

# SCIENTIFIC REPORTS

OPEN

## Two betweenness centrality measures based on Randomized Shortest Paths

Ilkka Kivimäki<sup>1,2</sup>, Bertrand Lebichot<sup>1</sup>, Jari Saramäki<sup>2</sup> & Marco Saerens<sup>1</sup>

Received: 21 August 2015

Accepted: 04 December 2015

Published: 01 February 2016

This paper introduces two new closely related betweenness centrality measures based on the Randomized Shortest Paths (RSP) framework, which fill a gap between traditional network centrality measures based on shortest paths and more recent methods considering random walks or current flows. The framework defines Boltzmann probability distributions over paths of the network which focus on the shortest paths, but also take into account longer paths depending on an inverse temperature parameter. RSP's have previously proven to be useful in defining distance measures on networks. In this work we study their utility in quantifying the importance of the nodes of a network. The proposed RSP betweenness centralities combine, in an optimal way, the ideas of using the shortest and purely random paths for analysing the roles of network nodes, avoiding issues involving these two paradigms. We present the derivations of these measures and how they can be computed in an efficient way. In addition, we show with real world examples the potential of the RSP betweenness centralities in identifying interesting nodes of a network that more traditional methods might fail to notice.

One of the most fundamental and popular topics in network science is determining the *centrality* of a node in a network according to the structure of the network. The concept of centrality can be interpreted in many ways and a vast number of measures have been proposed based on different interpretations. One commonly used interpretation is *betweenness centrality*, which reflects the extent to which a node lies in between pairs or groups of other nodes of the graph. This can be also stated as the extent to which a node is an intermediate in communication over the network. Different models have been proposed to measure the participation of a node in this communication ranging from the *shortest path betweenness centrality* of Freeman<sup>1–3</sup>, which considers communication flowing only along the shortest paths, to the *current flow betweenness centrality*<sup>4,5</sup>, which interprets communication flowing as electric current or as random walks in the network.

In this article we propose two families of betweenness centrality measures based on the *Randomized Shortest Paths* (RSP) framework<sup>6–8</sup>. The framework is based on Boltzmann probability distributions over paths between the nodes of a network which focus on short, optimal paths, but give some probability mass also to longer paths. The extent of focus on optimal paths is controlled by an inverse temperature parameter  $\beta$ . The RSP framework has previously been shown to function well when defining distance measures on networks for clustering and classification of network nodes<sup>8</sup>. In this work we extend the study of RSP's by showing their potential also in defining centrality measures. The two RSP betweenness centralities presented in this paper measure the involvement of each node in RSP's between the nodes of the graph. The first measure, which we call the simple RSP betweenness, measures the expected number of visits to a node during RSP's, while the second one, called the RSP net betweenness is the sum of expected net flows over the edges connected to a node.

The proposed RSP betweenness measures are attractive both theoretically as well as in practice. Theoretical interest is ensured by the fact that both measures can be seen as generalizations of classical betweenness measures. With large values of the parameter  $\beta$ , both RSP betweenness measures converge to a measure that we introduce as the shortest path likelihood betweenness, which is very closely related to the original betweenness centrality defined by Freeman<sup>1–3</sup> and its other similar variant, the load centrality<sup>9</sup>. The reason for defining two different betweenness measures with RSP's is in their behaviour as  $\beta$  is decreased. Namely, the simple RSP betweenness then converges to the stationary distribution of a random walk on the network (multiplied by a constant), whereas the RSP net betweenness converges to the current flow betweenness<sup>4,5</sup>. The definition, as well as the computation,

<sup>1</sup>Université catholique de Louvain, ICTEAM/LSM, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium. <sup>2</sup>Aalto University, Department of Computer Science, P.O.Box 15400, FI-00076 Aalto, Finland. Correspondence and requests for materials should be addressed to I.K. (email: ilkka.kivimaki@aalto.fi)

of the simple RSP betweenness are more straightforward than for its counterpart, the RSP net betweenness, which can also be stated of their corresponding limit functions. In addition, our experiments indicate that the simple RSP betweenness can in practice be more useful than the RSP net betweenness or the current flow betweenness. However, the choice of which definition to rely on in the end depends on the application domain.

Considering betweenness based on RSP's is motivated by the fact that measures based only on shortest paths or on random walks alone often involve undesirable features. Shortest paths in a complex network tend to pass through only a small fraction of the nodes of the network, which can cause highly skewed betweenness score distributions and fail in differentiating between the other nodes of the network<sup>9,10</sup>. Also, when considering communication or navigation in networks, it is not always even realistic to consider that they would occur along only the shortest, nor completely random paths. Instead, movement may follow a partly random route, with a drift towards a destination, for example when the navigating agent does not know the optimal way or wants to add secrecy and unpredictability to its route. As a result, the trajectories that are actually used in the network can be more spread over different nodes in areas of the network that contain many connections. Because of this, RSP's can help, for instance, in detecting bottlenecks of the network, where there exist no alternatives for the shortest path.

Measures based solely on random walks do take into account the abundance of connections between nodes. However, they may in many situations depend heavily on local features of a graph, especially for large graphs<sup>11–13</sup> instead of capturing its interesting global properties. Thus, some regularisation over the degree of randomness is needed, which in the RSP framework is controlled by the inverse temperature parameter  $\beta$ .

Models that find a compromise between the optimal shortest path and a random walk have recently received a lot of attention. Compared to these, the attractive aspect in using the RSP framework is that the compromise between the shortest and random paths is optimal by definition, as the Boltzmann distribution minimizes the expected cost of paths subject to a fixed relative entropy<sup>6,7</sup>. The minimization can also be expressed with respect to free energy<sup>8</sup>. In addition to the optimality aspect, the computation of quantities related to RSP's is fairly straightforward and efficient. A drawback of the algorithms presented in the paper is that as such they are not tractable with very large networks. However, in the future we plan to develop more specialized methods that will enable the RSP quantities to be computed with large networks as well.

Ideas similar to the RSP framework of interpolating between the two extremes can be found in the work of Alamgir and von Luxburg<sup>14</sup>, involving  $p$ -resistances and graph node distances based on them, of Chebotarev involving distances based on the matrix forest theorem<sup>15,16</sup>, of Zhang and Boley<sup>17</sup> with focus on routing schemes, of Estrada, who has defined the subgraph centrality<sup>18</sup> and the communicability betweenness<sup>19</sup>. Even more related to this paper are the recent works of Bavaud and Guex<sup>20</sup> and Lebichot *et al.*<sup>21</sup>. In fact, the two betweenness centrality measures presented in this paper can be shown equal to the betweenness measures proposed by Bavaud and Guex<sup>20</sup>, although the relation between the two works is not entirely obvious. We will discuss this relation and the relation between the RSP betweenness measures and other previously proposed betweenness measures in more detail later. To sum up, the contributions of this paper are:

- We define two betweenness centrality measures which form a spectrum between measures based on shortest paths and pure random walks, therefore integrating information about both the optimality and the abundance of paths between nodes of the network,
- We derive algorithms for computing these measures in a convenient and efficient way, and
- We demonstrate with example networks that the proposed RSP betweenness measures may provide a more interesting ranking of the network nodes than the classical measures that they generalize.

## Notation and terminology

In the paper we consider weighted directed graphs  $G = (V, E)$  with node set  $V = \{1, 2, \dots, n\}$  and edge set  $E = \{(i, j)\}$  of  $m$  edges. We define a path, or walk, interchangeably, as a sequence of nodes  $\varphi = (i_0, \dots, i_T)$ , where  $T > 0$  and  $(i_\tau, i_{\tau+1}) \in E$  for all  $\tau = 0, \dots, T-1$ . A path is *absorbing*, if the last node of the path appears on the path only once. We denote the set of all absorbing paths starting from node  $s$  and ending in node  $t$ , i.e.  $s$ - $t$ -paths or  $s$ - $t$ -walks, by  $\mathcal{P}_{st}$ .

The weights on edges,  $a_{ij}$ ,  $(i, j) \in E$ , reflect the similarity or strength of connection between adjacent nodes and form the *adjacency matrix*  $\mathbf{A}$  of the graph. The edge weights define the *reference transition probabilities* of the *unbiased random walk* as  $p_{ij}^{\text{ref}} = a_{ij} / \sum_j a_{ij}$ . The transition probabilities form the *reference transition probability matrix*  $\mathbf{P}^{\text{ref}}$ , which can be computed as  $\mathbf{P}^{\text{ref}} = \mathbf{D}^{-1} \mathbf{A}$ , where  $\mathbf{D}$  is the diagonal matrix containing the row sums of  $\mathbf{A}$ .

The reference path probability  $\tilde{P}_{st}^{\text{ref}}(\varphi)$  of a path  $\varphi \in \mathcal{P}_{st}$  is simply defined as the product of the transition probabilities along the path. In addition to the weights, the edges are also assigned costs  $c_{ij}$ , which, in contrast to weights, can be considered also as the dissimilarity or distance of adjacent nodes. The cost of a path  $\varphi$  is then simply defined as  $\tilde{c}(\varphi) = \sum_{(i,j) \in \varphi} c_{ij}$ . Accordingly, we use the term shortest path to mean the path between two nodes with the lowest cost over all paths between the nodes. We denote the set of shortest paths from node  $s$  to node  $t$  by  $\mathcal{P}_{st}^*$ , the total number of such paths by  $|\mathcal{P}_{st}^*|$  and the cost of the shortest path from  $s$  to  $t$  by  $\tilde{c}_{st}^*$ . The directed graph that consists only of the nodes and directed edges that belong to the shortest paths from  $s$  to  $t$  is called the *directed shortest path graph* from  $s$  to  $t$ .

