Einführung in Web- und Data-Science

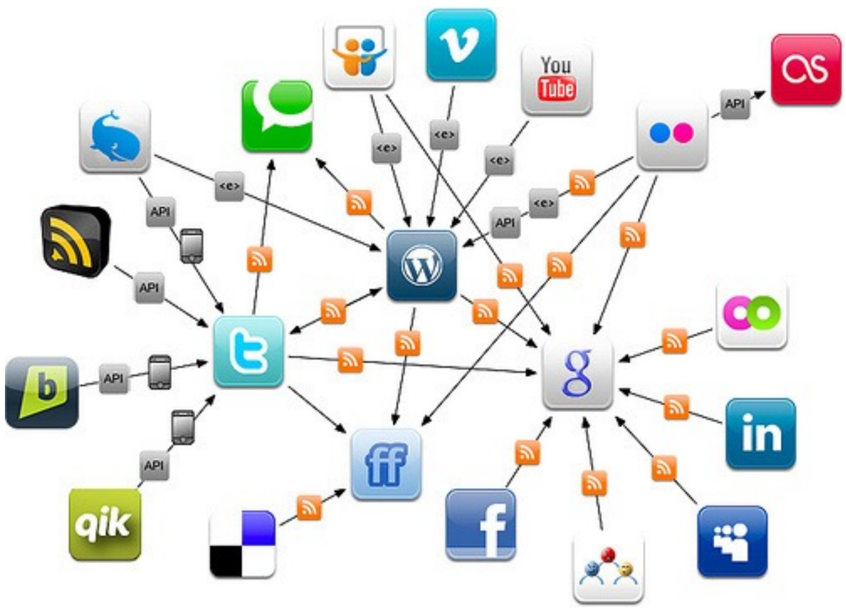
Link Prediction

Dr. Marcel Gehrke
Universität zu Lübeck
Institut für Informationssysteme

Acknowledgment

Hong Kong University of Science and Technology
Advanced Data Mining
COMP 4332 / RMBI 4310

Computer Science and Engineering
IIT Kharagpur
Link Prediction in Social Networks
Pabitra Mitra



University of Southern California
CS 599: Social Media Analysis
Social Ties and Link Prediction
Kristina Lerman

A Theoretical Justification of Link Prediction Heuristics
Deepayan Chakrabarti, Purnamrita Sarkar, Andrew Moore

Stanford University
Graph Representation Learning
Jure Leskovec

Applications of Link Prediction on Graphs

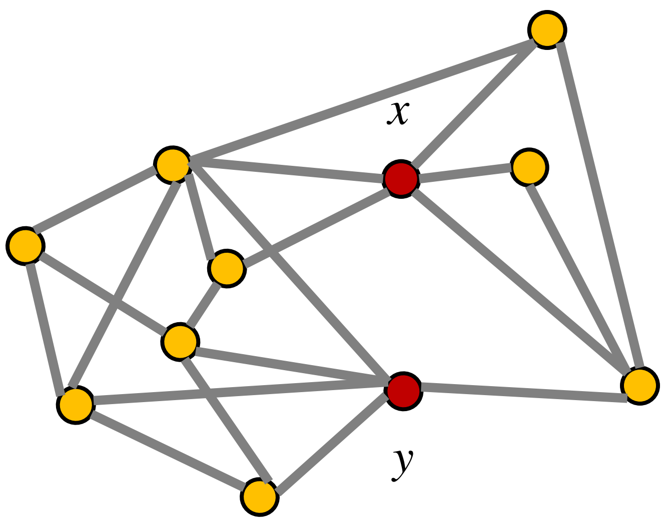
- Who are/will become friends?
- Who will collaborate in drug racketeering?
- Which products to recommend to which persons?
- Are there unknown commonalities between species?
- Where will new protein interactions show up?

