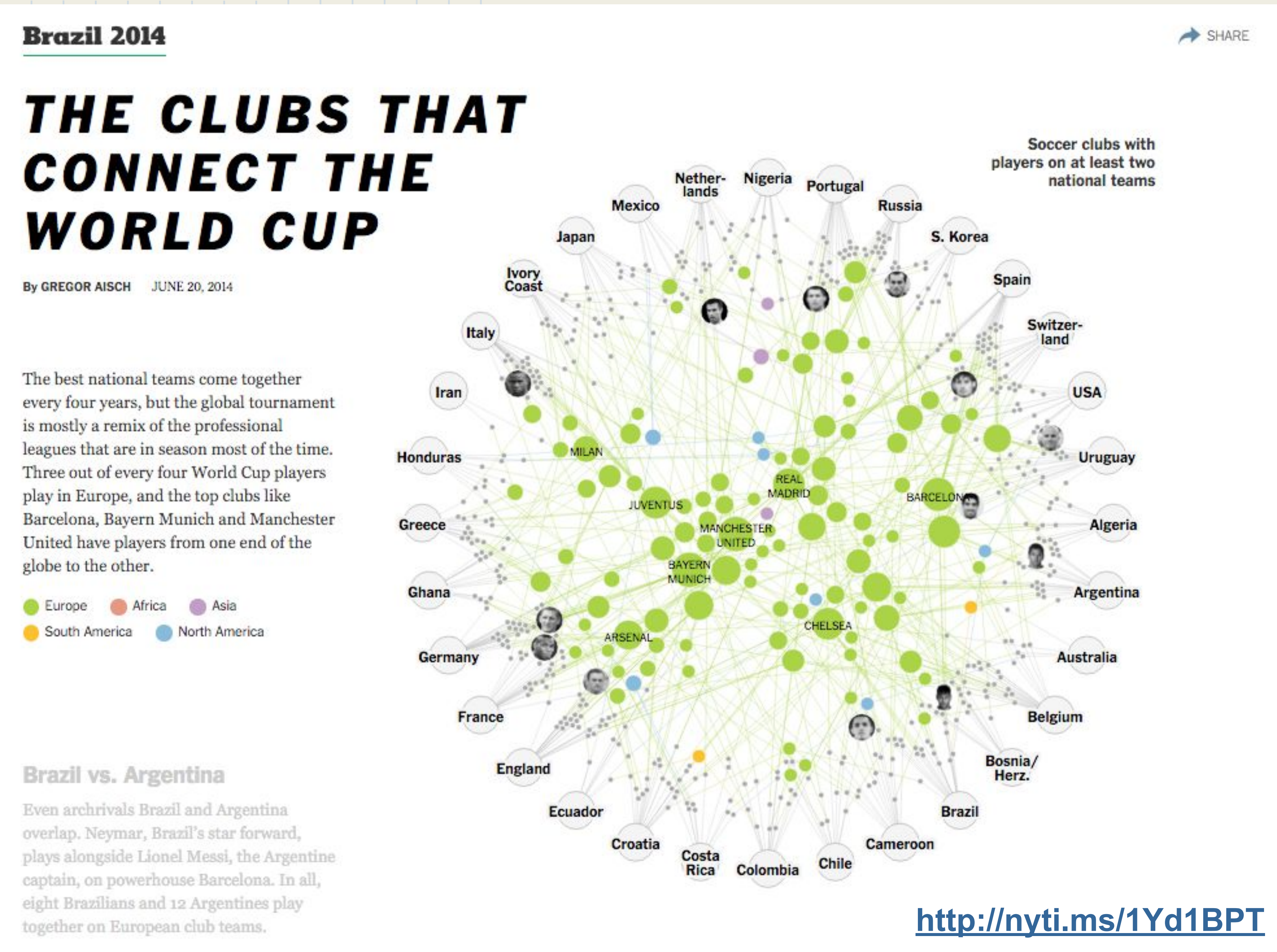
Graph Analysis



Brazil 2014

SHARE

THE CLUBS THAT CONNECT THE WORLD CUP

By GREGOR AISCH JUNE 20, 2014

The best national teams come together every four years, but the global tournament is mostly a remix of the professional leagues that are in season most of the time. Three out of every four World Cup players play in Europe, and the top clubs like Barcelona, Bayern Munich and Manchester United have players from one end of the globe to the other.

● Europe ● Africa ● Asia

● South America ● North America

Soccer clubs with players on at least two national teams

Brazil vs. Argentina

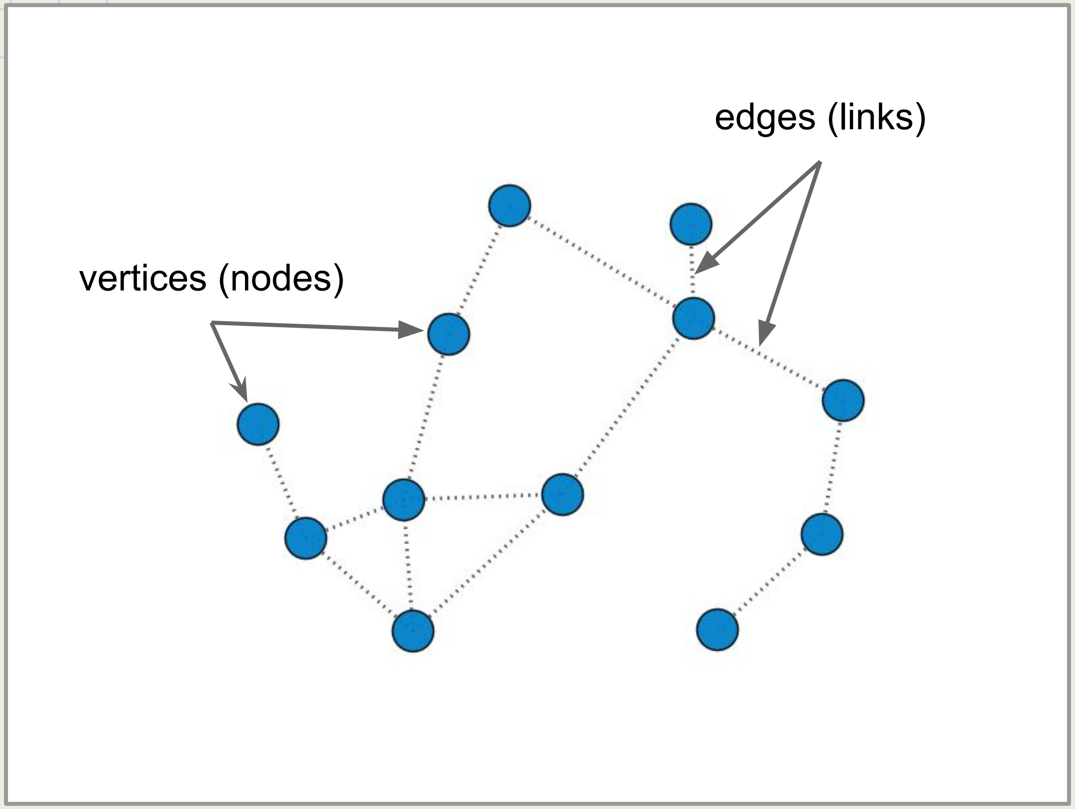
Even archrivals Brazil and Argentina overlap. Neymar, Brazil's star forward, plays alongside Lionel Messi, the Argentine captain, on powerhouse Barcelona. In all, eight Brazilians and 12 Argentines play together on European club teams.

http://nyti.ms/1Yd1BPT

Graph Theory

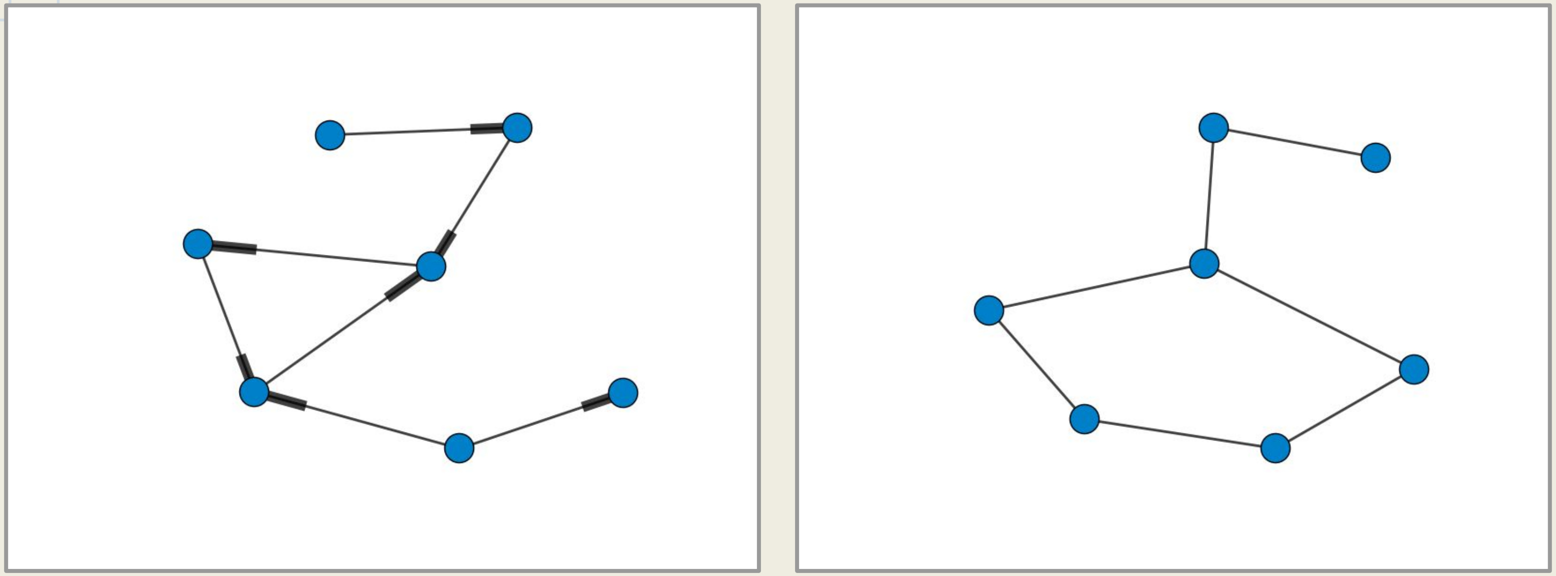
The Mathematical study of the application and properties of graphs, originally motivated by the study of games of chance.

