Differentially Private Network Data Release via Structural Inference

Qian Xiao, NUS
Rui Chen, HKBU
Kian-Lee Tan, NUS

KDD 2014

Idea Spotlight

- Perfect Queries ⟶ Perfect Answers

Idea Spotlight

- Perfect Queries ↛ Perfect Answers

Not always true
if under Differential Privacy !

Idea Spotlight

- Perfect Queries ~~→~~ Perfect Answers

  Not always true
  if under Differential Privacy !

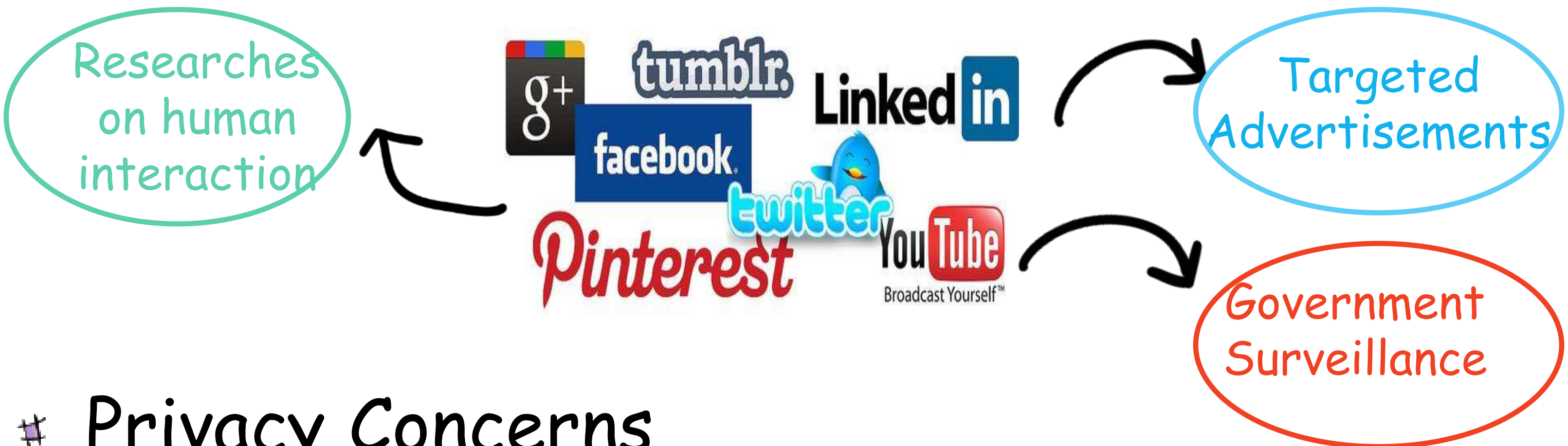
- Queries not that Perfect

  →

  Good Answers + Privacy + Social Good

Why Privacy-aware Network Data Release ???

- Increasing Demands on Network Data for Exploratory Data Analysis



- Privacy Concerns
  - Social Contacts
  - Personal opinions
  - Private communication records

Why Privacy-aware Network Data Release ???

- Emerging Privacy Standard :
  - Differential Privacy[Dwork06]
    - Resilient to attacks with **arbitrary** side information
    - **Worst case guarantee**
    - Rigorous mathematical formulation
- Prevalent Randomization Techniques to generate noisy results while satisfying DP:
  - Laplacian noise(for counting queries)
  - Exponential mechanism(for selecting discrete query outcomes)

Problem Statement

- Given an original simple graph $G = (V, E)$, find a random sanitized graph $\tilde{G}$ to release
- The goal is to
  - Approximate G's statistical properties of in $\tilde{G}$ as much as possible to preserve essential structural information
  - Satisfy edge Differential Privacy(ϵ-DP) to hide each user's connections to others

Problem Statement

- DP requires:

A randomized algorithm $\mathcal{A}$ *is* ϵ*-differential privacy if for any two neighboring graphs* G *and* G'*, and for any output* $O \in Range(\mathcal{A})$,

$$\Pr[\mathcal{A}(G) \in O] \le e^{\epsilon} \times \Pr[\mathcal{A}(G') \in O]$$

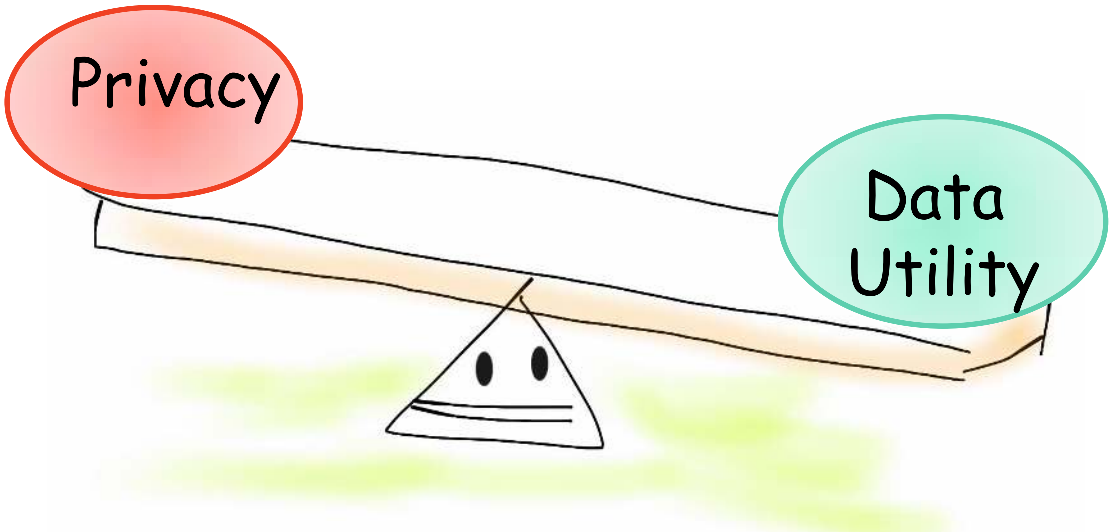
Outcome with my connection in G

Outcome without my connection in G'

! Output distribution shall not change much if any single edge is missing, that is, the sensitivity of $\mathcal{A}$ shall be limited.

Problem Statement

- To find a reasonable balance between privacy and data utility, we need to limit the query sensitivity (the dependence of noise required by DP on network size n)



State-of-the-art Approaches

- To satisfy ϵ-DP:
  - dK-2 series:
    Global sensitivity is $O(\mathrm{n})$ [Sala11, Wang13]
  - Spectral graph analysis:
    Global sensitivity is $O(\sqrt{n})$ [Wang13]

Our Approach:
Differentially Private Network Data Release via Structural Inference

- Transform edges to connection probabilities via Hierarchical Random Graph(HRG)

- Our approach's sensitivity is $O(\log n)$

Edges → Connection Probabilities

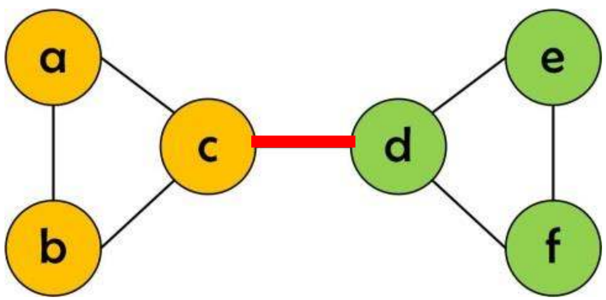
Highly sensitive !
=
Prohibitive noise

Not that sensitive
in a graph of
moderate or large size

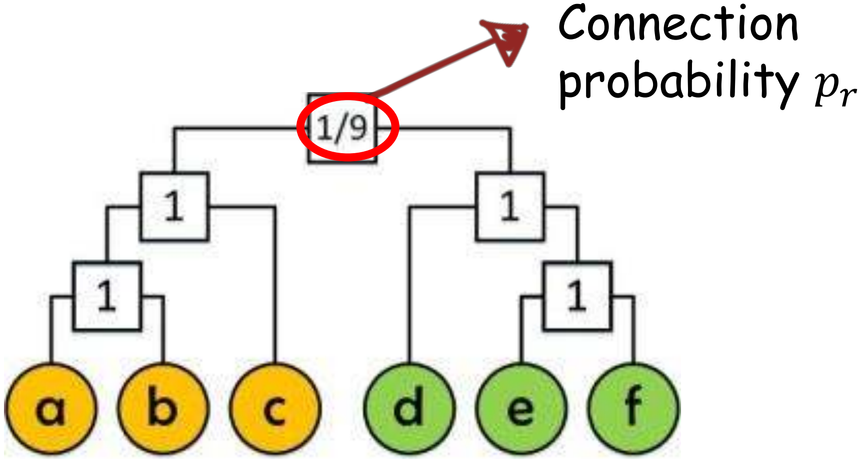
Outline

- Motivation
- Hierarchical Random Graph(HRG)
- Structural inference under DP with MCMC
- Sensitivity Analysis
- Experimental evaluation
- Conclusion