Informal Definitions

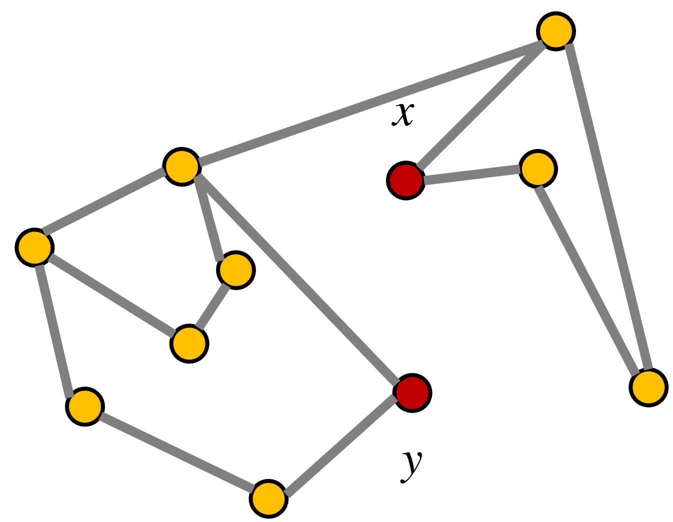
- Link Prediction Problem
  - Given a snapshot of a network, can we infer which new interactions among its nodes are likely to occur in the near future?
- Link Completion Problem
  - If the network is known to be incomplete, can we infer which interactions are possibly missing (and should be added)?
  - Then, solve link prediction problem on completed data
- Both problems to be formalized based on "proximity" of nodes in a network

The Intuition

- In many networks, people who are “close” belong to the same social circles and will inevitably encounter one another and become linked themselves.
- Link prediction heuristics measure how “close” people are



Red nodes are close to each other

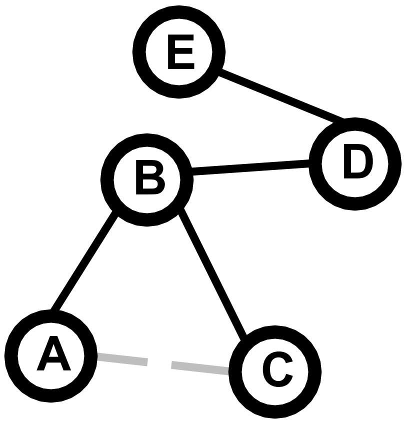


Red nodes are more distant

Challenges

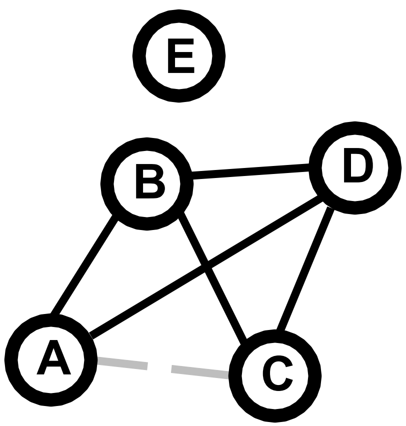
- Data is usually sparse
  - Missing data/relationships
- Imbalance
  - So many possibilities, so few choices
  - Ill-posed problem
  - Low accuracy in practice
- Accuracy vs. scalability
  - Modeling (unobserved/unknown factors)
  - Tasks of approximation/optimization

Graph distance & Common Neighbors



- Graph distance: (Negated) length of shortest path between x and y

(A, C)	-2
(C, D)	-2
(A, E)	-3



- Common Neighbors: A and C have 2 common neighbors, more likely to collaborate

$$score(x, y) := |\Gamma(x) \cap \Gamma(y)|$$

where $\Gamma(x)$ denotes the neighbors of x

Preferential Attachment

- Preferential Attachment: Probability that a new collaboration involves x is proportional to $\Gamma(x)$, the current neighbors of x

- score (x, y) := $|\Gamma(x)| \cdot |\Gamma(y)|$

Hitting time, PageRank

- Hitting time: expected number of steps for a random walk starting at x to reach y
- Commute time: $-(H_{x,y} + H_{y,x})$

