Reconstructing Graph Diffusion History from a Single Snapshot

Ruizhong Qiu† Dingsu Wang† Lei Ying‡ H. Vincent Poor§ Yifang Zhang¶ Hanghang Tong†

{rq5, dingsuw2, htong, zhang303}@illinois.edu leiying@umich.edu poor@princeton.edu

† UIUC ‡ UMich § Princeton ¶ C3.ai DTI

KDD2023 AUGUST 6 - 10 LONG BEACH, CA

HIGHLIGHTS

➢**Problem definition:** *Reconstructing Diffusion history from A Single snapsHot* (DASH).
- We **do not** assume knowing **true** diffusion parameters.
- We **do not** assume having real histories as **training data**.

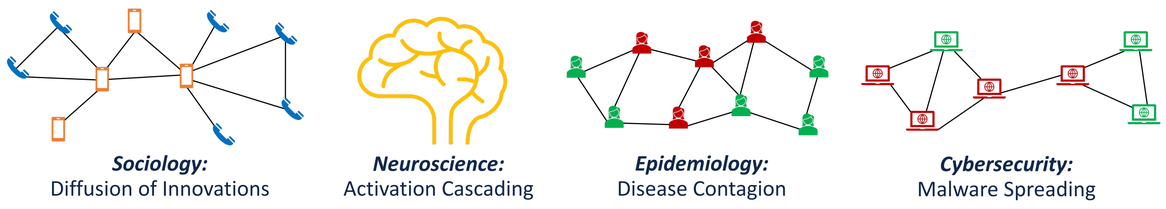
➢**Theoretical insights:** Fundamental limitation of the MLE formulation.
- Theorems 1 & 2 ⟹ **Unavoidable** estimation error of diffusion parameters.
- Theorem 3 ⟹ The MLE formulation is **sensitive** to that estimation error.

➢**Problem formulation:**
- A novel ***barycenter formulation*** based on *hitting times*.
- Provably **stable** against estimation error of diffusion parameters.

➢**Proposed method:** *DIffusion hiTting Times with Optimal proposal* (DITTO).
- Reducing the problem to estimating *posterior expected hitting times* via M–H MCMC;
- Using a GNN to learn an **optimal** proposal to accelerate convergence of M–H MCMC.

INTRODUCTION

Diffusion on Graphs



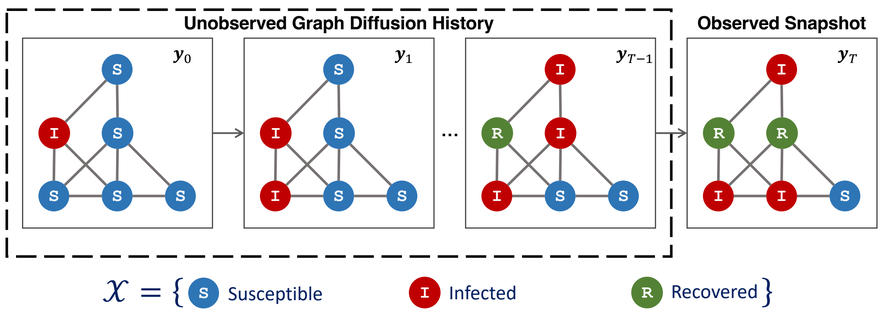
Sociology: Diffusion of Innovations

Neuroscience: Activation Cascading

Epidemiology: Disease Contagion

Cybersecurity: Malware Spreading

Problem Definition



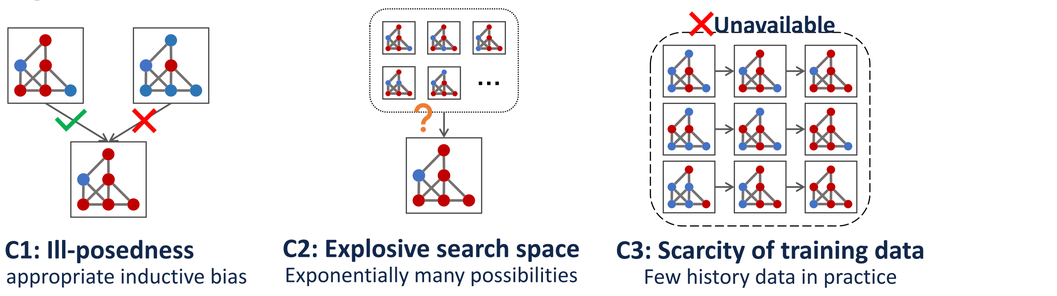
Problem (DASH): *Reconstructing Diffusion history from A single SnapsHot.*

Input: *(i) graph* $(\mathcal{V}, \mathcal{E})$; *(ii) timespan* T *of interest;* *(iii) final snapshot* $\boldsymbol{y}_T \in \mathcal{X}^{\mathcal{V}}$; *(iv) initial distribution* $P[\boldsymbol{y}_0]$.

