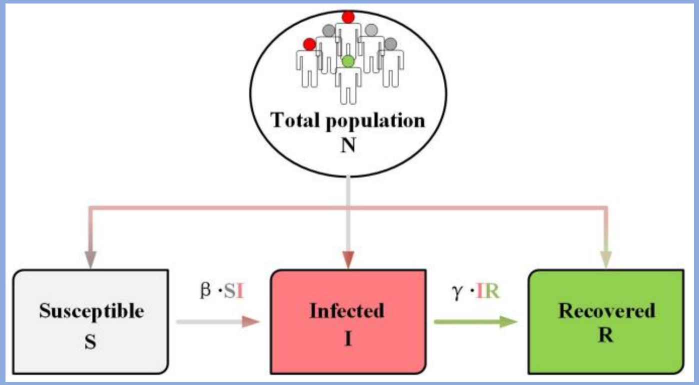
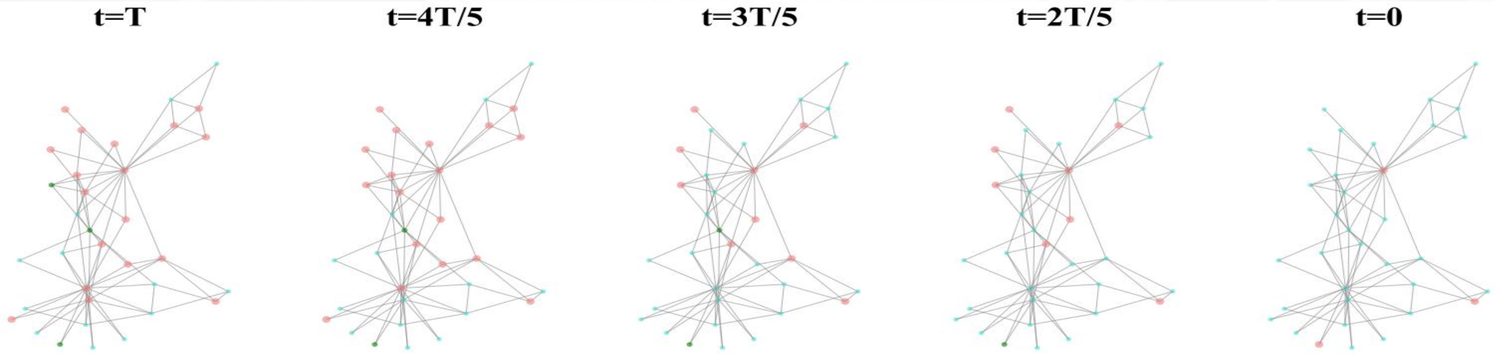
Diffusion Model for Graph Inverse Problems: Towards Effective Source Localization on Complex Networks

Xin Yan, Hui Fang, Qiang He

Task Description



SIR: A typical infectious disease model where information spreads in this form across a complex network.



How to infer the infection graph at each moment from the observation graph at time t=T?

