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

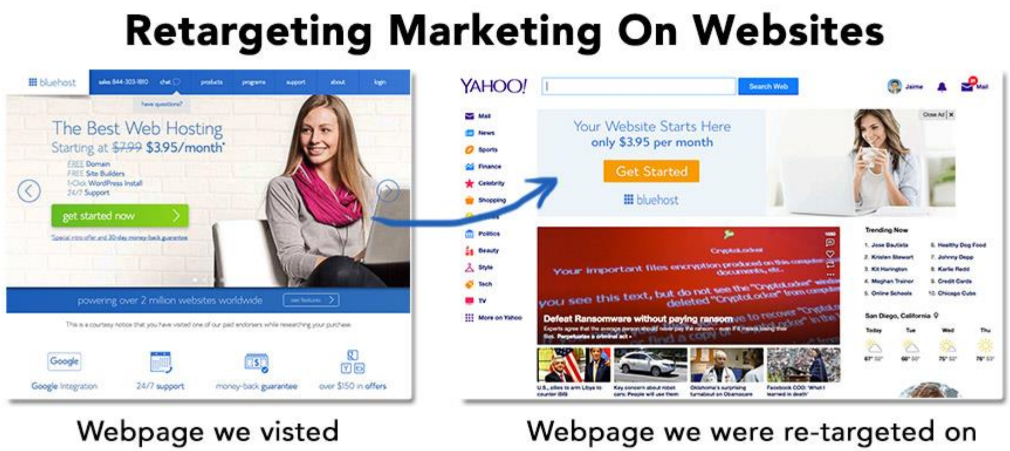
Chetan Kumar, Riazat Ryan, Ming Shao

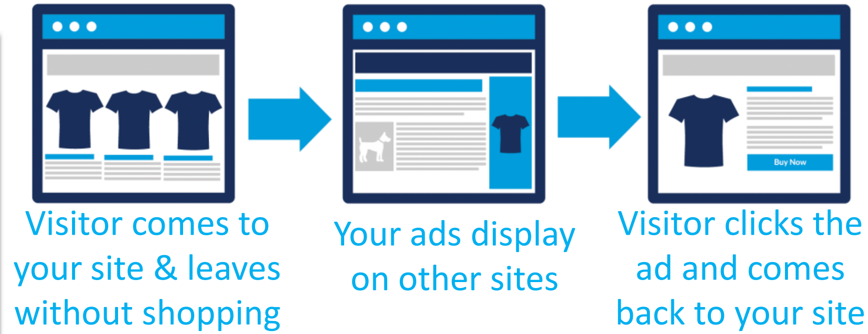
Department of Computer and Information Science,
University of Massachusetts, Dartmouth

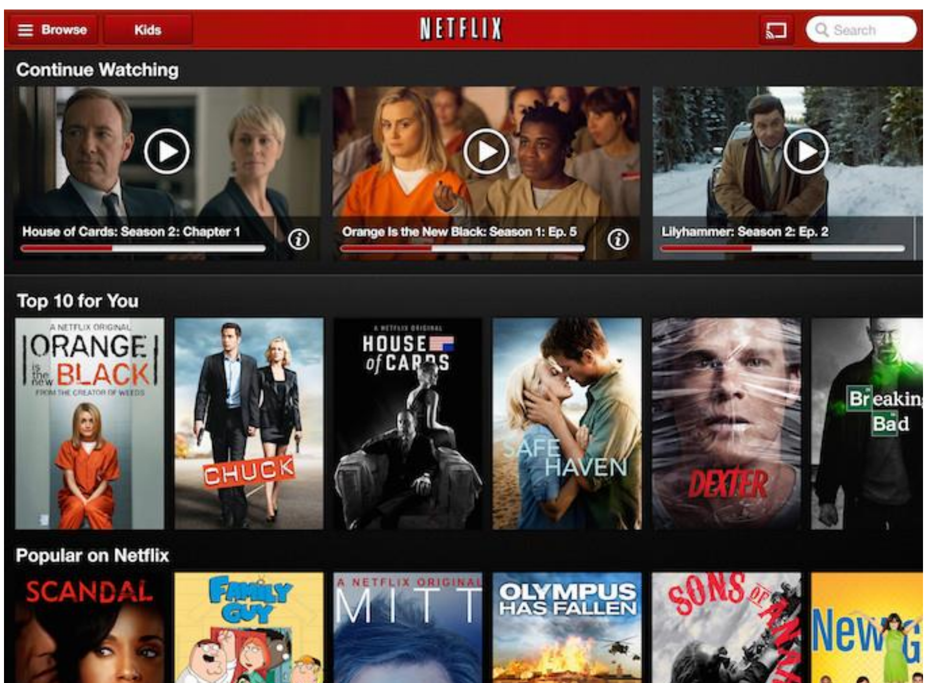
Data Leakage:

- Limited time to read Terms & Conditions
- Limited knowledge (especially children) to understand
- Unintentional leakage

Behavioral Targeting:





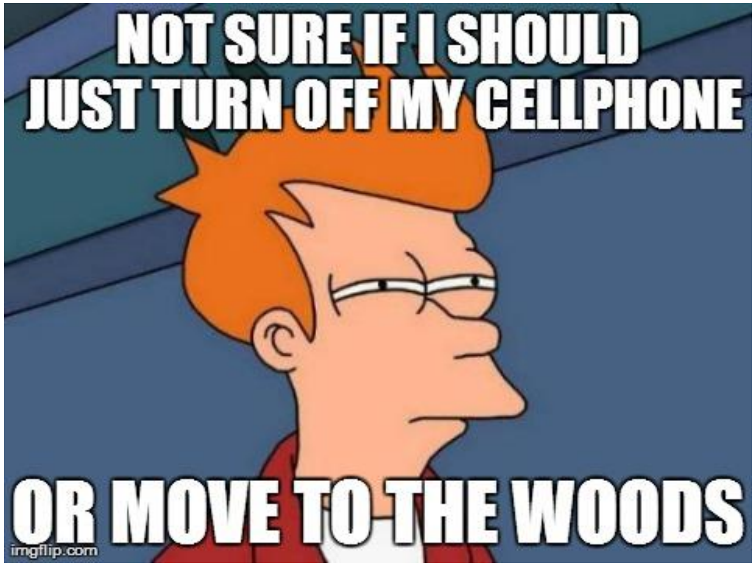


- Already developed Advanced Algorithms to analyze users' personal data and identity:
  - Shopping Habits
  - Movie Preferences
  - Reading Interests
  - etc.

