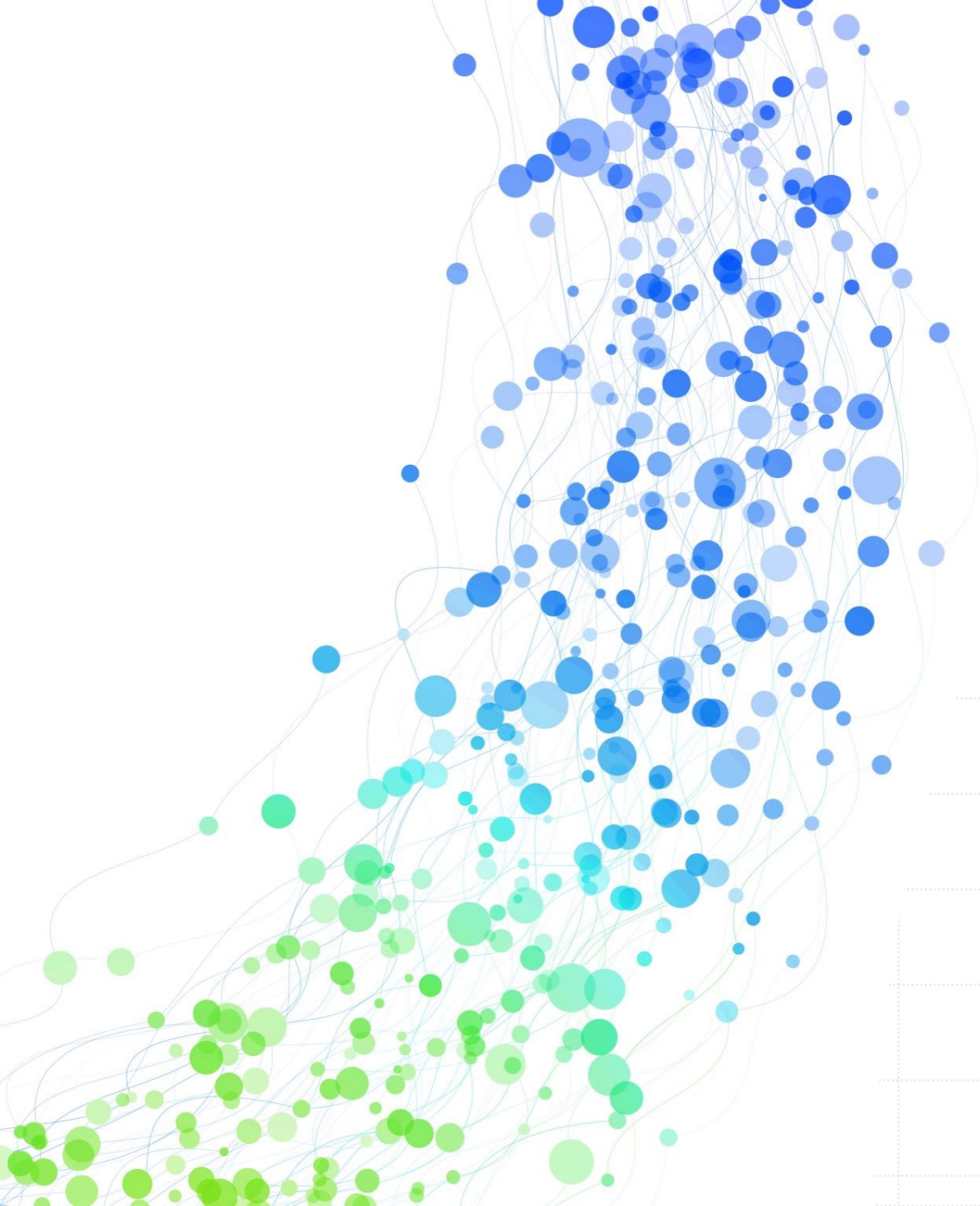


ICML International Conference On Machine Learning 40 Years

EMORY UNIVERSITY | Department of Computer Science

The 40th International Conference on Machine Learning (ICML 2023)

