PrivGraph: Differentially Private Graph Data Publication by Exploiting Community Information

Quan Yuan[1], Zhikun Zhang[2,3], Linkang Du[1],
Min Chen[3], Peng Cheng[1], and Mingyang Sun[1]

[1]Zhejiang University [2]Stanford University

[3]CISPA Helmholtz Center For Information Security

浙江大學
ZHEJIANG UNIVERSITY

Stanford
University

CISPA
HELMHOLTZ CENTER FOR
INFORMATION SECURITY

Outline

- Background
- Problem Definition
- Method
- Evaluation
- Conclusion

Outline

- Background
- Problem Definition
- Method
- Evaluation
- Conclusion

Big Data Era

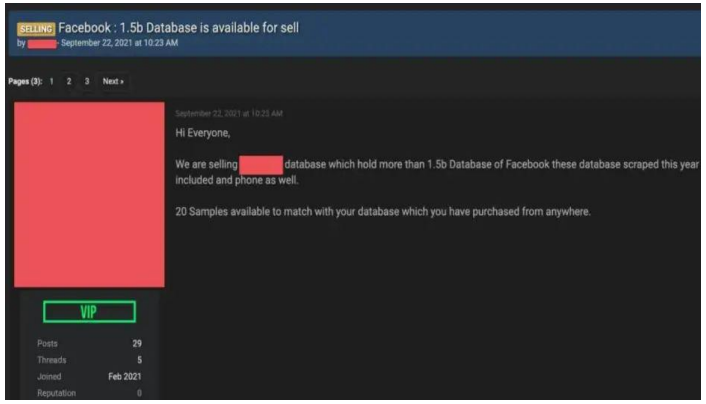
- Data collection
  - Browsing history, communication records, …
- Data analysis
  - Improving user experience, recommendation, …