Motivation:

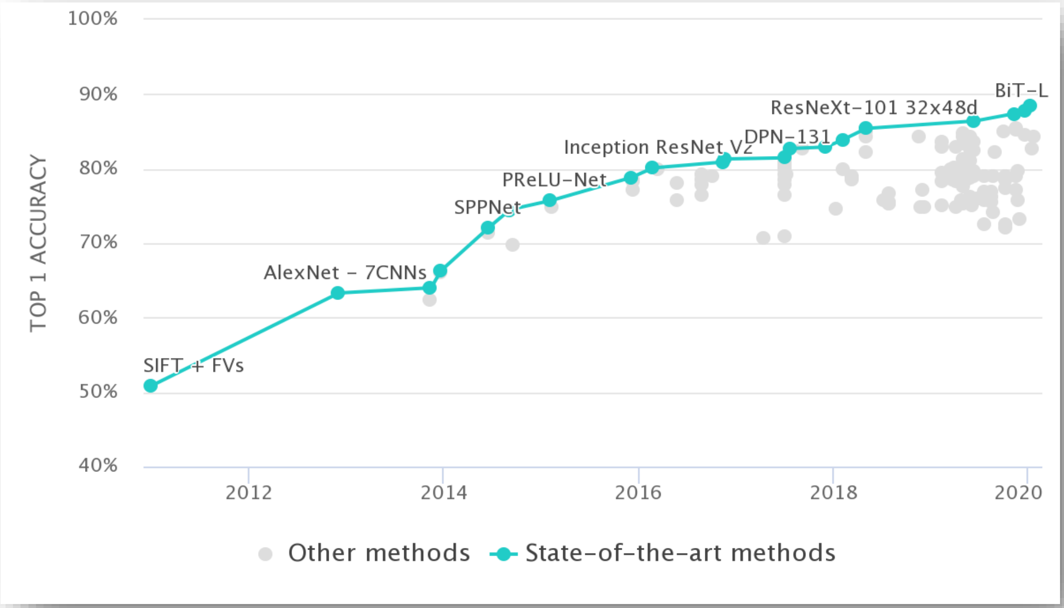
UMass | Dartmouth

- Generally, people have no willing to disclose personal data



- Image recognition has achieved significant process in the past decade

Image Classification on ImageNet



- Visual kinship understanding drawing more attention

Motivation:

UMass Dartmouth

- **Graph Neural Network (GNN)**
  - GNN provides a new perspective for learning with Graph
  - It may promote familial feature learning and understanding
- **Social Media**
  - Social Media is mainly featured by sharing photos and social connections (friend, relative, 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 a celebrity will be when their family members photos are exposed without their permission

Privacy Leakage over Social Media:

UMass Dartmouth

Photo Clicked by a Person

Privacy Leakage over Social Media:

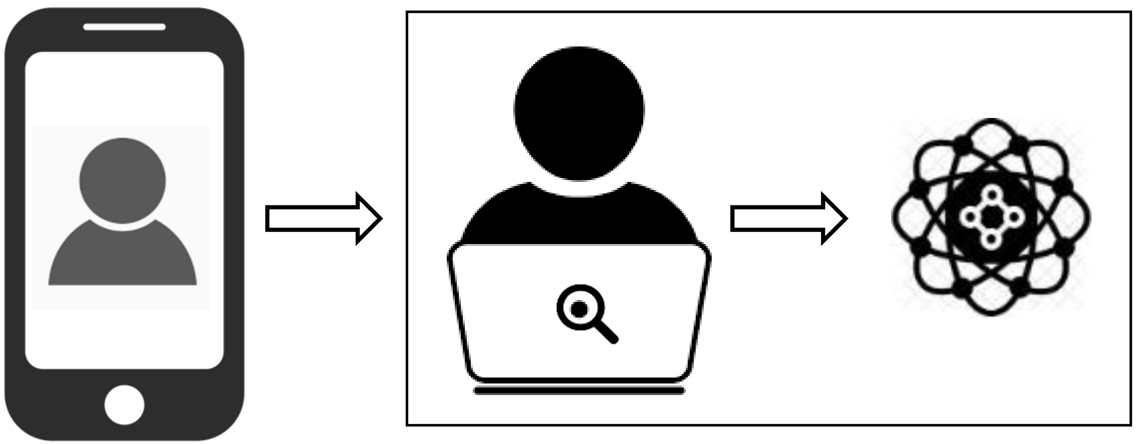


Photo Clicked by a Person

Family Information Searched over the Web

Privacy Leakage over Social Media:

UMass | Dartmouth

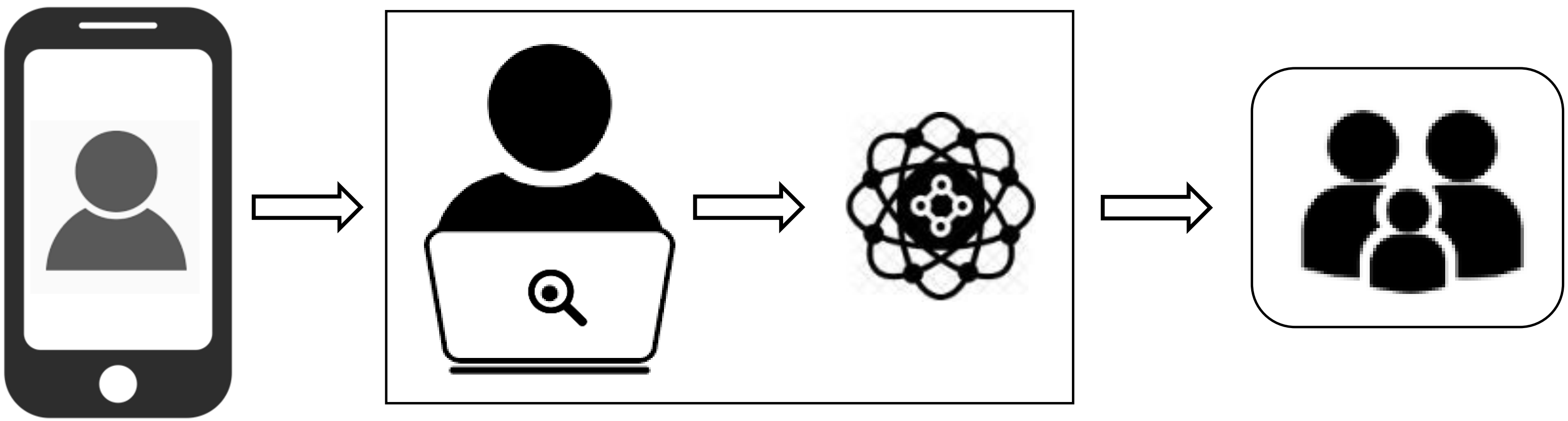


Photo Clicked by a Person

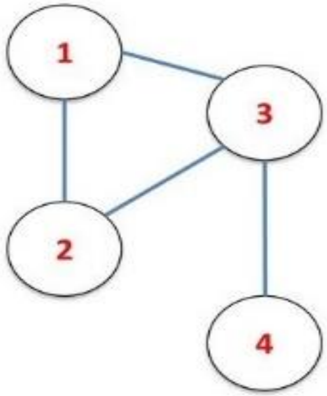
Family Information Searched over the Web

Family Data is Found

Family Recognition on the Graph:

UMass | Dartmouth

- $G = (V, E)$ an attributed and undirected graph



- The **adjacency matrix** $A \in \{0, 1\}^{N \times N}$

$$A_{ij} = \begin{cases} 1 & \text{if edge from vertex i to j} \\ 0 & \text{otherwise} \end{cases}$$

Adjacency matrix(A)

	1	2	3	4
1	0	1	1	0
2	1	0	1	0
3	1	1	0	1
4	0	0	1	0

- $X \in \mathbb{R}^{N \times D}$ represents **node features**
- $X_L \in \mathbb{R}^{D \times N_L}$ and $X_U \in \mathbb{R}^{D \times N_U}$ be the **labeled and unlabeled** image features
- $y_L \in \mathbb{R}^{N_L}$ is the **label vector**
- Goal is to find the mapping:

$$\boldsymbol{f_G}: ([\boldsymbol{X_L}, \boldsymbol{X_U}]) \to ([\boldsymbol{y_L}, \boldsymbol{y_U}])$$

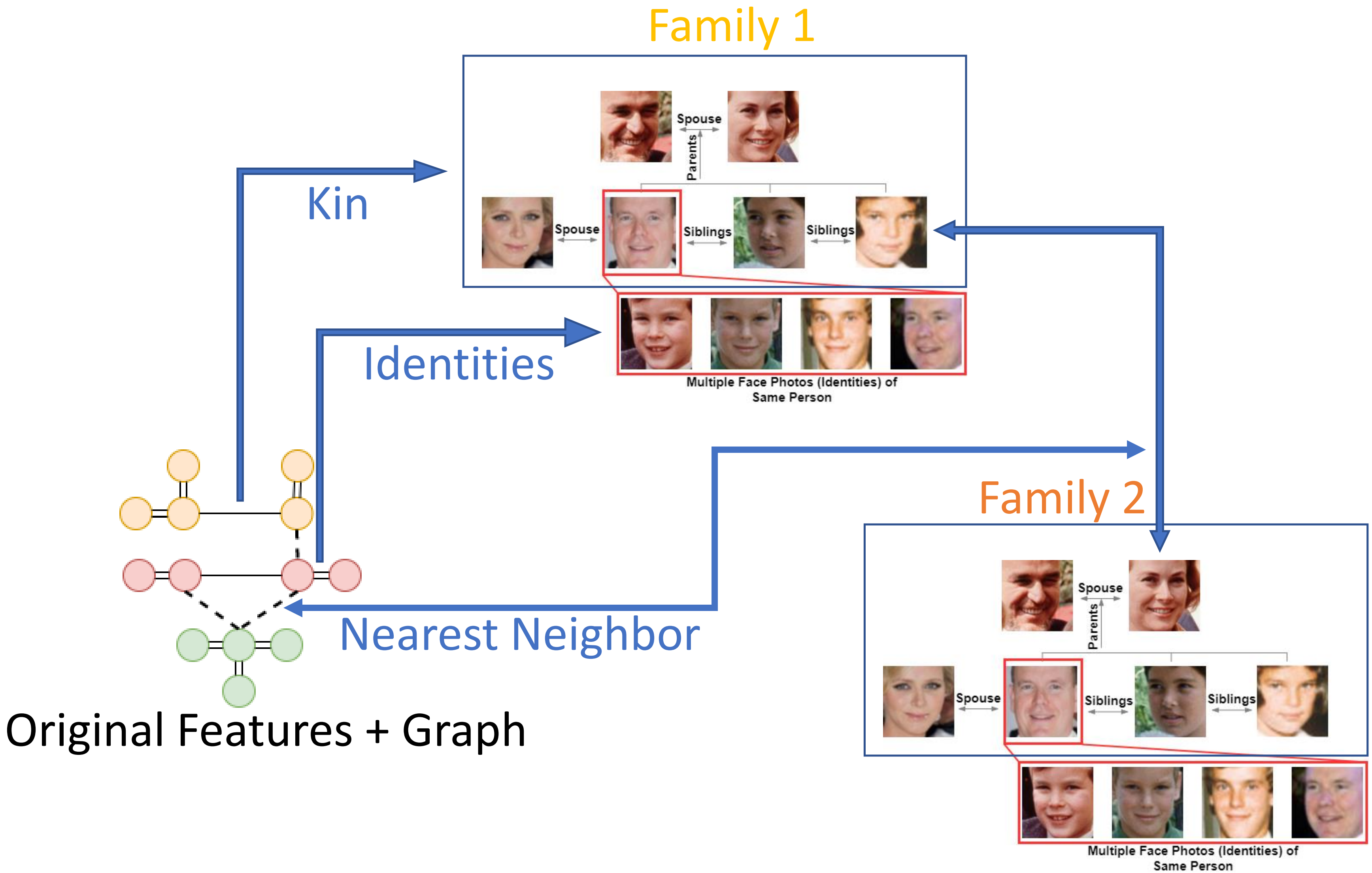
Degree matrix (D)

	1	2	3	4
1	2	0	0	0
2	0	2	0	0
3	0	0	2	0
4	0	0	0	1

Graph Construction:

UMass | Dartmouth

- IDs (Identities)
- Kin (Family Relation)
- NN (Nearest Neighbor)



Model Learning:

$$H^{(l)} = \sigma\,[\underbrace{D'^{-\frac{1}{2}} A' D'^{-\frac{1}{2}}}_{\text{Normalize Graph Structure}}\, \underbrace{H^{(l-1)} W^{(l-1)}}_{\text{Multiply Node Parameters and Weights}}]$$

Output to next layer/Result

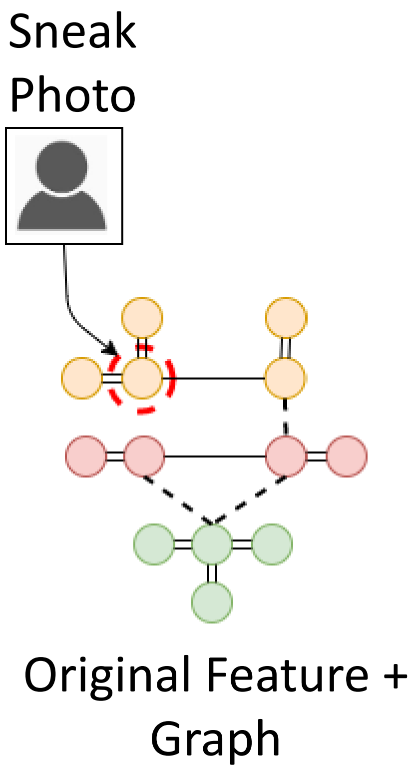
ReLU Function

Where,

- $A' = (A + I)$ to add self-loops
- D' is the Degree Matrix of A' to normalize large degree nodes
- $H^0 = X$

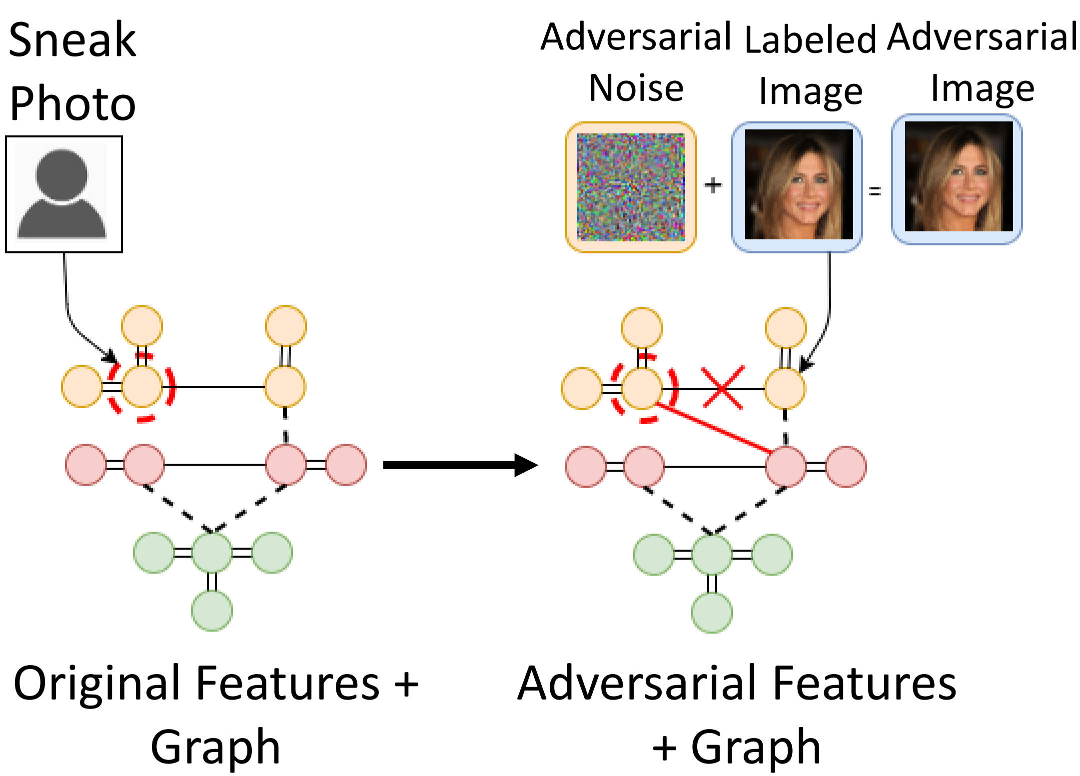
Model Framework:

- Privacy at Risk
  - Social media data may expose sensitive personal information
  - This can be leveraged and lead to information leakage without user's attention