Hierarchical Random Graph



G



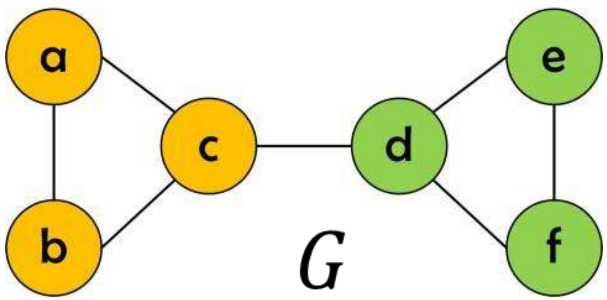
best-fitting HRG T_1, $\mathcal{L}(T_1)$=0.0433…

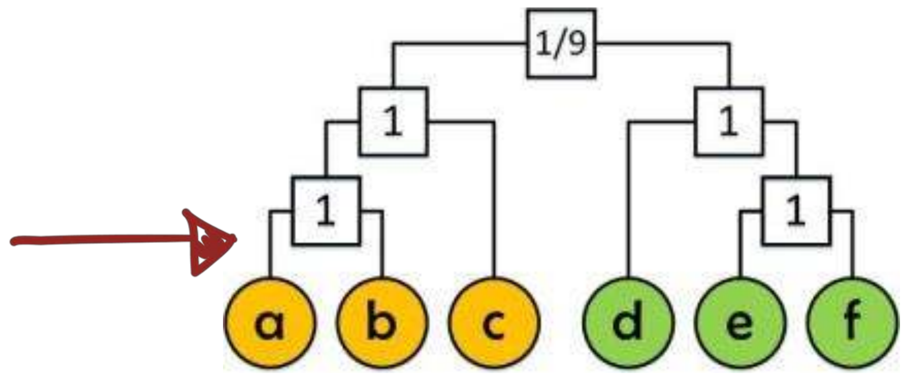
Likelihood of an HRG T:

$$\mathcal{L}(T, \{p_r\}) = \prod_{r \in T} p_r^{\,e_r} (1 - p_r)^{n_{Lr} n_{Rr} - e_r}$$

An HRG example in [Clauset07,08]

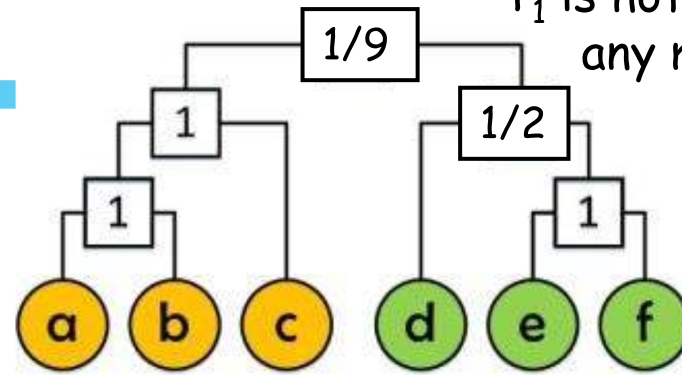
Why HRG ?





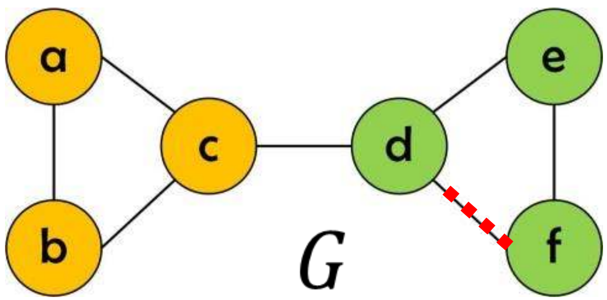
best-fitting HRG T_1,
$\mathcal{L}(T_1)$=0.0433...

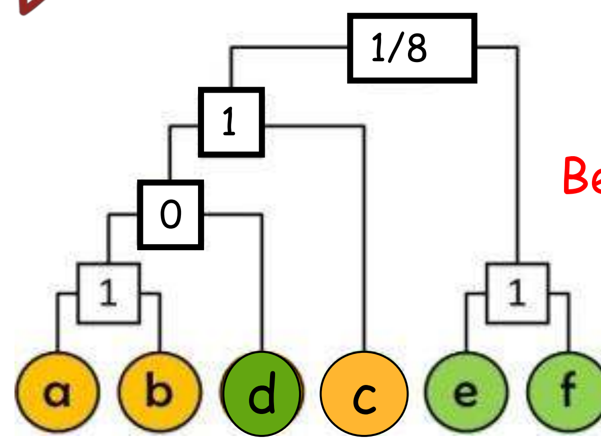
Why HRG ?



T_1 is not the best any more !

$\mathcal{L}(T_1)$=0.0108…





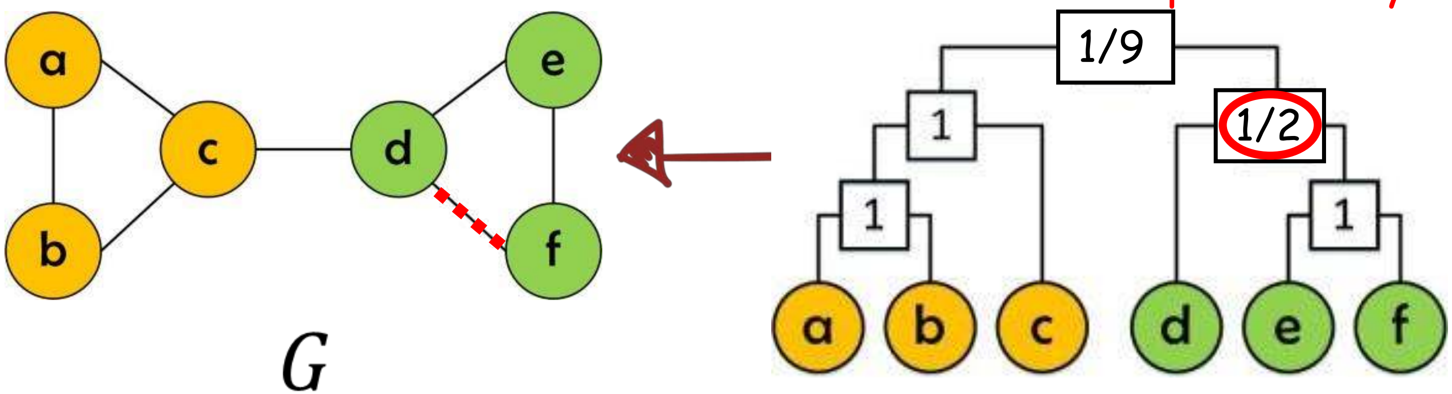
Best-fitting HRG T_2

$\mathcal{L}(T_2)$ =0.0491…

One edge missing
→
Completely different best-fitting HRG

Why HRG ?

One edge missing only affects one probability



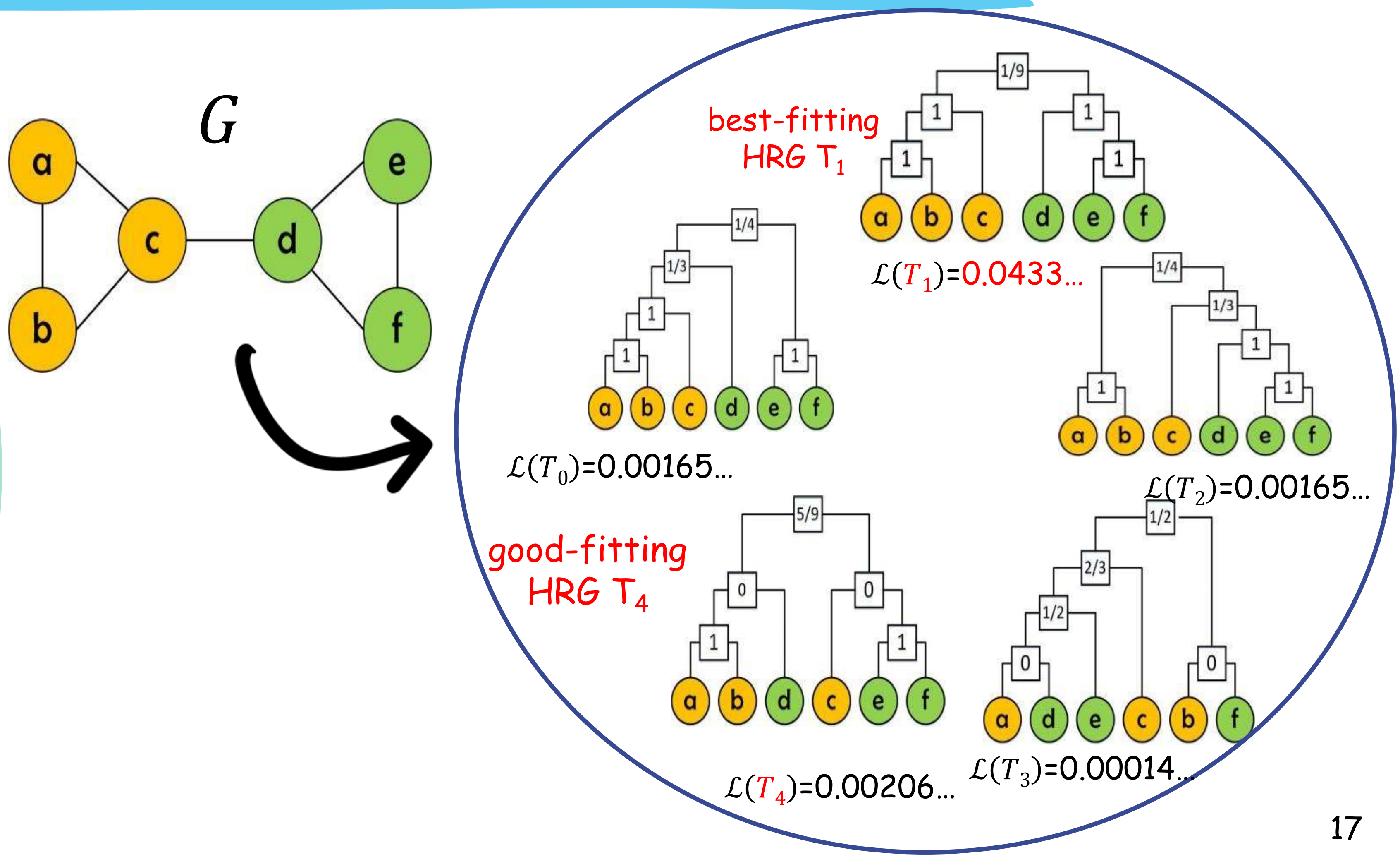
$\mathcal{L}(T_1)$ =0.0108...