Traced back to Euler’s work on the Konigsberg Bridges problem (1735), leading to the concept of Eulerian graphs.



$$G=(V,E)$$

A Graph, `G`, consists of a finite set denoted by `V` or `V(G)` and a collection `E` or `E(G)` of ordered or unordered pairs `{u,v}` where u and `v` ∈ `V`



Graphs can be directed or undirected

DiGraphs, the edges are ordered pairs: `(u,v)`

Network Definitions

$$G=(V,E) \quad E \subset V^2$$

$$\{(x,x)\,|\,x \in V\} \bigcap E = \varnothing$$

Adjacency Matrix

$$A_{ij} = \begin{cases} 1 \;\; if \;\; (i,j) \in E \\ 0 \;\; otherwise \end{cases}$$

Cardinality

$$O(G) = |V|$$ Order

$$S(G) = |E|$$ Size

Nodes

$$N_G(v_i) = (v_j \in A_{ij} \;\; if \;\; A_{ij} = 1)$$

$$K(v_i) = |N_G(v_i)|$$ Degree

Directed Networks

$$k_i^{out} = \sum_j A_{ij} \qquad k_i^{in} = \sum_j A_{ji}$$

$$k_i = k_i^{in} + k_i^{out}$$

Undirected Networks

$$k_i = \sum_j A_{ji} = \sum_j A_{ij}$$

Basic Notation and Terminology

Paths in a Network

$p = \langle v_i, \ldots, v_j \rangle \ (v_{k-1}, v_k \in E)$ a path from i to k

Length: # of nodes in path, set of paths from i to j: Paths(i,j)

Shortest (unweighted) path length

$$L(i,j) = \min(\{length(p) \mid p \in Paths(i,j)\})$$

Diameter: the “longest shortest path”

Classes of Graph Problems

- **Existence**

Does there exist [a path, a vertex, a set] within [constraints]?

- **Construction**

Given a set of [paths, vertices] is a [constraint] graph possible?

- **Enumeration**

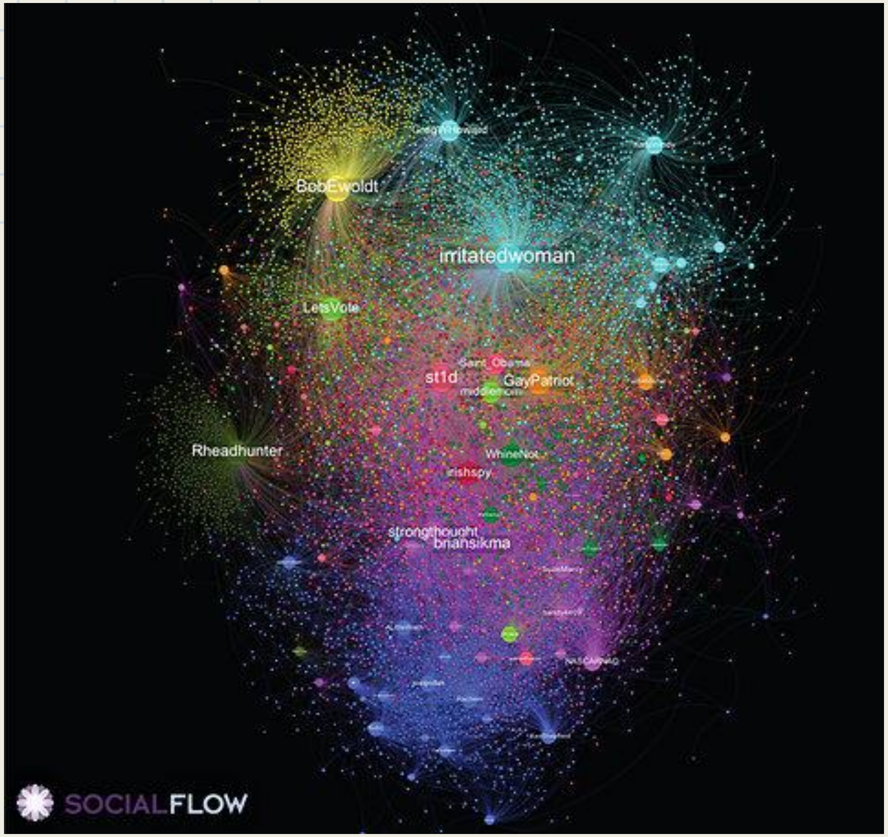
How many [vertices, edges] exist with [constraints]
Is it possible to list them?

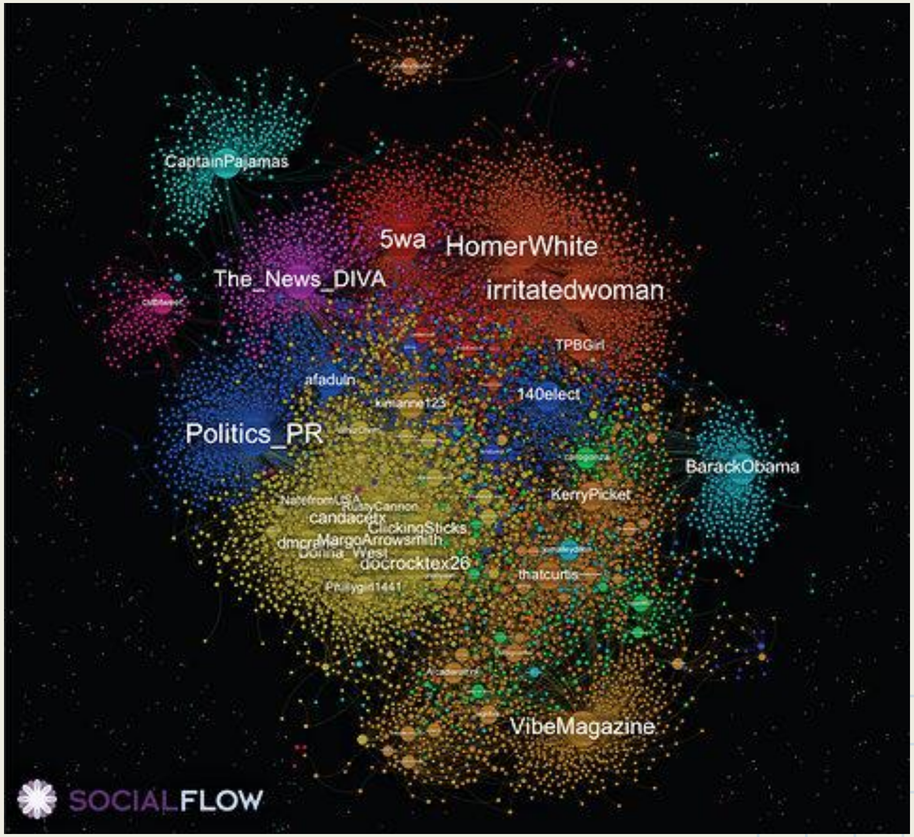
- **Optimization**

Given several [paths, etc.] is one the best?

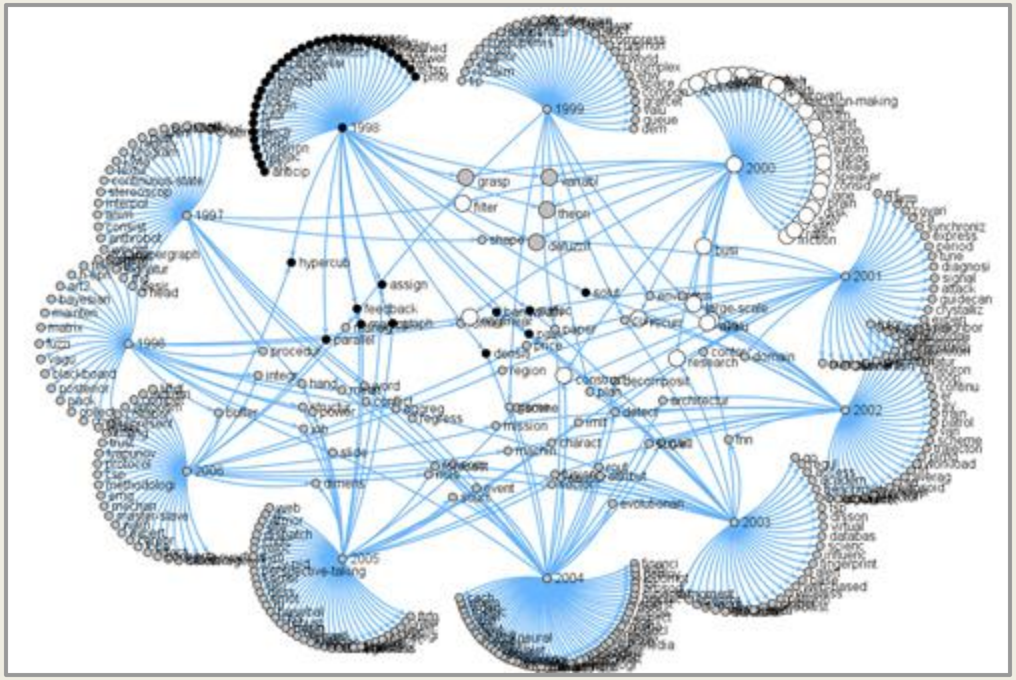
Why Graphs?

1. Graphs are abstractions of real life
2. Represent information flows that exist
3. Explicitly demonstrate relationships
4. Enable computations across large datasets
5. Allow us to compute locally to areas of interest with small traversals
6. Because everyone else is doing it (PageRank, Social Graph)