- If y has a large stationary probability, $H_{x,y}$ is small. To counterbalance, we can normalize

$$\text{score}(x, y) := -(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$$

- Rooted PageRank: to cut down on long random walks, walk can return to x with a probablity α at every step y

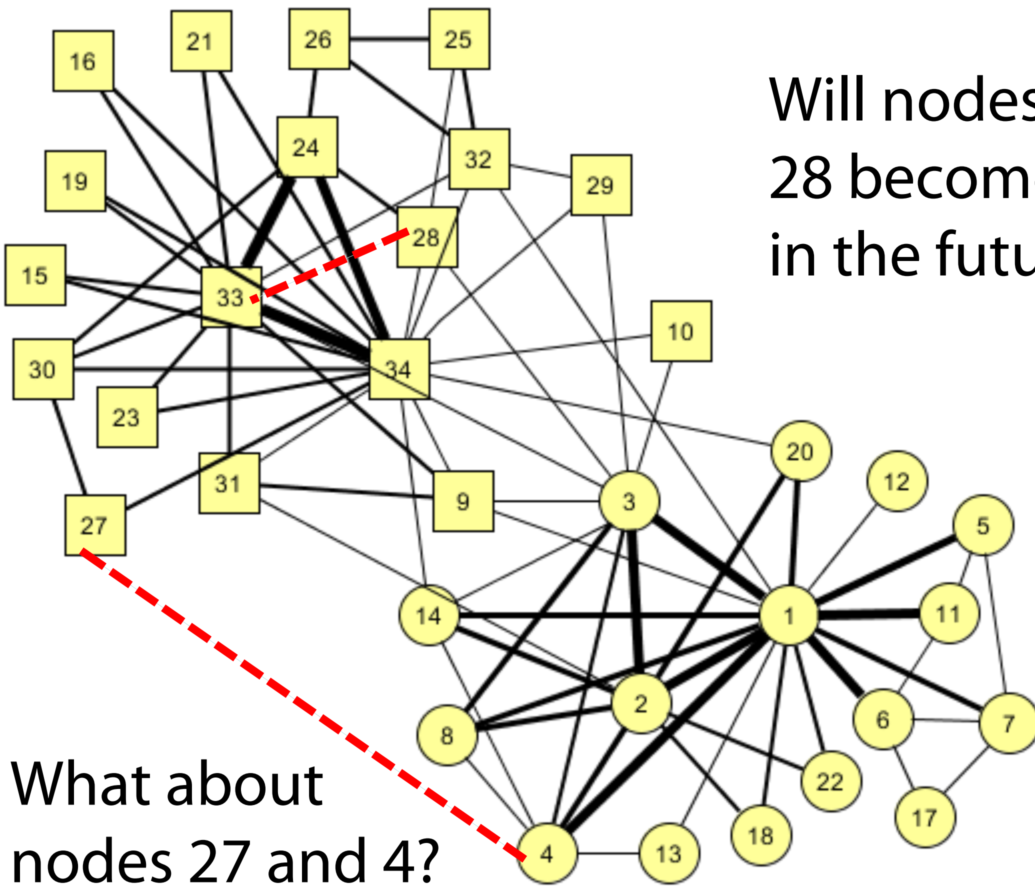
SimRank

Defined by this recursive definition: two nodes are similar to the extent that they are joined by similar neighbors

$$\text{similarity}(x, y) := \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{similarity}(a, b)}{|\Gamma(x)| \cdot |\Gamma(y)|}$$

$$\text{score}(x, y) := \text{similarity}(x, y)$$

Link Prediction

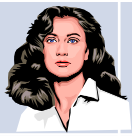
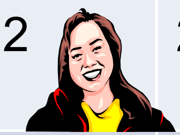


Will nodes 33 and 28 become friends in the future?

Does network structure contain enough information to predict what new links will form in the future?

What about nodes 27 and 4?