User Data **Data Collection** **Data Analysis**

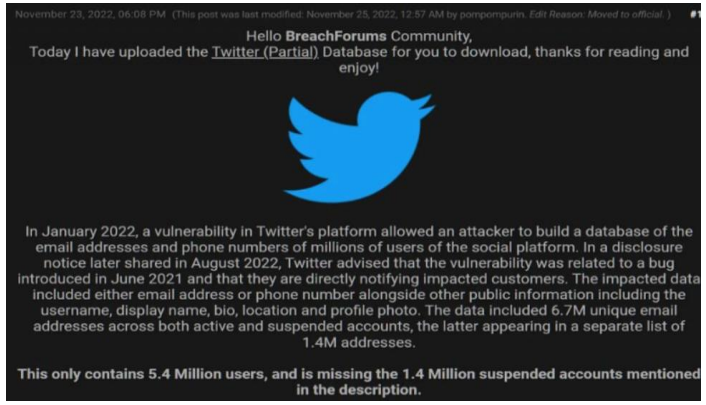
Privacy Accidents

2017, Yahoo
breached 3 billion user data

2020, Microsoft
exposed 250 million records

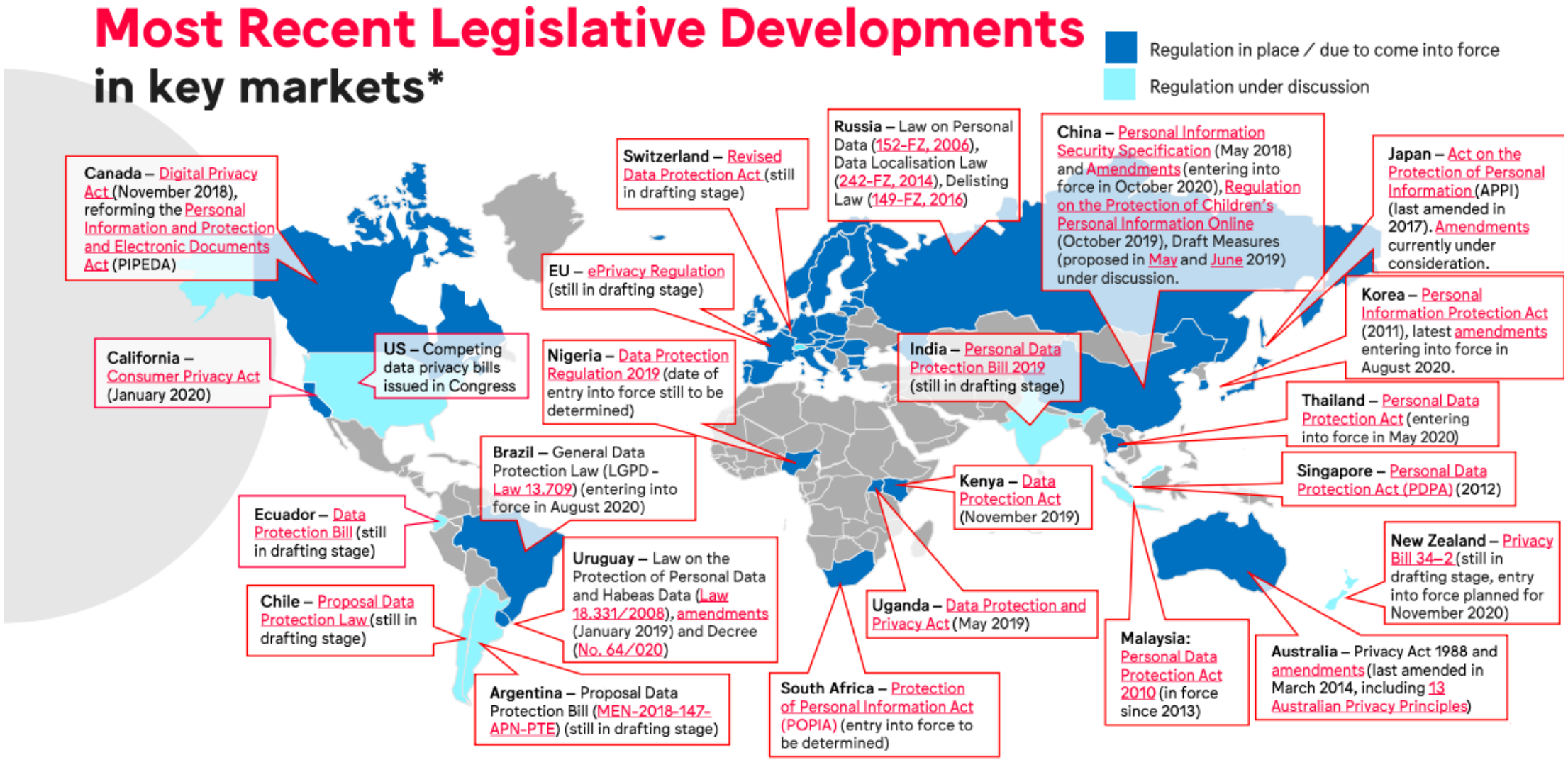


2021, Facebook
1.5 billion user data sold



2022, Twitter
breached 540 million user data

Laws for Privacy Protection



Outline

- Background
- Problem Definition
- Method
- Evaluation
- Conclusion

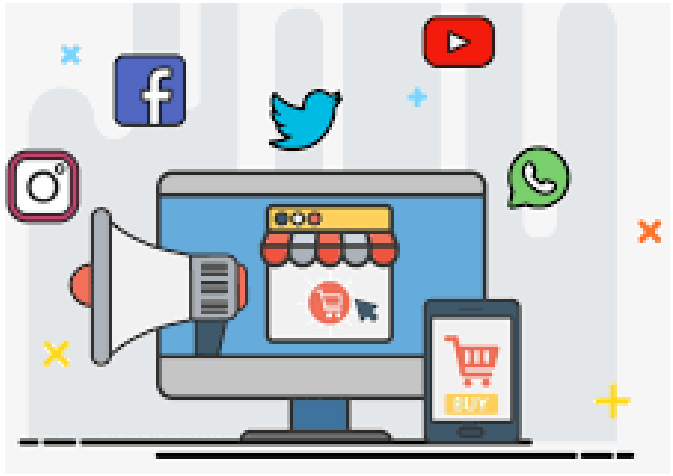
Problem Definition

Social Contacts **Vote Network** **Email Communication**

The edges of a graph may contain users' sensitive information.

Problem Definition

- **Related Tasks**

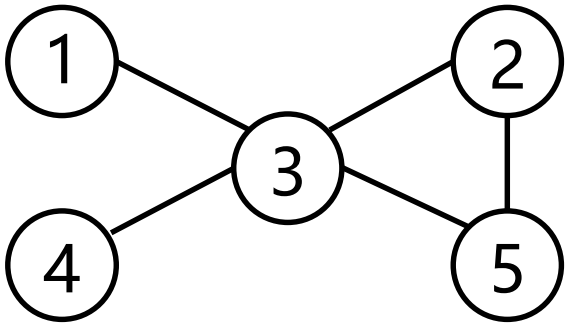


Advertising

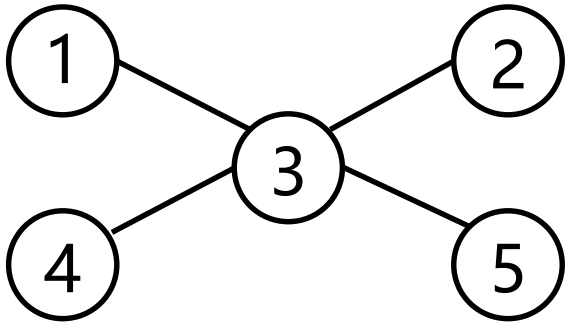
User Protrait

Epidemiological Study

Problem Definition



A graph G



An edge neighboring graph G'

