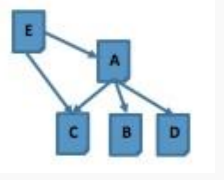
Exercise

2 Consider the following graph of web pages:



(a) Determine the most interesting hub.
(b) Determine the most important authority.
(c) Suppose we are looking for information about a car model X and page A contains that model, how would that change your previous results?

48

Larry Page and Sergey Brin

Co-founders of Google

Developers of the PageRank algorithm

49

PageRank vs HITS

- HITS was proposed in **January 1998** (Kleinberg)
- PageRank was proposed in **April 1998** and is used by Google (Sergey Brin and Larry Page)
- HITS and PageRank have many similarities, but they have very important differences, as **PageRank**:
  - Does not depend on the query
  - Is based on a single score
- The idea of **PageRank:**
  - To rank pages according to their prestige
  - prestige is (mainly) determined by the **in-links** and their respective prestige
  - I.e., *a page is important if it is pointed to by other important pages*.

50

PageRank: The idea

1. Given a network of n pages, assign to each page i a PageRank value $r(i) = 1/n$.
2. Until k-th iteration, or convergence, do:
   - update the PageRank value of each page i by:

$$r(i) = \sum_{\boxed{j \to i}} \frac{r(j)}{O_j}$$

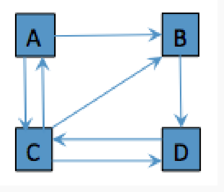
Page j pointing to i

where O_j is the number of out-links from page j.

- $r(i)$ calculates the **probability of getting to page *i*** when coming from each of the possible page's j that point to it.
- Most important pages, will have higher probability.
- The value of k depends on the size of the network.

51

PageRank: Example



*Assume **k=2***

i	$R_0(i)$	$R_1(i)$	$R_2(i)$	PageRank
A	1/4	1/12	1/8 = 0.125	4
B	1/4	5/24	1/6 = 0.167	3
C	1/4	3/8	3/8 = 0.375	1
D	1/4	1/3	1/3 = 0.333	2

Step 1:

$R_1(A) = R_0(C) \div O(C) = 1/4 \div 3 = 1/12$

$R_1(B) = R_0(A) \div O(A) + R_0(C) \div O(C) = 1/4 \div 2 + 1/4 \div 3 = 5/24$

$R_1(C) = R_0(A) \div O(A) + R_0(D) \div O(D) = 1/4 \div 2 + 1/4 \div 1 = 3/8$

$R_1(D) = R_0(B) \div O(B) + R_0(C) \div O(C) = 1/4 \div 1 + 1/4 \div 3 = 1/3$

Step2:

$R_2(A) = 3/8 \div 3 = 1/8$

$R_2(B) = 1/12 \div 2 + 3/8 \div 3 = 1/6$

$R_2(C) = 1/12 \div 2 + 1/3 \div 1 = 3/8$

$R_2(D) = 5/24 \div 1 + 3/8 \div 3 = 1/3$

52

PageRank

- We have seen an iterative approach, where we updated the values one by one.
- We have a system of n equations with n unknowns.
- We can use a matrix to represent all the equations and do all the calculations at the same time.
- Let $R = (r(1), r(2), \dots, r(n))^T$ be a n-dimensional column vector of PageRank values.
- Let A be the adjacency matrix of our network, with

$$A_{ij} = \begin{cases} 1/O_i & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$$

where O_i is the number of out-links from page i.

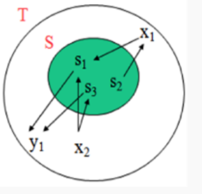
- Then, we can write the system of n equations as

$$R^{(k+1)} = A^T R^{(k)}$$

53

PageRank (problems and solutions)

- For $R = A^T R$ to have a unique solution, A must be:
  - **a)** **stochastic**, i.e. all rows must sum 1



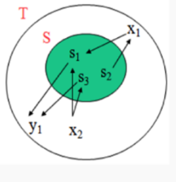
- often it is not: there are pages with no out-links (in the example: y1)
- solution 1: remove pages without out-links
- solution 2: artificially insert equal weights into a row with zeros

$$
\begin{array}{c|cccccc}
 & S_1 & S_2 & S_3 & X_1 & X_2 & Y_1 \\
\hline
S_1 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\
S_2 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 \\
S_3 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\
X_1 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\
X_1 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\
Y_1 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
\Rightarrow
\begin{array}{c|cccccc}
 & S_1 & S_2 & S_3 & X_1 & X_2 & Y_1 \\
\hline
S_1 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\
S_2 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 \\
S_3 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\
X_1 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\
X_1 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\
Y_1 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6
\end{array}
$$

PageRank (problems and solutions)

b) **irreducible**, i.e. in the graph there is a path from any node to any other node

- often it is not the case (there is no path from S1 to S2)



c) **aperiodic**, i.e. the greatest common divisor of all cycles for each node is 1

- *Example:* $A \rightarrow B$, $B \rightarrow C$, $C \rightarrow A$: the cycle has period 3
- No loop traps

Solution to deal with above two problems:

- Add a link from each page to every other page and give each link a small transition probability controlled by a parameter ***d*** .

