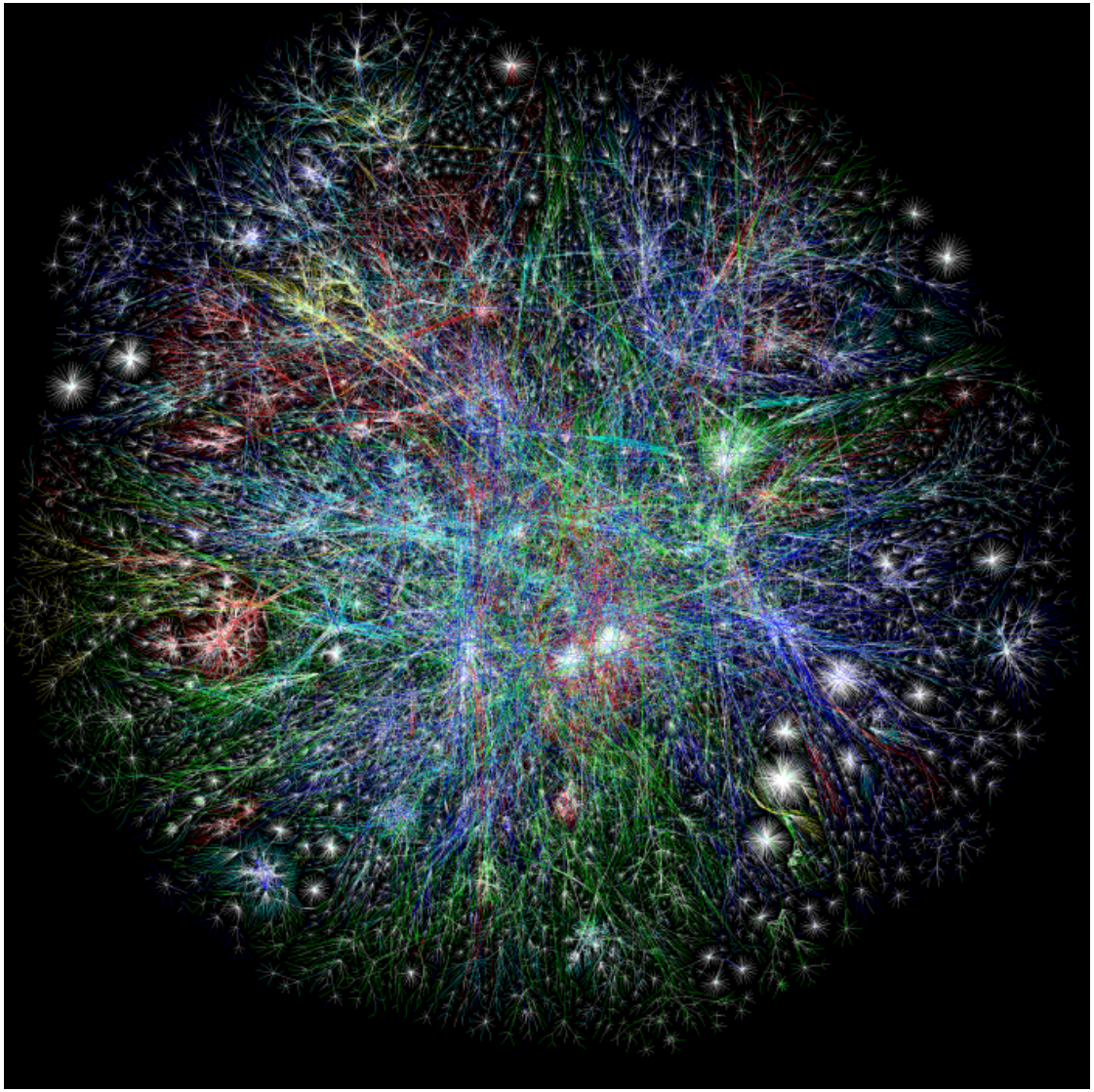
Structural Analysis of Networks

Russell Martin

NeST Hackathon

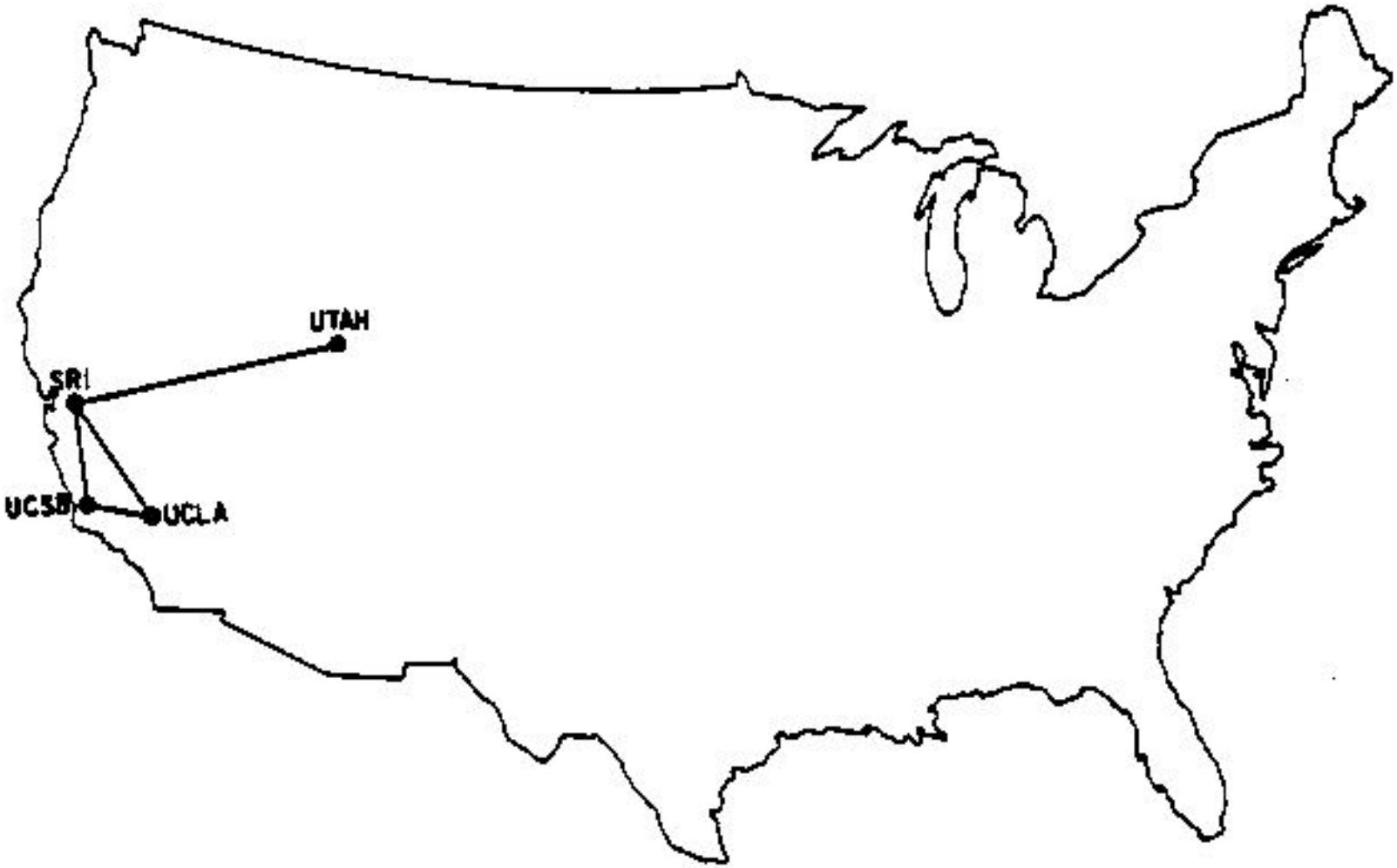
April 18-19, 2015

What is this?



