The Power of Quasi-Shortest Paths: ρ-Geodesic Betweenness Centrality

Dianne S. V. Medeiros, Miguel Elias Mitre Campista, Nathalie Mitton, Marcelo Dias de Amorim, Guy Pujolle

The Power of Quasi-Shortest Paths: ρ-Geodesic Betweenness Centrality

Dianne S. V. Medeiros, Miguel Elias M. Campista, Nathalie Mitton,
Marcelo Dias de Amorim, and Guy Pujolle

Abstract—Betweenness centrality metrics usually underestimate the importance of nodes that are close to shortest paths but do not exactly fall on them. In this paper, we reevaluate the importance of such nodes and propose the *ρ-geodesic betweenness centrality*, a novel metric that assigns weights to paths (and, consequently, to nodes on these paths) according to how close they are to shortest paths. The paths that are just slightly longer than the shortest one are defined as *quasi*-shortest paths, and they are able to increase or to decrease the importance of a node according to how often the node falls on them. We compare the proposed metric with the traditional, distance-scaled, and random walk betweenness centralities using four network datasets with distinct characteristics. The results show that the proposed metric, besides better assessing the topological role of a node, is also able to maintain the rank position of nodes overtime compared to the other metrics; this means that network dynamics affect less our metric than others. Such a property could help avoid, for instance, the waste of resources caused when data follow only the shortest paths and reduce associated costs.

Index Terms—Centrality metrics, betweenness, graph, static and dynamic networks.

1 Introduction

Identifying central nodes in a graph is a fundamental problem in network science [1], [2], [3], [4], [5]. In computer networking, for instance, central nodes may be useful to run a number of control functions or play the role of seeders to help disseminate content [6], [7], [8]. Determining the centrality of a node requires modeling the network as a graph and computing some sort of *centrality metric*, which usually associates the importance of a node with its relative position in the network [9], [10], [11].

One of the most popular centrality metrics is the *betweenness centrality*, which relates the importance of a node to the number of shortest paths it belongs to [9]. It is known that network protocols can greatly benefit from this metric [12], [13], [14], [15]. We argue, however, that using only such paths to assign importance to a node may underestimate other important nodes — in particular, those in the close vicinity of shortest paths but that do not belong to them. This happens when a node that falls on a certain number of shortest paths is classified as more important than another node that belongs to fewer shortest paths but is part of many more "a-little-bit-longer" paths. Yet, we should question why such nodes are neglected. In practice, they are good candidates to maintain the network connected in case a more important node fails. This situation is illustrated in Figure 1, where v_c is part of all shortest paths between networks N1 and N2. Hence, v_c is much more central than v_b according to the traditional betweenness centrality. In turn, v_b achieves a betweenness equal to 0 if we consider only the nodes in N1 and N2. Node v_b, however, is very close to all shortest paths between these networks, differing of 1 hop only from the shortest path. It is so close that, if v_c fails, all the shortest paths will be deviated to v_b and its betweenness will instantaneously grow. We believe that ignoring v_b is not reasonable, even prior to the failure, as the node is always close to the shortest paths between N1 and N2 and can be part of the backup paths between them.

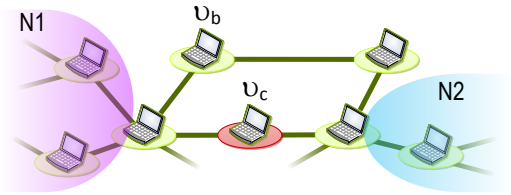


Fig. 1. Node v_c has a high betweenness because it falls on most shortest paths between networks N1 and N2. In opposition, v_b is almost completely forgotten, being assigned a very low betweenness, even though it will assume the role of v_c in case this node fails.

This work questions the use of shortest paths as the sole parameter to quantify the importance of nodes [16], [17], [18], [19], [20]. We propose a weighted betweenness centrality metric that we call *ρ-geodesic betweenness centrality*. The key idea is to extend the definition of the traditional betweenness to also consider the contribution of paths that are a little bit longer than the shortest ones, herein defined as *quasi*-shortest paths. In a nutshell, the ρ-geodesic betweenness of a node v_k is computed using the proportion of shortest and *quasi*-shortest paths that v_k falls on between all possible pair of nodes in the network. This proportion is weighted by the ratio between the cost of the shortest path

- *Dianne S. V. Medeiros and Miguel Elias M. Campista are with Universidade Federal do Rio de Janeiro (UFRJ) – GTA/PEE-COPPE/DEL-Poli – Rio de Janeiro/RJ, Brazil.*
  Email: {dianne, miguel}@gta.ufrj.br
- *Nathalie Mitton is with Inria Lille-Nord Europe – FUN Team – Villeneuve d'Ascq, France.*
  Email: nathalie.mitton@inria.fr
- *Marcelo Dias de Amorim and Guy Pujolle are with Université Pierre et Marie Curie – Paris VI (UPMC) – LIP6 – Paris, France.*
  Email: {marcelo.amorim, guy.pujolle}@lip6.fr

Manuscript received September 21st, 2016.

connecting a pair of nodes and the cost of the *quasi*-shortest path between the same pair of nodes passing through v_k. The search for *quasi*-shortest paths is limited by a parameter ρ, which defines the maximum extra path cost that the proposed ρ-geodesic betweenness can take into account. We will see in this paper that a small ρ is enough to capture well the idea of *quasi*-shortest paths while keeping the computational load low.

