Plan for Today's Lecture(s)

- XPath
- The Architectural Perspective on Relationships
- Structure between Resources
- Link Architecture
- Bibliometrics and Altmetrics

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

8 October 2013
Lecture 12.6 – XPath

XPath (1)

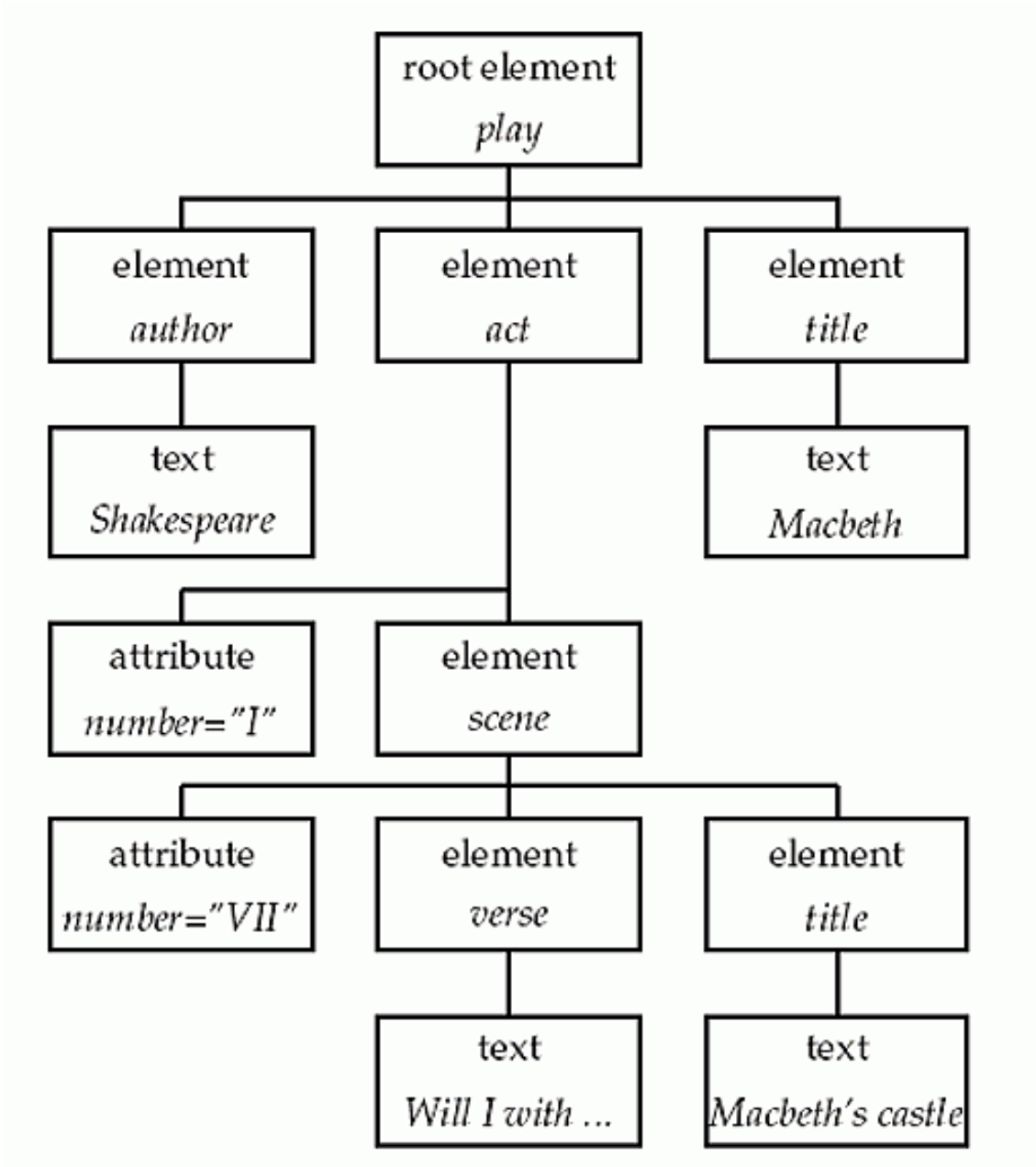
- A standard way of addressing parts of XML documents
- Defines the structures and patterns used by XML transformations, queries, and forms
- Similar in concept to addressing files on the filesystem, i.e. at a shell or command prompt, but much more general and powerful
- XPath lets you move in all sorts of different directions and multiple levels in a single step

XPath (2)

- Key idea is to view an XML document as a tree of information items called "nodes" - this is more abstract than thinking of it as a stream of marked-up text
- XPath lets us select a set of matching candidate documents for retrieval that might be further analyzed for relevance

The Node Tree

- XPath describes the locations of addresses of parts of XML documents by navigating through the "node tree" along a "node axis"
- There are seven types of nodes, corresponding to the different kinds of "stuff" in XML documents (most important : "element," "attribute," and "text")
- There are thirteen different axes that specify different ways of following relationships among the nodes (the default is "child" -- look down the tree at the nodes directly linked as children)

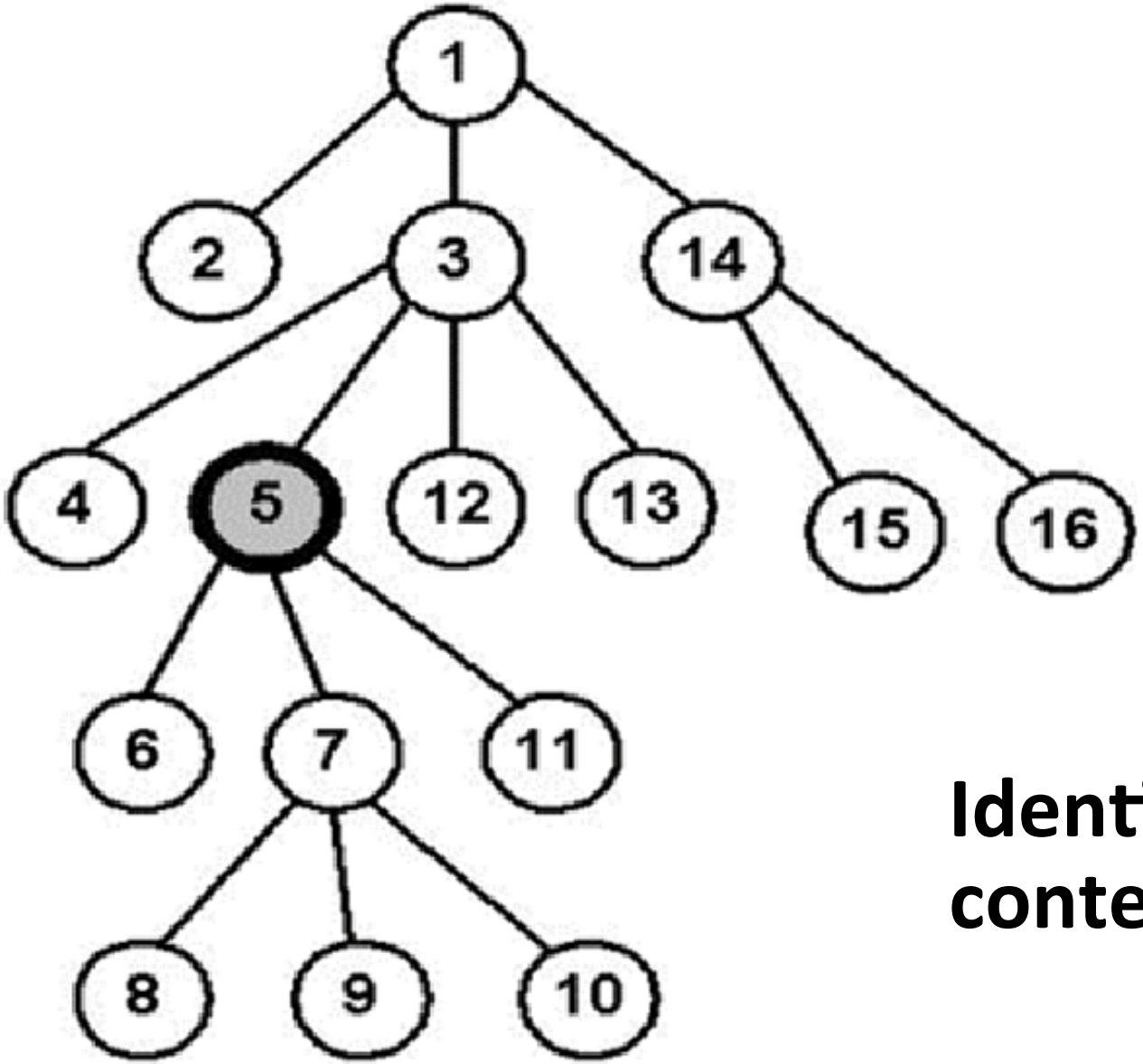


A Tree For A Shakespeare Play

Figure 10.2 From Manning (2008)

