GRAPH TRAVERSAL

Analysis of Large Scale Social Networks

Bart Thijs

In Social Network Analysis, graph traversals describe the flow of information across networks.

Formally, a traversal can be described as a sequence of nodes and edges (or depending on the definition used, just a sequence of nodes).

Based on the restrictions that are imposed on such a sequence, it is possible to identify different types of traversals

WHAT?

TYPES OF GRAPH TRAVERSALS

Wasserman and Faust (1994) introduces three different types of traversals or routes

- Walk: sequence of alternating nodes and edges, with a start and end node. Edges are connecting preceding and following nodes.
- Trail: A walk in which each edge only occurs once
- Path: A trail in which each node only occurs once.

SHORTEST PATH

A shortest path between two nodes is a path for which it is impossible to find a path with a shorter length between these nodes.

Shortest path problem is defined at the local level:

- Between individual pairs
- From one source to all/many others

At the global level:

- Between all possible pairs

BETWEENNESS AND SHORTEST PATH

Consider an unordered pair of points, {p_i, p_j }, (i<>j).

- Either p_i and p_j are unreachable from one another
- Or, there are one or more paths between them.

In the latter case, each of the paths has a length equal to the number of edges contained in it.

Among the paths connecting p_i and p_j one or more have the shortest length: the ***geodesics***.

(See Freeman 1977)

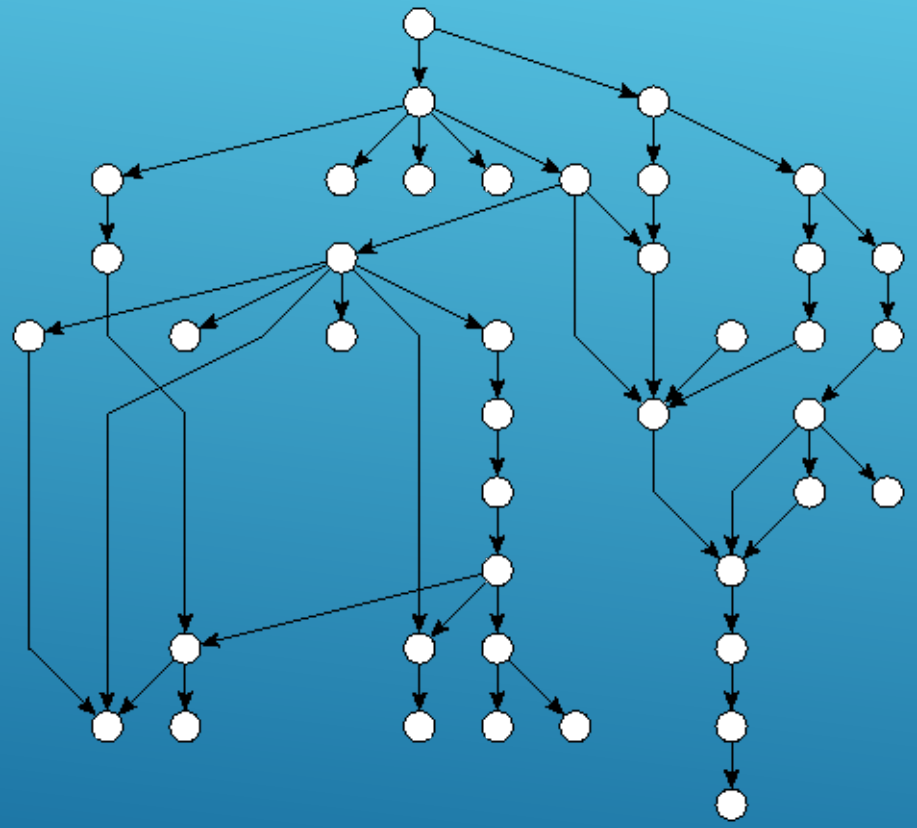
Freeman also states: '*A point is considered to be **central** here to the degree that it falls between other points on their shortest or geodesic communication paths. A point falling between two others can facilitate, block, distort or falsify communication between the two; it can more or less completely **control** their communication. But if it falls on some but not all of the geodesics connecting a pair of points, its potential for control is more **limited**.*'