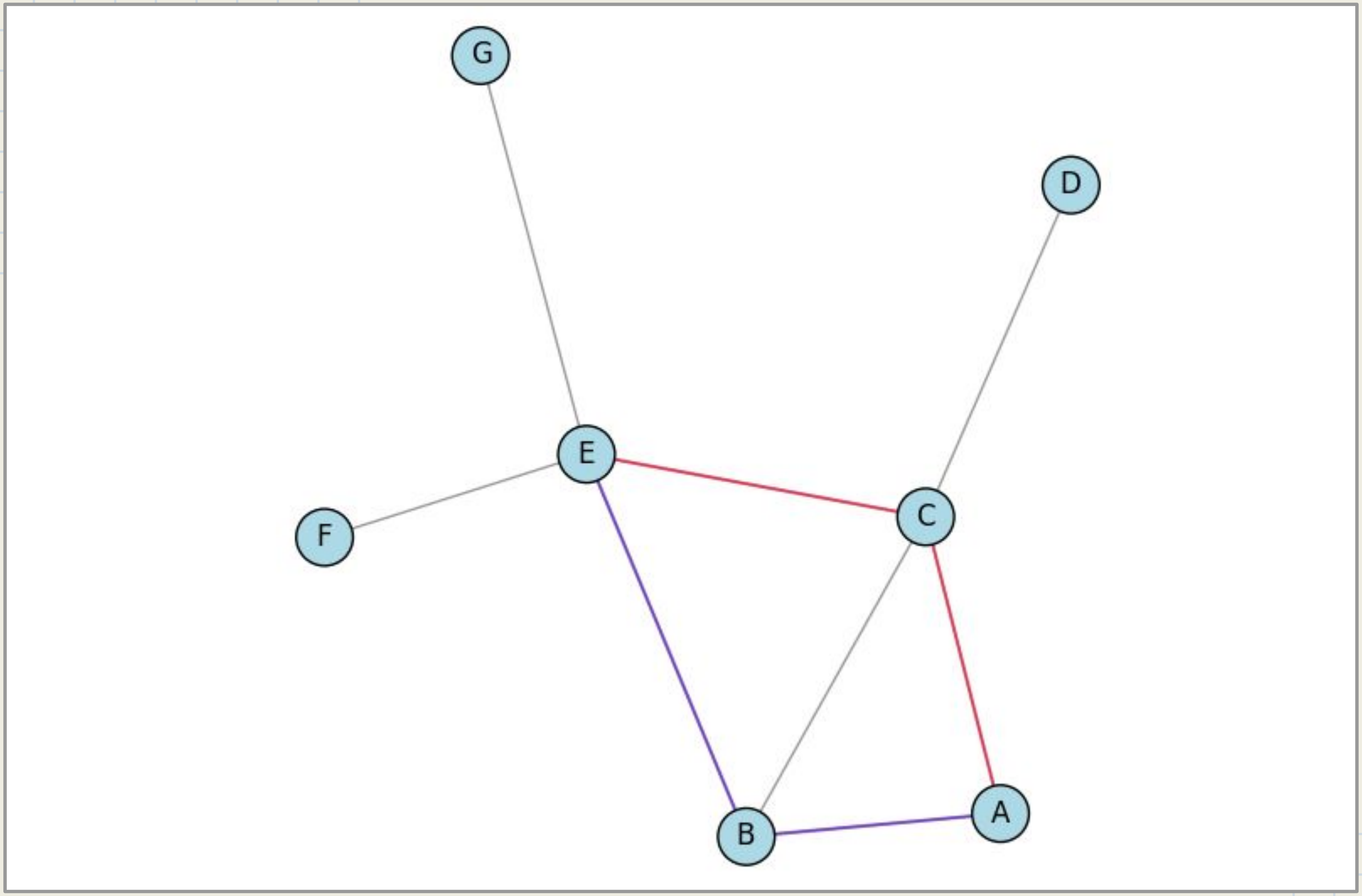


Ryan vs. Biden Debate (Twitter Reaction)
http://nyti.ms/1LZhVfY

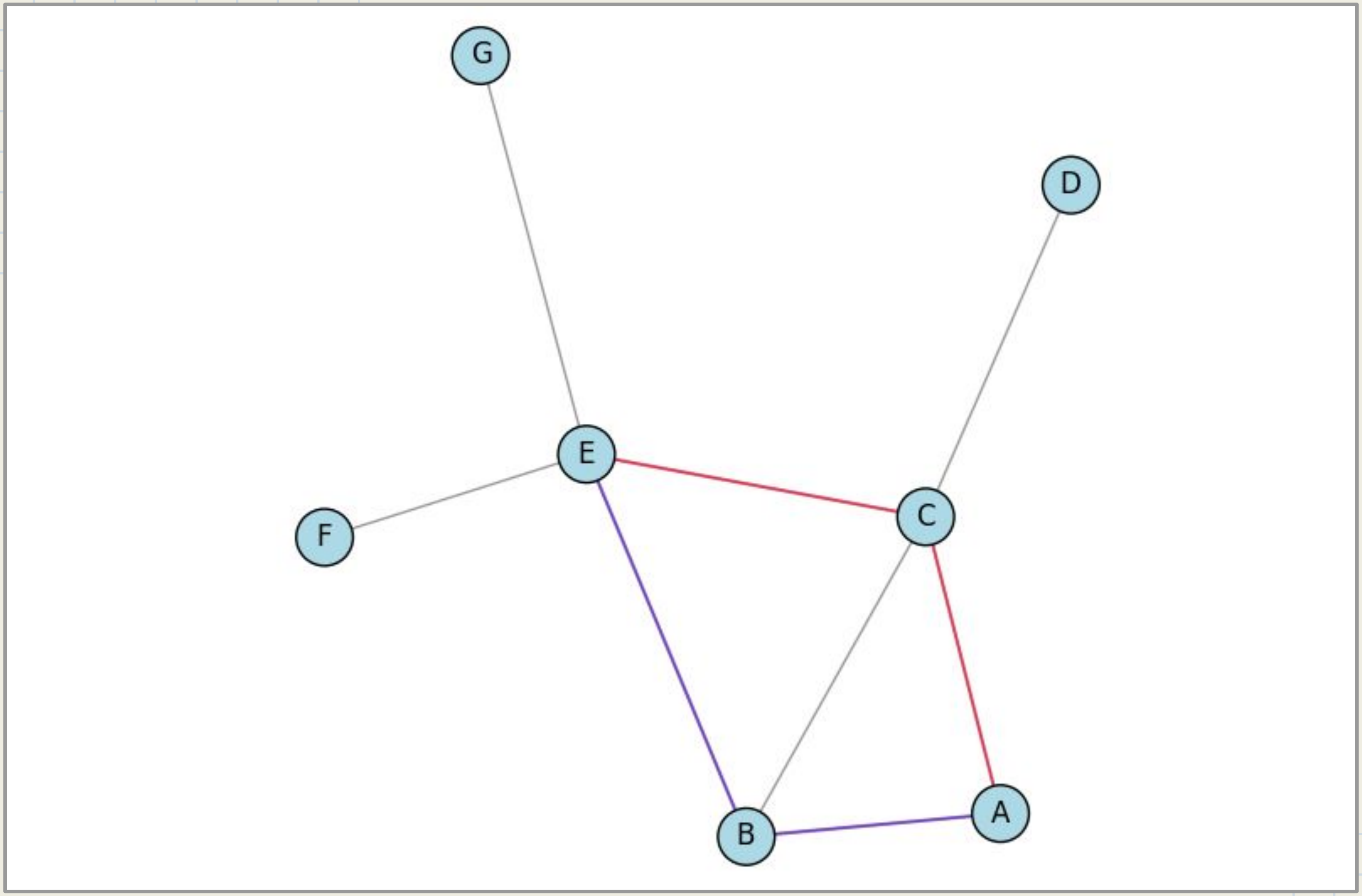


Topics shifting over time
http://bit.ly/1L7KNbh

Graph Analytics



Sample graph - note the shortest paths from D to F and A to E.



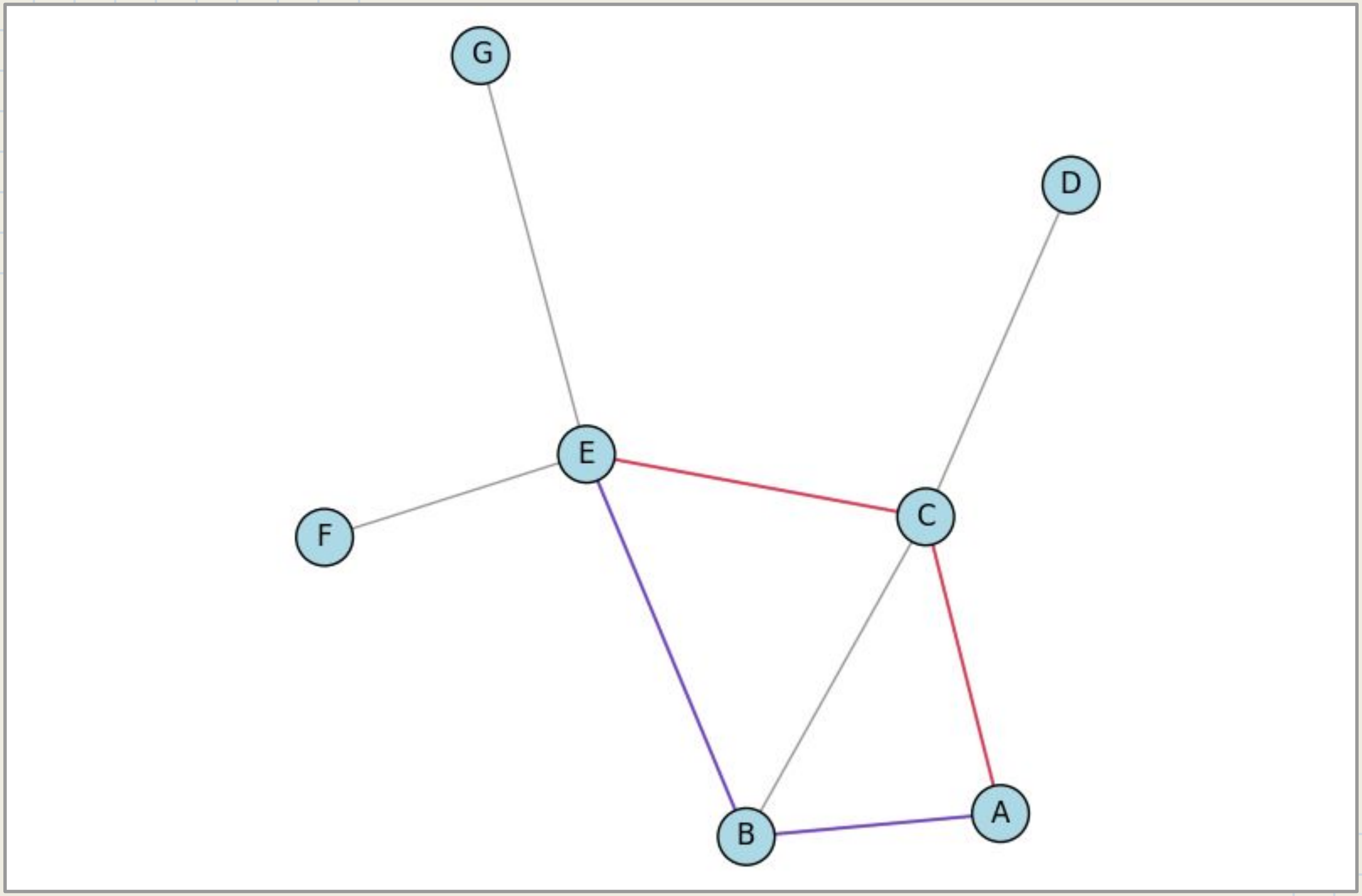
What is the most important vertex?

Centrality

Identification of vertices that play the most important role in a particular network (e.g. how close to the center of the core is the vertex?)