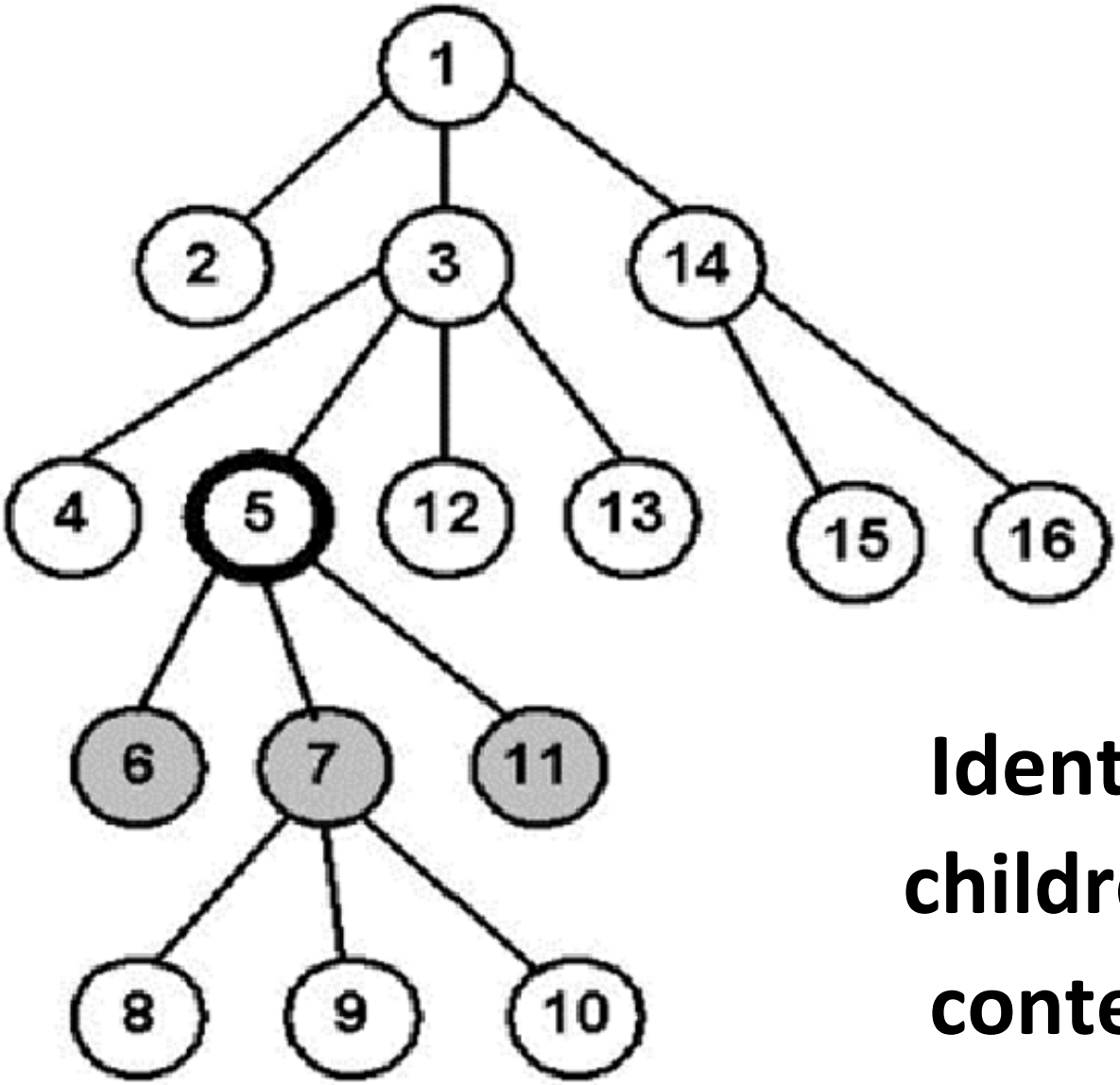
The Node Axes

- There are thirteen different axes that define different directions of "walking the node tree" depth-first starting from the context node;
- Depth-first means visiting all the children recursively throughout the entire document, shown using the numbering of the nodes in the following graphs
- The SELF Axis identifies the context node
- Other nodes are defined relative to the context node



