You are Who You Know and How You Behave: Attribute Inference Attacks via Users' Social Friends and Behaviors

Neil Zhenqiang Gong
Iowa State University

Bin Liu
Rutgers University

25th USENIX Security Symposium, Aug., 2016, Austin

Online Social Networks are Popular

1.71 billion users

111 million users

310 million users

OSNs Are a Mixture of Public and Private Information

- Public information
  - Friends
  - User behaviors
    - Like/share/review webpages and apps
  - Self-reported attributes
    - Education, employment, location
- Private information
  - Personal interests
  - Sexual orientation
  - Drug usage
  - Religious view
  - …

Attribute Inference Attacks

- Given public information of some users
- Infer private attributes of some target users

Existing Attribute Inference Attacks

- Friend based: *you are who you know*

- Behavior based: *you are how you behave*

Our Attack

Combine *both* friends and behaviors

Roadmap

- Threat model
- Our attack algorithm
- Evaluation
- Conclusion

Threat Model

- Attackers
  - Cyber criminal
  - OSN provider
  - Advertiser
  - Data broker

- Attack procedure
  - Attacker collects publicly available friends, user attributes, and behaviors
  - Use our algorithm to infer private attributes of target users

Threat Model

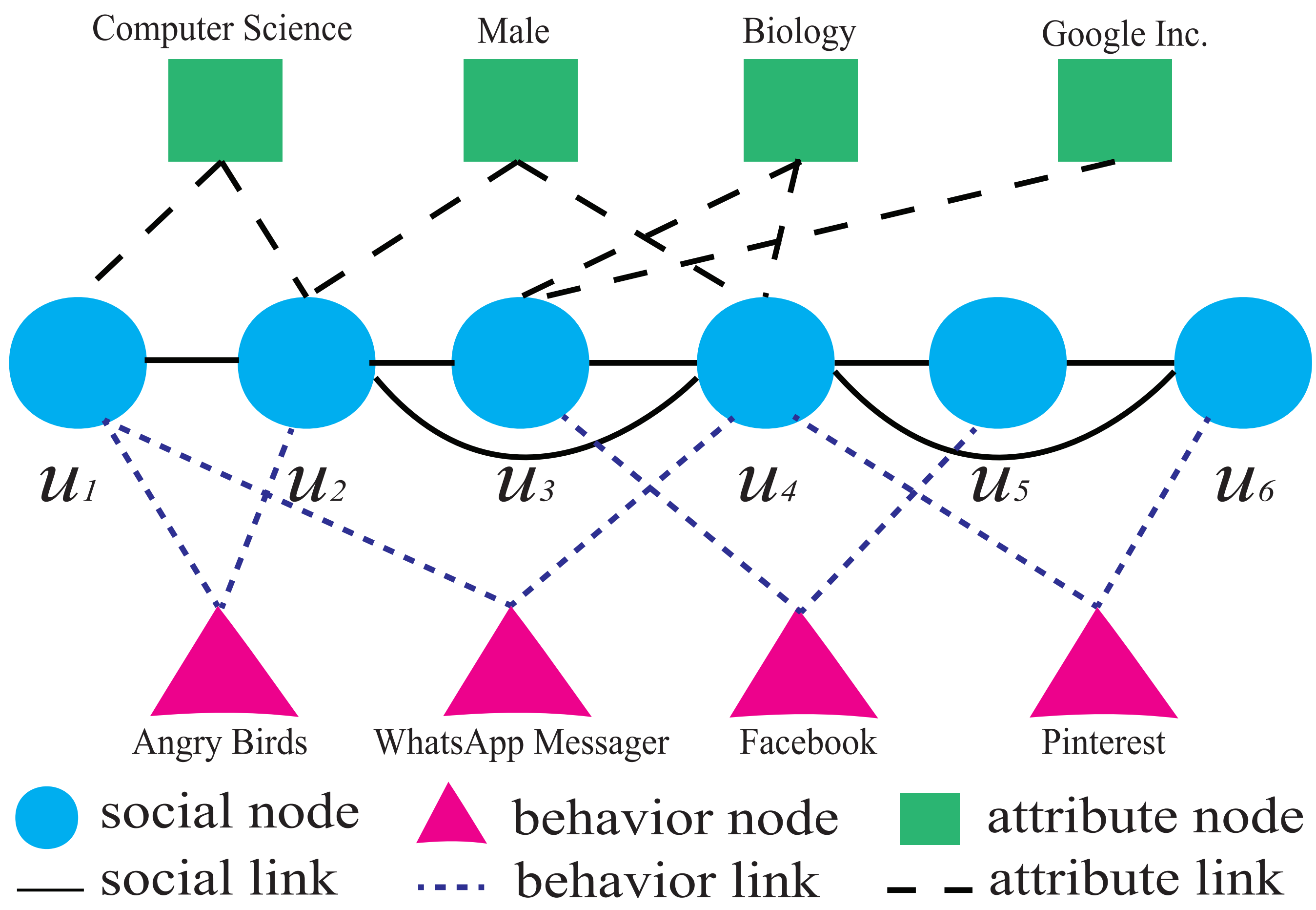
- Implication/Application of attribute inference attacks
  - Privacy threat
  - Targeted advertisement
  - Targeted phishing attacks
  - Breaking "security question" based user authentication
- Perform further attacks
  - Help profile users across social networks
  - Help combine online profile with offline data

