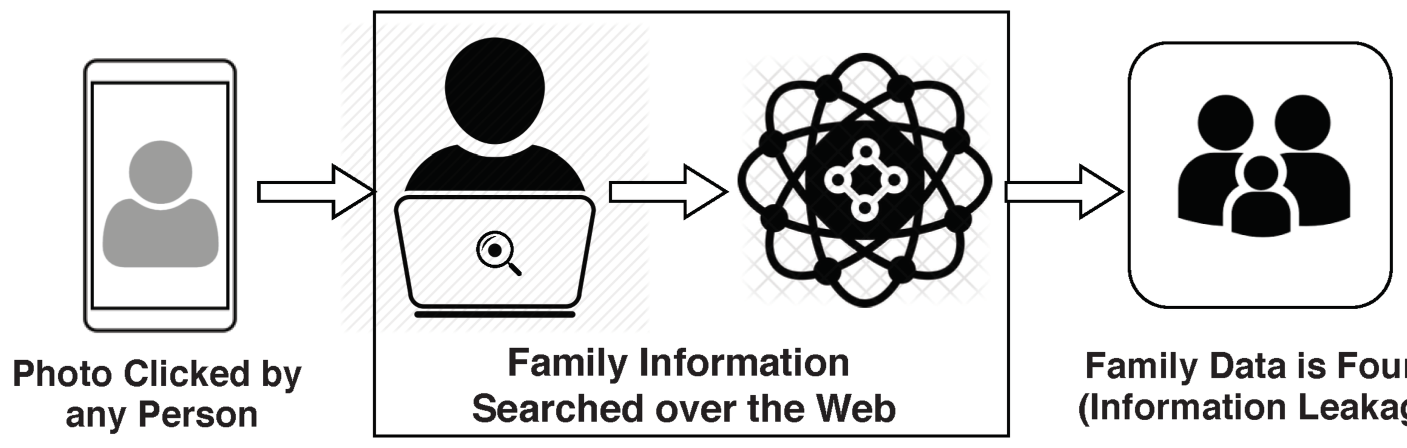
Adversary for Social Good: Protecting Familial Privacy through Joint Adversarial Attacks