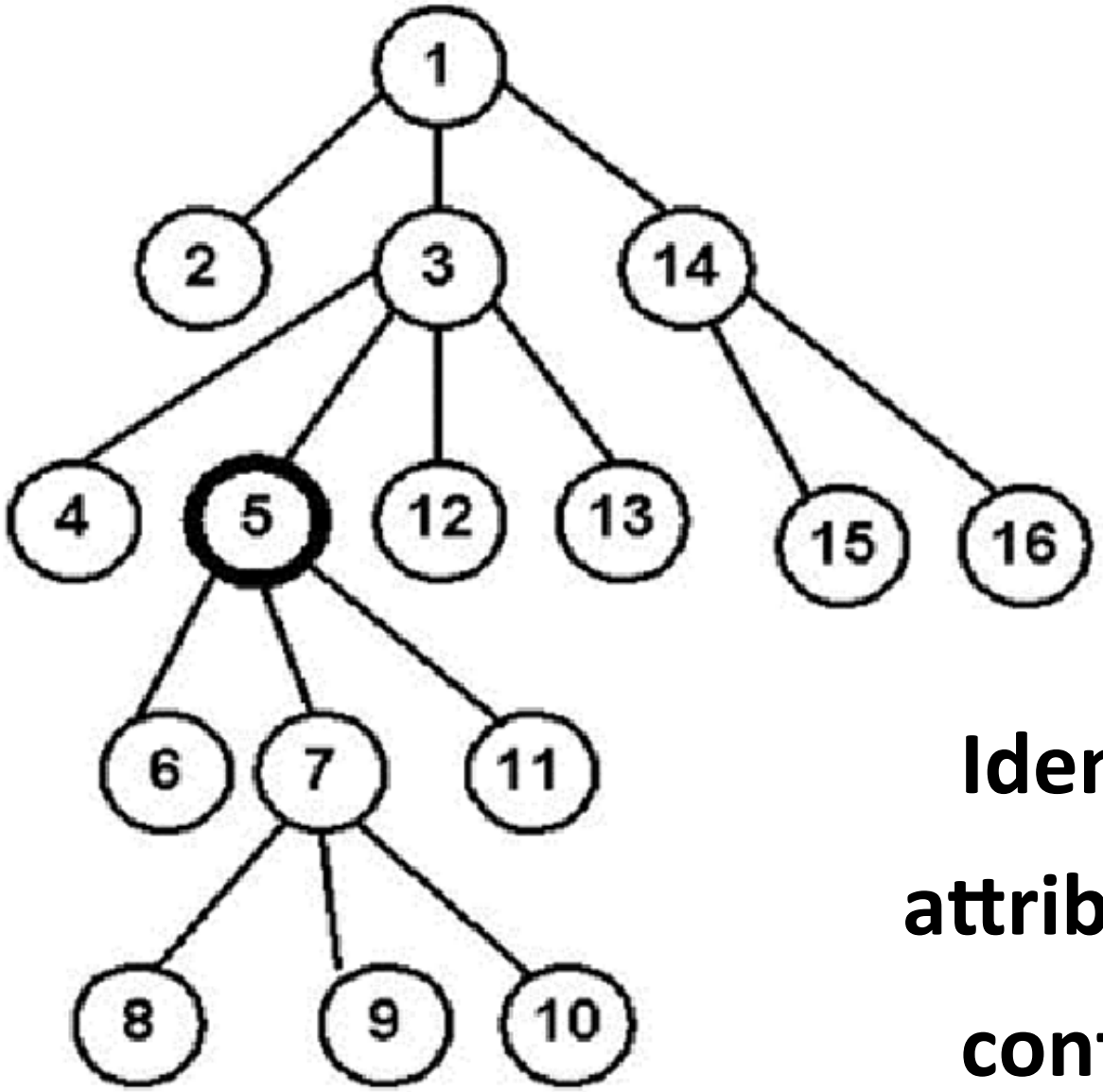
The SELF Axis

Identifies the context node



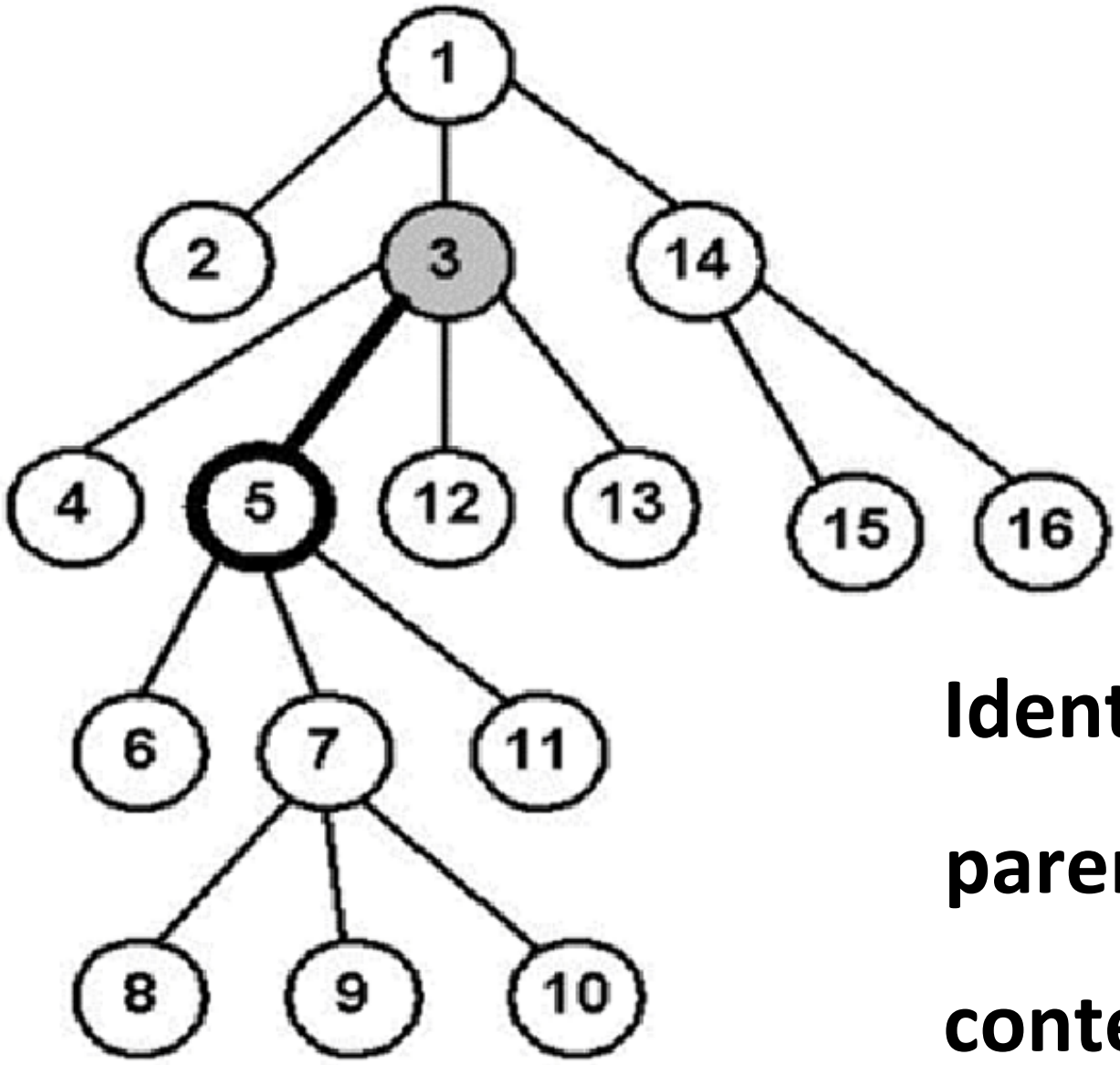
The CHILD Axis

Identifies the children of the context node



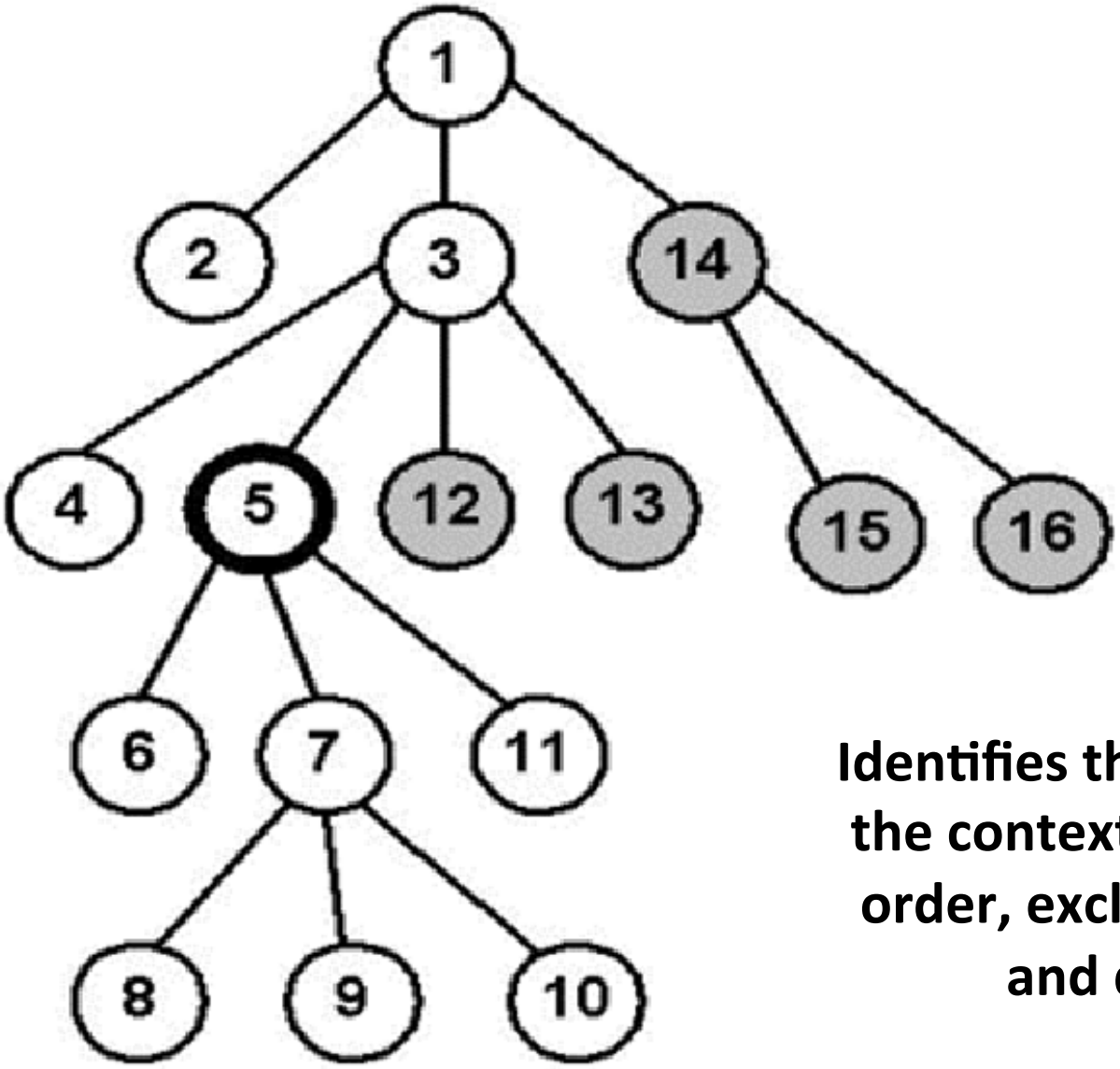
The ATTRIBUTE Axis

Identifies the attributes of the context node



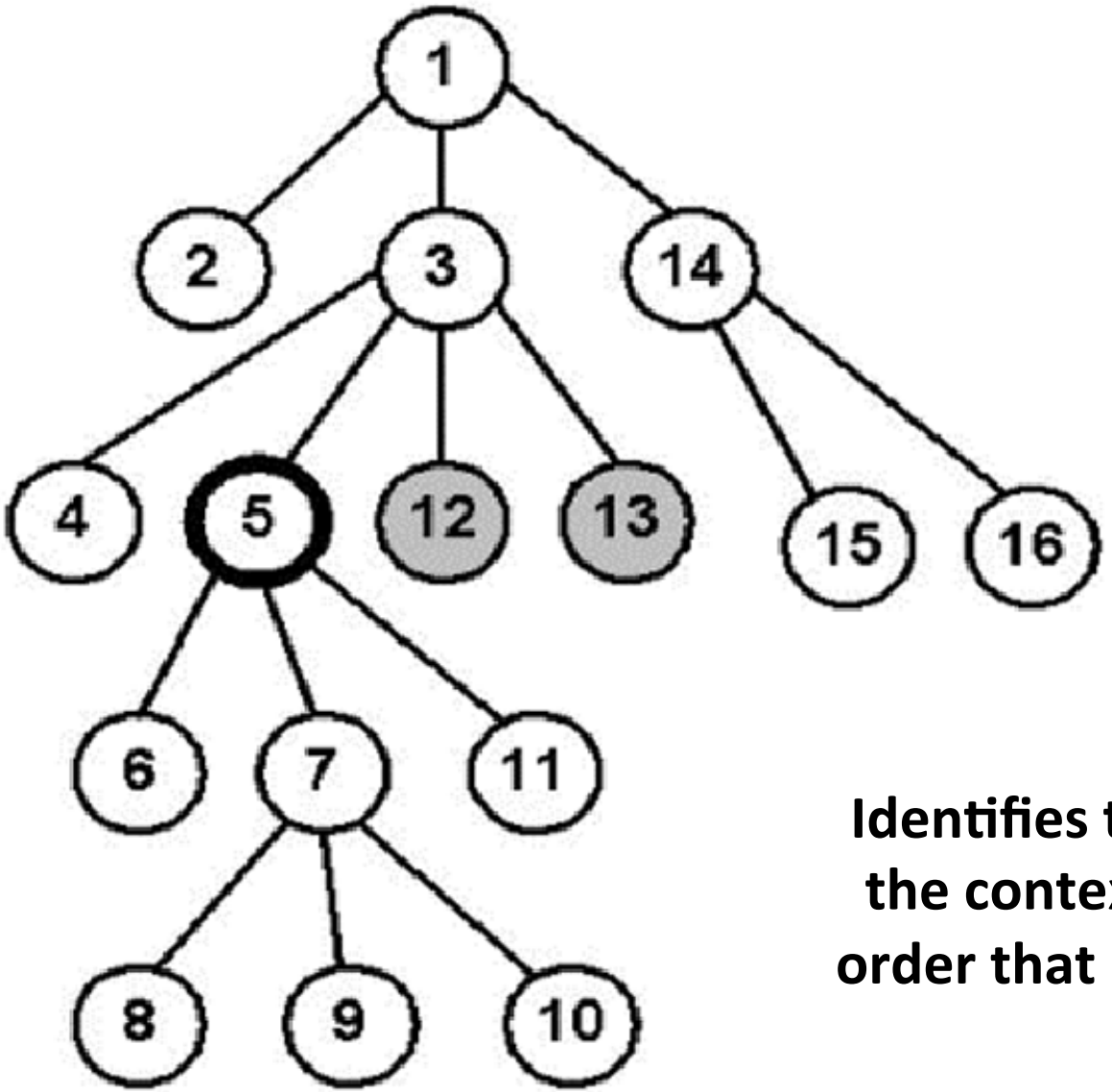
The PARENT Axis