Output: *reconstructed complete diffusion history* $\widehat{\boldsymbol{Y}} = [\widehat{\boldsymbol{y}}_0, \dots, \widehat{\boldsymbol{y}}_{T-1}, \boldsymbol{y}_T]^{\mathrm{T}} \in \mathcal{X}^{\mathcal{T}\times\mathcal{V}}$.

➢We **do not** assume knowing **true** diffusion parameters.

Challenges of the DASH Problem



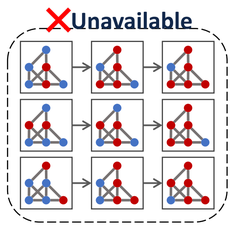
C1: Ill-posedness Need appropriate inductive bias

C2: Explosive search space Exponentially many possibilities

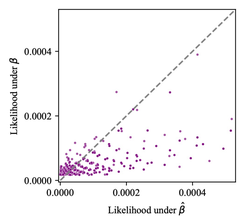
C3: Scarcity of training data Few history data in practice

Previous Methods & Their Limitations

Supervised time series imputation is **impractical** due to the scarcity of training data.



Maximum likelihood estimation (MLE) is **sensitive** to estimation error of diffusion parameters (our Theorems 1 & 2).



(a) $P_{\boldsymbol{\beta}}[Y]$ **vs** $P_{\widehat{\boldsymbol{\beta}}}[Y]$ **in the MLE formulation.**

REVISITING DIFFUSION HISTORY MLE

NP-Hardness of Diffusion Parameter Estimation

- To estimate diffusion parameters $\boldsymbol{\beta}$, a conventional approach is MLE:
$$\max_{\widehat{\boldsymbol{\beta}}} P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{y}_T]. \quad (\star)$$
- **Theorem 1** (informal): *Computing the probability* $P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{y}_T]$ *is NP-hard.*
➢$O\left(\binom{T+1}{2}^n (n+m)\right)$ time.

❖Think deeper: Is there an algo for $\widehat{\boldsymbol{\beta}}$ MLE **without** computing $P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{y}_T]$?

- **Theorem 2** (informal): *Diffusion parameter MLE* $(\star)$ *is NP-hard.*

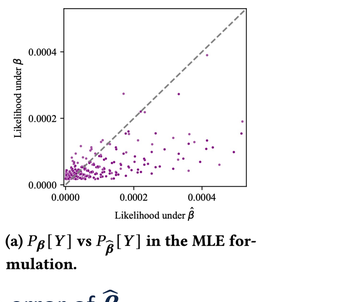
⟹ Implication: Estimation error of $\boldsymbol{\beta}$ is **unavoidable**.

Sensitivity to Estimation Error of Diffusion Parameters

- ***MLE formulation*** for diffusion history reconstruction:
$$\max_{\widehat{\boldsymbol{Y}} \in \mathrm{supp}(P|\boldsymbol{y}_T)} P_{\widehat{\boldsymbol{\beta}}}[\widehat{\boldsymbol{Y}}].$$
- **Theorem 3**. *Under the SIR model and mild conditions, for all possible history* $\boldsymbol{Y}$*, we have:*
$$\frac{\partial}{\partial \beta^{\mathrm{I}}} P_{\boldsymbol{\beta}}[\boldsymbol{Y}] = \Theta\left(\frac{1}{\beta^{\mathrm{I}}}\right) P_{\boldsymbol{\beta}}[\boldsymbol{Y}], \qquad \frac{\partial}{\partial \beta^{\mathrm{R}}} P_{\boldsymbol{\beta}}[\boldsymbol{Y}] = \Theta\left(\frac{1}{\beta^{\mathrm{R}}}\right) P_{\boldsymbol{\beta}}[\boldsymbol{Y}].$$

➢ **Large** for small $\boldsymbol{\beta}$

❖Real-world diffusion typically has **small** true $\boldsymbol{\beta}$.

