Scaling Epidemic Inference on Contact Networks: Theory and Algorithms

Guanghui Min, Yinhan He, Chen Chen

University of Virginia

Motivations & Contributions

Challenges of Epidemic Inference

1. **High Computational Cost of Monte Carlo Simulation.** Current epidemic inference relies heavily on Monte Carlo (MC) simulation, which is statistically reliable but **computationally expensive**.
2. **Limitations of Population-Level Models.** Population-level models (e.g., SIR ODEs) assume homogeneous mixing and **cannot capture individual-level infection risk** or the impact of network topology.
3. **Shortcomings of Graph-Based Approaches.** Existing graph-based methods (like message passing or centrality) either **ignore infection dynamics** or lack theoretical guarantees on convergence and accuracy. Also the forward iteration is **computationally expensive.**

Our Contributions

- **Theoretical Analysis of MC Variance.**

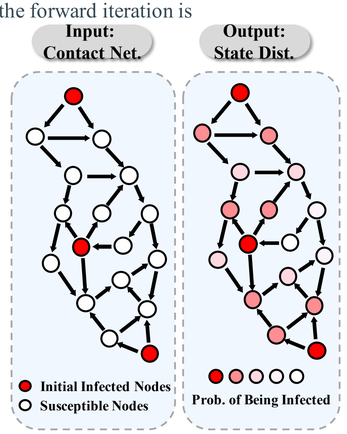
Theoretically analyze how **network topology and epidemic parameters** affect MC variance;

- **Development of the RAPID Algorithm.**

Develop a **residual-aware propagation algorithm, RAPID** that matches MC accuracy at runtime comparable to a single MC simulation through asynchronous local updates;

- **Comprehensive Empirical Validation.**

Conduct experiments across six real-world networks, demonstrating robust **accuracy**, **scalability**, and **theoretical consistency**.



Theoretical Insight

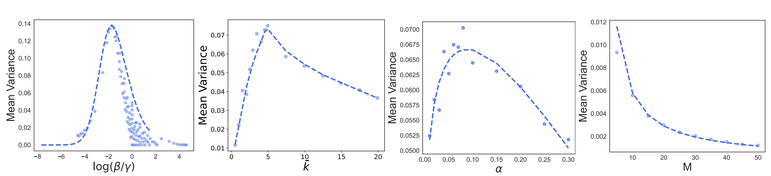
Theorem 3.1 quantifies how the variance of Monte Carlo (MC) estimators for node infection probability fundamentally depends on **epidemic parameters** (β, γ), **network structure** (average degree $\bar{k}$ and diameter D), **initial infection fraction** α, and **the number of simulations** M.

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.

RAPID Algorithm

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.

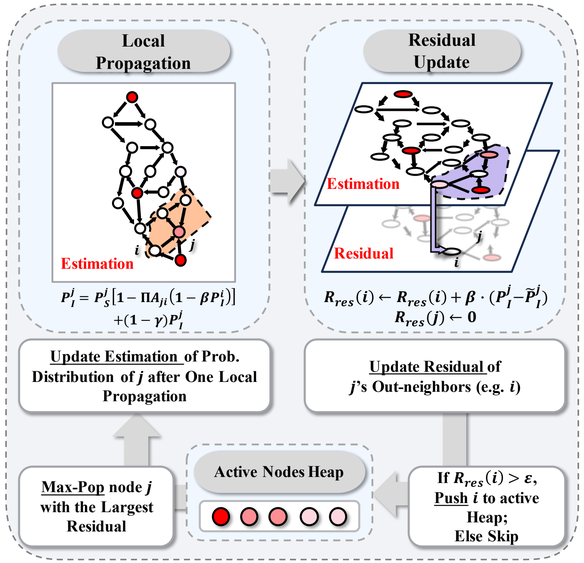


Illustration of a single iteration of RAPID.

Base: Message Passing Foundation

Each node i updates its infection probability P_I^i using local messages from its in-neighbors:

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

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

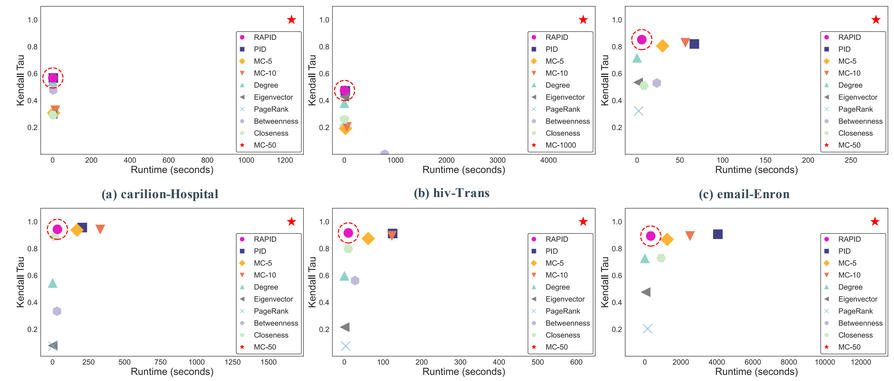
This standard PID update defines RAPID's computational base.

Residual-Driven Propagation

To quantify *"how much information remains to be propagated"*, we define the propagation residual at node i:

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

where $\tilde{P}_I^j$ is the cached infection probability before the last update. This residual measures the unbalanced local influence, which is a first-order **Jacobian approximation** of the difference between current and previously propagated messages from the in-neighbors.



(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 & Results

- **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.
- **Accuracy**: **RAPID** achieves the **lowest MAE** among baselines and **comparable Kendall-tau coefficient** to forward iteration baseline PID, benefiting from asynchronous updates that reduce the impact of small loops and accelerate convergence.
- **Efficiency**: On six datasets, **RAPID** achieves an average speedup of 5.11×, 10.67×, and 8.52× over MC-5, MC-10, and PID, respectively. The speedup is more pronounced on denser graphs and its runtime remains comparable to a single-run Monte Carlo iteration across all datasets.
- **Scalability**: Runtime grows sublinearly with network size, validating efficient inference.
- **Robustness**: Performance remains stable across wide ranges of, β, γ, and seed fractions α.

Takeaway

RAPID attains multi-run MC accuracy with consistent, dataset-wide runtime acceleration.

		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.

Paper Code