Publishing Attributed Social Graphs with Formal Privacy Guarantees

Zach Jorgensen (NSCU), Graham Cormode (Warwick), Ting Yu (QCRI)

Motivation

Social network analysis (SNA) studies graphs of entities and their relations.

Applications range from marketing to predicting disease spread.

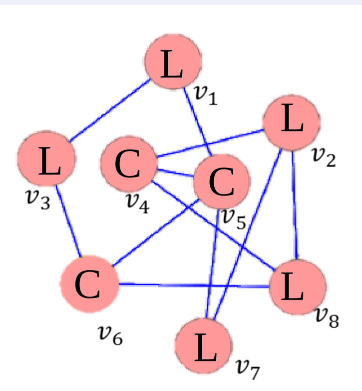
SNA data is sensitive: can we produce realistic data while preserving privacy?

Prior work has only studied graph structure, but real graphs also have (sensitive) attributes on nodes.

We describe a differentially private (DP) framework for realistic synthetic social graphs with attributes, and evaluate on real-world datasets.

Models and Overview

Graph Model: Graph G has n nodes N, m (undirected) edges E and attribute list X for nodes.



Example:

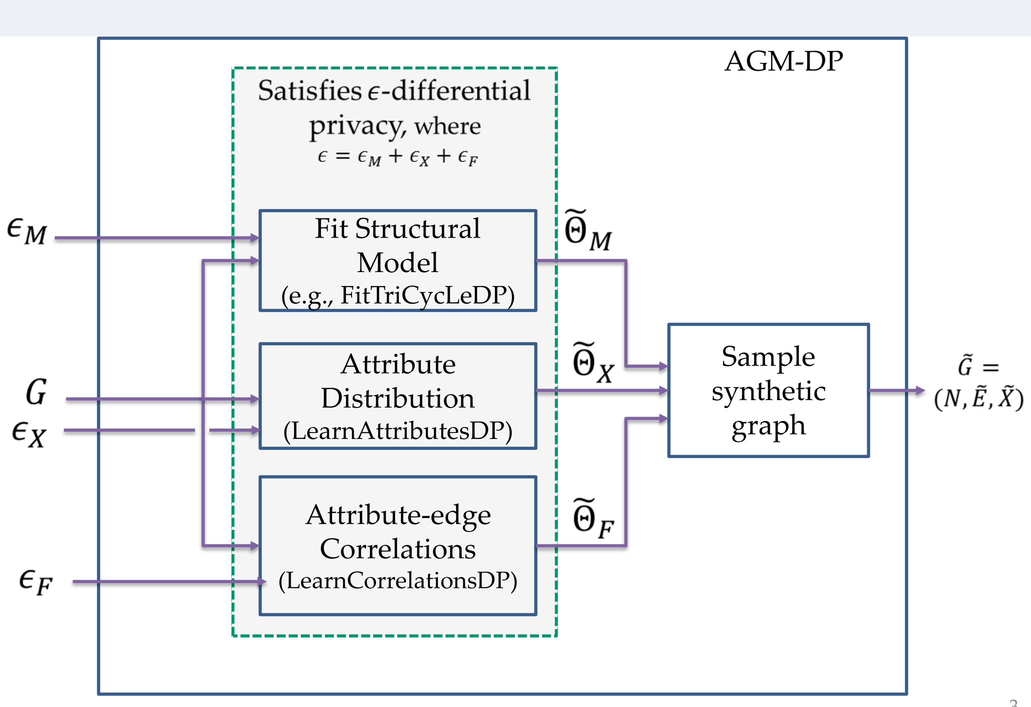
w = 1 attribute, *political views*
L = Liberal (0) C = Conservative (1)
$N = \{v_1, \ldots, v_9\}$
$E = \{e_{13}, e_{15}, e_{24}, e_{27}, e_{29}, \ldots\}$
$X = \{\langle 0\rangle, \langle 0\rangle, \langle 0\rangle, \langle 1\rangle, \ldots, \langle 0\rangle\}$

Privacy Model: Algorithm A satisfies ϵ-attributed graph differential privacy (adapting [2, 1]) if for any pair of neighboring graphs G, G' and properties O,

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

G and G' are *neighboring* if they differ by a single edge or in the attributes of a single node.

