TSINGHUA UNIVERSITY

Recommender System and Link Prediction

Present by *Feng Xie*

September 18, 2011

Content

- Recommender System (RS)
- Link Prediction (LP)
- Group Discovery

Recommender Systems

The world is an over-crowded place

They all want to get our attention

Who can help us?

- Can google help?
  - Yes, but only when we really know what we are looking for
- Can experts help?
  - Yes, but it won’t scale well
    - Everyone receives exactly the same advice!

Ok, here is RS

- To recommender to us something we may like
- How?
  - Based on our history of selection
  - Based on other people with similar interests

Example of RS

- GroupLens
- Amazon Recommendation
- Netflix ($ 1million 10%)
- 豆瓣

Some evidences

- Netflix
  - 2/3 rented movies are from recommendation
- Google news
  - 38% more click-through are due to recommendation
- Amazon
  - 35% sales are from recommendation

What do RS do, exactly?

- Predict how much you may like a certain product / service
- Compose a list of N best items for you
- Compose a list of N best users for a certain product / service
- Explain to you why these items are recommended to you
- Adjust the prediction and recommendation based on your feedback and other people

Approaches of RS

- Collaborative filtering
  - User-based
  - Item-based
- Content-based filtering
- Hybrid
  - Linear/Switching combination/Sequential
  - Information Quantity

Collaborative Filtering (1)

- User-based (1994, GroupLens)

	Taken	Titanic	Panda					
Alice	5	4	5		3			4
Lily		3	5			4		5
Jacky		4		5	4			
Bob	5	?	4	5		3	5	
	4				3	3		4
	5	2			3	5		
			1	4	2			
				5			4	3

Collaborative Filtering (2)

- Item-based (2001, Amazon)

	Taken	Titanic	Panda					
Alice	5	4	5		3			4
Lily		3	5			4		5
Jacky		4		5	4			
Bob	5	?	4	5		3	5	
	4				3	3		4
	5	2			3	5		
			1	4	2			
				5			4	3

Content-based (1)

- Web page: words, hyperlinks, images, tags, comments, titles, URL, topic
- Music: genre, rhythm, melody, harmony, lyrics, meta data, artists, bands, press releases, expert reviews, loudness, energy, time, spectrum, duration, frequency, pitch, key, mode, mood, style, tempo
- User: age, sex, job, location, time, income, education, language, family status, hobbies, general interests, Web usage, computer usage, fan club membership, opinion, comments, tags, mobile usage
- Context: time, location, mobility, activity, socializing, emotion

Content-based (2)

- Can we acquire those content pieces automatically?
  - Fairly easy for text
  - Difficult for music and video, except for digital signals

Similarity Measures

- Cosine-based

	Taken	Titanic	Panda					
Alice	5	4	5		3			4
Lily		3	5			4		5
Jacky		4		5	4			
Bob	5	?	4	5		3	5	
	4				3	3		4
	5	2			3	5		
			1	4	2			
				5			4	3

-

- Statistic-based & Orthogonalization

Evaluation

- How do we know the recommendation is good?
- Practice: training / testing split (80/20%)
- Metrics
  - MAE (Mean Absolute Error), RMSE (Root Mean Square Error)
  - Recall, precision