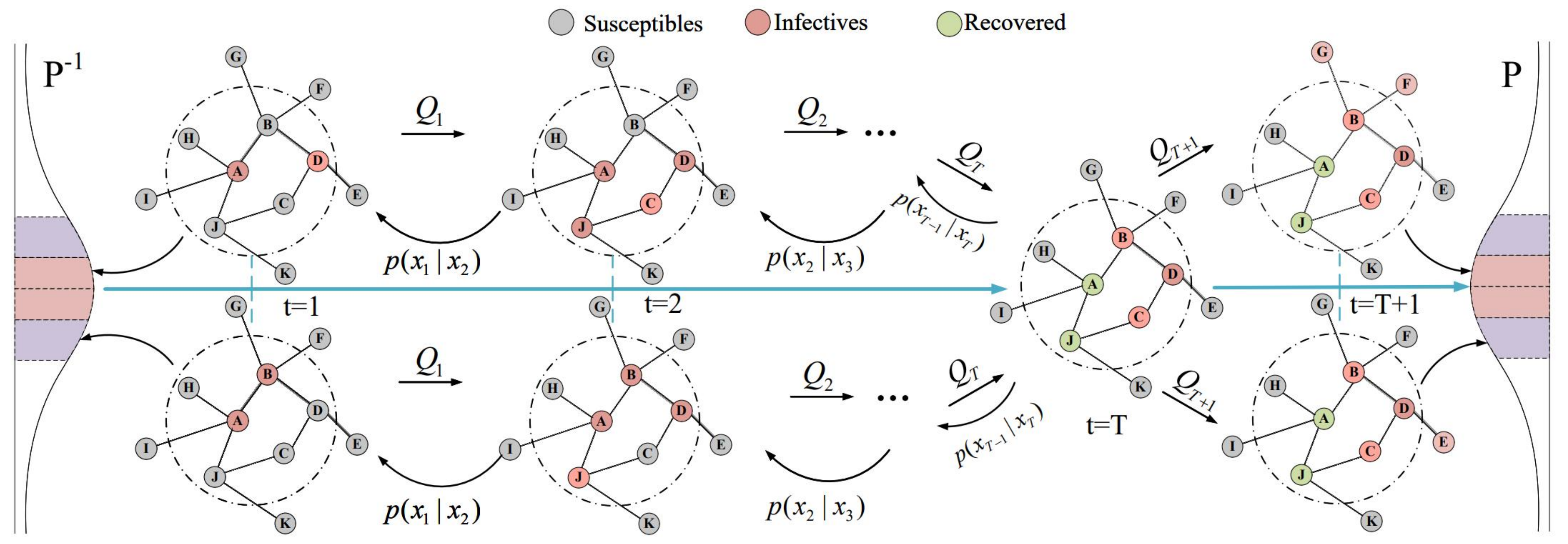


Fig1. The process of information dissemination on a graph.

Modeling the spread process of diseases using Markov chains. Assuming that the result at time t=T+1, caused by the diffusion of the source node at time t=1, follows the distribution P, the problem is defined as solving the distribution P^{-1}.

Reversible residual network

$$P_\Omega^i(t+1) = h(P_\Omega^i(t)) + \sigma\left(\sum_{j\in N(i)} h(P_\Omega^j(t))\right)$$

SIR Node status update process

Structurally equivalent

$$\begin{cases} m_\Omega^i(t+1) &= \sum_{j\in N(i)} M_t\left(h(P_\Omega^i(t)), h(P_\Omega^j(t)), e_{ij}\right) \\ P_\Omega^i(t+1) &= U_t\left(h(P_\Omega^i(t)), m_\Omega^i(t+1)\right) \\ h(P_\Omega^i(t)) &= \sigma\left(\mathbf{W}_\Omega P_\Omega^i(t) + \mathbf{b}_\Omega\right) \end{cases}$$

Message Passing Process

$$\begin{cases} \mathbf{h}_{i,t}^{(0)} = \mathbf{SN}(\mathbf{U}x_i^t) \text{ with } \mathbf{U} \in \mathbb{R}^{C\times M} \\ g(\mathbf{h}) = \sigma_g(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\cdot \mathbf{h}\cdot \mathbf{w} + \mathbf{b}) \\ \mathbf{h}_{i,t}^{(l+1)} = \mathbf{h}_{i,t}^{(l)} + \sigma\left(\mathbf{SN}\left(\mathbf{BN}\left(g\left(\mathbf{h}_{i,t}^{(l)}\right)\right)\right)\right) \end{cases}$$

Residual GCN Block

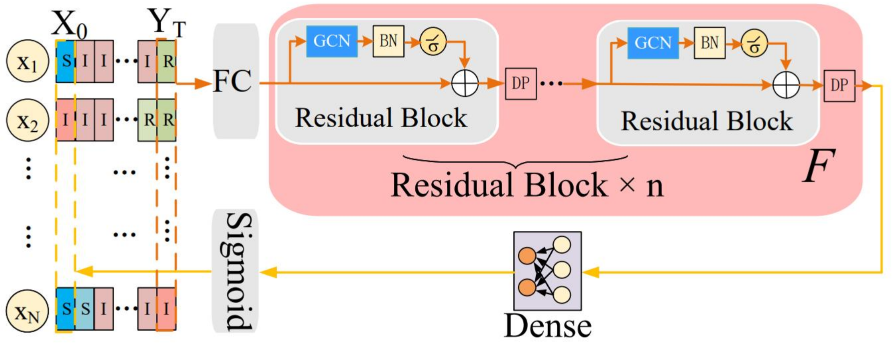


Fig2. Reversible residual graph convolutional network.