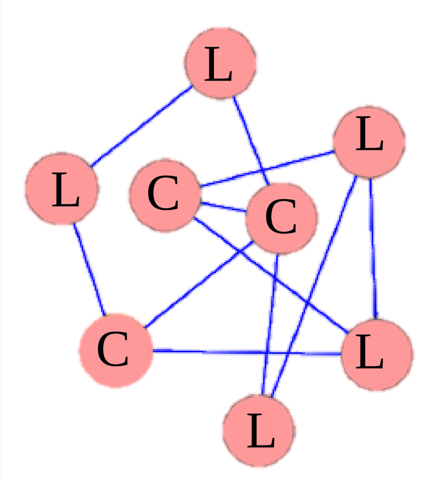
Overview: We (privately) model structure Θ_M, attributes Θ_X, and attribute-edge distributions Θ_F, to sample synthetic graphs.



Graph Distributions

Node-attribute distribution Θ_X: estimate prior distribution of attribute values.

Compute 2^w counts, add Laplace noise (histogram query).



Example:

$Q = \langle q_L, q_C \rangle = \langle 5,3 \rangle$

$\tilde{Q} = Q + \mathrm{Lap}^{2^w}(0, \frac{2}{\epsilon})$

$\tilde{\Theta}_X = \dfrac{\tilde{Q}}{\mathrm{sum}(\tilde{Q})}$

Attribute-Edge distribution Θ_F: estimate the probability of an edge given the two node values.

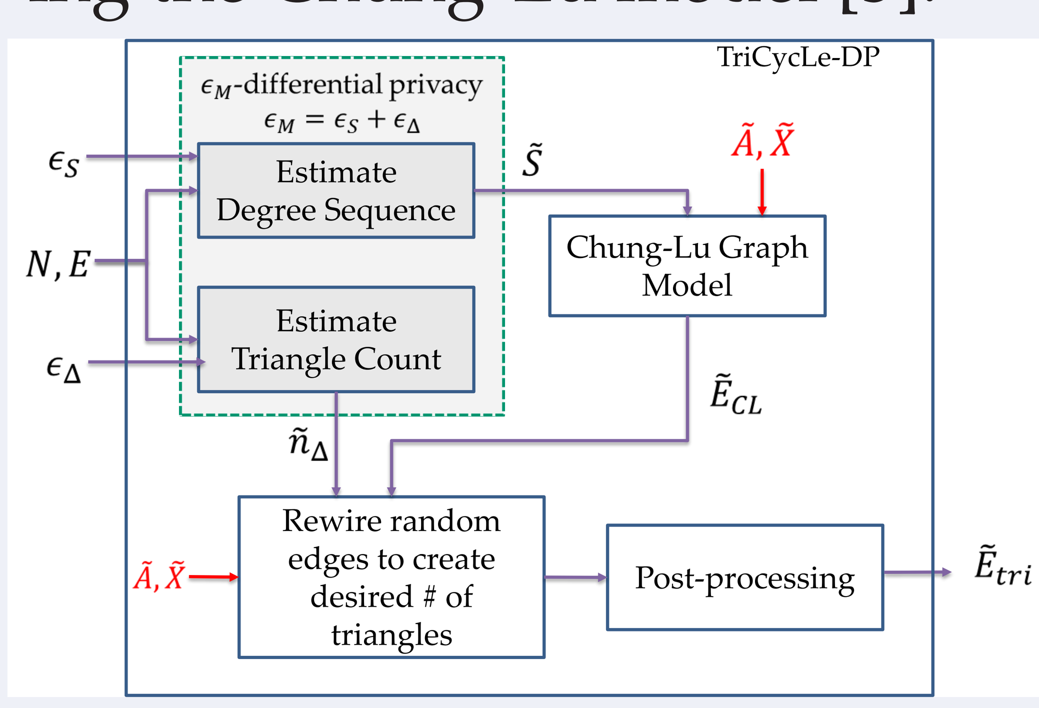
Query has high "sensitivity" if node degrees can be large.

Use edge truncation: bound the degree of nodes in the input graph to k. Heuristic: $k = \sqrt[3]{n}$.

Structural Model

Many social graph models have been proposed – but these don't combine well with privacy.

We propose TriCycle, a new privacy-friendly model, extending the Chung-Lu model [3].



The parameters Θ_M are the degree sequence and number of triangles. These can be found accurately under DP.