Deep Graph Representation Learning and Optimization for Influence Maximization

Presented by: Chen Ling

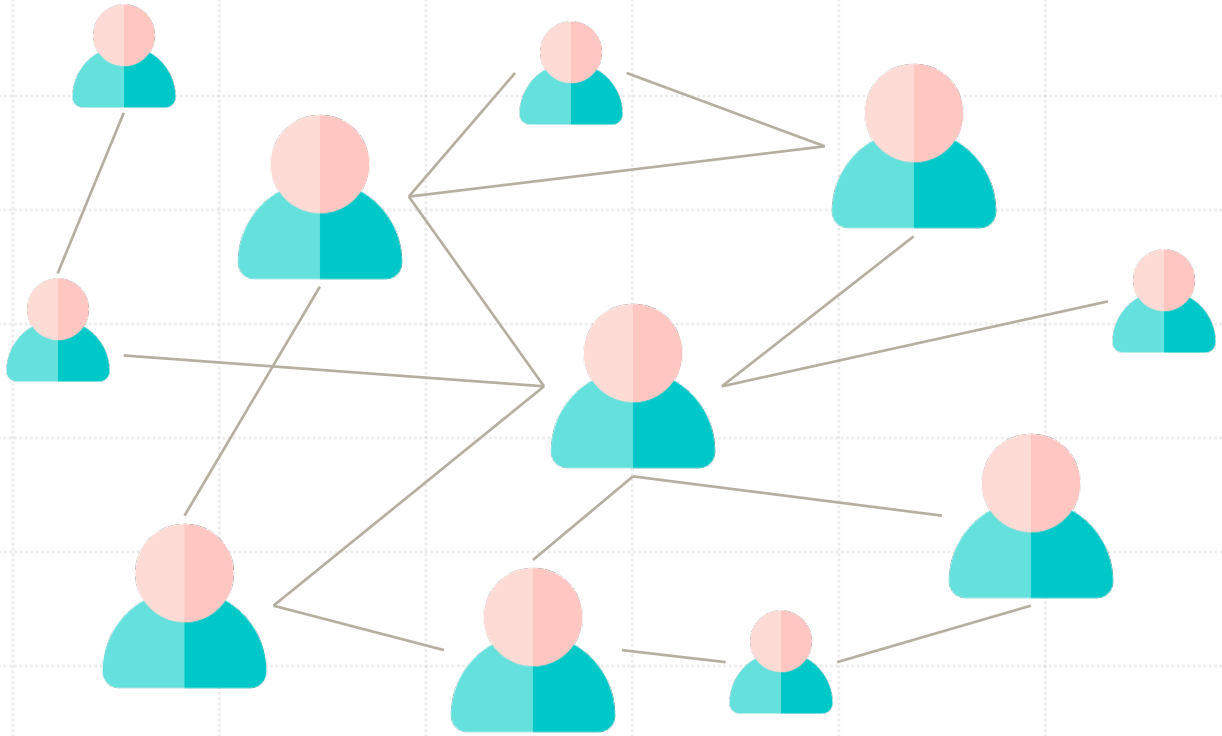
chen.ling@emory.edu

lingchen0331.github.io

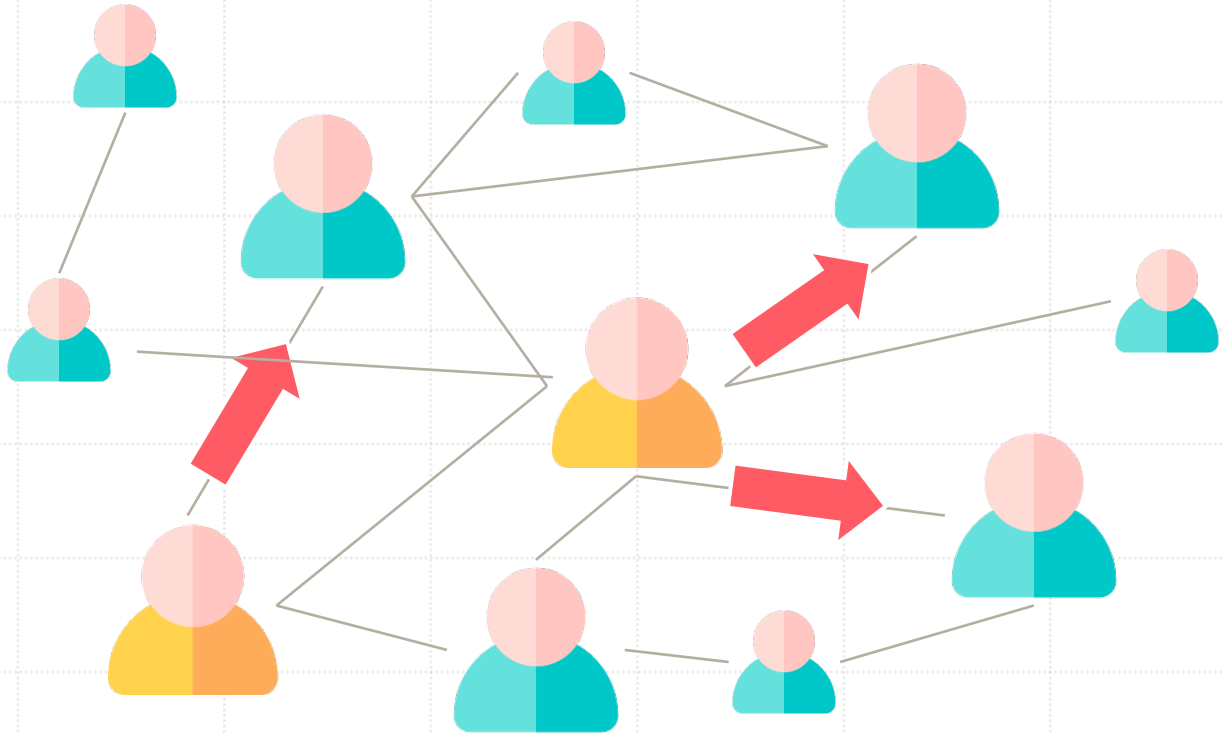
Joint work with Junji Jiang, Junxiang Wang, My T. Thai, and Liang Zhao

