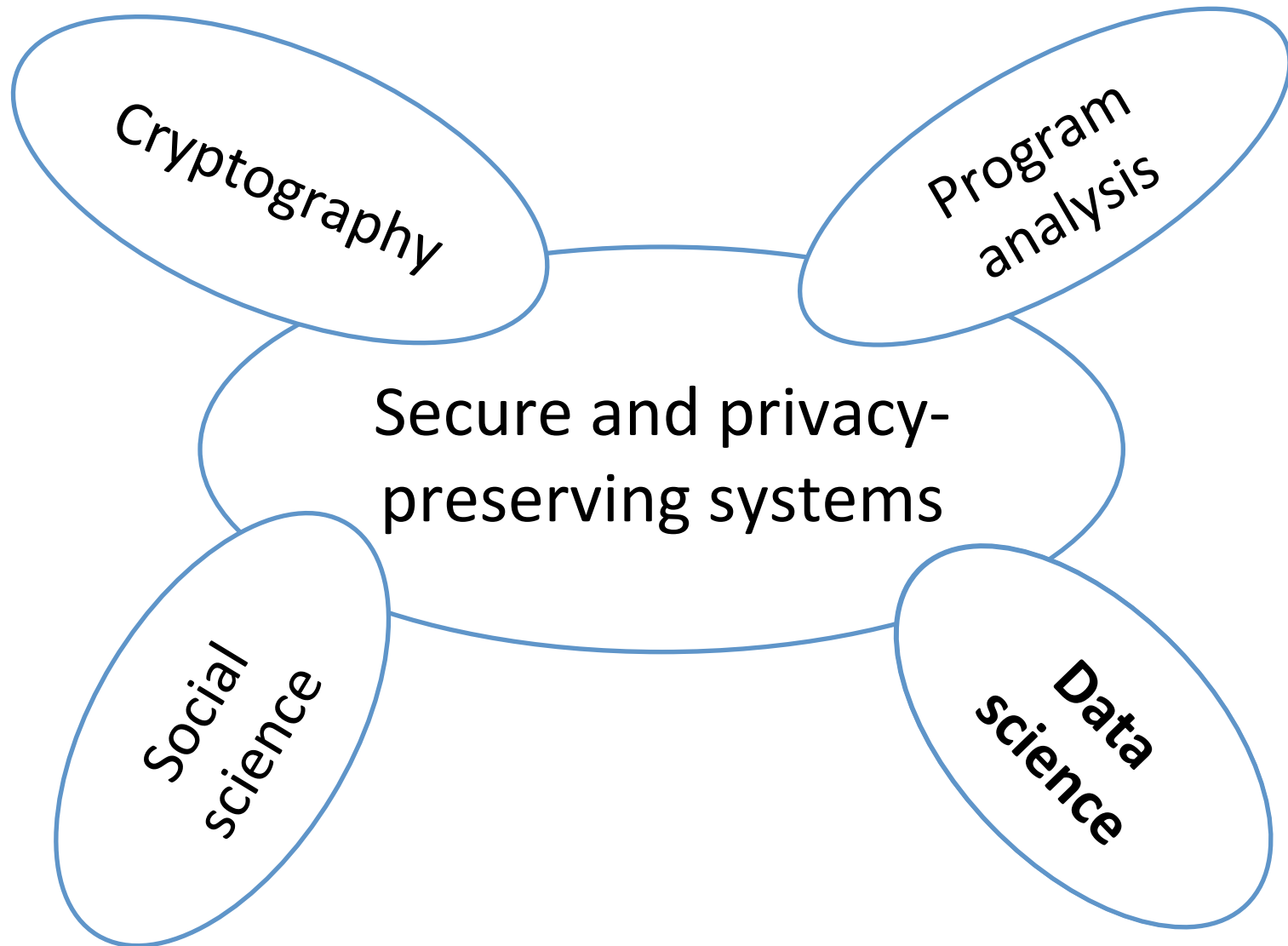


Towards Secure and Privacy-Preserving Social Web Services

Neil Gong
ECpE, Iowa State University
October 24, 2016

Overview of Our Research



Overview of Our Research

- Big data for security and privacy
 - Secure and privacy-preserving online social networks
 - Secure and usable authentication

Overview of Our Research

- Trustworthy machine learning/data mining

Towards Secure and Privacy-Preserving Social Web Services

What are Social Web Services?



Security: Fake Account Detection



1 in 10 Twitter accounts is fake, say researchers

BY KEITH WAGSTAFF | First published November 25th 2013, 4:35 pm



Yelp deems 20% of user reviews 'suspicious'

Published: Sept 27, 2013 8:36 a.m. ET

Privacy Issues

- Private information
 - User identity
 - Demographics
 - Interests
- Protecting user privacy--current paradigm
 - Privacy settings
 - Users not disclose
- How about machine learning techniques?