Degree Centrality

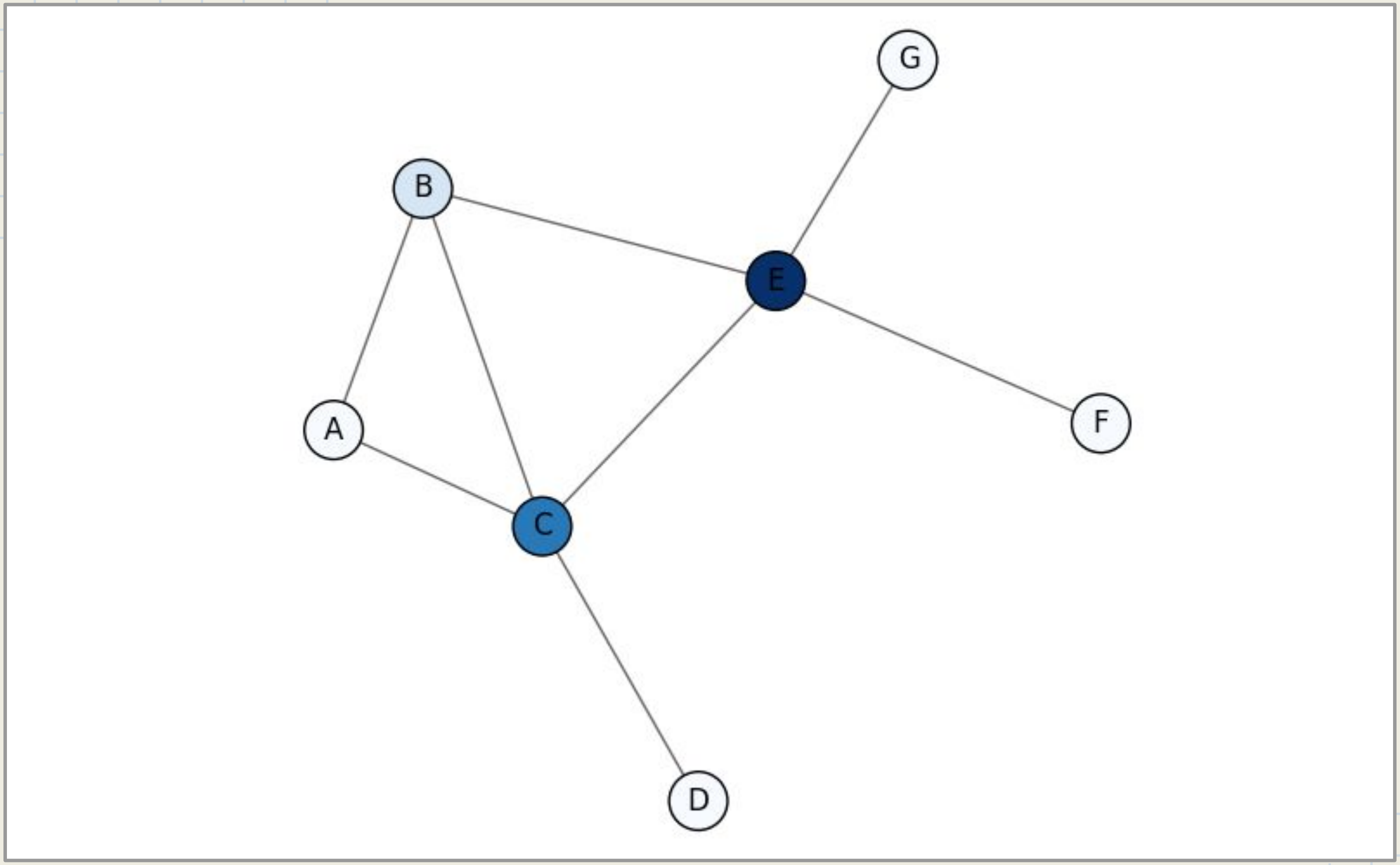
A measure of popularity, determines nodes that can quickly spread information to a localized area.



Who has the largest degree, is the most popular?

Betweenness Centrality

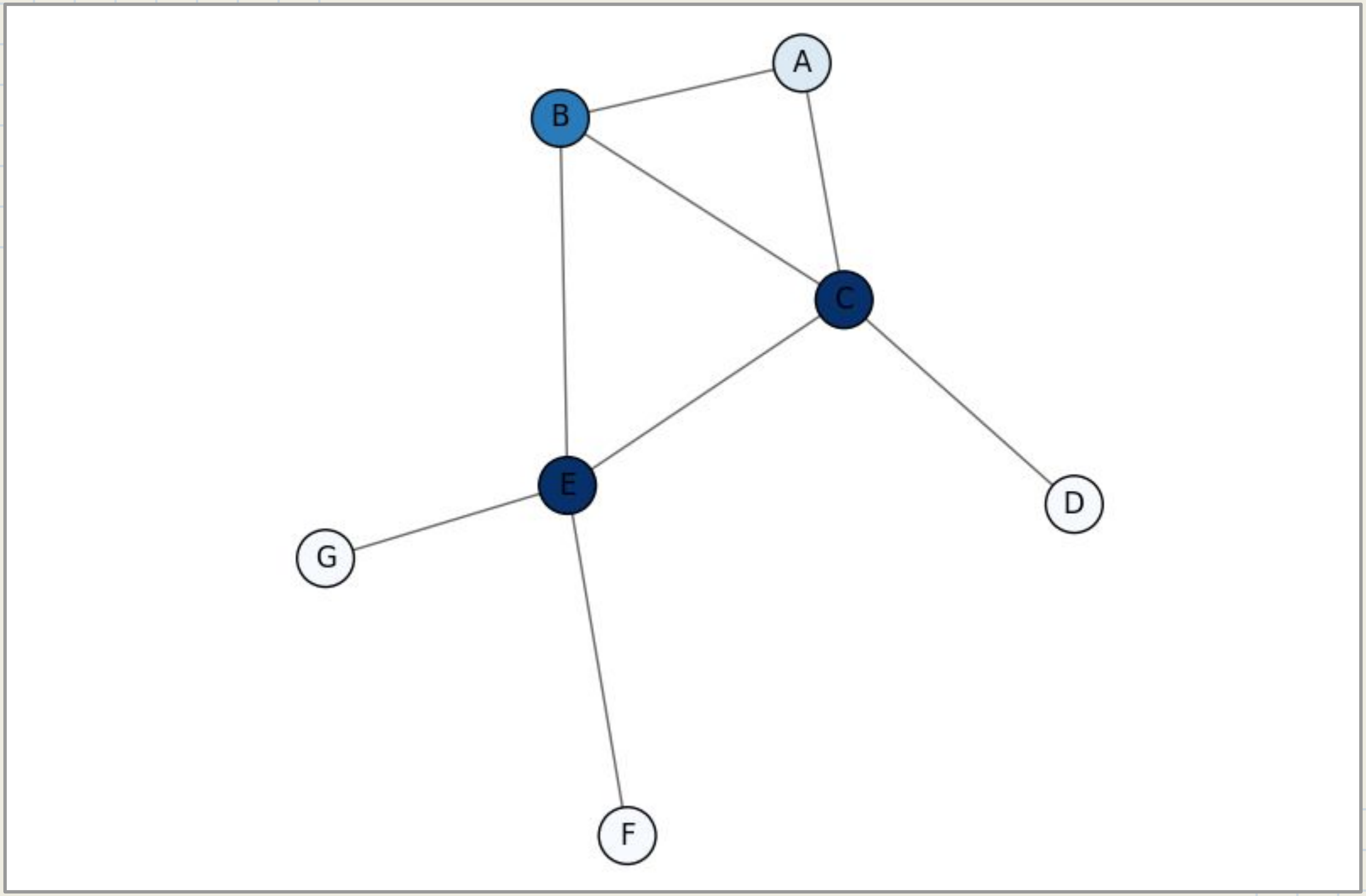
Shows which nodes are likely pathways of information and can be used to determine how a graph will break apart of nodes are removed.



Betweenness: the sum of the fraction of all the pairs of shortest paths that pass through that particular node. Can also be normalized by the number of nodes or an edge weight.

Closeness Centrality

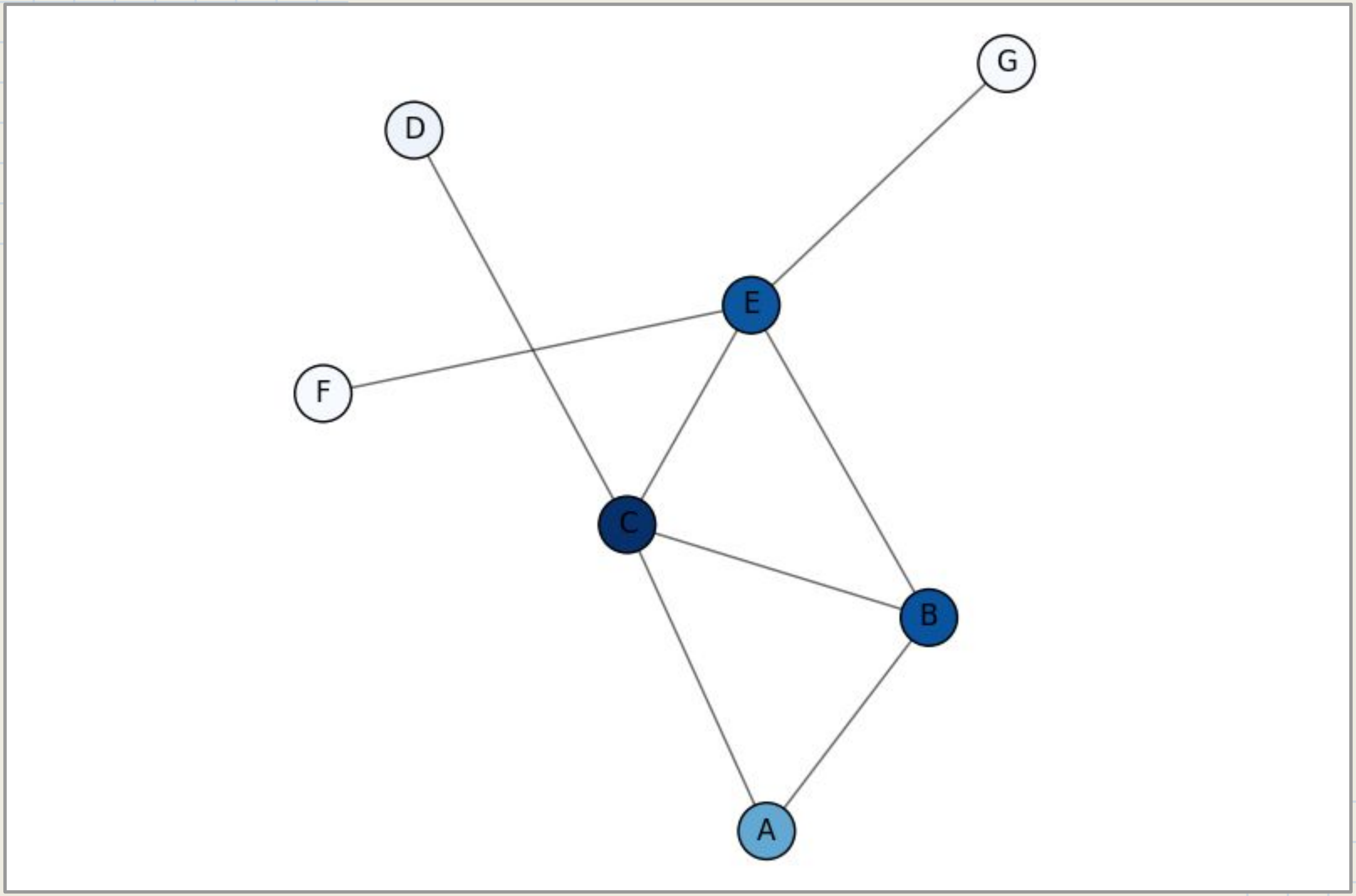
A measure of reach; how fast information will spread to all other nodes from a single node.