Jun 24, 2023

Background



Background



Viral Marketing

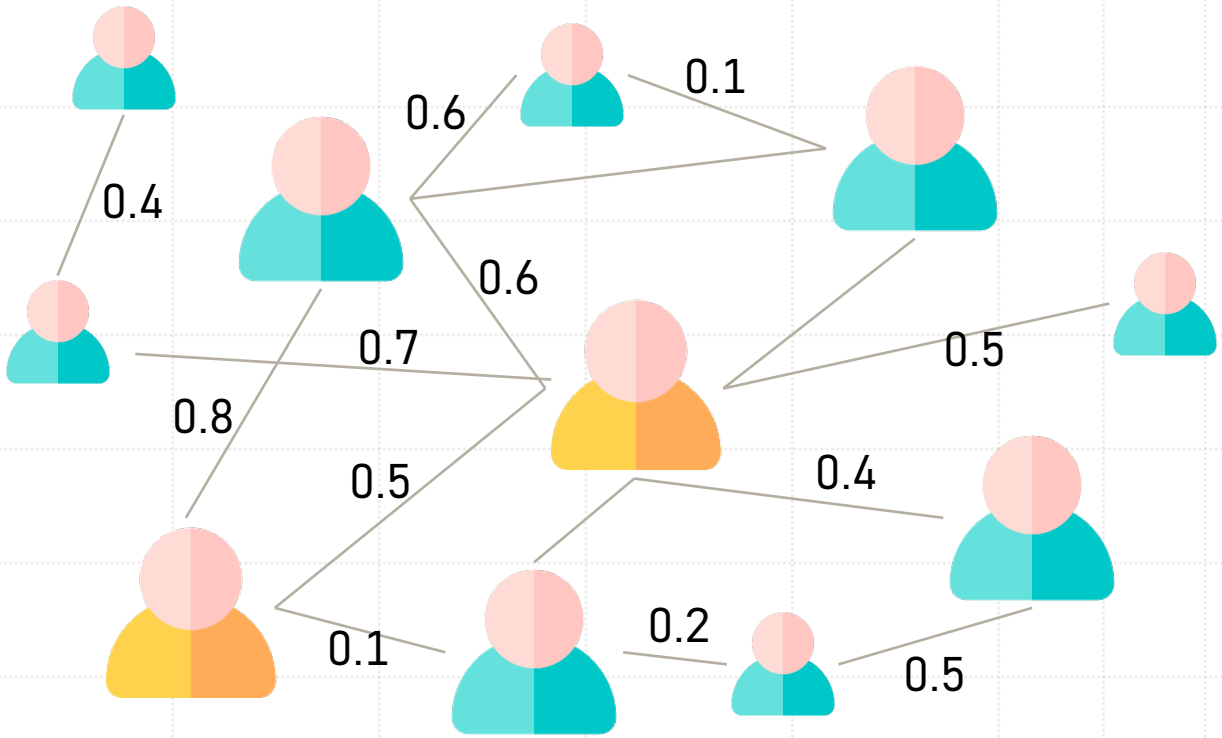
Outbreak Detection

… …

Background

- *Influence Maximization (IM)* aims at selecting a subset of users to maximize the spread of information in the network.

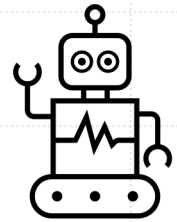
- Traditional IM solutions requires explicit information diffusion model (e.g., Independent Cascade) as model input, which limiting the real-world usage.



