Algorithms for Estimating Relative Importance in Networks

Scott White, Padhraic Smyth
(Information and Computer Science, University of California)

Presented by Martin Leginus

28th of November, 2012

Agenda

- Motivation
- Related work
- Relative importance using weighted paths
- Relative importance using Markov Chains
- Evaluation
- Discussion

Why Relative importance?

- Many datasets can be transformed into graph or network structures
- Need for quantitative tools for analysing graph properties
  - Centrality, latent Euclidean spaces, Hits, PageRank
  - Focus on ranking of relative importance of a node to all other nodes

How to measure a relative importance with respect to a set of root nodes?

Motivation/Contribution

- Given G and r and t, where $\{r, t\} \subset G$, compute the importance of t with respect to the root node r
- Rank all the nodes of a graph according to their importance to the root node r
- Compute importance for a set of root nodes

Related work

Research within several fields:

- Social network analysis: a global importance in the network expressed with centrality measure
- Web ranking: PageRank and HITS algorithms

Only a few works on relative importance with respect to some nodes:

- Personalized PageRank
- Personalized HITS

Notation

A directed graph $G = (V, E)$ consists of two sets: a set of nodes V and a set of edges E.
Each edge e is defined as an ordered pair of nodes (u, v) for directed connection from u to v.
A walk from from u to v is a sequence of edges $(u, u_1), (u_1, u_2) \dots (u_k, v)$.
A walk is a path if no nodes are repeated.

- k-short paths as a set of all paths shorter than k
- $P(u, v)$: a certain set of paths between u and v
- $s_{out}(u)$: a set of distinct outgoing edges from u
- $s_{in}(u)$: a set of distinct ingoing edges towards u
- and $d_{in}(u) = |s_{in}(u)|$ and $d_{out}(u) = |s_{out}(u)|$

Computing relative importance using weighted paths

Two nodes are related according to the paths that connect them. The longer the path is the less important is the relation between two nodes.

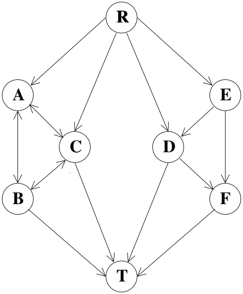
$$I(t|r) = \sum_{i=1}^{|P(r,t)|} \lambda^{|-p_i|} \quad (1)$$

- $P(r, t)$: a set of paths from r to p
- p_i: i-th path in $P(r, t)$
- λ: is a scalar coefficient

Importance decays with path length.
The choice of $P(r, t)$ affects the final importance.

Shortest Paths also called geodesics

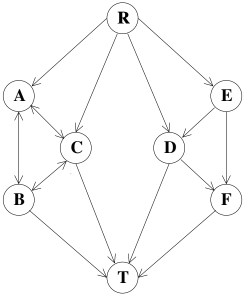
Useful when is possible to ignore all the vertices that do not lie on the geodesics between r and t.



$P(R, T) = \{R - C - T, R - D - T\}$. Ignores importance of A, B, E and F and their importance towards T relative to R.
Widely used in social network analysis for centrality measures as 'closeness' and 'betweness'.

k-Short Paths

A set of all paths from R to T that is shorter than k.

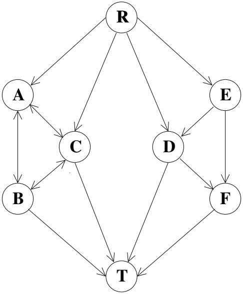


$P(R,T) = \{R-C-T, R-D-T, R-A-B-T, R-C-B-T, R-A-C-T, R-E-F-T, R-E-D-T, R-D-F-T\}$ (3-short paths).

Do not consider "capacity constrains" of nodes or edges.

k-Short Node-Disjoint Paths

A set $k - short$ paths that have neither edges nor nodes in common.



$P(R,T) = \{R - C - T, R - D - T, R - A - B - T, R - E - F - T\}$
(3-short node-disjoint paths).
Enforces "capacity constrains" on vertices and edges.