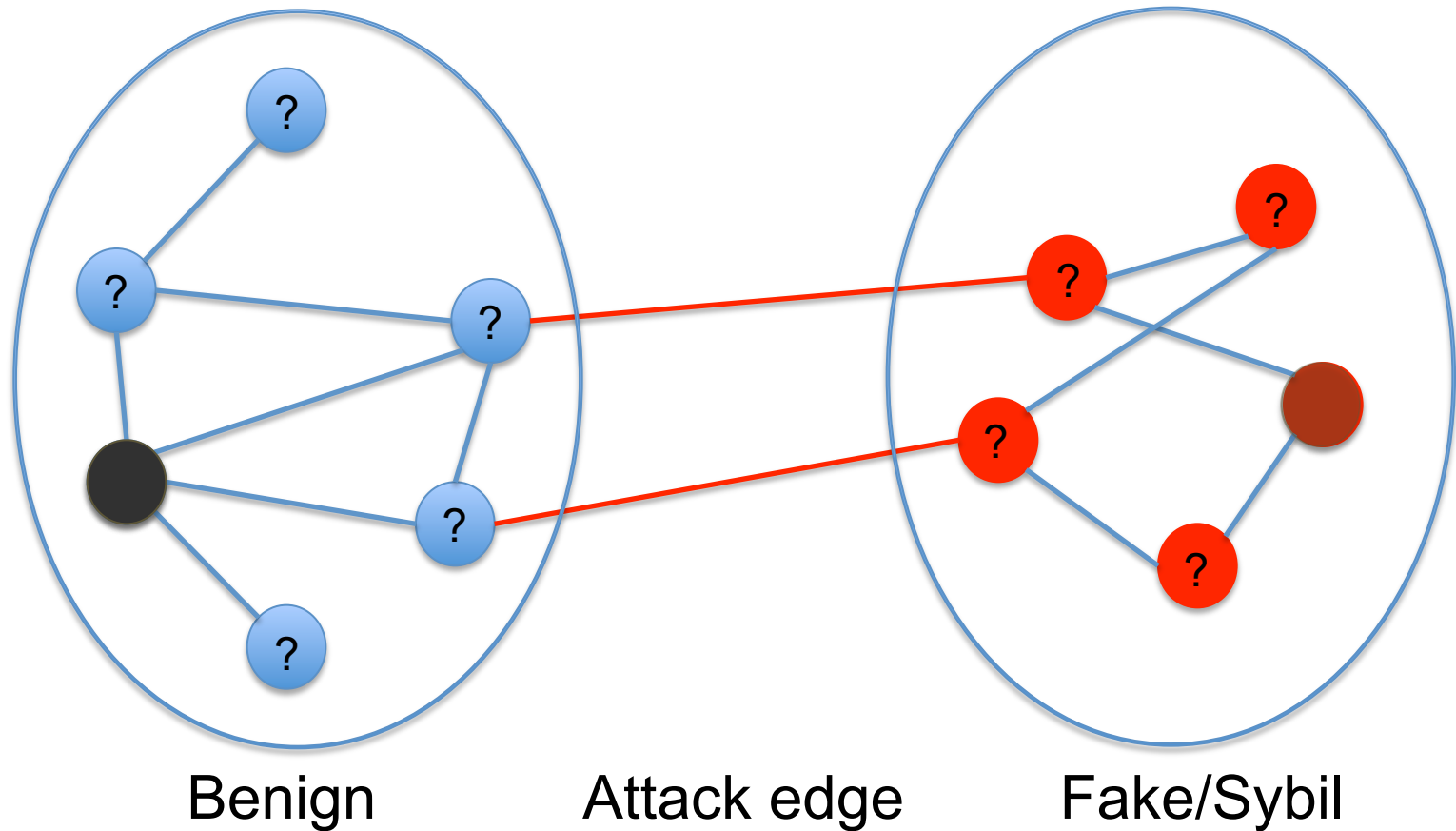
Outline

- I. Fake account detection via probabilistic graphical model techniques
- II. Private information inference: machine learning as new privacy attacks

Risks Brought by Fake Accounts

- Disrupting presidential election
- Influencing financial market
- Subvert personal security and privacy
 - Distribute malware or spam
 - Carry out phishing attacks
 - Steal users' private information
- Manipulate data analytics
 - Manipulate Google search via fake “+1” clicks

Social Structure based Detection



Existing Approaches

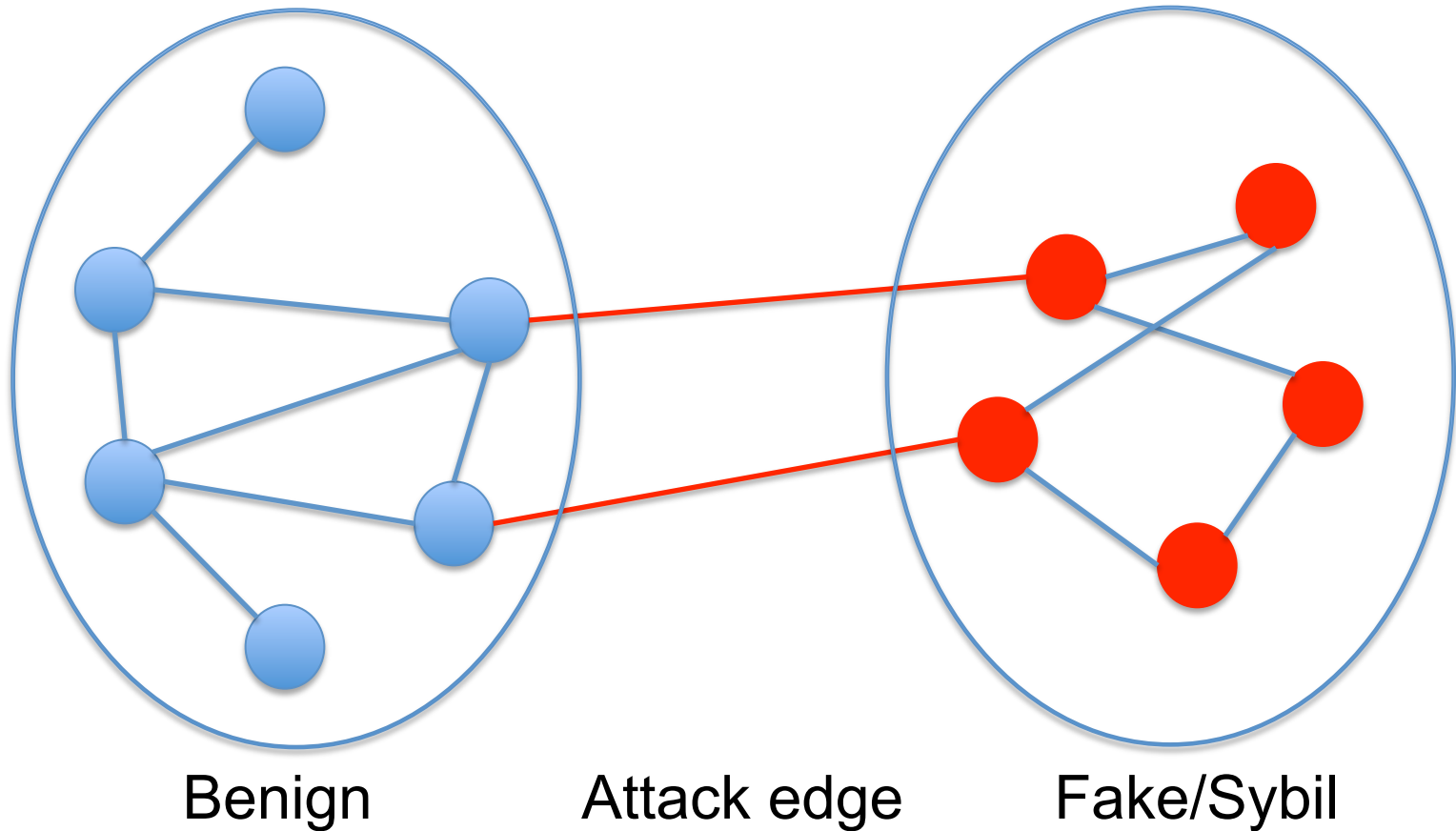
- Mathematical foundation
 - Random walks
 - Community detection
- One-class classification
 - Either labeled benign or labeled fake accounts in the training dataset

Our Approach

- SybilBelief: A scalable semi-supervised learning framework
 - Leverage both labeled benign and labeled fake accounts in the training dataset
- Mathematical foundation
 - Pairwise Markov Random Fields
 - Loopy Belief Propagation