Closeness: average number of hops to reach any other node in the network.
The reciprocal of the mean distance: n-1 / size(G) - 1 for a neighborhood, n

Eigenvector Centrality

A measure of related influence, who is closest to the most important people in the Graph? Kind of like “power behind the scenes” or influence beyond popularity.



Eigenvector: proportional to the sum of centrality scores of the neighborhood.
(PageRank is a stochastic eigenvector scoring)

Clustering

Detection of communities or groups that exist in a network by counting triangles.

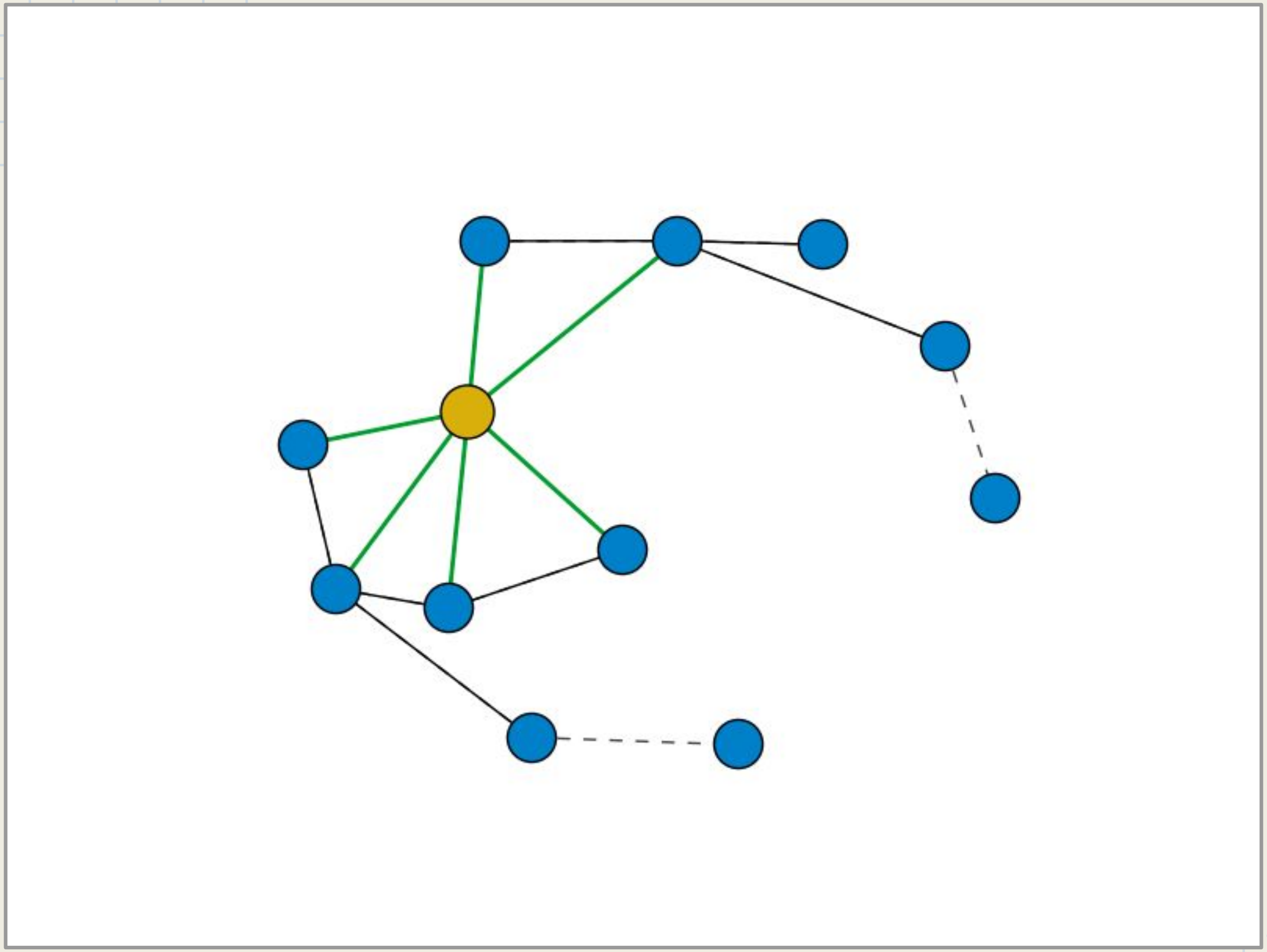
Measures “transitivity” - tripartite relationships that indicate clusters