Computing relative importance using Markov Chains

A graph represents a stochastic process - a 1st order Markov chain. *Imagine a token that stochastically traverses a graph for an infinitely long time. The probability of moving from the current node to a next node is conditioned by the properties of the current node. A time that a token spends at a particular node can be interpreted as a global importance of the node with respect to all other nodes.*

- An improved version is PageRank where a random-surfer approach is introduced
- Usually, a probability of moving from node i to j is defined as:

$$p(i|j) = \frac{1}{d_{out}(j)} \quad (2)$$

Markov Centrality

Inverse of the mean first-passage time in the Markov chain.

$$m_{r,t} = \sum_{n=1}^{\infty} n \cdot f_{r,t}^{(n)} \tag{3}$$

It can be interpreted as an expected number of steps taken until a first arrival to a node t from r.

- n is a number of taken steps
- $f_{r,t}^{(n)}$ is a probability that the chain returns to t from r in exactly n steps

Markov Centrality 1

The mean first passage matrix is defined as:

$$M = (I - Z + E \cdot Z_{dg}) \cdot D \quad (4)$$

- I is an identity matrix and E is a matrix of ones
- D is a diagonal matrix with elements $d_{v,v} = \frac{1}{\pi v}$ where πv is a stationary distribution of node v
- Z_{dg} is a diagonal matrix where elements are from fundamental matrix Z
- $Z = (I - A - e\pi^T)^{-1}$ where A is the Markov transition probability matrix, e is vector of 1 and π is a column vector of the stationary probabilities for the Markov chain
- $f_{r,t}^{(n)}$ is a probability that the chain returns to t from r in exactly n steps

Markov Centrality 2

The importance of a node t with respect to root nodes R is defined as:

$$I(t|R) = \frac{1}{\frac{1}{|R|}\sum_{r \in R} m_{r,t}} \quad (5)$$

A complexity is $O(V^3)$

It reflects the notion of how central a given node t is in a network relative to a root node r.

PageRank with priors

Relative importance to a root node is introduced through a vector of prior probabilities p_r.
A random surfer is assured with a back probability β - determines how often we jump back to a root node.

$$\pi(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1} d_{in}(v)p(v|u)\pi^{(i)}(u)\right) + \beta p_v \quad (6)$$

The resulting ranks biased towards r are considered as definition of importance after convergence i.e.;

$$I(v|R) = \pi(v) \quad (7)$$

HITS with priors

Relative importance to a root node is introduced through a vector p_r of prior probabilities.
A random surfer is assured with a back probability β - determines how often we jump back to a root node.

$$a(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1} d_{in}(v)\frac{h^{(t)}(u)}{H^{(i)}}\right) + \beta p_v \tag{8}$$

$$h(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1} d_{out}(v)\frac{a^{(t)}(u)}{A^{(i)}}\right) + \beta p_v \tag{9}$$

The resulting ranks (stationary distribution of each node) biased towards r are considered as definition of importance after convergence i.e.;

$$I(v|R) = \pi(v) \tag{10}$$

k-step Markov approach

A random surfer is assured with a path lenght limitation - determines how often we jump back to a root node. Relative importance to a root node is introduced through a vector p_r of prior probabilities.

$$I(v|R) = [A \cdot p_R + A^2 \cdot p_R \ldots A^K \cdot p_R] \tag{11}$$

The resulting ranks (stationary distribution of each node) biased towards r are considered as definition of importance after convergence i.e.;

Evaluation on simulated data

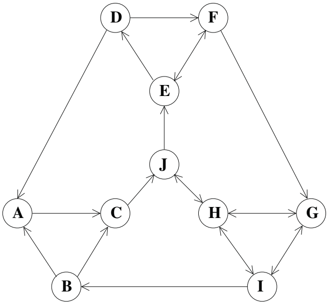


Table 1: Importance rankings for the nodes in Figure 3 with respect to nodes A and F.