Identifies the parent of the context node



The FOLLOWING Axis

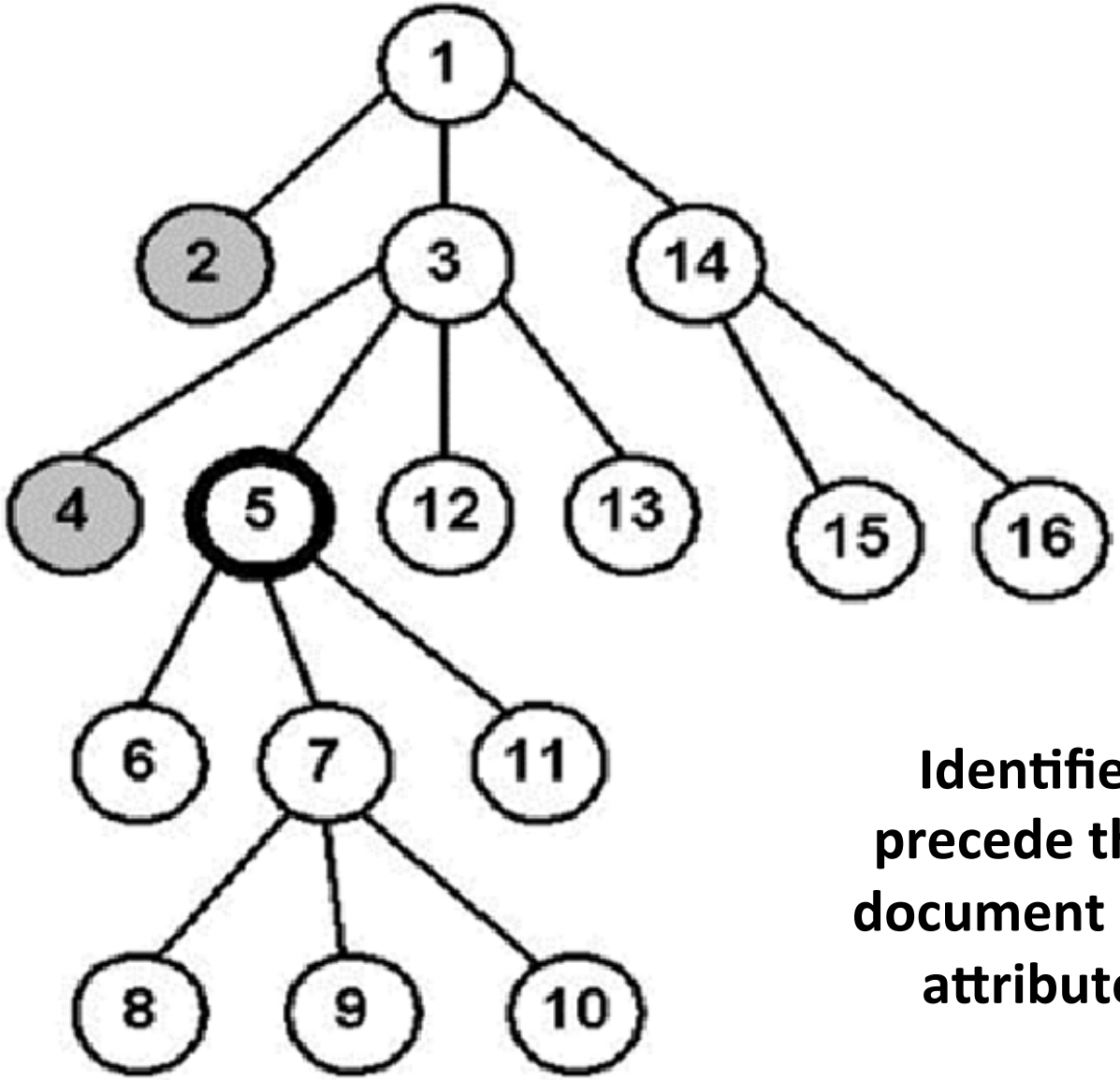
Identifies the nodes that follow the context node in document order, excluding its attributes and descrendants



The

FOLLOWING SIBLING Axis

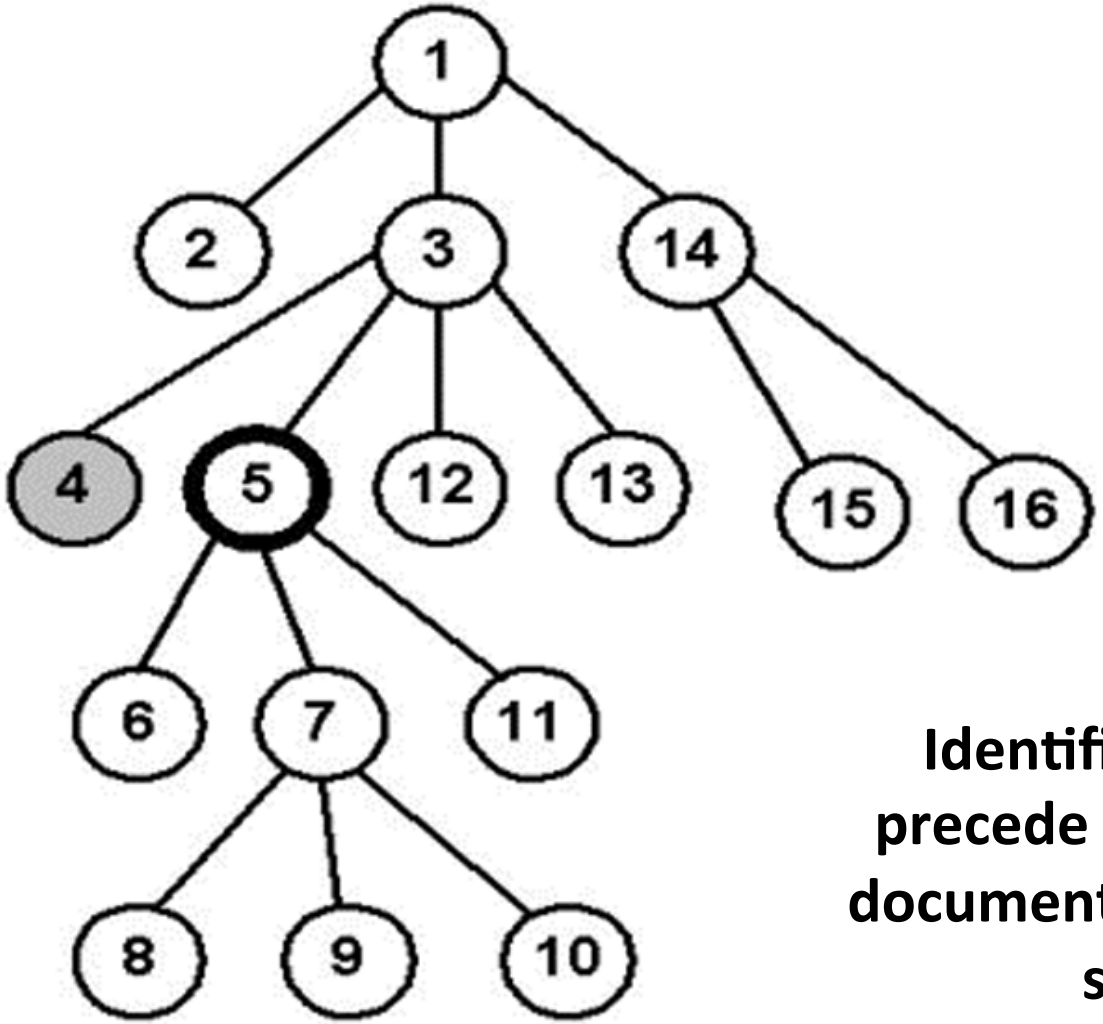
Identifies the nodes that follow the context node in document order that have the same parent