Randomization Algorithm

Statistical result

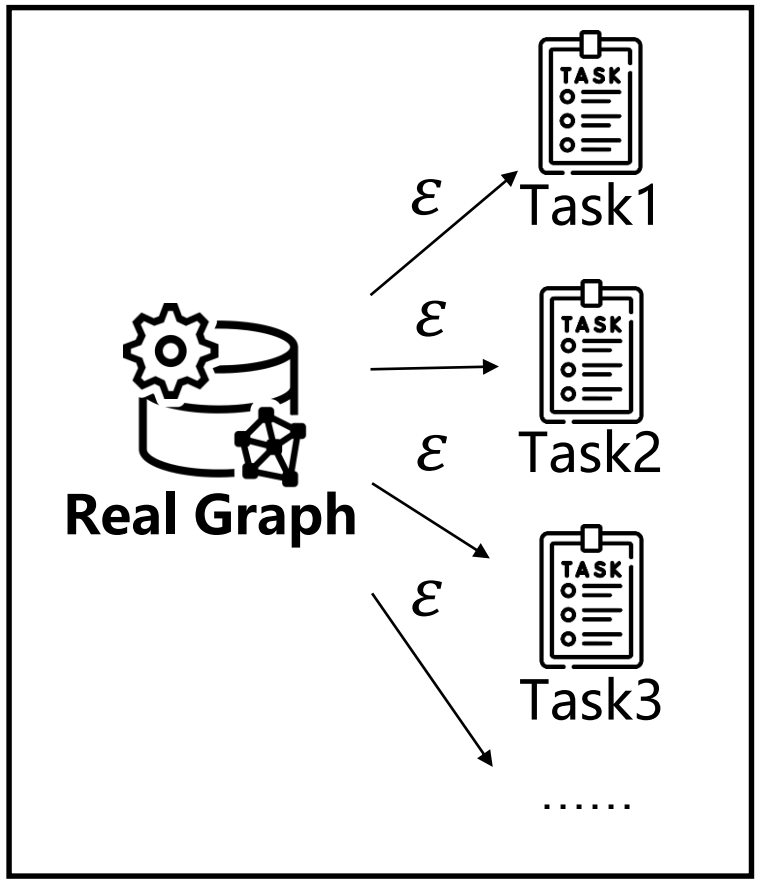
T satisfy

$$\frac{\Pr[A(G) \in T]}{\Pr[A(G') \in T]} \leq e^{\varepsilon}$$

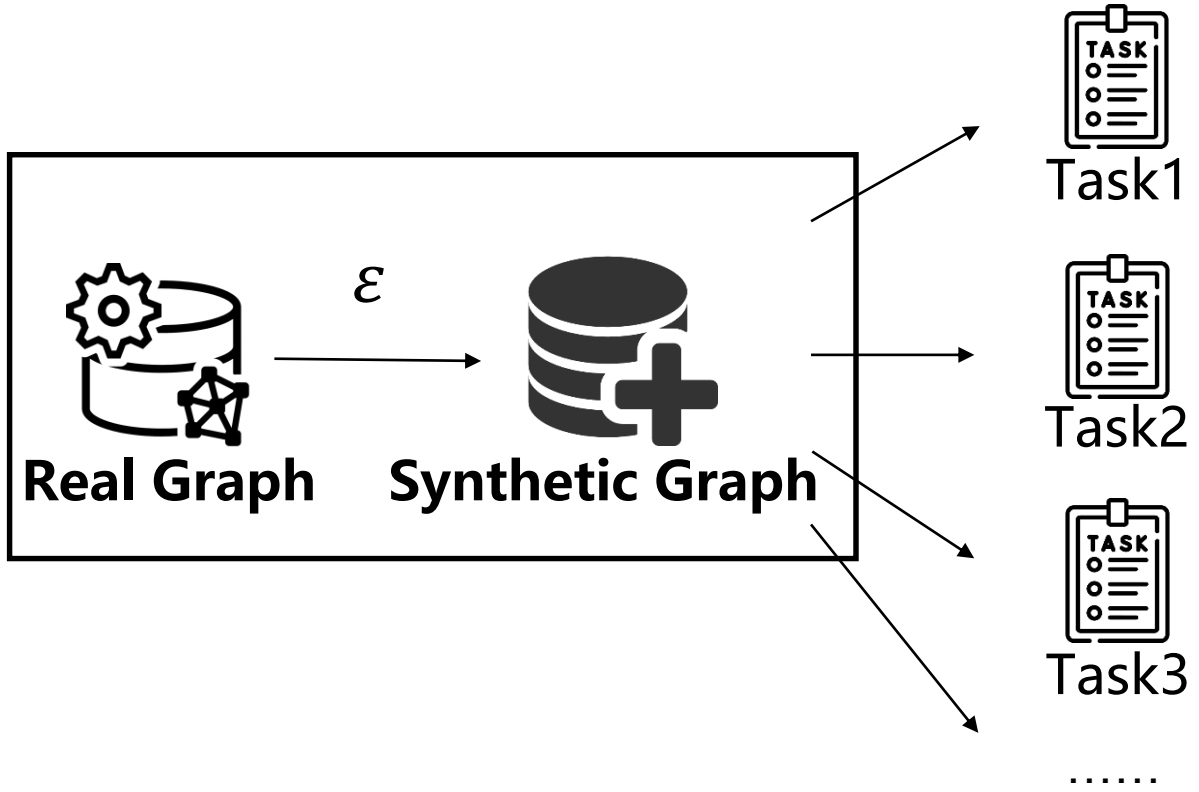
Edge-DP: Limit the impact of any edge in the graph on the output