Independent Cascade (IC) Model

Background

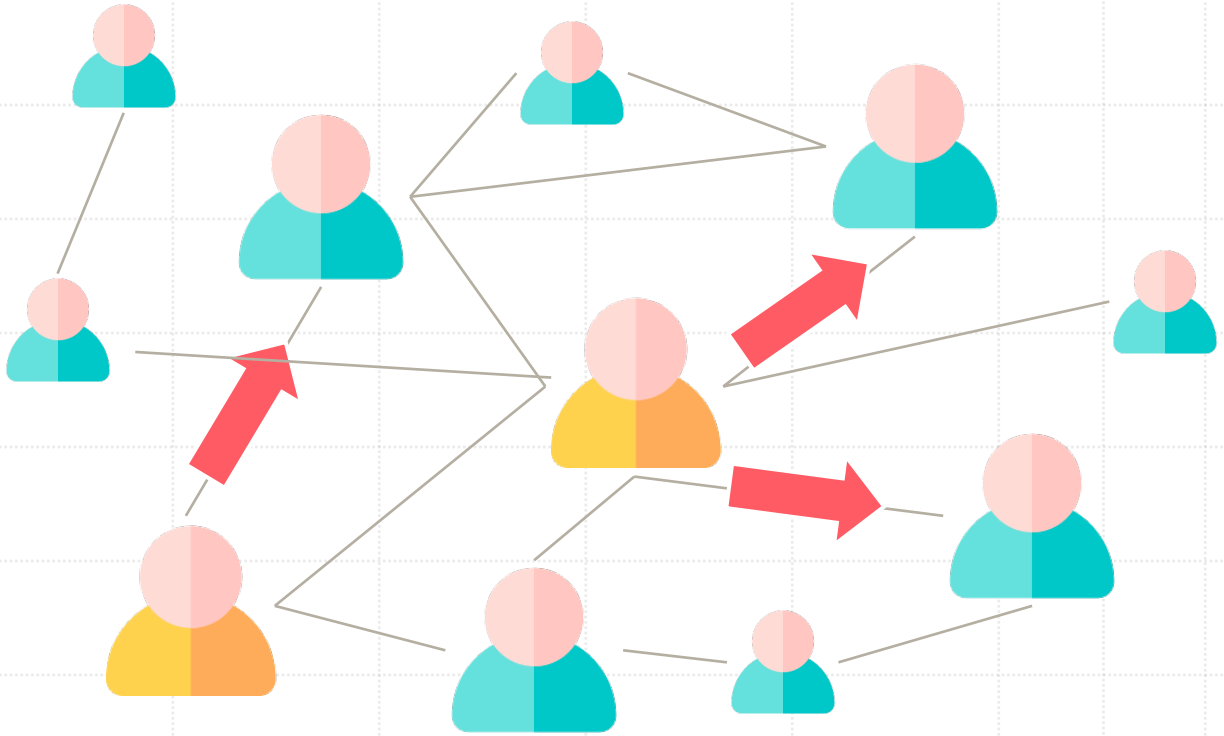
Representation Learning



Reinforcement Learning

- *Influence Maximization (IM)* aims at selecting a subset of users to maximize the spread of information in the network.

- Learning-based IM solutions improve their solution generalization ability but also bring:
  - Effectively and efficiently optimizing the objective function.
  - Automatically identifying and modeling the actual diffusion process.
  - Adapting solutions to various node-centrality-constrained IM problems.



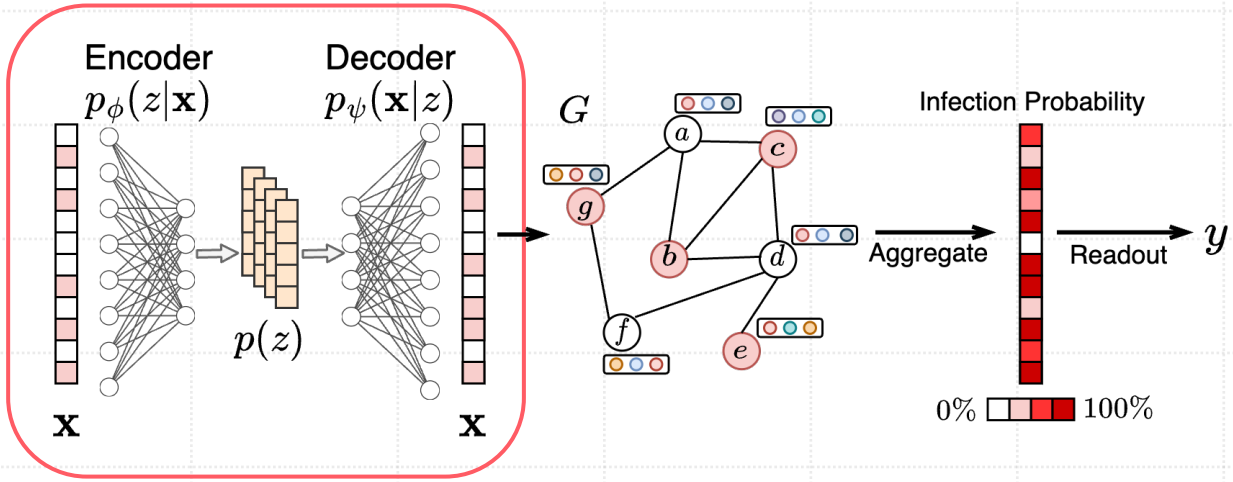
DeepIM – Basic Notations

- The *seed node set* is defined over V as $\boldsymbol{x} = \{0, 1\}^{|V|}$, $x_i = 1$ or 0 denotes seed node or not.
- The *total number* of infected nodes is denoted as $y \in \mathbb{R}_+$.
- IM aims at selecting $\tilde{\boldsymbol{x}} = \underset{|\boldsymbol{x}| \leq k}{\operatorname{argmax}} M(\boldsymbol{x}, G; \theta)$, where $y = M(\cdot)$ denotes a diffusion estimation function.