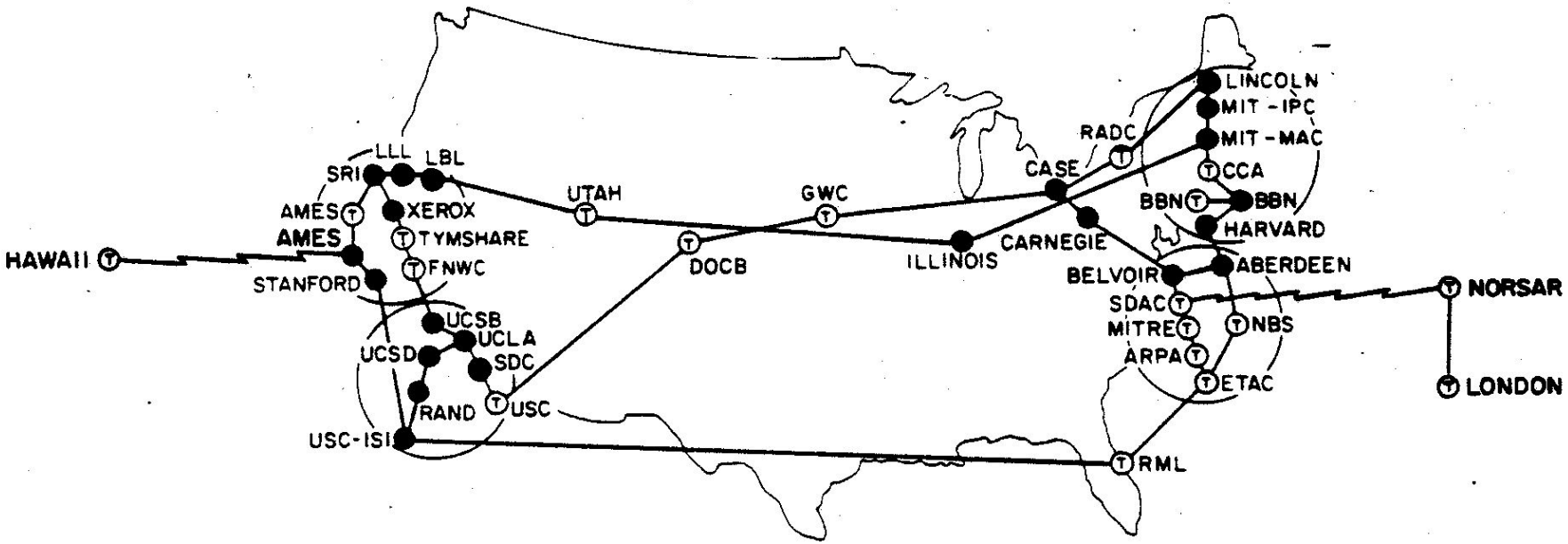
The Internet, circa 2003 (source: The Opte Project)

The ARPANET, Dec 1969 (Precursor to the Internet)



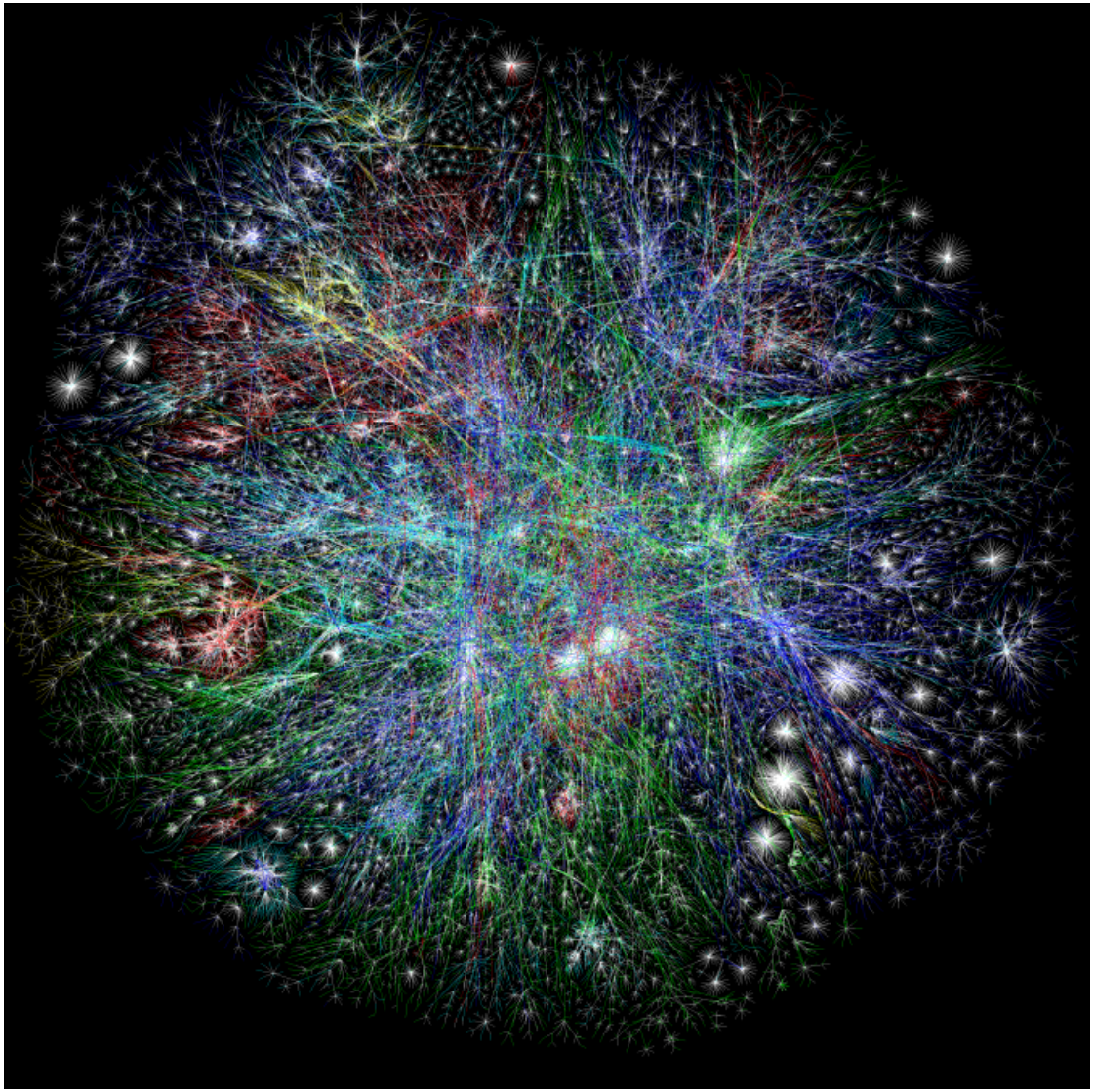
Source: F. Heart, A. McKenzie, J. McQuillian, and D. Walden, *ARPANET Completion Report*, Bolt, Beranek and Newman, Burlington, MA, January 4, 1978.