The PRECEDING Axis

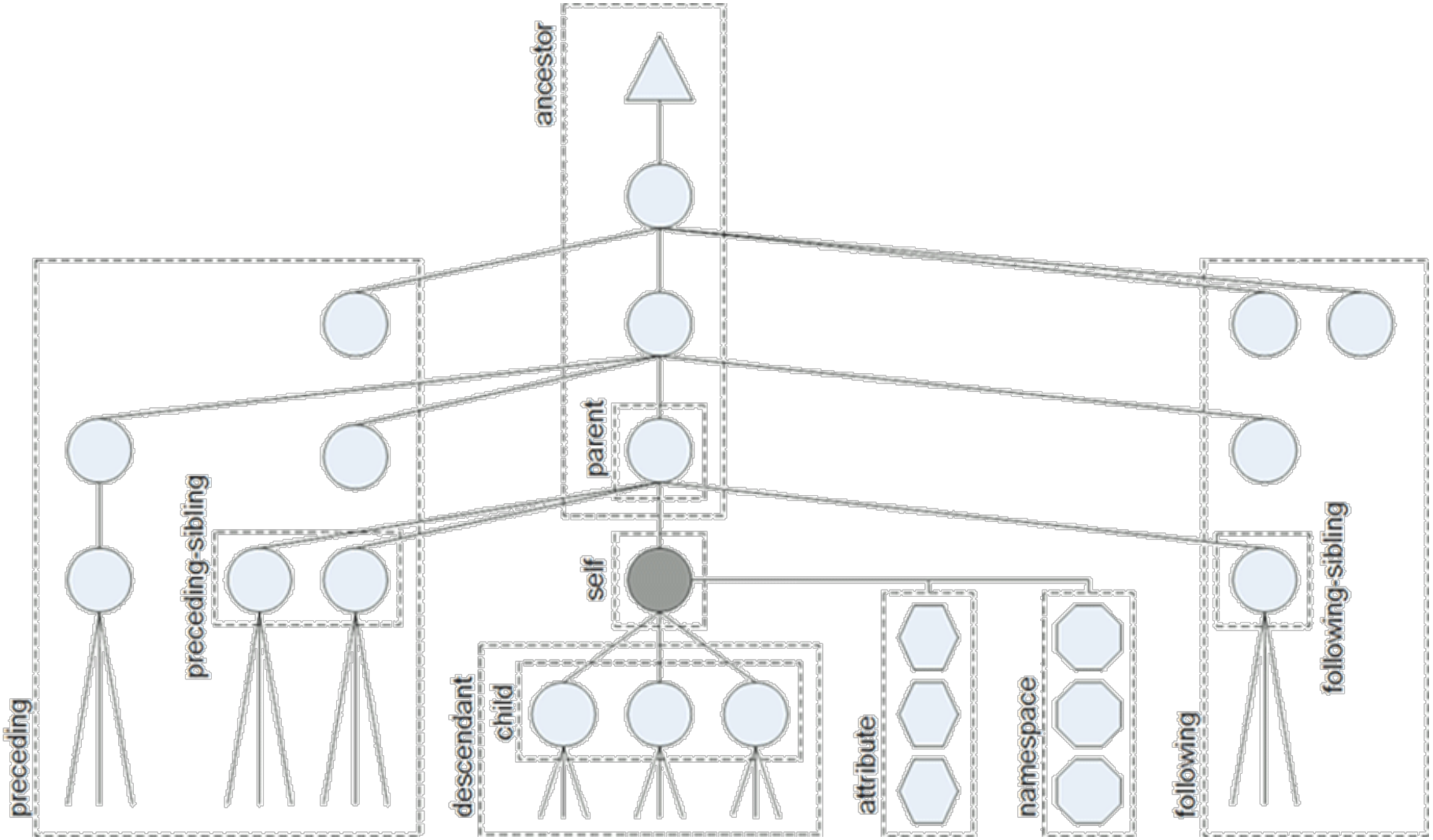
Identifies the nodes that precede the context node in document order, excluding its attributes and ancestors

The PRECEDING SIBLING Axis



Identifies the nodes that precede the context node in document order that have the same parent

All the Node Axes



Path Expressions

- Each Path Expression consists of Location Steps separated by "/" that specify:
  - ... the direction taken by the step (the "Axes")
  - ...the type and number of nodes selected
  - ... additional filters for selecting specific nodes ("Predicates")
- // means start at the root of the tree
- The Child access is the default

XPath Location Examples

- Find the slides in the lecture: //slide
- Find their titles: //slide/title
- Find the title of the first slide: //slide[1]/title

Path Expressions

- Each Path Expression consists of Location Steps separated by "/" that specify:
- ... the direction in which each step moves (the "Axes")
- ...the type and number of nodes selected by the step
- ... additional filters for selecting specific nodes ("Predicates")

Location Path Processing

- Start with a given context
- For each location step:
  - Based on the axis, select the nodes on the axis
  - Reduce the set to the nodes that satisfy the node test
  - Apply the selection predicate(s) to further filter the node set
  - Take the remaining node set as the context for the next location step

Tree In, Selection Out

- XPath also has numerous arithmetic, logical, and string operators and many built-in functions
- The result of the XPath evaluation is a selection
  - //img[not(@alt)] → select all images which have no alt attribute
  - count(//img) → return the number of images
  - /descendant::img[3]/@src → return the third image's src URI

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

10 October 2013
Lecture 13.1 – The Architectural Perspective on Relationships

The Architectural Perspective: Degree or Arity

- The architectural perspective embodies Kent's definition of "relation" as "A sequence of categories, that includes one thing from each category"
- The DEGREE or ARITY of a relationship is the number of different "entity types" or "resource categories" in the relationship
  - Husband is-married-to Wife is BINARY
  - Person is-married-to Person is UNARY

The Architectural Perspective: Cardinality

- The CARDINALITY is the number of instances that can be associated with each entity type
  - Husband is-married-to Wife is ONE-TO-ONE, because husbands have only one wife and vice versa (in monogamous societies)
  - Father is-parent-of Child is ONE-TO-MANY
  - Homer is-parent-of Bart AND Lisa AND Maggie is one-to-three

Modeling Relationships as Binary Ones

- Relationships can always be modeled as binary ones, but this makes some relationships implicit that were explicit
- Binary relationships are relationship "triples" with a "subject", "predicate," and "object"
- With binary relationships the reason for the relationship can often be interpreted in both directions (one is the inverse of the other)
- With triples we can combine relationships into a graph and "reason" over the set of relationships when they have common components

The Architectural Perspective: Directionality

- The DIRECTIONALITY of a relationship defines the order in which the arguments of the relationship are connected
- A ONE-WAY or UNI-DIRECTIONAL relationship can be followed in only one direction
- A BI-DIRECTIONAL one can be followed in both directions
- All symmetric relationships are bi-directional, but not all bi-directional relationships are symmetric

The Architectural Perspective {and,or,vs.} the Structural Perspective

- The architectural perspective is abstract and prescriptive
  - It defines what kinds of relationships can be created
- The structural perspective is concrete and descriptive
  - It says "this is what exists" and describes the actual patterns of association, arrangements, proximity, or connection between resources"