Forward Process

Write the following state transition matrix for each node:

$$Q_t^i = \begin{bmatrix} 1-\beta_I^i(t) & \beta_I^i(t) & 0 \\ 0 & 1-\gamma_R^i(t) & \gamma_R^i(t) \\ 0 & 0 & 1 \end{bmatrix}$$

The forward diffusion process is as follows:

$$q\left(\boldsymbol{x}_t^i \mid \boldsymbol{x}_{t-1}^i\right) = \boldsymbol{x}_{t-1}^i Q_t^i \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{i}\,T} \sim \operatorname{Cat}\left(\boldsymbol{x}_t^i ; \boldsymbol{p} = \boldsymbol{x}_{t-1}^i Q_t^i\right)$$

$$q\left(\boldsymbol{x}_t^i \mid \boldsymbol{x}_0^i\right) = \sum_{\boldsymbol{x}_{1:t-1}^i} \prod_{k=1}^{t} q\left(\boldsymbol{x}_k^i \mid \boldsymbol{x}_{k-1}^i\right) = \boldsymbol{x}_0^i \bar{Q}_t^i \boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{i}\,T} \sim \operatorname{Cat}\left(\boldsymbol{x}_t^i ; \boldsymbol{p} = \boldsymbol{x}_0^i \bar{Q}_t^i\right)$$

Reverse process

The following equation can be obtained from the Bayesian formula:

$$q\left(\boldsymbol{x}_{t-1}^{i} \mid \boldsymbol{x}_{t}^{i}, \boldsymbol{x}_{0}^{i}\right) \sim \operatorname{Cat}\left(\boldsymbol{x}_{t-1}^{i} ; \boldsymbol{p}=\frac{\left(\boldsymbol{x}_{t}^{i} {Q_{t}^{i}}^{T}\right) \odot\left(\boldsymbol{x}_{0}^{i} \bar{Q}_{t-1}^{i}\right)}{\boldsymbol{x}_{0}^{i} \bar{Q}_{t}^{i} {\boldsymbol{x}_{\boldsymbol{t}}^{\boldsymbol{i}}}^{T}}\right)$$

$$q\left(\boldsymbol{x}_{t-1}^{i} \mid \boldsymbol{x}_{t}^{i}, \boldsymbol{x}_{0}^{i}\right)=q\left(\boldsymbol{x}_{t-1}^{i} \mid \boldsymbol{x}_{t}^{i}\right)=\frac{\sum_{\boldsymbol{x}_{0}^{i}} q\left(\boldsymbol{x}_{t-1}^{i}, \boldsymbol{x}_{t}^{i}, \boldsymbol{x}_{0}^{i}\right)}{q\left(\boldsymbol{x}_{t}^{i}\right)}=\mathbb{E}_{q\left(\boldsymbol{x}_{0}^{i} \mid \boldsymbol{x}_{t}^{i}\right)} q\left(\boldsymbol{x}_{t-1}^{i} \mid \boldsymbol{x}_{t}^{i}, \boldsymbol{x}_{0}^{i}\right)$$