Link Prediction using Collaborative Filtering

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

Link Prediction using Collaborative Filtering

- Memory-based Approach
  - User-based approach [Twitter]
  - Item-based approach [Amazon & Youtube]
- Model-based Approach
  - Latent Factor Model [Google News]

- Hybrid Approach

Memory-based Approach

- Few modeling assumptions
- Few tuning parameters to learn
- Easy to explain to users
  - Dear Amazon.com Customer, We've noticed that customers who have purchased or rated *How Does the Show Go On: An Introduction to the Theater* by Thomas Schumacher have also purchased *Princess Protection Program #1: A Royal Makeover (Disney Early Readers)*.

Algorithms: User-Based Algorithms (Breese et al, UAI98)

- $v_{i,j}$ = vote of user *i* on item *j*
- I_i = items for which user *i* has voted
- Mean vote for *i* is

$$\overline{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

- Predicted vote for "active user" *a* is weighted sum

$$p_{a,j} = \overline{v}_a + \kappa \sum_{i=1}^{n} \underbrace{w(a,i)}_{\text{weights of } n \text{ similar users}} (v_{i,j} - \overline{v}_i)$$

normalizer

weights of *n* similar users

Algorithms: User-Based Algorithms (Breese et al, UAI98)

- K-nearest neighbor

$$w(a,i) = \begin{cases} 1 & \text{if } i \in \text{neighbors}(a) \\ 0 & \text{else} \end{cases}$$

- Pearson correlation coefficient (Resnick ' 94, Grouplens):

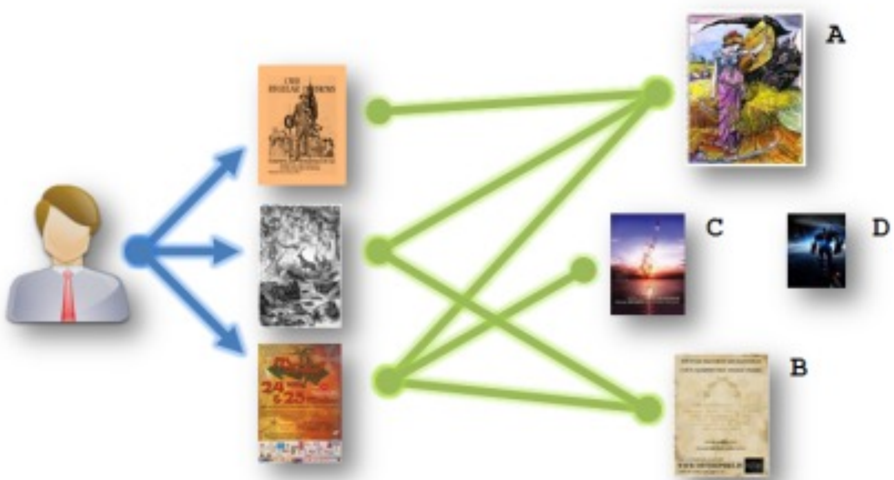
$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v}_a)(v_{i,j} - \overline{v}_i)}{\sqrt{\sum_j (v_{a,j} - \overline{v}_a)^2 \sum_j (v_{i,j} - \overline{v}_i)^2}}$$

- Cosine distance (from IR)

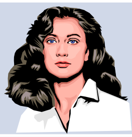
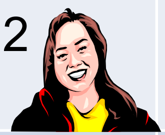
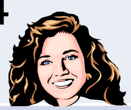
$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}}$$

Algorithm: Amazon’s Method

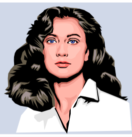
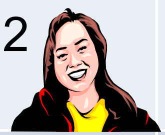
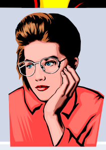
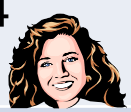
- Item-based Approach
  - Similar with user-based approach but is on the item side



Item-based CF Example: infer (user 1, item 3)

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

How to Calculate Similarity (Item 3 and Item 5)?