We evaluate the proposed metric by comparing it with three existing metrics: traditional betweenness [21], random walk betweenness [18], and distance-scaled betweenness [19]. We compute these metrics for four network datasets. Firstly, we verify if the metrics are capable of pinpointing nodes that should receive a different value for their centralities compared with the traditional betweenness. Secondly, we compare the concordance between the rankings obtained for each metric and assess the degree of differentiation between nodes. Thirdly, we verify the influence of the parameter ρ on the rank variation and, finally, the behavior of the rank is investigated over time, to verify the impact of the metric on the ability to intermediate flows. The results show that (i) the ρ-geodesic betweenness can identify nodes poorly classified by betweenness centralities based on shortest paths, already using low values for ρ. It is also useful to (ii) provide a wider range of rank positions, presenting a more fine-grained classification. Yet, the ρ-geodesic betweenness (iii) reduces the frequency with which a node loses its ability to intermediate flows, considering that flows follow the shortest path rule. In addition, (iv) our metric is able to keep nodes on the same rank position for longer time spans in networks with dynamic topologies.

As a summary, the contributions of this work are:

- We identify the need of a betweenness metric that better captures the positions of the nodes in a network, instead of focusing only on shortest paths.
- We propose the *ρ-geodesic betweenness*, a weighted centrality metric that better evaluates the importance of nodes that not necessarily fall on shortest paths but frequently participate in paths almost as short as the shortest ones.
- We compare our proposal with a number of related metrics, including the traditional betweenness centrality, and show that our solution does identify nodes that are underestimated by other metrics.

This paper is organized as follows. We introduce definitions used in this work in Section 2. Section 3 discusses the related works, presenting the formalization of the betweenness metrics. In Section 4 we present the proposal of this work and we discuss some possible application areas and the need for a new metric. Section 5 discusses the difference between our proposal and its main rival. The evaluation setup is discussed in Section 6, including the analysis guidelines and the dataset description. Section 7 discusses the results and Section 8 concludes this work, presenting future research directions.

2 Network Model, Notations and Definitions

Let us describe the network model we consider in our work as well as the main definitions that are necessary to lay down the basis of our proposal.

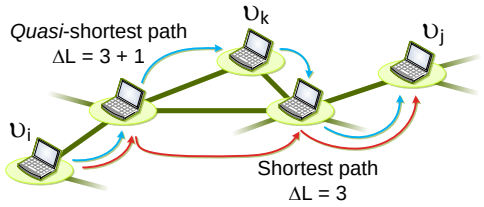


Fig. 2. The shortest path between v_i and v_j is $\Delta L^* = 3$ hops long. If $\rho = 1$, the *quasi*-shortest path of length $\Delta L^* = 4$ through v_k can be considered too.

2.1 Paths and costs

We consider that networks can be modeled as weighted graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega)$, where $\mathcal{V}$ and $\mathcal{E}$ are the sets of vertices and edges, respectively, and the weight ω represents the cost of an edge. Thus, neighbors v_i and v_j are connected by edge $\varepsilon_{i,j}$ whose cost is $\omega_{i,j} \in \mathbb{R}_+$. The edge $\varepsilon_{j,i}$ automatically exists if the graph is undirected. If it is undirected, $\varepsilon_{j,i}$ will exist only if v_j also is neighbor of v_i.

A *path* $p_{1,L}$ between source v_1 and destination v_L is an ordered sequence of distinct nodes in which any consecutive pair of nodes is connected by a link. A path does not contain any loops and any change in the sequence of nodes, either by switching or by shifting a node, originates a new path. We denote the length of path $p_{1,L}$ as $\Delta L = L - 1$, with $L \in \mathbb{N}^*$. The cost of this path is denoted by $\delta_{1,L}$, with $\delta_{1,L} \in \mathbb{R}^*_+$, and it is given by the sum of the individual costs of all links composing the path.

The *shortest path* $p^*_{1,L}$ between v_1 and v_L will be the one for which the cost is the smallest, denoted by $\delta^*_{1,L}$. This path is also known in the literature as the *least cost path*. In this work, we use both *shortest path* and *least cost path* interchangeably. We also consider, without loss of generality, the number of hops as the cost of a path, such that $\delta_{1,L} = \Delta L$, with $\delta_{1,L} \in \mathbb{N}^*$. In this case, the cost of the shortest path is given by $\delta^*_{1,L} = \Delta L^*$. Note that more than one shortest path may exist between the same pair of nodes. We denote the number of shortest paths between v_i, v_j as $n^*_{i,j}$. Yet, we denote the number of shortest paths between v_i, v_j passing through v_k as $n^*_{i,j}(v_k)$.

2.2 Taking nodes on quasi-shortest-path into account

We can now provide two conjugated definitions needed to understand our metric.

Definition 1. *Quasi*-shortest path: *The quasi-shortest path is a path $p_{1,L}$ for which $\delta_{1,L} - \delta^*_{1,L} \leq \rho$, where ρ is called the spreadness factor.*

Definition 2. Spreadness: *The spreadness ρ is the maximum tolerable difference between the costs $\delta_{1,L}$ and $\delta^*_{1,L}$, i.e., $\rho = \delta_{1,L} - \delta^*_{1,L}$, with $\rho \in \mathbb{R}_+$.*

The *quasi*-shortest path is the most important concept of this work. The idea behind it is illustrated in Figure 2, where $\rho = 1$. Such *quasi*-shortest paths are able to increase the importance of nodes that are ignored or underestimated when we consider only the shortest paths to compute the

betweenness – this is the case, for example, of node v_k (that does not fall on any shortest paths). Nevertheless, this node is very close to all shortest paths between both sides of the network, as represented by nodes v_i and v_j, respectively. Paths going through v_k differ from the shortest path by only one hop. Note that more than one *quasi*-shortest path with the same cost can exist between two nodes and more than one of the paths between these nodes can pass through the same intermediary node. Therefore, we represent the number of *quasi*-shortest paths between v_i, v_j as $n_{i,j}$ and, among those, the ones passing through v_k as $n_{i,j}(v_k)$.