BETWEENNESS AND SHORTEST PATH

SHORTEST PATH ALGORITHMS

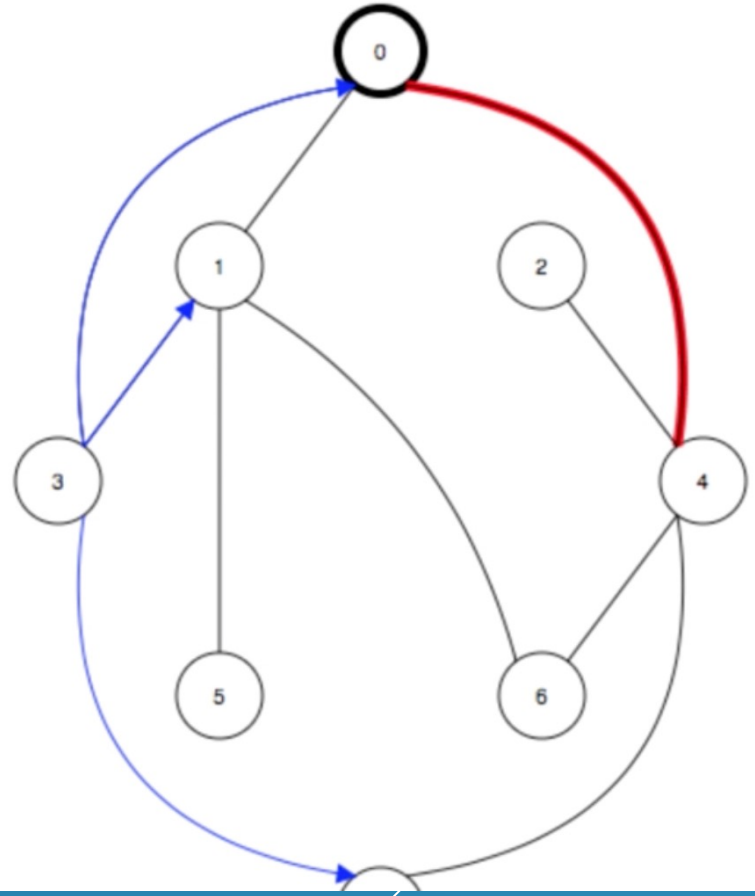
- Single Pair Shortest Path
  - A* algorithm
- Single Source / Destination Shortest Path
  - Breadth-First Search
  - Depth-First Search
  - Dijkstra (Weighted Edges, implements a priority queue)
  - Bellman-Ford (Negative Edge Weights)
- All Pairs Shortest Path
  - Floyd-Warshall (Negative Edge Weights)

See: https://www.cs.usfca.edu/~galles/visualization/Algorithms.html



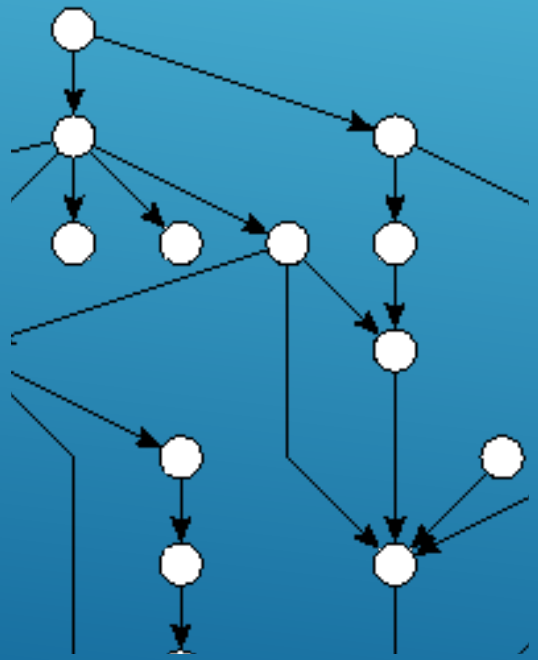
	Parent
0	3
1	3
2	
3	
4	
5	
6	
7	3

	Visited
0	T
1	T
2	f
3	f
4	f
5	f
6	f
7	T



BREADTH-FIRST SEARCH ON A FLOW DIAGRAM

Due to the ranking step in the creation of the processing queue many possible shortest paths are ignored and thus never counted for the calculation of the betweenness.