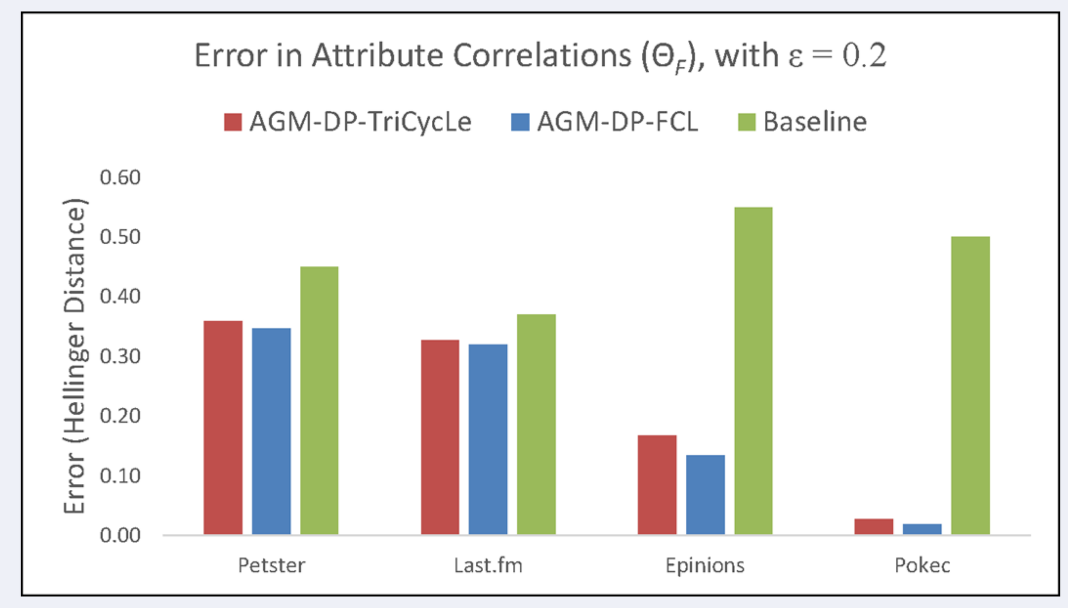
Datasets

We use four real-world social network datasets. We list max and average degree $d_{\max}$, d_{avg}, and number of triangles n_Δ.

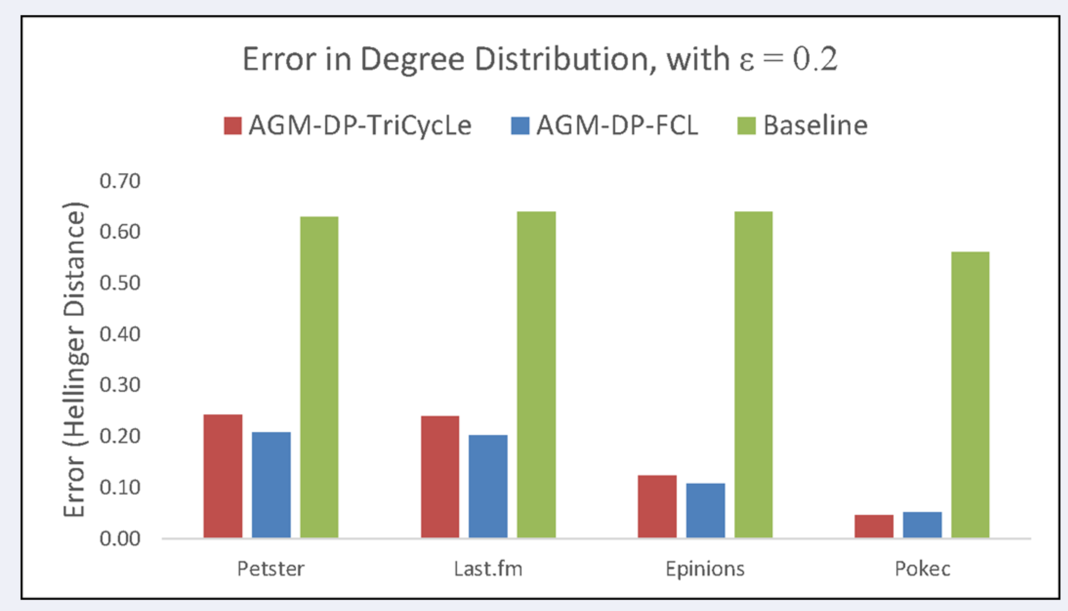
Dataset	n	m	$d_{\max}$	d_{avg}	n_Δ
Last.fm	1,843	12,668	119	6.9	19,651
Petster	1,788	12,476	272	7.0	16,741
Epinions	26,427	104,075	625	3.9	231,645
Pokec	592,627	3,725,424	1,274	6.3	2,492,216

