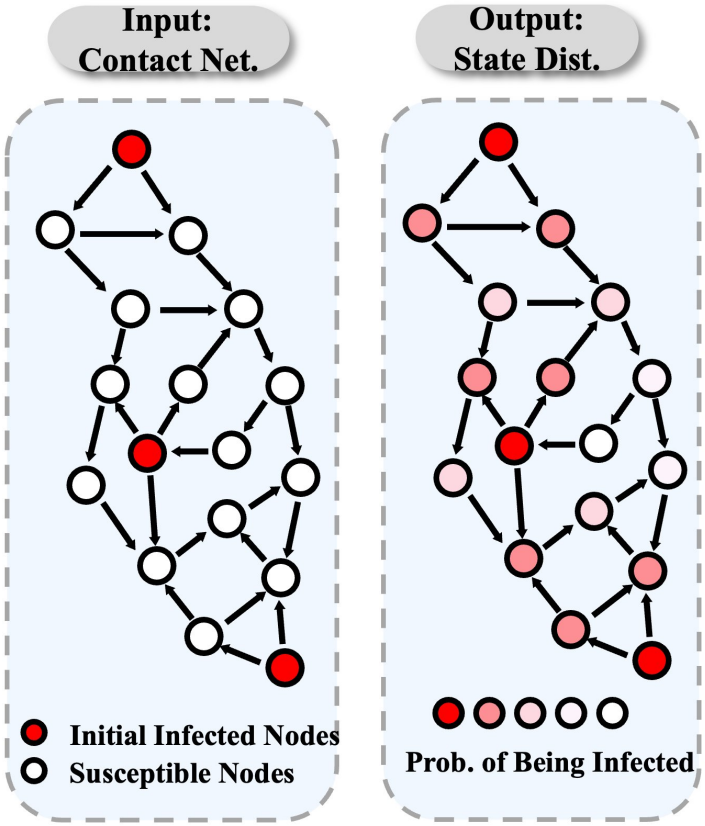
Scaling Epidemic Inference on Contact Networks: Theory and Algorithms

[1]Guanghui Min [1]Yinhan He [1]Chen Chen

[1]University of Virginia

Background Introduction

- Epidemic Modeling Importance
  - Large-scale outbreaks like COVID-19 have highlighted the need for accurate modeling and prediction of disease spread dynamics on contact networks.
- Epidemic Inference
  - The goal is to estimate the infection probability distribution of each individual (node) given the network structure, initial infections, and epidemic parameters.



Challenge

- Limitations of Traditional Models
  - Population-level ODE models (e.g., SIR, SEIR) assume homogeneous mixing, thus missing local heterogeneity and individual-level infection dynamics within the network.
- Monte Carlo (MC) Simulations as Standard Tool
  - MC simulations are widely used for epidemic inference because they make no structural or distributional assumptions, offering robust estimates. ⇒ Yet, they require hundreds to thousands of runs for statistical reliability, leading to prohibitively high computational costs on large networks.
- Research Gap
  - Despite extensive use, there is no theoretical understanding of how network topology and epidemic parameters influence MC variance and convergence behavior.

Theoretical Insight

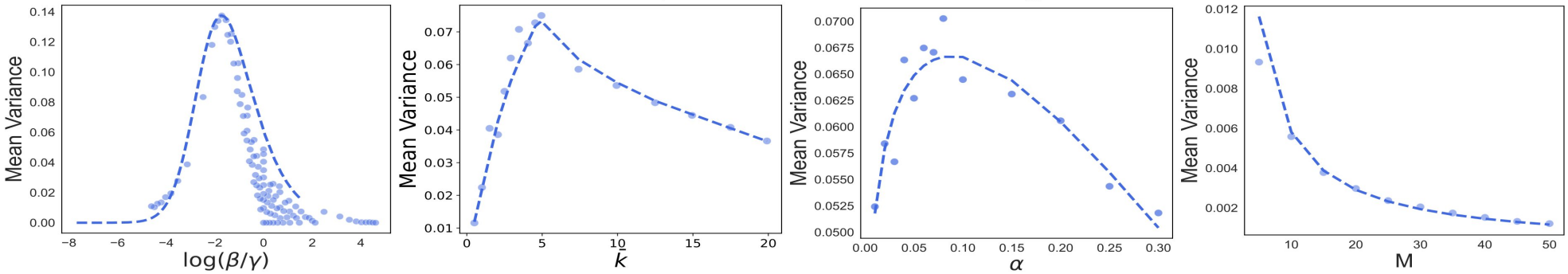
Theorem 3.1 quantifies how the variance of Monte Carlo (MC) estimators for node infection probability fundamentally depends on **epidemic parameters** (β, γ), **network structure** (average degree $\bar{k}$ and diameter D), **initial infection fraction** α, and **the number of simulations** M.

