Data Mining

Web Mining II – Web Structure Mining (Part B)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Where am I?

- Part A introduces the definitions, terms and concepts for connectivity analysis and Web structure mining.

- Part B presents the HITS algorithm for connectivity analysis.

Connectivity Analysis

1. Given a query or topic *Q*, a root-set $S = \{s_1, \ldots, s_n\}$ of *n* seed pages is collected by making a search request, based on *Q*, to a search engine.

Typically, only a fixed number of the pages returned by the search engine should be used.

2. The root-set *S* is then expanded to a larger set *T*, called a base set or neighborhood graph, by adding any page *p* that has a hyperlink to or from any page in *S*. That is,

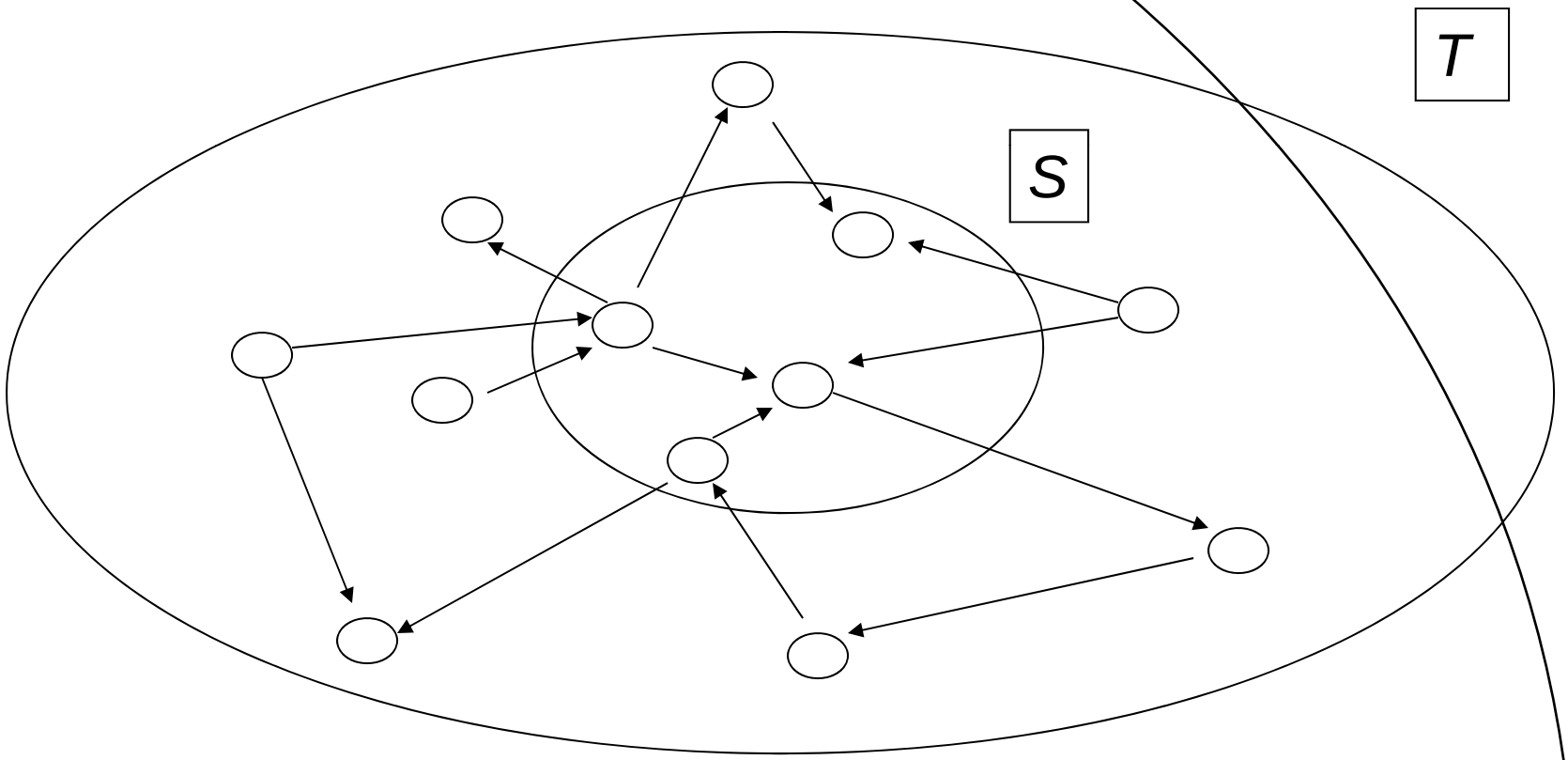
T = *S* U *N* where,

N = { ρ | ∃ δ in *S*

such that either

ρ $\longrightarrow$ δ or δ $\longrightarrow$ ρ }.

The relationship between *T* and *S* is depicted in the following figure.



Each page ρ ϵ T is initially assigned an authority weight and a hub weight of 1, denoted by $\alpha(\rho)$ and $\lambda(\rho)$, respectively.

3. Each page's α and λ are then iteratively updated as follows:

$$\alpha(\rho) = \sum_{\delta \to \rho} \lambda(\delta)$$

$$\lambda(\rho) = \sum_{\rho \to \delta} \alpha(\delta)$$

Thus, each iteration replaces $\alpha(\rho)$ by the sum of $\lambda(\delta)$, where δ links to ρ; and then replaces $\lambda(\rho)$ by the sum of $\alpha(\delta)$, where δ is linked by ρ.

Normalize $\alpha(\rho)$ and $\lambda(\rho)$ and repeat Step 3 until α and λ converge to stable states of authority and hub weights, which typically takes about 10 iterations.

4. The community is discovered by taking the top *k* pages with the highest α values and the top *k* pages with the highest λ values.

End of
Web Structure Mining Module
(Part B)