Rank	PRankP		HITSPa		HITSPh		WKPaths		MarkovC		KSMarkov	
1	F	0.200	A	0.252	F	0.225	F	0.206	J	0.180	H	.146
2	A	0.167	F	0.241	A	0.186	A	0.206	C	0.133	G	.142
3	C	0.122	G	0.128	D	0.162	E	0.116	G	0.130	E	.142
4	E	0.107	C	0.110	B	0.119	C	0.108	H	0.129	J	.140
5	J	0.105	E	0.099	E	0.090	G	0.095	E	0.111	C	.120
6	G	0.103	H	0.052	I	0.067	J	0.068	I	0.101	I	.098
7	H	0.086	D	0.032	H	0.061	H	0.066	F	0.069	F	.087
8	I	0.056	I	0.032	J	0.050	I	0.052	D	0.051	D	.061
9	D	0.037	J	0.025	G	0.028	D	0.052	A	0.047	A	.034
10	B	0.013	B	0.024	C	0.008	B	0.026	B	0.044	B	.024

September 11th Terrorist Network

The terrorist network graph consists of 63 nodes and 308 edges. It contains also 19 hijackers from 11th of September.

Table 2: Importance rankings for the terrorist network with respect to nodes Khemais and Beghal.

Rank	PRankP		HITSP		WKPaths		MarkovC		KSMarkov	
1:	Khemais	0.221	Khemais	0.173	Beghal	0.045	Atta	0.063	Khemais	0.115
2:	Beghal	0.218	Beghal	0.166	Khemais	0.045	Al-Shehhi	0.041	Beghal	0.108
3:	Moussaoui	0.044	Atta	0.038	Moussaoui	0.045	al-Shibh	0.037	Moussaoui	0.065
4:	Maaroufi	0.039	Moussaoui	0.029	Maaroufi	0.044	Moussaoui	0.036	Maaroufi	0.059
5:	Qatada	0.036	Maaroufi	0.026	Bensakhria	0.037	Jarrah	0.030	Qatada	0.052
6:	Daoudi	0.035	Qatada	0.025	Daoudi	0.037	Hanjour	0.028	Daoudi	0.049
7:	Courtaillier	0.032	Bensakhria	0.023	Qatada	0.036	Al-Omari	0.026	Bensakhria	0.045
8:	Bensakhria	0.031	Daoudi	0.023	Walid	0.031	Khemais	0.025	Courtaillier	0.045
9:	Walid	0.030	Courtaillier	0.022	Courtaillier	0.031	Qatada	0.025	Walid	0.040
10:	Khammoun	0.025	Khammoun	0.021	Khammoun	0.029	Bahaji	0.024	Khammoun	0.034

Biotech Collaborative Network

The biotech network data set contains 2700 nodes (companies & collaborators). Collaborations (8690 edges) include finance, R&D and commercial ventures.

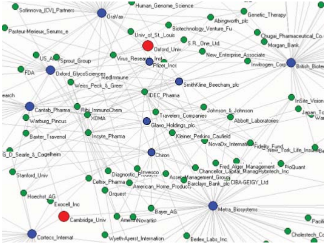


Figure 5: A portion of the biotechnology network.

The task was to find the most relevant authorities related to Oxford and Cambridge Universities.

Biotech Collaborative Network

The biotech network data set contains 2700 nodes (companies & collaborators). Collaborations (8690 edges) include finance, R&D and commercial ventures.

The task was to find the most relevant authorities related to Oxford and Cambridge Universities.

Table 3: Importance rankings for the biotechnology network with respect to nodes Cambridge University and Oxford University.