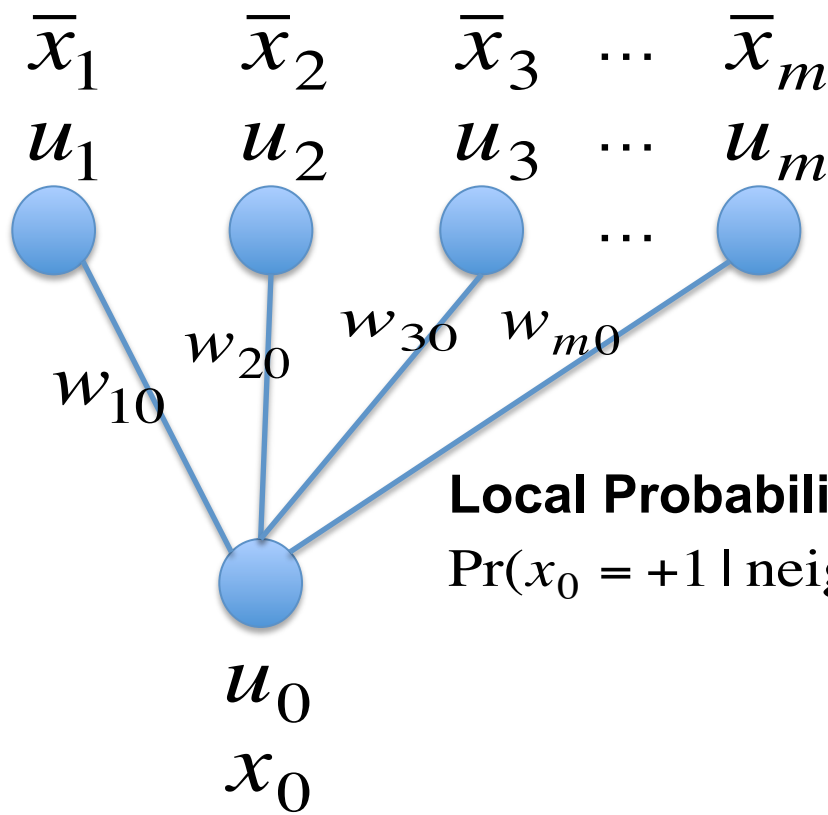
Key Observation: Homophily

Two connected accounts tend to have the same label



Modeling Homophily for One Account

binary random variable $x_i \in \{+1, -1\}$, +1 is benign and -1 is fake



$h_i > 0$: biased to be benign

$h_i = 0$: no bias

$h_i < 0$: biased to be fake

Prior knowledge about u_0

Local Probabilistic Rule:

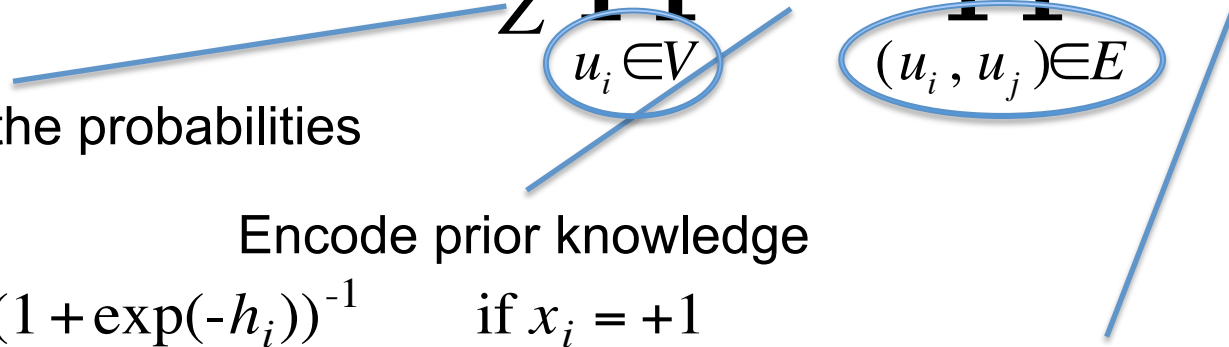
$$\Pr(x_0 = +1 \mid \text{neighbors' labels}) = \frac{1}{1 + \exp(-\sum w_{i0} \bar{x}_i - h_0)}$$

Homophily, $w_{ij} > 0$

Generalizing to the Entire Social Structure

Given $G = (V, E)$

Pairwise Markov Random Fields:

$$\Pr(x_0, x_1, \dots, x_{n-1}) = \frac{1}{Z} \prod_{u_i \in V} \phi(x_i) \prod_{(u_i, u_j) \in E} \varphi(x_i, x_j)$$


Normalize the probabilities

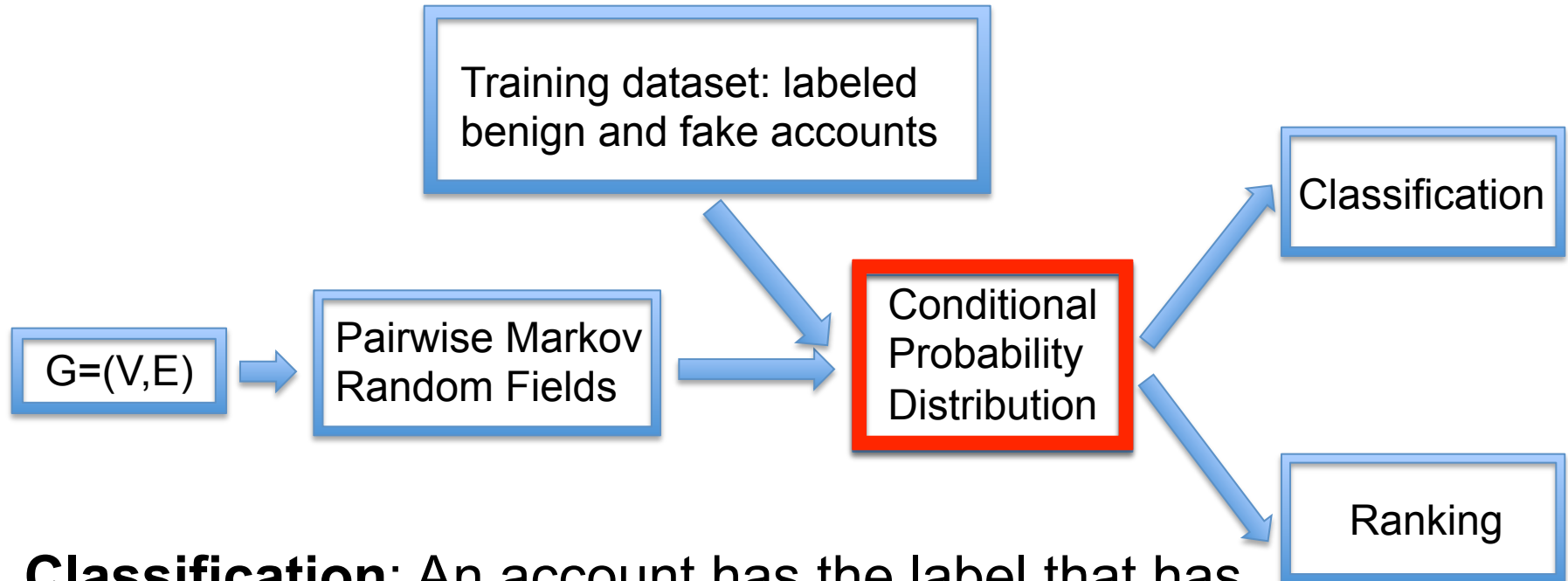
Encode prior knowledge

$$\phi(x_i) = \begin{cases} (1 + \exp(-h_i))^{-1} & \text{if } x_i = +1 \\ 1 - (1 + \exp(-h_i))^{-1} & \text{if } x_i = -1 \end{cases}$$