$$C_i = \binom{k_i}{2}^{-1} T(i)$$

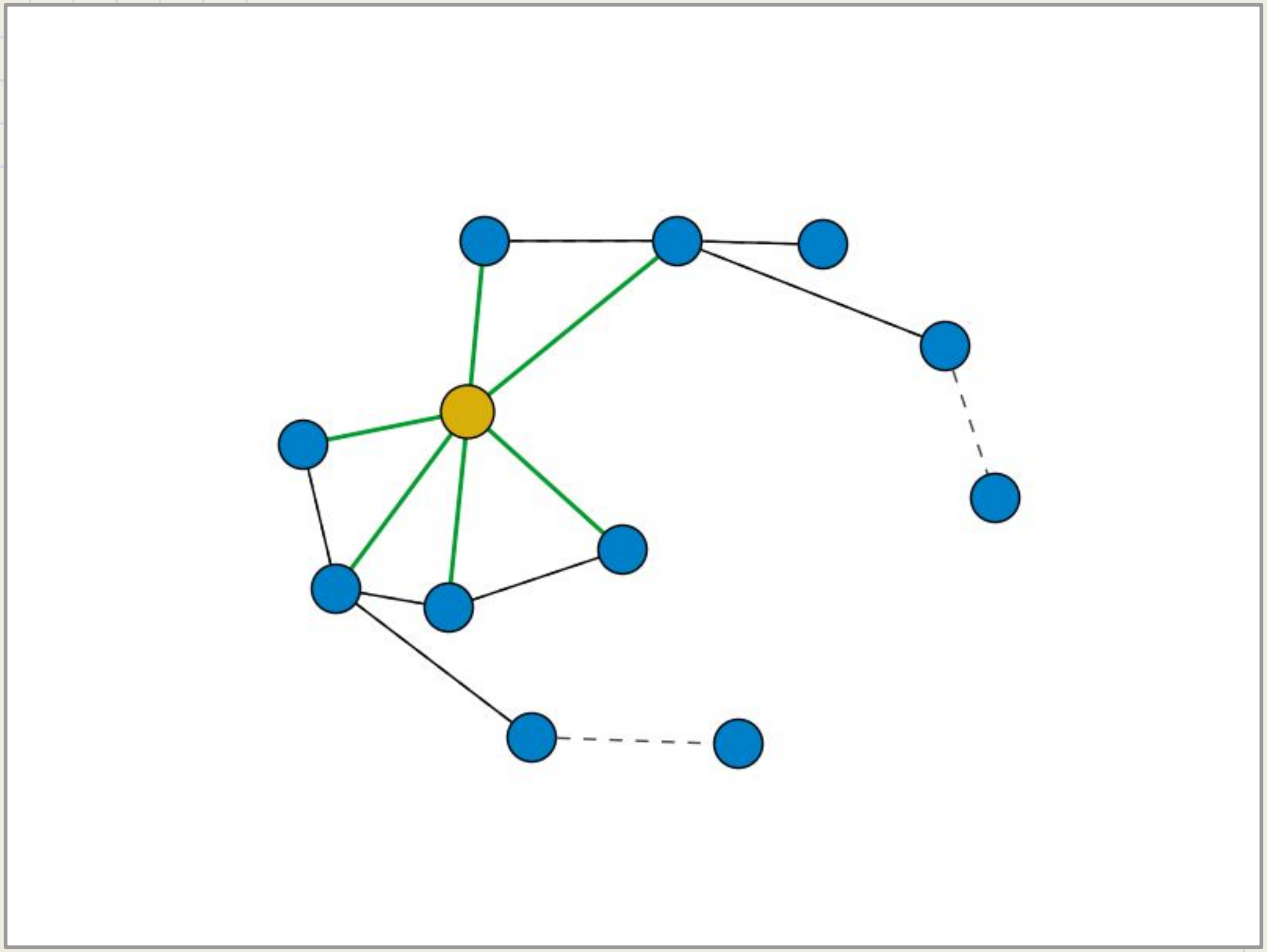
Local Clustering Coefficient

$$C = \frac{1}{n} \sum_{i \in V} C_i$$

Graph Clustering Coefficient



Green lines are connections from the node, black are the other connections



$k_i = 6$

$T(i) = 4$

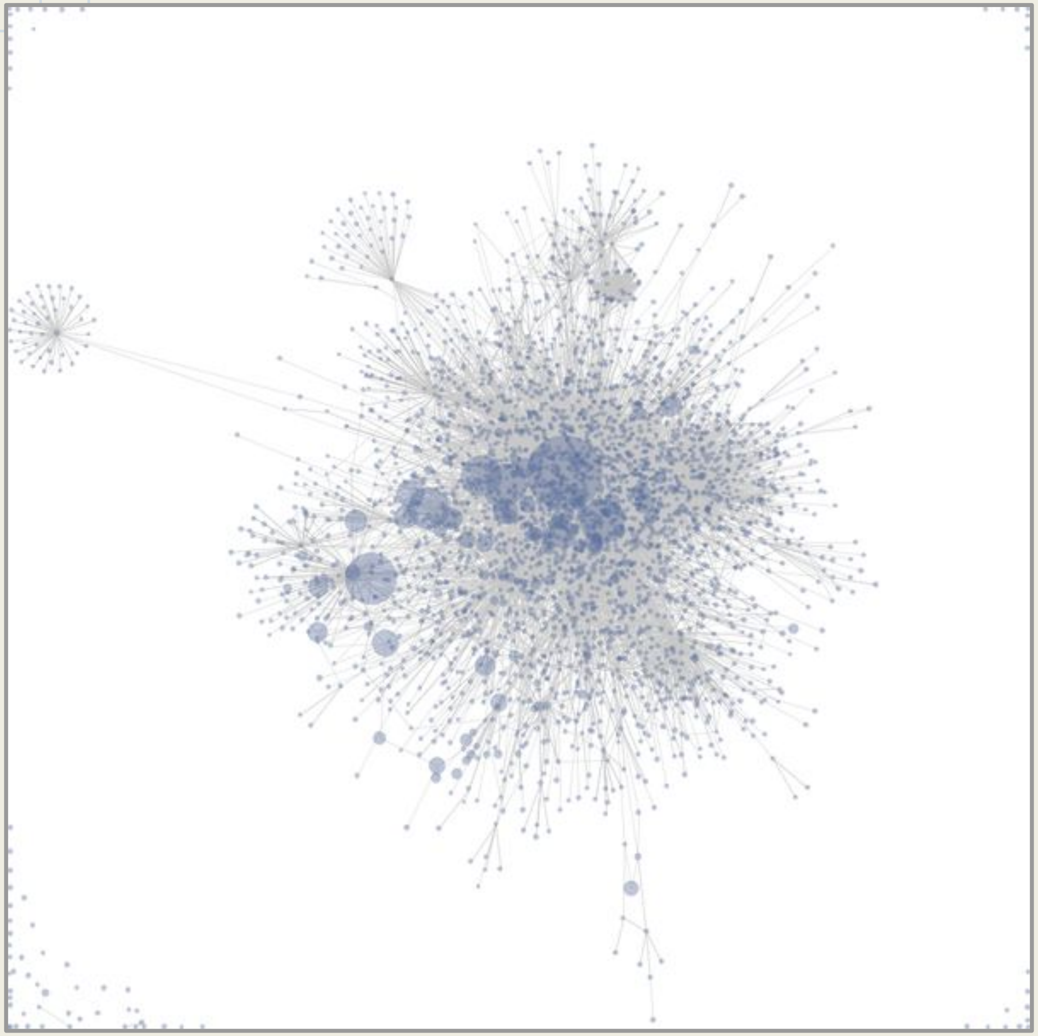
$C_i = (2*4) / (6*(6-1)) = 0.266$

Graph Visualization

Layouts

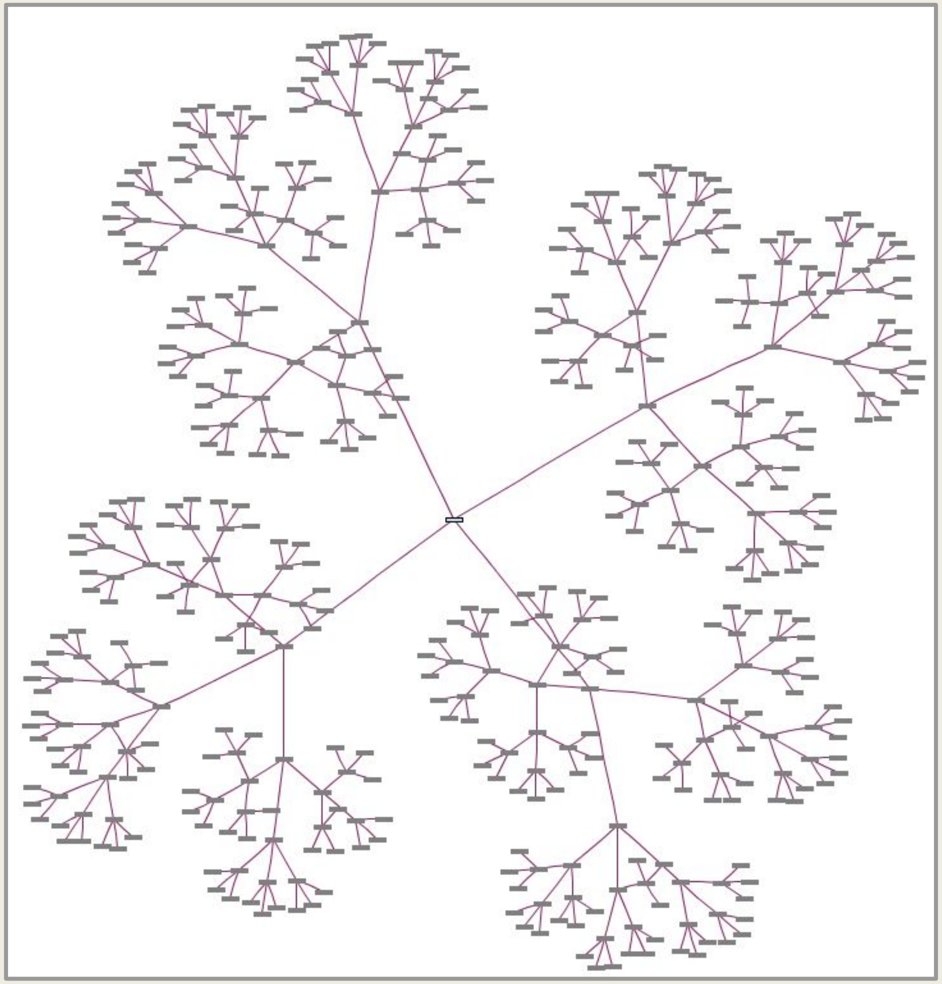
- Open Ord
  - http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=731088
  - Draws large scale undirected graphs with visual clusters
- Yifan Hu
  - http://yifanhu.net/PUB/graph_draw_small.pdf
  - Force Directed Layout with multiple levels and quadtree
- Force Atlas
  - http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0098679
  - A continuous force directed layout (default of Gephi)
- Fruchterman Reingold
  - http://cs.brown.edu/~rt/gdhandbook/chapters/force-directed.pdf
  - Graph as a system of mass particles (nodes are particles, edges are springs) This is the basis for force directed layouts