PageRank - the "damping factor"

- In this model, the "random surfer" at a page has two options:
  - with probability ***d*** , he randomly chooses an out-link to follow;
  - with probability **1 − *d*** , he "jumps" to a random page without a link, by typing its URL ("teleportation").

$$r(i) = (1-d) + d \times \sum_{j \to i} \frac{r(j)}{O_j}$$

- ***d*** is called the **damping factor**, which can be set between 0 and 1.

56

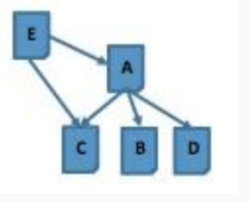
Observations about PageRank

- PageRank can be computed offline
  - the values of pages (and the implicit ordering amongst them) is query independent
  - **advantage**: at query time only, a lookup is needed
  - **disadvantage**: a page can be an authority in a topic but not in general
- PageRank is more robust to SPAM
  - importance of a page depends on in-links not on out-links
  - it is not easy to add in-links into a page from other important pages
- PageRank is more robust to perturbations in the input than HITS
- PageRank, however, does not consider time

57

Exercise

3 Consider the following graph, assuming a damping factor of 0.9



Suppose the PageRank of A and E is 1. What is the PageRank of B and C?

r(B) = (1 - 0.9) + 0.9 * 1/3 = 0.4
r(C) = (1 - 0.9) + 0.9 * (1/2 + 1/3) = 0.85

58

Community Discovery

59

Community Structure

- **Community:** group of entities (people, organizations) sharing common interests.
  - Users who like specific jazz music
  - People who speak Italian
  - Trekkies
  - …
- What for?
  - Source of resources for users with similar interests
  - Target advertising
  - Predictive analysis
  - Understand the sociology of the web
  - …

60

Community Structure (cont.)

- Given a set of entities $S = \{s_1, s_2, \ldots, s_n\}$ of the same type
- A community is a pair $C = (T, G)$, where
  - T is the community topic (usually represented with a set of keywords)
  - $G \subseteq S$ is the set of all entities in S that shares the topic T.
  - If $s_i \in G$, s_i is said to be a member of the community C.
- Example:
  - Users that are between 18 and 21 years old

61

Communities - how to find the common topic

- Web pages
  - Users in the same community are usually interconnected through hyperlinks
  - Pages may contain words that reveal the theme
- Emails
  - Members of a community tipically exchange emails
  - Emails contain words revealing the topic
- Documents
  - Members of a community are more likely to appear together in the same sentences or documents
  - Words indicate the community topic

62

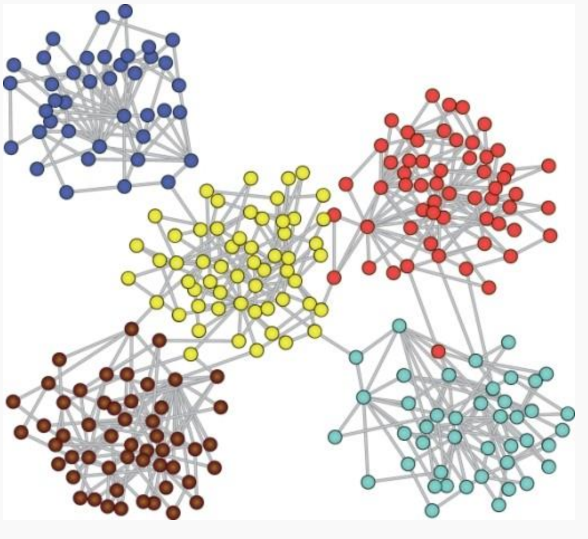
Communities Discovery

Community Discovery:

Discovering groups of nodes in a network where the nodes' group memberships **are not** explicitly given.

How to find that in a graph?

Look for **densely-knit** parts of the graph (e.g., a k-clique).