Our Attack Algorithm: High-Level Overview

- Construct a Social-Behavior-Attribute (SBA) network to unify friends, attributes, and behavior information

- For a target user, find the most "similar" attributes on the SBA network based on *homophily*
  - Homophily: users that have similar attributes share similar friends and behaviors

Social-Behavior-Attribute (SBA) Network



Vote Distribution Attack (VIAL) Algorithm

- Phase I:
  - Iteratively distribute a fixed vote capacity from the *targeted user v* to the rest of users

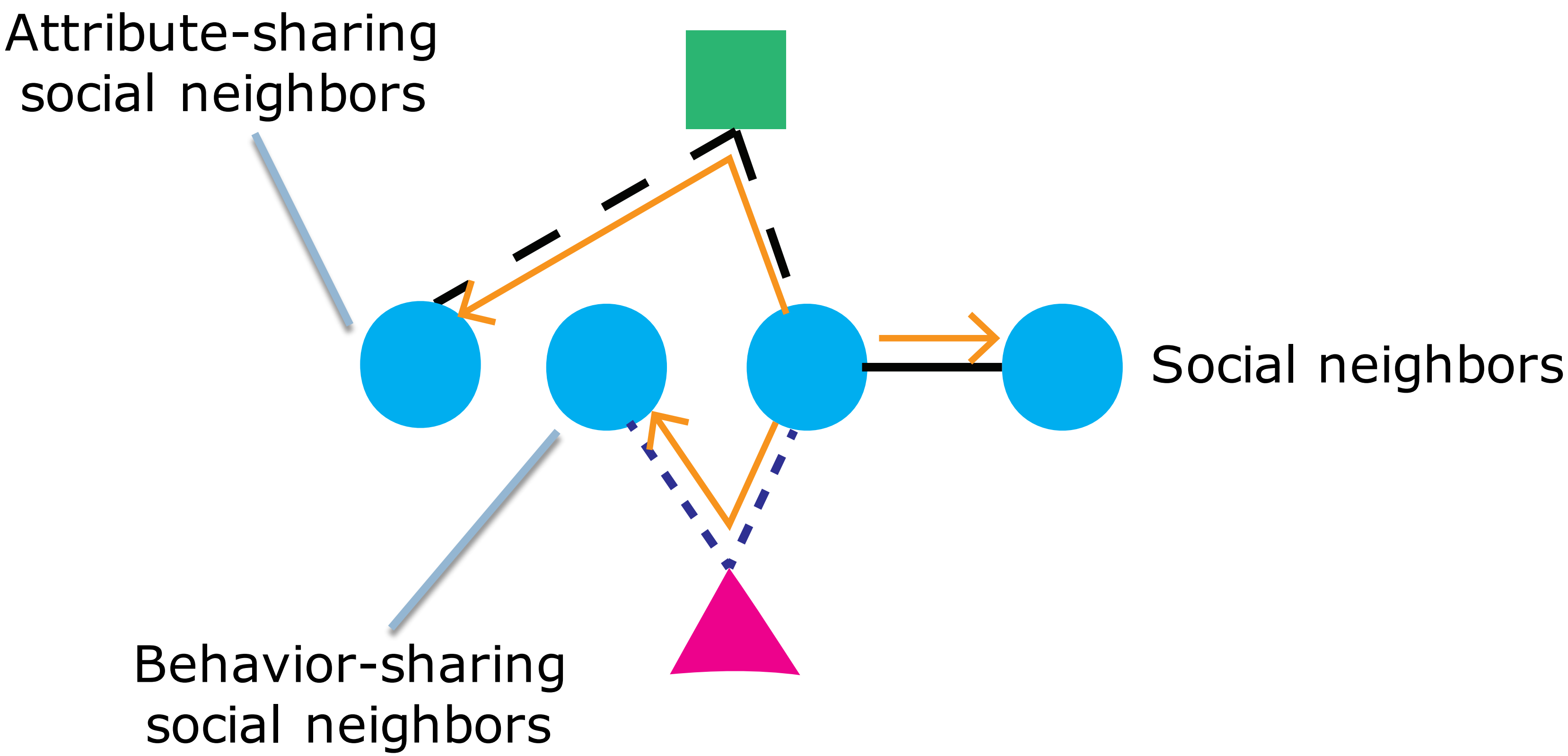
- Phase II:
  - Each user votes his/her own attributes using his/her vote capacity
  - The target user is predicted to have the attribute values that receive the highest votes