Others: circular, shell, neato, spectral, dot, twopi …



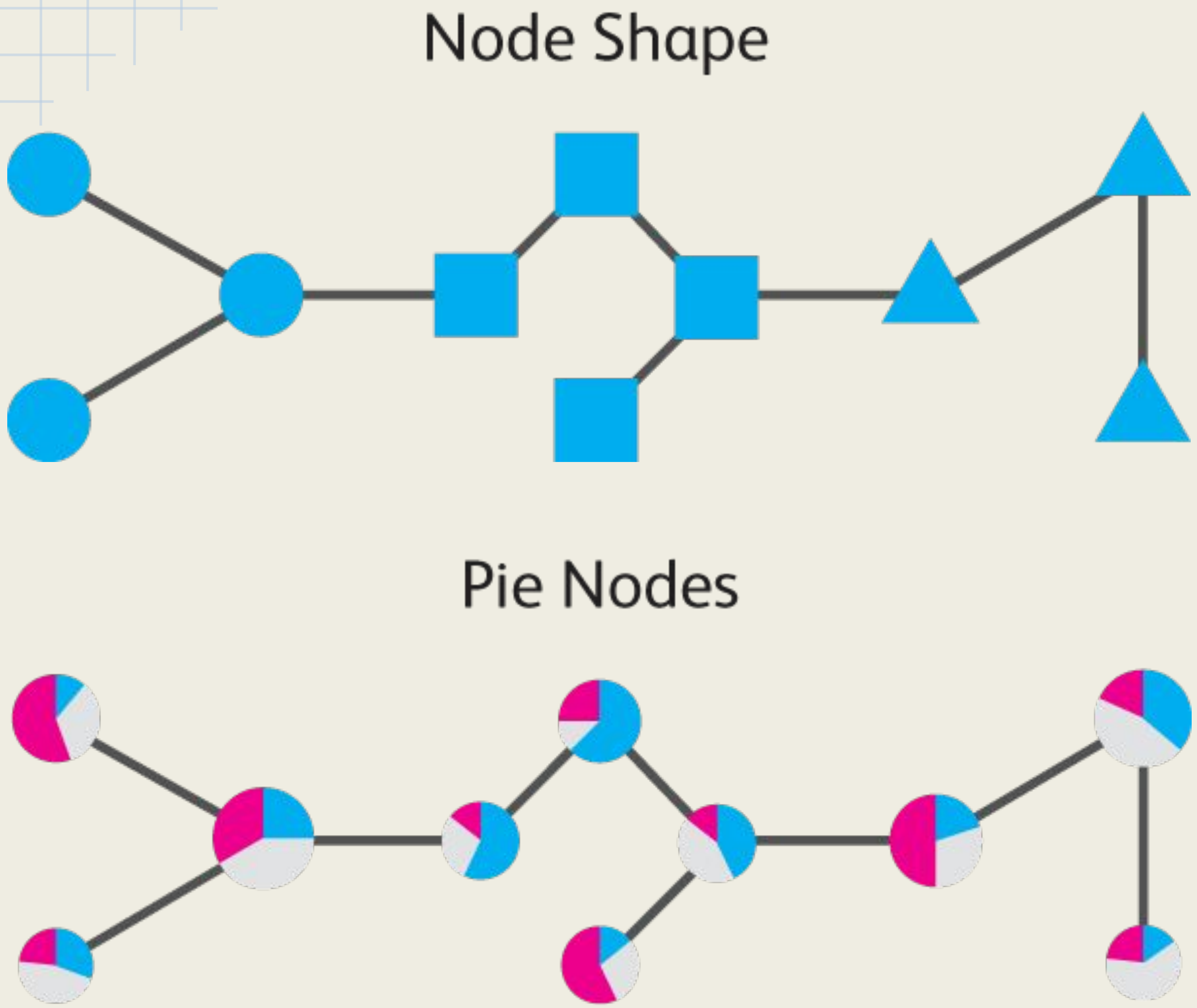
Force Directed

http://en.wikipedia.org/wiki/Force-directed_graph_drawing



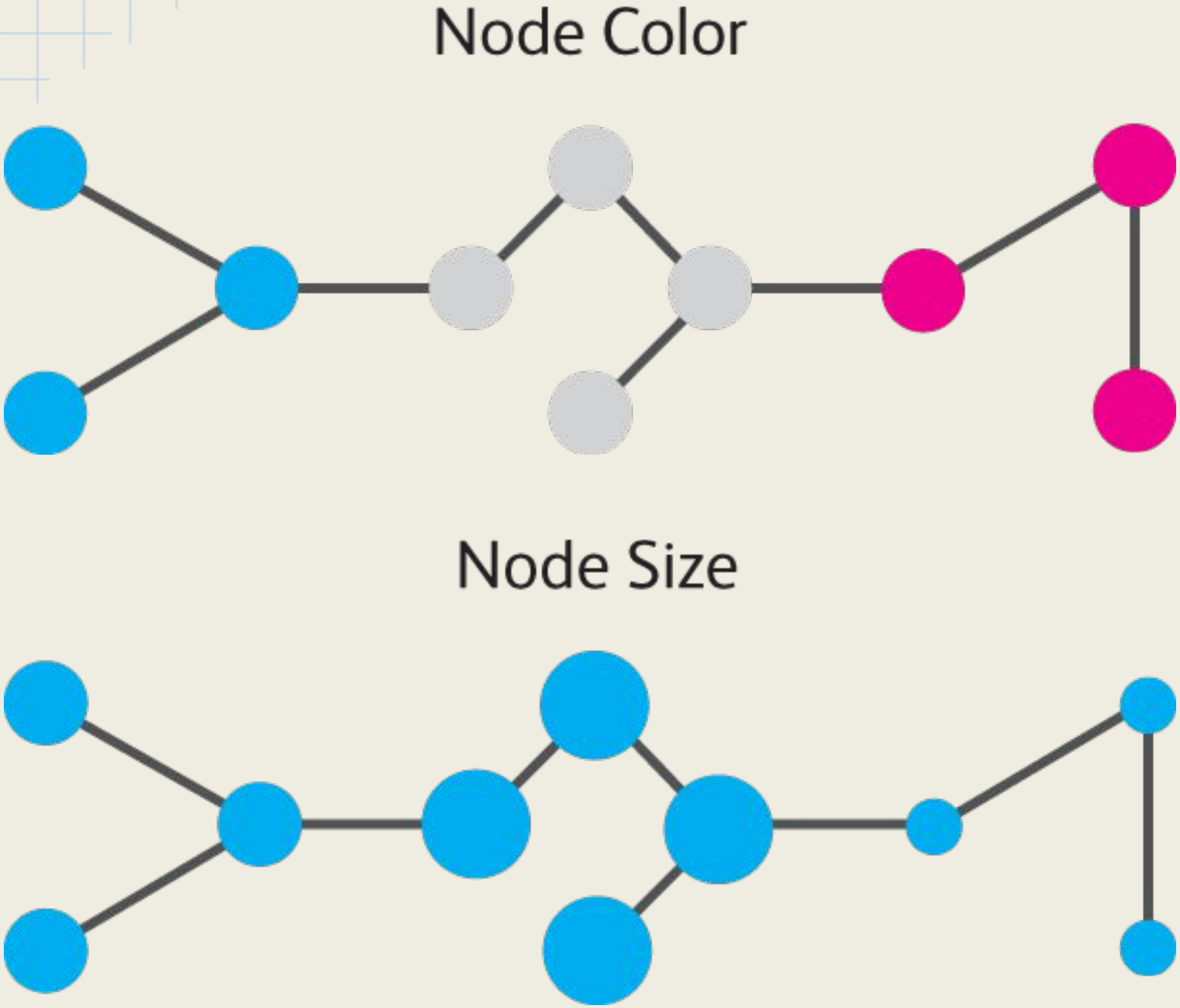
Hierarchical Graph Layout

https://seeingcomplexity.files.wordpress.com/2011/02/tree_graph_example.gif



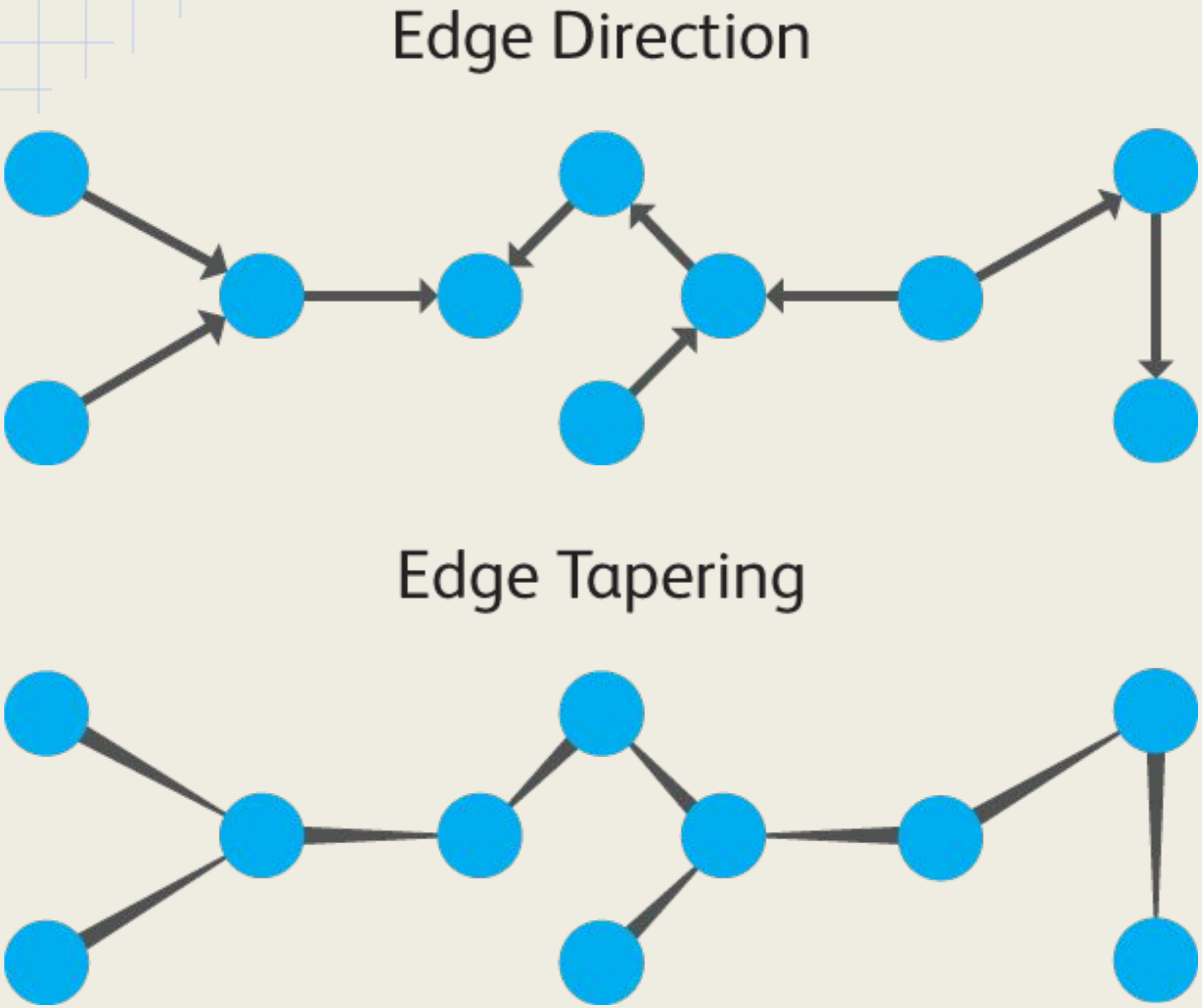
Lane Harrison, The Links that Bind Us: Network Visualizations

http://blog.visual.ly/network-visualizations



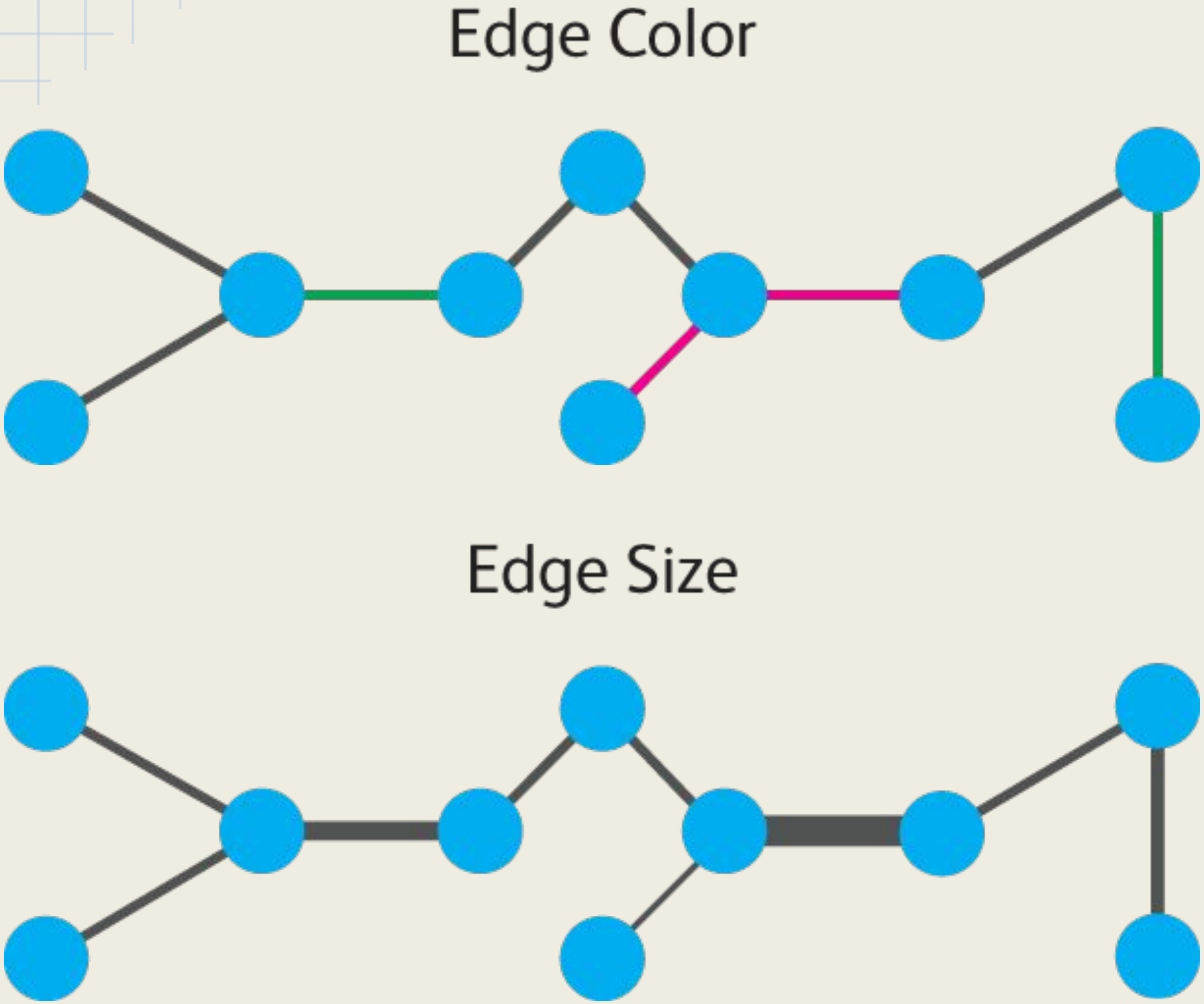
Lane Harrison, The Links that Bind Us: Network Visualizations