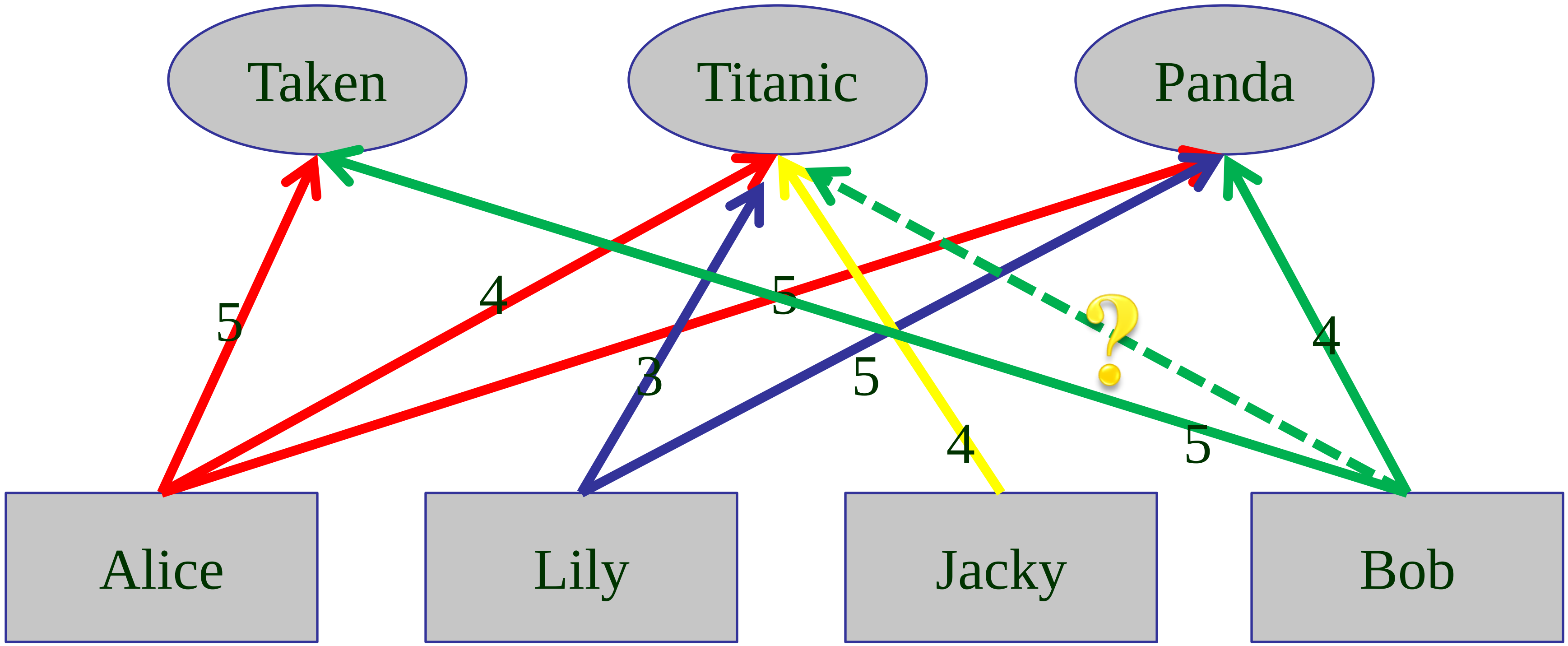
Problems with RS

- Scale
  - Netflix (2007): 5M users, 50K movies, 1.4B ratings
- Sparse data
  - I have rated only one book at Amazon!
- Cold-Start
  - New users and items do not have history
- Popularity bias
  - Everyone reads “Harry Potter”
- Trust

More State-of-the-arts

- Research in Recommender Systems is becoming a *mainstream*, evidenced from the recent conference ACM RecSys.
- Other conferences
  - *KDD, SDM, ICDM, PKDD, WSDM, RecSys*

RS & Graph



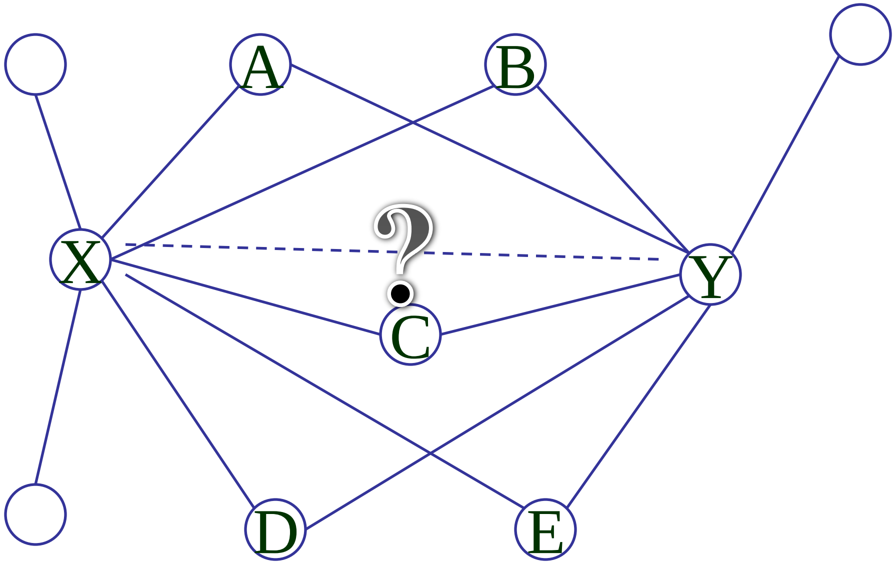
RS can be considered as a sub-problem of LP!