The spreadness ρ defines how much we can stretch the geodesic, i.e., how long the *quasi*-shortest paths can be. This limitation avoids the explosion of the number of possible paths. Although we defined $\rho \in \mathbb{R}_+$, in this work we consider the number of hops as cost metric and, thus, $\rho \in \mathbb{N}$. The spreadness limits the search depth to look only for *quasi*-shortest paths that are slightly longer than the shortest path. The idea is based on the fact that the throughput of information traveling through paths for which $\delta_{1,L} \gg \delta^*_{1,L}$ is expected to be low. Note that if $\rho = 0$, $\delta_{1,L} = \delta^*_{1,L}$, and only the shortest paths are considered.

3 Centrality metrics: Shortest paths and other alternatives

It is common to rely on the notion of centrality to quantify the importance of a node. Examples of centrality metrics are degree, closeness, and betweenness [9], [10], [11], [20]. The degree relates to the popularity of a node, the closeness to how quickly it can access or spread resources (e.g., information), and the betweenness to the control of a node over network flows [9]. We focus on the betweenness, formally introduced by Freeman in 1977 [21] based on the intuitions revealed in several previous works, using paths to determine the importance of a node.

3.1 Betweenness centrality

The idea behind the betweenness centrality is that the more a node v_k is centrally positioned, the more it falls on shortest paths between other nodes [22], [23], [24], [25]. Hence, such nodes are strategically positioned and can influence the network by controlling the flow of information. Considering a pair of nodes v_i, v_j, the control exerted by v_k over the flows between these nodes increases with the number of shortest paths between them that cross v_k.

Freeman assumes that the probability that a message passes through one of the existing shortest paths between v_i, v_j is $1/n^*_{i,j}$ [21]. Hence, we can randomly pick one of such paths passing through v_k with probability [21]:

$$b_{i,j}(v_k) = \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}, \tag{1}$$

which can be averaged for all pairs in the network, defining the overall *betweenness centrality* of v_k as:

$$B_{trad}(v_k) = \sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}, \tag{2}$$

where $i \neq k$, $j \neq i$, and $j \neq k$[1]. The betweenness can be normalized by the maximum possible value assigned to a node in a network. This is obtained for the central node in a star graph with the same number of nodes as the network in analysis. The betweenness for this central node is equal to the number of paths it falls on: $(1/2)(|\mathcal{V}|-1)(|\mathcal{V}|-2)$ [21], in undirected graphs; or $(|\mathcal{V}|-1)(|\mathcal{V}|-2)$, in directed graphs.

Freeman suggests that his metric is suitable for networks where node betweenness can potentially impact the examined process, such as in communication networks, where it is highly relevant to know the potential to control the communication for each node [21]. Nevertheless, Freeman's betweenness is limited to simple graphs, leaving aside the strength or cost of the relationship between adjacent nodes (weight). In the remainder of this paper, we refer to Freeman's betweenness as the *traditional betweenness* or B_{trad}.

Freeman also assumes that the information flow is always governed by the shortest-path rule, which may not be true in some cases. For instance, rumors and diseases spread randomly. Rumors can be, in addition, intentionally channeled through specific intermediaries [26]. Policies [27] and the placement of virtual machines, on the other hand, are neither necessarily ruled by randomness nor shortest paths. Instead, they usually follow previously defined requirements, e.g., to meet energy constraints or performance goals. Yet, in mesh networks, we should search for the highest capacity links, which do not always coincide with the shortest paths, it must account the weights of the links.

Many works already questioned the shortest-path rule, proposing new metrics to quantify the importance of a node [16], [17], [18], [19], [20], [28], [29]. Some of them also tried to tackle this issue in weighted networks [16], [20]. The most simple proposals were made by Borgatti and Everett [19] and Geisberger et al. [29], as discussed next.

3.2 Bounded-distance and distance-scaled betweenness

Borgatti and Everett still focus on shortest paths, but they argue that the length of the path should influence the betweenness because longer paths are less valuable to be controlled or may not be realistic for some networks, such as friendship. Based on these assumptions, Borgatti and Everett propose two approaches to lower the importance of longer paths. In the first one, they simply disregard the shortest paths longer than κ, defining the *bounded-distance betweenness* (B_κ) as formalized in Equation 3:

$$B_\kappa(v_k) = \sum_{\substack{i \in |\mathcal{V}| \\ \Delta L^*_{i,j} \leq \kappa}} \sum_{j \in |\mathcal{V}|} \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}. \tag{3}$$

The second approach considers all shortest paths but weights the betweenness with the inverse of the length of the path, which defines the *distance-scaled betweenness* (B_{dist}), formalized as:

$$B_{dist}(v_k) = \sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} \frac{1}{\Delta L^*_{i,j}} \cdot \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}. \tag{4}$$

1. We always consider these three constraints. Therefore, they will be omitted in the remainder of the manuscript.

3.3 Linearly-scaled betweenness

Geisberger et al. proposed a complementary variation of the distance-scaled betweenness, the *linearly-scaled betweenness* (B_{lin}), in which they also account for the distance between the source and the intermediary node, arguing that intermediary nodes closer to the destination should have more control on the communication. Their metric is formalized in Equation 5:

$$B_{lin}(v_k) = \sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} \frac{\Delta L^*_{i,k}}{\Delta L^*_{i,j}} \times \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}. \quad (5)$$

3.4 Betweenness for weighted networks

Changing the focus to weighted networks, Freeman et al. [16] and Opsahl et al. [20] studied how to quantify the importance of nodes in such scenarios, crucial to represent the strength or the cost of a relationship between two nodes. Freeman et al. interpret the weight of each edge in the graph as the capacity of a channel. Thus, the more distant the node, the narrower the channel and, consequently, the smaller the capacity (until the nodes are set apart, where no channel exists at all). The authors determine the maximum flow, $m_{i,j}$, between a pair of nodes and the maximum flow between these nodes that passes through v_k, $m_{i,j}(v_k)$. To this end, they use the concept of cut sets. If this cut set is removed from the graph, the pair of nodes will no longer be able to communicate. The overall dependency on v_k to the maximum flow defines the *flow betweenness* (B_{flow}), which measures the flow amount supported by a node when the maximum flow is pumped in the network. It is given by:

$$B_{flow}(v_k) = \sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} m_{i,j}(v_k). \quad (6)$$

This value can be normalized by the total flow between all pairs of nodes, given by $\sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} m_{i,j}$.

