodes to the ground truth compared with others.



Outline

Background

Proposed Method – SL-VAE

Experiment

Concluding Remarks

Key Takeaways

Full Paper Code

- As one important *graph inverse problem*, information diffusion source localization is an essential yet challenging task with numerous network science applications.
- SL-VAE utilizes the *deep generative models* to approximate the intrinsic patterns of diffusion sources directly, and leverage an *end-to-end optimization* way to locate diffusion sources based on the diffused observation.
- Extensive experiments demonstrate the effectiveness of the proposed method and empirically prove that SL-VAE can generalize under any information diffusion patterns.

chen.ling@emory.edu