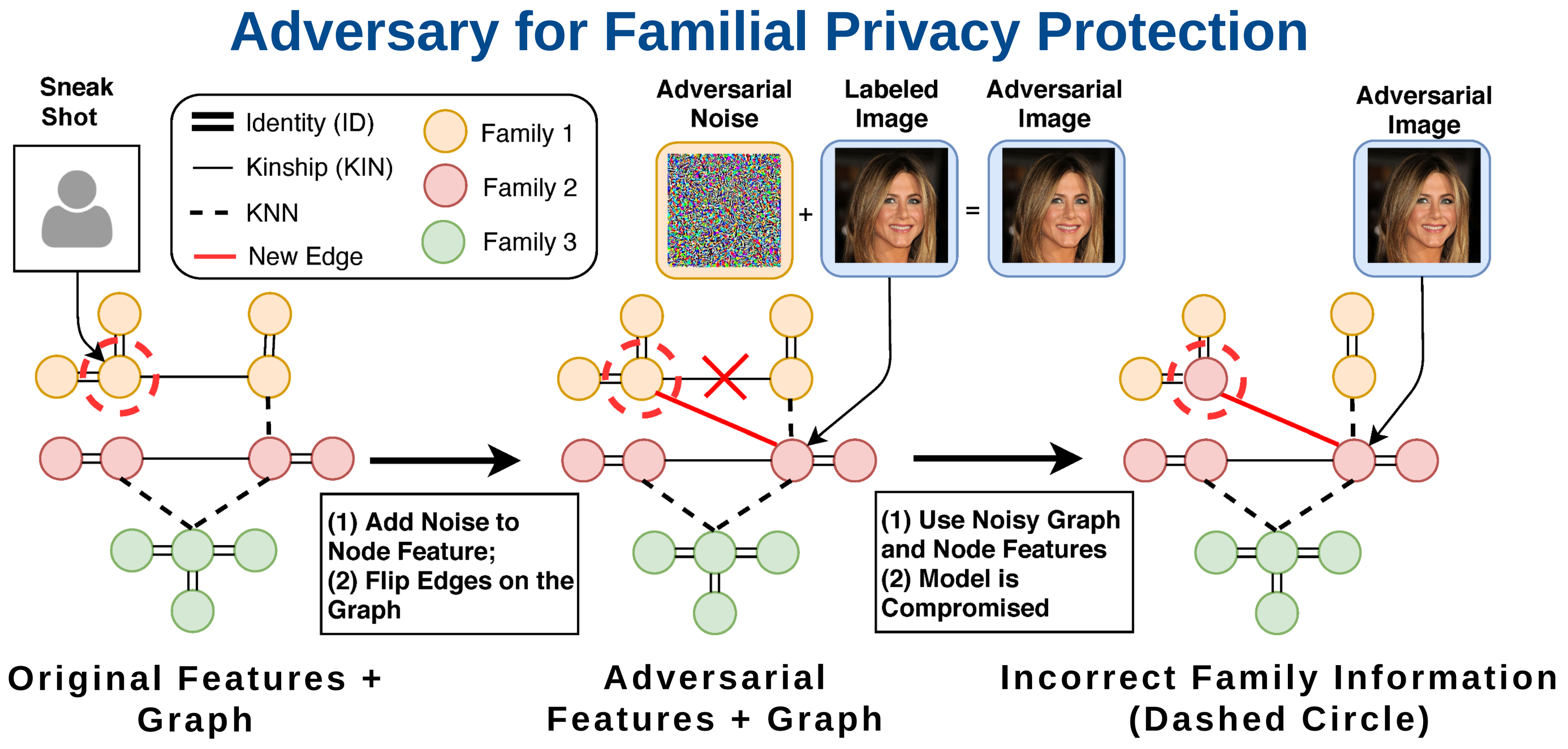
Chetan Kumar, Riazat Ryan, Ming Shao
Department of Computer and Information Science
University of Massachusetts, Dartmouth

Introduction



- **Social Media**
  - Social Media is mainly featured by sharing photos and social connections (friends, relatives, etc.)
  - Learning models with social media data can be developed towards various goals
  - Unfortunately, it may lead to information leakage and expose privacy w/ or w/o intention
  - You can imagine how furious the celebrity will be when their family members photos are exposed without their permission
- **Data Leakage**
  - Limited time to read Terms & Conditions
  - Limited knowledge (especially children) to understand
  - Unintentional leakage
  - Generally, people have no willingness to disclose personal data but it has already been out of our control, as long as people remain connected by the society and the Internet

Adversary for Familial Privacy Protection



Social Family Recognition (SFR)

- Family recognition can be addressed under the network environment by casting it to a semi-supervised learning problem on the social networks
- Conventional visual family recognition (VFR) is to train a multi-class classifier first, and then assign family labels to each probe image in the running time
- Even with the most recent deep features designed for visual kinship, e.g., SphereNet (Liu et al. 2017), the accuracy is far from acceptable