DeepIM – Learning Probability over Seed Sets

- The *seed node set* is defined over V as $\boldsymbol{x} = \{0,1\}^{|V|}$, $x_i = 1$ or 0 denotes seed node or not.
- The *total number* of infected nodes is denoted as $y \in \mathbb{R}_+$.
- IM aims at selecting $\tilde{\boldsymbol{x}} = \underset{|\boldsymbol{x}| \le k}{\operatorname{argmax}} M(\boldsymbol{x}, G; \theta)$, where $y = M(\cdot)$ denotes a diffusion estimation function.

- *To build an effective and efficient objective function:*
  → Characterize the probability of the seed node set $p(\boldsymbol{x})$ over $\boldsymbol{x}$ given the graph G.

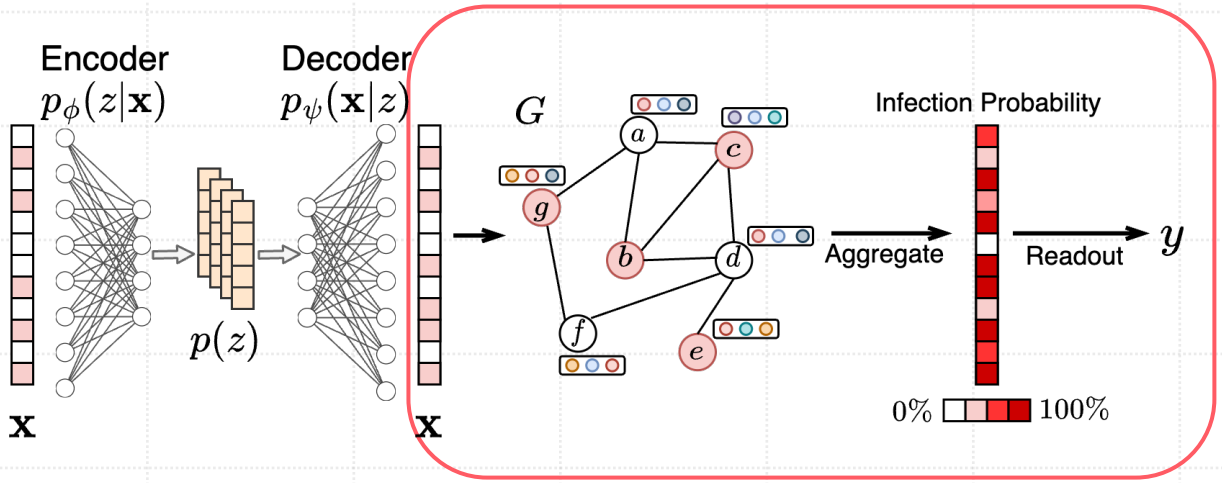


DeepIM – Learning End-to-end Diffusion Model

- The *seed node set* is defined over V as $\boldsymbol{x} = \{0,1\}^{|V|}$, $x_i = 1$ or 0 denotes seed node or not.
- The *total number* of infected nodes is denoted as $y \in \mathbb{R}_+$.
- IM aims at selecting $\tilde{\boldsymbol{x}} = \underset{|\boldsymbol{x}| \le k}{\operatorname{argmax}} M(\boldsymbol{x}, G; \theta)$, where $y = M(\cdot)$ denotes a diffusion estimation function.

- *To make the diffusion model applicable in various real-world diffusion scenarios:*

  → Design an end-to-end GNN-based diffusion estimation model $M(\cdot)$ with theoretical (i.e., monotonicity) guarantee.