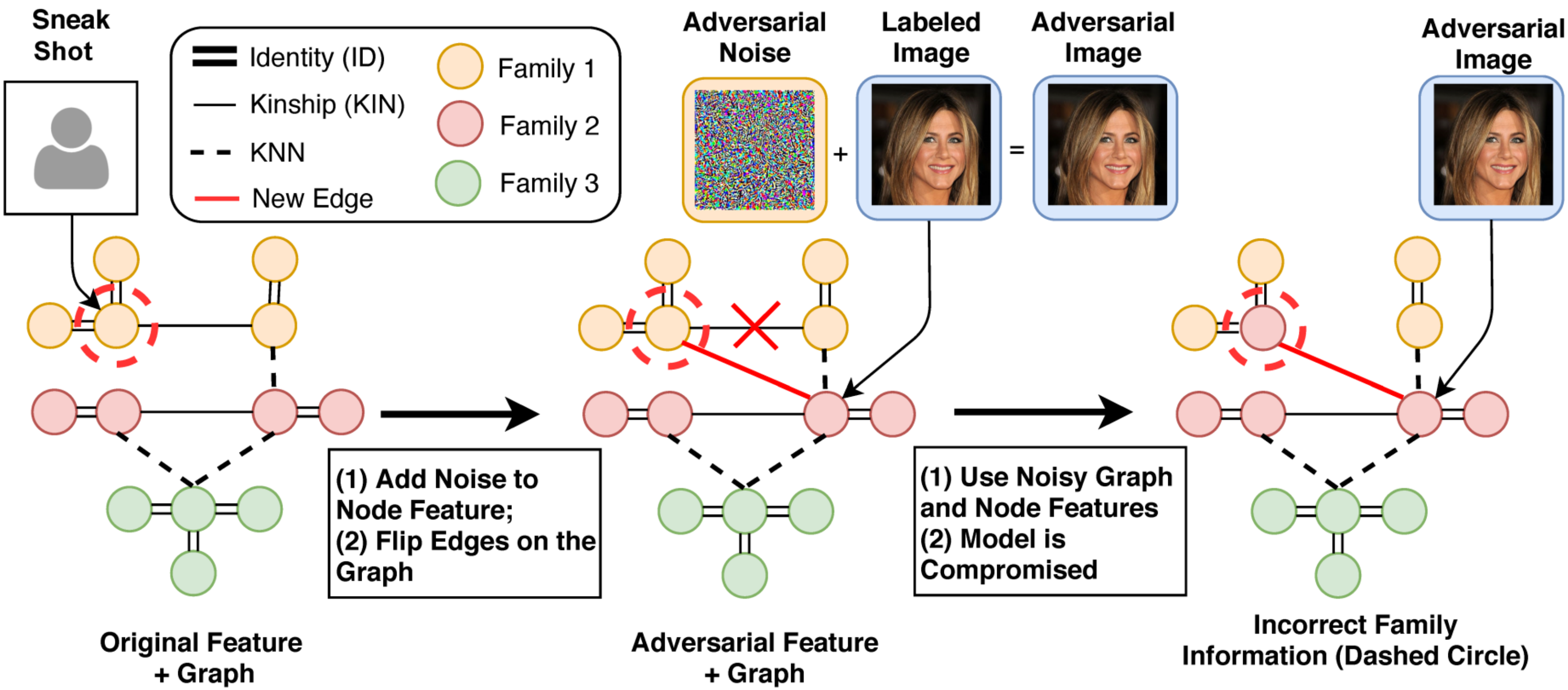
Model Framework:

- Adversarial Attack:
  - Added Noise to Node Features by calculating sign of the Gradient
  - Added/Removed edges (relationships) between nodes

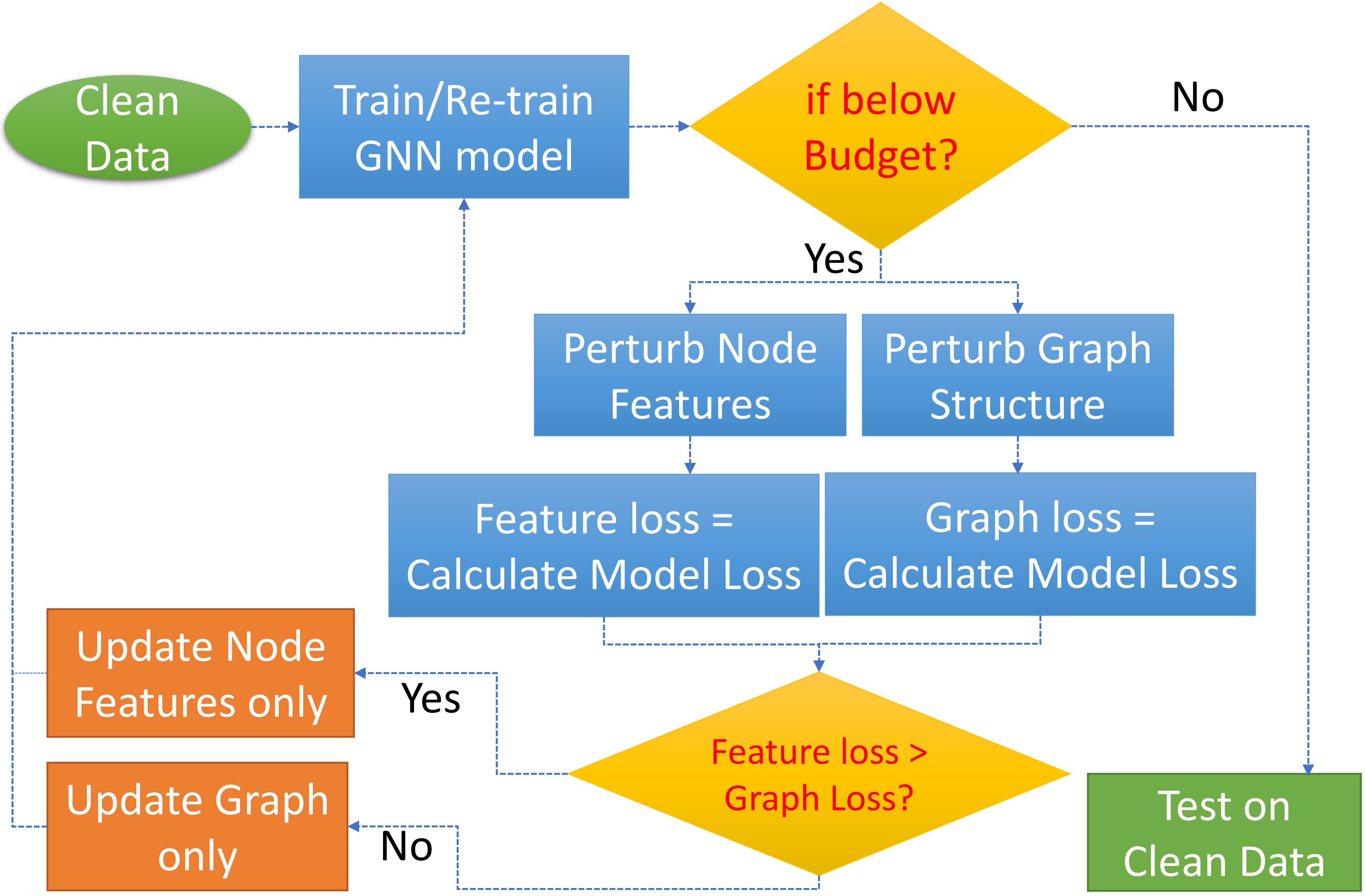


Model Framework:

- Model Compromised:
  - By using Noisy Features and Noisy Graph



Algorithm:


Clean Data
Train/Re-train GNN model
if below Budget?
No
Yes
Perturb Node Features
Perturb Graph Structure
Feature loss = Calculate Model Loss
Graph loss = Calculate Model Loss
Feature loss > Graph Loss?
Yes
Update Node Features only
No
Update Graph only
Test on Clean Data