In many situations the edge costs and edge weights can be defined based on one another, for example, as reciprocals  $c_{ij} = 1/a_{ij}$ , which corresponds to the interpretation of costs as resistances and weights as conductances in an electric circuit. This convention is also normally used when computing centrality measures on weighted networks. However, in general in the RSP framework the weights can be independent of the costs, as long as  $c_{ij} < \infty$  whenever  $a_{ij} > 0$ . Accordingly, the transition probabilities and thus the unbiased random walk can be independent of the costs. This means that the edge costs define the interpretation of shortest paths, i.e. the low temperature

behavior of the system, whereas the edge weights determine the interpretation of a random walk, i.e. the high temperature behavior. The interplay between weights and costs is thus similar to a tradeoff between *exploration*, based on local possible movements, and *exploitation*, based on long-term preferred movements.

## Graph node centrality

The concept of graph node centrality has many interpretations, and for most of the interpretations there exists a lengthy catalog of different proposed measures, which are often derived for different application purposes. In addition there have been efforts of stating axioms that a centrality measure should have<sup>22,23</sup>. Also, as discussed by Kolaczyk<sup>24</sup>, there have been attempts to define a typology of centrality measures, for instance by Borgatti<sup>25</sup>. In an interesting recent work, Brandes and Hildenbrand study the problem of finding minimal graphs for which different centrality measures rank different nodes as the most central<sup>26</sup>. In this section, we make a brief survey of the different centrality measures with special focus on betweenness centrality measures.

There are different possibilities in whether or not the source and target nodes,  $s$  and  $t$  should be considered also as intermediate nodes of a path for centrality considerations. In this paper we use the convention where the first and last node of a path do affect the betweenness scores. For betweenness measures based on shortest paths this choice only changes the overall betweenness scores by an additive constant in a strongly connected graph, and thus only affects the ranking of nodes when the network has several components. In contrast, when measuring betweenness based on random walks, further visits to the starting node  $s$  after the first step may increase the betweenness score of node  $s$  and thus may also affect the rankings. Betweenness measures are also often normalized, e.g., according to the number of possible node pairs, when considering shortest paths between nodes. However, in this paper we leave the normalization out of consideration in all the definitions, because it never affects the rankings of nodes within a strongly connected network.

## Betweenness centralities.

*Betweenness based on shortest paths.* Possibly the best-known centrality measure of all is the original *betweenness centrality* of Freeman<sup>1–3</sup>, which counts the fraction of shortest paths between a pair of nodes that an intermediate node lies on and sums these fractions over all node pairs. We will also refer to it as *shortest path betweenness*, for specificity. Formally, the shortest path betweenness centrality of node  $i$  can be expressed as

$$C_i = \sum_{s,t=1}^n \frac{n(i \in \mathcal{P}_{st}^*)}{|\mathcal{P}_{st}^*|}, \quad (1)$$

where  $n(i \in \mathcal{P}_{st}^*)$  means the number of paths that contain node  $i$ . Notice that if there are more than one shortest path connecting  $s$  to  $t$ , each of these paths will contribute a score of  $1/|\mathcal{P}_{st}^*|$  to the betweenness of the nodes on them.

There are several variations of the shortest path betweenness defined above. A thorough review of these variants and their efficient computation was provided by Brandes<sup>27</sup>. One variant, called the *load centrality*<sup>9,27,28</sup>, replaces the fractional term inside the sum in (1) with the *branching probability* of a path, i.e. the probability that a random walker moving in the directed shortest path graph from  $s$  to  $t$  follows a path that contains node  $i$ , when at each branching point in this graph it selects the edge to follow with uniform probability over all outgoing edges. For illustration of the difference between the shortest path betweenness and load betweenness, see<sup>27</sup>. With weighted graphs the two measures are usually equal, as there often exists a unique shortest path between all nodes, especially if the weights are real-valued.

In what follows, we will need to consider another variant of the shortest path betweenness, which we have not encountered in the literature previously. We call this measure the *shortest path likelihood betweenness*. Similarly to the load centrality, we define the likelihood betweenness by replacing the term inside the sum in (1) with the *normalized likelihood* of a shortest path containing node  $i$ . The *likelihood* of a shortest path  $\varphi^* \in \mathcal{P}_{st}^*$  is the product of the reference transition probabilities along that path, i.e. the same as its reference path probability  $\tilde{P}_{st}^{\text{ref}}(\varphi^*)$ . Note, that the likelihood is different from the branching probability which is based on transition probabilities in the directed shortest path graph from  $s$  to  $t$ , instead of the whole graph. The *normalized likelihood* of  $\varphi^*$  is then obtained by normalizing by the sum of likelihoods of all shortest paths,  $\tilde{P}_{st}^{\text{ref}}(\varphi^*) / \sum_{\varphi \in \mathcal{P}_{st}^*} \tilde{P}_{st}^{\text{ref}}(\varphi)$ , which is also the contribution of  $\varphi^*$  to the shortest path likelihood betweenness of all the nodes along it. The shortest path likelihood betweenness is introduced here for the sake of completeness because it is the limiting function of both RSP betweenness measures presented in this paper. However, the differences between the three variants of shortest path betweenness presented above are very small, and in practice they provide very similar rankings of the nodes of a network.

*Betweenness based on random walks.* Betweenness has also been considered with respect to other than just shortest paths, namely by considering random walks or flows on a network. The first such measure was proposed by Freeman<sup>29</sup>, who defined the *flow betweenness centrality* as the amount of flow through a node over maximum flows between all node pairs. The idea of considering flows was developed further by Newman<sup>4</sup>, who defined the *current flow betweenness centrality*, which measures the centrality of a node as the total sum of electrical current that flows through it, when considering all node pairs as source-sink pairs of a unit current flow. The current flow betweenness was also coined the *random walk betweenness centrality* because of the well-established connection between electric current flows and random walks<sup>30</sup>. Indeed, it can also be interpreted as the sum of expected net flows of a random walk over the edges connected to a node, meaning that the number of times that the walk enters or leaves the node along an edge cancel each other out. Brandes and Fleischer<sup>5</sup> developed an algorithm

which improves the efficiency of computing the current flow betweenness for all nodes of a network. The properties and computation of the current flow betweenness have also been studied by Bozzo and Franceschet<sup>31</sup>.

Instead of considering net flows, a more straightforward definition of betweenness based on random walks is simply the overall expected number of visits to a node during a random walk. For arbitrary, non-absorbing walks, this quantity is not well-defined. However, it is possible to compute the proportion of steps of such a walk that the walker spends in a node. This is equal to the probability of finding the walker at the node after a long walk, i.e. the stationary probability of the node. The stationary probabilities define the *stationary distribution*, which is the unique vector  $\pi$  that satisfies the equation  $\pi = (\mathbf{P}^{\text{ref}})^T \pi$ , given that the network is strongly connected and aperiodic. It is well known that for such graphs the stationary probabilities are proportional to the recurrence times and, if the graph is undirected, to the degree centralities (or strength, for weighted graphs)<sup>32</sup>. For other graphs, the stationary distribution is not necessarily unique or may not even exist. To overcome this issue, for instance, the PageRank algorithm<sup>33</sup> uses a teleportation probability which transforms any kind of a graph into a strongly connected approximation and computes the stationary distribution on the transformed graph.

Even though the expected number of visits to a node is not well-defined for arbitrary walks, it is well-defined for absorbing walks. A bit surprisingly, we have discovered that when computed over all absorbing random walks on a strongly connected aperiodic directed graph, this measure is equal to the stationary distribution, up to a multiplying constant. The multiplying constant is the sum, over all  $s$ - $t$ -pairs of the graph, of average hitting times, i.e. the expected number of steps along an absorbing path from  $s$  to  $t$ , denoted by  $\langle \bar{e} \rangle_{st}$ . This result is not completely obvious and we have not found it mentioned explicitly in the literature. The proof of the result is omitted from the paper, as it is not relevant for this work, although the result itself is relevant for considering the RSP-based betweenness measures.

The RSP betweenness centrality measures proposed in this paper are also based on random walks, more particularly on the RSP framework. The first betweenness measure is based on counting the expected number of passages through a node of a random walker moving according to the RSP probabilities, while the other computes the expected net flow of walkers going through the node.

In fact, the RSP betweenness measures coincide with betweenness measures proposed recently, independently of our work, by Bavaud and Guex<sup>20</sup>. They propose a framework which interpolates between shortest paths and random walks by the minimization of free energy in a similar fashion as the free energy derivation of the RSP framework of Kivimäki *et al.*<sup>8</sup>. One main difference in the work of Bavaud and Guex<sup>20</sup>, compared to the derivation of RSP's, is that a more general form of energy functionals, besides the expected length (or cost) of a path, is considered. In addition to that the relative entropy is considered with respect to transition probabilities instead of path probabilities, as is done in the definition of RSP's. The path distribution derived by Bavaud and Guex<sup>20</sup> can, however, be shown to equal the path distribution defined in the RSP framework, although this requires some lengthy and uninteresting derivations and is left out of this paper. The relation between the two approaches has also been studied by in the more recent work of Guex and Bavaud<sup>34</sup>. Compared to the work of Bavaud and Guex<sup>20</sup>, our work focuses more on the computational and practical aspects of the methodology. Thanks to the recent developments in the RSP framework<sup>8</sup>, we are also able to present efficient algorithms for computing the quantities in question and illustrate the use of the methodology with practical examples.

Also closely related to the idea behind the RSP betweenness centralities is the *bag-of-paths (BoP) betweenness centrality*<sup>21</sup>. It is based on the BoP framework, which also defines a Boltzmann distribution on the paths of a graph in a similar way as the RSP framework. The BoP betweenness is then defined as the a posteriori probability that a path selected according to the Boltzmann distribution visits an intermediate node  $i$ , when it walks from node  $s$  to node  $t$  according to the BoP probabilities. The BoP betweenness is also defined for groups and used for semi-supervised classification of graph nodes. A similar group betweenness measure can be defined from the RSP betweenness measures proposed here, but this is left out of the scope of this paper. The node classification task was also tackled by Devooght *et al.*<sup>35</sup> by using a modularity measure derived from the BoP framework.