The ARPANET, Sept 1973



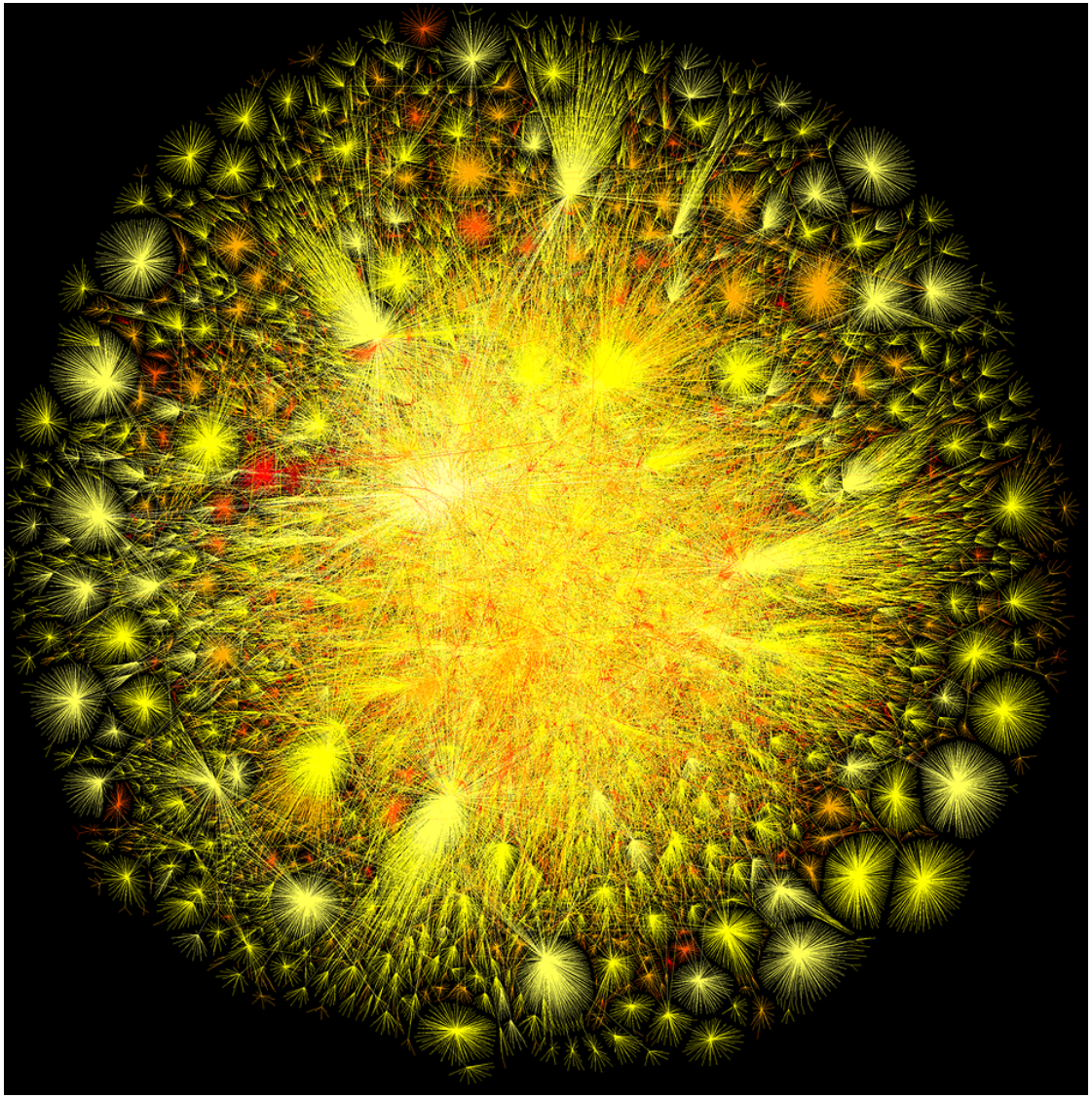
Source: F. Heart, A. McKenzie, J. McQuillian, and D. Walden, *ARPANET Completion Report*, Bolt, Beranek and Newman, Burlington, MA, January 4, 1978.

The Internet, circa 2003



(source: The Opte Project)

The Internet, circa 2010



Source: The Opte Project

Who? What? How?

How do we make sense of large networks?

Who or what is important in these networks?

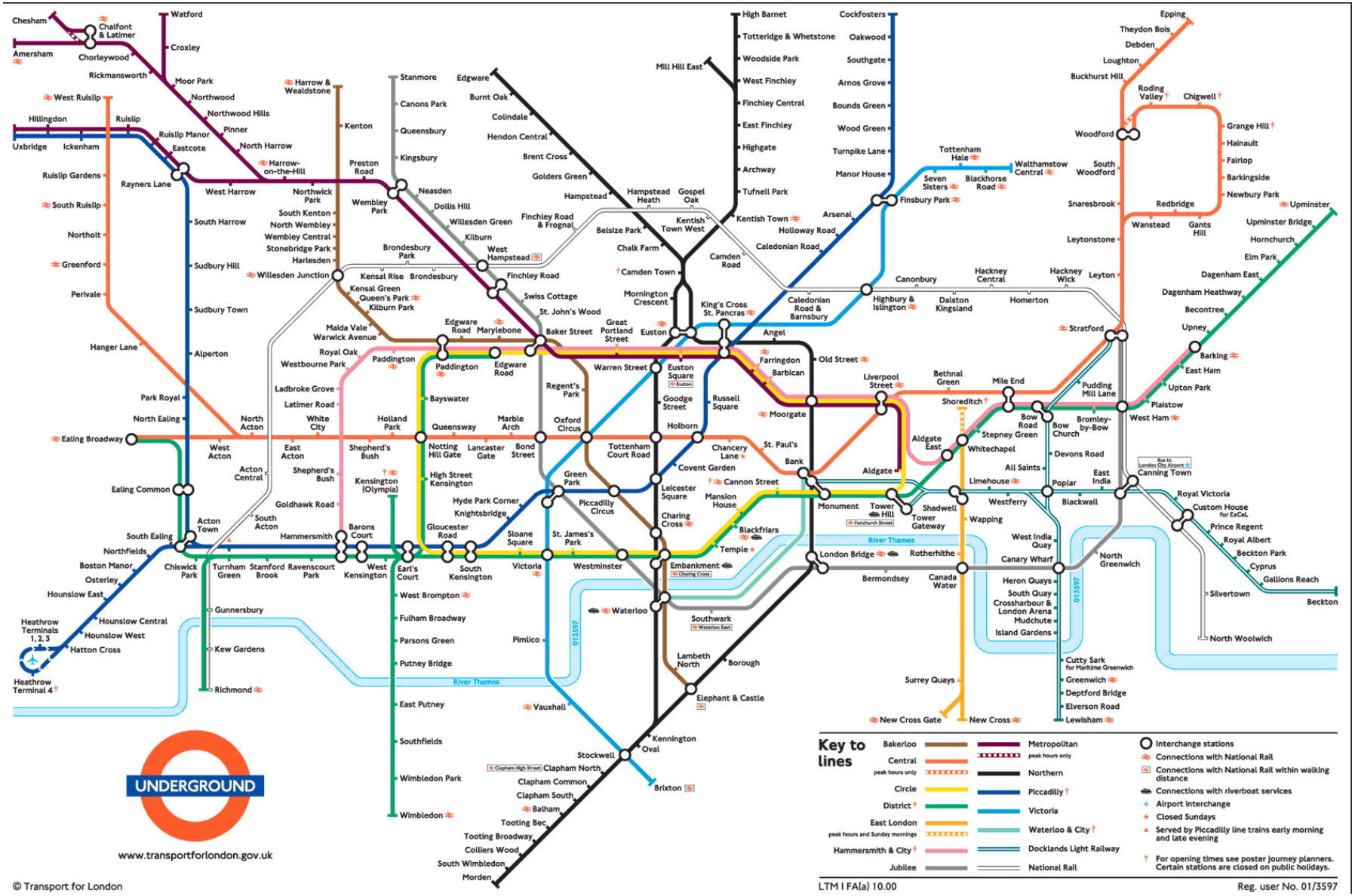
What is meant by "important"?

''Importance" will depend upon what we want to know, and (probably) where the network comes from.