Phase I- Distributing Vote Capacity

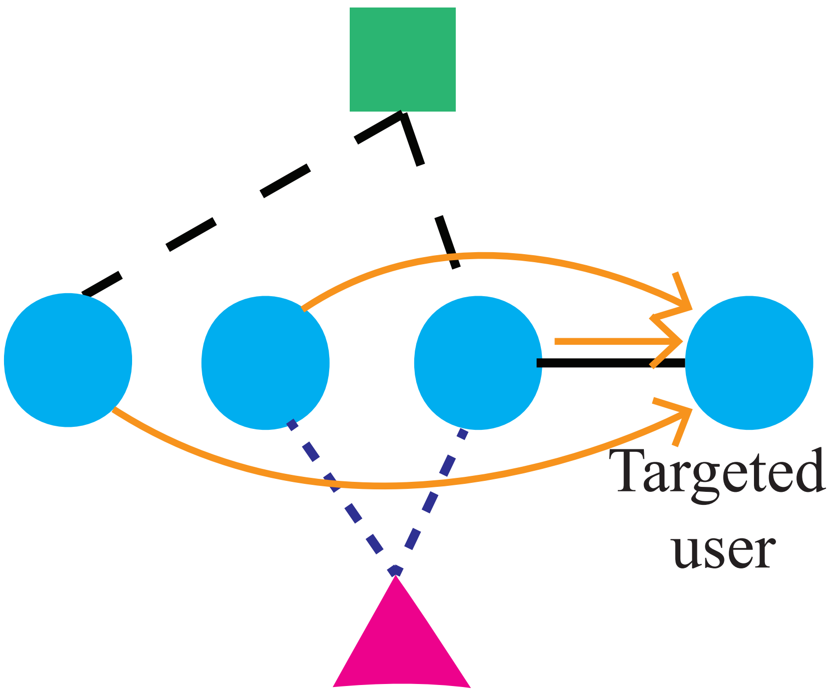
- A user receives a high vote capacity if the user and the targeted user are structurally similar

- Distribution via three local rules
  - Dividing
  - Backtracking
  - Aggregating