It establishes a non-zero lower bound on the average estimator variance:

$$\frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}(\hat{p}_i - p_i) \gtrsim \frac{1}{2M}\min\{1-(1-p_0)^{c\bar{k}\alpha}, (1-p_0)^{c\bar{k}\alpha}\},$$

where

$$p_0 := \left(\frac{\beta}{\beta+\gamma}\right)^{\ell}, \quad \ell := \min\{D, \frac{\log N}{\log \bar{k}}\}$$



Influence of key factors on MC estimator variance.

Methodology

- Core Idea
  - RAPID builds upon the Probabilistic Infection Dynamics (PID) **message-passing** equations and introduces a **residual-driven asynchronous** propagation mechanism that updates only where changes are significant.
- Base: Message Passing Foundation
  - Each node i updates its infection probability P_I^i using local messages from its in-neighbors:

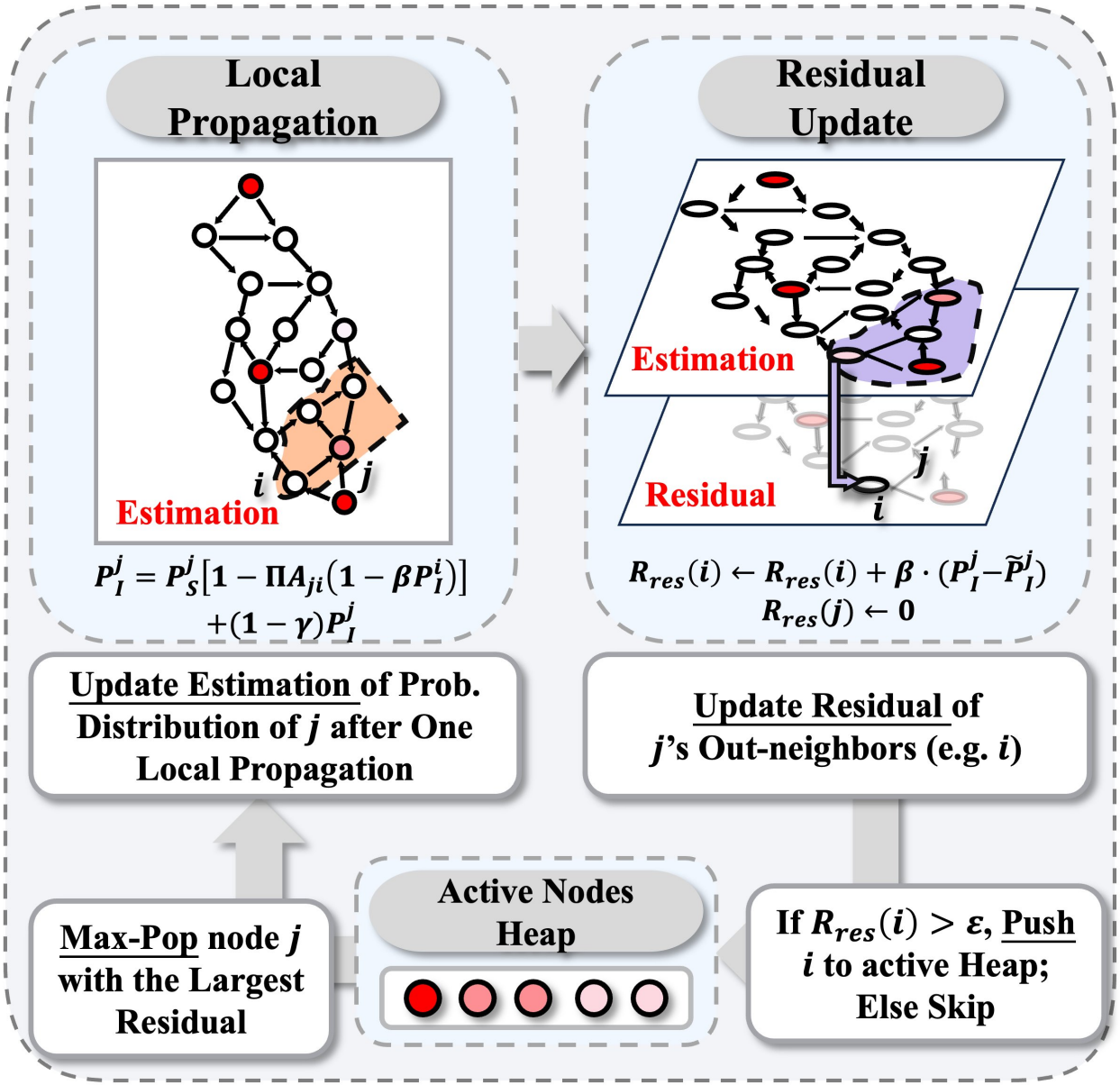
$$P_S^i(t+1) = P_S^i(t) \prod_{j \in \mathcal{V}} \mathbf{A}_{ji}(1 - \beta P_I^j(t))$$