The main drawback of the flow betweenness is the need to know all the independent sets between each pair of nodes in the network, which increases its time complexity. Opsahl et al. proposed a simpler workaround to handle weighted networks [20]. They extend the traditional concept to account both the number of paths and the strength of the relationship between the nodes. The authors use a modified implementation of Dijkstra's algorithm to find the shortest path using inverted weights tuned by a parameter $\alpha \in \mathbb{R}_+$, which determines the relative importance of the number of links compared to the link weights. The metric accounts for only the number of shortest paths ($\alpha = 0$), only the inverse of the weights ($\alpha = 1$), favors the length of the path over the cost ($0 < \alpha < 1$), or favors the cost of the path over the length ($\alpha > 1$). The metric is formalized as:

$$B^{\omega\alpha}(v_k) = \sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} \frac{n^{\omega\alpha}_{i,j}(v_k)}{n^{\omega\alpha}_{i,j}}, \quad (7)$$

where $n^{\omega\alpha}_{i,j}$ and $n^{\omega\alpha}_{i,j}(v_k)$ represent, respectively, the number of shortest paths and how many of them pass through v_k.

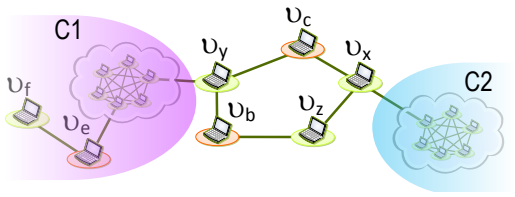


Fig. 3. Example of network where betweenness centrality metrics can fail to capture the importance of a critical node on a *quasi*-shortest path. The clouds represent any type of connected network topology.

3.5 Random walk betweenness

The aforementioned works consider that information in the network always follows some kind of ideal path. Newman claims that a realistic betweenness measure should include paths that are not necessarily the shortest [18]. Thus, he proposes to completely relax the idea of following ideal paths, suggesting that information in the network can wander around essentially at random until it finds its destination. Thus, we should include contributions from many paths that are not optimal in any sense. Hence, he proposes the *random walk betweenness* (B_{rnd}), which measures the number of times that a random walk starting at a source and ending at a destination passes through a node along the way, averaged over all pairs of nodes. Such a metric is computed using matrix methods, and it is proven [18], [28] to be equivalent to the *current flow betweenness* [17], [18]. The formal definition of random walk betweenness is given by:

$$B_{rnd}(v_k) = \sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} I^{(ij)}_k, \quad (8)$$

where $I^{(ij)}_k$ is the amount of information flowing through v_k, which is half the sum of the absolute values of the information flowing along the edges incident to this node.

Note that, in the traditional betweenness, the flow knows exactly where it is going to and which path is the best to arrive there; whereas in the random walk betweenness, it has no prior idea of where the destination is, wandering around at random until the destination is found. Hence, we can consider these metrics as the two extremes of a betweenness centrality spectrum with the other metrics based on shortest paths lying between them.

4 ρ-Geodesic Betweenness Centrality

We propose the *ρ-geodesic betweenness centrality* [2] (B_ρ), which aims to capture the potential of intermediary nodes neglected by shortest-path-based betweenness centralities. Such nodes can be crucial to the network, but are not accounted by typical betweenness metrics just because they do not fall on a sufficiently large number of shortest paths. This is illustrated in Figure 3, in which both v_c and v_b are important to maintain the network components `C1` and `C2` connected, and v_e connects an edge node to the rest of the

2. A preliminary version of this work was accepted at GIIS'16 [30]. Here, we propose a metric that improves the relation between the ratio of shortest and *quasi*-shortest paths and the cost ratio; furthermore, contrarily to the conference paper, we also consider additional paths other than only the *shortest quasi*-shortest paths.

TABLE 1
Comparison of traditional (B_{trad}), distance-scaled (B_{dist}) and random walk (B_{rnd}) betweenness metrics for the nodes highlighted in Figure 3.

Node	B_{trad}	B_{dist}	B_{rnd}
v_c	63.0	12.1	47.6
v_b	9.0	2.5	35.6
v_e	17.0	3.9	17.0

network. The traditional betweenness of v_e is higher than the one of v_b. This is counter-intuitive, because it seems that v_b is topologically more crucial than v_e. Indeed, it can assume a much more important role for the entire network connectivity, compared with v_e, mainly if v_c fails.

Table 1 compares the betweenness computed for v_e, v_c, and v_b using the traditional, distance-scaled, and random walk betweenness, considering two mesh networks composed of six nodes, one connected between v_e and v_y, and the other connected to v_x, as shown in Figure 3. We observe in Table 1 that the random walk betweenness is the only metric able to capture the importance of v_b to the network by giving it more weight than to v_e. This happens because this metric also accounts paths longer than the shortest one.