Encode homophily

$$\varphi(x_i, x_j) = \begin{cases} (1 + \exp(-w_{ij}))^{-1} & \text{if } x_i x_j = +1 \\ 1 - (1 + \exp(-w_{ij}))^{-1} & \text{if } x_i x_j = -1 \end{cases}$$

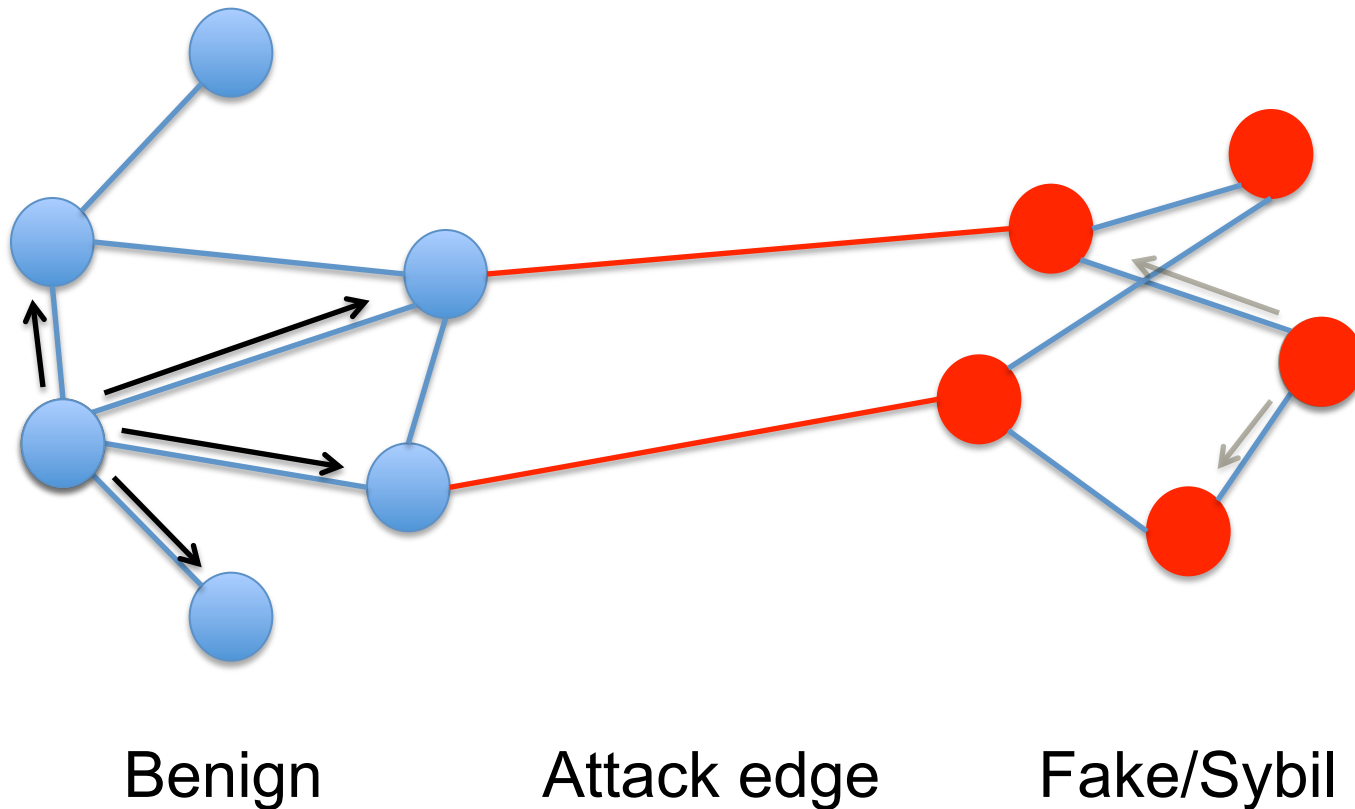
Detecting Fake Accounts



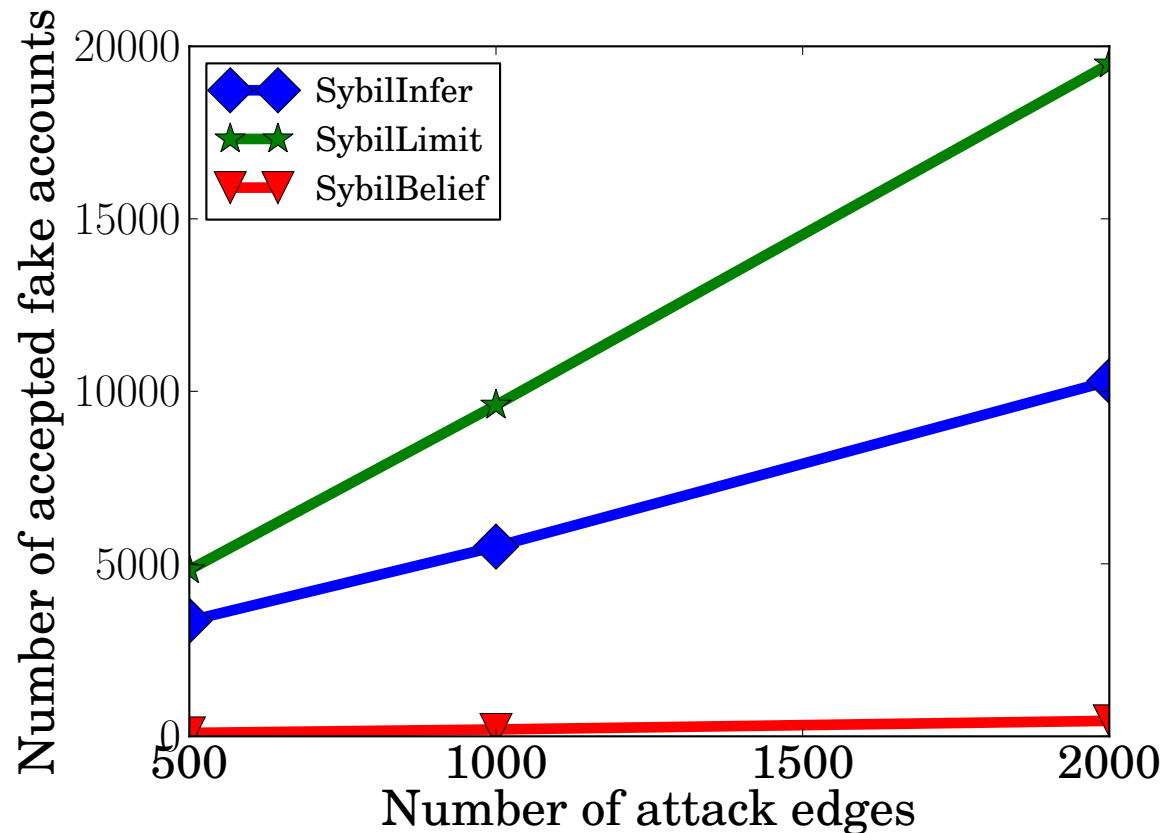
Classification: An account has the label that has the higher conditional probability