Likelihood of an HRG T:

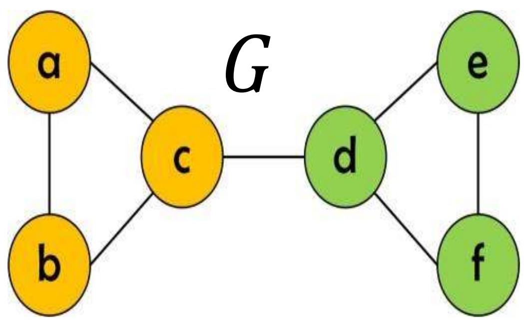
$$\mathcal{L}(T,\{p_r\}) = \prod_{r \in T} p_r{}^{e_r}(1-p_r)^{n_{Lr}n_{Rr}-e_r}$$

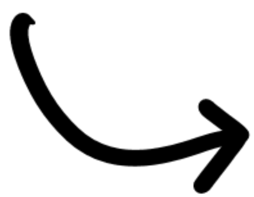
An HRG example in [Clauset07,08]

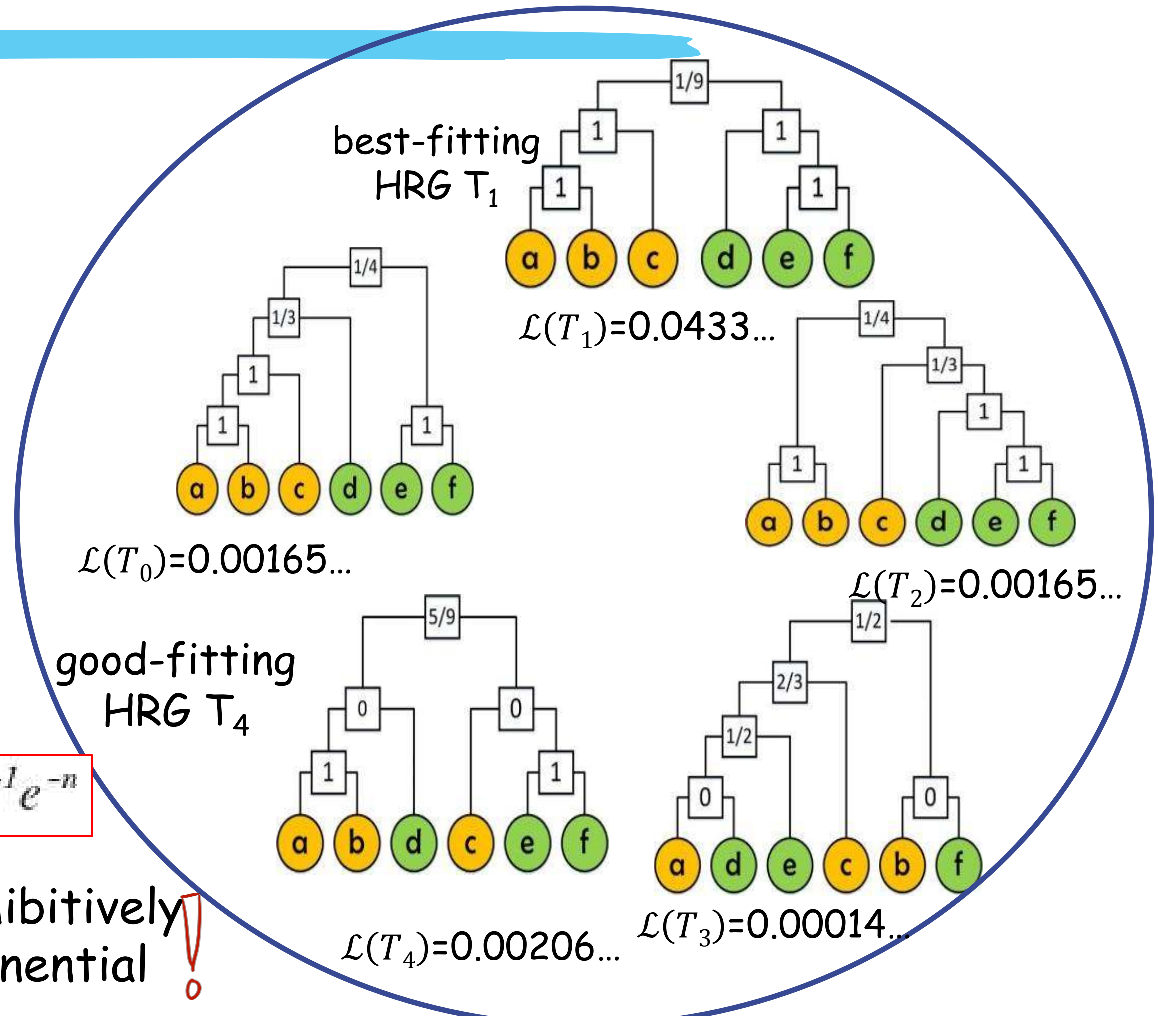
HRG space $\mathbb{T}$



HRG space $\mathbb{T}$







$|\mathbb{T}|$ is

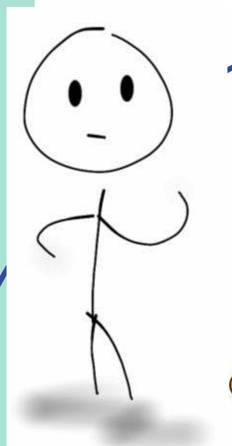
$$(2n-3)!! \approx \sqrt{2}\,(2n)^{n-1}e^{-n}$$

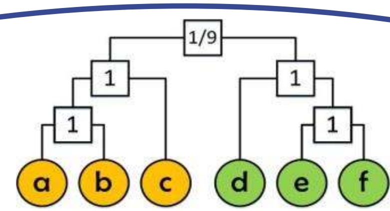
Super-exponential, prohibitively expensive to apply Exponential Mechanism directly !

Outline

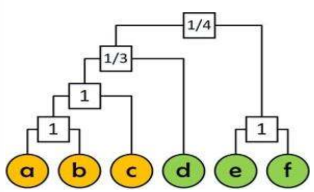
- Motivation
- Hierarchical Random Graph(HRG)
- Structural inference under DP with MCMC
- Sensitivity Analysis
- Experimental evaluation
- Conclusion

What to do with HRG ? MCMC process - 1



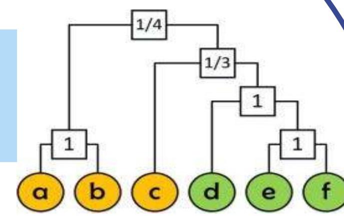


$\mathcal{L}(T_1)$=0.0433…

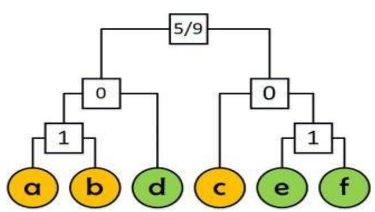


$\mathcal{L}(T_0)$=0.00165…

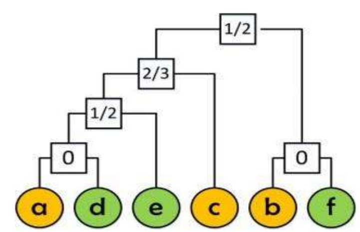
- Randomly pick an arbitrary HRG as the initial state T_0