4.1 Metric formalization

Our approach differs from the aforementioned works by considering in a single metric the number of shortest and *quasi*-shortest paths between all pairs of nodes, as well as the cost of each path. These costs are introduced as a ratio between the cost of the shortest path, $\delta^*_{i,j}$, and the cost of the *quasi*-shortest path through v_k, given by $\delta_{i,j} = \delta_{i,k}+\delta_{k,j}$. Hence, the ρ-geodesic betweenness weights the paths proportionally to their costs, assigning higher importance to nodes on shorter paths. The *maximum* cost of the *quasi*-shortest path depends on the spreadness factor ρ. For instance, if $\rho = C$, the maximum cost considered for the *quasi*-shortest path will be $\delta^*_{i,j} + C$, hence if a certain path costs $\delta^*_{i,j} + (C + \varphi)$, where $\varphi \in \mathbb{R}^*$, it will be ignored in the computation. Note that we account for all the paths with cost $\delta_{i,j} \leq \delta^*_{i,j} + C$, including the shortest one ($C = 0$).

The concept of the ρ-geodesic betweenness (B_ρ) is quite similar to the one of the traditional betweenness, which can be understood as the frequency with which v_k falls on shortest paths between all pairs of nodes in the network. Analogously, the proposed metric measures the frequency with which v_k falls on paths that cost less than or equal to $\delta^*_{i,j}+\rho$. The idea behind the limitation imposed by ρ is based on the fact that the throughput of information traveling through paths for which $\delta_{i,j} \gg \delta^*_{i,j}$ is expected to be low. The proposed metric is formalized in Equation 9.

$$B_\rho(v_k) = \sum_{\substack{i\in|\mathcal{V}|\\ \delta_{i,k}+\delta_{k,j}-\delta^*_{i,j}\leq\rho}} \sum_{j\in|\mathcal{V}|} \frac{n^*_{i,j}(v_k) + n_{i,j}(v_k)}{n^*_{i,j} + n_{i,j}} \times \frac{\delta^*_{i,j}}{\delta_{i,k} + \delta_{k,j}}. \tag{9}$$

Again, if $\rho = 0$, $\delta_{i,j} = \delta^*_{i,j}$, and only the shortest paths are accounted for. In addition, the metric is computed for source-destination nodes that lie in the same component, such that each partial value is equal to zero if these nodes are in different components.

4.2 Properties

The ρ-geodesic betweenness centrality has the following properties:

- It considers the number of multiple paths, both shortest and *quasi*-shortest.
- It increases with the participation of v_k in both shortest and *quasi*-shortest paths.
- It prioritizes low cost paths by decreasing the contribution of expensive paths through a cost ratio.
- It grows with the centrality of the node.

Note that, in this work, the node is considered more central if it participates on multiple paths, either shortest or *quasi*-shortest. The reason behind this consideration is that nodes that participate in several *quasi*-shortest paths should not be discarded just because they are not on the shortest path, as they could be important in many situations. For instance, such nodes that are so close to the shortest path could serve as backup nodes during a network failure. In Figure 3, for example, v_b is part of a possible backup path between both sides of the network. The ρ-geodesic betweenness of nodes v_c, v_b, and v_e for $\rho = 3$ are equal to 117.10, 39.56 and 35.62, respectively. Thus, we note that v_b is now given the importance we intuitively believe it should have when compared to the other highlighted nodes.

The upper and lower limits of each partial term of the metric depend on the proportion of shortest and *quasi*-shortest paths that v_k participates. In addition, these limits depend on the ratio between the costs of such paths. The value of ρ can modify the proportion of node participation on shortest and *quasi*-shortest paths. As a consequence, it can influence the limits of each partial term, being able to decrease the lower limit down to 0 if $\delta_{i,k} + \delta_{k,j} - \delta^*_{i,j} > \rho$.

Higher values of ρ allow to find more *quasi*-shortest paths and if ρ is sufficiently high to account at least one of these paths, the lower limit will tend to 0 if the cost of the *quasi*-shortest paths is much greater than the cost of the shortest path between the same pair of nodes, i.e., $\delta_{i,k} + \delta_{k,j} \gg \delta^*_{i,j}$. Note that if the cost is ∞, the nodes are considered as not reachable, meaning that the contribution to the ρ-geodesic betweenness is null. The lower limit will also tend to 0 if the value of ρ provides too many *quasi*-shortest paths, such that the number of existing paths between v_i, v_j is much greater than the number of such paths that v_k falls on, i.e., $n^*_{i,j} + n_{i,j} \gg n^*_{i,j}(v_k) + n_{i,j}(v_k)$. In the best case scenario, the upper limit of each term is equal to 1, when v_k only falls on shortest paths and participates in all shortest paths connecting v_i, v_j, meaning that $n^*_{i,j}(v_k) + n_{i,j}(v_k) = n^*_{i,j} + n_{i,j}$ and $\delta^*_{i,j} = \delta_{i,k} + \delta_{k,j}$

Another important characteristic of the ρ-geodesic betweenness is its intrinsic higher variance, compared to other shortest-path-based centrality metrics, such as the traditional and distance-scaled betweenness. As so, we can have a broader spectrum to classify nodes according to their importance and, thus, we achieve a more fine-grained node ranking. This is specially true for higher values of ρ. Further, the ρ-geodesic betweenness is able to assign importance to nodes even if their ego network density is unitary, whereas the aforementioned metrics cannot, as we observe in Figure 4. If we consider only the set of nodes $\{v_a, v_b, v_c, v_d, v_e\}$, it is clear that using only shortest

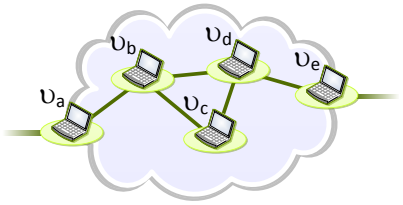


Fig. 4. The nodes in the example can be divided in 3 sets according to their ability to intermediate flows: $\{v_a, v_e\}$ cannot intermediate them, $\{v_b, v_d\}$ intermediate the great majority, and $\{v_c\}$ intermediates if necessary. Nevertheless, shortest-path-based centralities, such as the traditional and the distance-scaled betweenness, classify v_c in the same set of v_a and v_e.