Local Rule I: Dividing

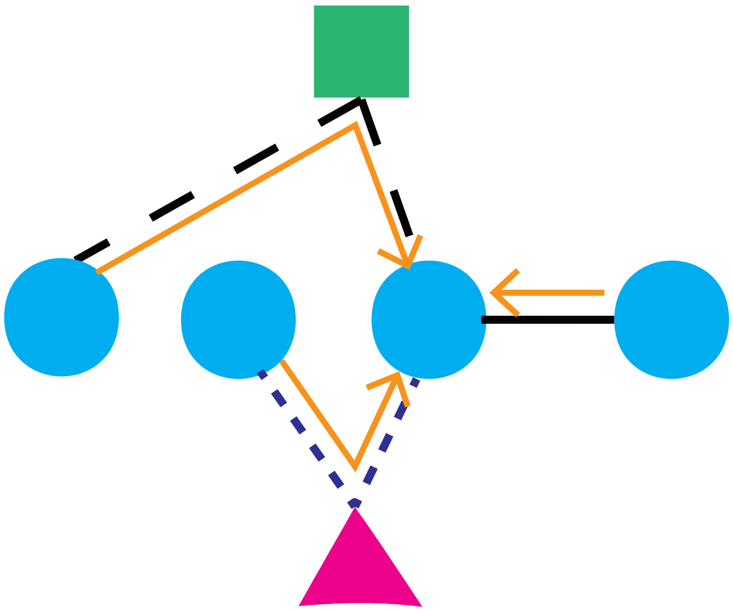


Local Rule II: Backtracking



Take a portion of *a user*'s vote capacity back to the targeted user

Local Rule III: Aggregating



Compute a new vote capacity for *a user* by aggregating the vote capacities from its neighbors

Phase II:

- In the end of Phase I, each user has a certain vote capacity
- Each user divides its vote capacity to its own attributes
- Each attribute sums the received votes
- Attributes with the highest votes are predicted to belong to the targeted user

Evaluation Data

- One snapshot of Google+ from Gong et al. (IMC'12)
  - Friends
  - Publicly available attributes
- Collect behaviors from Google Play
  - Liked/reviewed apps, movies, books, etc.

Evaluation Data

- Considered attributes
  - Major (62)
  - Employer (78)
  - Cities lived (70)

- Construct a SBA network

#nodes			#links		
social	behavior	attri.	social	behavior	attri.
1,111,905	48,706	210	5,328,308	3,635,231	269,997

Evaluation Setting

- Randomly sample a set of users
- Remove their attributes as ground-truth
- Treat them as targeted users
- Predict top-K attributes for each targeted user
- Measure Precision, Recall, and F-Score

Comparing with (Best) Friend-based and Behavior-based Attacks

Absolute performance gain
Relative performance gain

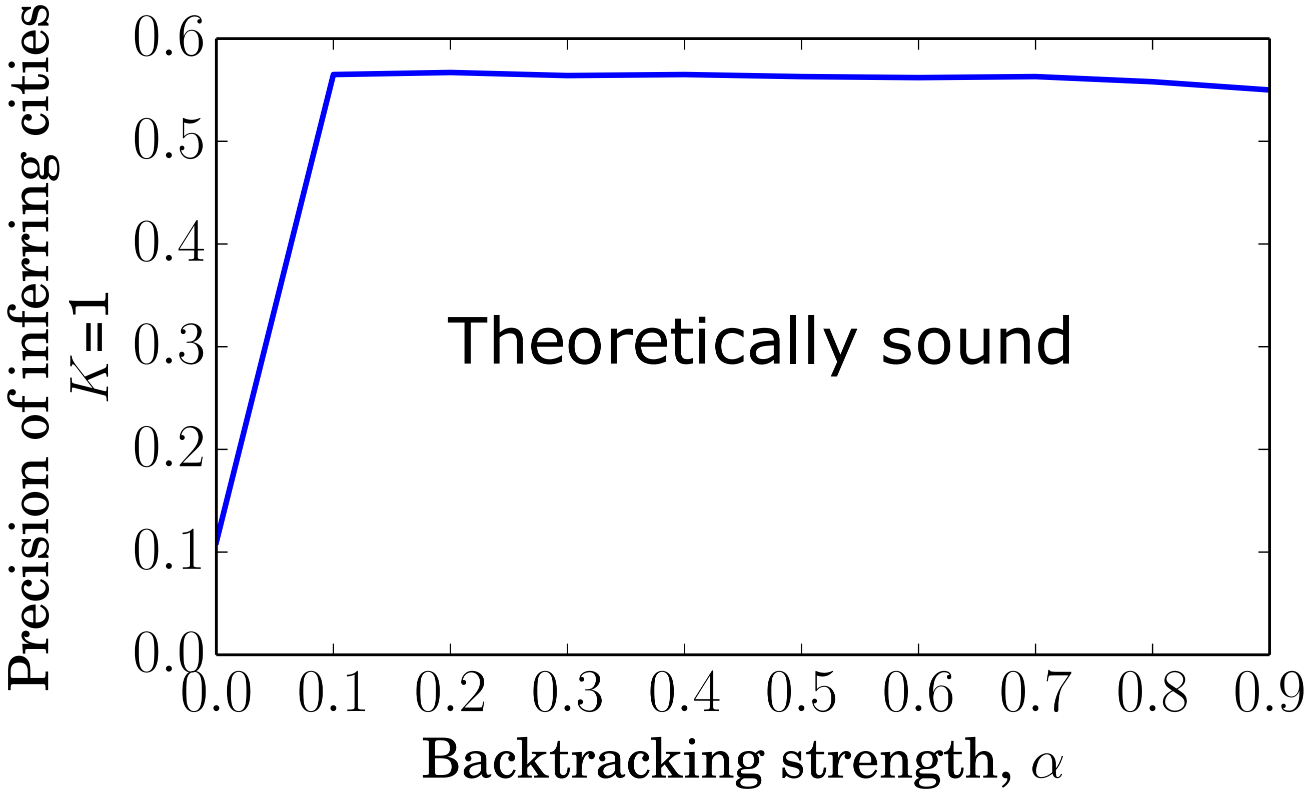
Attack	ΔP	ΔP%	ΔR	ΔR%	ΔF	ΔF%
Random	0.36	526%	0.22	535%	0.27	534%
RWwR-SAN	0.07	20%	0.05	23%	0.06	22%
VIAL-B	0.22	102%	0.13	99%	0.16	100%