Family Recognition on the Graph

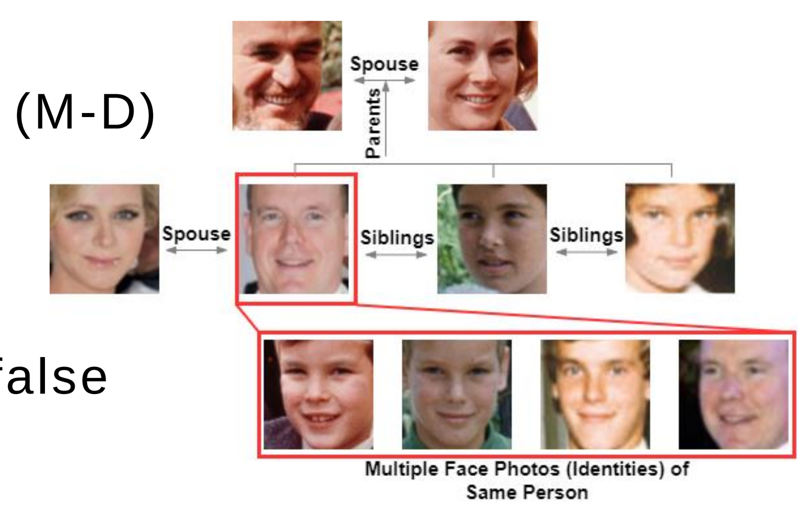
- In our graph, each node represents visual features generated by the state-of-the-art kinship descriptors
- Edges encode the relation between two nodes
- Three types of relations are considered i.e.,
  - Identity (ID): Link nodes of the same person
  - Kinship (KIN): Link nodes of the same family label
  - *k*-NN: Link nodes between different families, to avoid isolated nodes

Proposed Framework

- **Privacy at Risk**
  - Social media data could be handy to develop a model
  - This model could be used against finding private information
- **Adversarial Attack:**
  - Added Noise to Node Features by calculating sign of the Gradient
  - Added/Removed edges (relationships) between nodes
- **Model Compromised:**
  - By using Noisy Features and Noisy Graph

Dataset

- **Families in the Wild (FIW)**
  - 11 types of relationships
  - Same generation (S-S) to first (M-D) to third (GM-GD)
  - Consists of 1000 families with average 12 images/family
  - Pairs are labeled with true or false kin relationship



- **Created two social networks**

Family-100
- Contains 502 subjects
- 2758 facial images
- 502 nodes for training
- 2256 nodes for validation and testing

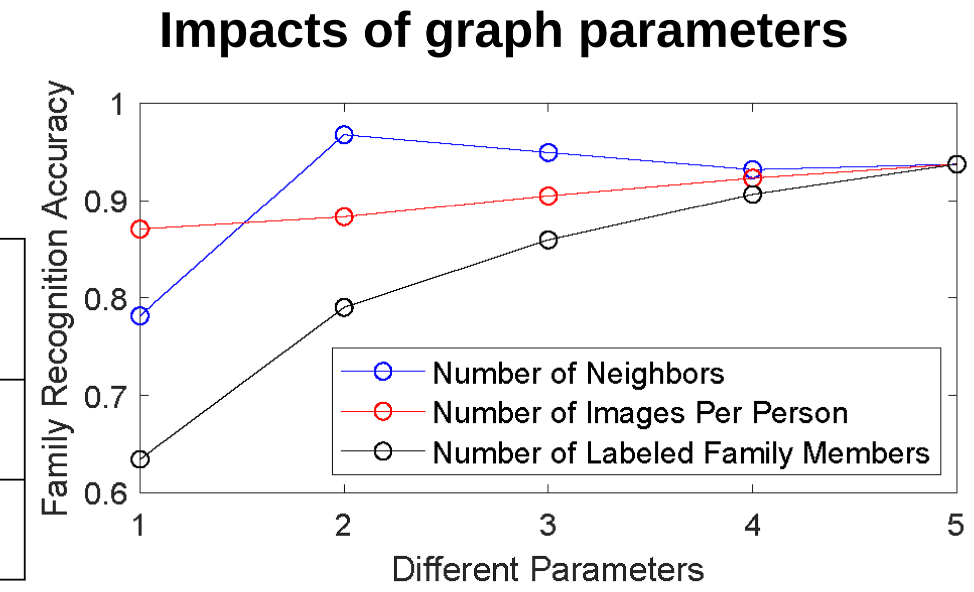
Family-300
- Contains 1712 subjects
- 10255 facial images
- 1712 nodes for training
- 8543 nodes for validation and testing

