Personalized privacy protection in social networks through adversarial modeling

THE UNIVERSITY OF ILLINOIS AT CHICAGO UIC

Sachin Biradar, Elena Zheleva

Department of Computer Science, University of Illinois at Chicago

Contact: sachinbiradar9@gmail.com, ezheleva@uic.edu

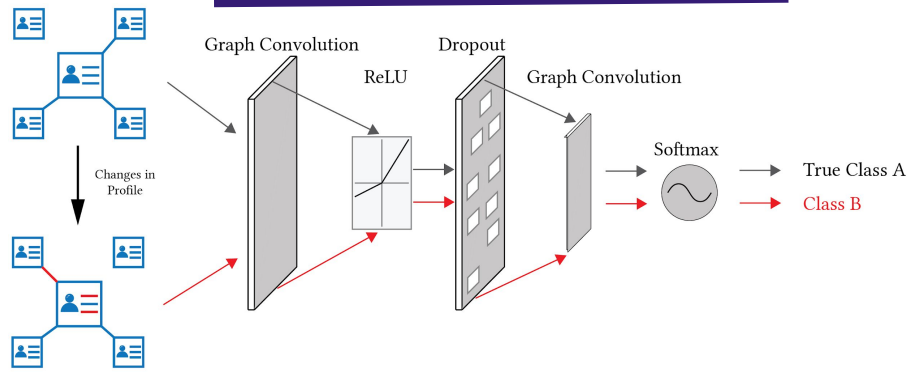
Problem description

- **Motivation**: Machine learning classifiers can predict sensitive attributes of social media users without their permission
- **Goal:** Create an algorithm that can help a user defend themselves from privacy-invasive classifiers by suggesting changes to their social media profile that would make a target classifier misclassify the user's sensitive attribute value while satisfying utility constraints about how much the user is willing to change

Contributions

- Frame a new privacy problem of user-centric adversarial perturbations with utility constraints
- Define a novel "grey-box" scenario in which the target classifier type is known but the classifier parameters are not and can be estimated using publicly available data.

Outwit Algorithm

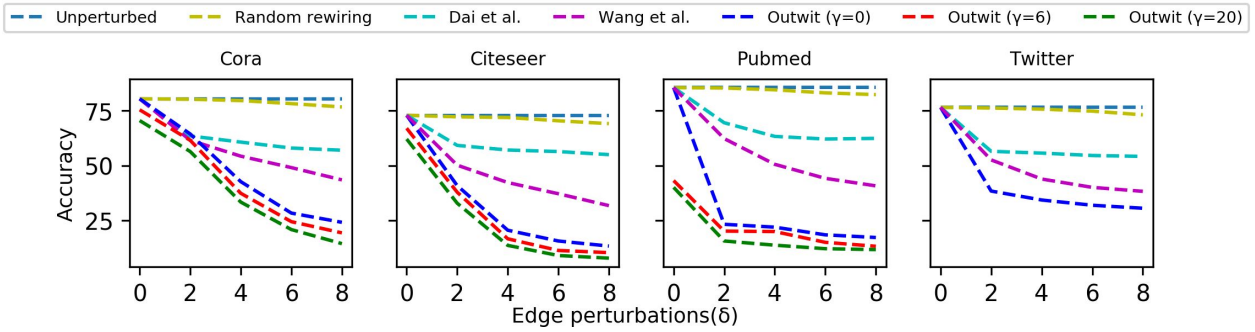


Target Classifier: Graph Convolutional Network (Kipf and Welling 2017)
Outwit: gradient-based algorithm which finds the minimum number of node attribute and edge changes necessary to get the sensitive attribute misclassified

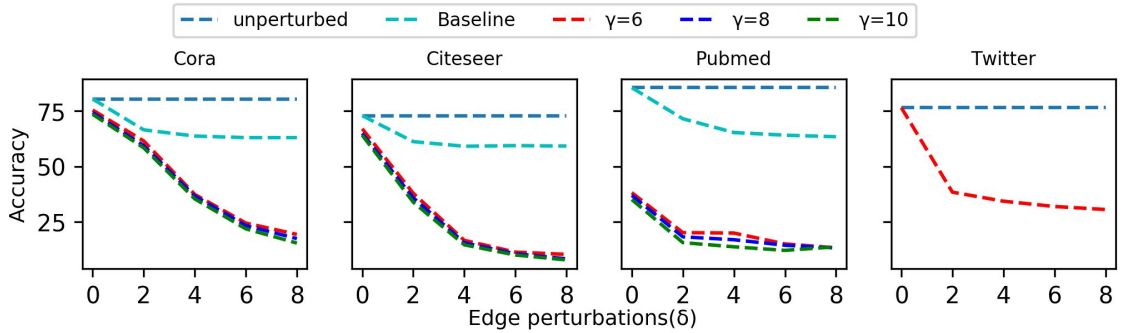
Results

Outwit achieves a much better user privacy protection with significant decrease in target classifier accuracy (8-47%) compared to state-of-the-art adversarial algorithms for graphs

- δ: maximum number of edge perturbations allowed
- γ: maximum number of attribute perturbations allowed



Baselines: Dai et al. 2018a and Wang and Gong 2019 white-box models



Acknowledgments: Supported by NSF RAPID Grant #1801644.