Obviously $\boldsymbol{x}_{0}^{i}$ and $q\left(\boldsymbol{x}_{0}^{i} \mid \boldsymbol{x}_{t}^{i}\right)$ are unknown, we use the reversible residual graph convolutional network mentioned above to infer the distribution:

$$\begin{aligned} q\left(\boldsymbol{x}_{t-1}^{i} \mid \boldsymbol{x}_{t}^{i}\right) \approx p_{\theta}\left(\boldsymbol{x}_{t-1}^{i} \mid \boldsymbol{x}_{t}^{i}\right) &=\mathbb{E}_{\boldsymbol{x}_{0}^{i} \sim p_{\theta}\left(\boldsymbol{x}_{0}^{i} \mid \boldsymbol{x}_{t}^{i}\right)} q\left(\boldsymbol{x}_{t-1}^{i} \mid \boldsymbol{x}_{t}^{i}, \boldsymbol{x}_{0}^{i}\right) \\ &=\frac{q\left(\boldsymbol{x}_{t}^{i} \mid \boldsymbol{x}_{t-1}^{i}\right)\left[\sum_{j} q\left(\boldsymbol{x}_{t-1}^{i} \mid {\boldsymbol{x}_{\mathbf{0}}^{\boldsymbol{i}}}^{(j)}\right) p_{\theta}\left({\boldsymbol{x}_{\mathbf{0}}^{\boldsymbol{i}}}^{(j)} \mid \boldsymbol{x}_{t}^{i}\right)\right]}{q\left(\boldsymbol{x}_{t}^{i} \mid \boldsymbol{x}_{0}^{i}\right)} \end{aligned}$$

Loss Function

1.Variational lower bound loss function

$$L_{\mathrm{vb}}^{i}=\mathbb{E}_{q\left(\boldsymbol{x}_{0}^{i}\right)}\underbrace{\left[D_{\mathrm{KL}}\left[q\left(\boldsymbol{x}_{T}^{i}\mid\boldsymbol{x}_{0}^{i}\right)\|p\left(\boldsymbol{x}_{T}^{i}\right)\right]\right.}_{L_{T}}+$$

$$\sum_{t=2}^{T}\underbrace{\mathbb{E}_{q\left(\boldsymbol{x}_{t}^{i}\mid\boldsymbol{x}_{0}^{i}\right)}\left[\mathrm{KL}\left[q\left(\boldsymbol{x}_{t-1}^{i}\mid\boldsymbol{x}_{t}^{i},\boldsymbol{x}_{0}^{i}\right)\|p_{\theta}\left(\boldsymbol{x}_{t-1}^{i}\mid\boldsymbol{x}_{t}^{i}\right)\right]\right]}_{L_{t-1}}-\underbrace{\mathbb{E}_{q\left(\boldsymbol{x}_{1}^{i}\mid\boldsymbol{x}_{0}^{i}\right)}\left[\log p_{\theta}\left(\boldsymbol{x}_{0}^{i}\mid\boldsymbol{x}_{1}^{i}\right)\right]\Big]}_{L_{0}}$$

2.Constrained reconstruction propagation loss function

To ensure that the reconstructed diffusion path conforms to the SIR propagation rules, for nodes with a state of I at the previous time, the state at the next time can only be I or R.

$$\begin{cases} L_{constrain1}=Relu\left(\mathbf{X_{t-1}}-(\mathbf{A}+\mathbf{I})\mathbf{X_{0}}\right) \\ L_{constrain2}=\left\|\max\left(\mathbf{0},\mathbf{X_{t-1}^{(j)}}-\mathbf{X_{t-1}^{(i)}}\right)\right\|_{2}^{2},\forall\mathbf{X_{0}^{(i)}}\supseteq\mathbf{X_{0}^{(j)}} \end{cases}$$

Datasets

Datasets	#Nodes	#Edges	#Avg(degree)	#Average clustering coefficient	#Density	#Diameter
Karate	34	78	2.29	0.57	0.14	5
Jazz	198	2,742	27.70	0.62	0.14	6
Cora_ml	2,810	7,981	5.68	0.28	0.002	17
Power Grid	4,941	6,594	1.33	0.08	0.005	46
PGP	10,680	24,316	4.55	0.27	0.0004	24

We conducted simulation synthesis datasets of SIR diffusion pattern on 5 real networks for training and testing. In addition, two real datasets, Douban and Twitter, were used for testing.

Experimental Result

Our method (DDMSL) performs the best in traceability, with F1 averaging about 25% ahead of the baseline method.