Ranking: Ranking accounts using their conditional probability of being benign

Inferring Conditional Probability via Loopy Belief Propagation



Comparison with Classification Methods

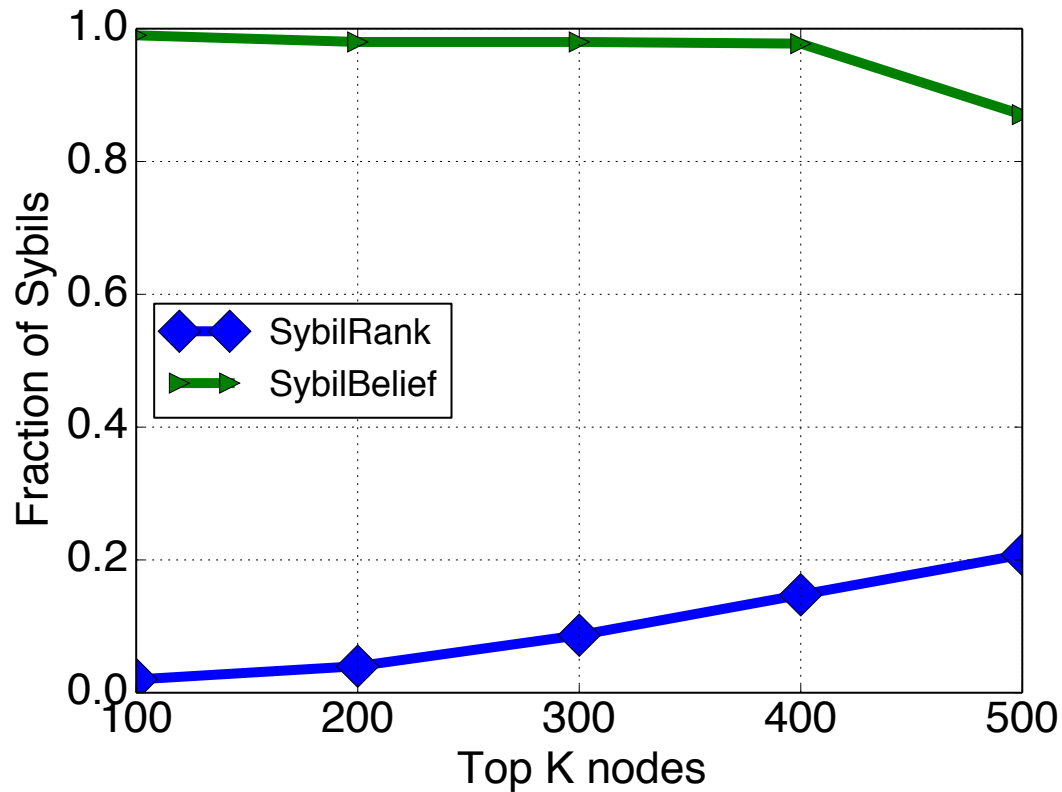


SybilBelief performs orders of magnitude better than previous methods

Comparison with Ranking Methods

- Twitter dataset
 - 10K benign accounts
 - 1K fake accounts (spammers)

Ranking Results on Twitter



SybilBelief detects significantly more fake accounts than SybilRank

Outline

- I. Fake account detection via probabilistic graphical model techniques
- II. Private information inference: machine learning as new privacy attacks

N. Z. Gong, B. Liu. “You are Who You Know and How You Behave: Attribute Inference Attacks via Users' Social Friends and Behaviors”. In Usenix Security Symposium, 2016

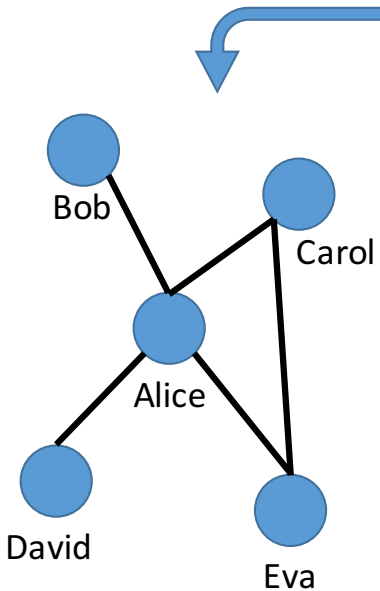
Mixture of public and private information

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

Attribute Inference Attacks

- Given public information of some users
 - Friends
 - Behaviors
 - Attributes
- Infer private attributes of some target users

An Example



User	Alice	Bob	Carol	David	Eva	Page A	Page B	Page C	Page D	Sexual Orientation
Alice		✓	✓	✓	✓	✓	✓		✓	-
Bob	✓						✓	✓		heterosexual
Carol	✓				✓	✓	✓		✓	homosexual
David	✓					✓		✓		bisexual
Eva	✓		✓			✓		✓	✓	homosexual

Friend lists

Behaviors (Page likes)