Results

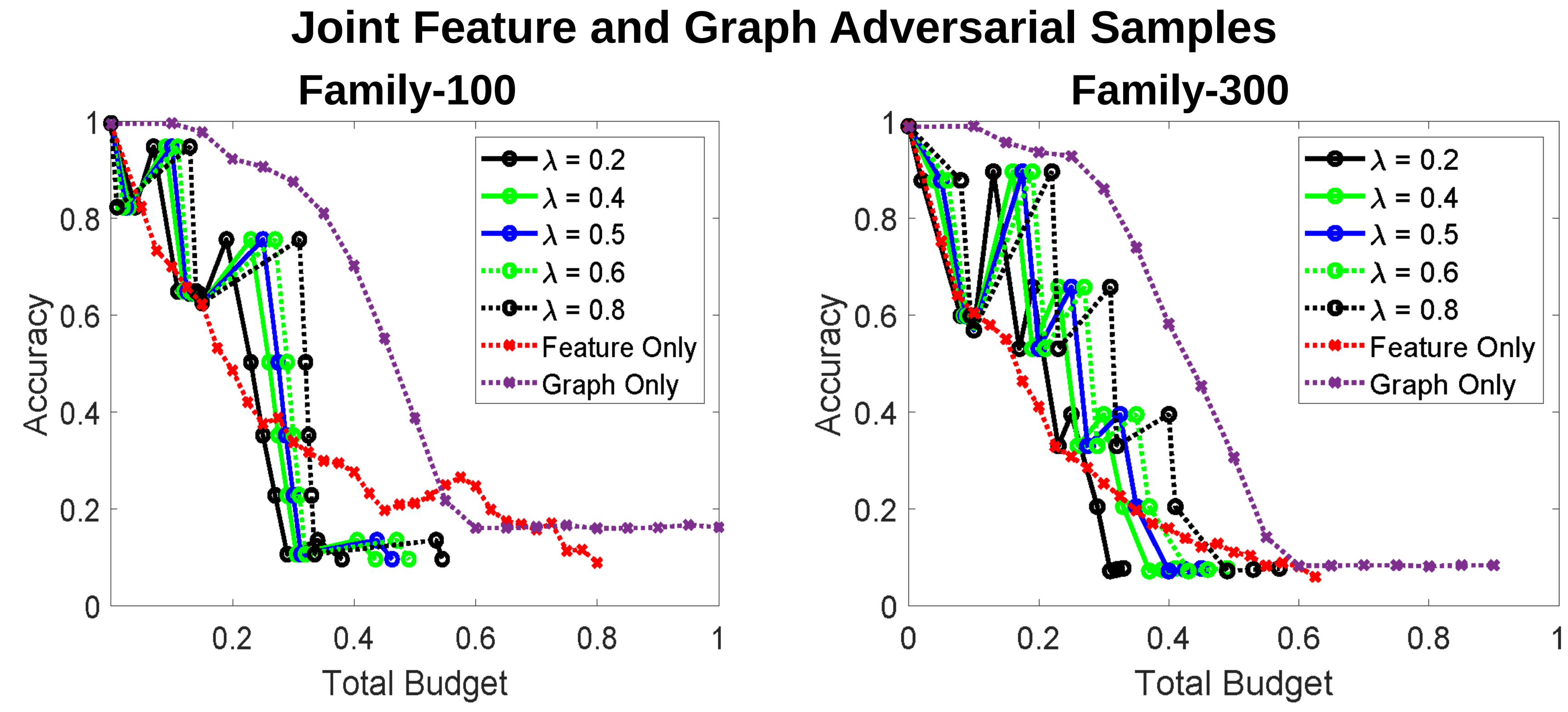
Family recognition on Facial Images only vs. Images + Graph