**Other centralities.** Besides betweenness, centrality has also been characterized with other additional terms such as *closeness*, *feedback* and *vitality*, to name a few<sup>36</sup>. There have also been many efforts for stating axioms that would define centrality starting from the work of Sabidussi<sup>22</sup> and recently by Boldi and Vigna<sup>37</sup>. The concept of centrality can also be considered with respect to edges, instead of nodes. Although in this work we focus on centrality according to the betweenness interpretation, we nonetheless mention some of the other interpretations in this section, as the RSP framework could be combined with many of them too.

Closeness centrality measures are based on some interpretation of overall proximity of a node to the other nodes of a network. The *shortest path closeness centrality* of node  $i$  is classically defined as  $C_i = 1/\sum_{j=1}^n \Delta_{ij}$ , where  $\Delta_{ij}$  is the shortest path distance between nodes  $i$  and  $j$ <sup>22,38</sup>. Considering communication on a network, betweenness centrality can be interpreted as the amount of *control* of a node, whereas closeness centrality measures the *efficiency* of the communication of the node<sup>2,39</sup>. Instead of using the shortest path distance to define closeness centrality, other distances can be used too. For instance, White and Smyth<sup>40</sup> define *Markov centrality* by replacing the shortest path distance in the closeness centrality definition with the average hitting time of an unbiased random walker, i.e., essentially the (unsymmetrized) commute time distance. Related to that, Brandes and Fleischer<sup>5</sup> considered a current flow analogy of closeness centrality by replacing the distance between two nodes with their potential difference. They managed to show that the current flow closeness centrality is equivalent to the information centrality defined by Stephenson and Zelen<sup>41</sup>. The equivalence has been confirmed with another proof by Bozzo and Franceschet<sup>31</sup>. One subject for future research will be to extend closeness centrality by replacing the shortest path distance with RSP-based distance measures.

Feedback centrality measures are in many ways related to random walk based centrality measures. The *eigenvector centrality*<sup>42</sup> is the archetype feedback measure and is based on the idea that the centrality of a node should

be the sum of the centralities of its neighbors. The solution of the formulation is the eigenvector of the adjacency matrix corresponding to the largest eigenvalue. The previously mentioned PageRank<sup>33</sup> can also be considered in this sense and formulated as an eigenvector. The Katz centrality<sup>43</sup> is based on the same idea of feedback, but considers also longer dependencies than only the neighbors of a node. The effect of longer dependencies decays according to the distance between nodes. The *subgraph centrality*<sup>18</sup> and *communicability betweenness*<sup>19</sup> are based on similar ideas using the matrix exponential of the adjacency matrix. The relationship between all the feedback centrality measures listed above has been studied by Benzi and Klymko<sup>44</sup>.

### Randomized shortest paths betweenness centralities

In this section we introduce two betweenness measures based on the randomized shortest paths (RSP) framework<sup>6–8</sup>. These measures both generalize the shortest path likelihood betweenness measure defined in the section “Betweenness based on shortest paths”. In addition to that, one of the two measures also generalizes the stationary distribution of the graph while the other generalizes the current flow betweenness centrality. The RSP framework has previously been used for defining distance measures between graph nodes, which has proven useful for many data analysis tasks such as clustering and classification of graph nodes<sup>6,8</sup>.

**Randomized shortest paths.** In its core, the RSP framework<sup>6–8,45</sup> is based on a probability distribution over paths between two nodes of a graph. The framework can also be formulated considering all paths, containing non-absorbing paths, but in this work, for simplicity, we restrict the framework to absorbing paths. The RSP probability distribution can be defined through either minimization of expected cost subject to a relative entropy constraint, or minimization of free energy<sup>8,20</sup>. Here we recall the former definition of minimization of expected cost, which is perhaps a more intuitive one.

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 walks is minimal, when constrained with a fixed relative entropy with respect to the reference path probability distribution. Formally, we seek for the solution to the following problem

$$\text{Minimize } \sum_{\varphi \in \mathcal{P}_{st}} \tilde{P}_{st}(\varphi) \tilde{c}(\varphi) \quad \text{subject to} \quad \begin{cases} J(\tilde{P}_{st} || \tilde{P}_{st}^{\text{ref}}) = J_0 \\ \sum_{\varphi \in \mathcal{P}_{st}} \tilde{P}_{st}(\varphi) = 1 \end{cases} \quad (2)$$

where  $\tilde{P}_{st}^{\text{ref}}(\varphi)$  is the reference path probability,  $\tilde{c}(\varphi)$  the overall cost of path  $\varphi$  and  $J(\tilde{P}_{st} || \tilde{P}_{st}^{\text{ref}})$  is the relative entropy, or Kullback-Leibler divergence, which is set to a desired level  $J_0$ .

The solution of this minimization is a Boltzmann distribution (for details, see<sup>6,7</sup>):

$$\tilde{P}_{st}(\varphi) = \frac{\tilde{P}_{st}^{\text{ref}}(\varphi) \exp(-\beta \tilde{c}(\varphi))}{\mathcal{Z}_{st}}, \quad (3)$$

where

$$\mathcal{Z}_{st} = \sum_{\varphi \in \mathcal{P}_{st}} \tilde{P}_{st}^{\text{ref}}(\varphi) \exp(-\beta \tilde{c}(\varphi)) \quad (4)$$

is the *partition function* of absorbing walks from  $s$  to  $t$  and the inverse temperature parameter  $\beta$  controls the divergence from the unbiased random walk probabilities. When applying the framework, the user is supposed to input the value for  $\beta$ , instead of the relative entropy  $J_0$ . Low and high values of  $\beta$  correspond, respectively, to low and high values of  $J_0$  and, inversely, to high and low temperature. In other words, for high values of  $\beta$  (low temperature), the path distribution  $\tilde{P}_{st}$  focuses on shortest paths, whereas for low values of  $\beta$  (high temperature) more random paths are also preferred. It is also possible to find the model corresponding to a particular value of  $J_0$ , for instance, with a binary search over different values of  $\beta$ .

The partition function  $\mathcal{Z}_{st}$  plays an important role in the derivation of the computation of many quantities related to the RSP framework, as can be seen in the following. Concerning the computation of  $\mathcal{Z}_{st}$ , we refer to earlier work<sup>8,46</sup> which shows that it can be expressed as

$$\mathcal{Z}_{st} = \frac{z_{st}}{z_{tt}}, \quad (5)$$

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

$$\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}, \quad \text{with} \quad \mathbf{W} = \mathbf{P}^{\text{ref}} \circ \exp(-\beta \mathbf{C}), \quad (6)$$

where  $\circ$  and  $\exp$  denote the element-wise matrix multiplication and exponential, respectively. The matrix  $\mathbf{W}$ , defined from the reference transition probability matrix  $\mathbf{P}^{\text{ref}}$  and the cost matrix  $\mathbf{C}$ , is a substochastic matrix and can be interpreted as defining a *killed random walk*. Consequently, the partition function  $\mathcal{Z}_{st}$  can then be interpreted as the probability of a walker surviving the walk from  $s$  to  $t$  (see<sup>8,46</sup> for details).

**The simple RSP betweenness.** We first define the simple RSP betweenness centrality of a node  $i$  with respect to absorbing paths from  $s$  to  $t$ , as the expected number of visits through  $i$  over all  $s$ - $t$ -walks, denoted by



$\bar{n}_i(s, t)$ , with respect to the RSP probabilities of Equation (3). This can further be expressed based on the expected number of passages through edges leaving from node  $i$  over all  $s$ - $t$ -walks, denoted by  $\eta_{ij}(s, t)$ , as:

$$\text{bet}_i^{\text{RSP}}(s, t) = \bar{n}_i(s, t) = \sum_{j: (i,j) \in E} \bar{\eta}_{ij}(s, t) \quad (7)$$

The function  $\text{bet}_i^{\text{RSP}}$  can be useful for visualizing path distributions and for path planning tasks between two fixed nodes of the graph<sup>45,47</sup>. Moreover, it can be used to investigate how central an intermediate actor is with respect to, for instance, the communication between two other actors in a social network.

We then define the overall *simple RSP betweenness centrality* of node  $i$  as the sum of contributions over all source-target pairs, on the graph:

$$\text{bet}_i^{\text{RSP}} = \sum_{s=1}^n \sum_{t=1}^n \text{bet}_i^{\text{RSP}}(s, t). \quad (8)$$

Next, we derive the method for computing this quantity in closed form. Let us denote by  $\eta(i \rightarrow j \in \varphi)$  the number of times that the edge  $(i, j)$  is traversed along the path  $\varphi$ . Then, by writing out the expression of  $\eta_{ij}(s, t)$ , the use of the partition function in the computation of the RSP quantities becomes evident:

$$\begin{aligned} \eta_{ij}(s, t) &= \sum_{\varphi \in \varphi} \tilde{P}_{st}(\varphi) \eta(i \rightarrow j \in \varphi) = \frac{\sum_{\varphi \in \varphi} \tilde{P}^{\text{ref}}(\varphi) \exp[-\beta \tilde{c}(\varphi)] \eta(i \rightarrow j \in \varphi)}{\sum_{\varphi' \in \varphi} \tilde{P}^{\text{ref}}(\varphi') \exp[-\beta \tilde{c}(\varphi')]} \\ &= \frac{\sum_{\varphi \in \varphi} \tilde{P}^{\text{ref}}(\varphi) \exp[-\beta \tilde{c}(\varphi)] \frac{\partial \tilde{c}(\varphi)}{\partial c_{ij}}}{\sum_{\varphi' \in \varphi} \tilde{P}^{\text{ref}}(\varphi') \exp[-\beta \tilde{c}(\varphi')]} = -\frac{1}{\beta} \frac{\partial \log \mathcal{Z}_{st}}{\partial c_{ij}} \end{aligned} \quad (9)$$

Thus, the expected number of transitions  $i \rightarrow j$  over  $s$ - $t$ -walks can be computed by differentiating the logarithm of the partition function, which is common knowledge in statistical physics (see, e.g.,<sup>48</sup>). Note that (9) holds only if there exists a path from  $s$  to  $t$ . Otherwise, naturally,  $\eta_{ij}(s, t) = 0$ .