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

10 October 2013
Lecture 13.2 – Introduction to
Describing Structure Between Resources

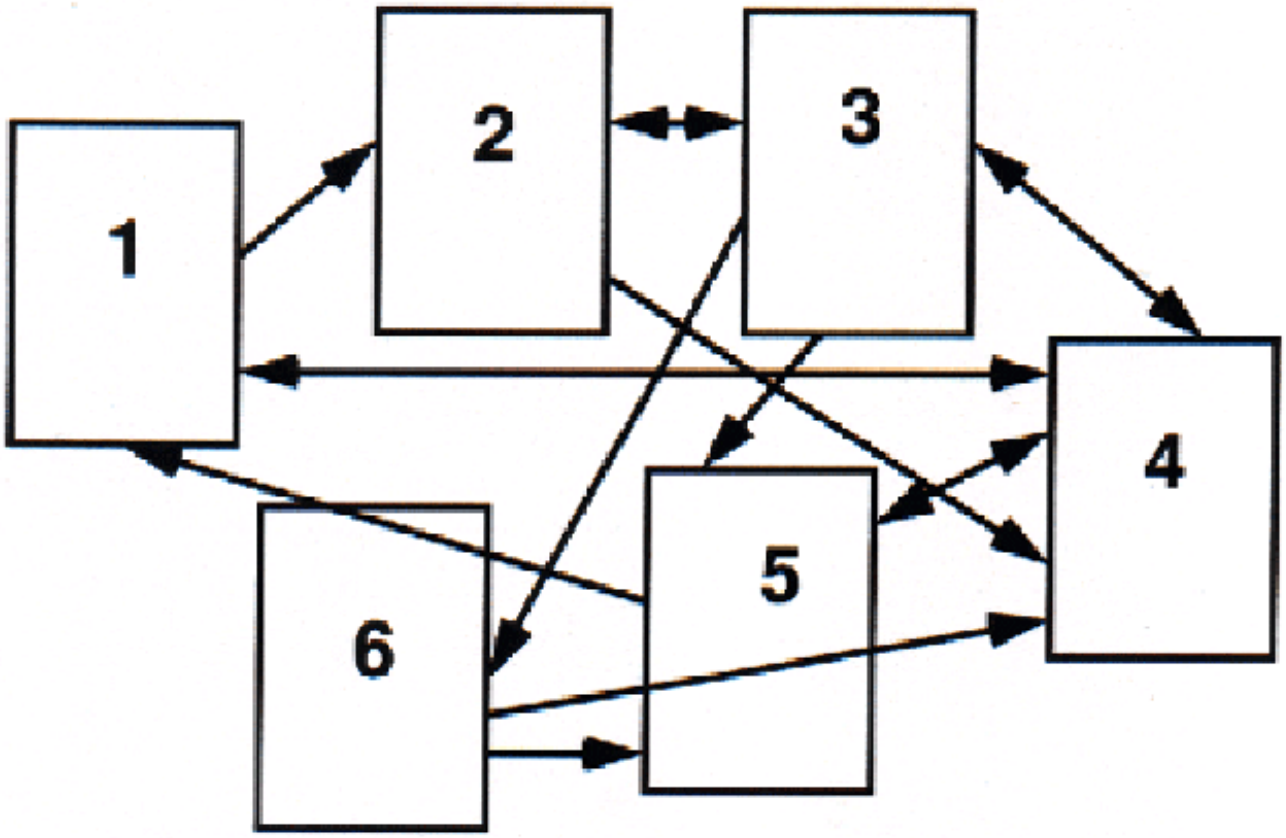
Between-Resource Structure

- Links between printed or digital documents – citations, cross refs, notes
- Links between web pages
- Links – communication and information flows - between people, organizations, any other kind of interacting “actors” or resources – social networks
- We can analyze all of these with some common concepts and abstractions, which we will introduce in as gentle a way as possible

Links Between Documents

- We can distinguish:
  - The starting point of the link (the ANCHOR)
  - The end point of the link (the DESTINATION)
  - How the starting point of the link is presented (the LINK MARKER)
  - How (if at all) the reason for the link is indicated (the LINK TYPE)

Link Network – Graphical View



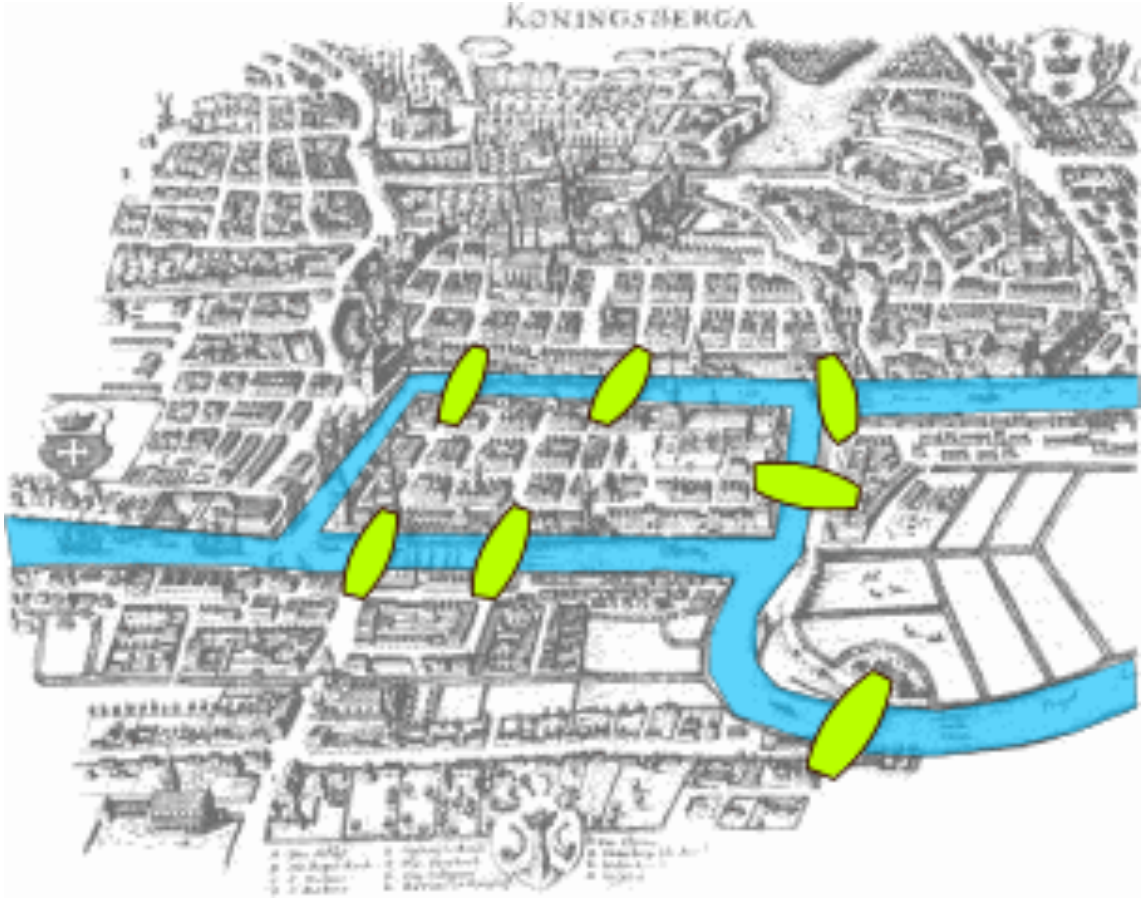
Link Network – Matrix View

	1	2	3	4	5	6
1		x				
2			x	x		
3		x		x	x	x
4	x		x		x	
5	x			x		
6				x	x	

Graph Theory

- We can apply graph theory to understanding relationships from a structural perspective
  - A GRAPH treats resources as VERTICES or NODES
  - The pairwise relationships between resources are represented by the EDGES that connect them
  - If the edges have an associated direction, this is a DIRECTED graph
  - A WEIGHT can be assigned to each edge if the relationship has a numerical aspect (distance, cost, time, etc.)

The Origins of Graph Theory (Euler 1735)



- 2 islands, 7 bridges – can you visit all 4 land masses without crossing any bridge more than once? Euler (1735) proved you couldn't and invented much of graph theory to explain it

Euler's Seven Bridges of Königsberg Problem

Computing the Properties of Graphs

- Reachability – is there a path between any two nodes in the graph?
- Shortest path – if there are multiple paths between two nodes, which is the shortest?
- Centrality – which nodes are the most connected or have the average shortest paths to the other nodes?
- Subgraph discovery – are there sub-graphs that are completely contained in a larger graph?

Reachability and Transitive Closure

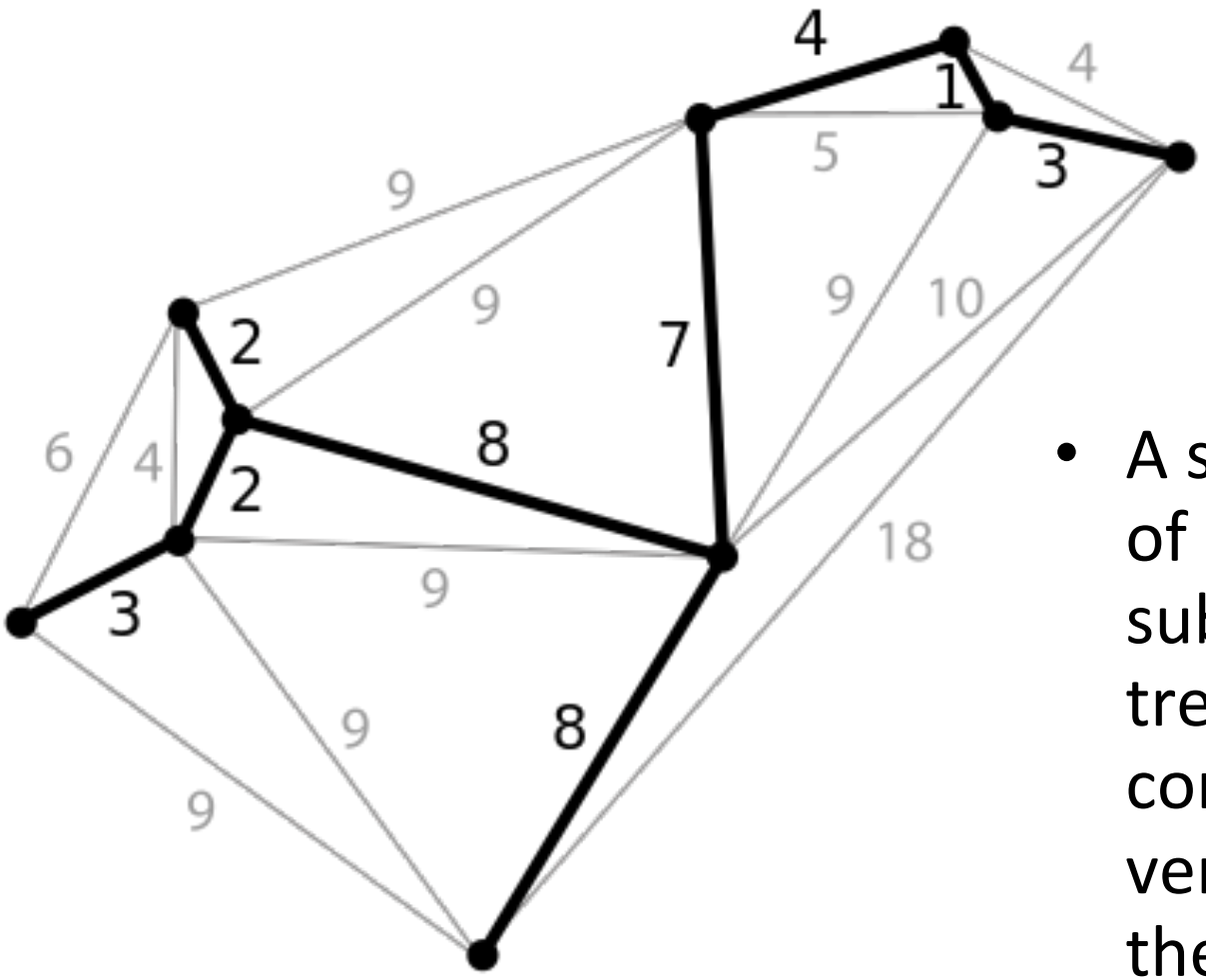
- The REACHABILITY property of a graph is "can you get there from here”
- We can determine whether a path exists between any two nodes in a graph by calculating the transitive closure of the graph; the most commonly used approach is Warshall's algorithm