Model	Accuracy (%)
SphereNet	17
Ours	98.89

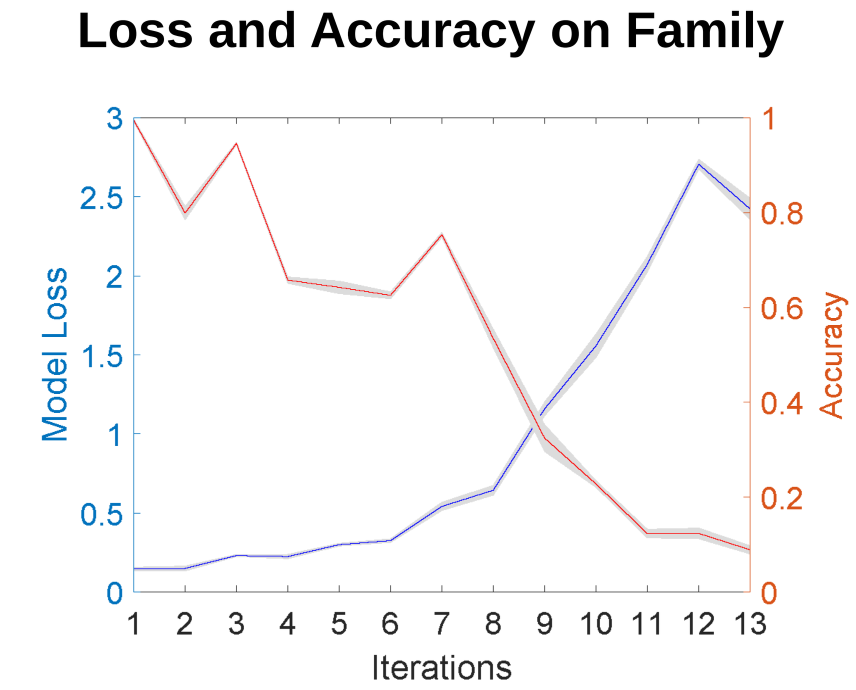
Impacts of graph parameters



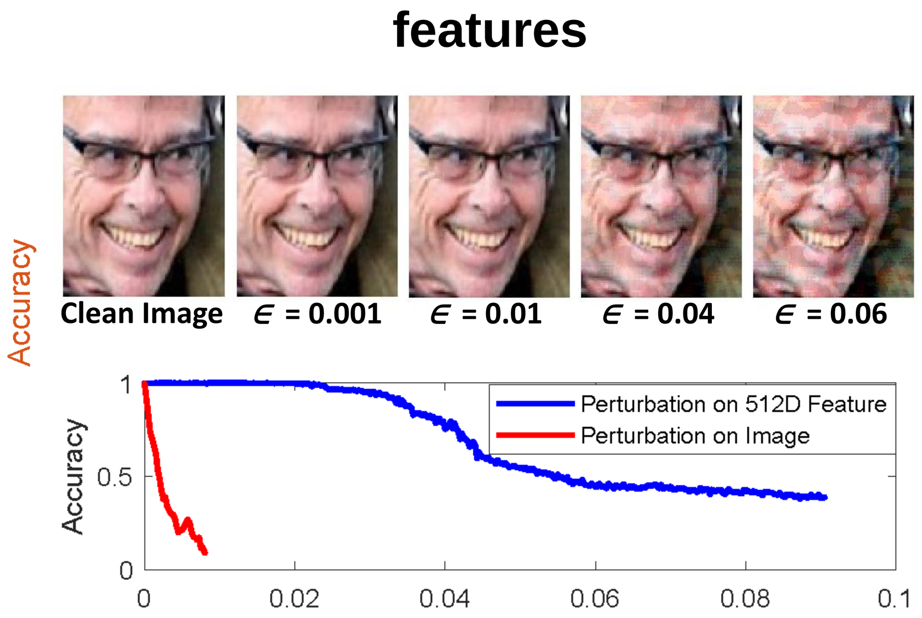
Joint Feature and Graph Adversarial Samples



Loss and Accuracy on Family



Impacts of ϵ on visual and node features



References

1. Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
2. Bojchevski, A., and Gunnemann, S. 2019. Adversarial attacks on node embeddings via graph poisoning. In International Conference on Machine Learning, 695–704