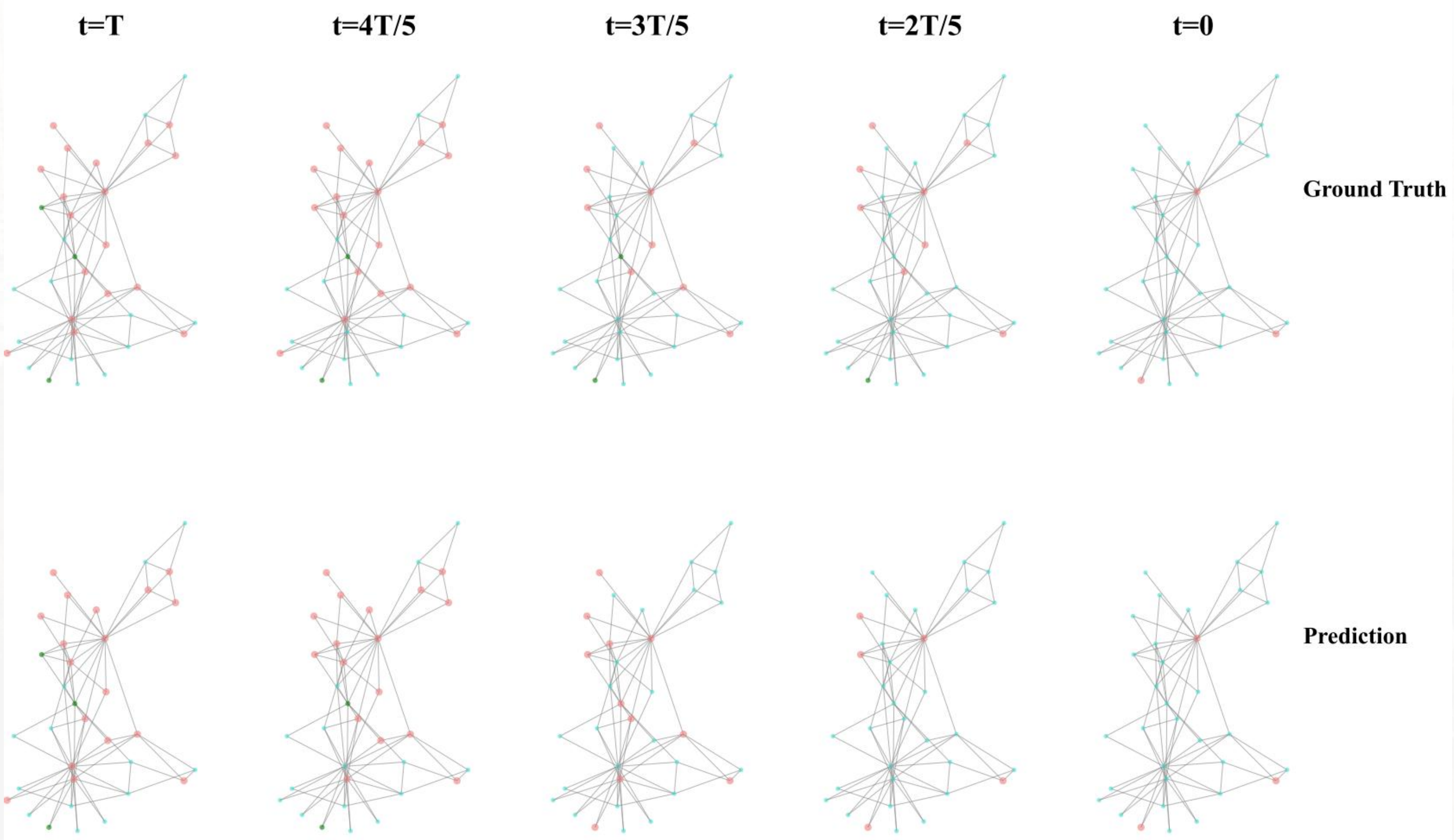
Methods	Karate				Jazz				Cora Ml				Power Grid				PGP			
	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC
DDMSL	**0.708**	**0.736**	**0.722**	**0.853**	**0.817**	**0.881**	**0.848**	**0.930**	**0.894**	**0.867**	**0.880**	**0.928**	0.833	**0.879**	**0.855**	**0.930**	**0.856**	**0.903**	**0.879**	**0.943**
GCNSI	0.275	0.410	0.329	0.671	0.301	0.363	0.330	0.641	0.247	0.273	0.260	0.591	0.165	0.182	0.173	0.540	0.554	0.543	0.549	0.748
LPSI	0.211	0.393	0.274	0.646	0.400	0.098	0.158	0.543	0.246	0.026	0.048	0.509	0.193	0.012	0.022	0.503	0.518	0.437	0.474	0.696
SLVAE	0.552	0.400	0.464	0.696	0.750	0.576	0.651	0.778	0.815	0.721	0.765	0.852	**0.908**	0.719	0.803	0.856	0.817	0.721	0.766	0.851
OJC	0.178	0.265	0.213	0.594	0.147	0.161	0.154	0.535	0.114	0.114	0.114	0.508	0.109	0.109	0.109	0.505	0.128	0.128	0.128	0.516
DDMIX	0.289	0.234	0.258	0.308	0.215	0.197	0.205	0.238	0.162	0.273	0.204	0.250	0.333	0.253	0.287	0.346	0.172	0.194	0.182	0.213
Improve.	28.3%	84.3%	55.7%	22.5%	8.9%	52.9%	30.1%	19.6%	9.7%	20.2%	15.0%	8.9%	-8.3%	22.2%	6.6%	8.7%	4.8%	25.3%	14.8%	10.8%
Significance	***	***	***	***	**	***	**	*	***	*	**	*	***	***	***	***	***	***	***	***