Rank	PRankP		HITSP		WKPaths		KSMarkov	
1:	CambridgeU	0.1537	OxfordU	0.1510	OxfordU	0.0020	Cortecs	0.0616
2:	OxfordU	0.1531	CambridgeU	0.1510	CambridgeU	0.0020	Cantab	0.0559
3:	Cortecs	0.0480	Metra	0.0088	OxfordGlyco	0.0016	BritishBio	0.0550
4:	Cantab	0.0453	BritishBio	0.0084	Cantab	0.0016	Metra	0.0532
5:	BritishBio	0.0451	OraVax	0.0080	OraVax	0.0016	OraVax	0.0510
6:	Metra	0.0443	Cantab	0.0075	BritishBio	0.0016	OxfordGlyco	0.0428
7:	OraVax	0.0432	OxfordGlyco	0.0072	Glaxo	0.0015	Pfizer	0.0069
8:	OxfordGlyco	0.0395	Cortecs	0.0072	Metra	0.0015	Glaxo	0.0066
9:	Pfizer	0.0046	NIH	0.0068	SmithKline	0.0014	Incyte	0.0066
10:	Glaxo	0.0044	Chiron	0.0055	Pfizer	0.0014	CambridgeU	0.0056

The CITESEER Co-Authorship Network

Data set consists of 387703 papers from period 1991 till 2002.

Table 4: Importance rankings for the coauthorship network with respect to the Tom Mitchell node.

Rank	PRankP		HITSP		WKPaths		KSMarkov	
1	Mitchell	0.342	Mitchell	0.322	Mitchell	0.005	McCallum	0.070
2	Freitag	0.054	Thrun	0.038	Thrun	0.004	Freitag	0.067
3	McCallum	0.054	McCallum	0.038	Freitag	0.003	Mitchell	0.067
4	Thrun	0.051	Freitag	0.035	McCallum	0.003	Thrun	0.064
5	Joachims	0.050	Nigam	0.034	Nigam	0.002	Joachims	0.061
6	Armstrong	0.046	Blum	0.032	Joachims	0.002	Armstrong	0.054
7	Nigam	0.040	Joachims	0.031	Armstrong	0.002	Nigam	0.046
8	Blum	0.036	Armstrong	0.031	Blum	0.002	Blum	0.041
9	O'Sullivan	0.035	O'Sullivan	0.030	O'Sullivan	0.002	O'Sullivan	0.038
10	Seymore	0.011	Seymore	0.006	Caruana	0.001	Seymore	0.019

Table 5: Importance rankings for the coauthorship network with respect to nodes Brin, Page, and Kleinberg.

Rank	PRankP		HITSP		WKPaths		KSMarkov	
1:	Brin	0.2014	Brin	0.1119	Kleinberg	0.0023	Brin	0.1045
2:	Page	0.1352	Kleinberg	0.1107	Brin	0.0019	Motwani	0.0627
3:	Kleinberg	0.1137	Page	0.1087	Motwani	0.0017	Ullman	0.0536
4:	Motwani	0.0474	Motwani	0.0184	Raghavan	0.0016	Silverstein	0.0467
5:	Ullman	0.0429	Raghavan	0.0147	Page	0.0014	Page	0.0394
6:	Silverstein	0.0392	Ullman	0.0136	Silverstein	0.0014	Kleinberg	0.0194
7:	Raghavan	0.0111	Silverstein	0.0119	Ullman	0.0014	Raghavan	0.0138
8:	Lynch	0.0086	Williamson	0.0113	Williamson	0.0012	Zhang	0.0109
9:	Kedem	0.0086	Papadimitriou	0.0110	Vempala	0.0012	Guibas	0.0106
10:	Williamson	0.0085	Lynch	0.0108	Indyk	0.0010	Robertson	0.0101

Correlations of ranked lists

Table 6: Correlations of top-10 rankings in Table 2.

	PRankP	HITSP	WKPaths	MarkovC	KSMarkov
PRankP	1	0.80	0.87	0.47	0.98
HITSP	0.80	1	0.76	0.52	0.82
WKPaths	0.87	0.76	1	0.44	0.89
MarkovC	0.47	0.52	0.44	1	0.43
KSMarkov	0.98	0.82	0.89	0.43	1

Conclusion and future work

- General framework for importance estimation of nodes in a graph relative to some root nodes
- How weighted edges can be incorporated into models?
- Usage for my PhD project
  - Graph based tag cloud generation
  - Fraud detection for SKAT project