PREDICTION AND CLASSIFICATION

Analysis of Large Scale Social Networks

Bart Thijs

Link Prediction:

Given the topology of a network at time *t*, it is the aim to predict the emergence of additional edges in the interval *t* – *t'*

Local level: will there be a link between two unconnected nodes?
Global level: What is the set of node pairs that will get connected in the near future?

Assumption: prediction is purely based on structure of the network itself

WHAT?

Work together as a group of students to solve the following tasks:

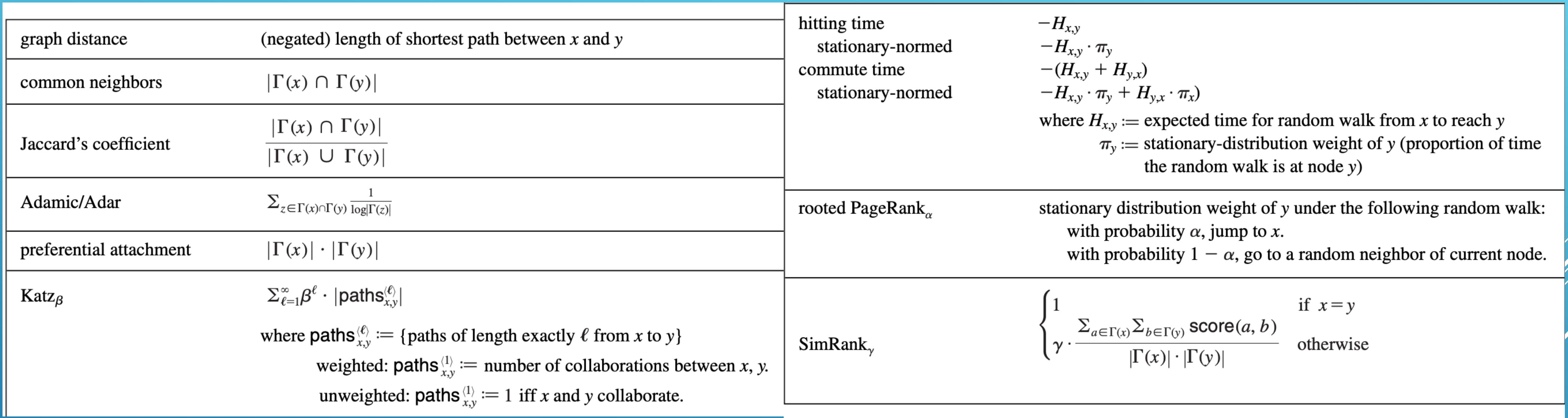
1. Define an applications where link prediction can be used
2. Describe the topology of the underlying network, possible data and graph data model
3. Find a methodology to perform the link prediction
   a. Find literature that reviews these techniques
   b. link that to other courses in the MAI program

Time: 20 minutes in group

Outcome:
Present Application, Data and possible methodology in main session
Short description in the discussion forum

TASK 1

Liben-Nowell & Kleinberg (2007)



graph distance	(negated) length of shortest path between x and y
common neighbors	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$
Adamic/Adar	$\Sigma_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log\lvert\Gamma(z)\rvert}$
preferential attachment	$\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$
Katz$_\beta$	$\Sigma_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\mathsf{paths}_{x,y}^{\langle\ell\rangle}\rvert$ where $\mathsf{paths}_{x,y}^{\langle\ell\rangle} :=$ {paths of length exactly ℓ from x to y} weighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle} :=$ number of collaborations between x, y. unweighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle} := 1$ iff x and y collaborate.
hitting time stationary-normed commute time stationary-normed	$-H_{x,y}$ $-H_{x,y} \cdot \pi_y$ $-(H_{x,y} + H_{y,x})$ $-H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$ where $H_{x,y} :=$ expected time for random walk from x to reach y $\pi_y :=$ stationary-distribution weight of y (proportion of time the random walk is at node y)
rooted PageRank$_\alpha$	stationary distribution weight of y under the following random walk: with probability α, jump to x. with probability $1 - \alpha$, go to a random neighbor of current node.
SimRank$_\gamma$	$\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\Sigma_{a\in\Gamma(x)}\Sigma_{b\in\Gamma(y)} \mathsf{score}(a,b)}{\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert} & \text{otherwise} \end{cases}$

FIG. 2. Values for $\mathsf{score}(x, y)$ under various predictors; each predicts pairs $\langle x, y\rangle$ in descending order of $\mathsf{score}(x, y)$. The set $\Gamma(x)$ consists of the neighbors of the node x in G_{collab}.

PREDICTING RELATIONSHIPS

Return to your group and proceed with this question:
What is the effect of these three common features of real world networks:

- Small world properties
- Preferential Attachment
- High Clustering

Describe the possible existence of these three features in your dataset, how this might compromise the results obtained with your proposed methodology and how you would solve that.

Timing and outcome: 20 min, main group presentation and forum

TASK 2

Return to your group and proceed with this question:
Consider the link prediction problem as a classic binary classification problem.

How would you implement a different approach to solve the prediction problem? What kind of information would you require?

Timing and outcome: 20 min, main group presentation and forum

TASK 3