Best behavior-based attack

Best friend-based attack

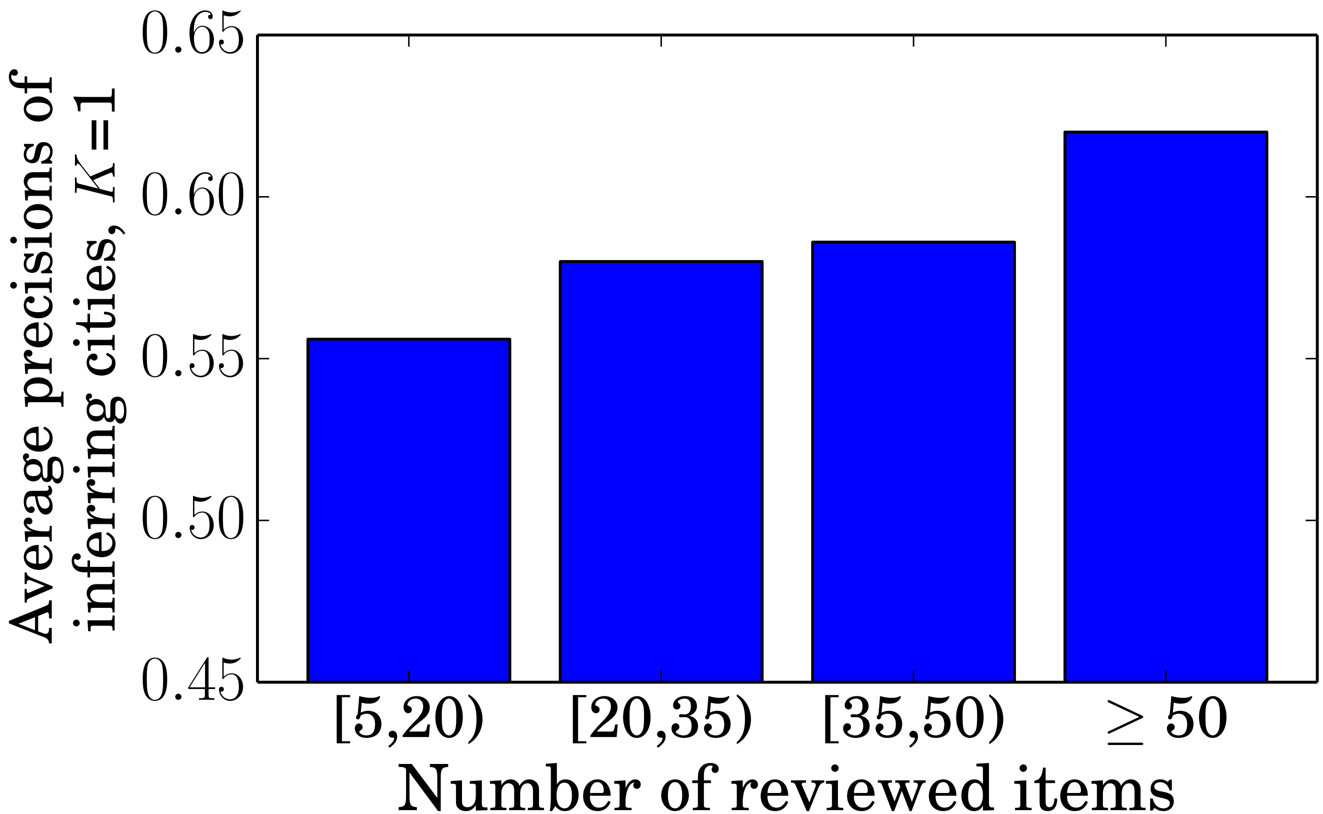
Our attacks are significantly more accurate than existing ones