Attributes

Public data

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

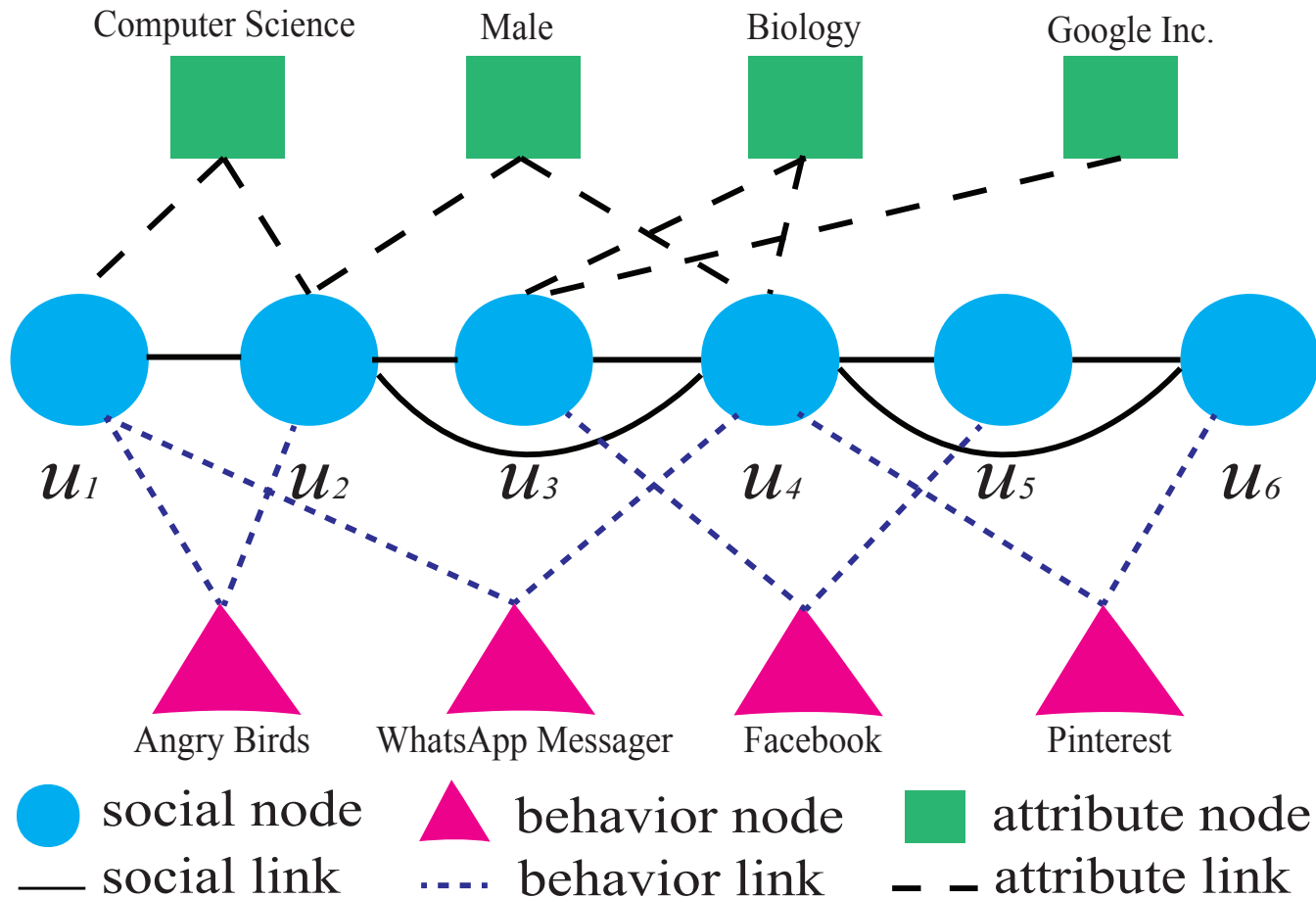
Threat Model

- Implication/Application of attribute inference attacks
 - Privacy threat
 - Targeted phishing attacks
 - Breaking “security question” based user authentication
 - Targeted advertisement
- 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



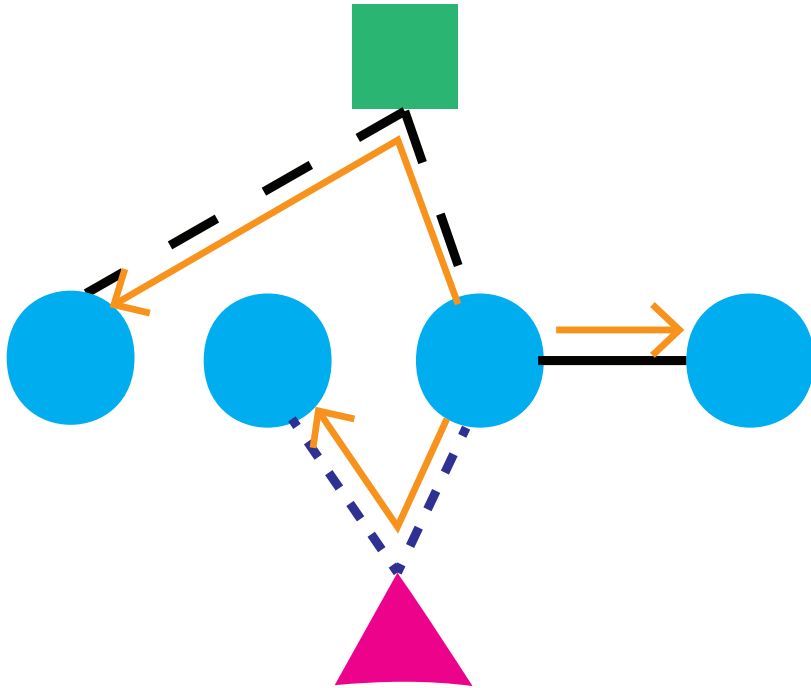
Vote Distribution Attack (VIAL) Algorithm

- 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

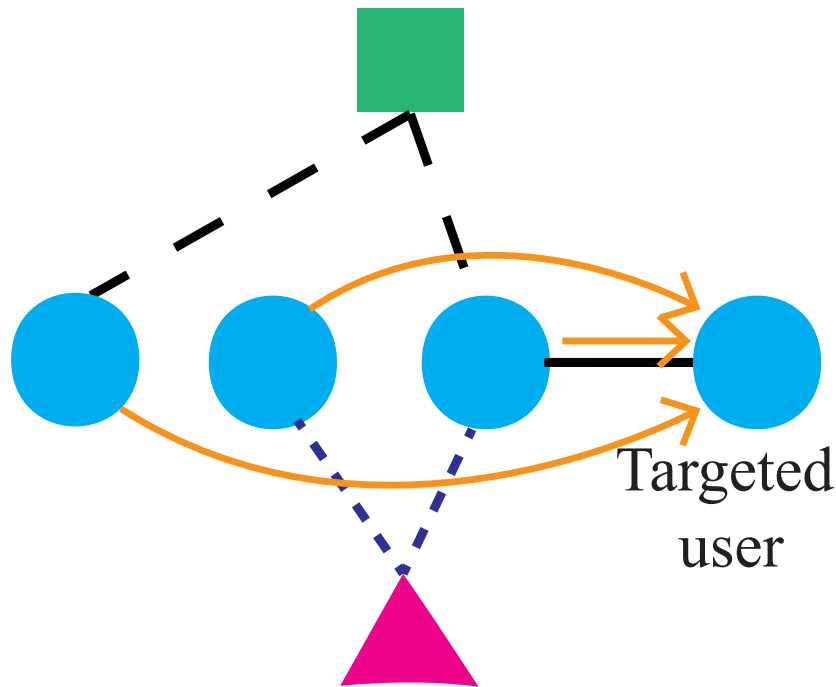
- A user receives a high vote capacity if the user and the targeted user are structurally similar
- Distribution via three local rules
 - Dividing
 - Backtracking
 - Aggregating