Joint Feature and Graph Adversarial Samples

The proposed joint attack model can be formulated as:

$$\max_{\{X',A'\}} \mathcal{L}_{\mathcal{AD}}(X', A') \triangleq \max_{\{X',A'\}} \ln Z^*_{\text{pert}} - \ln Z^*_{\text{clean}},$$

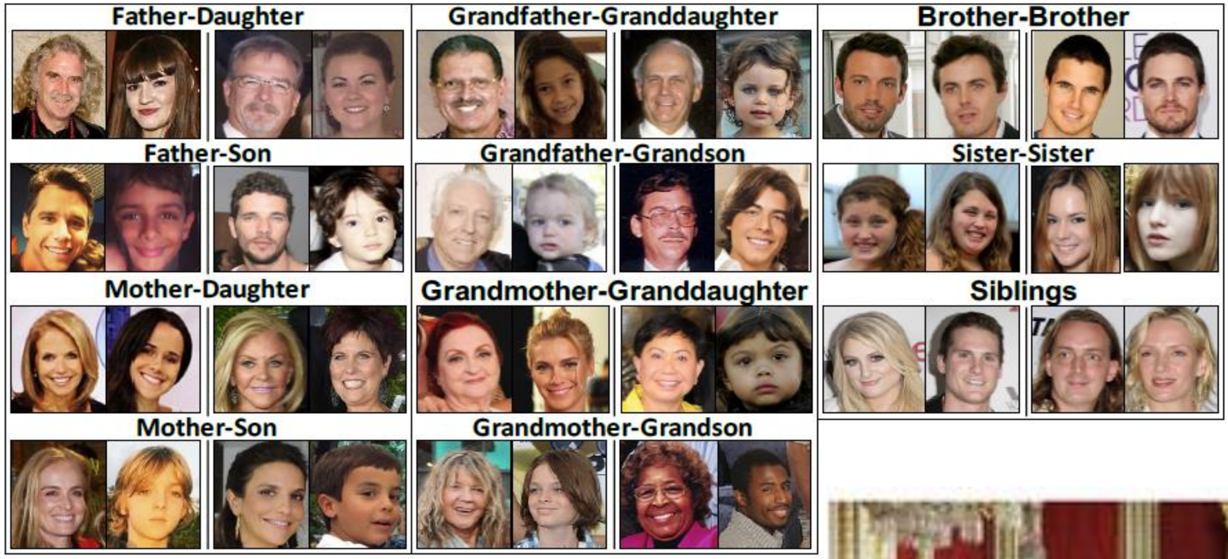
$$\text{s.t., } \lambda\|A - A'\|_0 + (1-\lambda)\|X - X'\|_F \leq \theta$$

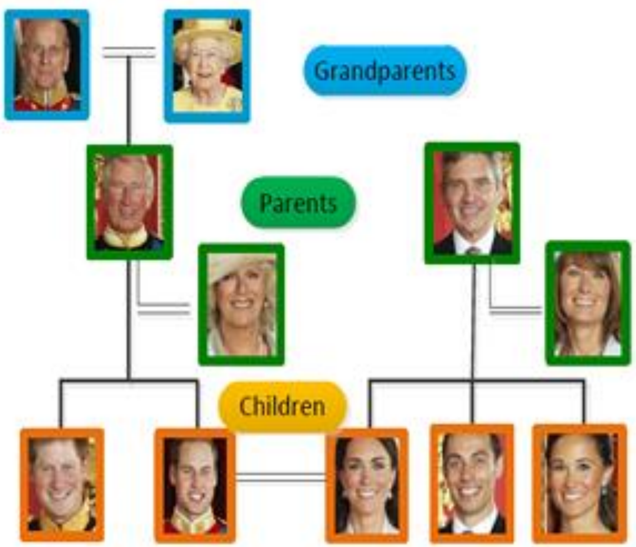
Here,

- L_{AD} is the loss function of the joint attack
- $||.||_F$ is the matrix Frobenius norm
- λ is the balancing parameter
- Z^*_{pert} is the softmax output of the perturbed labeled data
- Z^*_{clean} is based on clean features and graph

Datasets:

Families in the Wild (FIW)