TABLE 2
Comparison of the betweenness of nodes in Figure 4. Both the random walk (B_{rnd}) and ρ-geodesic betweenness (B_ρ) are able to broaden the node ranking. The distance-scaled (B_{dist}) and traditional (B_{trad}) betweenness cannot capture the importance of node v_c.

Node	B_{trad}	B_{dist}	B_{rnd}	$B_{\rho=1}$
v_a	0.0	0.0	0.0	0.0
v_b	3.0	1.3	3.7	2.7
v_c	0.0	0.0	1.3	1.3
v_d	3.0	1.3	3.7	2.7
v_e	0.0	0.0	0.0	0.0

paths will lead us to a condensed ranking, with only two positions, occupied by two groups of nodes: $\{v_b, v_d\}$ and $\{v_a, v_c, v_e\}$. Nevertheless, v_c is clearly different from v_a and v_e in the sense that it can obviously intermediate communications, if necessary, while the former cannot because they are endpoints. Therefore, we argue that the second group should not be composed by $\{v_a, v_c, v_e\}$. Instead, v_c should be reclassified as more important than the other two nodes in this group, broadening the ranking. This reclassification is achieved by both the random walk and the ρ-geodesic betweenness, as we observe in the results of Table 2. Note, however, that the ρ-geodesic betweenness metric assigns lower importance to v_c proportionally to v_b and v_d (1.3 *vs.* 2.7) than the random walk betweenness (1.3 *vs.* 3.7).

4.3 Implementation

The proposed metric is implemented using the algorithm described in Algorithm 1. In this work, we use the number of hops as cost metric and, thus, $\rho \in \mathbb{N}$. Therefore, we use $\Delta_{max} = \rho + 1$ vectors $D^{\Delta}_{\texttt{src}}$ and $N^{\Delta}_{\texttt{src}}$ for each `src` to account for the paths that cost from 0- to ρ-hops more than the shortest one. Vectors $D^{\Delta}_{\texttt{src}}$ and $N^{\Delta}_{\texttt{src}}$ are composed by `numNodes` elements each and D^{Δ}_{src} represents the path cost between `src` and all other nodes, while N^{Δ}_{src} has the number of paths between these nodes. Hence, for a given ρ, we have $D^{\Delta}_{\texttt{src}} = [\delta_{\texttt{src},1}, \ldots, \delta_{\texttt{src},\texttt{numNodes}}]$ and $N^{\Delta}_{\texttt{src}} = [n_{\texttt{src},1}, \ldots, n_{\texttt{src},\texttt{numNodes}}]$. Note that $0 \leq \Delta \leq \rho$, where $\Delta = 0$ refers to arrays concerning the shortest paths, and $\Delta > 0$, to the ones regarding the *quasi*-shortest paths of cost $\delta^*_{i,j} + 1 \leq \delta_{i,j} \leq \delta^*_{i,j} + \rho$. Yet, $N\_k^{\Delta}_{\texttt{src}}$ represents $\Delta_{max} = \rho + 1$ matrices with size `numNodes` $\times$ `numNodes`, where each matrix represents one source node. In these matrices, each element is the number of paths between `src` and all other nodes that v_k falls on. Hence, $N\_k^{\Delta}_{\texttt{src}}$ for a given ρ is represented by:

Algorithm 1 BASIC ρ-GEODESIC BETWEENNESS

Input: ρ, G
Output: ρ-GB
1: **for** `src` $\leftarrow 1,$ `numNodes` **do**
2: $D^{\Delta}_{\texttt{src}}, N^{\Delta}_{\texttt{src}}, N\_k^{\Delta}_{\texttt{src}}, \texttt{T}_{\texttt{src}} \leftarrow$ INITIALIZE(G, ρ)
3: $D^{0}_{\texttt{src}}, N^{0}_{\texttt{src}}, \texttt{T}_{\texttt{src}} \leftarrow$ FIND_SP$(\texttt{src}, \rho)$
4: $D^{\forall\Delta>0}_{\texttt{src}}, N\_k^{\forall\Delta>0}_{\texttt{src}}, N^{\forall\Delta>0}_{\texttt{src}} \leftarrow$ FIND_QSP$(src, \texttt{T}_{\texttt{src}})$
5: ρ-GB$\leftarrow$ACCUMULATE$(\rho, \texttt{src}, D^{\Delta}_{\texttt{src}}, N^{\Delta}_{\texttt{src}}, N_k^{\Delta}_{\texttt{src}}, \rho\text{-GB})$

