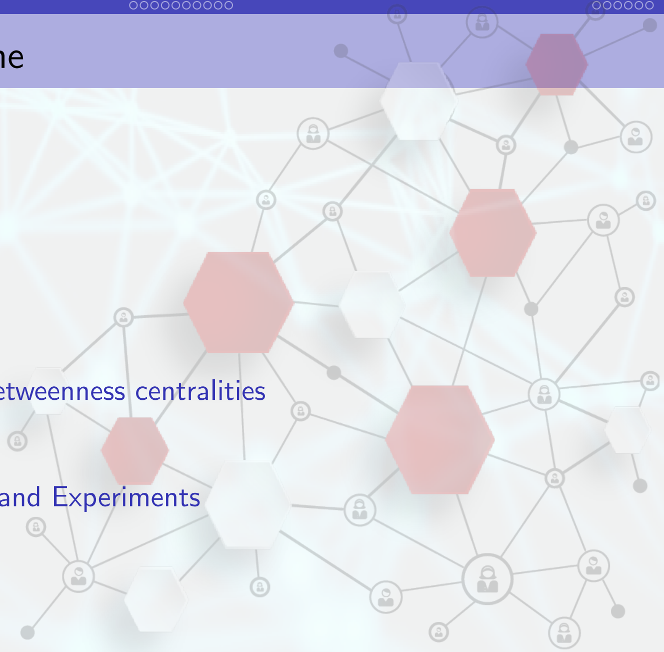
Two betweenness centrality measures based on Randomized Shortest Paths

Prova Integrativa - Complessità nei Sistemi e nelle Reti

Matteo Bonfadini
Politecnico di Milano

Outline



1. Goal
2. RSP betweenness centralities
3. Limits and Experiments

RSP framework (1/2)

One of the most fundamental topics in network science is determining the *centrality* of a node in a network accordingly to its structure. The concept of centrality can be interpreted in many ways and a vast number of measures have been proposed.

We aim to introduce two new closely related betweenness centrality measures based on the Randomized Shortest Paths (RSP) framework:

- the simple RSP betweenness
- the RSP net betweenness

RSP framework (2/2)

The RSP framework is based on *Boltzmann probability distributions* over paths between the nodes of a network which focus on short, optimal paths, but give some probablity mass also to longer paths.

The Boltzmann probabilities describe the probability that a thermodynamic system is in a particular state, given a certain *energy* value of that system.

Rather than the energy, we'd like to control the focus on optimal paths with an inverse temperature parameter β.

Usefulness (1/2)

Measures based only on shortest paths or random walks alone often involve undesirable features:

- highly skewed betweenness score distribution in complex networks;
- not always realistic to consider that navigating agents would occur along only the shortest paths;
- random walks may depend heavely on local features of a graph, especially for large graphs.

Usefulness (2/2)

Because of this, RSP's measures can help:

- in detecting bottlenecks of the networks, where there exist no alternatives for the shortest path;
- in introducing some regularisation over the degree of randomness, which is controlled by the inverse temperature parameter β.

In addiction, the RSP framework has previously been used for defining distance measures in many data analysis tasks such as clustering and classification of graph nodes.

Notation and terminology (1/2)

- We consider weighted directed graphs $G = (V, E)$ with node set $V = \{1, 2, \ldots, n\}$ and edge set $E = \{(i, j)\}$ of m edges.
- Each edge of the graph is associated with a **weight** a_{ij} and a **cost** c_{ij}:
  - The weights are collected in the non-symmetric $n \times n$ *adjacency matrix* $\mathbf{A}$ and they reflect the strength of connection between adjacent nodes.
  - The edge weights define the *reference transition probability matrix* $\mathbf{P}^{\text{ref}}$, which can be computed as

$$\mathbf{P}^{\text{ref}} = \mathbf{D}^{-1}\mathbf{A}, \qquad \mathbf{D} = \text{diag}(k_1, \ldots, k_n)$$

or

$$p_{ij}^{\text{ref}} = \frac{a_{ij}}{\sum_j a_{ij}} = \frac{a_{ij}}{s_i^{\text{out}}}$$

Notation and terminology (2/2)

- The edge costs, in contrast to weights, describe the distance of adjacent nodes. The cost of a path $\mathscr{P}$ is

$$\widetilde{c}(\mathscr{P}) = \sum_{(i,j)\in\mathscr{P}} c_{ij}$$

In many situations the edge costs and edge weights can be defined based on one another, for istance, as $c_{ij} = 1/a_{ij}$.
However, in the RSP framework the weights can be independent of the costs, thus the transition probabilities do not depend on the costs. This means that:

- ▷ the edge costs define the interpretation of shortest paths, i.e. the *low temperature behavior* of the system;
- ▷ the edge weights determine the interpretation of a random walk, i.e. the *high temperature behavior*.