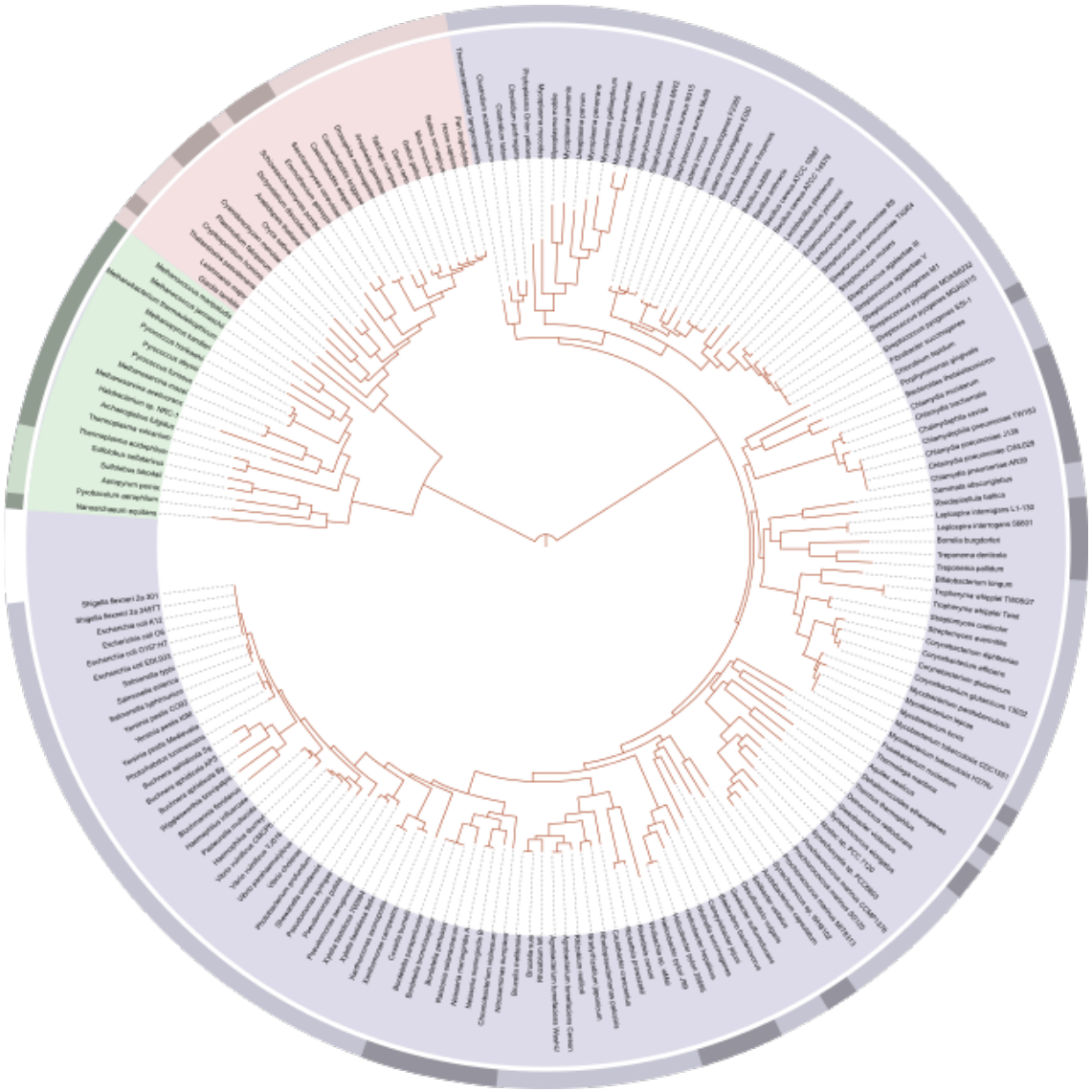
Networks are everywhere

- Transportation networks
  - Road, train, and shipping networks
  - Airline traffic
  - London Underground and other subway systems
- Technological networks
  - The Internet
  - The World Wide Web
  - Phone system (landline and/or mobile)
  - Electrical power supply grid
- Biological networks
  - Protein-protein interactions
  - Nervous system
  - Animal migration patterns
  - Evolutionary relationships
- Social networks
  - Facebook, Twitter, LinkedIn
  - Spread of disease
  - Who knows whom (Stanley Milgram's famous "sixth degrees of separation" social experiments in the 1960s)