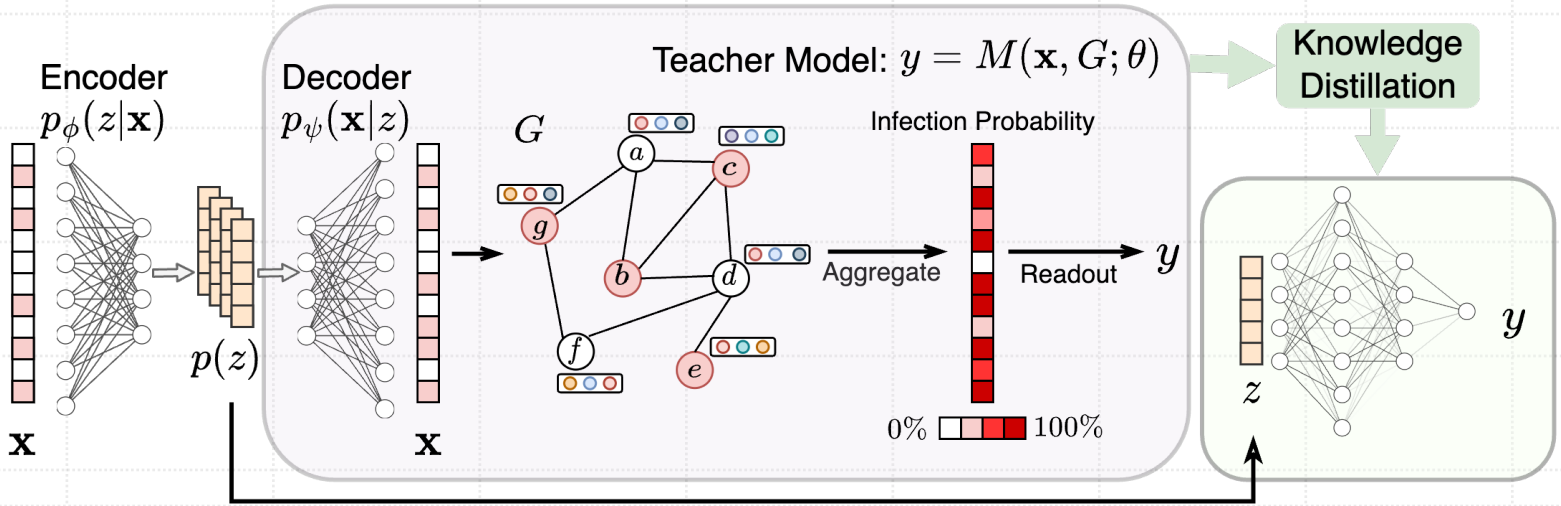


DeepIM – Knowledge Distillation Module

- The *seed node set* is defined over V as $\boldsymbol{x} = \{0,1\}^{|V|}$, $x_i = 1$ or 0 denotes seed node or not.
- The *total number* of infected nodes is denoted as $y \in \mathbb{R}_+$.
- IM aims at selecting $\tilde{\boldsymbol{x}} = \underset{|\boldsymbol{x}| \le k}{\operatorname{argmax}} M(\boldsymbol{x}, G; \theta)$, where $y = M(\cdot)$ denotes a diffusion estimation function.

- *To improve the efficiency in the influence estimation stage:*

  → Design a knowledge distillation module $M_s(\cdot)$ based on MLP structure that can estimate influence directly from the latent variable z.



DeepIM – Training Objective

- The *seed node set* is defined over V as $\boldsymbol{x} = \{0,1\}^{|V|}$, $x_i = 1$ or 0 denotes seed node or not.
- The *total number* of infected nodes is denoted as $y \in \mathbb{R}_+$.
- IM aims at selecting $\tilde{\boldsymbol{x}} = \underset{|\boldsymbol{x}| \le k}{\operatorname{argmax}} M(\boldsymbol{x}, G; \theta)$, where $y = M(\cdot)$ denotes a diffusion estimation function.

- *Final Objective Function for jointly training three components:*

GNN-based influence Estimation Model (Efficacy)

MLP-based influence Estimation Model (Efficiency)

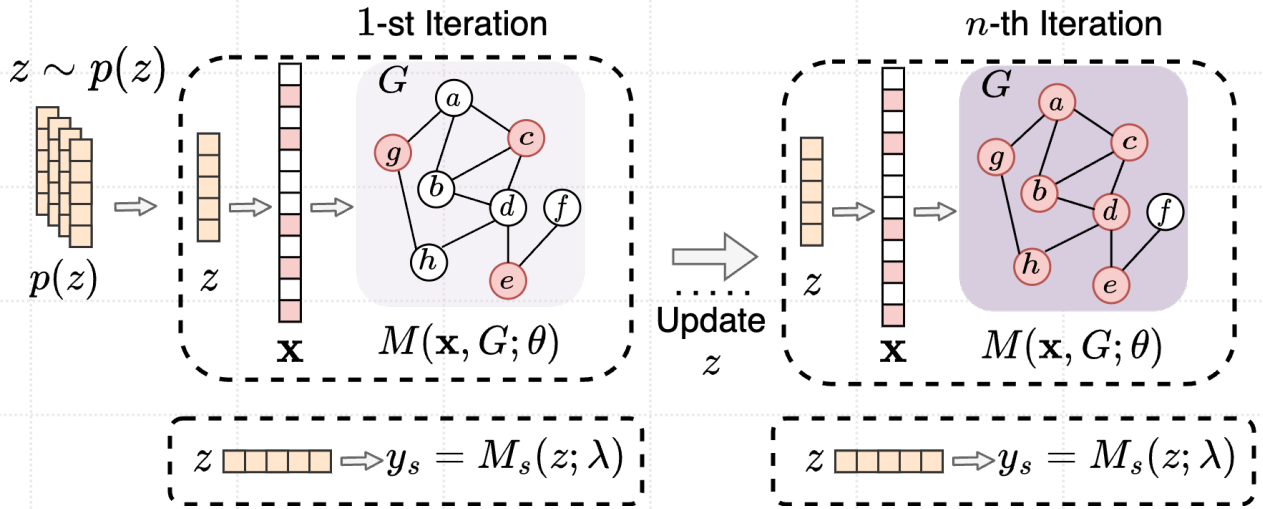
Seed Node Set Probability Quantification