Experiments

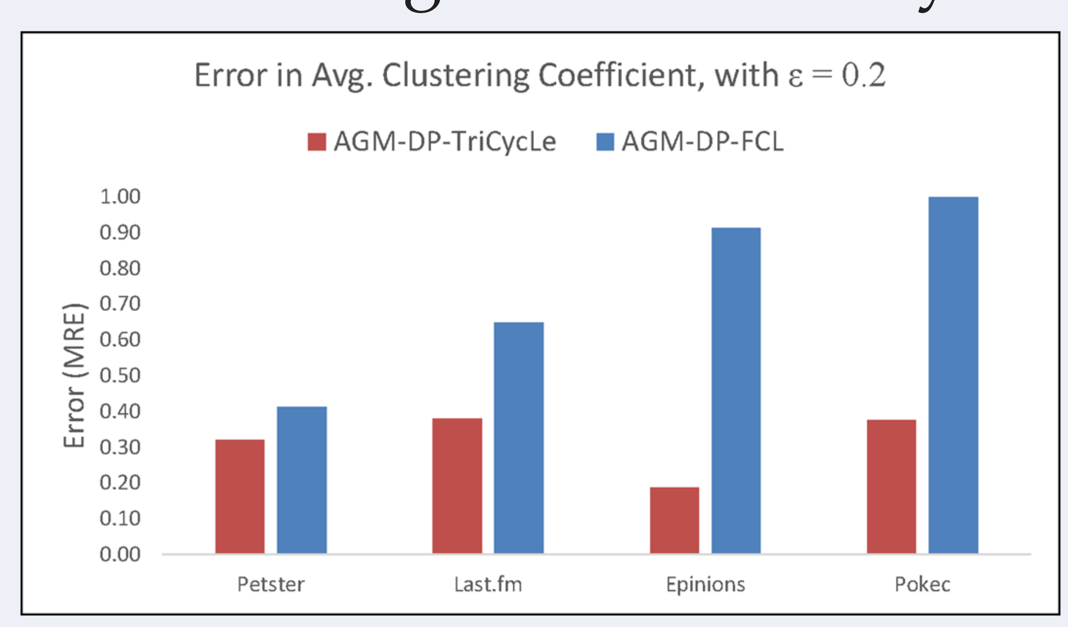
We generated multiple synthetic graphs from each model and measured accuracy of various properties compared to Fast Chung-Lu (FCL).



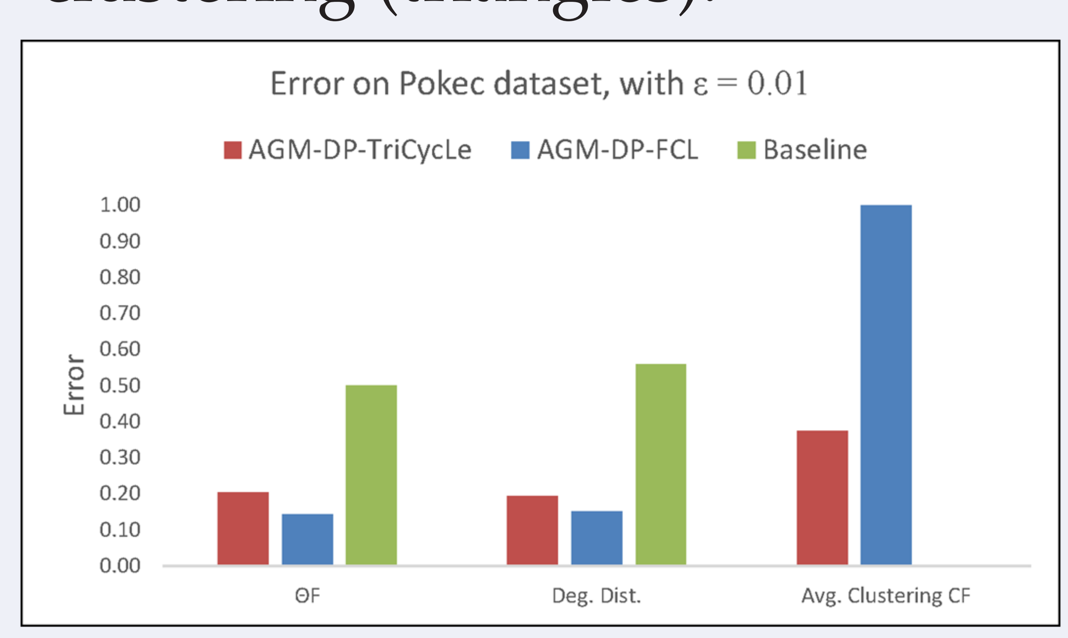
Baseline: uniform probabilities.



Baseline: edges set uniformly.



FCL does not explicitly model clustering (triangles).



Pokec is a very large dataset so allows a very strong privacy guarantee, $\epsilon = 0.01$.

References

[1] J. Blocki, A. Blum, A. Datta, and O. Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In *ITCS*, 2013.

[2] C. Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 2011.

[3] J. J. Pfeiffer, T. La Fond, S. Moreno, and J. Neville. Fast generation of large scale social networks while incorporating transitive closures. In *PASSAT*, 2012.