Algorithm 2 ACCUMULATE CONTRIBUTIONS FROM `src` TO ALL NODES

Input: $\rho, \texttt{src}, D^{\Delta}_{\texttt{src}}, N^{\Delta}_{\texttt{src}}, N\_k^{\Delta}_{\texttt{src}}, \rho$-GB
Output: ρ-GB
1: **for** `dest` $\leftarrow 1,$ `numNodes` **do**
2: **for** $k \leftarrow 1,$ `numNodes` **do**
3: **if** $v_k \neq v_{src}$ & $v_k \neq v_{\texttt{dest}}$ & $v_{\texttt{src}} \neq v_{\texttt{dest}}$ **then**
4: **for** $\Delta \leftarrow 0, \rho$ **do**
5: **if** $\exists$ SP through v_k $\|$ $\exists$ QSP through v_k **then**
6: $\rho\text{-GB}_k \leftarrow \rho\text{-GB}_k + \frac{N\_k^{0}_{\texttt{src,dest}} + N\_k^{\Delta}_{\texttt{src,dest}}}{N^{0}_{\texttt{src,dest}} + N^{\Delta}_{\texttt{src,dest}}} \cdot \frac{D^{0}_{\texttt{src,dest}}}{D^{\Delta}_{\texttt{src,dest}}}$

$$N\_k^{\Delta}_{\texttt{src}} = \begin{bmatrix} n_{1,1}(v_k) & \ldots & n_{1,\texttt{numNodes}}(v_k) \\ \ldots & \ldots & \ldots \\ n_{\texttt{numNodes},1}(v_k) & \ldots & n_{\texttt{numNodes},\texttt{numNodes}}(v_k) \end{bmatrix}.$$

Vector $\texttt{T}_{\texttt{src}}$ is composed by `numNodes` elements and it contains the maximum allowed cost for the *quasi*-shortest path between `src` and all other nodes, i.e., $\texttt{T}_{\texttt{src}} = [\delta_{\texttt{src},1} + \rho, \ldots, \delta_{\texttt{src},\texttt{numNodes}} + \rho]$. Finally, ρ-GB is a vector composed by `numNodes` elements, where each element is the ρ-geodesic betweenness of an intermediary node v_k.

Algorithm 1 uses as input the matrix representation of the network, G, and the spreadness factor, ρ. It returns the vector ρGB, which contains the ρ-geodesic betweenness of every node. The function `INITIALIZE` is responsible to create the arrays and initialize them with the proper values. As we use the number of hops as the cost of the path, the function `FIND_SP` implements the Breadth-First Search (BFS) algorithm to find the shortest paths, while the function `FIND_QSP` implements the Depth-First Search (DFS) algorithm, constrained in depth by vector `T`, to find the *quasi*-shortest paths. Note that `FIND_SP` uses ρ as input only to compute the maximum allowed length of the *quasi*-shortest paths. If a real cost was used, for instance, these functions must be changed. A possible candidate is to modify the Dijkstra algorithm to compute paths that costs more than the shortest one.

The `ACCUMULATE` function is described in Algorithm 2, and it is responsible for summing up the contribution of each pair of nodes v_i, v_j to the betweenness of the intermediary node v_k. Each time this function is called it updates the ρ-geodesic betweenness of the intermediary nodes that participate in the paths from `src` to all the other nodes. Consequently, in the end of the loop of the basic algorithm, vector ρ-GB will contain the ρ-geodesic betweenness of every node due to every single pair of nodes in the network.

4.4 Application

We saw in Table 2 that both the random walk and our metric are able to assign more importance to nodes that are not on shortest paths. This will be more broadly confirmed in Section 7, using real datasets. One may ask, then, why bother to create another metric if the random walk betweenness proposed by Newman [18] does the same job gracefully. The main reason for that is the basis conjecture behind the random walk betweenness itself, which cannot be applied in some cases. Newman states in his work [18] that his metric suits well situations where information may follow random paths until it finds its destination, and he considers that information may not know where it is going to. This may be true if no global knowledge about the network structure exists, or if we are trying to model the natural spread of diseases, for instance. As a counter example, in a computer network where end-to-end paths do exist, and the source of information knows exactly who is the target of the message, it will always *try* to use the most efficient path. Similarly, in transport networks, a driver or a delivery vehicle will always be more interested in using the shortest path. In both networks, the packet delivery fits better to model the flow process in the network. In these networks, and others, it is not always that the shortest path will be the best option. That is why we also need to consider longer paths.

We argue that the utilization of *quasi*-shortest paths, or at least considering them as reasonable alternatives, is a choice that can be driven by (i) reactive or (ii) proactive situations. In the first case, entities try to escape from the common-sense, a.k.a., the shortest path, to avoid unwanted consequences that are already expected to happen. For instance, a packet may be sent through a *quasi*-shortest path if the shortest path between two nodes in a computer network is congested, or a node, or link, in this path is expected to fail. Also, the driver in a transport network can chose a-little-bit-longer paths during rush hours to avoid jammed shortest paths. The idea is that it is better to take a little extra time to arrive, when compared to the normal shortest path, than to either risk being blocked on the normal shortest path or being forced to take alternative paths on-the-fly. As for the proactive situations, the idea is to avoid, beforehand, to damage the shortest path in the near future. This situation can happen whenever multiple alternatives exist and any one of them could be picked according to a given criteria. For instance, each packet flow can be sent through different paths so as to prevent congestions in computer networks. In the same sense, the audience of a soccer game may also follow different trajectories using different gates to enter a stadium. Even in social networks we can observe the situation where information may occasionally follow a path that is neither shortest nor random, e.g., the act of a friend telling a secret of a third person to a common friend. In both proactive and reactive situations, the entity arbitrarily choses a slightly longer path when there is a high chance that the shortest path is damaged or will be damaged in the near future. We use the spreadness factor to denote the additional cost the entity is willing to pay to arrive at the destination the fastest possible, considering that the shortest path can be damaged in some sense. This can be better understood using an analogy. Suppose that we have a set of pipes, with different diameters, ending in a container. The shorter pipes are also the larger ones, whereas thin pipes are long. We want to fill the container the fastest we can with some kind of solid particle. We cannot push all the particles through the shorter pipe because at some point it will be clogged. Hence, the fastest way to fill the container is to push the particles through all the pipes. Nevertheless, we cannot use some thinner pipes because the solid particles do not fit into them. In this case, the spreadness factor could model the diameter of the pipe, so that only the ones into which the particles fit can be used.