Link Prediction

- Estimating the likelihood of the existence of a link between two nodes, based on the observed topology
- Prediction of *existed yet unknown links* for sampling networks, such as food webs, protein-protein interaction networks and metabolic networks
- Prediction of *future links* for evolving networks, like on-line friendship networks

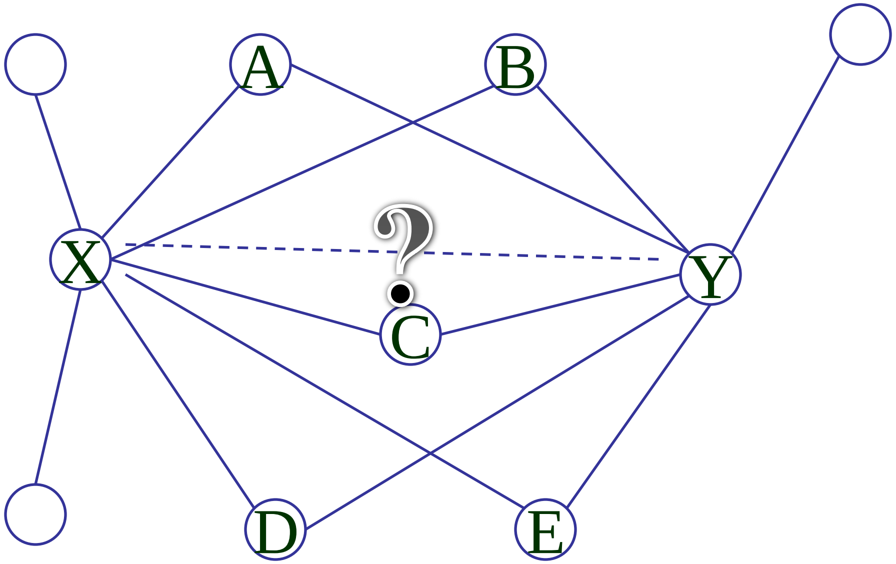
LP algorithms

- Attributes
- Network structure
  - Node-based
  - Path-based



Node-based (1)

- Common neighbor based



$$s_{xy} = \left|\Gamma(x) \cap \Gamma(y)\right|$$

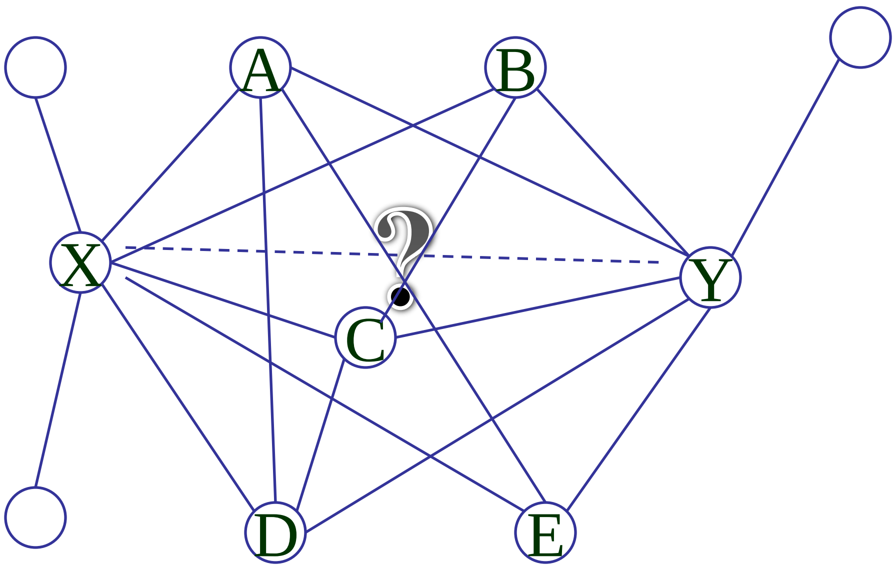
$$s_{xy} = \frac{\left|\Gamma(x) \cap \Gamma(y)\right|}{\left|\Gamma(x) \cup \Gamma(y)\right|}$$