http://www.cs.sunysb.edu/~skiena/combinatorica/animations/graphpower.html

Shortest Path

- The shortest path between two nodes in a graph can be calculated using Dijkstra's algorithm
- The shortest path problem is ubiquitous and its solution has obvious value – travel, transportation, financial arbitrage, grocery shopping…
- If the graph edges aren't weighted and we only care about connectivity, we have the familiar "degrees of separation" situation
- For mathematicians; for actors

Minimum Spanning Tree



- A spanning tree of a graph is a subgraph that is a tree that connects all the vertices – what's the shortest one, and why would we care?

Moving Beyond "Reachability"

- We've been treating relationships in purely structural terms - is one thing connected to another - but we can refine that into two perspectives:
- RELATIONAL analysis treats links as indicators of the amount of connectedness or the direction of flow between documents, people, groups, journals, disciplines, domains, organizations, or nations
- EVALUATIVE analysis treats links as indicators of the level of quality, importance, influence or performance of documents, people, groups

Analysis of Large-Scale Social & Information Networks (Kleinberg)

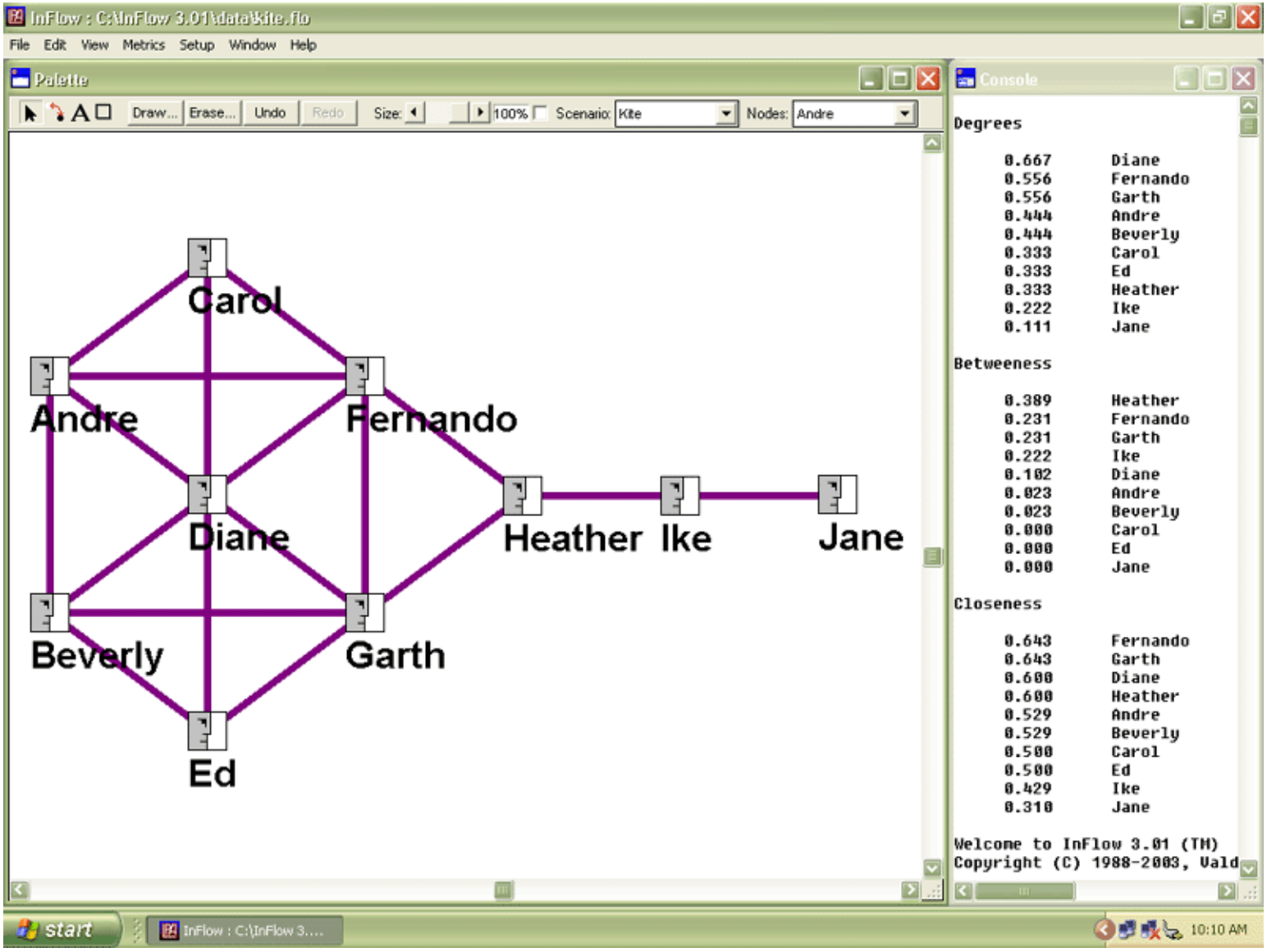
- "The Web lets us observe, measure and analyse social interaction at a level of scale and resolution that was previously unimaginable, observing social phenomena that had previously remained essentially unrecorded and invisible"
- Social social network firms exploit "triadic closure" – the increased tendency for two people to form a relationship when they people in common
- People who are the boundaries of one's social networks – "weak ties" – can be crucial sources of information

Analysis of Large-Scale Social & Information Networks (Kleinberg)

- Many social networks are directed graphs because of asymmetries in status, power, or values
- Identifying "influencers" and predicting their influence is the billion dollar question for web-based social networks

- If you liked this paper, you'll love 203 next semester. If you didn't like this paper, never mind
- See Coursera course on Social Network Analysis

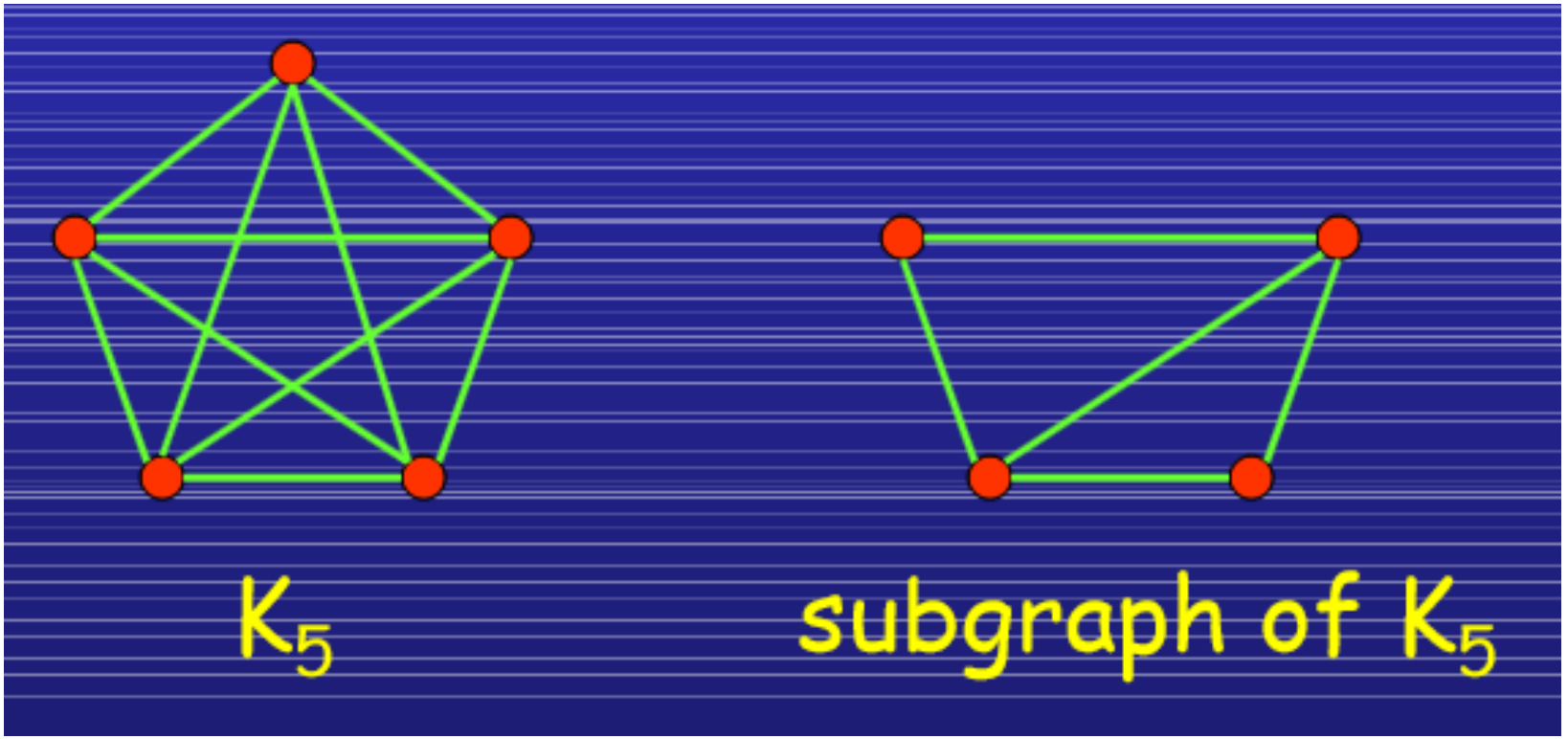
“Kite Network” Example from http://www.orgnet.com/sna.html