$\mathcal{L}(T_2)$=0.00165…

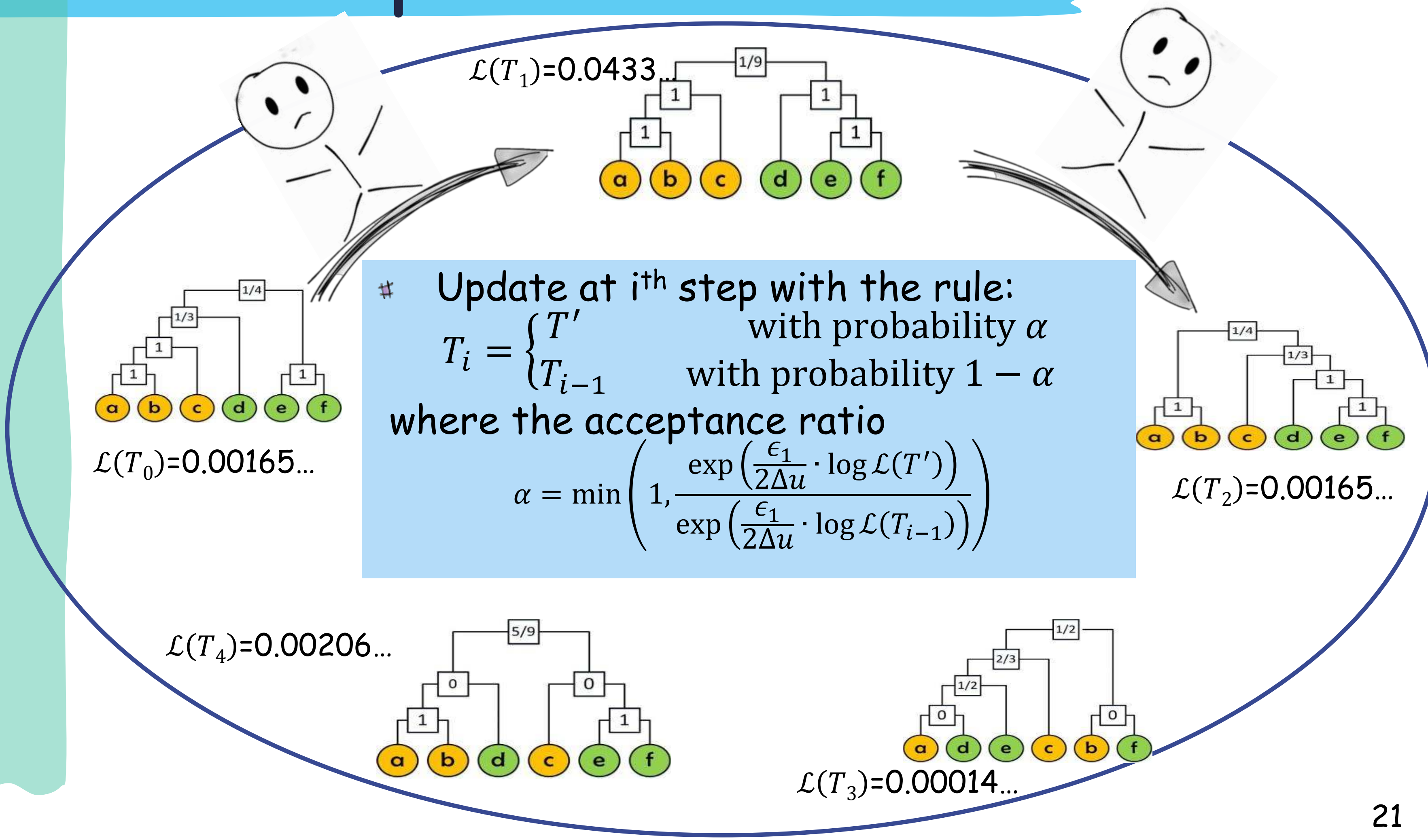


$\mathcal{L}(T_4)$=0.00206…



$\mathcal{L}(T_3)$=0.00014…

What to do with HRG ? MCMC process - 2



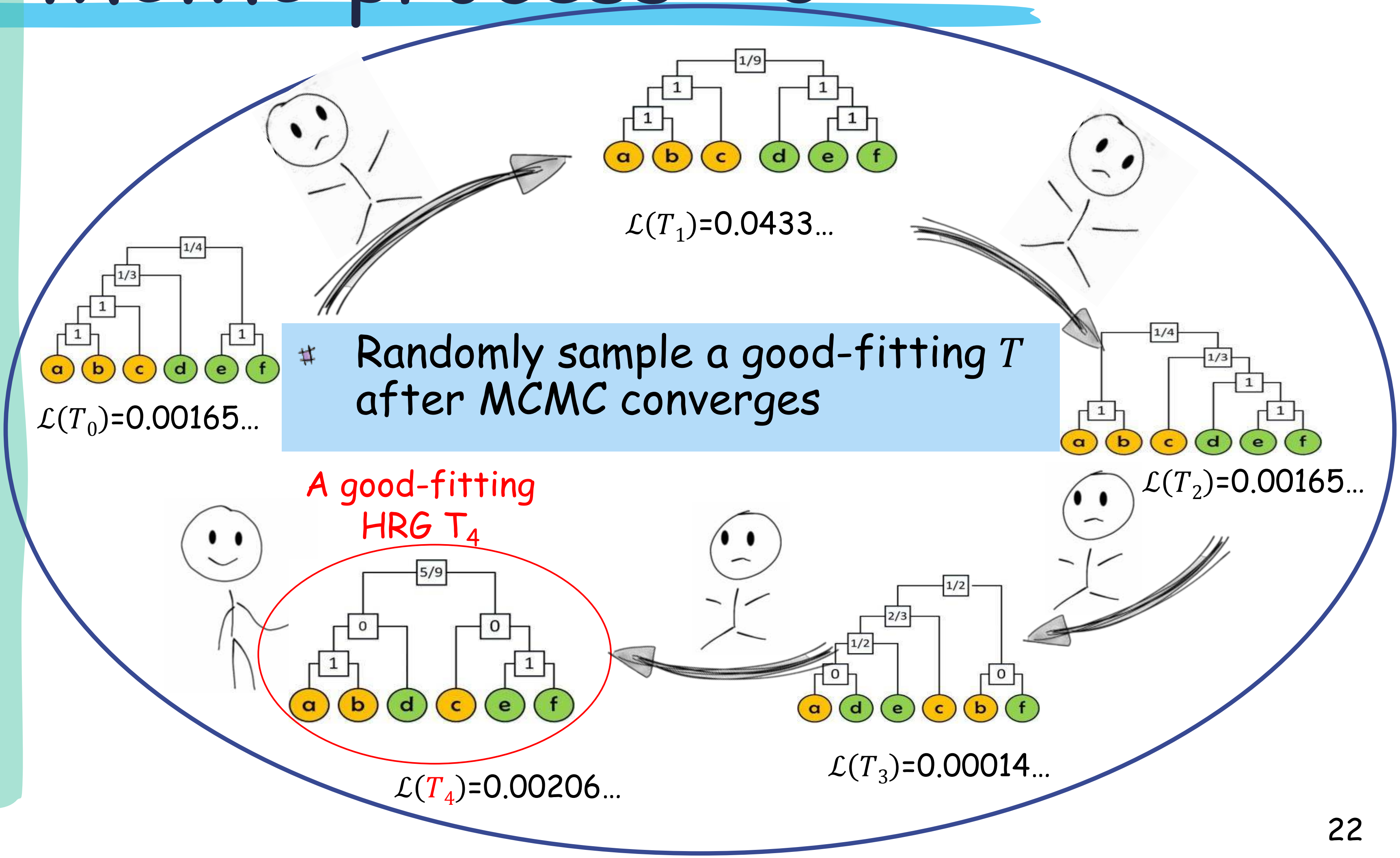
- Update at i[th] step with the rule:

$$T_i = \begin{cases} T' & \text{with probability } \alpha \\ T_{i-1} & \text{with probability } 1-\alpha \end{cases}$$

where the acceptance ratio

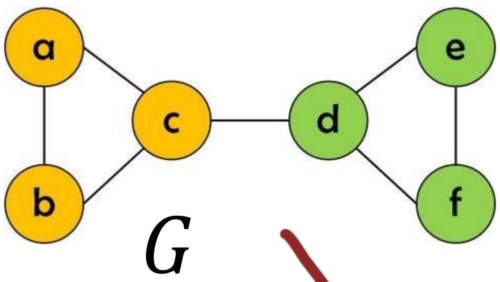
$$\alpha = \min\left(1, \frac{\exp\left(\frac{\epsilon_1}{2\Delta u} \cdot \log \mathcal{L}(T')\right)}{\exp\left(\frac{\epsilon_1}{2\Delta u} \cdot \log \mathcal{L}(T_{i-1})\right)}\right)$$

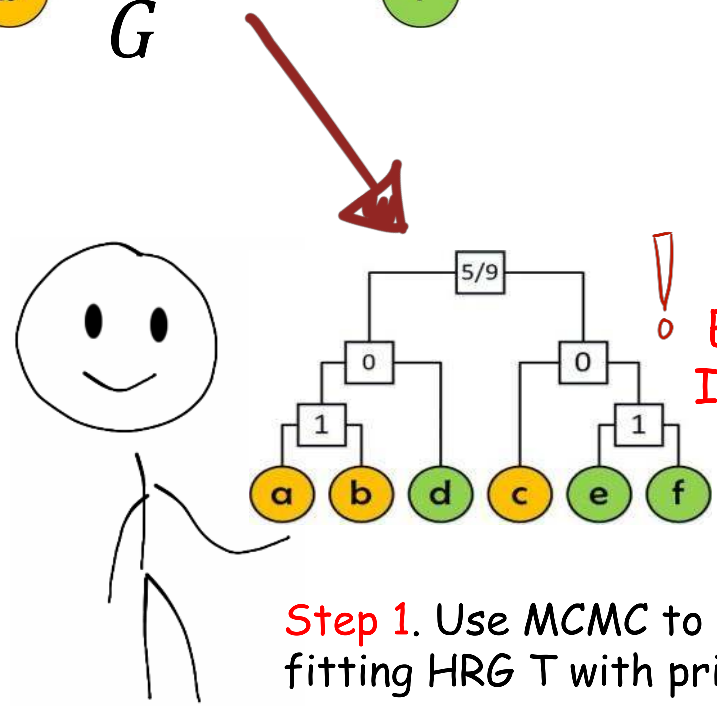
What to do with HRG ?
MCMC process - 3



- Randomly sample a good-fitting T after MCMC converges

Structure Inference under DP with MCMC

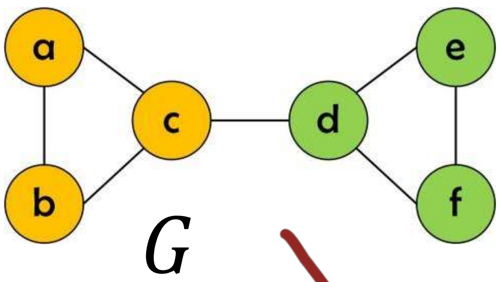




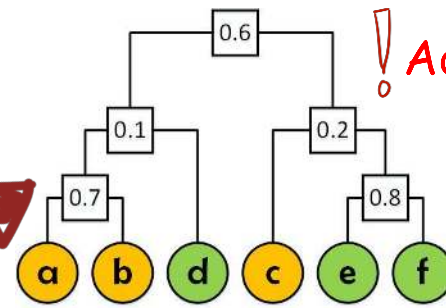
MCMC does the job of Exponential Mechanism. It satisfies DP. [Shen13]