Problem Definition

Private Tasks



Private Data Synthesis

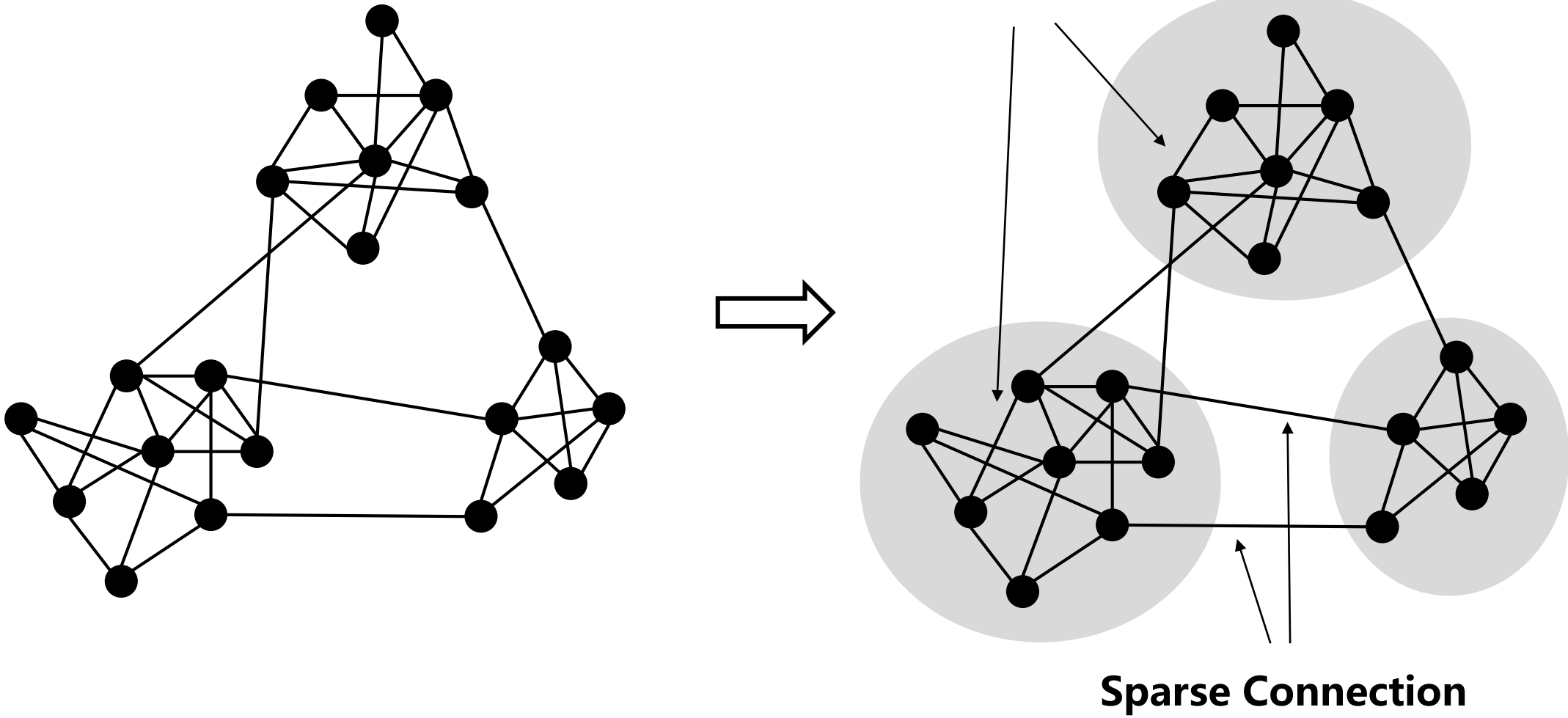


Our goal: Synthesize a graph under edge-DP while ensuring high utility

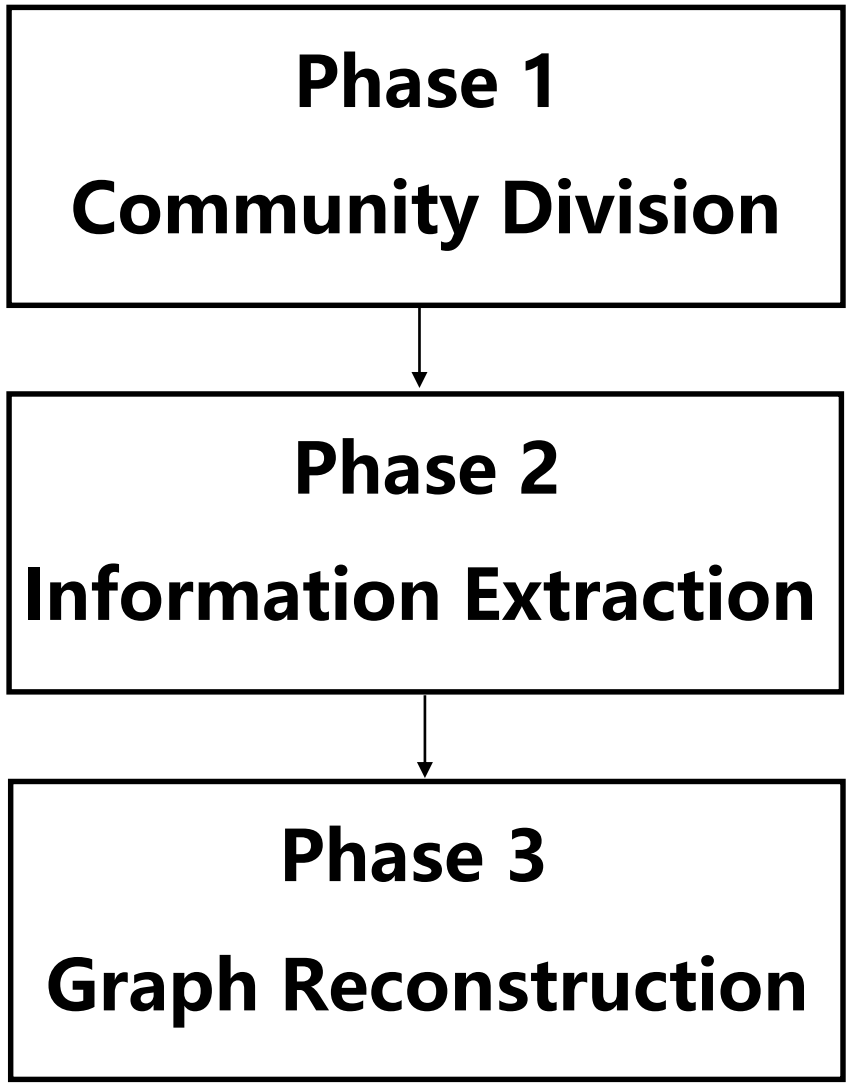
Outline

- Background
- Problem Definition
- Method
- Evaluation
- Conclusion

Intuition

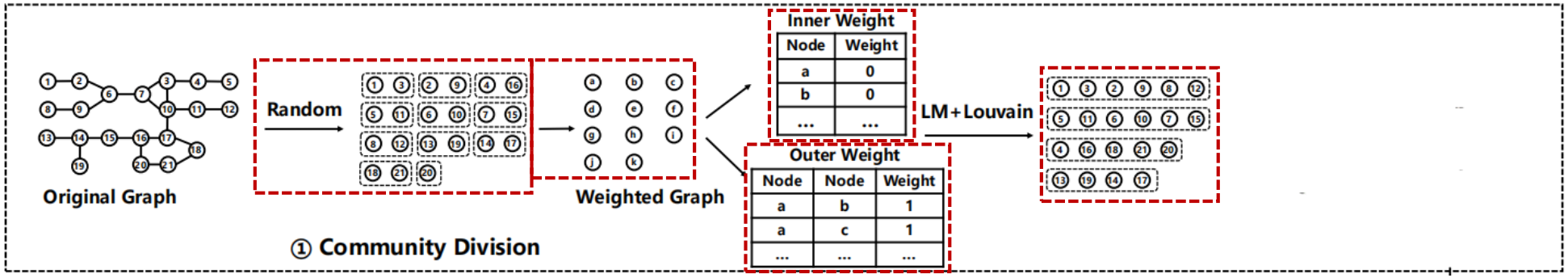


Workflow of PrivGraph