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

Similarity between Items

Item 3	Item 4	Item 5
?	2	7
5	7	5
7	4	7
7	3	8
4	6	?
8	3	7

- How similar are items 3 and 5?
  - How to calculate their similarity?

Similarity between items

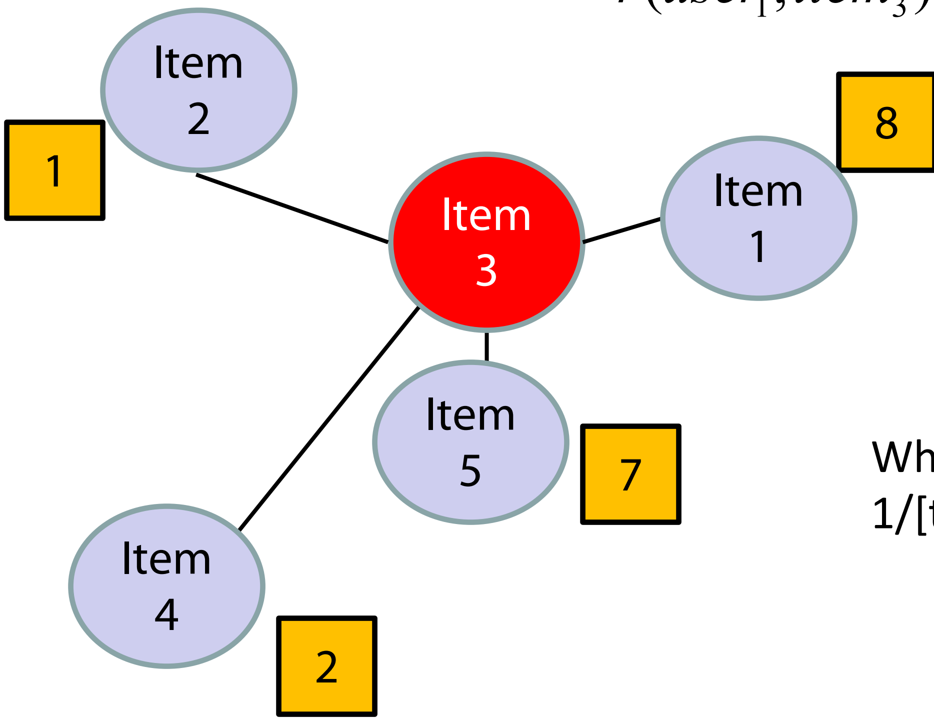
Item 3	Item 5
?	7
5	5
7	7
7	8
4	?
8	7

- Only consider users who have rated both items
- For each user: Calculate difference in ratings for the two items
- Take the average of this difference over the users

sim(item 3, item 5) = cosine((5, 7, 7, 8), (5, 7, 8, 7))
= (5*5 + 7*7 + 7*8 + 8*7)/
(sqrt($5^2+7^2+7^2+8^2$)* sqrt($5^2+7^2+8^2+7^2$))

- Can also use Pearson Correlation Coefficients as in user-based approaches

Prediction: Calculating ranking r(user1,item3)



$$r(user_1, item_3) = \alpha * \{r(user_1, item_1) sim(item_1, item_3) + r(user_1, item_2) sim(item_2, item_3) + r(user_1, item_4) sim(item_4, item_3) + r(user_1, item_5) sim(item_5, item_3)\}$$

Where α is a normalization factor, which is 1/[the sum of all sim($item_i$,$item_3$)].

Algorithm: Youtube's Method

- Youtube also adopt item-based approach

- Adding more useful features
  - Num. of views
  - Num. of likes
  - etc.

Link Prediction: Summary

- Link prediction is the underlying problem in many applications
- No methods fits all purposes
- Need to carefully evaluate a method in a practical setting
- Methods are hard to analyze theoretically, but see

Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W. Moore.
Theoretical justification of popular link prediction heuristics.
In: Proc. IJCAI-11. pp. 2722–2727. **2011**.