By combining Equations (5) and (9), the computation of  $\bar{\eta}_{ij}(s, t)$  can be written as:

$$\eta_{ij}(s, t) = -\frac{1}{\beta} \frac{\partial \log \mathcal{Z}_{st}}{\partial c_{ij}} = -\frac{1}{\beta} \frac{\partial \log(z_{st}/z_{tt})}{\partial c_{ij}} = -\frac{1}{\beta} \left( \frac{1}{z_{st}} \frac{\partial z_{st}}{\partial c_{ij}} - \frac{1}{z_{tt}} \frac{\partial z_{tt}}{\partial c_{ij}} \right). \quad (10)$$

Therefore, we need to compute  $\partial z_{st}/\partial c_{ij}$ , which can be achieved using matrix formalism. If we denote by  $\mathbf{e}_i$  the  $(n \times 1)$ -vector whose element  $i$  is 1 and others are 0, then

$$\begin{aligned} \frac{\partial z_{st}}{\partial c_{ij}} &= \frac{\partial (\mathbf{e}_s^T \mathbf{Z} \mathbf{e}_t)}{\partial c_{ij}} = \frac{\partial (\mathbf{e}_s^T (\mathbf{I} - \mathbf{W})^{-1} \mathbf{e}_t)}{\partial c_{ij}} \\ &= \mathbf{e}_s^T \frac{\partial (\mathbf{I} - \mathbf{W})^{-1}}{\partial c_{ij}} \mathbf{e}_t = \mathbf{e}_s^T \left( -\mathbf{Z} \frac{\partial (\mathbf{I} - \mathbf{W})}{\partial c_{ij}} \mathbf{Z} \right) \mathbf{e}_t \\ &= \mathbf{e}_s^T \left( \mathbf{Z} \frac{\partial \mathbf{W}}{\partial c_{ij}} \mathbf{Z} \right) \mathbf{e}_t = -\beta p_{ij}^{\text{ref}} \exp[-\beta c_{ij}] \mathbf{e}_s^T \mathbf{Z} \mathbf{e}_i \mathbf{e}_j^T \mathbf{Z} \mathbf{e}_t = -\beta w_{ij} z_{si} z_{jt} \end{aligned} \quad (11)$$

where we used  $\frac{\partial \mathbf{X}^{-1}}{\partial x} = -\mathbf{X}^{-1} \frac{\partial \mathbf{X}}{\partial x} \mathbf{X}^{-1}$  (see, e.g.,<sup>49,50</sup>). Equation (10) can therefore be rewritten as

$$\eta_{ij}(s, t) = \frac{1}{z_{st}} z_{si} w_{ij} z_{jt} - \frac{1}{z_{tt}} z_{ti} w_{ij} z_{jt} = \left( \frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}} \right) w_{ij} z_{jt} \quad (12)$$

Furthermore, the total flow transiting through node  $i$ , given that  $i \neq t$ , is

$$\bar{n}_i(s, t) = \sum_{j=1}^n \eta_{ij}(s, t) = \left( \frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}} \right) \sum_{j=1}^n w_{ij} z_{jt} \quad (13)$$

The expression on the right-hand side of Equation (13) can be further simplified in the following way. We know that  $(\mathbf{I} - \mathbf{W})^{-1}(\mathbf{I} - \mathbf{W}) = \mathbf{I}$ , which implies that  $\mathbf{Z}(\mathbf{I} - \mathbf{W}) = \mathbf{I}$  and therefore that  $\mathbf{Z} = \mathbf{Z}\mathbf{W} + \mathbf{I}$ , or element-wise,  $z_{it} = \sum_{j=1}^n w_{ij} z_{jt} + \delta_{it}$ , where  $\delta_{it}$  is the Kronecker delta. However, the term  $\delta_{it}$  can be discarded, when considering Equation (13), because when  $i = t$ , we anyway have  $(z_{si}/z_{st} - z_{ti}/z_{tt}) = 0$ . Thus, Equation (13) simplifies to

$$\bar{n}_i(s, t) = \left( \frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}} \right) z_{it} \quad (14)$$

Finally, the simple RSP betweenness centrality (Eq. (8)) can be computed with matrix manipulation as

$$\begin{aligned}
 \text{bet}_i^{\text{RSP}} &= \sum_{s,t=1}^n \bar{\eta}_i(s, t) = \sum_{s,t=1}^n \left( \frac{z_{si}}{z_{st}} - \frac{z_{ti}}{z_{tt}} \right) z_{it} \\
 &= \sum_{s,t=1}^n z_{it} \frac{1}{z_{ts}} z_{si} - \sum_{s,t=1}^n z_{it} \frac{1}{z_{tt}} z_{ti} \\
 &= \sum_{s,t=1}^n z_{it} \frac{1}{z_{ts}} z_{si} - n \sum_{t=1}^n z_{it} \frac{1}{z_{tt}} z_{ti} \\
 &= [\text{diag}(\mathbf{Z}(\mathbf{Z}^\dagger)^T \mathbf{Z}) - n \text{diag}(\mathbf{Z} \text{Diag}(\mathbf{Z}^\dagger) \mathbf{Z})]_i \\
 &= [\text{diag}(\mathbf{Z}(\mathbf{Z}^\dagger - n \text{Diag}(\mathbf{Z}^\dagger))^T \mathbf{Z})]_i
 \end{aligned} \tag{15}$$

where  $\text{diag}(\mathbf{X})$  and  $\text{Diag}(\mathbf{X})$  are, respectively, a column vector and a diagonal matrix containing the diagonal of  $\mathbf{X}$ ,  $\mathbf{X}^\dagger$  denotes the element-wise reciprocal matrix, i.e.,  $x_{ij}^\dagger = 1/x_{ij}$  and the superscript  $t$  denotes elements from the transposed matrix, i.e.,  $z_{ij}^t = z_{ji}$ . The vector  $\text{bet}^{\text{RSP}}$  of all betweenness values is computed accordingly.

The pseudocode for computing the simple RSP betweenness for all nodes is presented in Algorithm 1. In conclusion, the simple RSP betweenness scores of all nodes can be computed by performing the matrix inversion  $\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}$  and then simple matrix operations according to Equation (15). Thus, the computational bottleneck of the algorithm is the matrix inversion, which, in general, has time complexity  $\mathcal{O}(n^3)$  and space complexity  $\mathcal{O}(n^2)$ , because of which the method is currently not practical with very large networks.

**Algorithm 1**-Computing the simple RSP betweenness vector of a graph  $G$ .

**Input:**

- A directed strongly connected graph  $G$  with  $n$  nodes.
- The  $n \times n$  reference transition probability matrix  $\mathbf{P}^{\text{ref}}$  (defined from the adjacency matrix as  $\mathbf{P}^{\text{ref}} = \mathbf{D}^{-1} \mathbf{A}$ )
- The  $n \times n$  non-negative cost matrix  $\mathbf{C}$
- The inverse temperature parameter  $\beta$ .

**Output:**

- The  $n \times 1$  simple RSP betweenness vector  $\text{bet}^{\text{RSP}}$ .

1.  $\mathbf{W} \leftarrow \mathbf{P}^{\text{ref}} \circ \exp[-\beta \mathbf{C}]$
2.  $\mathbf{Z} \leftarrow (\mathbf{I} - \mathbf{W})^{-1}$
3.  $\mathbf{Z}^\dagger \leftarrow \mathbf{e} \mathbf{e}^T \div \mathbf{Z}$
4.  $\text{bet}^{\text{RSP}} \leftarrow \text{diag}(\mathbf{Z}(\mathbf{Z}^\dagger - n \text{Diag}(\mathbf{Z}^\dagger))^T \mathbf{Z})$
5. **return**  $\text{bet}^{\text{RSP}}$

**RSP net betweenness.** Instead of only considering the overall outgoing flow of random walkers, as in the definition of Equation (8), it may in some cases make more sense to compute the net outgoing flow<sup>6</sup>, 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<sup>4,5,31</sup>. According to this approach, we define the *RSP net betweenness centrality* of node  $i$  as

$$\text{bet}_i^{\text{RSPnet}} = \sum_{s=1}^n \sum_{t=1}^n \sum_{j:(i,j) \in E} \left| \eta_{ij}(s, t) - \eta_{ji}(s, t) \right| \tag{16}$$

The computation of this quantity for all nodes at once is a bit more involved than in the case of the simple RSP betweenness, because of the absolute value in the expression. A naïve algorithm would loop over all  $s$ - $t$ -pairs and compute the contribution of the corresponding paths to each intermediate node. However, a more efficient solution is to perform a loop over each edge of the network, compute separately the net flow through that edge over all  $s$ - $t$ -paths, and to add this net flow to the betweenness score of the starting node of the edge. This change in looping strategy improves the complexity by a factor from  $\mathcal{O}(n^2)$  to  $\mathcal{O}(m)$ , which makes a big difference, when dealing with a sparse network. The faster computation can be achieved by writing out the matrix  $\mathbf{N}^{\text{ij}}$ , whose element  $(s, t)$  is  $\eta_{ij}(s, t)$  from Equation (12):

$$[\mathbf{N}^{\text{ij}}]_{st} = \eta_{ij}(s, t) = \left( \frac{z_{si} z_{jt}}{z_{st}} - \frac{z_{ti} z_{jt}}{z_{tt}} \right) w_{ij} = \left[ \left( \mathbf{z}_i^c (\mathbf{z}_j^r)^T \div \mathbf{Z} - \mathbf{e} (\mathbf{z}_i^c \circ \mathbf{z}_j^r \div \text{diag}(\mathbf{Z}))^T \right) \right]_{st} w_{ij}, \tag{17}$$

where  $\circ$  and  $\div$  denote elementwise matrix multiplication and division, respectively,  $\mathbf{z}_i^c$  and  $\mathbf{z}_j^r$  denote the  $(n \times 1)$ -vectors corresponding, respectively, to the  $i$ -th column and  $j$ -th row of matrix  $\mathbf{Z}$  and  $\mathbf{e}$  is the  $(n \times 1)$ -vector whose all elements are 1.