London Tube Map

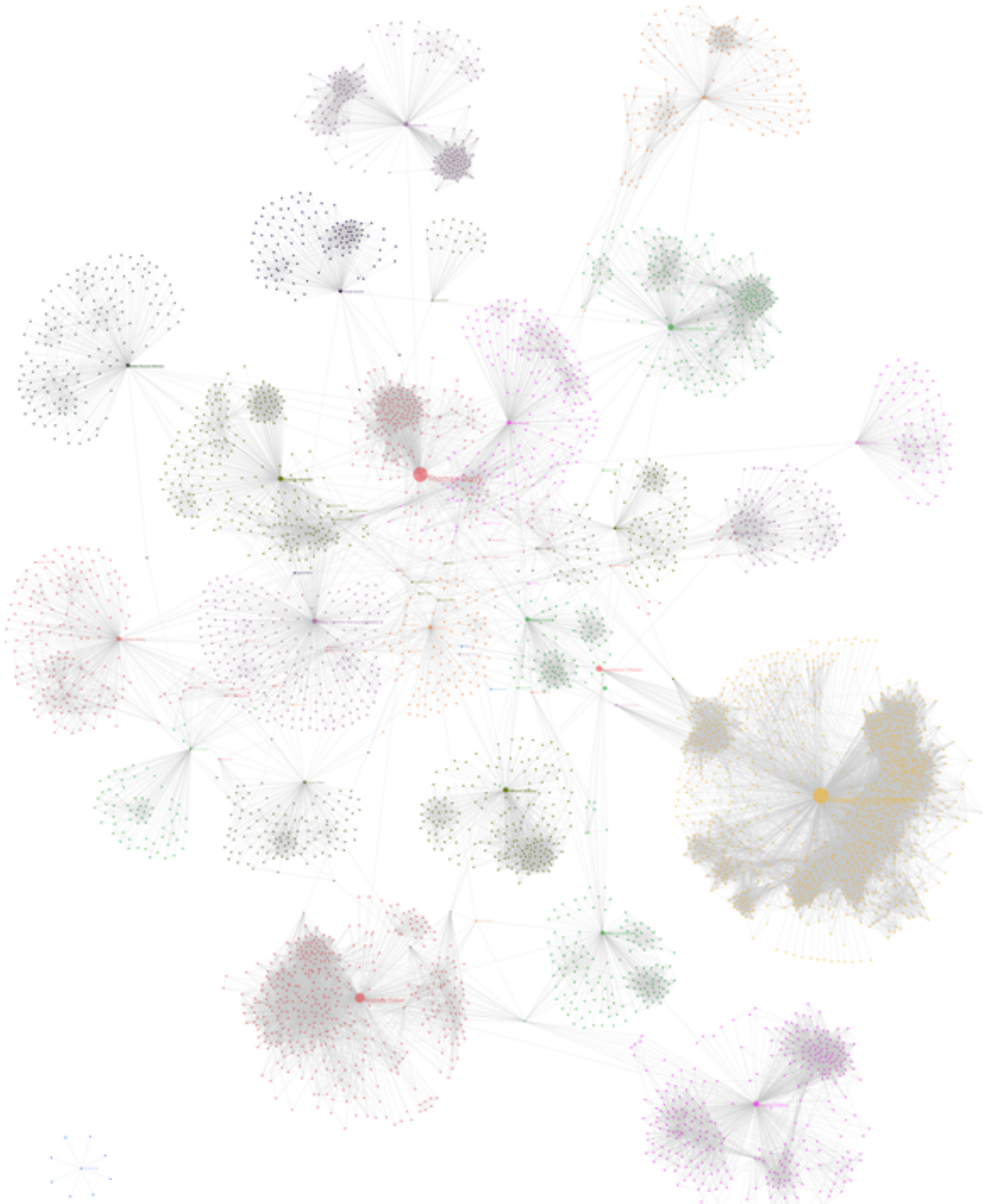


Phylogenetic tree of life



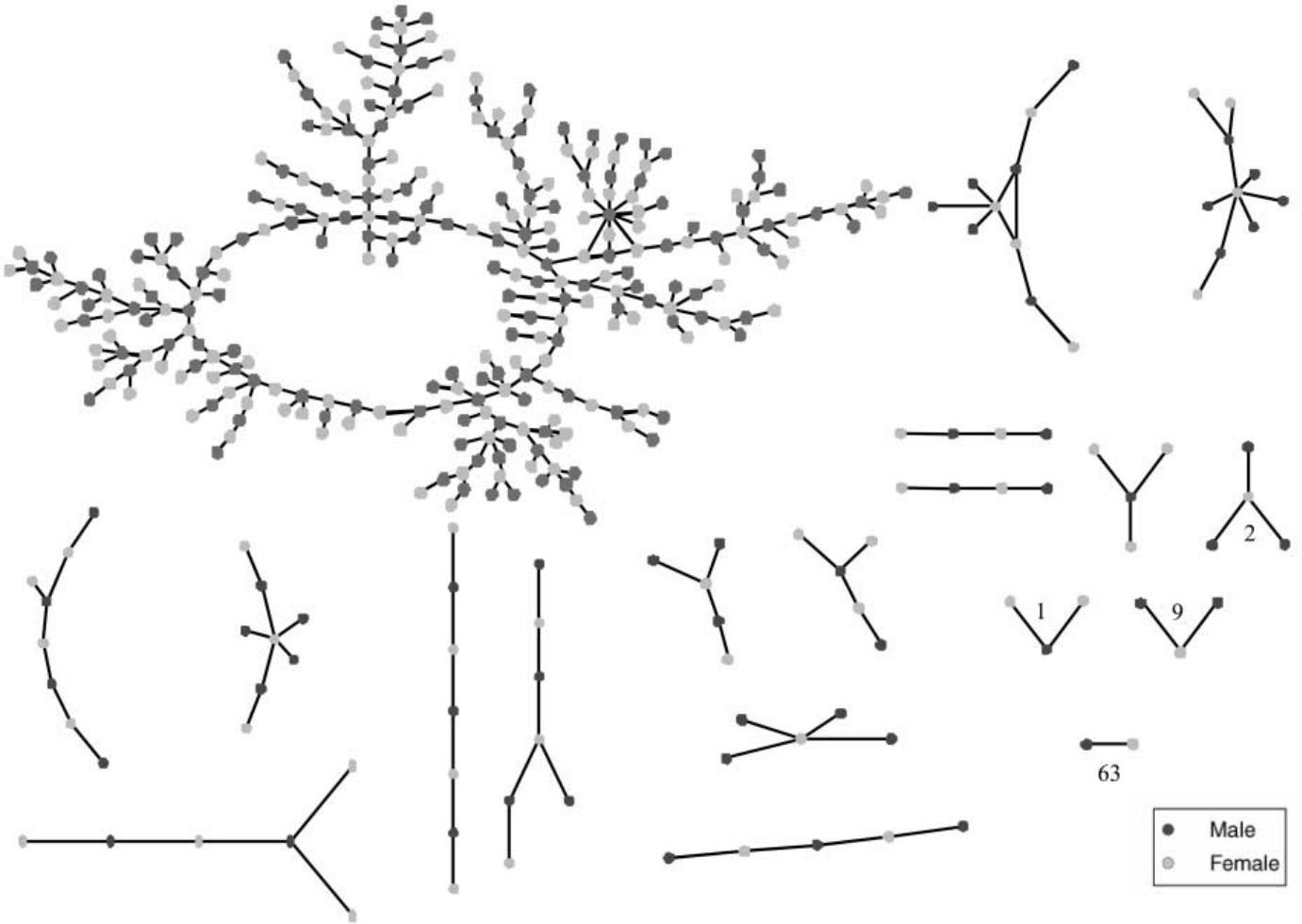
Author: Ivica Letunic Retraced by Mariana Ruiz Villarreal

Social network example (Facebook)



Social network example

Romantic relationships in a Chicago-based high school (18 months over 1993-1995).



Source: P. Bearman, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology* 110(1), 2004.

How do we make sense of these networks?

With large networks, pictures don't necessarily help us make sense of the structure.

How can we find what nodes (or edges) are important in the network?

What do we mean by "important"?

That meaning may depend upon the problem we are considering or the context of our study.

What other structural properties of networks are "interesting" or "informative" to us?

Structural analysis

I will discuss two measures that can be used to define "importance" of nodes in a network.

1. PageRank
2. Betweenness centrality

PageRank

PageRank, named for Larry Page (one of the founders of Google), is an algorithm that is used to rank websites in the search engine results of Google.

According to Google:

> *PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.*

PageRank (cont.)

A rough idea behind the PageRank algorithm is the following:

1. A user starts at any node (webpage) in the network.
2. Randomly select an outgoing link from that webpage, and go to the new webpage that the link points to.
3. Instead of step 2, every so often (how "often" needs to be defined), the user types in a new webpage address, selected at random from all possible webpages.
4. Goto step 2 and repeat.

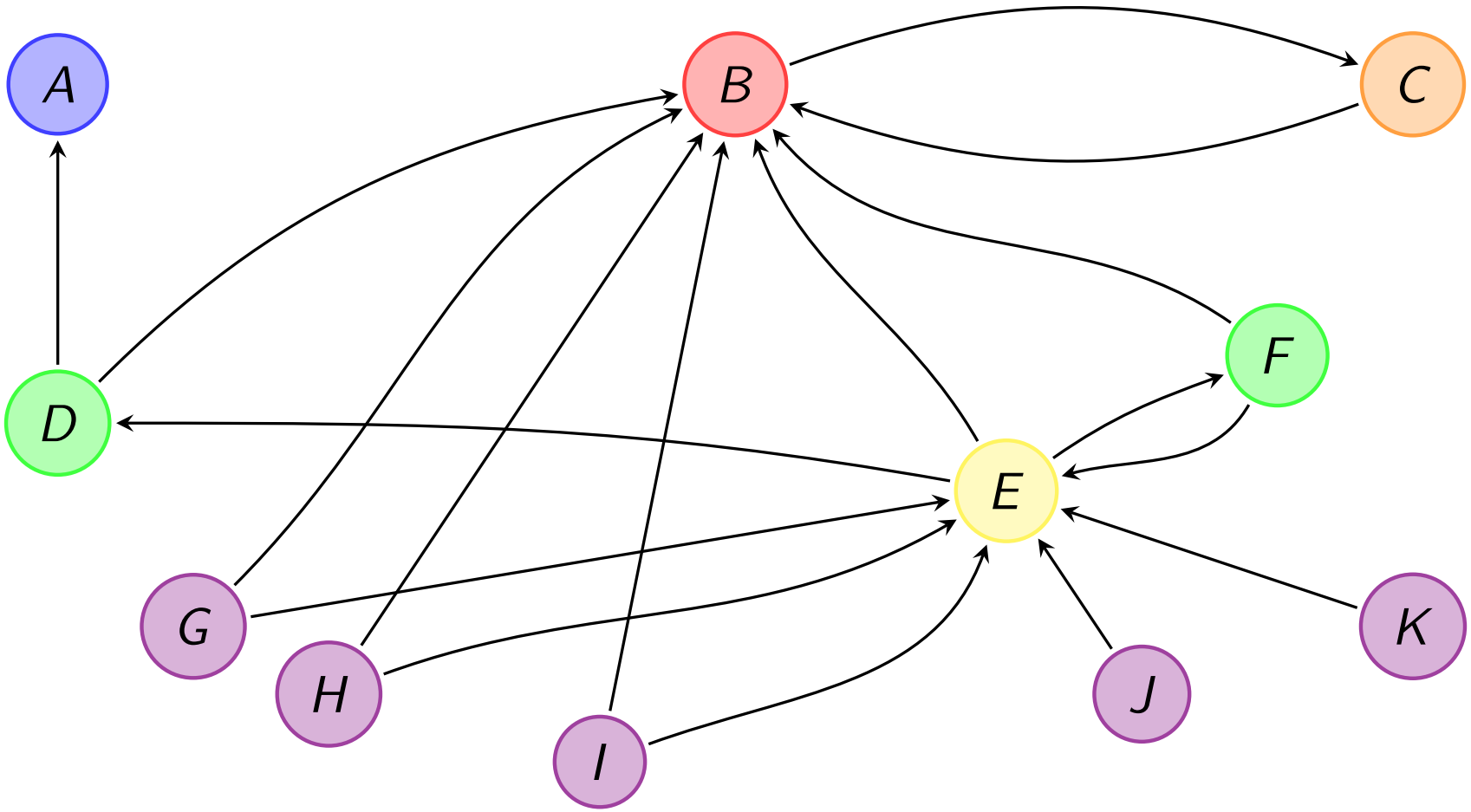
The PageRank of a webpage is the chance of landing on that webpage after a "large" number of steps of this procedure.

