Semi-Supervised Learning and Social Networks

- Semi-supervised learning algorithm combined with pseudo–random walk
    - Used to fill in missing attribute data for a node in a social network graph
- Nodes in social networks (LinkedIn, Google+, Facebook) assigned to people
- A lot of missing data associated with these nodes

## Profile Inference: Facebook Example

Profile

Hometown:	Palo Alto
High school:	Gunn
College:	Stanford
Employer:	Facebook
Current city:	Sunnyvale
Hobbies:	Politics, Music

Benefits of a Complete Profile

- People more easily searchable
- Can tailor news recommendations
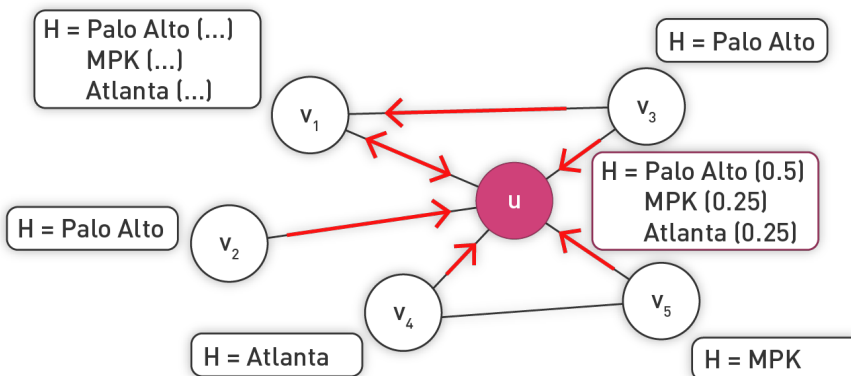- Can group recommendations
- Better ad targeting

How can we fill in missing profile fields?

Profile Inference
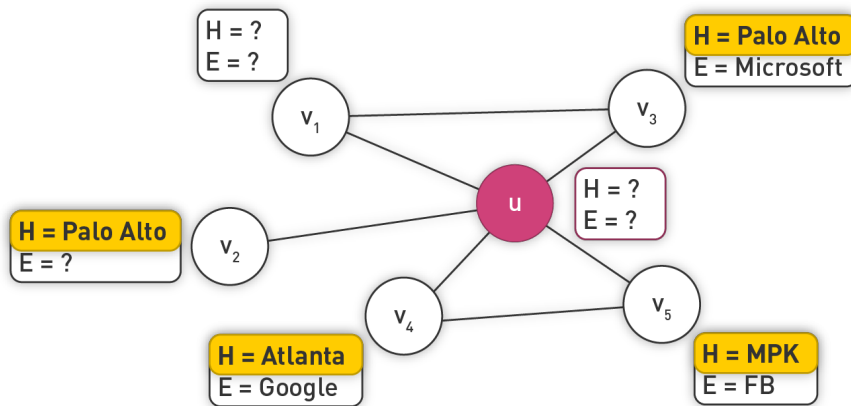
- Use social network and assumption of homophily.
  - Friendships form between similar people.
  - Infer missing labels to maximize similarity.

Previous Work on Profile Inference
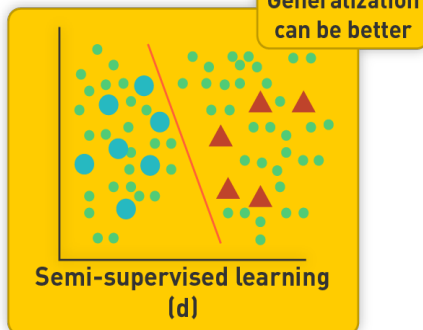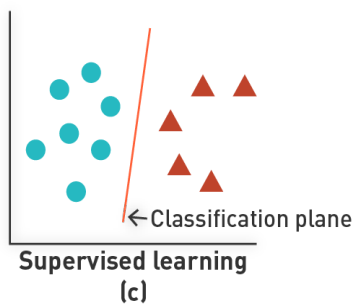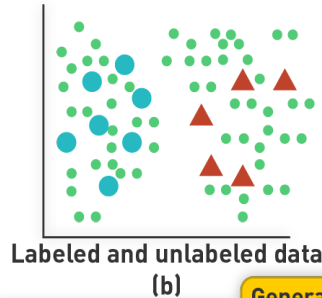

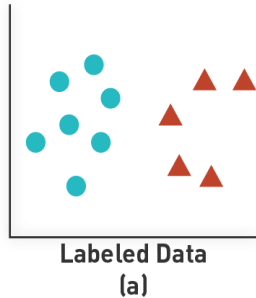
Label propagation (Zhu+/02, Macskassy+/07)

- "Propagate" labels through the network
- Probability(I have hometown H) = Fraction of my friends with hometown H
- Iterate until convergence

Label propagation in graphs falls into the category of semi-supervised learning.

Semi-Supervised Learning

- A class of supervised learning tasks and techniques that makes use of both labeled and unlabeled data for training.
    - Typically a small amount of labeled data with a large amount of unlabeled data.

- May refer to transductive learning (transductive SVMs).

- Is an attractive idea because of high cost or impossibility of labeling.

Semi-Supervised Learning

Labeled Data
(a)

Labeled and unlabeled data
(b)

Generalization can be better

←Classification plane

Supervised learning
(c)

Semi-supervised learning
(d)

**Semi-supervised learning falls between
unsupervised and supervised learning.**