Node-based (2)

- Resource Allocation (RA)

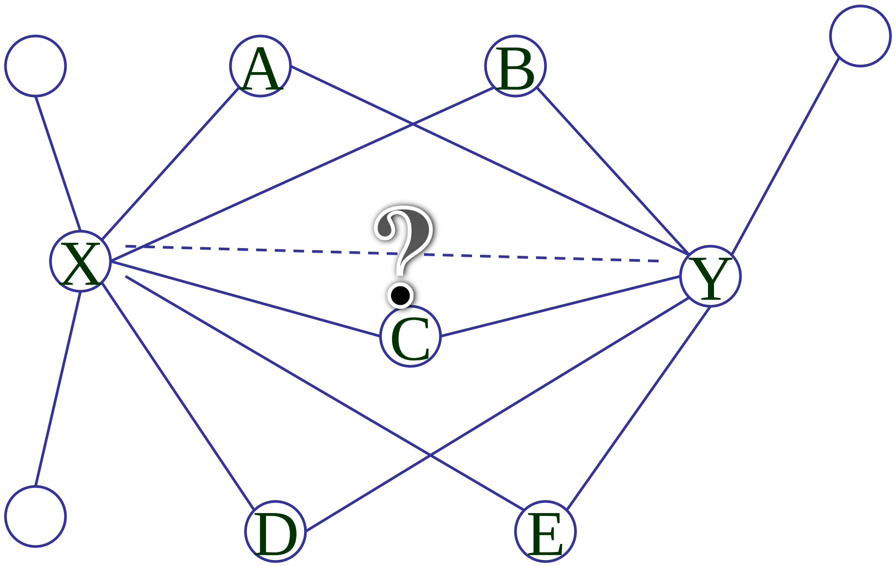
The node x can send some resource to y with their common neighbors playing the role of transmitters. Assume that each transmitter has a unit of resource, and will equally distribute it between all its neighbors. Then S(x,y) is defined as the amount of resource y received from x.

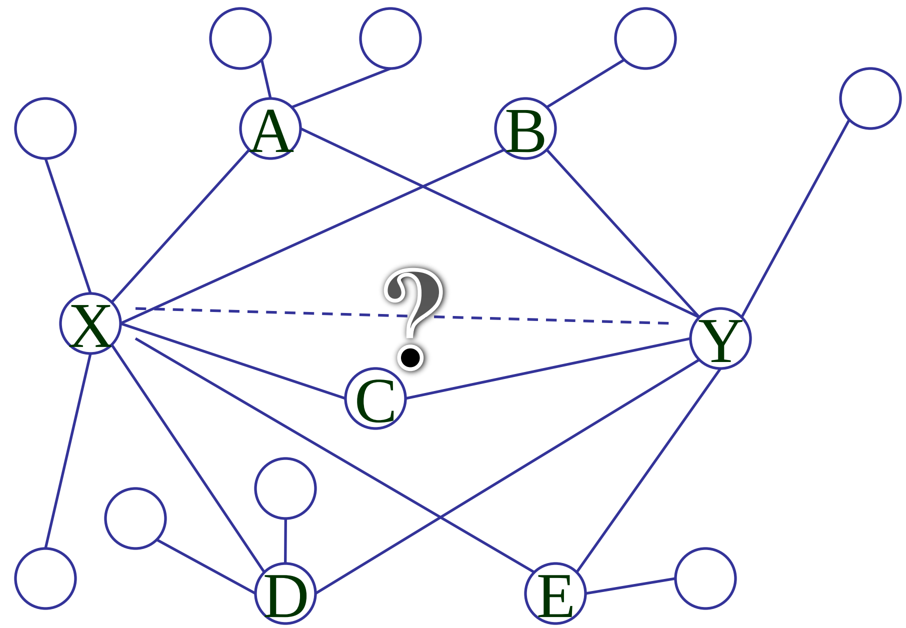
$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}$$

