

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 $S = \{s_1, \dots, s_n\}$ 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,

$T = S \cup N$ where,

$N = \{ p \mid \exists \delta \text{ in } S$

such that either

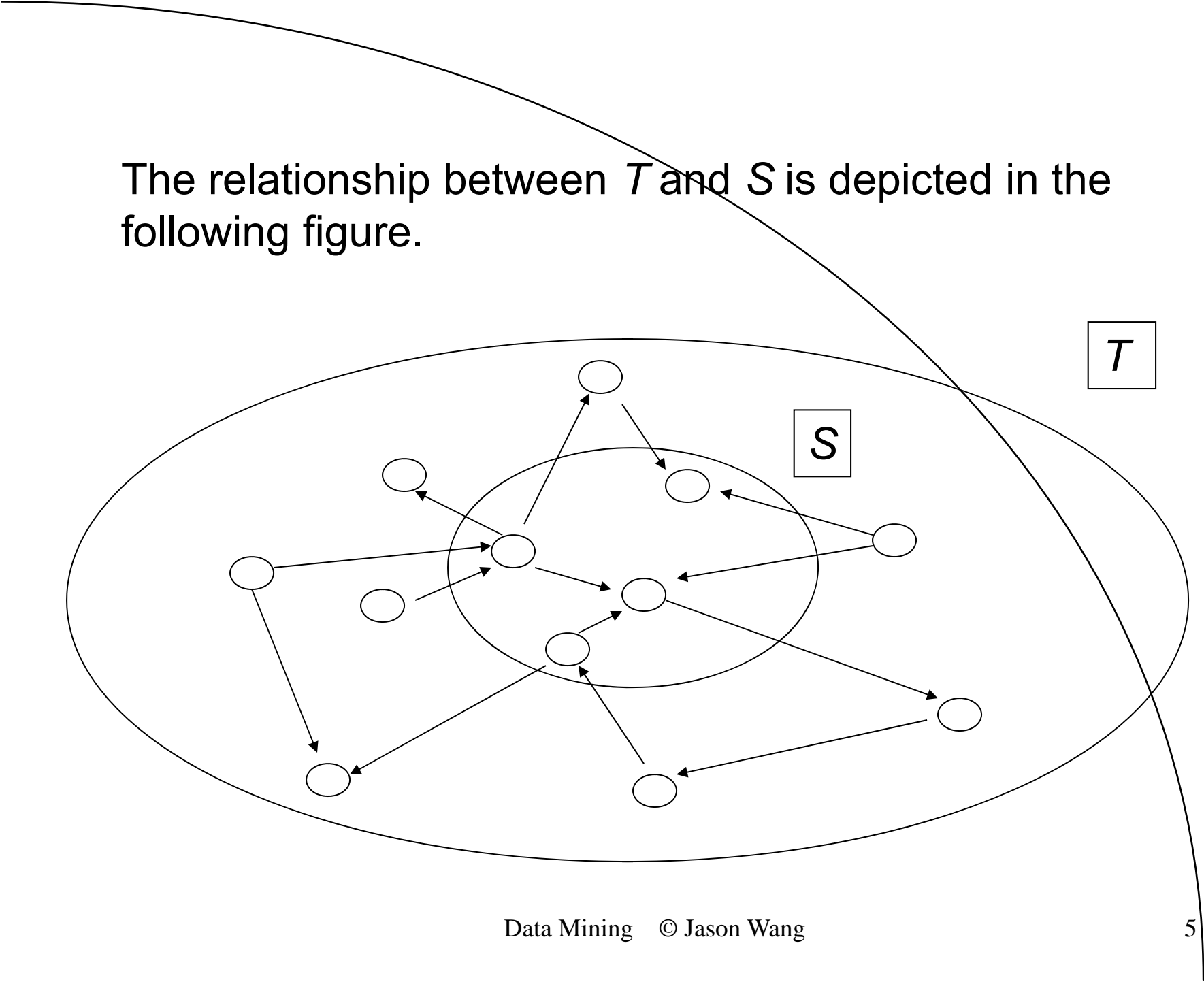
$p \longrightarrow \delta \text{ or } \delta \longrightarrow p \}$.

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

The diagram illustrates a directed graph structure. A large outer oval labeled T contains the entire graph. Inside T is a smaller oval labeled S which contains 6 nodes. The nodes in S are interconnected with arrows, forming a complex web. Nodes outside S but within T also have arrows pointing to or from nodes within S .

Data Mining © Jason Wang

5



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

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

$$\alpha(p) = \sum_{\delta \rightarrow p} \lambda(\delta)$$

$$\lambda(p) = \sum_{p \rightarrow \delta} \alpha(\delta)$$

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)