Local Rule I: Dividing



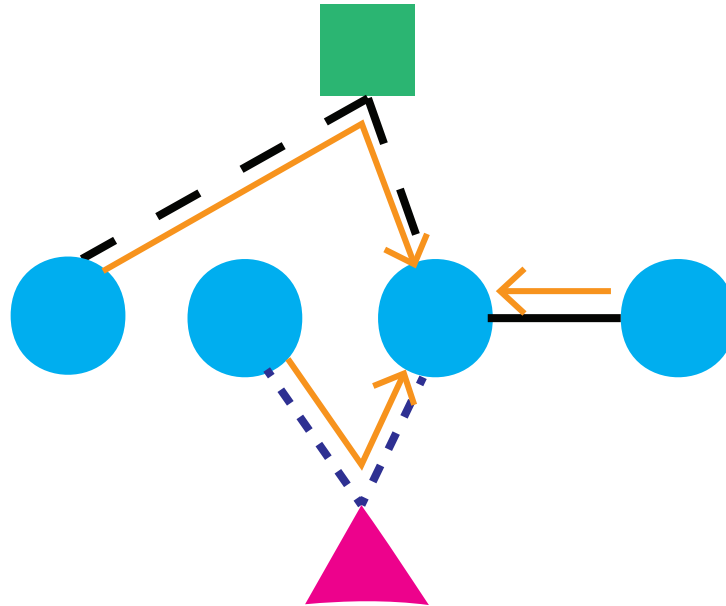
- Social neighbors
- Behavior-sharing social neighbors
- Attribute-sharing social neighbors

Local Rule II: Backtracking



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

Local Rule III: Aggregating

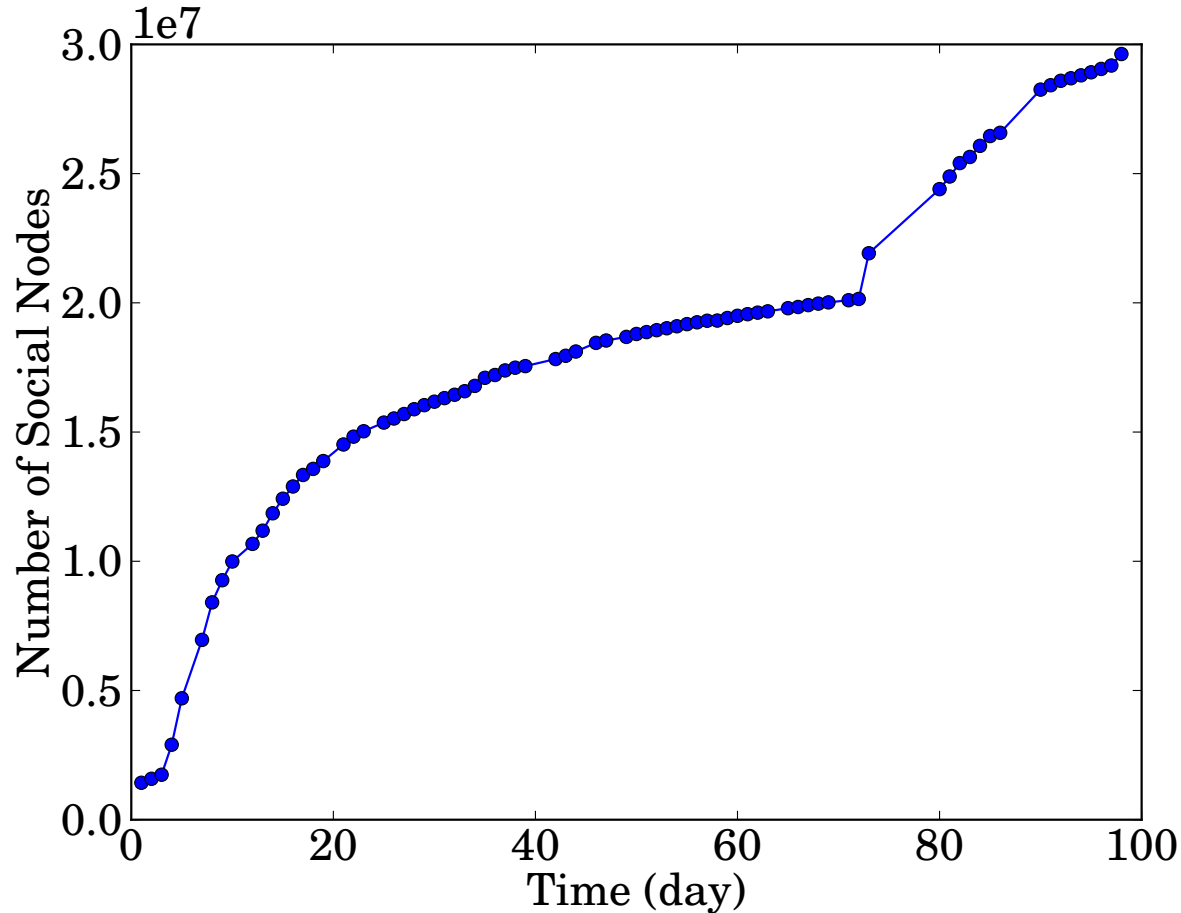


Compute a new vote capacity for *a user* by aggregating the vote capacities from its neighbors

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 - Google+



Social graph
User attributes
Publicly available
Downloaded by
~200 research
groups

Gong et al. "Evolution of Social-Attribute Networks: Measurements, Modeling, and Implications using Google+". In IMC'12.

Evaluation Data - Google Play

- 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

- Sample a set of users uniformly at random
- Remove their attributes as groundtruth
- Treat them as targeted users
- Predict top-K attributes for each targeted user
- Measure Precision, Recall, and F-Score