BREADTH-FIRST SEARCH ON A FLOW DIAGRAM

Solving this issue requires two additions:

1. The algorithm should record all possible parents with the same path lengths. To be able to do this an additional vector with cost has to be added.
2. For the calculation of betweenness all combinations of possible parents have to be considered.

BREADTH-FIRST SEARCH ON A FLOW DIAGRAM

In fact, the Dijkstra implementation shown in the clip suffers from the same problem. The nodes are ranked based on lowest cost and index of the node.

This touches upon the very nature of network-based algorithms. They can be very efficient through the use of traversals but ignore many other solutions. But looking at all combinations implies serious computational costs.

DIJKSTRA ON A FLOW DIAGRAM

FLOYD-WARSHALL

The Floyd-Warshall algorithm is looking at the shortest path betweenn all pairs of nodes in the network. But it is not able to identify *all possible shortest paths with the same cost* between two nodes.
But it comes already with a high computational cost.

Eg. directed network of 107 nodes requires 107 * 106 * 105 steps

The issue of similarity is easily solved by taking the inverse of the distance or dissimilarity.

RANDOM WALK

A random walk in a graph starting from node and of length *k* is a sequence of nodes and edges starting with node *i* and randomly chosen subsequent connecting edges and nodes. These walks can be considered as Markov Chains, each step is independent from the previous one.

$$Probability\ of\ step\ from\ node\ i\ to\ j\ at\ time\ t\ in\ a\ random\ walk\ (X_0, X_1, X_2, \dots, X_k, \dots)\ is\ equal\ to$$

$$p(X_k = j) = p(X_{k-1} = i)A_{ij}\frac{1}{d_i}$$

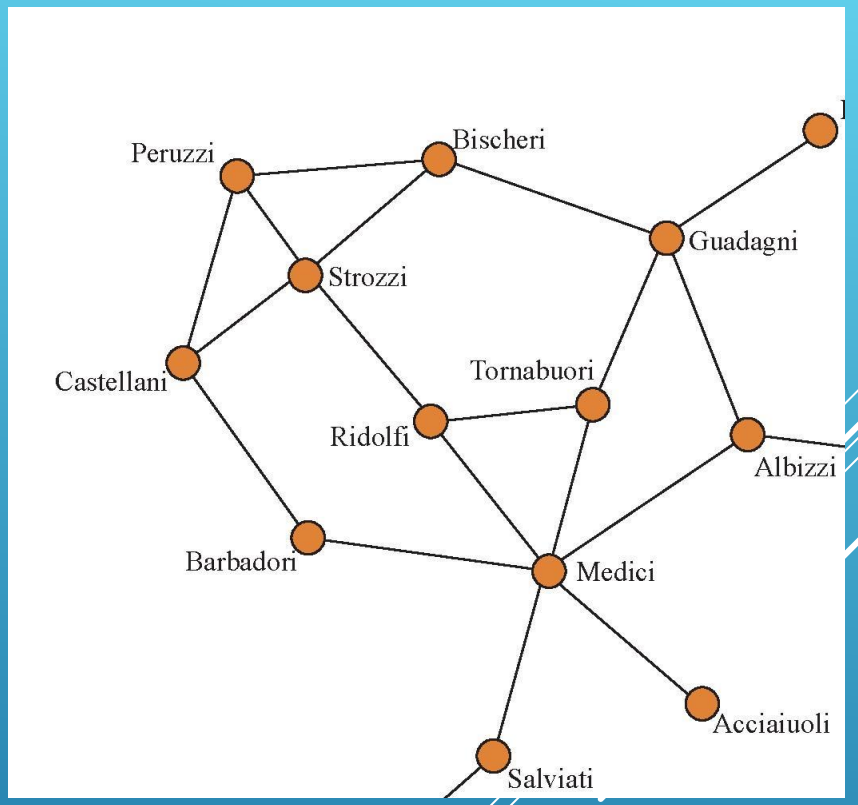
Probability of next step from Strozzi to Bischeri