$$\mathcal{L}_{\text{train}} = \min_{\theta,\lambda,\psi,\phi} \mathbb{E}\Big[-\log\left[p_\theta(y|\mathbf{x}, G)\right] - \log\left[p_\lambda(y_s|z)\right] - \log\left[p_\psi(\mathbf{x}|z) \cdot (p_\phi(z|\mathbf{x})\right]\Big], \text{s.t. } \theta \ge 0.$$

DeepIM – Iterative Optimization for Seed Set Inference

- We propose to search the *optimal seed node set* $\tilde{\boldsymbol{x}}$ in the lower-dimensional latent space $p(z)$ by iteratively optimizing the latent variable z sampled from $p(z)$.
- We further want to adapt our solution to different *IM Variants with Node Centrality Constraints* (e.g., there is a budget associated on each node).

$$\mathcal{L}_{\text{pred}} = \max_{z} \mathbb{E}\left[p_\theta(y|\mathbf{x}, G) \cdot p_\psi(\mathbf{x}|z)\right], \quad \text{s.t.} \sum_{i=0}^{|V|} \mathcal{F}(v_i, G) \cdot x_i \leq k,$$



Experiment

	Cora-ML				Network Science				Power Grid				Jazz				Synthetic				Digg				Weibo			
Methods	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%
IMM	1.7	34.8	52.2	66.4	2.5	11.9	18.1	33.6	4.6	19.9	31.7	56.9	1.4	5.7	13.4	24.5	1.1	5.2	13.1	66.9	2.4	10.8	37.4	55.6	1.6	6.7	19.3	45.2
OPIM	2.3	36.9	51.2	71.5	1.6	12.0	18.8	34.1	4.4	21.6	29.4	55.5	1.4	6.9	12.6	20.9	1.3	5.2	12.6	62.1	2.1	11.3	38.2	57.1	1.8	6.1	18.7	46.6
SubSIM	1.7	33.6	54.7	70.1	1.8	10.4	19.2	34.1	4.5	21.1	31.2	57.4	1.4	5.9	11.4	21.2	1.4	5.5	13.1	69.6	2.4	11.3	37.9	56.9	1.7	6.7	19.2	46.8
IMINFECTOR	2.1	33.9	51.3	70.6	2.1	11.8	18.7	34.5	4.2	21.3	31.6	56.2	1.4	6.2	13.5	22.8	1.3	5.2	12.9	67.4	2.2	11.1	38.9	58.7	1.8	6.4	18.6	47.5
PIANO	2.1	33.5	53.3	69.8	2.1	11.3	19.1	33,9	4.3	21.3	31.4	57.1	1.1	6.2	12.1	22.4	1.2	5.2	12.9	67.4	-	-	-	-	-	-	-	-
ToubleGDD	2.3	36.2	54.5	70.9	2.8	12.4	19.8	34.6	4.8	21.9	32.6	58.1	1.4	6.5	12.9	23.6	1.3	5.5	13.4	70.2	-	-	-	-	-	-	-	-
DeepIM$_s$	10.7	65.6	75.1	85.2	3.5	14.6	23.8	37.8	5.1	22.9	40.3	65.1	1.4	6.5	14.2	85.3	1.5	6.0	14.2	90.3	3.1	13.3	39.2	67.9	2.5	7.1	32.6	68.4
DeepIM	**13.4**	**69.2**	**83.5**	**94.1**	**4.1**	**16.6**	**26.7**	**41.5**	**6.3**	**24.4**	**46.8**	**71.7**	**1.9**	**6.5**	**16.4**	**99.1**	**1.5**	**6.5**	**15.5**	**99.9**	**3.5**	**15.9**	**41.3**	**76.2**	**3.1**	**7.6**	**39.3**	**72.4**

Table 3. Performance comparison under LT diffusion pattern. — indicates out-of-memory error. (Best is highlighted with bold.)

