AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

19CSE337 Social Networking and Security

Lecture 19

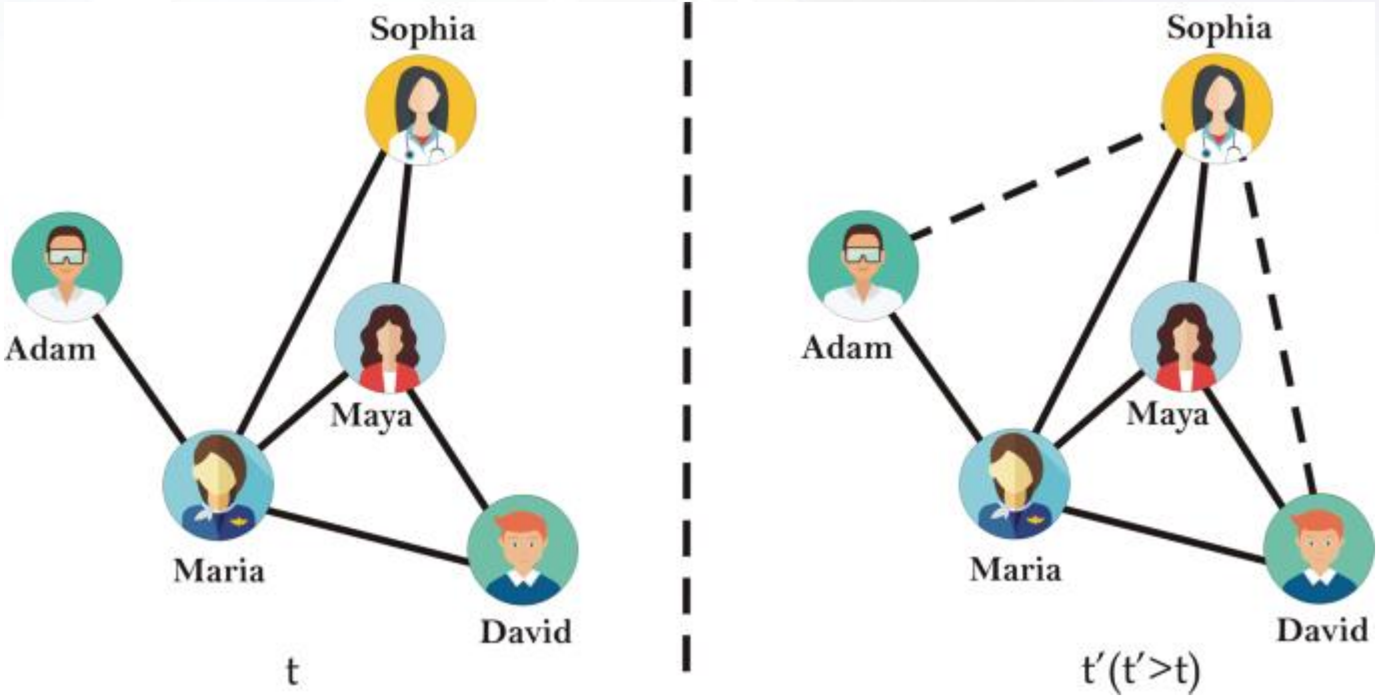
Topics to Discuss

- Link Prediction
- Applications
- Techniques

Link Prediction

- Mining and analyzing social network data is non-trivial but challenging.
- Two main challenges: Incompletion (real world data may include missing data) and dynamic (rapidly evolving which produces new connections or removes existing one).
- Therefore predicting missing or unobserved links in social network is very important.
- This problem is called link prediction.
- The objective of link prediction is to identify pairs of nodes that will either form a link or not in the future.

Link Prediction



Applications

- To design recommender systems in information retrieval and e-commerce.
  - Predict which customers are likely to buy what products on online marketplaces like Amazon.
  - It can help in making better product recommendations.
- Suggest interactions or collaborations between employees in an organization.
  - Friend suggestion, suggesting cross-domain partners, suggesting experts etc.