The overall net flow through edge  $(i, j)$  can then be computed as

$$\bar{\eta}_{ij}^{\text{net}} = \sum_{s,t=1}^n [|\mathbf{N}^{st}_{ij} - \mathbf{N}^{ji}|]_{st} = \mathbf{e}^T |\mathbf{N}^{ij} - \mathbf{N}^{ji}| \mathbf{e} \quad (18)$$

after which the betweenness score of each node can be simply computed by summing up the contributions of each edge connected to the node:

$$\text{bet}_i^{\text{RSPnet}} = \sum_{j:(i,j) \in E} \bar{\eta}_{ij}^{\text{net}}. \quad (19)$$

Algorithm 2 contains the pseudocode for computing the RSP net betweenness. In principle, the algorithm can be used with directed graphs and the result can be interpreted according to the net flow of random walks, even though the electric current interpretation only makes sense with undirected graphs. With undirected graphs, Algorithm 2 should be altered so that it only considers each undirected edge only once and increments also the betweenness of node  $j$  in addition to node  $i$  at step 10. This reduces the computation time on undirected networks by a half. Algorithm 2 also contains the same matrix inversion as Algorithm 1 of time complexity  $\mathcal{O}(n^3)$ . In addition to this, the other consuming task is the loop over all  $m$  edges of the graph in steps 4.–11., inside which elementary matrix operations of time complexity  $\mathcal{O}(n^2)$  have to be performed. Thus, the total time complexity of the algorithm is  $\mathcal{O}(n^3 + mn^2)$ .

Although the RSP net betweenness can be computed for directed graphs, we have not found a good use case for this purpose. The definition on directed graphs is not as intuitive as the simple betweenness and, moreover, current flows and the current flow betweenness are normally defined only for undirected graphs. Nevertheless, it is possible to use Algorithm 2 with directed graphs, in addition to which it is possible to derive a similar algorithm for computing the directed version of the current flow betweenness based on the pseudoinverse of the Laplacian matrix of the graph.

**Algorithm 2**–The RSP net betweenness vector of a graph  $G$ .

**Input:**

Same as Algorithm 1.

**Output:**

– The  $n \times 1$  RSP net betweenness vector  $\text{bet}^{\text{RSPnet}}$ .

1.  $\mathbf{W} \leftarrow \mathbf{P}^{\text{ref}} \circ \exp[-\beta \mathbf{C}]$
2.  $\mathbf{Z} \leftarrow (\mathbf{I} - \mathbf{W})^{-1}$
3.  $\text{bet}^{\text{RSPnet}} \leftarrow \mathbf{0}$
4. **for**  $i = 1$  to  $n$  **do**
5.      $\mathbf{Z}_i^c \leftarrow \mathbf{Z} \mathbf{e}_i, \mathbf{Z}_j^r \leftarrow \mathbf{Z}^T \mathbf{e}_j$
6.     **for all**  $j$  such that  $(i, j) \in E$  **do**
7.          $\mathbf{Z}_j^c \leftarrow \mathbf{Z} \mathbf{e}_j, \mathbf{Z}_j^r \leftarrow \mathbf{Z}^T \mathbf{e}_j$
8.          $\mathbf{N}^{ij} \leftarrow w_{ij} [(\mathbf{Z}_i^c (\mathbf{Z}_j^r)^T \div \mathbf{Z}) - \mathbf{e} ((\mathbf{Z}_i^c \circ \mathbf{Z}_j^r) \div \text{diag}(\mathbf{Z}))^T]$
9.          $\mathbf{N}^{ji} \leftarrow w_{ji} [(\mathbf{Z}_j^c (\mathbf{Z}_i^r)^T \div \mathbf{Z}) - \mathbf{e} ((\mathbf{Z}_j^c \circ \mathbf{Z}_i^r) \div \text{diag}(\mathbf{Z}))^T]$
10.          $\text{bet}_i^{\text{RSPnet}} \leftarrow \text{bet}_i^{\text{RSPnet}} + \mathbf{e}^T |\mathbf{N}^{ij} - \mathbf{N}^{ji}| \mathbf{e}$
11.     **end for**
12. **end for**
13. **return**  $\text{bet}^{\text{RSPnet}}$

**RSP betweenness centralities at the limit  $\beta \rightarrow \infty$ .** The simple RSP betweenness centrality counts the expected number of visits to each node during RSP's between all source-target pairs of the graph. In the low temperature limit, i.e. when  $\beta \rightarrow \infty$ , the RSP probability distribution of Equation (3) focuses solely on the shortest paths of the graph and the expected number of visits to a node on a shortest path approaches the probability of following that particular path. Moreover, when  $\beta \rightarrow \infty$ , for all paths  $\varphi \in \mathcal{P}_{st}$ , whose cost  $\tilde{c}(\varphi) > \tilde{c}_{st}^*$ ,  $\exp(-\beta \tilde{c}(\varphi)) \ll \exp(-\beta \tilde{c}_{st}^*)$ , because of which the partition function  $\mathcal{Z}_{st}$  of Equation (4) becomes dominated by the terms determined by the shortest paths. As a result,

$$\tilde{\mathbf{P}}_{st}(\varphi) = \frac{\tilde{\mathbf{P}}_{st}^{\text{ref}}(\varphi) \exp(-\beta \tilde{c}(\varphi))}{\mathcal{Z}_{st}} \xrightarrow{\beta \rightarrow \infty} \begin{cases} 0, & \text{if } \varphi \notin \mathcal{P}_{st}^* \\ \frac{\tilde{\mathbf{P}}_{st}^{\text{ref}}(\varphi)}{\sum_{\varphi \in \mathcal{P}_{st}^*} \tilde{\mathbf{P}}_{st}^{\text{ref}}(\varphi)}, & \text{if } \varphi \in \mathcal{P}_{st}^* \end{cases} \quad (20)$$



In other words, the RSP probability of a shortest path approaches the normalized likelihood of the path, which is also the contribution to the betweenness scores of the nodes along the path. Thus, the simple RSP betweenness converges to the shortest path likelihood betweenness.

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 in the net flow in Equation (16) becomes zero, as the walker will only move in one direction along each edge for a given  $s$ - $t$ -pair. Thus, the RSP net betweenness also approaches the shortest path likelihood betweenness, as  $\beta \rightarrow \infty$ .

**RSP betweenness centralities at the limit  $\beta \rightarrow 0^+$ .** In the high temperature limit, as  $\beta \rightarrow 0^+$ ,  $\exp(-\beta \tilde{c}(\varphi)) \rightarrow 1$  for all  $\varphi$ , and the RSP probabilities of Equation (3) converge to the unbiased random walk probabilities, determined by the reference transition probabilities, i.e.

$$\tilde{P}_{st} \xrightarrow{\beta \rightarrow 0^+} \tilde{P}_{st}^{\text{ref}}. \quad (21)$$

This means that the simple RSP betweenness converges to the expected number of visits to a node over all absorbing walks with respect to the unbiased random walk probabilities. As presented in the section “Betweenness based on random walks”, this measure is proportional to the stationary distribution, if the network is strongly connected and aperiodic, where the multiplicative factor is the sum of average hitting times between all  $s$ - $t$ -pairs. Thus, for undirected networks, the measure becomes proportional to the degree (or strength) centrality. In the same limit, as  $\beta \rightarrow 0^+$ , the RSP net flow converges to the current flow betweenness centrality<sup>4,5</sup>, as the edge flows  $\eta_{ij}(s, t)$  converge to the potential differences of adjacent nodes.

## Experiments

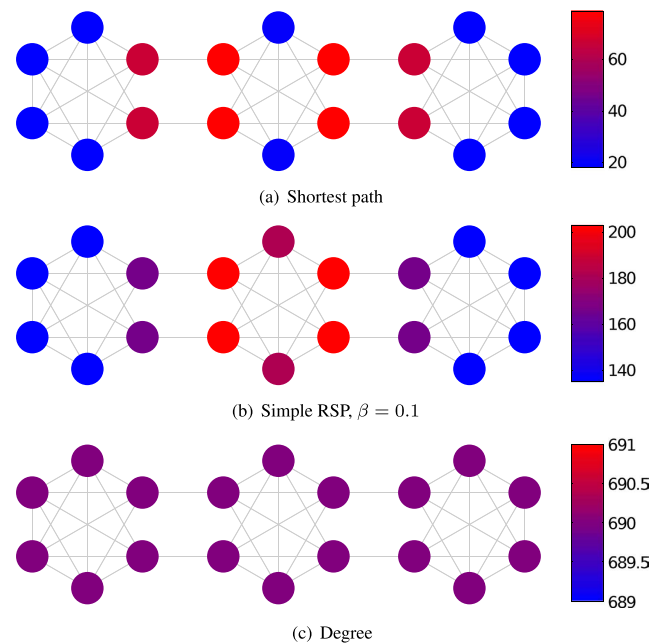
The interpolation between common centrality measures already makes the simple RSP and RSP net betweenness centralities interesting. Furthermore, there are also cases in which the RSP betweenness centralities can be more relevant than their limit functions, the shortest path likelihood betweenness, the current flow betweenness and the stationary distribution. In this section these benefits will be illustrated, first with artificially generated networks, and later with two real networks of very different nature. The artificial examples show the behavior of the RSP betweenness measures in a network consisting of communities. The first real network is the street network of Lower and Midtown Manhattan, which serves as an example of an undirected network. The second is a subset of the directed Wikipedia hyperlink network.