http://blog.visual.ly/network-visualizations



Lane Harrison, The Links that Bind Us: Network Visualizations

http://blog.visual.ly/network-visualizations

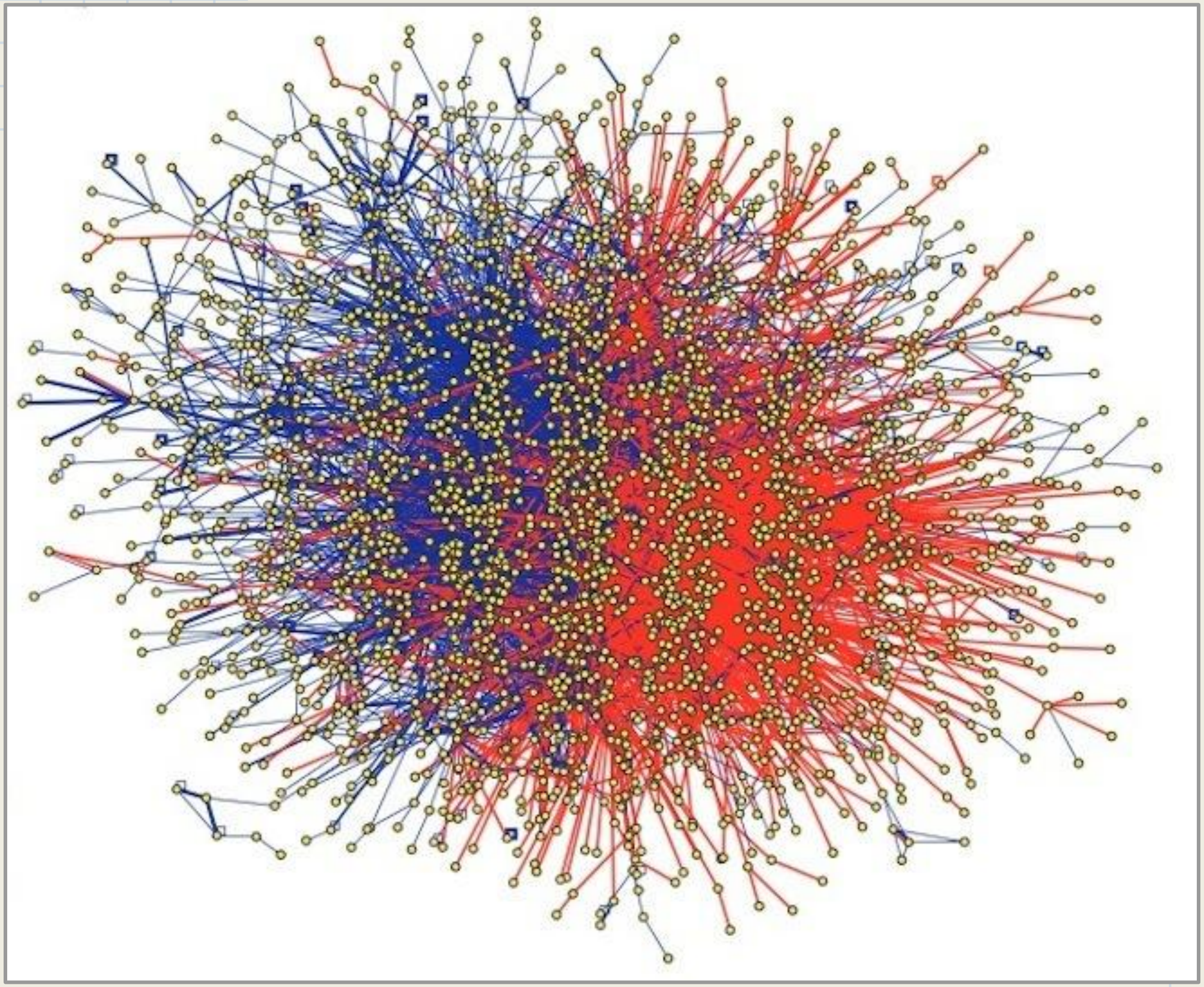


Edge Color

Edge Size

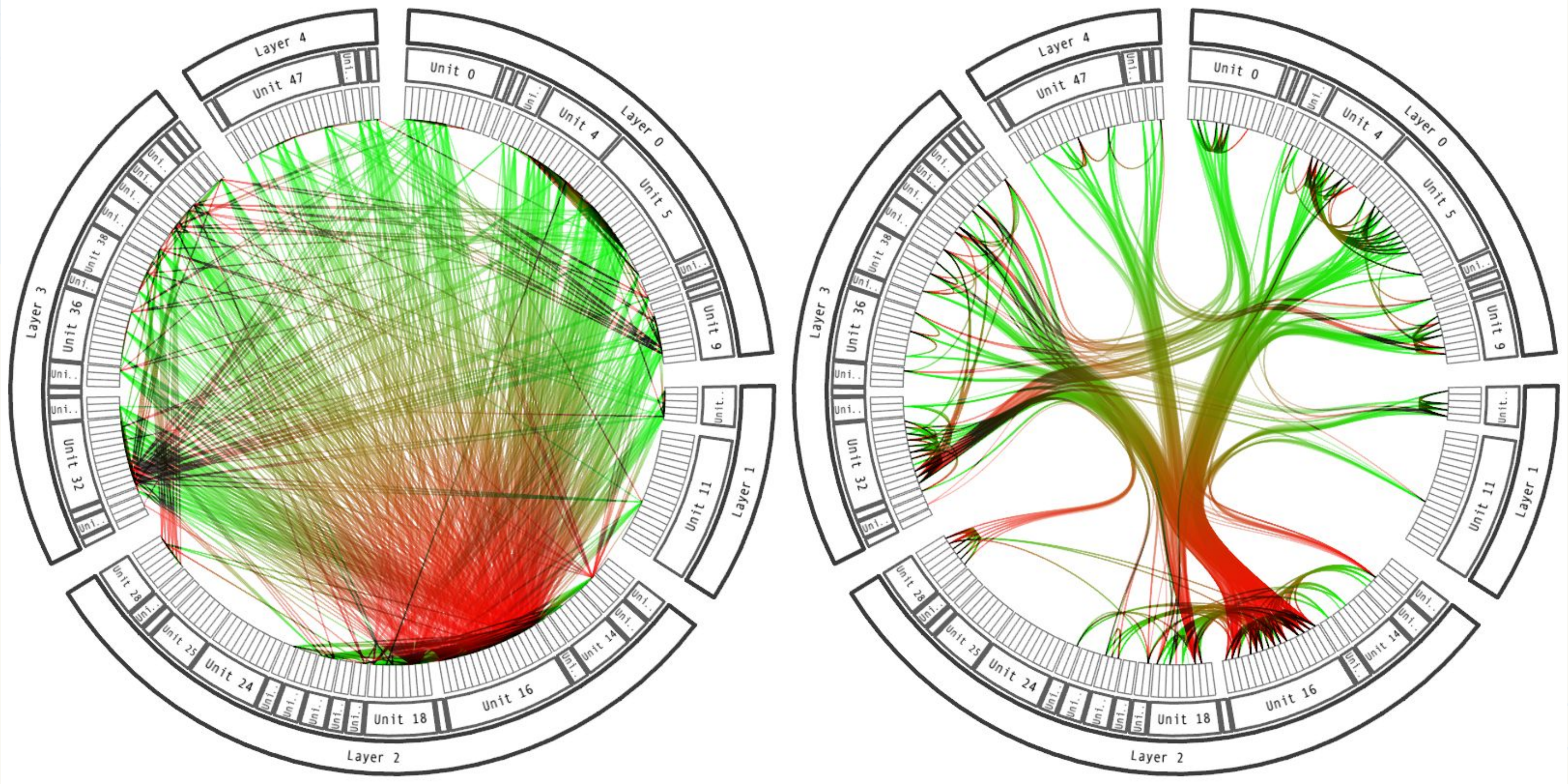
Lane Harrison, The Links that Bind Us: Network Visualizations

http://blog.visual.ly/network-visualizations



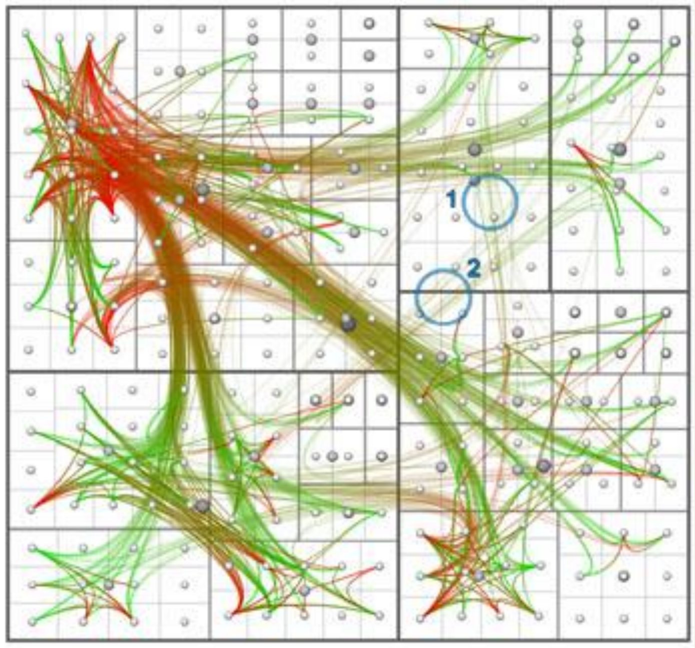
The Hairball

http://www.slideshare.net/OReillyStrata/visualizing-networks-beyond-the-hairball



Edge Bundling

https://seeingcomplexity.wordpress.com/2011/02/05/hierarchical-edge-bundles/



Region Bundling

http://infosthetics.com/archives/2007/03/hierarchical_edge_bundles.html

Tools for Graph Visualization