$$p(X_{k+1} = Bischeri \mid X_k = Strozzi) = p(X_k = Strozzi) A_{ij} \frac{1}{d_i}$$

$$p(X_{k+1} = Bischeri \mid X_k = Strozzi) = p(X_k = Strozzi) \frac{1}{4}$$

Probability of being at Strozzi at current step

$$p(X_k = Strozzi) = p(X_{k-1} = Bischeri) \frac{1}{3} +$$
$$p(X_{k-1} = Ridolfi) \frac{1}{3} + p(X_{k-1} = Castellani) \frac{1}{3} +$$
$$p(X_{k-1} = Peruzzi) \frac{1}{3}$$



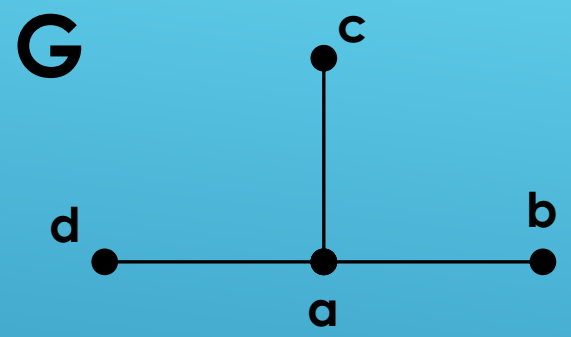
LAPLACIAN MATRIX

Suppose D is diagonal degree matrix and A is adjacency matrix then the Laplacian matrix L is defined as:

$$L = D - A$$

for an undirected and unweighted network where

$$L_{ij} = \begin{cases} \deg(v_i) \ \ if \ i = j \\ -1 \ if \ i \neq j \ and \ \ v_i \ and \ v_j \ are \ adjacent \\ 0 \ otherwise \end{cases}$$



$$L(G) = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

Properties:

- Singular, determinant is 0
- First eigenvalue is 0 with eigenvector $v_0 = (1)^n$
- All eigenvalues are nonnegative (0,1,1,4)
- The multiplicity of 0 is the number of connected components
- The second smallest value is the Algebraic Connectivity

(Connectivity: minimum number of nodes/edges to be removed to break connected graph

LAPLACIAN MATRIX: EXAMPLE

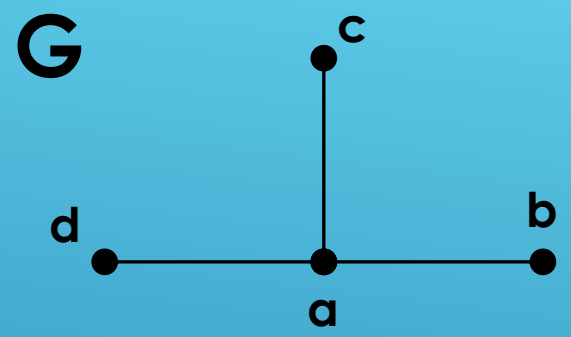
RANDOM WALK NORMALIZED LAPLACIAN

Suppose D^{-1} is diagonal matrix with inverse of degree, then

$$L^{RW} = D^{-1}L$$

With

$$L_{ij}^{RW} = \begin{cases} 1 \; if \; i = j \; and \; \deg(i) > 0 \\ -\frac{1}{\deg(i)} \; if \; i \neq j \; and \;\; i \; and \; j \; are \; adjacent \\ 0 \end{cases}$$



$$L(G) = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

$$L^{RW}(G) = \begin{pmatrix} 1 & -1/3 & -1/3 & -1/3 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

$L^{RW}(G) = I - P$ with $P = D^{-1}A$

RANDOM WALK NORMALIZED LAPLACIAN

MARKOV TRANSITION MATRIX

Matrix that holds the probabilities of moving from one note (state) to the next. Applicable on directed, weighted graph