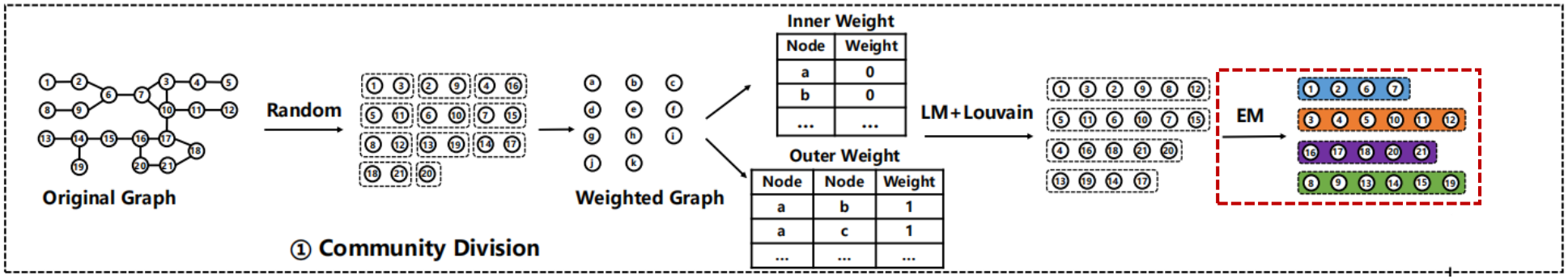
Workflow of PrivGraph

- Phase 1: Community Division (CD)
- Phase 2: Information Extraction (IE)
- Phase 3: Graph Reconstruction (GR)



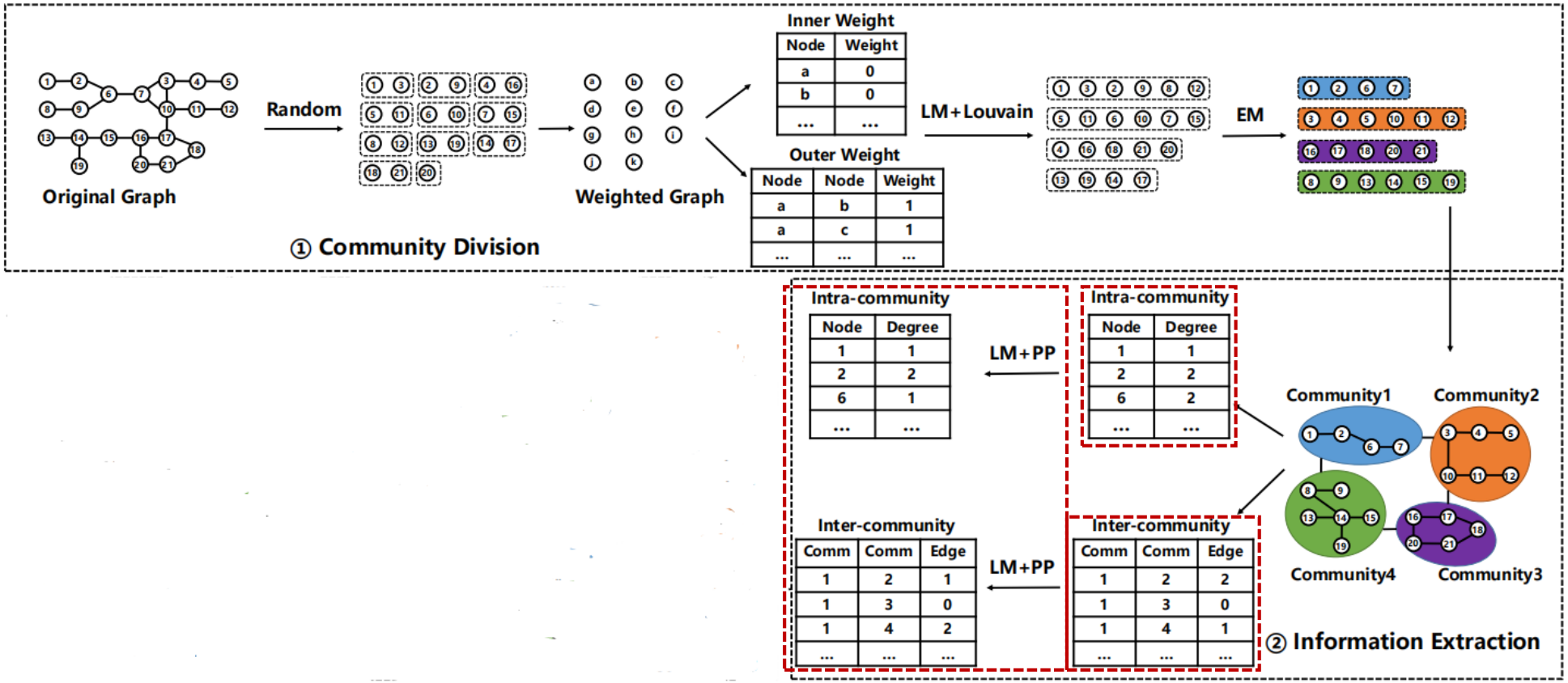
Workflow of PrivGraph

- Phase 1: Community Division (CD)
- Phase 2: Information Extraction (IE)
- Phase 3: Graph Reconstruction (GR)