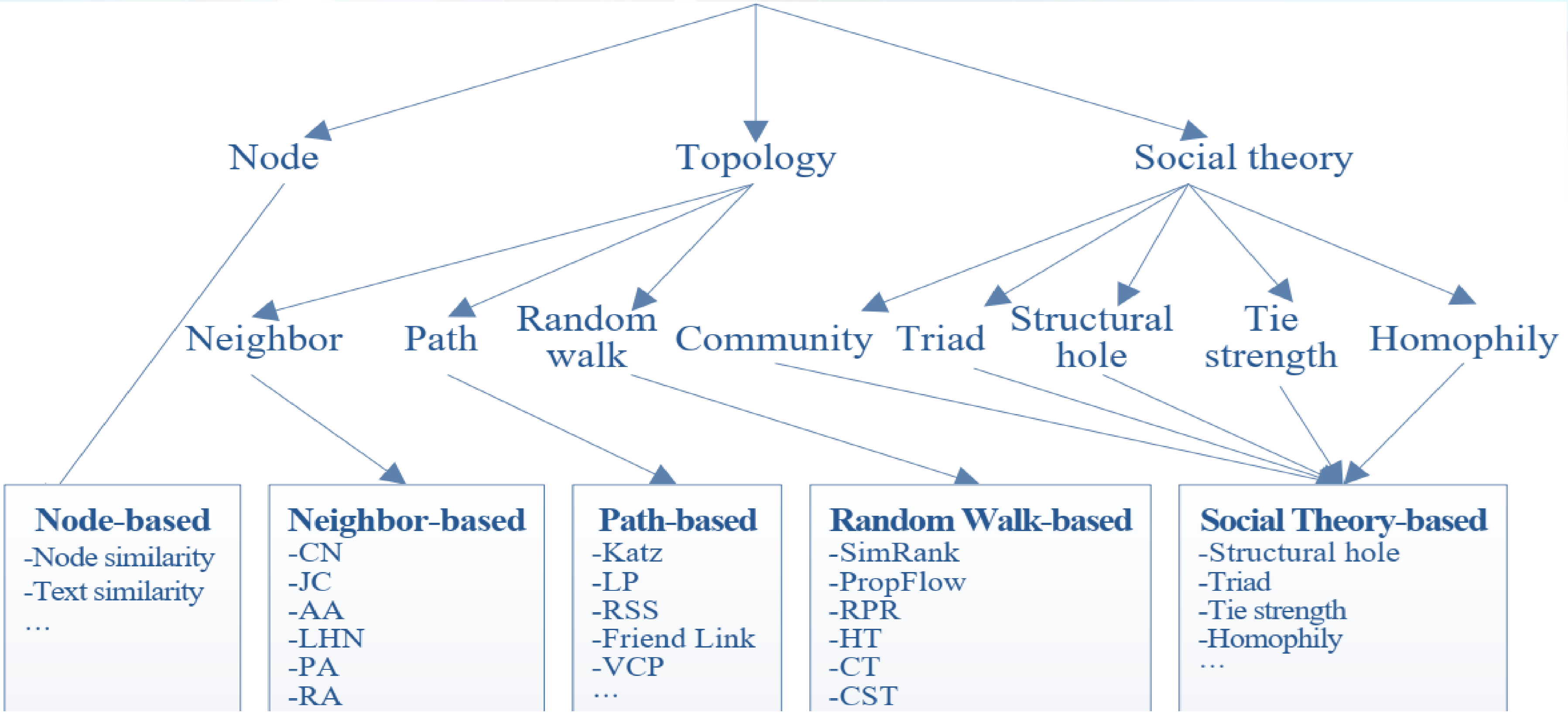
Applications

- Finding missing relationships in a social graph.
- Predicting possible contacts in a communication networks etc.
- Finding protein-protein interactions, can be applied in other biological networks.
- In security domain, useful to find abnormal or unusual communications.

Techniques

- Two main approaches.
  - Similarity based approaches.
  - Learning based approaches.
- Similarity based approaches: compute the similarity scores of non-connected pairs of nodes in a network. If the scores are higher, more the chances to connect them in future.
- Learning based approaches: use some ML models such as binary classifier to solve the problem. If a link possible then classified as +ve else –ve. It also uses some similarity features and other attributes to improve classification accuracy.