63

Some criteria to identify Communities

Main community detection approaches

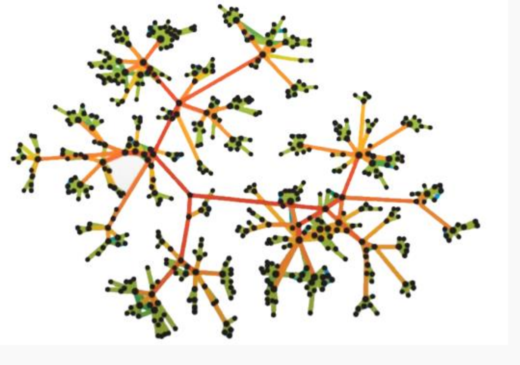
- Node-centric: each node in a group satisfies certain properties
  - Complete mutuality (cliques)
  - Reachability of members (e.g., k-clique, paths)
  - Node degrees
  - Relative frequency of ‘within community’ vs. ‘outside community’ links
- Group-centric: the whole group must satisfy certain properties regardless of the node-level properties
- Network-centric: partition the whole network into several disjoint sets
- Hierarchy-centric: construct a hierarchical structure of communities

64

Some algorithms for Communities Detection

Divisive Hierarchical Clustering

- Uses a topological similarity measure between each pair of nodes
- May consider **Single-Linkage Clustering**
  - two groups are considered separate communities if and only if all pairs of nodes in different groups have similarity lower than a given threshold
- Or **Complete-Linkage Clustering**
  - all nodes within every group have similarity greater than a threshold

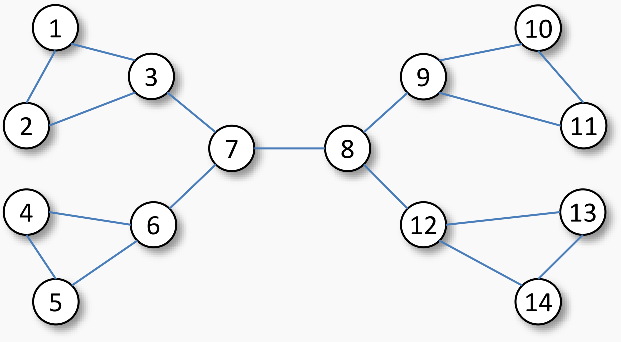


65

Algorithms for Communities Detection (Girvan-Newman)

- Girvan-Newman algorithm
  - Uses the betweenness centrality measure
  - Removes edges with the largest **edge betweenness** in every iteration
  - Repeat until there are no edges left

Edge Betweenness: Number of shortest paths passing through the edge.

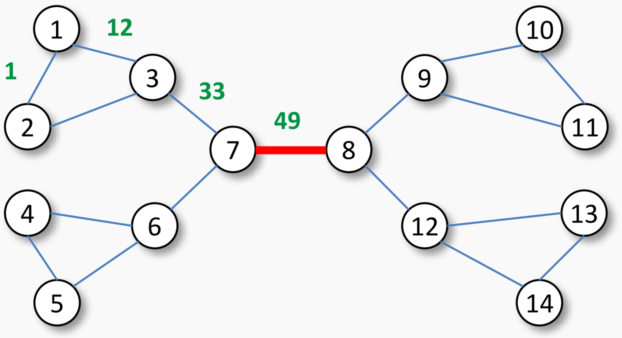


66

Algorithms for Communities Detection (Girvan-Newman)

- Girvan-Newman algorithm
  - Uses the betweenness centrality measure
  - Removes edges with the largest **edge betweenness** in every iteration
  - Repeat until there are no edges left

Edge Betweenness: Number of shortest paths passing through the edge.

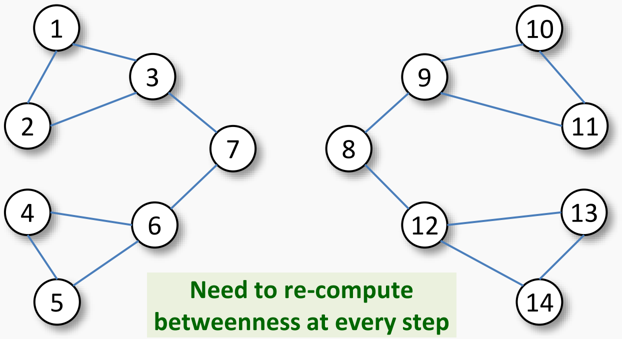


67

Algorithms for Communities Detection (Girvan-Newman)