Step 1. Use MCMC to sample a good-fitting HRG T with privacy budget ϵ_1

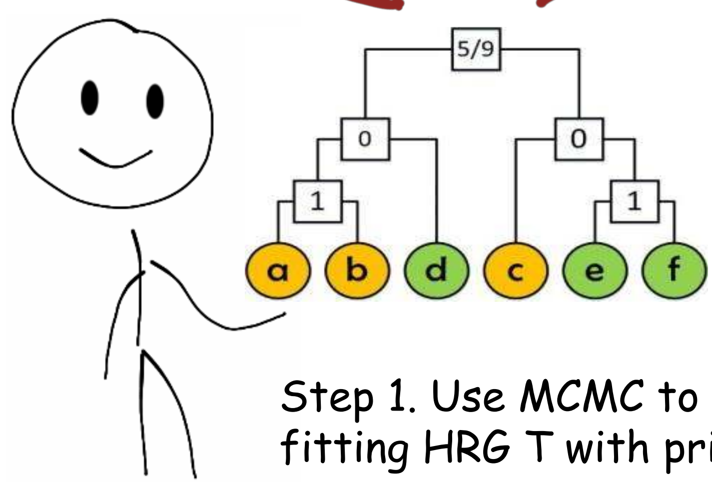
Structure Inference under DP with MCMC



Step 2. Perturb connection probabilities with privacy budget ϵ_2

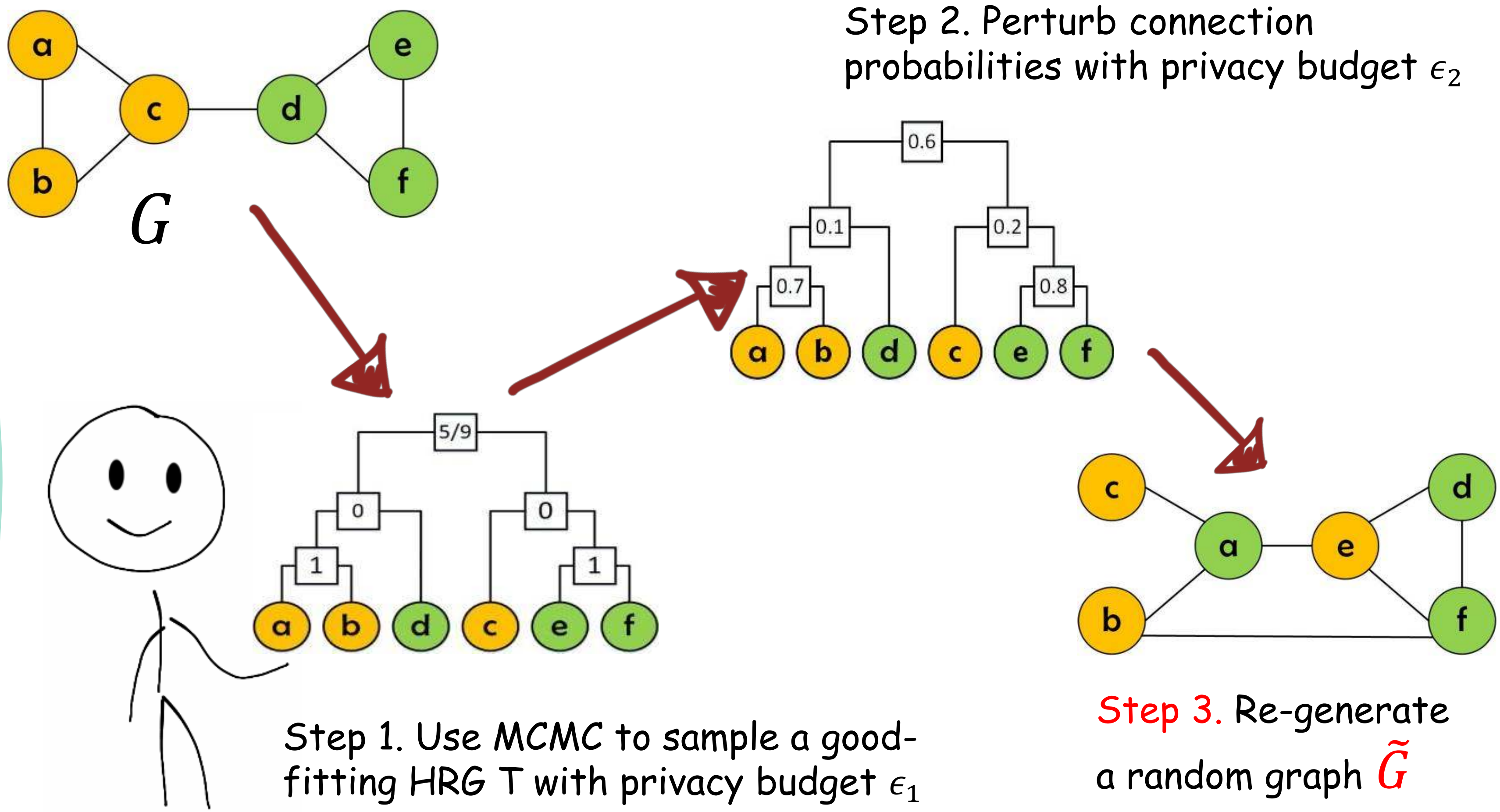


! Add Laplacian noise



Step 1. Use MCMC to sample a good-fitting HRG T with privacy budget ϵ_1

Structure Inference under DP with MCMC

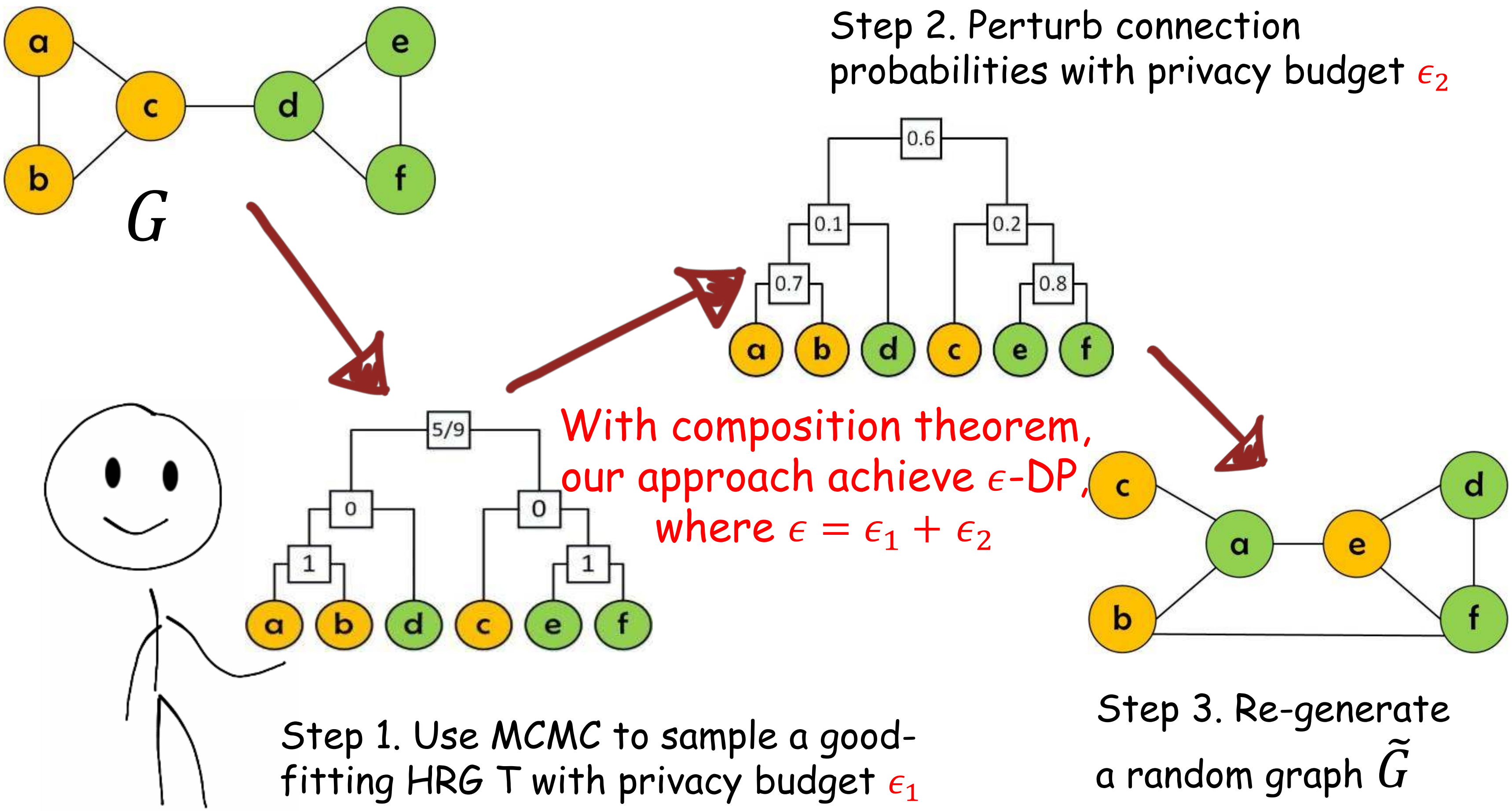


Step 2. Perturb connection probabilities with privacy budget ϵ_2

Step 1. Use MCMC to sample a good-fitting HRG T with privacy budget ϵ_1

Step 3. Re-generate a random graph $\tilde{G}$

Structure Inference under DP with MCMC



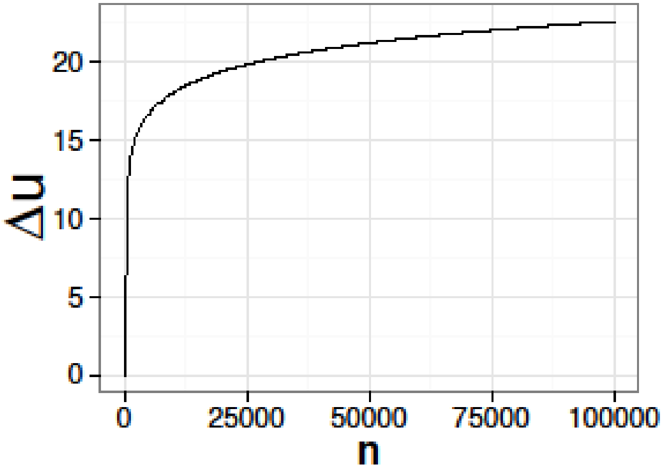
Outline