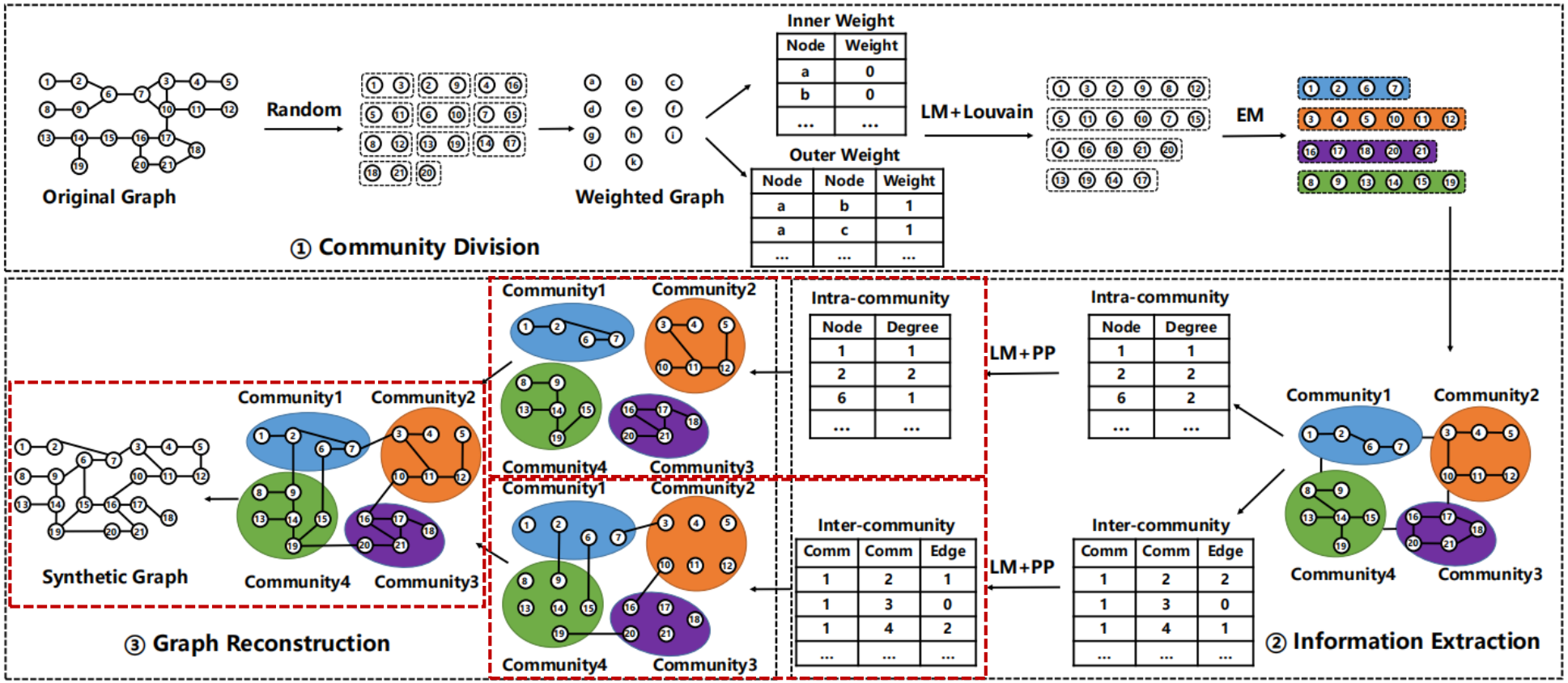
Workflow of PrivGraph

- Phase 1: Community Division (CD)
- Phase 2: Information Extraction (IE)
- Phase 3: Graph Reconstruction (GR)



Workflow of PrivGraph

- Phase 1: Community Division (CD)
- Phase 2: Information Extraction (IE)
- Phase 3: Graph Reconstruction (GR)



Outline

- Background
- Problem Definition
- Method
- Evaluation
- Conclusion

Experiment Setup

Dataset

- 6 real world datasets

Dataset	Nodes	Edges	Density	Type
Chamelon [65]	2,277	31,421	0.01213	Web page
Facebook [46]	4,039	88,234	0.01082	Social
CA-HepPh [44]	12,008	118,521	0.00164	Collaboration
Enron [63]	33,696	180,811	0.00032	Email
Epinions [62]	75,879	405,740	0.00014	Trust
Gowalla [13]	196,591	950,327	0.00005	Social

Experiment Setup

Metrics

- Community Discovery: Normalized Mutual Information
- Node Information: Eigenvector Centrality Score
- Degree Distribution
- Path Condition: Diameter
- Topology Structure: Clustering Coefficient, Modularity

Competitors

- LDPGen[1]
- TmF[2]
- PrivHRG[3]
- DER[4]