Minimization of expected cost (1/2)

For semplicity, we restrict the RSP framework to absorbing paths (for istance, take a directed strongly connected network).

The RSP framework is based on the probability distribution over the set $\mathcal{P}_{st}$ of absorbing s-t-walks for which the expected cost of the walk is minimal when constrained with a fixed relative entropy w.r.t. the reference path probability distribution. Formally, we seek for the solution to the following problem:

$$\min_{\widetilde{P}_{st}} \sum_{\mathscr{P}\in\mathcal{P}_{st}} \widetilde{P}_{st}(\mathscr{P})\,\widetilde{c}(\mathscr{P}) \quad \text{s.t.} \quad \begin{cases} J\left(\widetilde{P}_{st}\,||\,\widetilde{P}^{\text{ref}}_{st}\right) = J_0 \\ \displaystyle\sum_{\mathscr{P}\in\mathcal{P}_{st}} \widetilde{P}_{st}(\mathscr{P}) = 1 \end{cases}$$

Minimization of expected cost (2/2)

The solution is a Boltzmann distribution:

$$\widetilde{P}_{st}(\mathscr{P}) = \frac{\widetilde{P}^{\text{ref}}_{st}(\mathscr{P})\, e^{-\beta \widetilde{c}(\mathscr{P})}}{\mathcal{Z}_{st}}$$

where

$$\mathcal{Z}_{st} = \sum_{\mathscr{P} \in \mathcal{P}_{st}} \widetilde{P}^{\text{ref}}_{st}(\mathscr{P})\, e^{-\beta \widetilde{c}(\mathscr{P})}$$

is the *partition function* of absorbing *s*-*t*-walks.

Matrix formulation (1/1)

Concerning the computation of $\mathcal{Z}_{st}$, we have that

$$\mathcal{Z}_{st} = \frac{z_{st}}{z_{tt}}$$

where z_{st} is the (s,t)-element of the *fundamental matrix of non-absorbing paths*

$$\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1} \quad \text{with} \quad \mathbf{W} = \mathbf{P}^{\text{ref}} \circ e^{-\beta \mathbf{C}}$$

The matrix $\mathbf{W}$ is a substochastic matrix and can be interpreted as defining a *killed random walk*. Hence, one can see the partition function $\mathcal{Z}_{st}$ as the probability of a walker surviving the walk from s to t.

Simple RSP betweenness (1/3)

The simple RSP betweenness centrality of a node *i* is

$$\mathrm{bet}_i^{\mathrm{RSP}} = \sum_{s,t=1}^{n} \mathrm{bet}_i^{\mathrm{RSP}}(s,t)$$

where $\mathrm{bet}_i^{\mathrm{RSP}}(s,t)$ is the RSP simple betweenness of the node *i* w.r.t. absorbing paths from *s* to *t*, i.e. the expected number of visits through *i* over all *s*-*t*-walks w.r.t. the RSP solution probabilities:

$$\mathrm{bet}_i^{\mathrm{RSP}}(s,t) = \begin{cases} 0 & \text{if } \nexists\ s\text{-}t\text{-path} \\ \sum_{j=1}^{n} \overline{\eta}_{ij}(s,t) & \text{otherwise} \end{cases}$$

Simple RSP betweenness (2/3)

where

$$\overline{\eta}_{ij}(s,t) = -\frac{1}{\beta}\frac{\partial \log \mathcal{Z}_{st}}{\partial c_{ij}} = -\frac{1}{\beta}\frac{\partial \log\left(z_{st}/z_{tt}\right)}{\partial c_{ij}} = \left(\frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}}\right) w_{ij} z_{jt}$$