- Motivation
- Hierarchical Random Graph(HRG)
- Structural inference under DP with MCMC
- Sensitivity Analysis
- Experimental evaluation
- Conclusion

Sensitivity Analysis

- Global sensitivity:
  $$\Delta u = \max_{T \in \mathbb{T}, G, G'} |\log \mathcal{L}(T, G') - \log \mathcal{L}(T, G)|$$

- Δu is $O(\log n)$



Outline

- Motivation
- Hierarchical Random Graph(HRG)
- Structural inference under DP with MCMC
- Sensitivity Analysis
- Experimental evaluation
- Conclusion

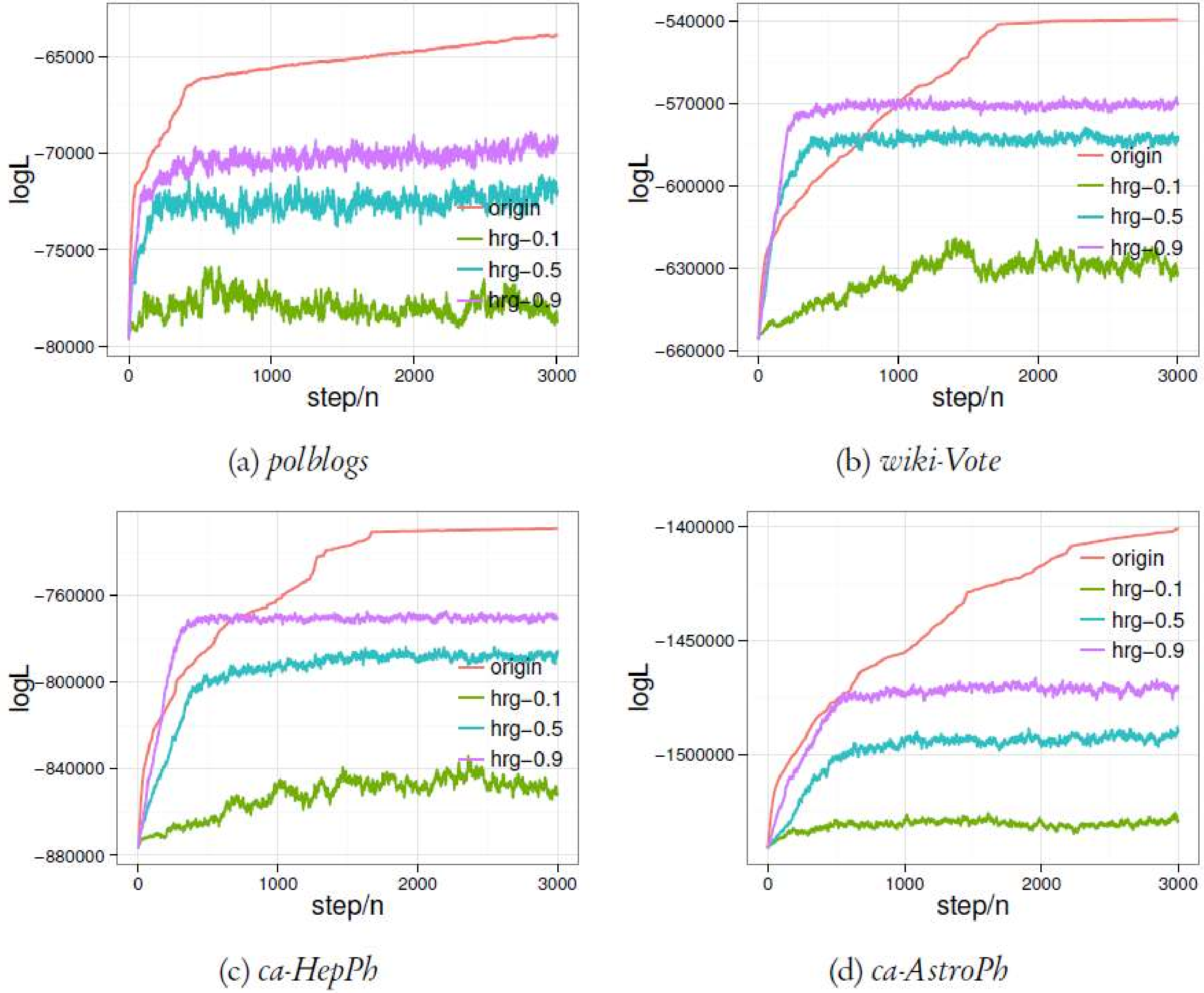
Datasets

Network dataset statistics

Dataset	#Nodes	#Edges	Max Degree Pair
polblogs	1,224	16,715	(351, 277)
wiki-Vote	7,115	100,762	(1065, 773)
ca-HepPh	12,008	118,489	(491, 486)
ca-AstroPh	18,772	198,050	(504, 420)