All of the example cases presented here indicate benefits of using the RSP betweenness measures over the shortest path and random walk based betweenness measures. The benefits are clearer in the case of the simple RSP betweenness than the RSP net betweenness. The simple RSP betweenness is also more preferable in terms of interpretability and computational efficiency compared to the net betweenness. However, the decision of which approach to use depends on the actual application and its premises. We have also experimented with several other types of graphs, including, for instance, examples presented by Brandes and Hildenbrand<sup>26</sup>, which have been designed to differentiate centrality measures. However, for these, and other simple cases, there is no clear difference in considering RSPs, but rather the ranking of nodes with intermediate values of  $\beta$  are in essence the same as with the limit values of  $\beta$ .

**Overall betweenness of an in-between community.** One possible use for the RSP betweenness measures is the detection of groups of nodes that are central in a network. Consider a network consisting of three disjoint communities, A, B and C which are highly intraconnected but loosely interconnected. In addition, community B is connected to communities A and C, which, however, do not share any edges with each other. In other words, community B is in between communities A and C and all paths between nodes of communities A and C have to go through community B. Such an organization is possible, for instance, in a hierarchical social network, where community B could represent a directoral board of a company. If, moreover, the graph is in general sufficiently sparse, then the shortest paths between communities A and C will run through only a few of the nodes of community B. Thus, betweenness measures based on shortest paths will only highlight those nodes of community B, whereas the other nodes of community B will get no contribution from the connections between communities A and C. For some applications, however, the nodes of community B should, in general, be considered more in-between than the nodes of communities A and C. Defining betweenness based on random walks, and especially RSPs can help in this matter, as will be demonstrated next.

We first consider a simple example of a regular graph with communities organized in the order described above. The example is depicted in Fig. 1 which shows a 5-regular graph containing three communities of 6 nodes, where one of the communities is connected to the other two. The graph has been constructed by considering three cliques of 6 nodes, and by removing and adding appropriate links to obtain the desired structure. Figure 1 contains the heat plots of the betweenness values of the nodes with the simple RSP betweenness (Fig. 1(b)) and its limit functions, i.e. the shortest path likelihood betweenness (Fig. 1(a)), which in this example equals the standard shortest path betweenness, and the stationary distribution multiplied by the sum of average hitting times (see section “RSP betweenness centralities at the limit  $\beta \rightarrow 0^+$ ”), which – as the network is undirected – corresponds to the degree centrality (Fig. 1(c)) up to a scaling factor.

The heat plots show that the simple RSP betweenness highlights the nodes of the central community more than its limit functions. The low temperature limit function, i.e. the shortest path likelihood betweenness highlights the nodes connecting the different communities, but the betweenness scores of the two other nodes in the central community are of the same magnitude as the scores of the other nodes in the peripheral communities. In the high temperature limit the simple RSP betweenness converges to the degree centrality, which is constant for



**Figure 1.** A 5-regular graph with three communities, and the heat plots of its betweenness values with the shortest path likelihood betweenness (a), the simple RSP betweenness (b), and the degree centrality (c).

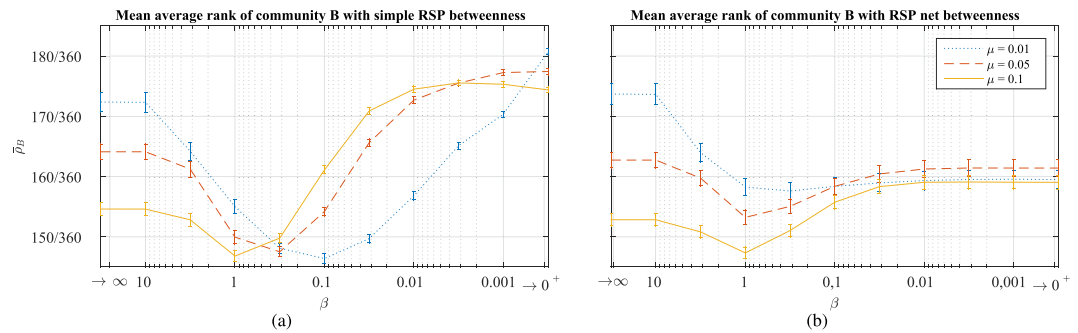
all nodes, as the graph is regular. Although the heat plots show the actual betweenness scores, and not the rankings of the nodes according to them, the rankings also comply with the above findings.

Using the RSP net betweenness (for which the results are not illustrated here), however, brings no benefit in this example, when compared to its limit functions. Namely, the current flow betweenness ranks the nodes of the central community a bit higher than the shortest path likelihood betweenness, but the RSP net betweenness does not increase those ranks with any intermediate values of  $\beta$ . On the other hand, the current flow betweenness values of the central community nodes are relatively much lower than the simple RSP betweenness values of Fig. 1(b). However, the RSP net betweenness, can be beneficial in a similar setting, but with a bit of more complexity, which will be shown next.

For a more complex example, we generate random networks with the LFR algorithm of Lancichinetti *et al.*<sup>51</sup> designed to construct scale-free networks with a community structure. This experiment confirms further the usefulness of the simple RSP, as well as the RSP net betweenness measure. We generated graphs consisting of three communities, A, B and C, and then simply removed the edges between two of the communities A and C. The size of the communities was set to 120 nodes, resulting in networks with 360 nodes, the average degree of the network was set to 10, the maximum degree to 120 and the power-law exponent of the degree distribution to  $-2$ . We tested three different values of the mixing coefficient  $\mu = \{0.01, 0.05, 0.1\}$ , which essentially determines the probability of having an edge between two communities after the degrees of nodes have been fixed. For each generated network, we computed the shortest path likelihood betweenness, the degree centrality, the current flow betweenness as well as the simple RSP and RSP net betweenness scores with several different values of the parameter  $\beta$ . We then rank the nodes according to each list of betweenness scores (with rank 1 naturally meaning the node with the highest score) and compute the average rank of the nodes in the central community B. We repeat the graph generation and the computation of average ranks 200 times and report the mean average rank of the nodes in the in-between community over these 200 runs.

The results are plotted in Fig. 2, which shows that in most cases both the simple RSP and the RSP net betweenness, with some intermediate values of  $\beta$ , rank the nodes of the in-between community more central than their limiting functions. The plots show the mean average rank, as well as the standard error of the mean, of the nodes of the central community B with different values of the inverse temperature  $\beta$ . The values at the extremes of the plots correspond to the results obtained with the limiting functions, with the left end corresponding to the low-temperature (high  $\beta$ ) case, i.e. the shortest path likelihood betweenness in both plots, and the right end to the high-temperature (low  $\beta$ ) case, i.e. the degree centrality in Fig. 2(a), and the current flow betweenness Fig. 2(b). It is evident from the figures that with some intermediate value of  $\beta$  the RSP betweenness measures in this setting often manage to rank the nodes of the central group B higher than the limit functions by taking into account other connections besides only the shortest ones. Note that this does not mean that the RSP betweenness rankings are lower than the rankings with the limit functions on each individual network, but that the result holds on average over the 200 generated networks. The fluctuations of the rankings are indicated by the error bars showing the standard error of the mean over the 200 networks.

**Manhattan street network.** One promising application area for RSP's are routing or path planning problems. RSP's allow the modeling of routing in situations that include an element of randomness, such as navigation



**Figure 2.** The mean average rank  $\bar{p}_B$  of the nodes of community B which lies in between two other communities, A and C, based on the nodes' RSP betweenness (a) and RSP net betweenness values (b) over 200 networks of 360 nodes generated using the LFR algorithm, as described in the body text (with low rank meaning a high betweenness score). The results are plotted for varying values of  $\beta$  and for three values of the mixing parameter  $\mu$  with error bars indicating the standard error of the mean over the 200 runs. In both plots, the values at the left end of the curves (as  $\beta \rightarrow \infty$ ) show the results with the shortest path likelihood betweenness and at the right end (as  $\beta \rightarrow 0^+$ ) the results with the degree centrality in (a), and the current flow betweenness in (b).

of people or animals in an environment. On the other hand, by definition, RSP's can be used for planning paths in an optimal way while keeping the predictability of the path at a desired level. RSP's could also be used for avoiding congestion problems in transportation and traffic networks.

We illustrate the use of RSP's for routing in a network by analyzing the street network of Midtown and Lower Manhattan. We have extracted this network from <http://www.openstreetmap.org><sup>52</sup> (see Materials). The nodes in the network correspond to intersections and the edges are the street segments between the intersections. We treat the network as undirected, and analyze the network using both the simple RSP and the RSP net betweenness.