Similarity Based Approaches



Node

Topology

Social theory

Neighbor

Path

Random walk

Community

Triad

Structural hole

Tie strength

Homophily

Node-based
-Node similarity
-Text similarity
…

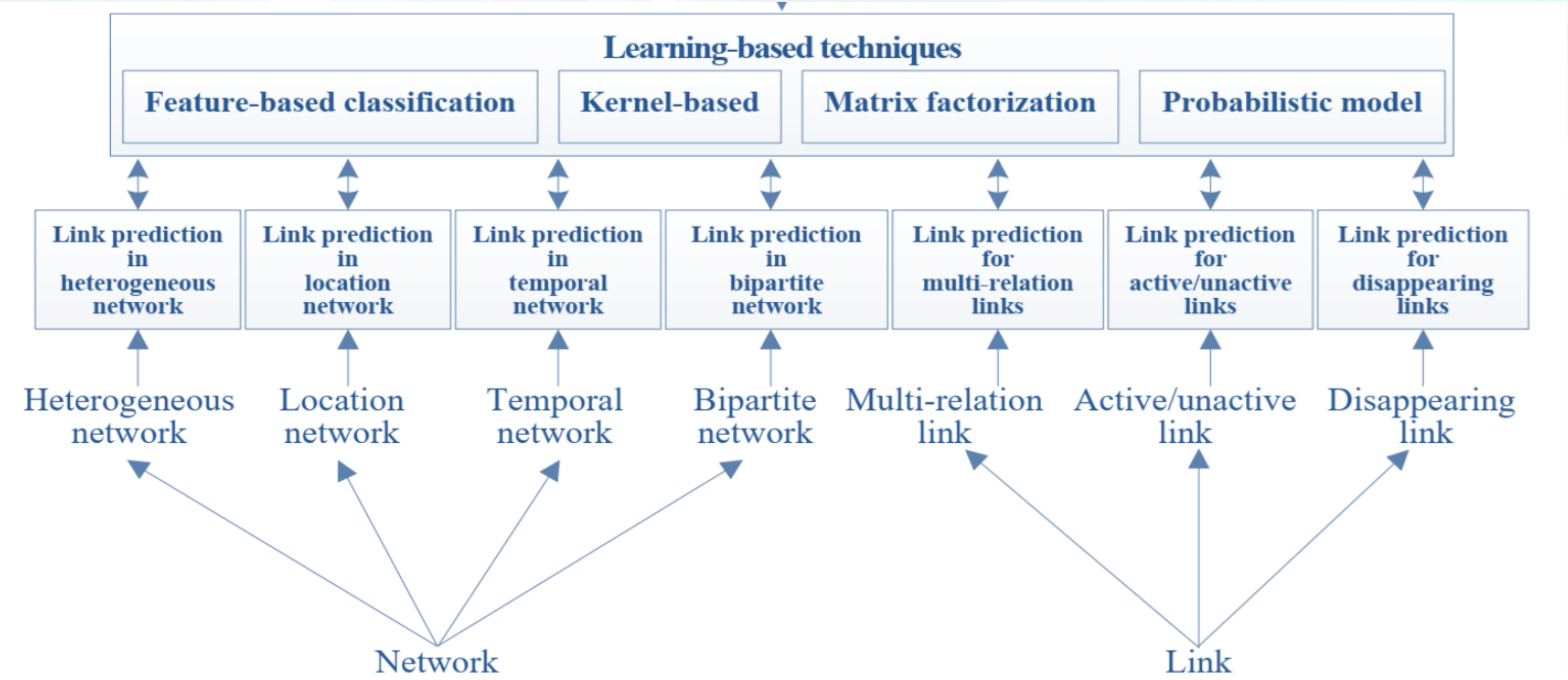
Neighbor-based
-CN
-JC
-AA
-LHN
-PA
-RA

Path-based
-Katz
-LP
-RSS
-Friend Link
-VCP
…

Random Walk-based
-SimRank
-PropFlow
-RPR
-HT
-CT
-CST

Social Theory-based
-Structural hole
-Triad
-Tie strength
-Homophily
…

Learning Based Approaches



Learning-based techniques

Feature-based classification

Kernel-based

Matrix factorization

Probabilistic model

Link prediction in heterogeneous network

Link prediction in location network

Link prediction in temporal network

Link prediction in bipartite network

Link prediction for multi-relation links

Link prediction for active/unactive links

Link prediction for disappearing links

Heterogeneous network

Location network

Temporal network

Bipartite network

Multi-relation link

Active/unactive link

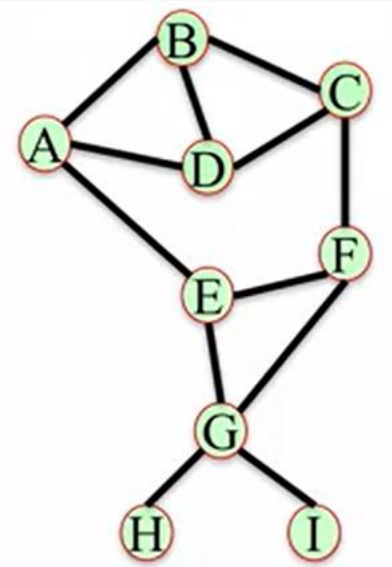
Disappearing link

Network

Link

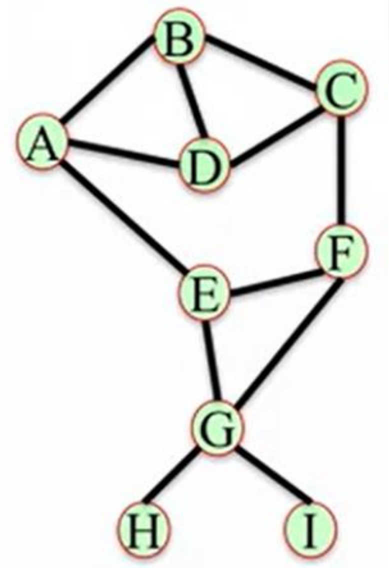
Common Neighbors

- The common-neighbors predictor captures the notion that two strangers who have a common friend may be introduced by that friend.
  - CN(A,C)=N(A)∩N(C)={B,D}=2
  - CN(A,H)= N(A)∩N(H)={}=0