Comparing with Friend-based and Behavior-based Attacks



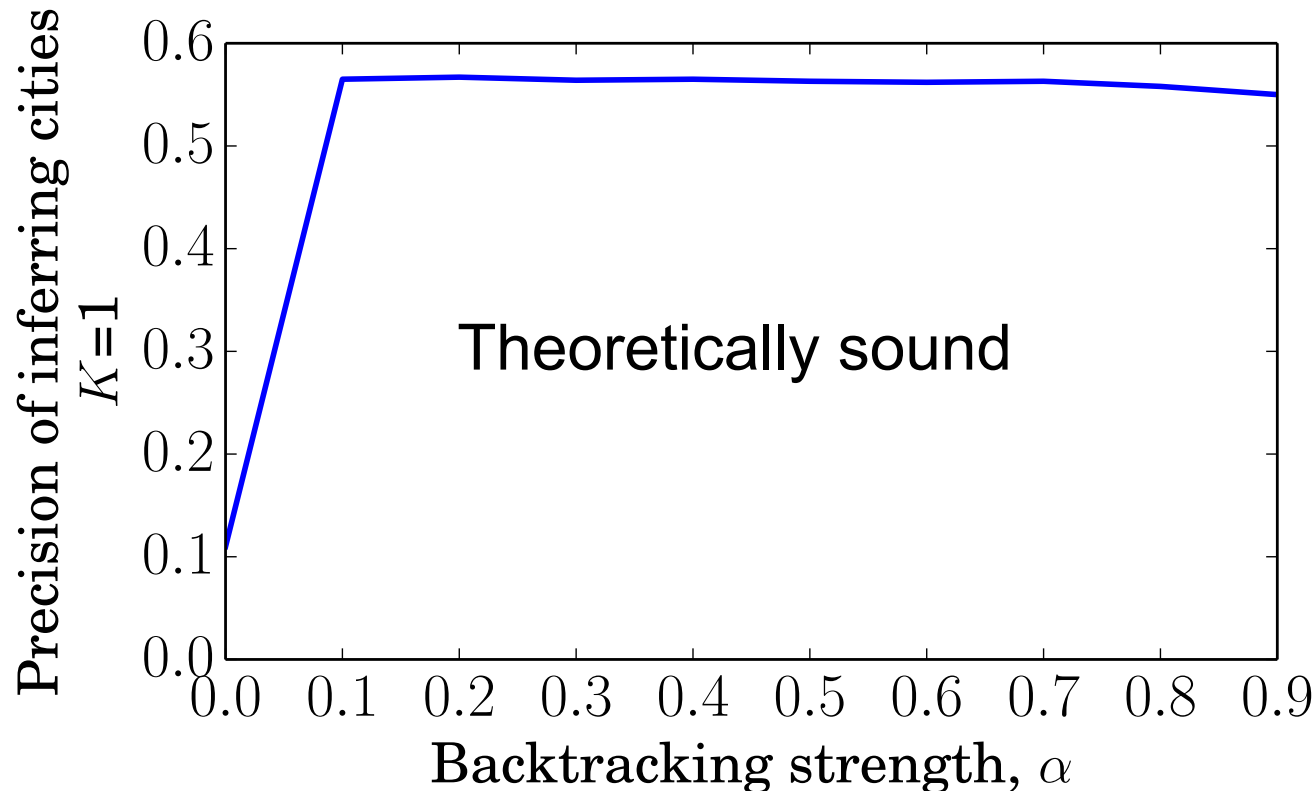
Attack	ΔP	$\Delta P\%$	ΔR	$\Delta R\%$	ΔF	$\Delta F\%$
Random	0.36	526%	0.22	535%	0.27	534%
RWwR-SAN	0.07	20%	0.05	23%	0.06	22%
VIAL-B	0.22	102%	0.13	99%	0.16	100%

Best behavior-based attack

Best friend-based attack

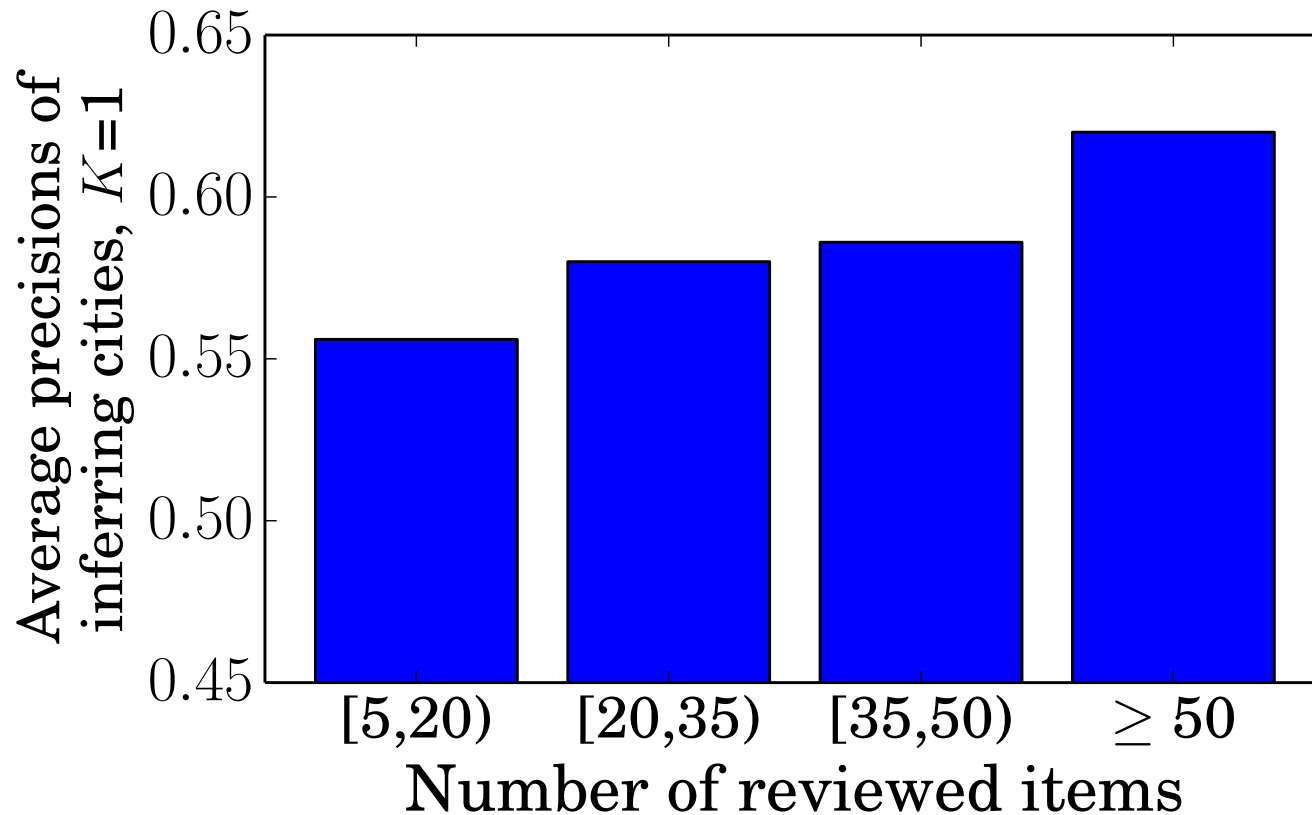
Our attacks are significant more accurate than existing ones

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

Other Inference Attacks

- Inferring author identity using writing styles [IEEE S & P 2012]
- De-anonymizing social networks [NDSS2015]
- Inferring user interests [WSDM2015]

Summary

- Private information can be inferred from public data via machine learning techniques
- Fundamental reason: private information is correlated with public information
- How to defend against inference attacks?