PageRank (cont.)

Somewhat more mathematically, PageRage can be defined as follows, in a recursive fashion:

The PageRank of a webpage is defined in terms of the number of links to that webpage, and the PageRank of all of those incoming links. A page that is linked to by many pages with high PageRank receives a high PageRank itself.

PageRank example



(Taken from Wikipedia PageRank description.)

PageRank example (cont.)

"Damping factor"

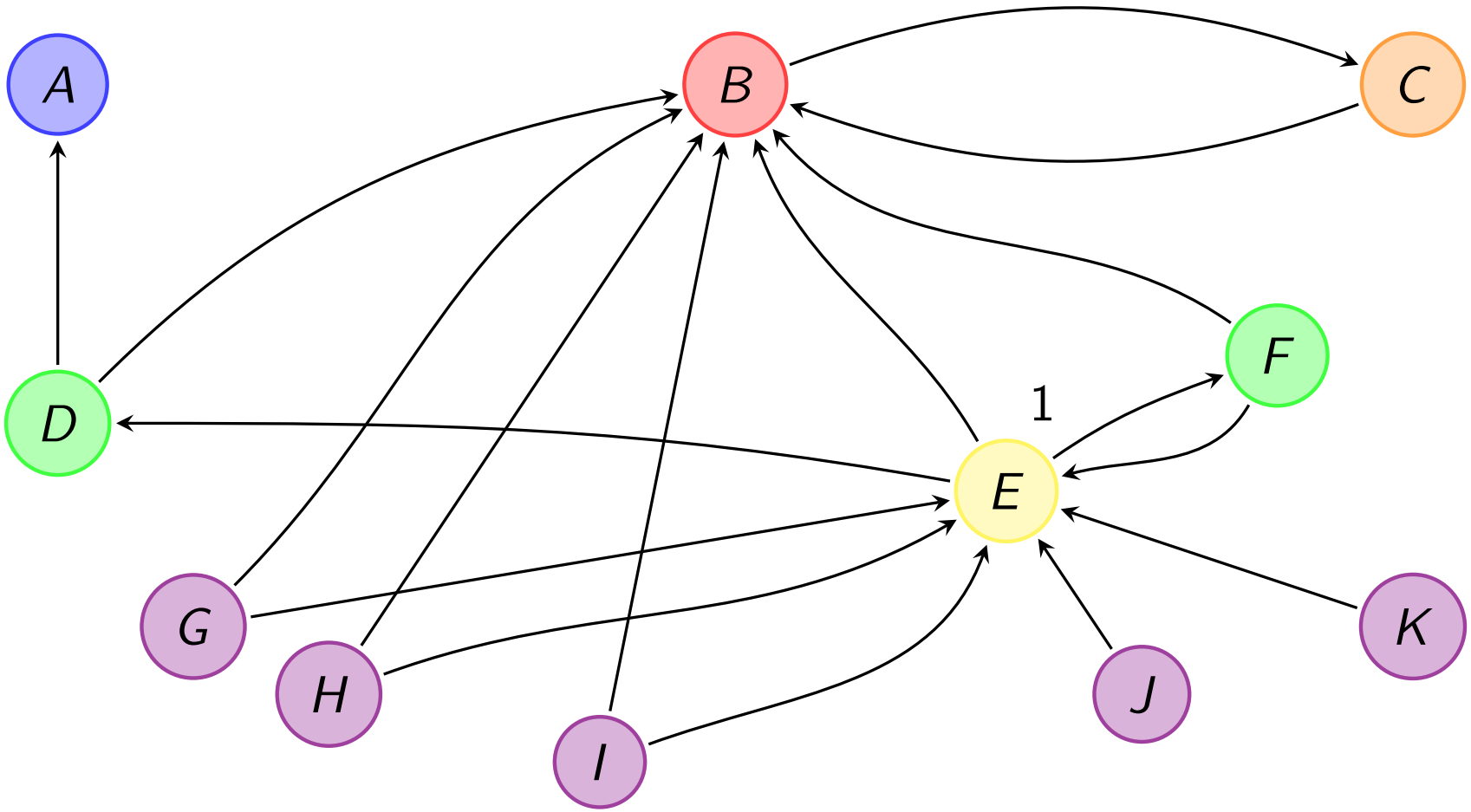
Suppose we take the *damping factor* to be equal to 85%.
This means that 85% of the time, a user follows a link from the current page he/she is viewing, i.e. he/she selects an outgoing edge at random from the current page.

The other 15% of the time, the user selects a node at random from the entire network.

(Note: If the current node has no outgoing edges, the user simply selects a node at random from the entire graph.)

Suppose that we start the process at node E, i.e. node E starts with PageRank $= 1$ and all other nodes have PageRank $= 0$.

PageRank example (cont.)



PageRank example (cont.)

Now we update the PageRank values according to the following method:

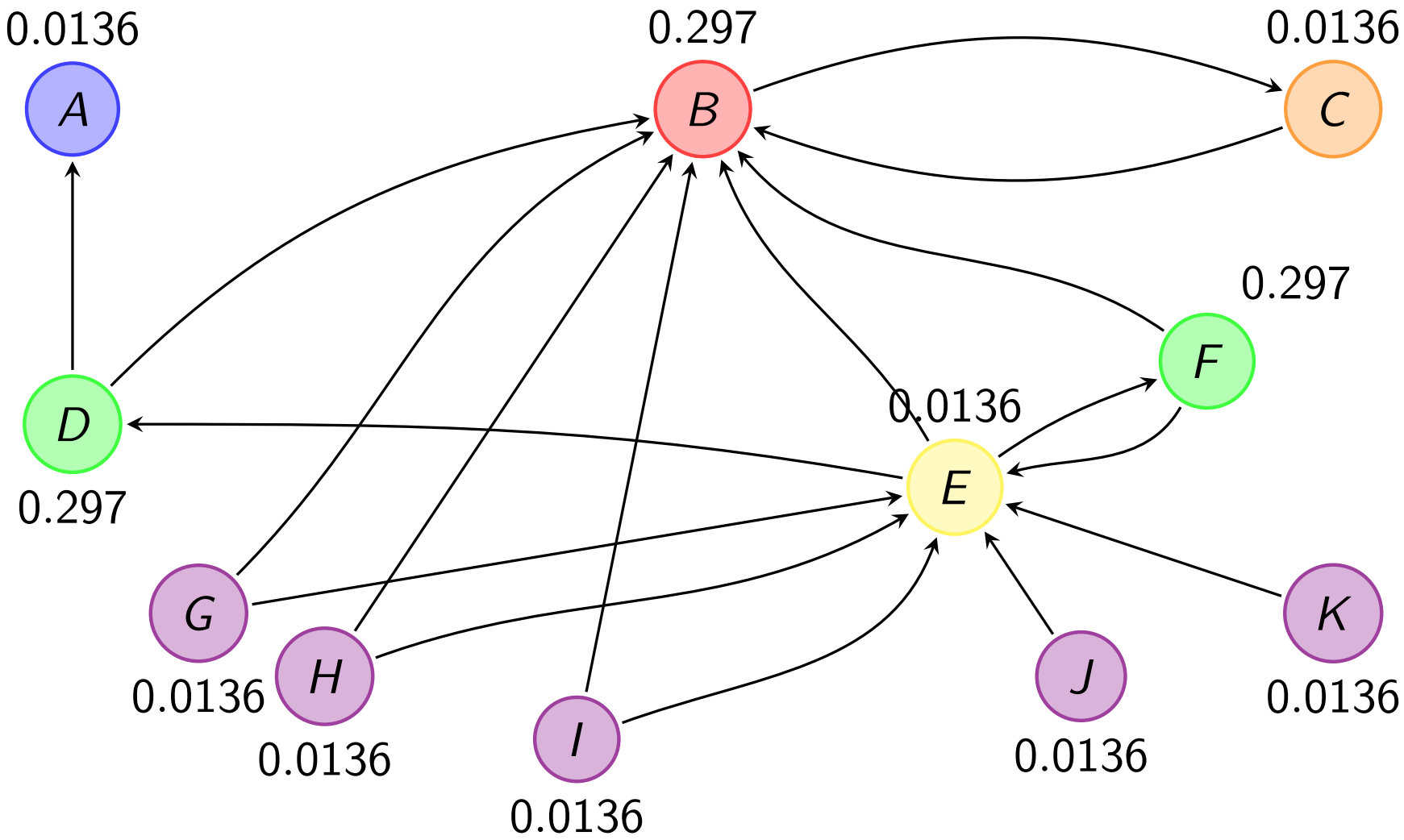
Because of the damping factor of 85%, each node receives a value of $(0.15) \cdot (1/11) \approx 0.0136$, since the network has 11 nodes in it.