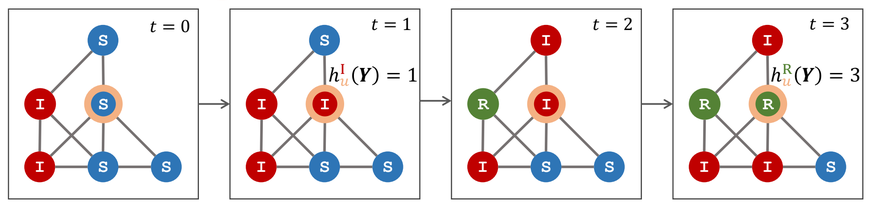


(a) $P_{\boldsymbol{\beta}}[Y]$ **vs** $P_{\widehat{\boldsymbol{\beta}}}[Y]$ **in the MLE formulation.**

⟹ Implication: MLE formulation is **sensitive** to estimation error of $\widehat{\boldsymbol{\beta}}$.

PROPOSED METHOD: DITTO

Hitting Times



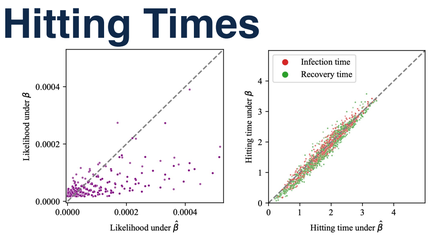
- First infection time: $h_u^{\mathrm{I}}(\boldsymbol{Y}) \coloneqq \min\{T+1, \min\{t \geq 0 : y_{t,u} \geq \mathrm{I}\}\}$.
- First recovery time: $h_u^{\mathrm{R}}(\boldsymbol{Y}) \coloneqq \min\{T+1, \min\{t \geq 0 : y_{t,u} \geq \mathrm{R}\}\}$.

Stability of Posterior Expected Hitting Times

(**Key** theoretical observation)

- **Theorem 4**. *Under SIR model and mild conditions, for any possible snapshot* $\boldsymbol{y}_T$*,*
$$\nabla_{\boldsymbol{\beta}} \mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\boldsymbol{\beta}}|\boldsymbol{y}_T} [h_u^{\mathrm{I}}(\boldsymbol{Y})] = O(1),$$
$$\nabla_{\boldsymbol{\beta}} \mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\boldsymbol{\beta}}|\boldsymbol{y}_T} [h_u^{\mathrm{R}}(\boldsymbol{Y})] = O(1).$$

➢**Stable** even for small $\boldsymbol{\beta}$



(a) $P_{\boldsymbol{\beta}}[Y]$ **vs** $P_{\widehat{\boldsymbol{\beta}}}[Y]$ **in the MLE formulation.** **(b)** $\mathbb{E}_{P_{\boldsymbol{\beta}}|\boldsymbol{y}_T}[h_u^x(Y)]$ **vs** $\mathbb{E}_{P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T}[h_u^x(Y)]$ **in the barycenter formulation.**

Figure 2: Sensitivity of the MLE formulation vs stability of the barycenter formulation.

❖ Proof Idea: Use Lemma 8 again to characterize $P_{\boldsymbol{\beta}}[\boldsymbol{Y}]$ and $P_{\boldsymbol{\beta}}[\boldsymbol{y}_T]$.

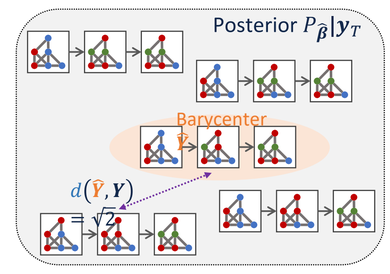
MLE Formulation → Barycenter Formulation

- Estimation error of $\boldsymbol{\beta}$ is **unavoidable**.

✗ The MLE formulation is **sensitive** to estimation error of $\boldsymbol{\beta}$.

Posterior expected *hitting times* are **stable** against estimation error of $\boldsymbol{\beta}$.

- Our solution: A novel ***barycenter formulation*** based on *hitting times.*
- ***History distance*** d***:***
(Euclidean distance with *hitting times* as coordinates)
$$d(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) \coloneqq \sqrt{\sum_{u\in\mathcal{V}} \sum_{x=\mathrm{I},\mathrm{R}} \left(h_u^x(\widehat{\boldsymbol{Y}}) - h_u^x(\boldsymbol{Y})\right)^2}.$$
- ***Barycenter formulation:*** (Finding the *barycenter* $\widehat{\boldsymbol{Y}}$ of the posterior distribution $P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T$ w.r.t. the history distance d)
$$\min_{\widehat{\boldsymbol{Y}}} \mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T} \left[d(\widehat{\boldsymbol{Y}}, \boldsymbol{Y})^2\right].$$