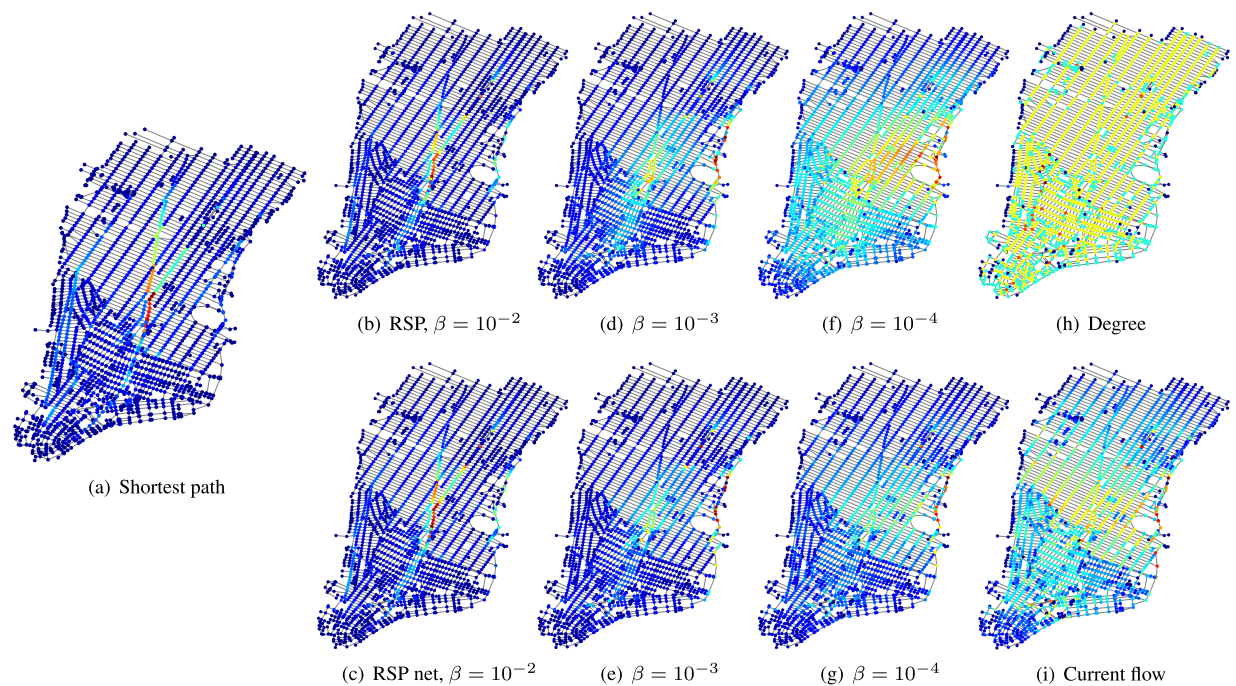
The length of each street segment is assigned as the cost of the corresponding edge. Accordingly, the overall cost of a path is its overall length. However, we define here the reference transition probabilities of the random walk according to the degree of each node,  $p_{ij}^{\text{ref}} = 1/d_i$ , i.e. only according to the number of edges connected to the node and independent of the edge costs. This seems like a reasonable choice for moving in a street network, if we consider that the decision of direction of the random walker in an intersection is not affected by the lengths of the street segments. Remember that this means that the shortest path likelihood betweenness on the graph is based on the edge costs, whereas the random walk based betweenness measures do not depend on the costs, but only the degrees of nodes.

The heat plots of the simple RSP and the RSP net betweenness measures and their limit functions on the Manhattan street network are depicted in Fig. 3. At low temperatures (large  $\beta$ ), both RSP based betweenness measures converge to the shortest path likelihood betweenness, shown in Fig. 3(a). Figure 3(b,d,f) show the betweenness values obtained with the simple RSP betweenness, and Fig. 3(c,e,g) the values obtained with the RSP net betweenness with different values of  $\beta$ . As the network is undirected and connected, and because we use reference probabilities based only on degrees, instead of costs, the limit function of the simple RSP betweenness, when  $\beta \rightarrow 0^+$ , is equal to the degree centrality multiplied by a constant, shown in Fig. 3(h). Finally, the limit function of the RSP net betweenness is the current flow betweenness, which is presented in Fig. 3(i).

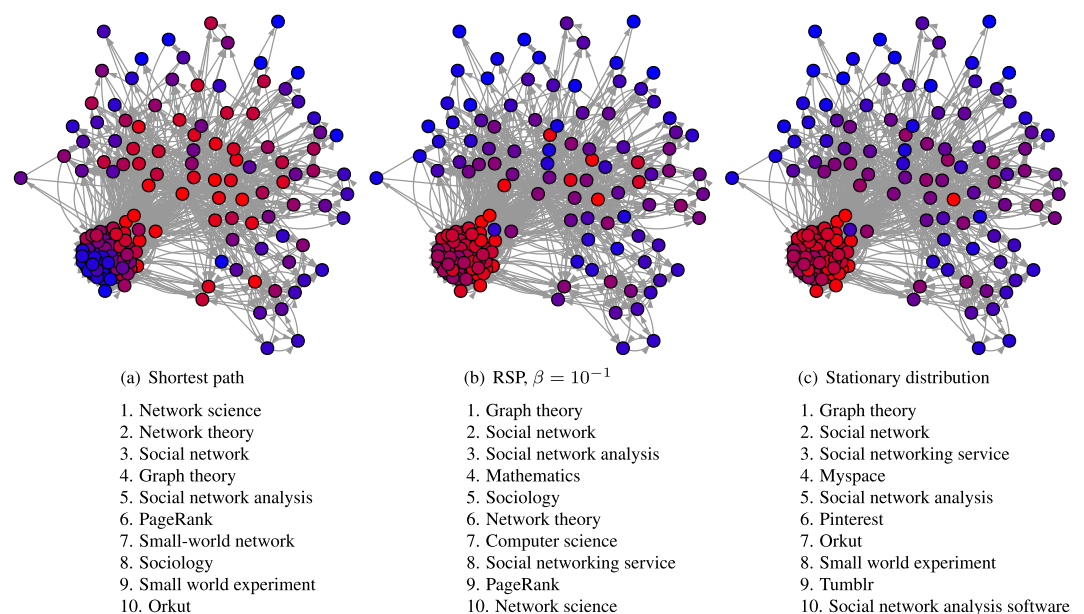
This example shows one strength of both of the RSP betweenness measures. It is evident from the plot that the shortest path likelihood betweenness is focused on Broadway, which functions as a diagonal shortcut in many routes on the grid-like Midtown. However, when  $\beta$  is decreased, both RSP based measures rank highest the intersections along the FDR Drive on the eastern shore. This is mainly caused by the sparsity of streets on the east shore close to the residential areas of Stuyvesant Town and Peter Cooper Village. As a result, the FDR Drive is a vital connection between the upper and lower eastern parts of the map. This aspect is not clear from the shortest path likelihood betweenness, but becomes apparent by computing the RSP-based betweenness values. The current flow betweenness also ranks high the intersections of FDR Drive, but as a drawback the importance of Broadway is not as apparent as it perhaps deserves.

Thus, it seems that the RSP based betweenness measures, by assuming suboptimal navigation between points in the network, can highlight bottlenecks such as the FDR Drive on the Manhattan network better than the deterministic shortest path betweenness or the unbiased current flow betweenness. Comparing between the two RSP betweenness, it is hard to find any major differences. However, in a street network, when considering the movement of people or vehicles, the simple RSP betweenness has a more sensible physical interpretation than the net betweenness. Moreover, again, the computation of the simple RSP betweenness is much more efficient and its interpretation clearer than the net betweenness, which make it a more preferable candidate. On the other hand, there may also appear applications, for which the net flow interpretation is more relevant, in which case the net betweenness should be used.

**A subnetwork of Wikipedia.** Here we illustrate the behavior of the simple RSP betweenness on a directed real network. The network is a subnetwork of the hyperlink network of Wikipedia. It consists of the Wikipedia page on Network Science and the pages that contain a hyperlink to it, or are linked to from it. We only consider the largest strongly connected component of this network which contains 151 nodes. We only report the results



**Figure 3.** The interpolation between the shortest path and random walk betweenness measures with the simple RSP and the RSP net betweenness measures on the Midtown and Lower Manhattan street network. The network has been extracted from <http://www.openstreetmap.org><sup>52</sup> (for the data and code, see Materials).



**Figure 4.** The interpolation between the shortest path likelihood betweenness the stationary distribution with the simple RSP betweenness on the subnetwork of Wikipedia. The color of a node in the plots indicates its rank w.r.t. the betweenness measure; red and blue indicate high and low rank (i.e. high and low betweenness values), respectively. The dense cluster of nodes in the lower left corner consists mostly of nodes related to social networks. Below the plots, the 10 highest-ranked nodes are listed.

obtained with the simple RSP betweenness, and not the RSP net betweenness, because the net flow interpretation is not particularly suitable when studying the World Wide Web, and in this example it does not provide any interesting results.

Figure 4 shows the change in the rankings of the nodes of the Wikipedia subnetwork with the simple RSP betweenness centrality (Fig. 4(b)), as well as with its limiting functions, the shortest path likelihood betweenness



(Fig. 4(a)), and the stationary distribution (Fig. 4(c)). In addition to the heat plots, there are lists of the top ten nodes according to each betweenness centrality. The plots directly illustrate the general structure of the network, with a division to a tightly intra-connected cluster, appearing on the lower left corner of the plots, and a more sparsely connected peripheral part. The dense cluster comprises of nodes related mostly to social networks, whereas the other nodes correspond to more general concepts.

The shortest path likelihood betweenness ranks quite high many nodes from both of the two groups of nodes, i.e. general nodes such as the seed node 'Network science', whereas the stationary distribution focuses mostly on the dense cluster of nodes related to social networks, highlighting especially particular social networking web-sites, such as 'Myspace' and 'Pinterest'. Interestingly, even the seed node 'Network science' is not on the top ten nodes according to the stationary distribution. Here, again, the simple RSP betweenness makes an interesting compromise between these two extremes by respecting the high connectivity of the social networks cluster while also highlighting important, general nodes, such as 'Mathematics' and 'Computer Science' from the peripheral group.

This example indicates the potential of the simple RSP betweenness in analyzing and exploiting semantic and associative networks, as in<sup>53</sup>. Moreover, the example can help in applications of web design and marketing, for instance, considering situations where a user tries to find a certain page on a web site by following hyperlinks of that site. This can happen, for example by browsing videos on <http://www.youtube.com>, or when playing the <http://www.thewikigame.com>, in which the purpose is to find a target Wikipedia page from a starting page only by clicking the hyperlinks on the pages.

## Conclusion

We have presented two new graph node betweenness centrality measures based on Randomized Shortest Paths. The first measure, the simple RSP betweenness centrality, counts the expected number of visits to a node, while the second, the RSP net betweenness, is based on the overall net flow over edges connected to a node. Both of these measures are parametrized generalizations of more traditional betweenness centrality measures. The RSP net betweenness and its high-temperature limit function, the current flow betweenness seem theoretically more elaborate than the simple RSP betweenness and its limit function, the stationary distribution. However, based on our experiments, the simple RSP betweenness seems to provide more satisfying and practical results than the net betweenness, in addition to which (and perhaps partly because) it is easier to interpret. In general, our experiments have shown that the RSP betweenness centralities can provide interesting insight into the role and importance of the nodes in a network in ways that the more traditional betweenness measures based on either shortest paths or unbiased random walks can not achieve.