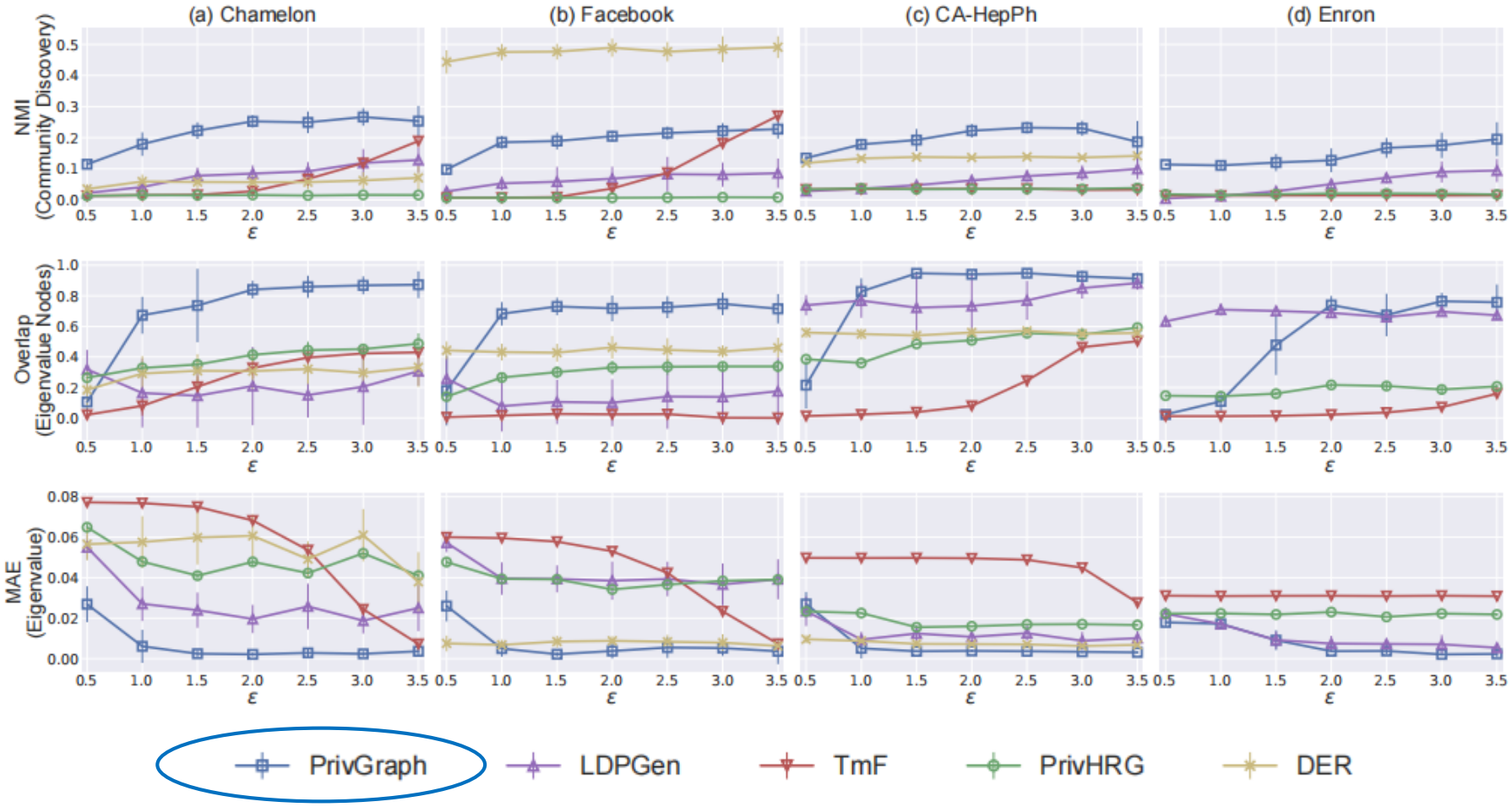
[1] 2017 CCS Generating Synthetic Decentralized Social Graphs with Local Differential Privacy

[2] 2015 ASONAM Differentially Private Publication of Social Graphs at Linear Cost

[3] 2015 SIGKDD Differentially Private Network Data Release via Structural Inference

[4] 2014 VLDBJ Correlated Network Data Publication via Differential Privacy

Performance

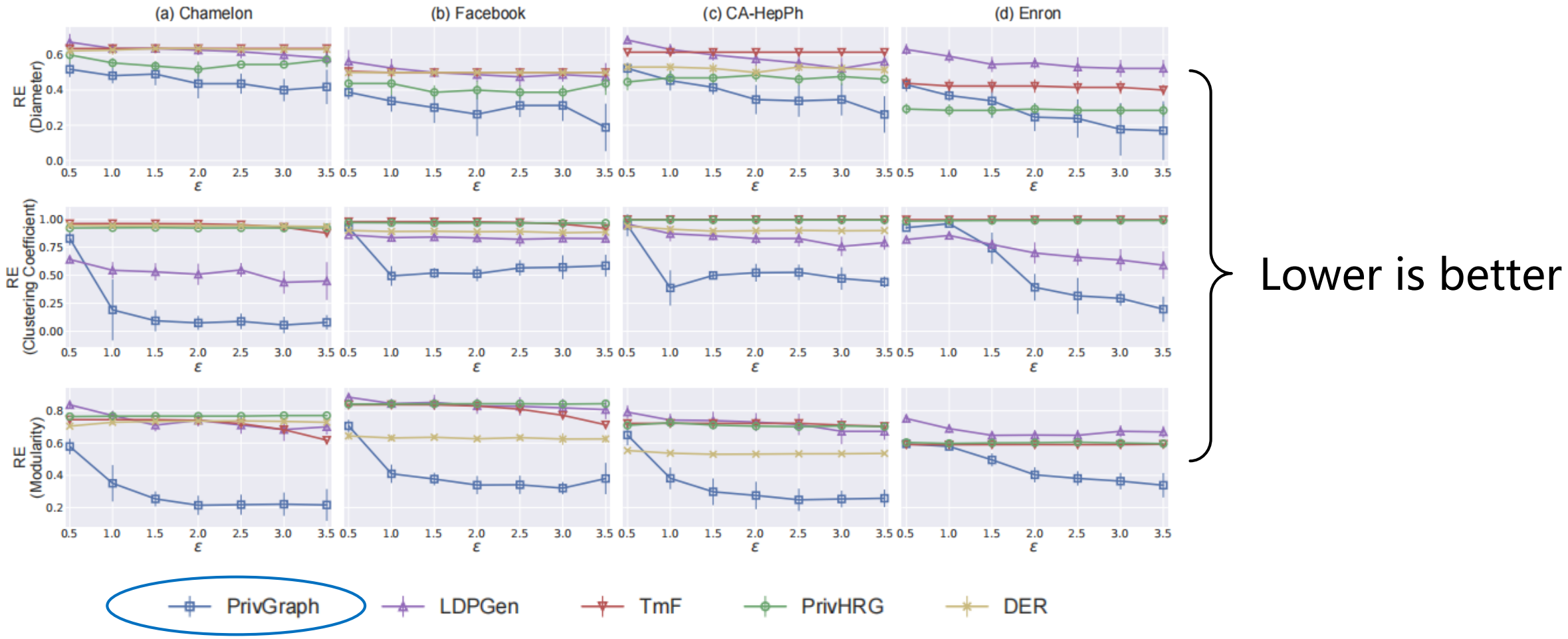


Higher is better

→ Lower is better

PrivGraph outperforms other methods in most cases.