Nodes B, D, and F, receive an additional value of $(0.85) \cdot (1) \cdot (1/3)$ from E.
This is because node E has three outgoing edges.

So nodes B, D, and F will have a new PageRank value of $(0.15) \cdot (1/11) + (.85) \cdot (1) \cdot (1/3) \approx 0.297$.

Each of the other nodes (including E) has a new PageRank value of $(0.15) \cdot (1/11) \approx 0.0136$.
The updated values are shown on the new graph.

PageRank example (cont.)



PageRank example (cont.)

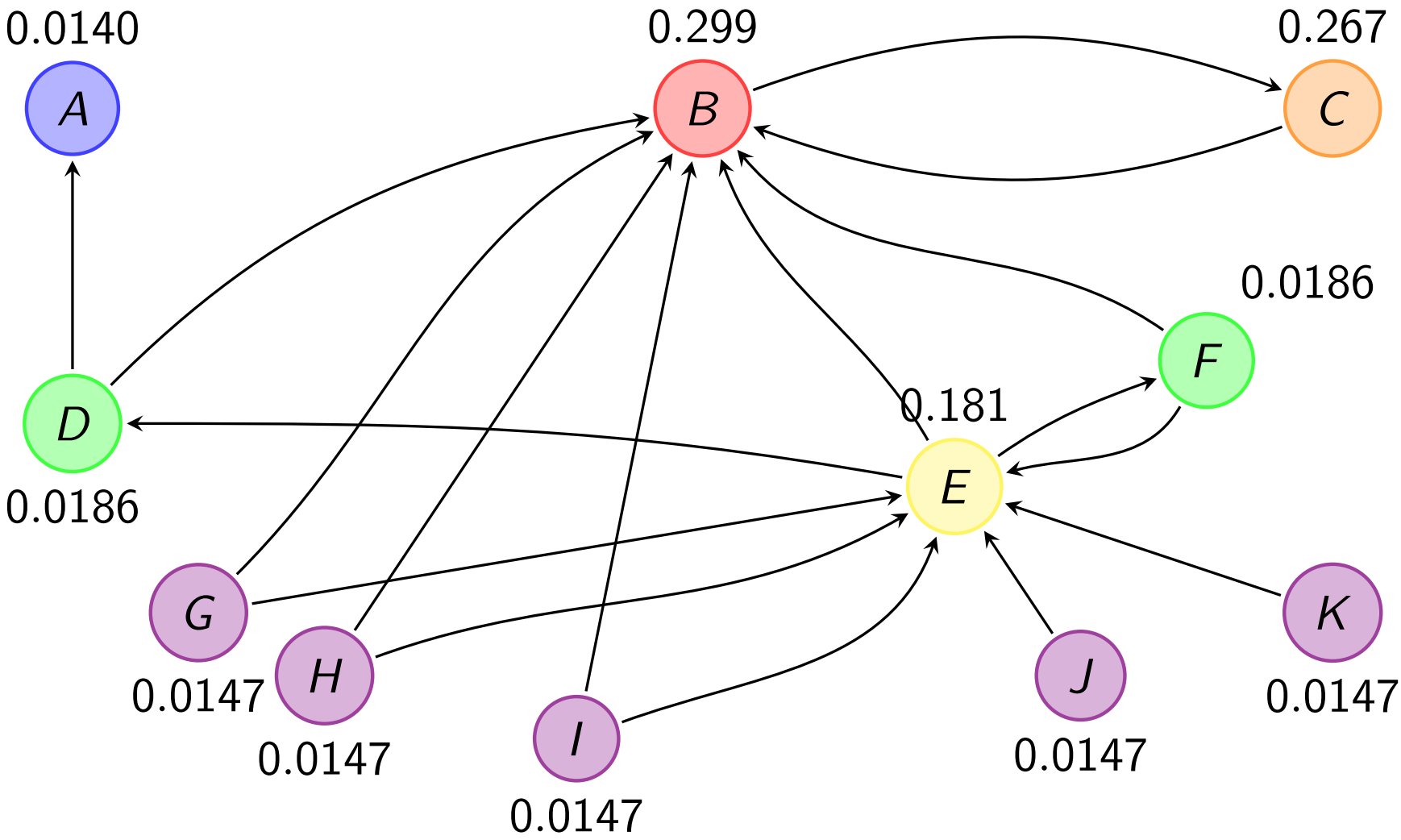
Using these new values of the PageRank, we re-calculate to again update the values of PageRank for each node.