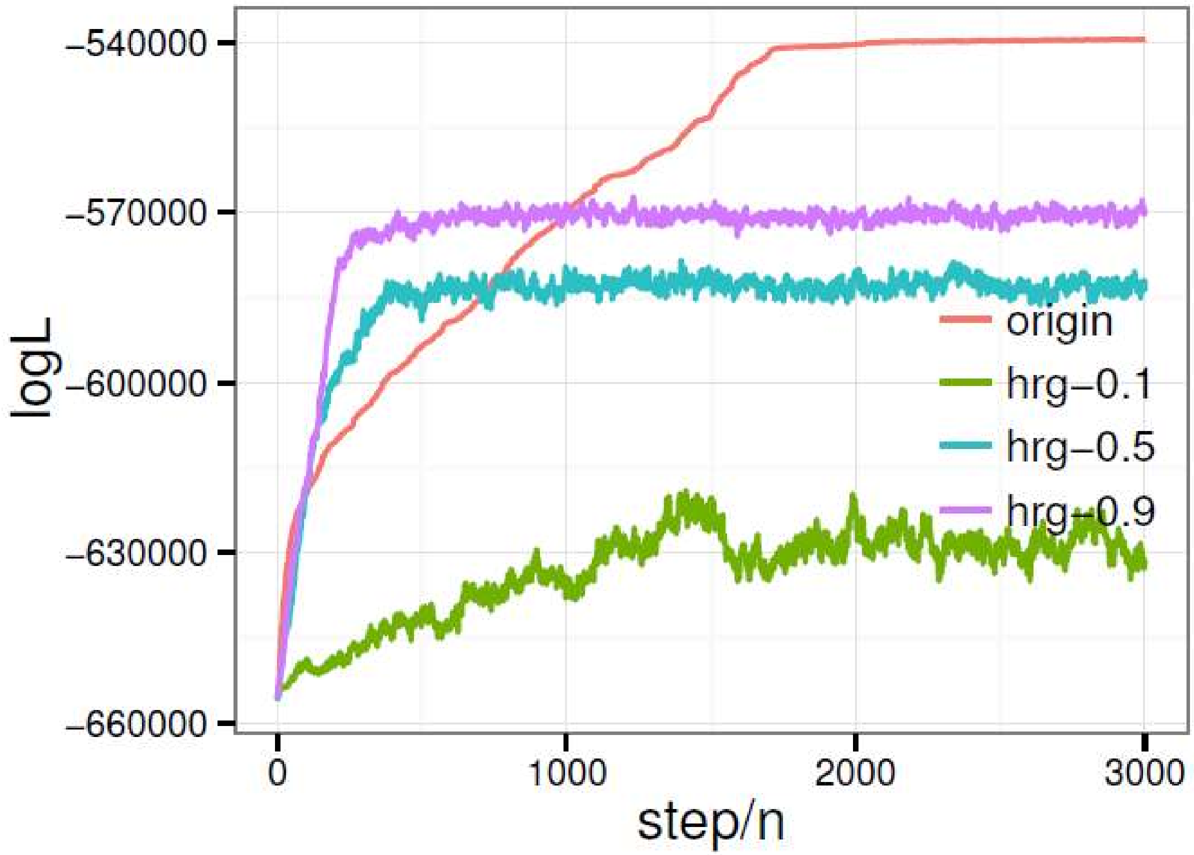
- All are real-life data

MCMC Convergence Study on $\log \mathcal{L}$



Trace of $\log \mathcal{L}$ as a function of the number of MCMC steps, normalized by n

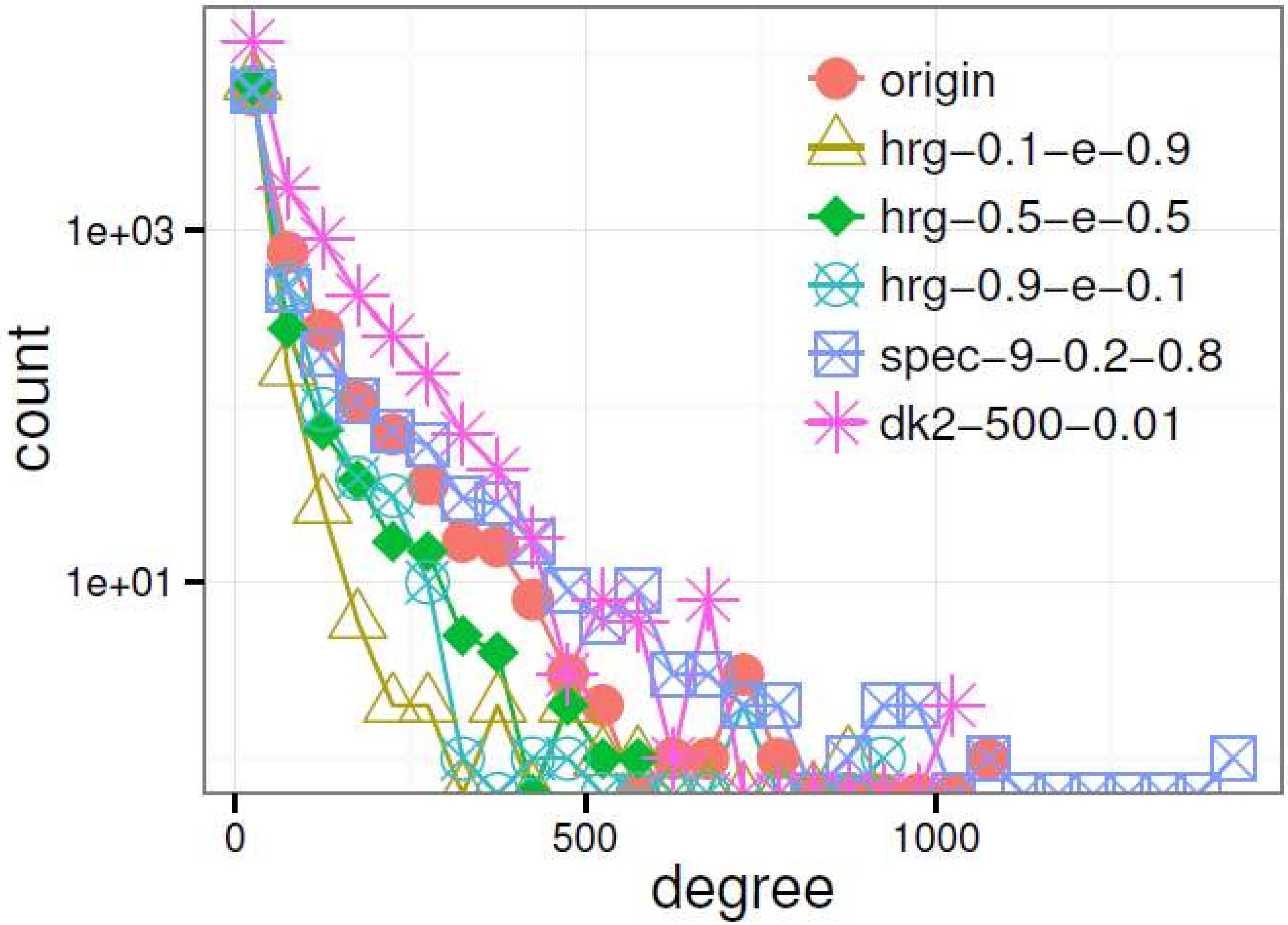
MCMC Convergence Study on $\log \mathcal{L}$



Wiki-Vote

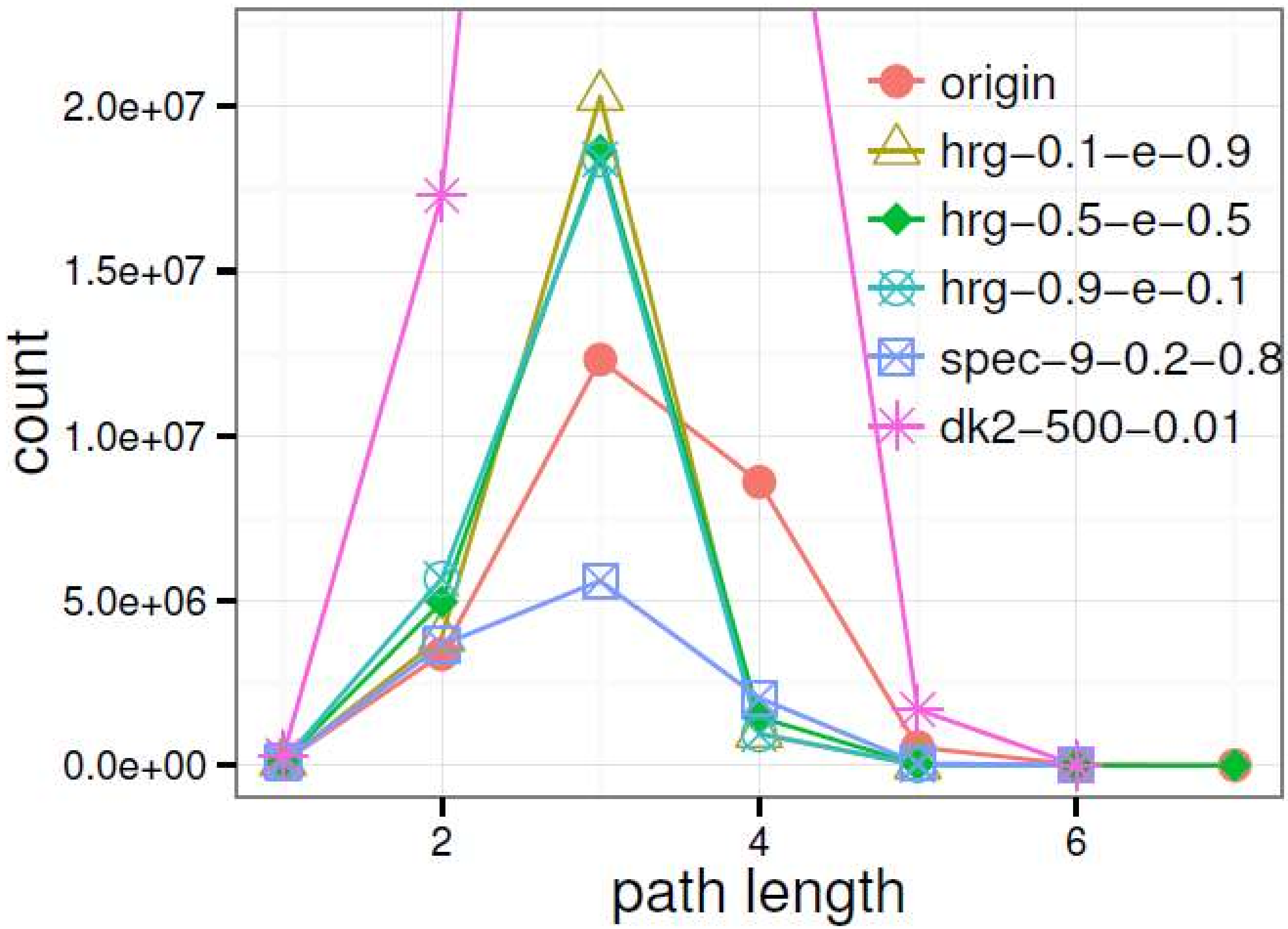
Trace of $\log \mathcal{L}$ as a function of the number of MCMC steps, normalized by n