- Here, A and C will establish a link in future but the chances for A and H is zero.
- In NetworkX, nx.common_neighbors(G,u,v)



Jaccard Coefficient

- Is a normalized score of common neighbors.
- JC(X,Y)=N(X)∩N(Y)/N(X)UN(Y)
- JC(H,I)=N(H)∩N(I)/N(H)UN(I)=1/1=1
- JC(A,C)={B,D}/{B,D,E,F}=2/4=0.5
- For all other pairs ,its less than 1.
- In NetworkX, nx.jaccard_coefficient(G)

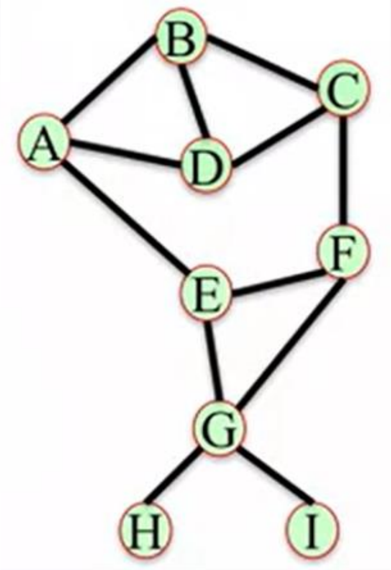


Adamic-Adar Coefficient

- The AA metric was proposed by Adamic and Adar for computing similarity between two web pages first, later it has been widely used in social networks.
- Related to Jaccard coefficient.
- Also known as frequency weighted common neighbors.
- Common neighbors which have fewer neighbors are weighted more heavily.
- It is defined as: $AA(X,Y)=\Sigma_{u\epsilon CN(X,Y)}\ 1/\log(N(u))$

Adamic-Adar Coefficient

- AA(A,C)=$\Sigma_{u\epsilon\{B,D\}}$ 1/log(N(u))
- AA(A,C)= 1/log(3)+1/log(3)=4.2
- In NetworkX, nx.adamic_adar_index(G)



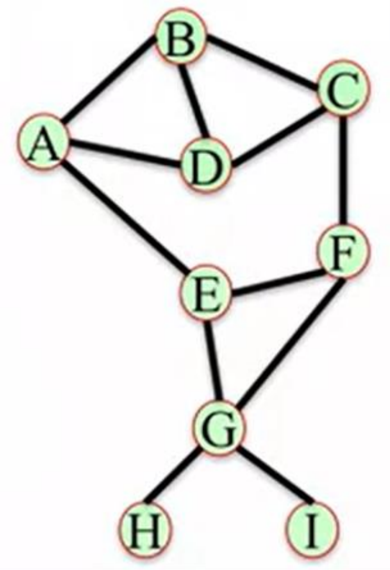
Preferential Attachment

- One well-known concept in social networks is that users with many friends tend to create more connections in the future. This is due to the fact that in some social networks, like in finance, the rich get richer.
- The PA metric indicates that new links will be more likely to connect higher-degree nodes than lower ones.