Table1.Performance of DDMSL in SIR diffusion mode.

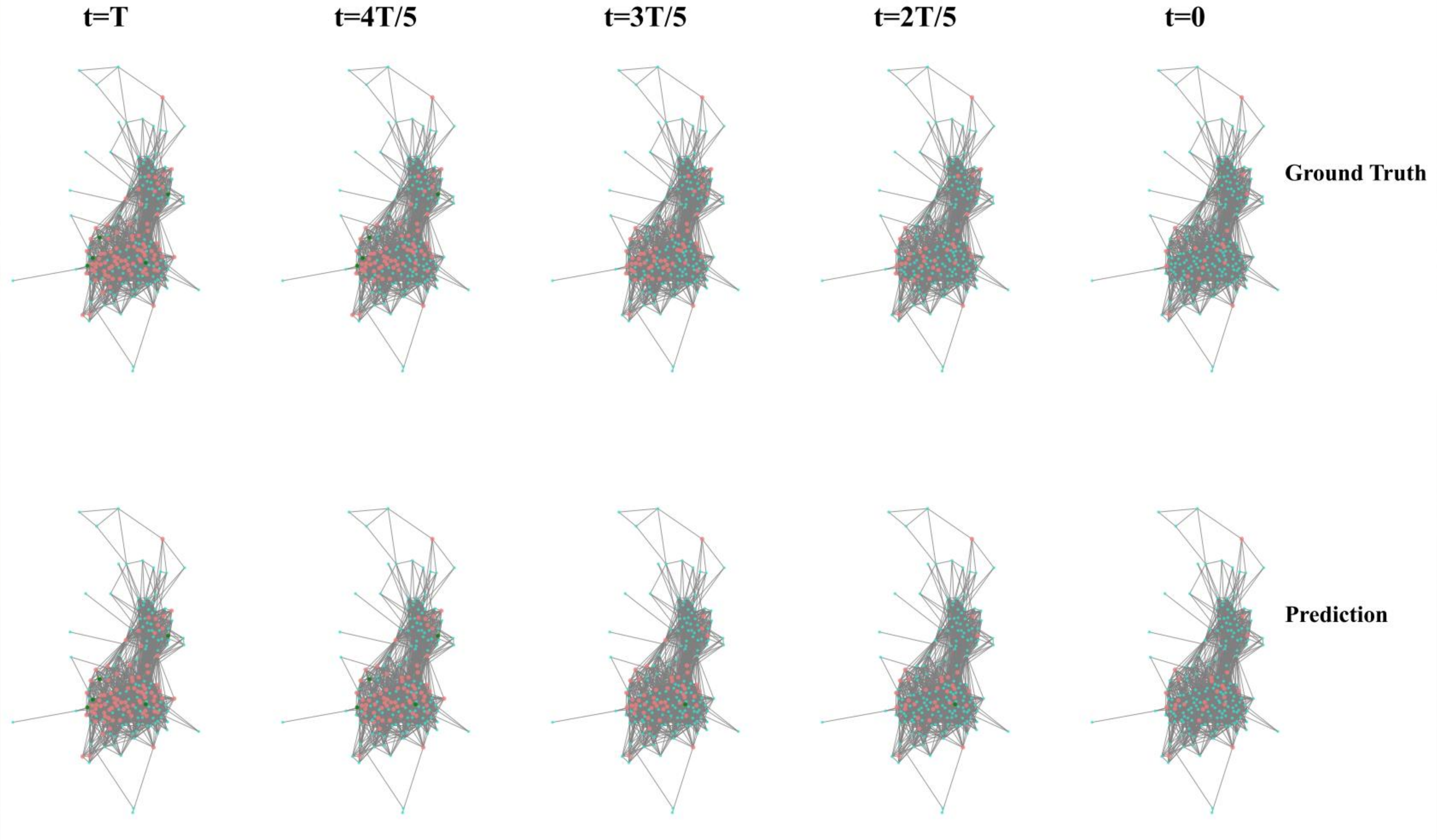
Methods	Karate				Jazz				Cora Ml				Power Grid				PGP			
	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC	PR	RE	F1	AUC
DDMSL	**0.706**	**0.980**	**0.798**	**0.972**	0.782	**0.853**	**0.813**	**0.914**	0.790	**0.908**	**0.845**	**0.941**	0.763	**0.966**	**0.852**	**0.966**	0.754	**0.887**	**0.815**	**0.928**
GCNSI	0.357	0.456	0.401	0.687	0.366	0.426	0.394	0.676	0.321	0.354	0.337	0.636	0.335	0.325	0.330	0.639	0.487	0.370	0.421	0.665
LPSI	0.339	0.414	0.351	0.681	0.474	0.097	0.156	0.544	0.494	0.207	0.291	0.592	0.343	0.277	0.306	0.609	0.453	0.284	0.349	0.623
SLVAE	0.591	0.477	0.503	0.733	**0.888**	0.579	0.691	0.785	**0.841**	0.728	0.780	0.856	**0.815**	0.780	0.797	0.880	**0.844**	0.633	0.723	0.810
OJC	0.267	0.396	0.318	0.663	0.120	0.127	0.123	0.517	0.125	0.125	0.125	0.514	0.178	0.178	0.178	0.544	0.118	0.118	0.118	0.510
DDMIX	0.253	0.377	0.303	0.654	0.244	0.133	0.172	0.212	0.220	0.222	0.221	0.247	0.345	0.235	0.280	0.340	0.189	0.186	0.187	0.207
NetSleuth	0.239	0.339	0.279	0.634	0.216	0.252	0.233	0.580	0.217	0.229	0.223	0.569	0.206	0.216	0.211	0.562	0.200	0.210	0.205	0.558
Improve.	19.4%	105.5%	58.5%	32.5%	-11.9%	47.3%	17.6%	16.3%	-6.0%	24.8%	8.3%	9.9%	-6.3%	23.8%	7.0%	9.8%	-10.6%	40.2%	12.7%	14.5%
Significance	***	***	***	***	***	***	***	***	***	***	***	***	*	*	**	**	***	***	***	***

Table2.Performance of DDMSL in SI diffusion mode.

Visualization1



Visualization2



Thank you

医学与生物信息工程学院