In other words, the total flow transiting through node i is

$$\text{bet}_i^{\text{RSP}}(s,t) = \underbrace{\left(\frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}}\right)}_{=0 \text{ if } i=t} \underbrace{\sum_j w_{ij} z_{jt}}_{\substack{=z_{it} \text{ if } i\neq t \\ \mathbf{Z}=(\mathbf{I}-\mathbf{W})^{-1} \\ \mathbf{Z}(\mathbf{I}-\mathbf{W})=\mathbf{I} \\ \mathbf{Z}=\mathbf{Z}\mathbf{W}+\mathbf{I}}} = \left(\frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}}\right) z_{it}$$

Simple RSP betweenness (3/3)

Finally, the simple RSP betweenness of node i is

$$\begin{aligned}
\mathrm{bet}_i^{\mathrm{RSP}} &= \sum_{s,t} \mathrm{bet}_i^{\mathrm{RSP}}(s,t) = \\
&= \sum_{s,t} \left(\frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}} \right) z_{it} = \\
&= \left[\mathbf{diag} \left(\mathbf{Z} \left(\mathbf{Z}^{\div} - n\mathbf{Diag}\left(\mathbf{Z}^{\div} \right) \right)^T \mathbf{Z} \right) \right]_i
\end{aligned}$$

The vector $\mathbf{bet}^{\mathrm{RSP}}$ of all betweenness values is computed accordingly.

Computing the simple RSP betweenness (1/1)

Input:

- a directed strongly connected graph G with n nodes
- the $n \times n$ reference transition probability matrix $\mathbf{P}^{\text{ref}} = \mathbf{D}^{-1}\mathbf{A}$
- the $n \times n$ cost matrix $\mathbf{C}$
- the inverse temperature parameter β

Output:

1. $\mathbf{W} = \mathbf{P}^{\text{ref}} \circ e^{-\beta \mathbf{C}}$
2. $\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}$
3. $\mathbf{Z}^{\div} = \mathbf{e}\mathbf{e}^{T} \div \mathbf{Z}$
4. return $\mathbf{bet}^{\text{RSP}} = \mathbf{diag}\left(\mathbf{Z}\left(\mathbf{Z}^{\div} - n\mathbf{Diag}\left(\mathbf{Z}^{\div}\right)\right)^{T}\mathbf{Z}\right)$

This highlight the computational bottleneck of the algorithm: the matrix inversion, which, in general, has complexity $\mathcal{O}\left(n^3\right)$.

RSP net betweenness (1/1)

Instead of only considering the overall outgoing flow of random walkers it may in some cases make more sense to compute the net outgoing flow, i.e. so that the outgoing and ingoing flows through one edge neutralize each other. This corresponds to the random walk interpretation of the *current flow betweenness* in undirected graphs.

According to this approach, we define the *RSP net betweenness centrality* of node i as

$$\text{bet}_i^{\text{RSPnet}} = \sum_{s,t} \sum_{j} \left| \overline{\eta}_{ij}(s,t) - \overline{\eta}_{ji}(s,t) \right|$$

Limits (1/2)

$$\widetilde{P}_{st}(\mathscr{P}) = \frac{\widetilde{P}^{\mathrm{ref}}_{st}(\mathscr{P})\, e^{-\beta\widetilde{c}(\mathscr{P})}}{\mathcal{Z}_{st}}, \qquad \mathcal{Z}_{st} = \sum_{\mathscr{P}\in\mathcal{P}_{st}} \widetilde{P}^{\mathrm{ref}}_{st}(\mathscr{P})\, e^{-\beta\widetilde{c}(\mathscr{P})}$$

- At the limit $\beta \to \infty$. In the low temperature limit, the RSP probability distribution focuses solely on shortest paths. Thus

$$\widetilde{P}_{st}(\mathscr{P}) \xrightarrow{\beta\to\infty} \begin{cases} 0, & \text{if } \mathscr{P} \notin \mathcal{P}^*_{st} \\ \dfrac{\widetilde{P}^{\mathrm{ref}}_{st}(\mathscr{P})}{\displaystyle\sum_{\mathscr{P}\in\mathcal{P}^*_{st}} \widetilde{P}^{\mathrm{ref}}_{st}(\mathscr{P})} & \text{if } \mathscr{P} \in \mathcal{P}^*_{st} \end{cases}$$

In other words, the simple RSP betweenness converges to the *shortest path likelihood betweenness*.