Preferential Attachment

- It is calculated as: PA(X,Y)=N(X)*N(Y)
  - PA(A,G)=N(A)*N(G)=3*4=12
  - PA(C,E)=N(C)*N(E)=3*3=9
- In NetworkX,

nx.preferential_attachment(G)

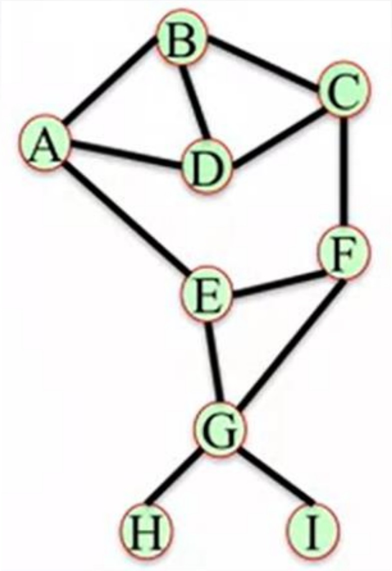


Resource Allocation

- This metric is motivated by the physical resource allocation.
- RA is similar to AA.
- Both suppress the contribution of high degree common neighbors.
- Therefore, both RA and AA have very close prediction results for networks with small average degrees. But RA performs better for networks with high average degree.
- Both RA and AA use direct neighbors and neighbors of neighbors.

Resource Allocation

- It is defined as: $RA(X,Y)=\Sigma_{u\epsilon CN(X,Y)}1/(N(u))$
- $RA(A,C)=1/3+1/3=2/3$
- In NetworkX, nx.resource_allocation_index(G)



Katz Index

- Path based technique.
- This heuristic defines a measure that directly sums over collection of paths, exponentially damped by length to count short paths more heavily.
- The Katz-measure is a variant of the shortest-path measure.
- The idea behind the Katz-measure is that the more paths there are between two vertices and the shorter these paths are, the stronger the connection.

Katz Index

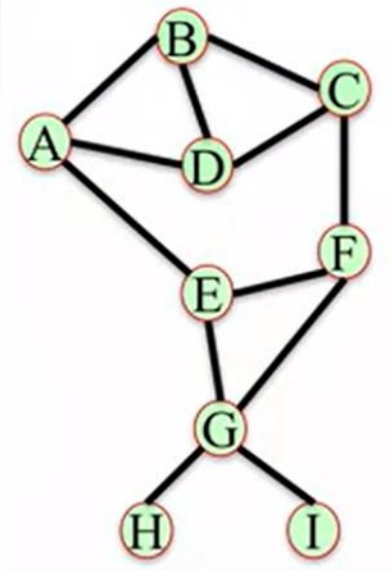
- It is calculated as:

$$\mathrm{Katz}(x,y)=\sum_{l=1}^{\infty}\beta^{l}\cdot|path_{x,y}^{l}|=\beta A+\beta^{2}A^{2}+\beta^{3}A^{3}+\cdots$$

- $\text{Path}^{l}{}_{x,y}$ Number of l length paths from x to y and β>0.

Katz Index

- To calculate Katz(A,C)
- $path^2_{A,C}$=2, $path^3_{A,C}$=3
- β=0.5 (very small value will cause Katz metric much like CN metric because paths of long length contribute very little to final similarities).
- Katz(A,C)=$0.5*0+0.5^2*2+0.5^3*3$=0.5+0.375=0.875

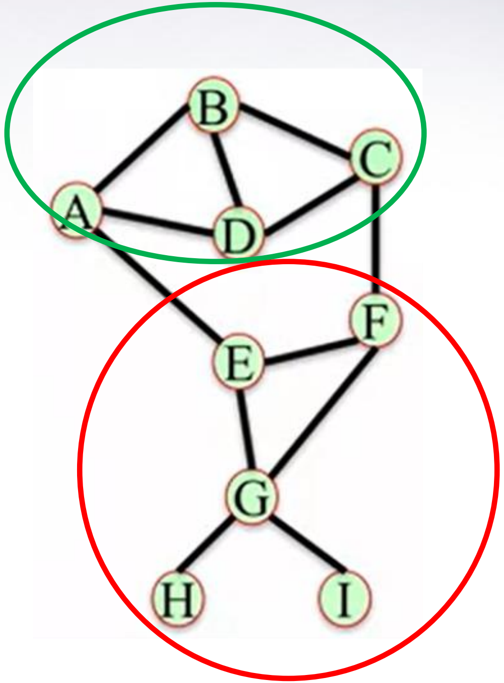


Community Common Neighbors

- Community based mechanism.
- If nodes belongs to same community, more common neighbors and more likely to establish links, but less probable if they are in different communities.
- Common Neighbor Soundarajan Hopcroft score.
- It is defined as: $\text{cn\_soundarajan\_Hopcroft}(X,Y)=|N(X)\cap N(Y)|+\Sigma_{u\in|N(X)\cap N(Y)|}$ f(u), where f(u)=1, if u in same community as of X and Y else 0.

Soundarajan Hopcroft

- Score(A,C)=2+1+1=4
- Score(E,I)=1+1=2
- Score(A,G)=1+0=1
- Score(A,H)=0+0=0



Thanks...........