With

$$P_{ij} = \begin{cases} \dfrac{w_{ij}}{\text{outdeg}(i)} \; if \; i \neq j \; and \; \; directed \; edge \; from \; i \; to \; j \\ 0 \end{cases}$$

APPLICATIONS OF RANDOM WALK

- Newman:
  - Betweenness Centrality
  - Clustering / Community Detection
- Rosvall and Bergstrom
  - Clustering / Community Detection

NEWMAN: BETWEENNESS

Questions on
Newman, M.E.J., (2005) A measure of betweenness centrality based on random walks, *Social Networks*, 27(1), 39-54:

- Explain the problems that Freeman (1991) seems to address with flow betweenness and how this solutions is not appropriate according to Newman.
- What are the implications of only using the *net* number of passes through a node?
- Why calculate betweenness. It often correlates strongly with degree. Think of an example of why it matters.

NEWMAN: BETWEENNESS

- Explain the problems that Freeman (1991) seems to address with flow betweenness and how this solutions is not appropriate according to Newman.

According the Newman, this approach still assumes that the ‘ideal route’ is known and that the network tries to maximize the flow by taking additional routes. However, these routes are limited as the additional flow cannot pass through the edges on the ideal route as they are already used for their full capacity.

Newman’s solution ignores this ‘ideal route’ completely through the use of ‘Random Walk’

NEWMAN: BETWEENNESS

- Explain the problems that Freeman (1991) seems to address with flow betweenness and how this solutions is not appropriate according to Newman.

In some sense, our random-walk betweenness and the shortest path betweenness of Freeman (1977) are at opposite ends of a spectrum of possibilities, one end representing information that has no idea of where it is going and the other information that knows precisely where it is going (page 4).

CALCULATION OF RANDOM WALK BETWEENNESS

- *Calculate D-A*
- *Remove single row and corresponding column (eg last)*
- *Inverse resulting matrix*
- *Add back row/column at same index as removed one*
- *Calculate betweenness*

NEWMAN: BETWEENNESS

- What are the implications of only using the *net* number of passes through a node?

Using the net number cancels out two problems:

- The random walker could pass through a node in both directions and thus going back and forth.
- Using multiple runs of a random walk could show passes in either direction.

Both problems are cancelled through the use of net passes.

NEWMAN: BETWEENNESS

- Why calculate betweenness. It often correlates strongly with degree. Think of an example of why it matters.

The interesting property is that the existence of such a correlation among betweenness and degree allows the detection of individual nodes that deviate from this pattern. Network analysis is as much about the global network as it is about detecting properties and roles of individual nodes.

RANDOM WALKS EXTENDED

Cooper, Frieze & Radzik: Multiple Random Walks

- Multiple particles travel random across network.
- Particles travel independently from each other
- Particles can only meet in nodes
- Particles can interact at meetings
- Meetings can come with a time cost

RANDOM WALKS EXTENDED

Cooper, Frieze & Radzik: Multiple Random Walks

- Multiple walks without interaction at meeting.
- Talkative Particles share information and proceed
- Predator-Prey: Preys get eaten at a meeting point. Expected time to extinction can be calculated
- Annihilating Particles destroy each other pairwise
- Coalescing Particles will travel together after meeting

RANDOM WALKS EXTENDED

Alamgir & von Luxburg: Multi-Agent Random Walk for local clustering on graphs

- Random walks are performed by multiple walkers (agents) pairwise connected by a ‘rope’ of fixed length

- The probability that connected walkers travel over a bridge/bottleneck from one cluster to another is small.

If transition probability is p for Random Walk
than the probability becomes p^a for RW with a agents

RANDOM WALKS EXTENDED

Gallesco, Müller & Popov (2010): Spiders in Random Environments

- A Spider is a multi-agent walker with *k* legs.
- At each transition, one leg moves to an adjacent node
- The spider can be bounded by a maximal span, which cannot be exceeded by the any distance between legs