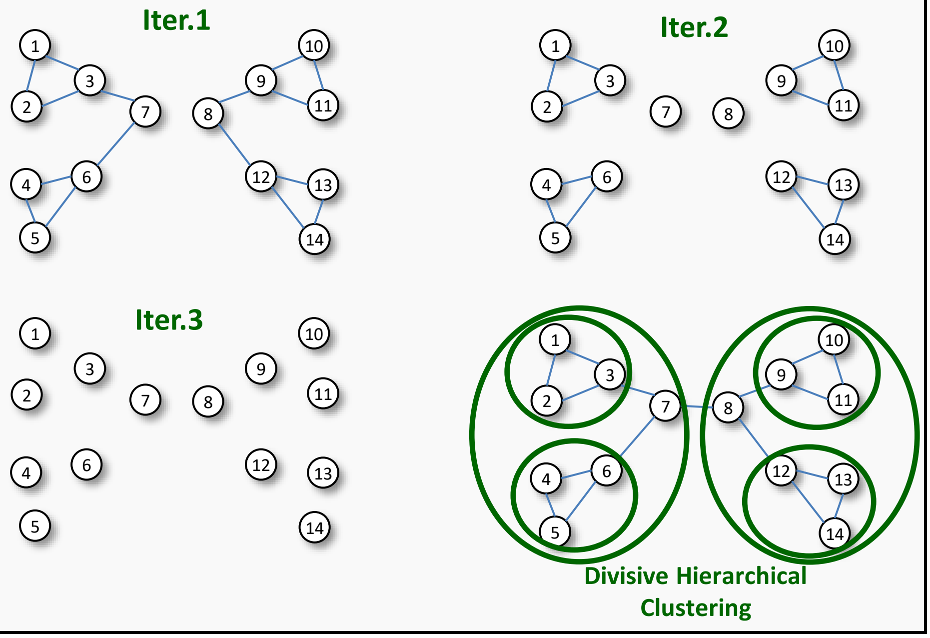
- **Girvan-Newman algorithm**
  - Uses the betweenness centrality measure
  - Removes edges with the largest **edge betweenness** in every iteration
  - Repeat until there are no edges left

Edge Betweenness: Number of shortest paths passing through the edge.



Need to re-compute betweenness at every step

Algorithms for Communities Detection (Girvan-Newman)



References

70

References

- Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005).
  Incorporating contextual information in recommender systems using a multidimensional approach.
  ACM Trans. Inf. Syst., 23(1):103-145.
- Adomavicius, G. and Tuzhilin, A. (2005).
  Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.
  IEEE Trans. on Knowl. and Data Eng., 17(6):734-749.
- Aggarwal, C. C. (2015).
  Data Mining, The Texbook.
  Springer.
- Breese, J. S., Heckerman, D., and Kadie, C. (1998).
  Empirical analysis of predictive algorithms for collaborative filtering.
  In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 43-52, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Craven, P.
  Google's pagerank explained and how to make the most of it.
  http://www.webworkshop.net/pagerank.html.

71

References (cont.)

- Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J. (1999).
  Combining collaborative filtering with personal agents for better recommendations.
  In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 439-446, Menlo Park, CA, USA. American Association for Artificial Intelligence.
- Han, J., Kamber, M., and Pei, J. (2011).
  ***Data Mining: Concepts and Techniques*.**
  Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.
- Huang, Z., Chen, H., and Zeng, D. (2004).
  Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering.
  ACM Trans. Inf. Syst., 22(1):116-142.
- Jayanthi, S. (2013).
  Web mining issues.
  http://webminingissues.blogspot.pt/.
- Jorge, A. (2016).
  Web mining.
  Slides.

72

References (cont.)

- Kammergruber, W. C., Viermetz, M., Ehms, K., and Langen, M. (2010).
  Using association rules for discovering tag bundles in social tagging data.
  In *2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM, Krakow, Poland, October 8-10, 2010*, pages 414-419.
- Kohavi, R., Henne, R. M., and Sommerfield, D. (2007).
  Practical guide to controlled experiments on the web: Listen to your customers not to the hippo.
  In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 959-967, New York, NY, USA. ACM.
- Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009).
  Controlled experiments on the web: Survey and practical guide.
  Data Min. Knowl. Discov., 18(1):140-181.
- Liu, B. (2011).
  ***Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*.**
  Springer, 2nd edition.
- Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2001).
  Effective personalization based on association rule discovery from web usage data.
  In *Proceedings of the 3rd International Workshop on Web Information and Data Management*, WIDM '01, pages 9-15, New York, NY, USA. ACM.

73

References (cont.)

- Oliveira, M. D. B. and Gama, J. (2012).
  A framework to monitor clusters evolution applied to economy and finance problems.
  Intell. Data Anal., 16(1):93-111.
- Palmisano, C., Gorgoglione, M., and Tuzhilin, A. (2008).
  Using context to improve predictive modeling of customers in personalization applications.
  IEEE Transactions on Knowledge & Data Engineering, 20:1535-1549.
- Resnick, P. and Varian, H. R. (1997).
  Recommender systems.
  Commun. ACM, 40(3):56-58.
- Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K., and Chakraborty, A. (2013).
  ***Practical Graph Mining with R*.**
  Chapman & Hall/CRC.
- Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001).
  Item-based collaborative filtering recommendation algorithms.
  In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285-295, New York, NY, USA. ACM.

74

References (cont.)

- Shani, G. and Gunawardana, A. (2011).
  ***Evaluating Recommendation Systems*, pages 257–297.**
  Springer US, Boston, MA.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2005).
  ***Introduction to Data Mining*.**
  Addison Wesley.

75