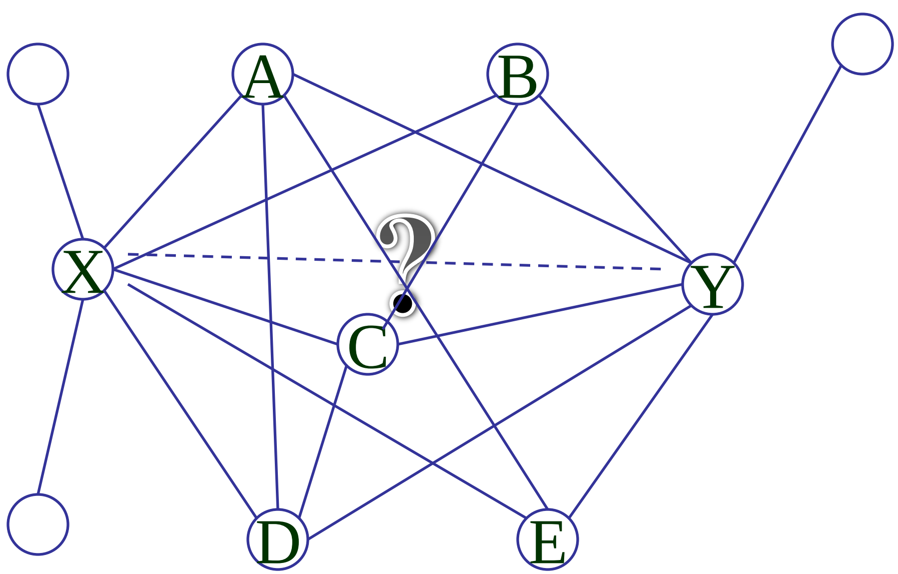


$$s_{xy} = \frac{1}{4} + \frac{1}{3} + \frac{1}{4} + \frac{1}{4} + \frac{1}{3}$$

Observation







Path-based

- Katz Index

$$s_{xy} = \sum_{l=1}^{\infty} \beta^l \cdot \left| paths_{xy}^{\langle l \rangle} \right|$$

- Local Path (LP)

$$S = A^2 + \varepsilon A^3$$

Group Discovery

- Related Paper

  Newman M E J and Girvan M, 2004*Phys. Rev. E* ***69026113***

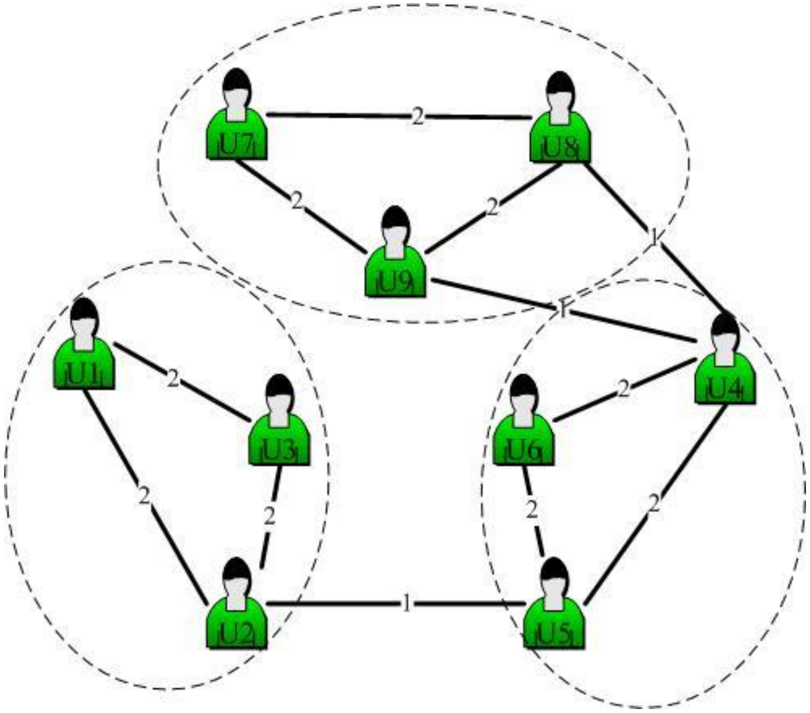
- Modularity

$$\|A\| = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}$$

$$\|A_{pq}\| = \sum_{i \in V_p}\sum_{j \in V_q} a_{ij}$$

$$e_{pq} = \|A_{pq}\| / \|A\|$$

$$Q = \sum_{p=1}^{m}\left[e_{pp} - \left(\sum_{q=1}^{m} e_{pq}\right)^2 \right]$$



Thank you!

Questions?