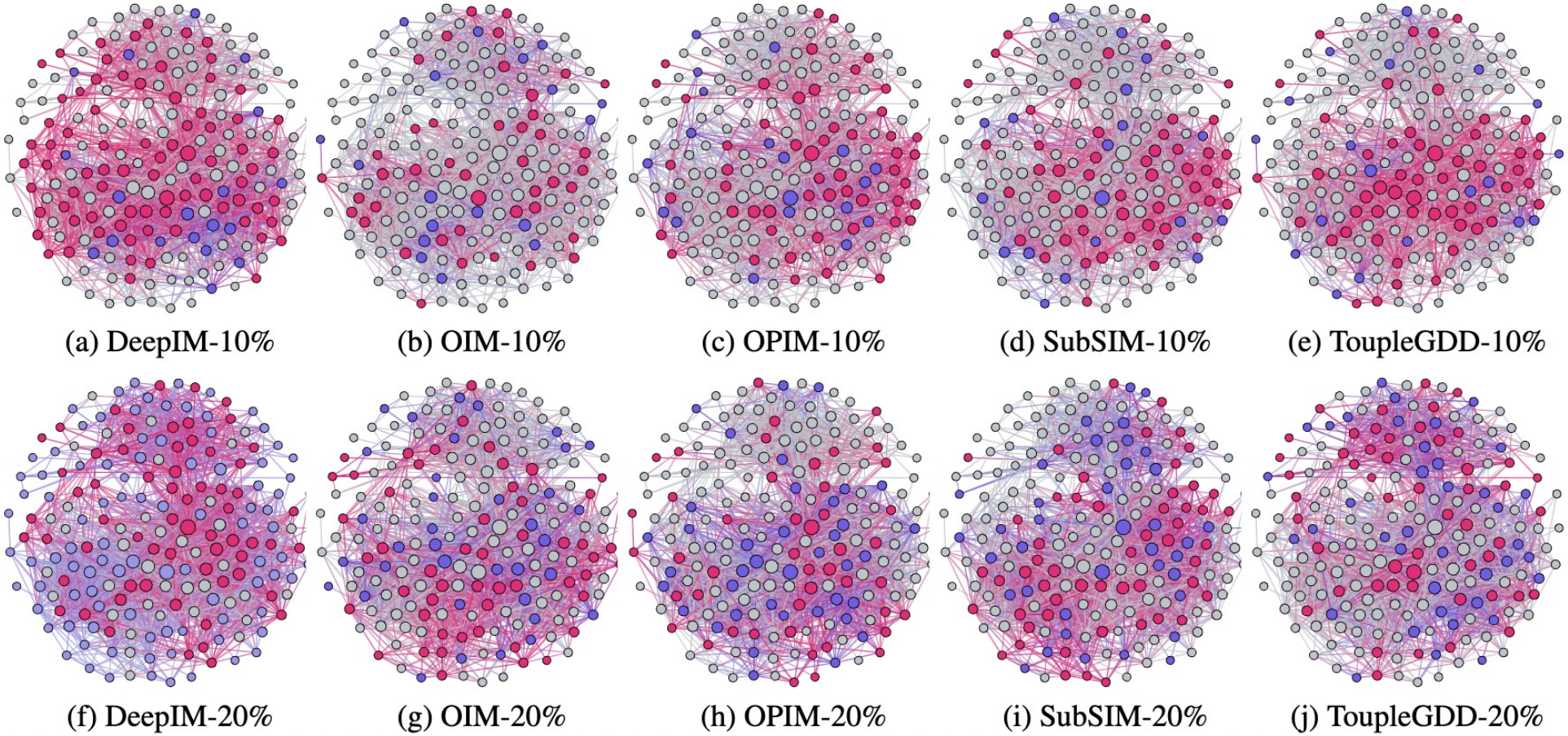
1. The primary purpose is to evaluate the number of influence spread y.
2. We compare to both traditional and learning-based IM solutions.
3. Both variants of DeepIM surpass other methods under Linear Threshold and Independent Cascade diffusion patterns.

	10,000	20,000	30,000	50,000	50,000 (Training)
IMINFECTOR	3.478s	7.842s	12.376s	16.492s	4753.67s
PIANO	5.948s	10.532s	16.575s	28.437s	14732.63s
ToupleGDD	10.476s	19.583	32.792s	58.985s	–
DeepIM$_s$	0.312s	0.616s	0.847s	1.275s	503.12s
DeepIM	1.402s	2.798s	5.124s	12.882s	1244.56s

Table 4. The average inference runtime (in seconds) with regard to the increase of node size (10,000, 20,000, 30,000, and 50,000). We also demonstrate the average training time by using 50,000 nodes graph. We select 10% of nodes as the seeds uniformly.

Visualization

Red: Infected Nodes
Blue: Selected Seeds
Grey: Uninfected Nodes



(a) DeepIM-10% (b) OIM-10% (c) OPIM-10% (d) SubSIM-10% (e) ToupleGDD-10%

(f) DeepIM-20% (g) OIM-20% (h) OPIM-20% (i) SubSIM-20% (j) ToupleGDD-20%

Key Takeaways

- A novel solution to tackle IM problem with generatively characterize the complex nature of the seed node set.
- An end-to-end way to jointly model the latent diffusion pattern without the need of prescribed diffusion model.
- A novel objective function that can be coupled with multiple node-centrality-based constraints for seed node set inference.

Further Questions?  chen.ling@emory.edu