$$P_I^i(t+1) = P_S^i(t)[1 - \prod_{j \in \mathcal{V}} \mathbf{A}_{ji}(1 - \beta P_I^j(t))] + (1 - \gamma) P_I^i(t)$$

  This standard PID update defines **RAPID**'s computational base.
- Residual-Driven Propagation
  - To quantify "*how much information remains to be propagated*", we define the propagation residual at node i:

$$R_{\text{res}}(i) = \beta \sum_{j \in \mathcal{V}} \mathbf{A}_{ji} \, (P_I^j - \tilde{P}_I^j),$$

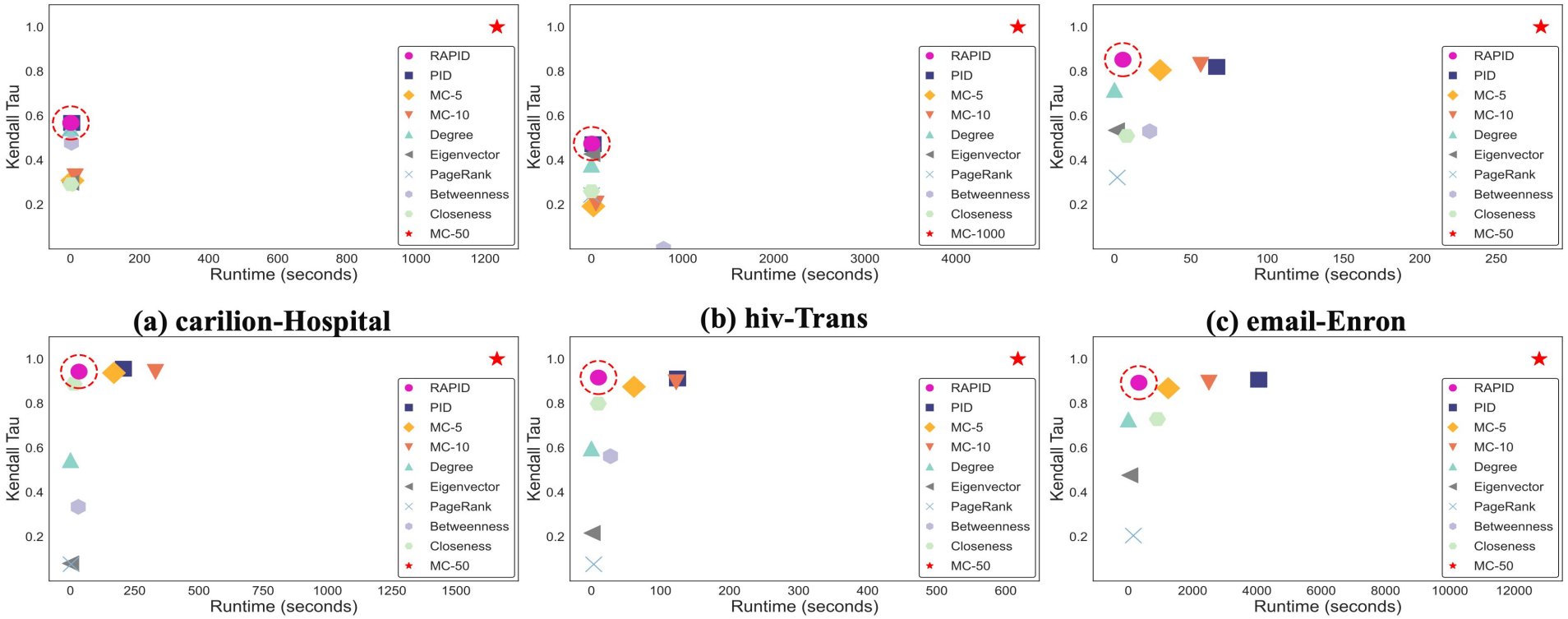
Methodology



Experiments

- Setup
  - Six real-world directed networks (carilion-Hospital, hiv-Trans, soc-Pokec, etc.).
- Baselines
  - MC-5/10/50, PID, and centrality heuristics.
- Metrics
  - Kendall-tau Coefficient, Mean Absolute Error, Precision/Recall/F1, Runtime.

Experiments



(a) carilion-Hospital (b) hiv-Trans (c) email-Enron

(d) email-EuAll (e) soc-Epinions (f) soc-Pokec

Trade-off between Kendall-Tau and Runtime across six datasets.

Experiments

Dataset	MC-5	MC-10	PID	RAPID
carilion-Hospital	13.01±0.35	10.78±0.44	5.67±0.51	**2.64**±0.49
hiv-Trans	5.12±0.12	3.24±0.11	6.43±0.52	**1.27**±0.47
email-Enron	7.60±0.02	5.98±0.03	8.70±0.04	**4.66**±0.00
email-EuAll	2.09±0.01	1.63±0.01	1.36±0.02	**1.03**±0.01
soc-Epinions	5.48±0.03	4.31±0.02	4.92±0.03	**2.77**±0.00
soc-Pokec	4.50±0.00	3.54±0.00	3.32±0.00	**2.32**±0.00

MAE comparison (lower is better). All values are scaled by 10^{-2}. Best results are in bold.

Experiments

		carilion-Hospital[2]	hiv-Trans[2]	email-Enron	email-EuAll	soc-Epinions	soc-Pokec
MC-5	t	5.81±0.51	21.87±3.15	29.84±1.43	169.46±8.23	59.69±1.26	1241.00±18.79
	Δ	5.43×	5.16×	5.41×	5.06×	5.82×	3.78×
MC-10	t	13.46±1.14	49.94±1.74	56.45±0.79	330.97±6.81	122.31±2.44	2506.60±45.46
	Δ	12.58×	11.78×	10.24×	9.88×	11.91×	7.64×