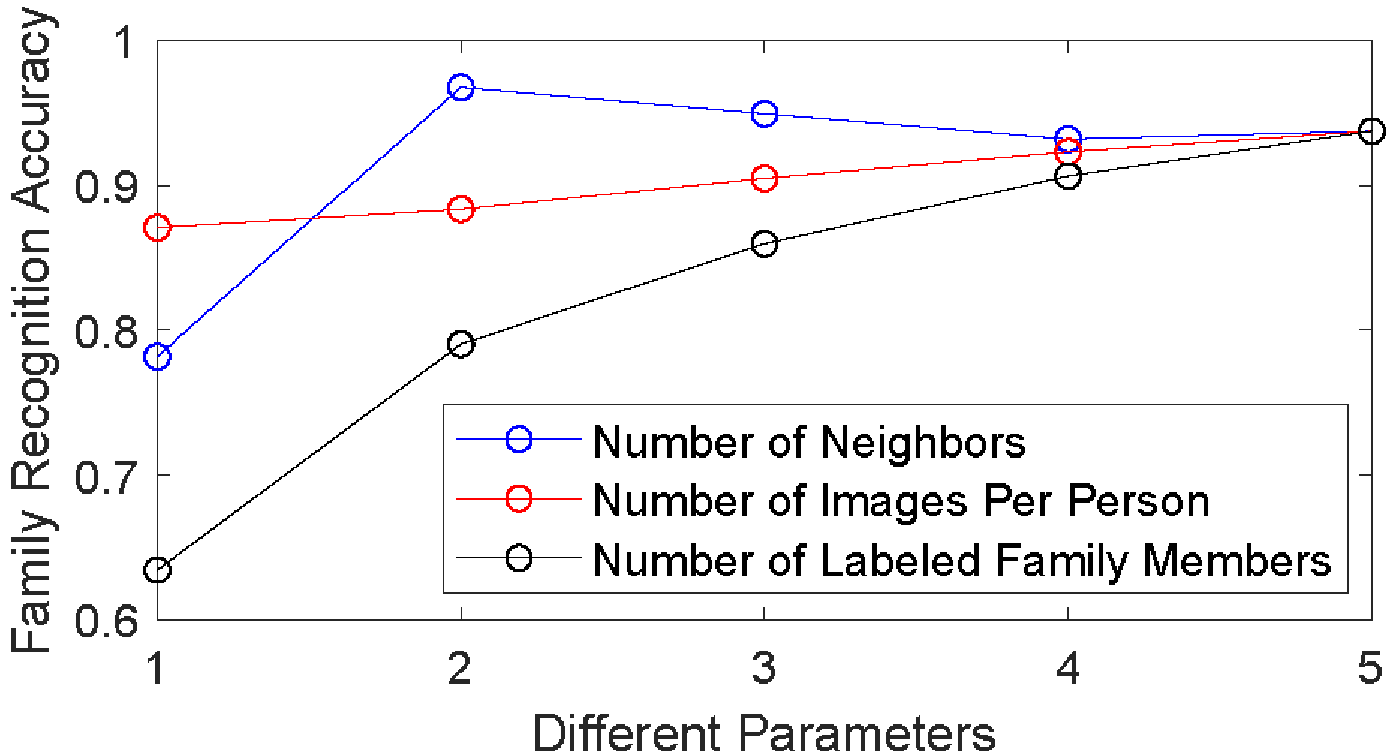


Datasets:

- **Pre-processing**
  - Extracting image features using pre-trained SphereNet
  - Constructed the social graph (IDs, Kin, *k*-NN)
  - Created two social networks
    - Family-100
      - Contains 502 subjects
      - 2758 facial images
      - 502/2758 nodes for training
      - 2256 for validation and testing
    - Family-300
      - Contains 1712 subjects
      - 10255 facial images
      - 1712/10255 for training
      - 8543 for validation and testing

Results:

- **Impacts of graph parameters**
  - Best value for *k* = 2
  - Best value for ID and Kin= 5



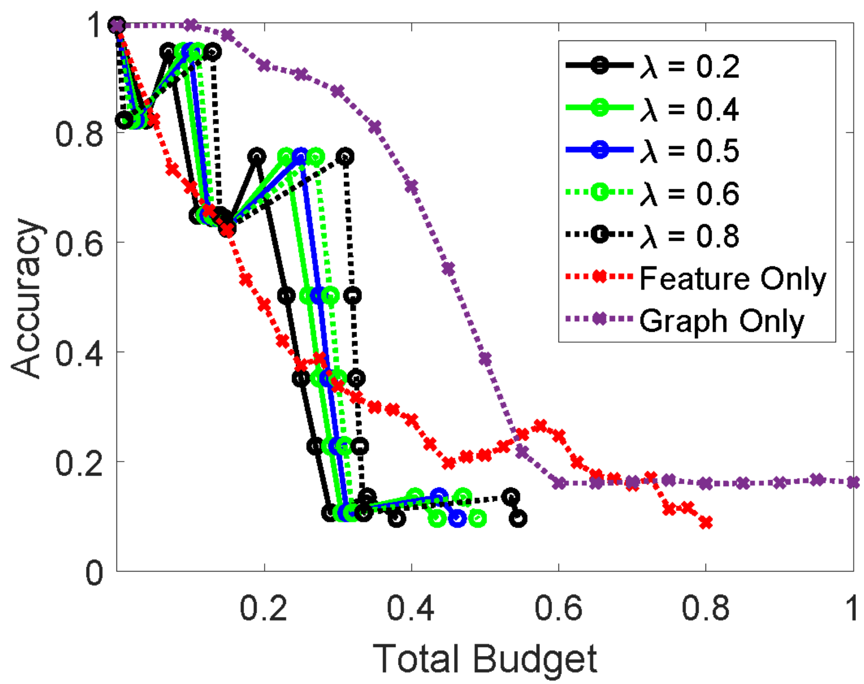
Results:

Joint Feature and Graph Adversarial Samples

$$Total\text{–}Budget = \lambda * Edge\text{–}Flipping\text{–}Ratio + (1\text{–}\lambda) * 100 * \epsilon$$

Family-100

- Single Attack
  - Feature only and graph only attacks are implemented
  - But excessive use of any particular attack compromises the ***data*** largely, i.e., perceivable visual change
- Joint Attack
  - We propose a joint attack which proves more cost-efficiency

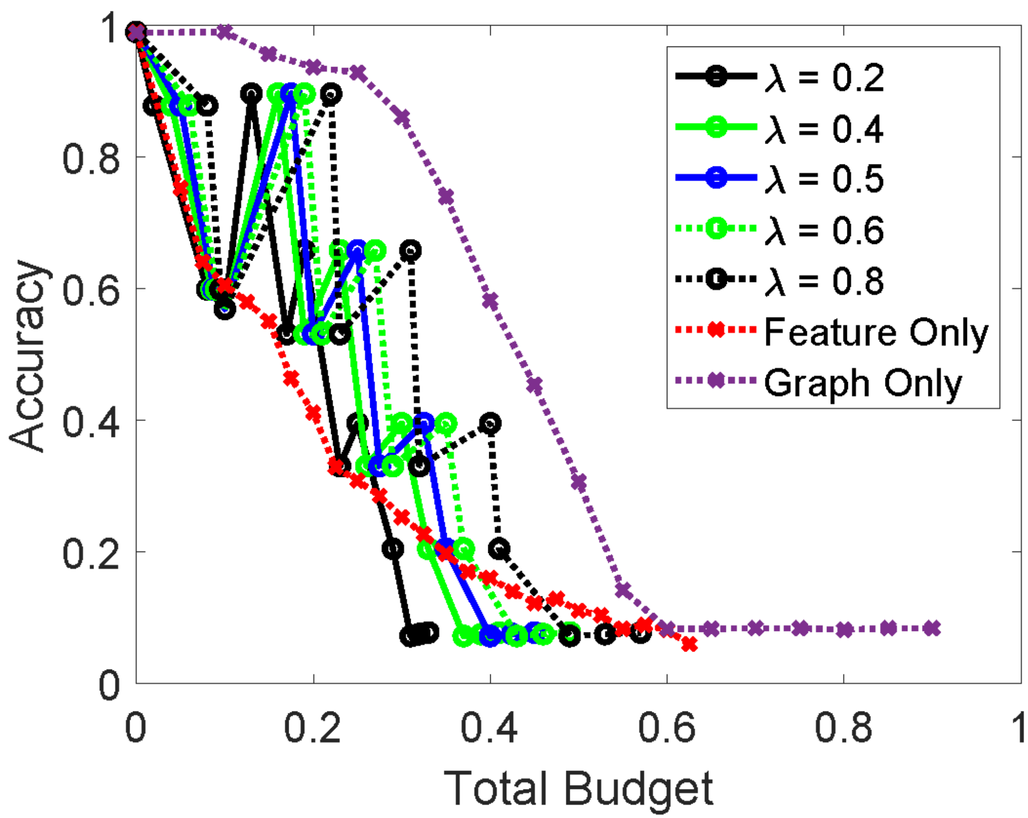


Results:

Joint Feature and Graph Adversarial Samples

Family-300

- Single Attack
- Joint Attack

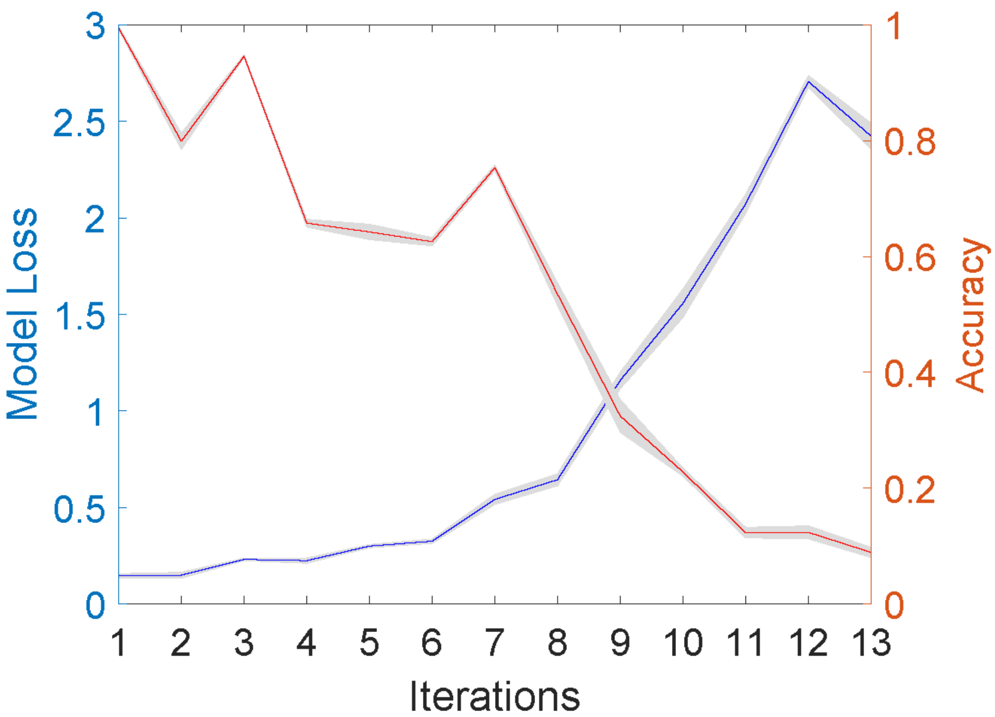


Results:

UMass Dartmouth

Loss and Accuracy on Family-100

- Run the ***Joint Attack Algorithm*** for 13 iterations
- Average result for 5 trials
- ***Accuracy*** decreased with more iterations
- And ***Model Loss*** is increasing



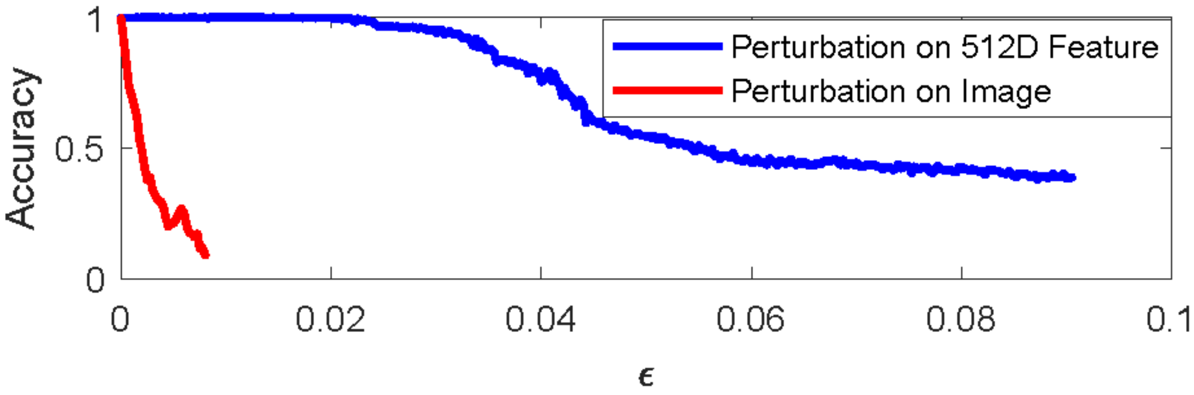
Qualitative Evaluation:

UMass Dartmouth

Impacts of $\in$ on image and node features

- High-dimensional raw image data require weak noise to fool the model
- Low-dimensional visual features require relatively strong noise to fool the model

Clean Image	$\epsilon = 0.001$	$\epsilon = 0.01$	$\epsilon = 0.04$	$\epsilon = 0.06$



Conclusion:

- Demonstrated the family information was at risk on social network through plain graph neural networks
- Proposed a joint adversarial attack modeling on both features and graph structure for family privacy protection
- Qualitatively showed the effectiveness of our framework on networked visual family datasets

- **Future extension:** Adapt our modeling to different types of data and other privacy related issues

Acknowledgement:

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References:

1. https://techcrunch.com/2014/05/19/netflix-neil-hunt-internet-week/
2. https://www.business2community.com/marketing/multiple-benefits-retargeting-ads-01561396
3. https://blog.ladder.io/retargeting-ads/
4. https://reelgood.com/movie/terms-and-conditions-may-apply-2013
5. https://clclt.com/charlotte/cucalorus-report-part-3/Content?oid=3263928
6. https://www.capitalxtra.com/terms-conditions/general/
7. https://paperswithcode.com/sota/image-classification-on-imagenet

Q & A

Thank you

www.chetan-kumar.com
http://www.cis.umassd.edu/~rryan2/