The RSP betweenness measures could be further compared with other centrality measures, which could be a subject for future work. However, the main purpose of this paper is only to focus on betweenness centrality measures, to introduce the RSP based methods and their computation and to provide some examples where one could benefit from using them. Also, the paper only considers betweenness as a global measure on nodes, but the methods can easily be extended to other uses such as edge betweenness, betweenness w.r.t. a group of nodes or betweenness between groups of source and target nodes, all of which have relevant applications. One drawback of the RSP framework is that it often requires a full matrix inverse, because of which it is not currently practical for very large networks. One topic for future research is to develop methods that would allow estimating the RSP-based quantities for large networks either by using more specialized computational methods or by approximation.

Although the computational complexity of the RSP-based methods can be too high for very large problems, the implementation and the interpretation of the computations is quite straightforward. Moreover, the framework lies on solid theoretical grounds, and considering the generalization of shortest paths by randomization makes sense for many application scenarios. In our examples we have shown them to give promising results in highlighting nodes that belong to a central group of nodes, in detecting possible bottlenecks in street networks for navigation modeling and also in evaluating the visit rate of pages on the World Wide Web.

**Materials.** The Matlab code for the algorithms presented in the paper and the real world networks used in the experiments, along with a parser of the OpenStreetMap data, are available online at <https://github.com/ikivimak/RSP-betweenness>.

## References

- Freeman, L. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41 (1977).
- Freeman, L. Centrality in social networks conceptual clarification. *Soc. Networks* **1**, 215–239 (1978–1979).
- Brandes, U. A faster algorithm for betweenness centrality. *J. Math. Sociol.* **25**, 163–177 (2001).
- Newman, M. A measure of betweenness centrality based on random walks. *Soc. Networks* **27**, 39–54 (2005).
- Brandes, U. & Fleischer, D. Centrality measures based on current flow. In *Proceedings of the 22nd Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, 533–544 (2005).
- Yen, L., Mantrach, A., Shimbo, M. & Saerens, M. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, 785–793 (2008).
- Saerens, M., Achbany, Y., Fouss, F. & Yen, L. Randomized shortest-path problems: Two related models. *Neural Comput.* **21**, 2363–2404 (2009).
- Kivimäki, I., Shimbo, M. & Saerens, M. Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A* **393**, 600–616 (2014).
- Goh, K.-I., Kahng, B. & Kim, D. Universal behavior of load distribution in scale-free networks. *Phys. Rev. Lett.* **87**, 278701 (2001).
- Barthelemy, M. Betweenness centrality in large complex networks. *Eur. Phys. J. B* **38**, 163–168 (2004).
- Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019–1031 (2007).



12. Nadler, B., Srebro, N. & Zhou, X. Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. & Culotta, A. (eds.) *Advances in Neural Information Processing Systems* 22, 1330–1338 (Curran Associates, Inc., 2009).
13. Luxburg, U. V., Radl, A. & Hein, M. Getting lost in space: Large sample analysis of the resistance distance. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. & Culotta, A. (eds.) *Advances in Neural Information Processing Systems* 23, 2622–2630 (Curran Associates, Inc., 2010).
14. Alamgir, M. & Luxburg, U. V. Phase transition in the family of p-resistances. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F. & Weinberger, K. (eds.) *Advances in Neural Information Processing Systems* 24, 379–387 (Curran Associates, Inc., 2011).
15. Chebotarev, P. A class of graph-geodesic distances generalizing the shortest-path and the resistance distances. *Discrete Appl. Math.* **159**, 295–302 (2011).
16. Chebotarev, P. & Shamis, E. The matrix-forest theorem and measuring relations in small social groups. *Autom. Remote Control* **58**, 1505–1514 (1997).
17. Li, Y., Zhang, Z.-L. & Boley, D. The routing continuum from shortest-path to all-path: A unifying theory. In *Proceedings of the 2011 31st International Conference on Distributed Computing Systems, ICDCS '11*, 847–856 (IEEE Computer Society, Washington, DC, USA, 2011).
18. Estrada, E. & Rodriguez-Velazquez, J. A. Subgraph centrality in complex networks. *Phys. Rev. E* **71**, 056103 (2005).
19. Estrada, E., Higham, D. J. & Hatano, N. Communicability betweenness in complex networks. *Physica A* **388**, 764–774 (2009).
20. Bavaud, F. & Guex, G. Interpolating between random walks and shortest paths: a path functional approach. In K. A. et al. (ed.) *SocInfo 2012*, vol. 7710 of *LNCS*, 68–81 (Springer, 2012).
21. Lebichot, B., Kivimäki, I., François, K. & Saerens, M. Semisupervised classification through the bag-of-paths group betweenness. *IEEE Trans. Neural Netw. Learn. Syst.* **25**, 1173–1186 (2014).
22. Sabidussi, G. The centrality index of a graph. *Psychometrika* **31**, 581–603 (1966).
23. Brandes, U. & Erlebach, T. (eds.) *Network analysis: methodological foundations* (Springer-Verlag, 2005).
24. Kolaczyk, E. D. *Statistical Analysis of Network Data* (Springer, 2009).
25. Borgatti, S. P. & Everett, M. G. A graph-theoretic perspective on centrality. *Soc. Networks* **28**, 466–484 (2006).
26. Brandes, U. & Hildenbrand, J. Smallest graphs with distinct singleton centers. *Netw. Sci.* **2**, 416–418 (2014).
27. Brandes, U. On variants of shortest-path betweenness centrality and their generic computation. *Soc. Networks* **30**, 136–145 (2008).
28. Newman, M. E. J. Erratum: Scientific collaboration networks ii. shortest paths, weighted networks, and centrality [phys. rev. e **64**, 016132 (2001)]. *Phys. Rev. E* **73**, 039906 (2006).
29. Freeman, L. C., Borgatti, S. P. & White, D. R. Centrality in valued graphs: A measure of betweenness based on network flow. *Soc. Networks* **13**, 141–154 (1991).
30. Doyle, P. G. & Snell, J. L. *Random Walks and Electric Networks* (The Mathematical Association of America, 1984).
31. Bozzo, E. & Franceschet, M. Resistance distance, closeness, and betweenness. *Soc. Networks* **35**, 460–469 (2013).
32. Grinstead, C. M. & Snell, J. L. *Introduction to probability* (American Mathematical Soc., 1997).
33. Page, L., Brin, S., Motwani, R. & Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999–0120, Computer Science Department, Stanford University (1999).
34. Guex, G. & Bavaud, F. Flow-based dissimilarities: Shortest path, commute time, max-flow and free energy. In *Data Science, Learning by Latent Structures, and Knowledge Discovery*, 101–111 (Springer, 2015).
35. Devooght, R. et al. Random walks based modularity: application to semi-supervised learning. In *Proceedings of the 23rd international conference on World wide web*, 213–224 (International World Wide Web Conferences Steering Committee, 2014).
36. Koschützki, D. et al. Centrality indices. In Brandes, U. & Erlebach, T. (eds.) *Network Analysis*, vol. 3418 of *Lecture Notes in Computer Science*, 16–61 (Springer, Berlin/Heidelberg, 2005).
37. Boldi, P. & Vigna, S. Axioms for centrality. *Internet Math.* **10**, 222–262 (2014).
38. Bavelas, A. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.* **725**–730 (1950).
39. Tutzauer, F. Entropy as a measure of centrality in networks characterized by path-transfer flow. *Soc. networks* **29**, 249–265 (2007).
40. White, S. & Smyth, P. Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2003)*, 266–275 (2003).
41. Stephenson, K. & Zelen, M. Rethinking centrality: Methods and examples. *Soc. Networks* **11**, 1–37 (1989).
42. Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.* **2**, 113–120 (1972).
43. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).
44. Benzi, M. & Klymko, C. A matrix analysis of different centrality measures. *arXiv preprint arXiv:1312.6722* (2013).
45. Garcia-Diez, S., Vandenbussche, E. & Saerens, M. A continuous-state version of discrete randomized shortest-paths, with application to path planning. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, 6570–6577 (2011).
46. François, K., Kivimäki, I., Mantrach, A., Rossi, F. & Saerens, M. A bag-of-paths framework for network data analysis. *arXiv preprint arXiv:1302.6766* (2013).
47. Panzacchi, M. et al. Predicting the continuum between corridors and barriers to animal movements using step selection functions and randomized shortest paths. *J. Anim. Ecol.* In press (2015).
48. Peliti, L. *Statistical Mechanics in a Nutshell*. In a Nutshell (Princeton University Press, 2011).
49. Harville, D. A. *Matrix algebra from a statistician's perspective* (Springer-Verlag, 1997).
50. Seber, G. A *matrix handbook for statisticians* (Wiley, 2008).
51. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**, 046110 (2008).
52. (c) OpenStreetMap Contributors. Licensed under the Open Data Commons Open Database License (ODbL).
53. Kivimäki, I. et al. A graph-based approach to skill extraction from text. *Graph-Based Methods for Natural Language Processing* 79–87 (2013).

## Acknowledgements

This work has been funded by Emil Aaltonen Foundation, InnovIris and the Région Wallonne. I.K. and J.S. also acknowledge support from the Academy of Finland, project no. 260427. I.K. would like to thank Jean-Charles Delvenne and Rainer Kujala for useful discussions and advice and Dries Verdegem for providing the parsed Wikipedia hyperlink network.

## Author Contributions

I.K. B.L. and M.S. together developed the theory behind the paper, I.K. wrote the manuscript, and J.S. took part in planning of the paper. All authors have reviewed the manuscript.

## Additional Information

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Kivimäki, I. *et al.* Two betweenness centrality measures based on Randomized Shortest Paths. *Sci. Rep.* **6**, 19668; doi: 10.1038/srep19668 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>