Limits (2/2)

The same result holds for the RSP net betweenness. Intuitively, as the path distribution focuses more and more on the shortest paths, one of the two terms of in and outgoing flows becomes zero, as the walker will only move in one direction.

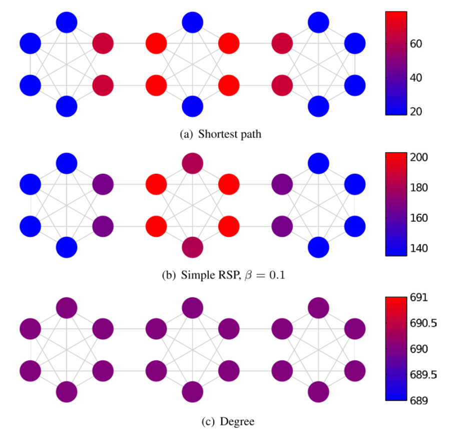
- At the limit $\beta \to 0^+$. In the high temperature limit the RSP probabilities converge to the *unbiased random walk probabilities*, determined by the reference transition probabilities, i.e.

$$\widetilde{P}_{st} \xrightarrow{\beta \to 0^+} \widetilde{P}_{st}^{\text{ref}} \qquad \forall \mathscr{P}$$

This means that the simple RSP betweenness is proportional to the stationary distribution π s.t. $\pi = \left(\mathbf{P}^{\text{ref}}\right)^T \pi$.

Artificial example (1/1)

One possible use for the RSP betweenness measures is the detection of groups of nodes that are central in a network.



(a) Shortest path

(b) Simple RSP, $\beta = 0.1$

(c) Degree

Manhattan street network. (1/3)

One promising application area for RSP's are path planning problems.

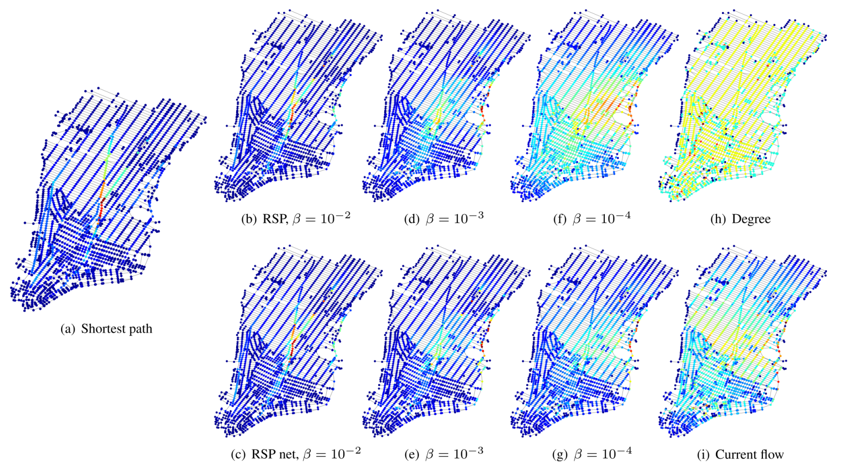
We illustrate the use of RSP's for routing in a network by analyzing the street network of Midtown and Lower Manhattan.

Manhattan street network. (2/3)

The nodes in the network correspond to intersections and the edges are the street segments between the intersections; we treat the network as undirected.

The length of each street segment is assigned as the cost of the corresponding edge. Accordingly, the overall cost of a path is its overall length. However, we define here the reference transition probabilities of the random walk according to the degree of each node, $p_{ij}^{\text{ref}} = 1/k_i$, i.e. only according to the number of edges connected to the node and independent of the edge costs.

Manhattan street network. (3/3)



(a) Shortest path

(b) RSP, $\beta = 10^{-2}$

(d) $\beta = 10^{-3}$

(f) $\beta = 10^{-4}$

(h) Degree

(c) RSP net, $\beta = 10^{-2}$

(e) $\beta = 10^{-3}$

(g) $\beta = 10^{-4}$

(i) Current flow