Solution to the Barycenter Formulation

- Bias–variance decomposition:
$$\mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T} \left[d(\widehat{\boldsymbol{Y}}, \boldsymbol{Y})^2\right] = \sum_{u\in\mathcal{V}} \sum_{x=\mathrm{I},\mathrm{R}} \left(\left(h_u^x(\widehat{\boldsymbol{Y}}) - \mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T} [h_u^x(\boldsymbol{Y})]\right)^2 + \mathop{\mathrm{Var}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T} [h_u^x(\boldsymbol{Y})]\right).$$
- Variances are **constant** w.r.t. $\widehat{\boldsymbol{Y}}$ ⟹ Optimal solution $\widehat{\boldsymbol{Y}}$:
$$h_u^x(\widehat{\boldsymbol{Y}}) = \mathrm{round}\left(\mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T} [h_u^x(\boldsymbol{Y})]\right), \qquad x = \mathrm{I}, \mathrm{R};$$
$$\hat{y}_{t,u} = \begin{cases} \mathrm{S}, & \text{for } 0 \leq t < h_u^{\mathrm{I}}(\widehat{\boldsymbol{Y}}); \\ \mathrm{I}, & \text{for } h_u^{\mathrm{I}}(\widehat{\boldsymbol{Y}}) \leq t < h_u^{\mathrm{R}}(\widehat{\boldsymbol{Y}}); \\ \mathrm{R}, & \text{for } h_u^{\mathrm{R}}(\widehat{\boldsymbol{Y}}) \leq t \leq T. \end{cases}$$

M–H MCMC for Posterior Expectation Estimation

- How to estimate $\mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T} [h_u^x(\boldsymbol{Y})]$? ➢Recall: **Intractable** to compute $P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{Y}|\boldsymbol{y}_T]$.
- Our solution: ***M–H MCMC*** [1, 2].
1. Design a ***proposal*** distribution $Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)[\cdot]$ over **possible** histories.
2. Each step of M–H MCMC samples L histories $\boldsymbol{Y}^{(s,i)} \sim Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)$, $i = 1, \dots, L$.
3. Each previous history $\boldsymbol{Y}^{(s-1,i)}$ is replaced by the new history $\boldsymbol{Y}^{(s,i)}$ with probability:
$$\min\left\{1, \frac{P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{Y}^{(s,i)}|\boldsymbol{y}_T] Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)[\boldsymbol{Y}^{(s-1,i)}]}{P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{Y}^{(s-1,i)}|\boldsymbol{y}_T] Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)[\boldsymbol{Y}^{(s,i)}]}\right\} = \min\left\{1, \frac{P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{Y}^{(s,i)}] Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)[\boldsymbol{Y}^{(s-1,i)}]}{P_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{Y}^{(s-1,i)}] Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)[\boldsymbol{Y}^{(s,i)}]}\right\}.$$ ➢**Tractable** to compute

➢The Markov chain $\langle \boldsymbol{Y}^{(s,i)} \rangle$ **provably converges** to the posterior distribution $P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T$.

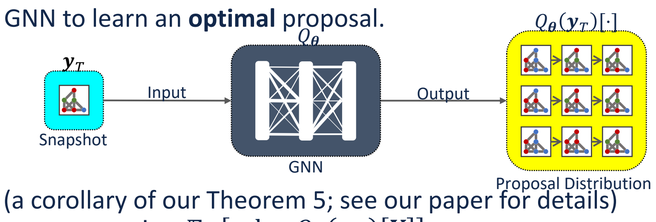
$$\mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T} [h_u^x(\boldsymbol{Y})] \approx \frac{1}{L} \sum_{i=1}^{L} h_u^x(\boldsymbol{Y}^{(s,i)}), \qquad s \to +\infty.$$

Learning an Optimal Proposal for M–H MCMC

- The ***convergence rate*** of M–H MCMC depends critically on the proposal $Q_{\boldsymbol{\theta}}$.

➢$Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)$ **closer** to $P_{\widehat{\boldsymbol{\beta}}}|\boldsymbol{y}_T$ ⟹ **Higher** rate of convergence.

- Our solution: Use a GNN to learn an **optimal** proposal.



- Objective function: (a corollary of our Theorem 5; see our paper for details)
$$\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{\boldsymbol{Y}\sim P_{\widehat{\boldsymbol{\beta}}}} [-\log Q_{\boldsymbol{\theta}}(\boldsymbol{y}_T)[\boldsymbol{Y}]].$$

MAIN EXPERIMENTS

Performance for Real-World Diffusion

Table 4: Results for real-world diffusion. "OOM" indicates "out of memory."