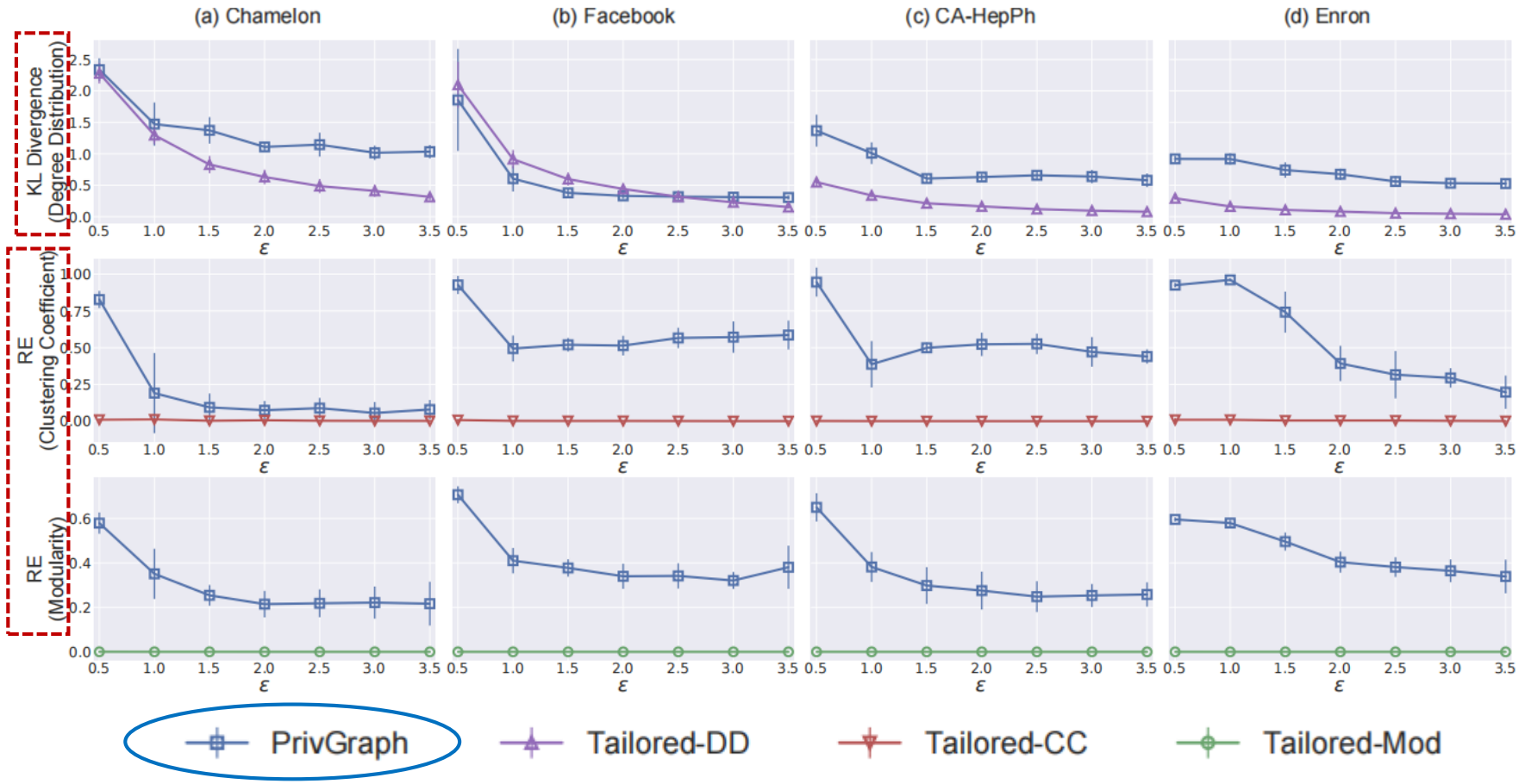
Performance



Lower is better

PrivGraph outperforms other methods in most cases.

Comparison with Tailored Methods

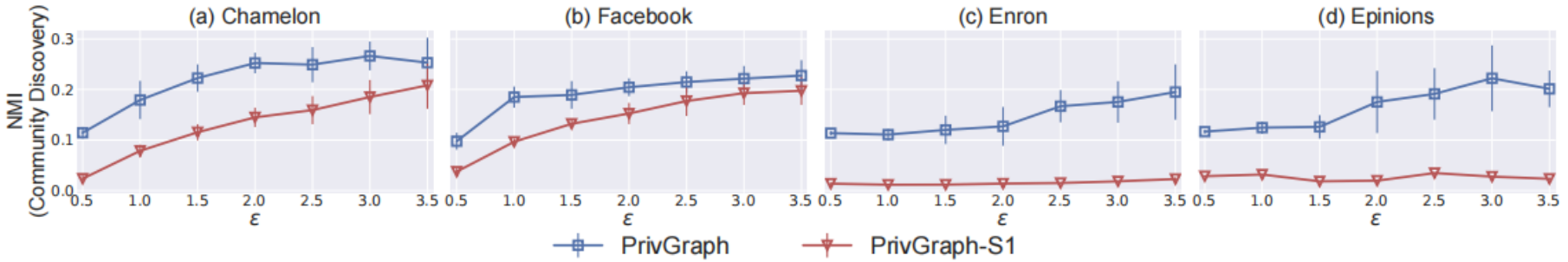


Lower is better

PrivGraph achieves competitive performance on the degree distribution.

Preservation for Small Communities

- Louvain adopted in phase 1 might miss the small communities[1] during the modularity optimization process.



→ Higher is better

PrivGraph can compensate the shortcoming of Louvain since the information extraction and graph reconstruction processes help to recover the small communities.

[1] S. Fortunato and M. Barthelemy. Resolution limit in community detection. PNAS, 104(1):36–41, 2007.

Outline

- Background
- Problem Definition
- Method
- Evaluation
- Conclusion

Conclusion

- A deep analysis of existing solutions on differentially private graph synthesis
- A practical method PrivGraph to generate a synthetic graph under DP
- An extensive evaluation on multiple datasets and metrics to illustrate the superiority of PrivGraph

Thank you for your attention

Q & A

Email: yq21@zju.edu.cn