For example, node A has new PageRank value of

$$(0.15) \cdot (1/11) + (0.85) \cdot (0.297)(1/2) \approx 0.140.$$

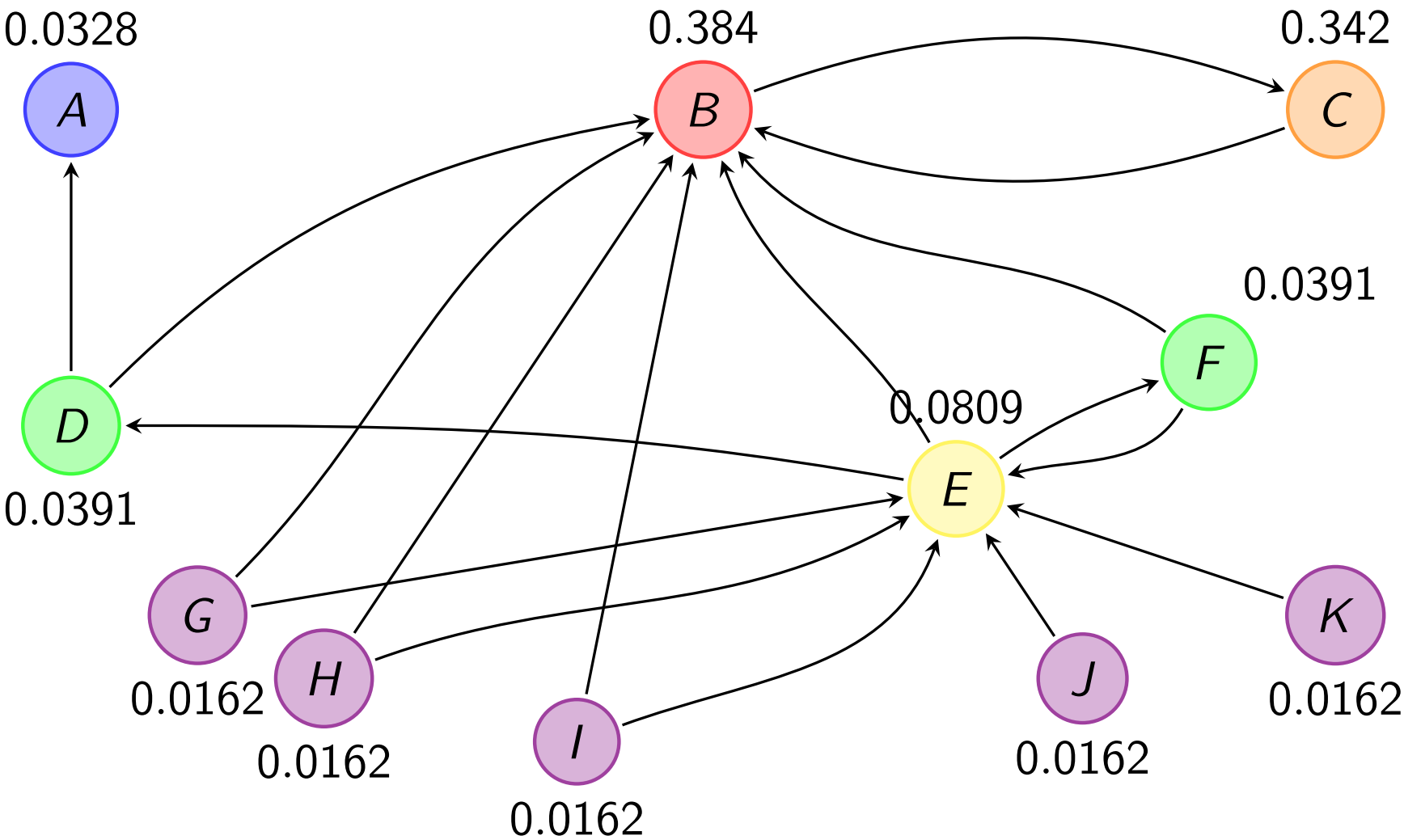
PageRank example (cont.)

After two steps (starting from E):



PageRank example (cont.)

Continue this process for a "large" number of steps, and the PageRank values converge:



Betweenness centrality

Betweenness centrality captures how much a node is "important" to a network, by measuring how many shortest paths go through that node.

Betweenness centrality has a large influence on the traffic and the transfer of items, information, news, gossip, etc through a network, under the assumption that the traffic will follow a shortest path between any two points.

The definition of betweenness centrality is attributed to sociologist Linton Freeman in 1977.

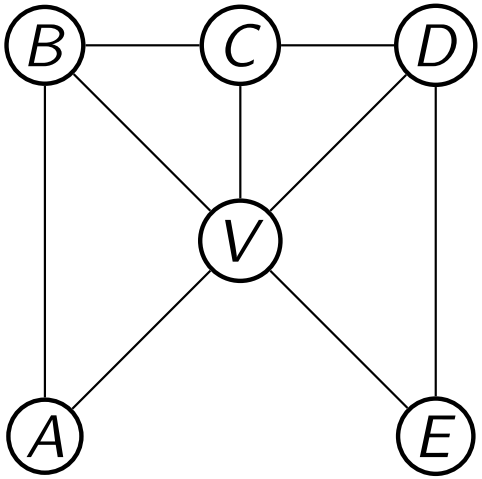
How to calculate it?

Betweenness centrality of a node v is given by the following expression:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those shortest paths that pass through v.

An example

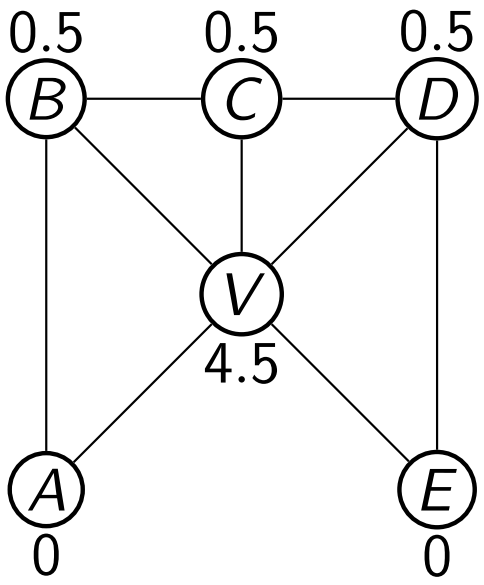


Consider all pairs not involving V.

A, B	One shortest path (not involving v)
A, C	Two paths (one using v)
A, D	One path (using v)
A, E	One path (using v)
B, C	One path (not involving v)
B, D	Two paths (one using v)
B, E	One path (using v)
C, D	One path (not involving v)
C, E	Two paths (one using v)
D, E	One path (not involving v)

So $g(V) =$
$0+1/2+1+1+0+1/2+1+0+1/2+0 = 9/2.$

An example (cont.)



We can calculate
the remaining betweenness centrality values
in a similar fashion (Brandes' algorithm
can be used to do this in time $O(|V| \cdot |E|)$).

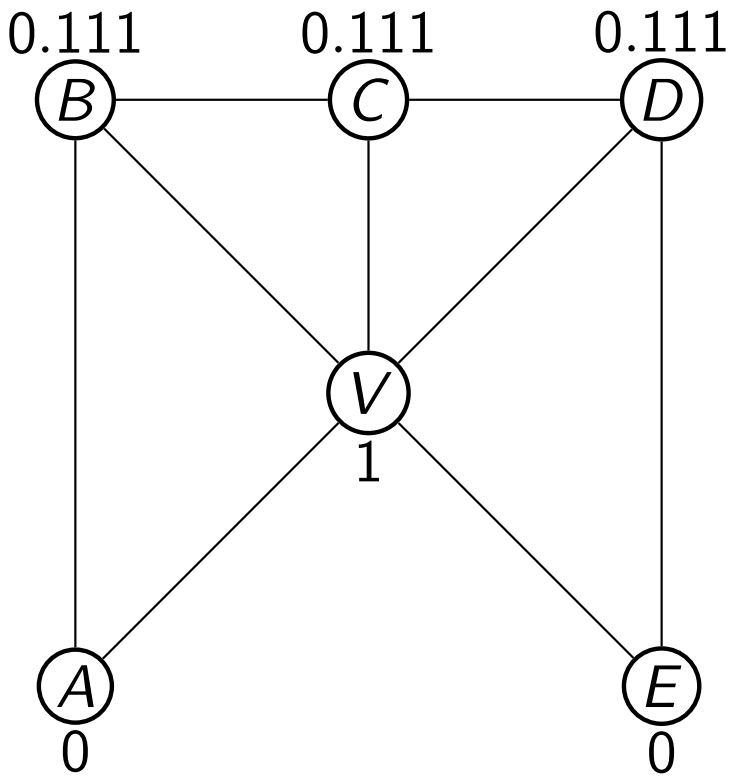
Often, the values are re-scaled so
that they are between 0 and 1, by defining

$$bc(v) = \frac{g(v) - min}{max - min}$$

where *min* is the minimum value of all the betweenness centrality values and *max* is the maximum.
This allows us to more meaningfully compare values amongst different networks (as otherwise the values depend upon the size of the network, i.e. the number of nodes).

An example (cont.)

Normalized betweenness centrality values



Other structural measures

- Degree sequence
- Diameter
- Triadic closure
- Clustering coefficients
- Homophily (how alike nodes are to their neighbours)
- Closeness centrality
- Eigenvector centrality
- Robustness (resistance to "overall" failure)
- ...
- ...
- Visualization of networks