Type	Method	BrFarmers F1↑	BrFarmers NRMSE↓	Pol F1↑	Pol NRMSE↓	Covid F1↑	Covid NRMSE↓	Hebrew F1↑	Hebrew NRMSE↓
Supervised (w/ estimated $\widehat{\boldsymbol{\beta}}$)	GCN	.5409	.6660	.4458	.4946	.3162	.5214	.3350	.6070
	GIN	.4548	.6565	.5203	.4767	.3226	.4951	.3704	.7816
	BRITS	.5207	.3995	OOM		.3524	.5333	.3120	.6584
	GRIN	.8003	.2425	.6518	.3731	.5448	.3040	.5916	**.2212**
	SPIN	**.8268**	**.2084**	OOM		.5917	.2932	.5178	.3330
MLE	DHREC	.6131	.4150	.7023	.3398	.3540	.6023	.6251	.4169
	CRI	.6058	.4444	.7468	.2942	.4170	.5487	.5344	.3552
Barycenter	DITTO (ours)	.8206	.2142	**.7471**	**.2903**	**.6240**	**.2637**	**.6411**	.2983

➢ *BrFarmers* is very close to SI.

DITTO: Consistently strong performance across all datasets.

MLE/Supervised: Bad when real diffusion deviates from SI/SIR.

Comparison with MLE-Based Methods

Table 5: Comparison with MLE-based methods on synthetic SI and SIR diffusion. *We use GRIN trained with true $\boldsymbol{\beta}$ as the ideal performance and calculate *Gap* w.r.t. this ideal performance.

Type	Method	BA-SI F1↑	BA-SI Gap↓	BA-SI NRMSE↓	BA-SI Gap↓	ER-SI F1↑	ER-SI Gap↓	ER-SI NRMSE↓	ER-SI Gap↓	Oregon2-SI F1↑	Oregon2-SI Gap↓	Oregon2-SI NRMSE↓	Oregon2-SI Gap↓	Prost-SI F1↑	Prost-SI Gap↓	Prost-SI NRMSE↓	Prost-SI Gap↓
Ideal	GRIN	.8404*	–	.2123*	–	.8317*	–	.2166*	–	.8320*	–	.2249*	–	.8482*	–	.2155*	–
MLE	DHREC	.6026	28.30%	.4644	118.75%	.6281	24.48%	.4495	107.53%	.6038	27.43%	.4101	82.35%	.6558	22.68%	.4138	92.02%
	CRI	.7502	10.73%	.3012	41.87%	.7797	6.25%	.2744	26.69%	.8183	1.65%	.2438	8.40%	.8083	4.70%	.2491	15.59%
Barycenter	DITTO (ours)	**.8384**	**0.24%**	**.2139**	**0.75%**	**.8269**	**0.58%**	**.2225**	**2.72%**	**.8280**	**0.48%**	**.2289**	**1.78%**	**.8327**	**1.83%**	**.2317**	**7.52%**
Type	Method	BA-SIR F1↑	BA-SIR Gap↓	BA-SIR NRMSE↓	BA-SIR Gap↓	ER-SIR F1↑	ER-SIR Gap↓	ER-SIR NRMSE↓	ER-SIR Gap↓	Oregon2-SIR F1↑	Oregon2-SIR Gap↓	Oregon2-SIR NRMSE↓	Oregon2-SIR Gap↓	Prost-SIR F1↑	Prost-SIR Gap↓	Prost-SIR NRMSE↓	Prost-SIR Gap↓
Ideal	GRIN	.7867*	–	.1692*	–	.7626*	–	.2484*	–	.8024*	–	.1651*	–	.8067*	–	.1652*	–
MLE	DHREC	.5080	35.43%	.4722	179.08%	.5500	27.88%	.4423	78.06%	.6044	24.68%	.4478	171.23%	.6268	22.30%	.4326	161.86%
	CRI	.5994	23.81%	.3356	98.35%	.6129	19.63%	.3109	25.16%	.5761	28.20%	.3576	116.60%	.5738	28.87%	.3406	106.17%
Barycenter	DITTO (ours)	**.7783**	**1.07%**	**.1633**	**−3.49%**	**.7734**	**−1.42%**	**.1679**	**−32.41%**	**.7928**	**1.20%**	**.1707**	**3.39%**	**.7929**	**1.71%**	**.1690**	**2.30%**

DITTO: Stably achieves the strongest performance.

MLE: Performance vary largely across datasets due to **sensitivity**.

ACKNOWLEDGEMENTS

NSF amazon DARPA USDA NIFA DEPARTMENT OF HOMELAND SECURITY DEVCOM ARMY RESEARCH LABORATORY C3.ai Digital Transformation Institute IBM