MC-50	t	1234.73±13.13	4678.18±8.96	279.26±3.09	1659.57±23.55	614.58±2.36	12782.37±237.30
	Δ	1153.95×	1103.34×	50.66×	49.52×	59.86×	38.93×
PID	t	3.56±0.01	17.91±0.14	66.95±0.29	206.18±0.65	132.60±0.62	4056.89±6.40
	Δ	3.33×	4.22×	12.14×	6.15×	12.91×	12.36×
RAPID	t	**1.07**±0.00	**4.24**±0.03	**5.51**±0.04	**33.50**±0.05	**10.27**±0.09	**328.28**±0.66

Runtime comparison across datasets (seconds, lower is better). Δ indicates the speedup factor relative to RAPID, computed as $\Delta = {}^{Baseline\ time}/_{RAPID\ time}$. On *carilion-Hospital* and *hiv-Trans*, we adopt 1000-run MC simulations as the ground truth for acceptable estimator variance.

Conclusions

- Theoretical Analysis
  - We systematically analyze the variance of Monte Carlo (MC) simulations in modeling disease spread on contact networks.
- Proposed Framework: RAPID
  - A residual-driven inference framework that estimates node-level infection probability distributions with high accuracy and low computational cost.
- Empirical Results
  - On six real-world networks, RAPID achieves the accuracy of multi-run MC while maintaining the runtime of a single simulation.
- Future Directions
  - Extend the framework to handle reinfection, time-varying parameters, and dynamic networks.

Acknowledgements

- This work was supported in part by the NVIDIA Academic Grant Program, the Commonwealth Cyber Initiative (CCI) under Award No. HV-2Q25-032, and the National Science Foundation under Grant No. 2331315. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

- We would like to thank the following funding agencies:

Commonwealth Cyber Initiative

The End

VIRGINIA DATA SCIENCE AND ANALYTICS (VAST) LAB

Thanks for listening!