Backtracking is Important



Backtracking substantially improves attack success rates

Sharing More Behaviors Makes You More Vulnerable

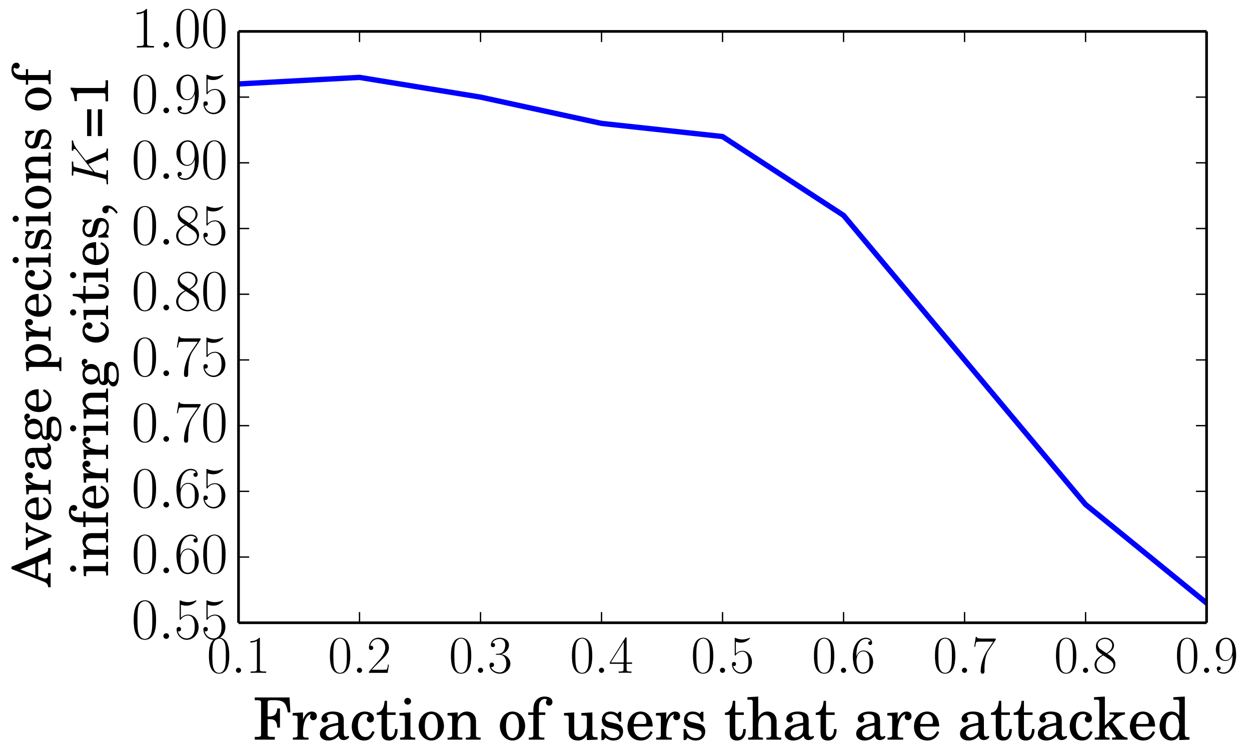


Attack success rates are higher when more behaviors are available

Confidence Estimation

- Produce a confidence score for each targeted user to measure how confident we are about its inference
- Choose to attack targeted users whose confidence scores are higher than a *threshold*
- Trade-off between attack success rates and #attacked targeted users
  - Higher threshold -> higher success rates & less attacked users

Trade-off Result



Success rates can be significantly improved when selectively attacking half of targeted users

Conclusion

- Attribute inference attacks for online social network users are feasible at large scale

- Fundamental reasons
  - Private attributes and public information are correlated
  - Machine learning/Data mining algorithms can capture such correlations