Degree distribution



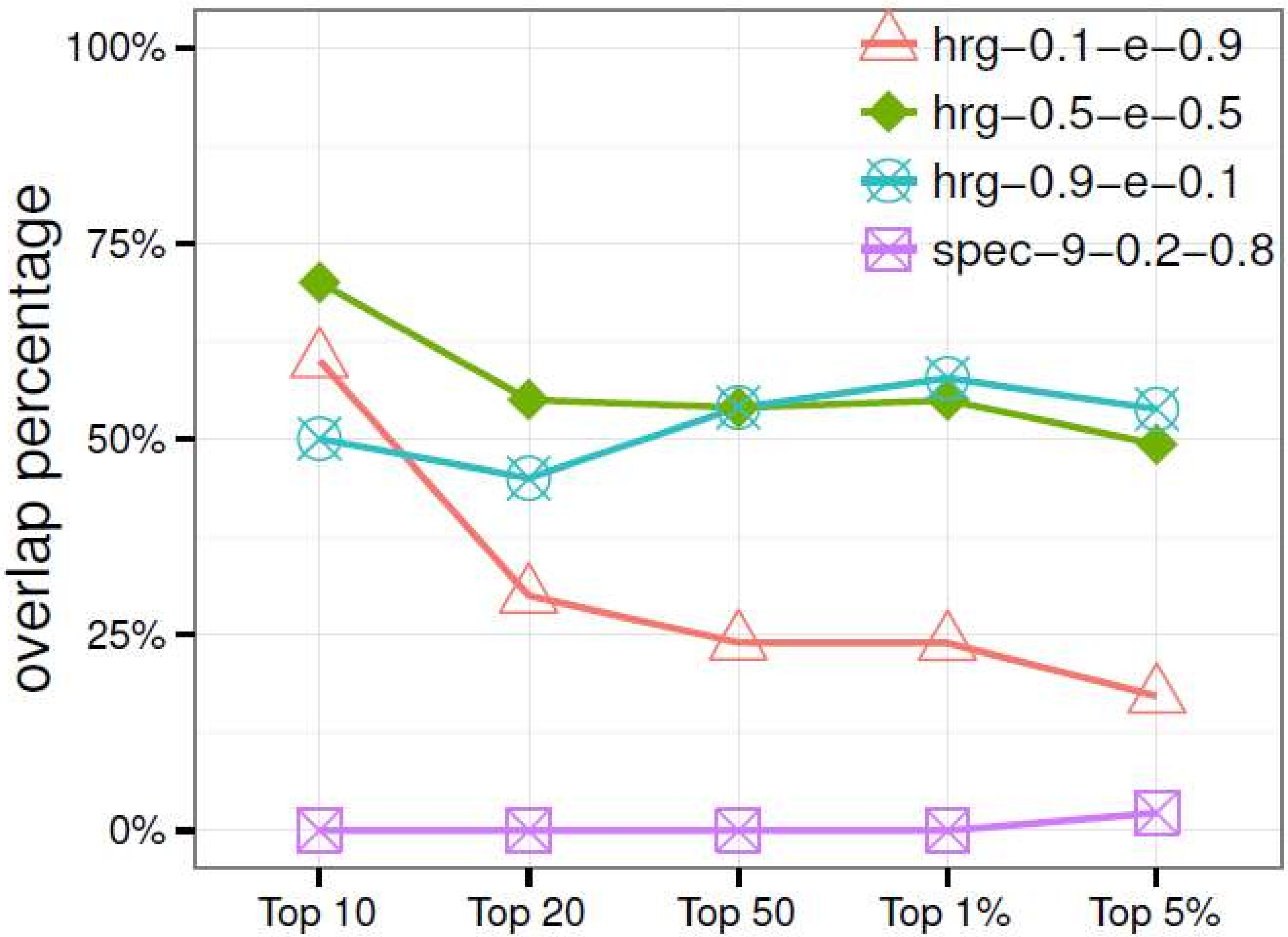
Wiki-Vote

Shortest path length distribution



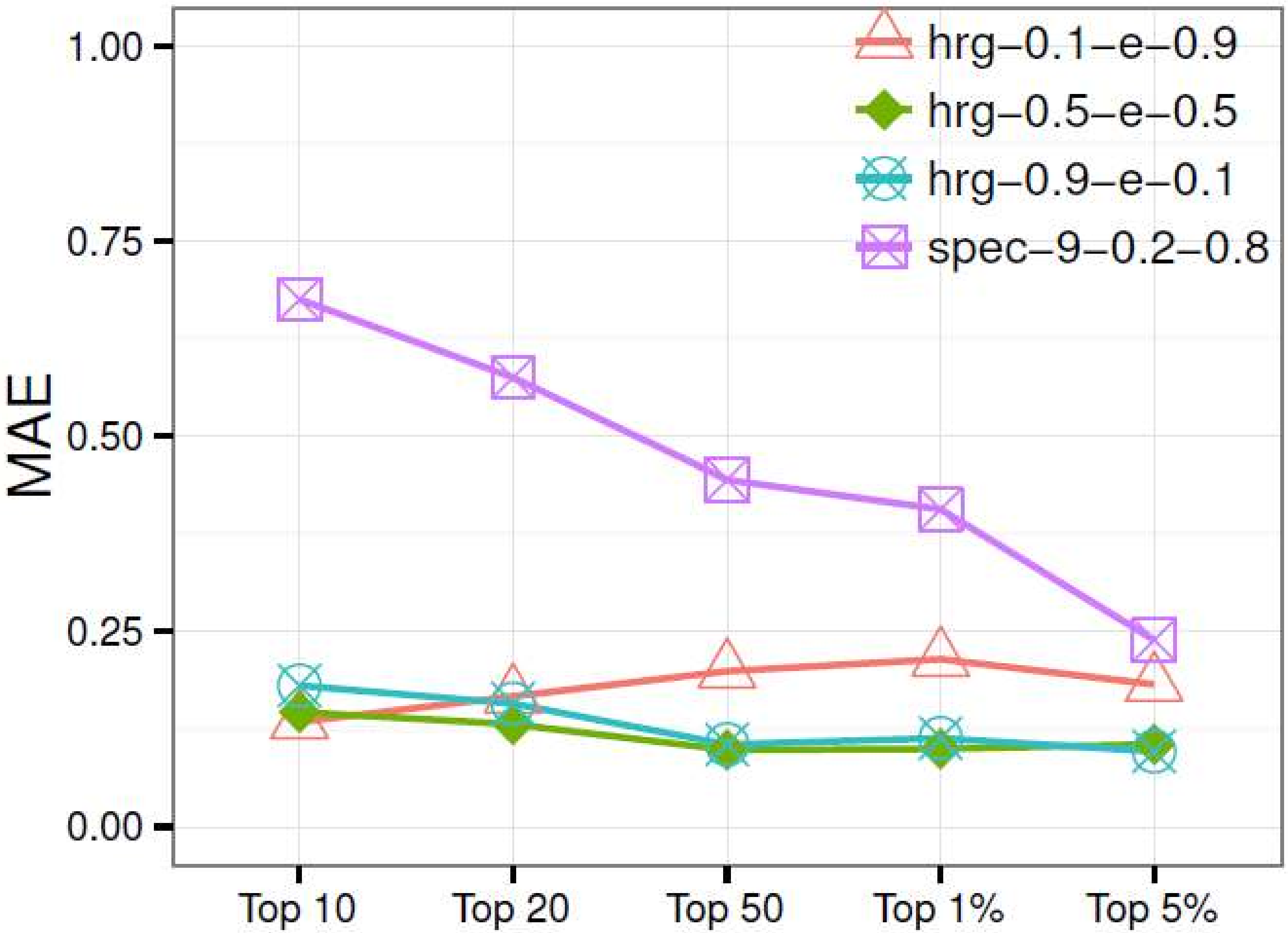
Wiki-Vote

Overlap of top-k vertices



Wiki-Vote

Mean absolute error of top-k vertices



Wiki-Vote

Outline

- Motivation
- Hierarchical Random Graph(HRG)
- Structural inference under DP with MCMC
- Sensitivity Analysis
- Experimental evaluation
- Conclusion

Conclusion

- We propose to infer connection probabilities with HRG for data sanitization under DP
- Our approach's sensitivity is $O(\log n)$
- Direct applying exponential mechanism on the huge space of HRG is prohibitively expensive. We overcome this challenge via doing sampling HRG space via MCMC
- Empirical experiments show our approach can effectively preserve many statistical properties in the network data

References

- C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In TCC, 2006.
- A. Sala, X. Zhao, C. Wilson, H. Zheng, and B. Y. Zhao. Sharing graphs using differentially private graph models. In IMC, 2011.
- Y. Wang, X. Wu, and L. Wu. Differential privacy preserving spectral graph analysis. In PAKDD, 2013.
- Y. Wang and X. Wu. Preserving differential privacy in degree-correlation based graph generation. TDP, 6(2), 2013.
- E. Shen and T. Yu. Mining frequent graph patterns with differential privacy. In SIGKDD, 2013.
- A. Clauset, C. Moore, and M. E. J. Newman. Structural inference of hierarchies in networks. In ICML on Statistical Network Analysis, 2007.
- A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. Nature, 453:98-101, 2008.

Thank you !

Q&A