Measures of Centrality

- Degree Centrality – number of direct connections
- Betweenness – a location in the network that interconnects different clusters
- Closeness – low average path length

Subgraphs



In a social network setting, connection subgraphs can identify the few most likely paths of transmission for a disease (or rumor, or information-leak, or joke) or spot whether an individual has unexpected ties to any members of a list of individuals.

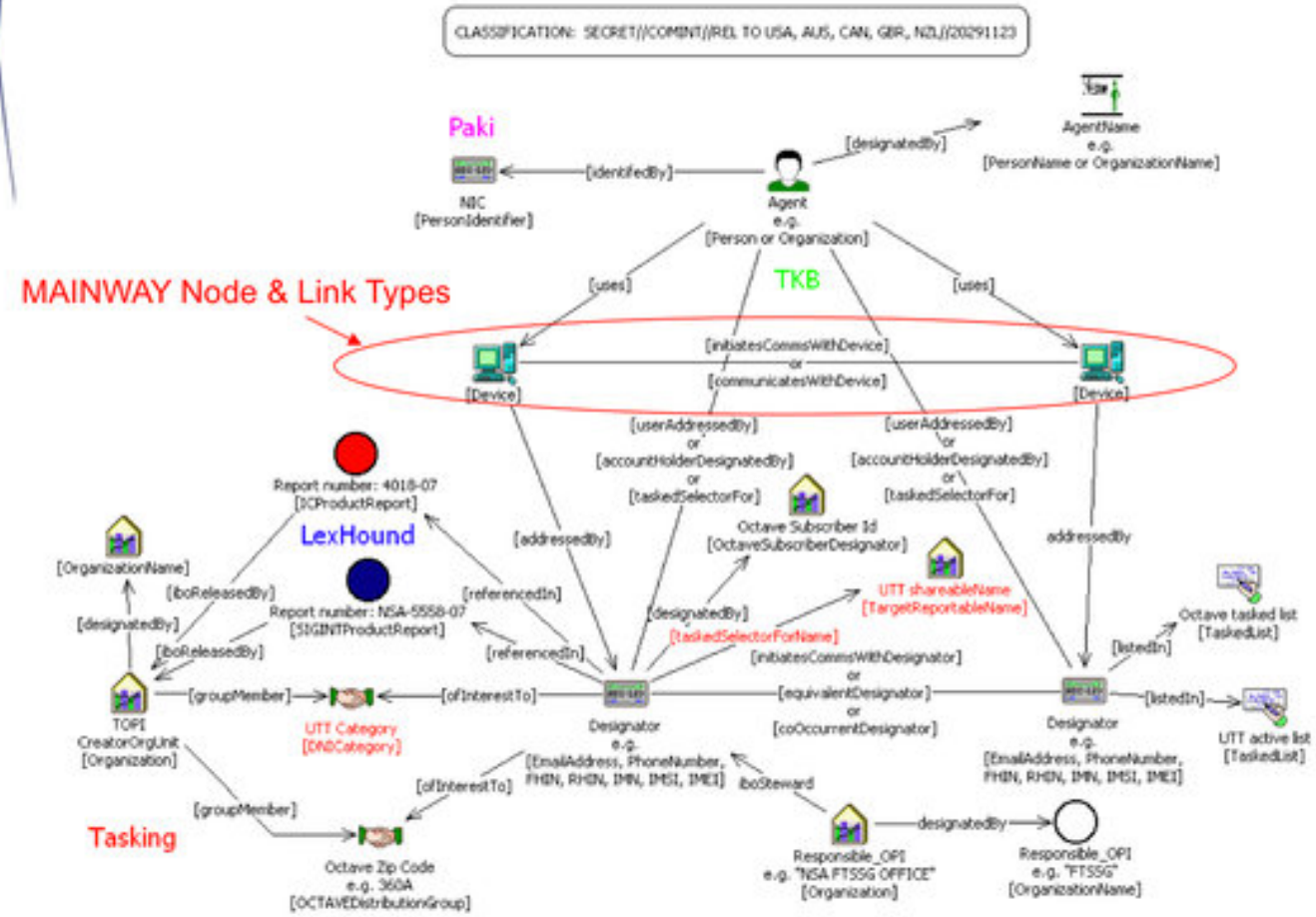
Faloutsos, Christos, Kevin S. McCurley, and Andrew Tomkins. "Connection subgraphs in social networks" 2004.

NSA's Social Network Analysis

- The agency was authorized to conduct "large-scale graph analysis on very large sets of communications metadata without having to check foreignness" of every e-mail address, phone number or other identifier, the document said.
- The agency can augment the communications data with material from public, commercial and other sources, including bank codes, insurance information, Facebook profiles, passenger manifests, voter registration rolls and GPS location information, as well as property records and unspecified tax data, according to the documents.

NSA Gathers Data on Social Connections of US Citizens.
NY Times 28 Sept 2013

(S//SI//REL USA, FVEY) SYANPSE Data Model



- “The Web lets us observe, measure and analyse social interaction at a level of scale and resolution that was previously unimaginable, observing social phenomena that had previously remained essentially unrecorded and invisible”

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

10 October 2013
Lecture 13.3 – Bibliometrics and Altmetrics

Citation Signals and Polarity (1)

- When one resource cites another there is often a lexical signal that indicates how a writer views the relationship of a citation to the text from which the citation is made
- This concept comes from rhetoric and critical theory but is relevant to document engineering
- Pedantic folks call this "genre of invocation" -- distinguishing things like "cite, mention, acknowledged”

Citation Signals and Polarity (2)

- A citation or link without a signal suggests by default that the citation supports the current text
- Explicit signals that indicate positive polarity include "See," "See also," "See generally," and "Cf."
- Signals that indicate negative polarity include "But see" and "Contra"

Bibilometrics (or "Scientometrics"): Structure of Scientific Citation

- Analysis of scientific citation began in the 1920s as a way to quantify the influence of specific documents or authors in terms of their "impact factor"
- It can also identify "invisible colleges" of scientists whose citations are largely self-referential
- It can recognize the emergence of new scientific disciplines

Some Problems for Citation Analysis

- People lie, are lazy, stupid, biased...
- Most papers cite only a small proportion of the sources that influenced them
- Secondary sources are cited more than primary ones, because people don't know the literature, and informal ones aren't cited
- People cite their friends and themselves more than is justified

Altmetrics

- Measures based on total output are biased against young scholars; one of them proposed the H-Index as a better predictor of scientific impact
- The "Altmetrics" movement is trying to make non-traditional contributions count for academic evaluations
  - Publishing in "open publications" – measuring downloading
  - Sharing "raw science" like datasets, code, and experimental designs
  - Blogging, microblogging, and comments or annotations on existing work

Adapting Citation Analysis to the Web

- The concepts and techniques of citation analysis seem applicable to the web since we can view it as a network of interlinked articles
- But not everything applies because the web is different in numerous ways
- We'll return to this issue with web search and structure-based retrieval

Readings for Next Lecture

- TDO 5.8, 8.4
- Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." Scientific American 284, no. 5 (2001): 28-37.
- Heath, Tom, and Christian Bizer. "Linked data: Evolving the web into a global data space." Synthesis lectures on the semantic web: theory and technology 1, no. 1 (2011). Chapters 1 and 2, pages 1-29.
- Byrne, Gillian, and Lisa Goddard. "The Strongest Link: Libraries and Linked Data." D-Lib Magazine 16, no. 11/12 (2010).