5 Random Walk vs. ρ-Geodesic Betweenness

The random walk betweenness considers *all* existing paths between any pair of nodes in the network, no matter the path length. The contribution of each path to the importance of a node is proportional to the probability of using the path. This probability, in turn, varies simultaneously with the length of the path and the degree of the nodes on it. If successive nodes have high degree or if the path is long, the contribution will be lower. Unlike random walk betweenness, in ρ-geodesic betweenness, we weight the contribution only as a function of the path length. In addition, we reduce the number of paths according to the spreadness factor. Hence, the contribution of longer paths tends to decrease more quickly for the random walk betweenness, as long as the nodes on the path have degree greater than 2.

To incorporate the several paths considered in the random walk betweenness we use mainly two approaches. The first one simulates several random walks between pairs of nodes. This method allows for the computation of approximated values of the random walk betweenness in a distributed fashion [31], [32]. Either sequential or distributed algorithms, however, require extra attention to not allow the random walks to loop over the same sequence of nodes, which would erroneously increase the importance of nodes that are traversed many times. Moreover, we need to be able to stop the simulation at a step where the values computed for the random walk betweenness approximate the exact value given by Newman's algorithm [18]. Note that the convergence time can be unfeasible for some applications when using this method [31]. The second approach computes the metric using Newman's algorithm, which applies a matrix approach in a very elegant fashion. This approach, however, is not appropriate for disconnected or directed graphs, due to the generation of null determinants that prevents further computation. The complexity of this algorithm is $O((m+n)n^2)$, which is roughly $O(n^3)$ in sparse and $O(n^4)$ in dense networks. The ρ-geodesic betweenness, in turn, does not have any restriction regarding the structure of the network. Additionally, since it only considers paths *up to* a length and not all the paths as the random walk betweenness, it is less time consuming. Note that, in some applications, it is reasonable to exclude all paths longer than a threshold to compute the importance of a node, as these paths are much likely neglected. Taking a look at Algorithms 1 and 2, we observe that the functions `INITIALIZE`, `FIND_SP` and `FIND_QSP`, and `ACCUMULATE` are, respectively, $O(n^2)$, $O(m+n)$, and $O(\rho n^2)$, where ρ is a

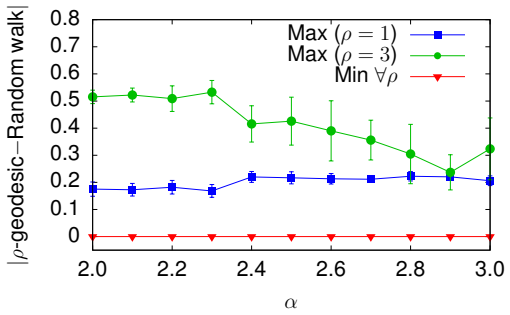


Fig. 5. Averaged results for the maximum and minimum absolute differences between the random walk and the ρ-geodesic betweenness, for $\rho = \{1, 3\}$. The minimum difference is independent of ρ, while the maximum difference becomes more significant for higher ρ.

constant. Hence, the complexity of the ρ-geodesic betweenness metric is reduced compared with Newman's algorithm and it can be computed in $O(n^2)$ or $O(n^3)$, depending if the network is sparse or dense, respectively. The complexity of our metric can be further reduced if the algorithm is parallelized, which is a matter of parallelizing the single-source shortest paths (SSSP) and the accumulation functions in Brandes' algorithm [33], considering unweighted networks. This is feasible [34], [35], [36], [37] and the graph traversal performed in the SSSP needs to be run $\rho+1$ times to find all the paths we need to compute the ρ-geodesic betweenness. In addition, if only local knowledge is available, it is possible to modify a distributed algorithm as the one proposed by Lehman and Kaufman [38] to compute our metric.

As discussed in this section, the random walk and the ρ-geodesic betweenness are quite different, even though their purpose is to account non-ideal paths. The main differences between them is the number of paths considered in the computation and the weight assigned to each one of them. We investigate the impact of this difference on the importance of nodes in synthetic random networks with power law degree distribution ($P \propto degree(v_i)^{-\alpha}$), generated by the Havel-Hakimi [39], [40] algorithm. We chose the power law distribution because it is the most common in real networks [41]. We were able to generate graphs with scaling factor within $1.5 \leq \alpha \leq 4.9$. This is not a problem, because researchers claim that for most real networks α falls approximately between 2 and 3 [41], [42]. Thus, we show the results for this range, without loss of generality.

We generate 10 random graphs for each α and compute the absolute maximum and minimum differences between the values assigned by the random walk and the ρ-geodesic betweenness, considering all nodes in the graph. Figure 5 shows the averaged results for each α. We observe that the minimum difference is always close to zero, while the maximum difference depends on the value of ρ. Note that increasing α means that many more nodes will have very low degree. Particularly, the Havel-Hakimi algorithm originates star-like graphs for higher α. This effect is shown in Figure 6, where the star graph is depicted in Figure 6(e) for comparison. Considering the 1-geodesic betweenness ($\rho = 1$) we note that the maximum difference between the metrics remains almost constant for all α. For the 3-geodesic betweenness ($\rho = 3$), the maximum difference becomes more significant for $2 \leq \alpha \leq 3$. We believe that within this interval the number of highly weighted paths considered by the ρ-geodesic betweenness becomes much greater than the ones for the random walk betweenness. As such paths can be used as backup or offloading paths, nodes that participate on them should be valuable. Hence, our metric is able to predict better which nodes are more important to increase network resilience. As a consequence, the network can achieve better throughput when our metric is used to determine node importance.

6 Evaluation Setup

We analyze the impact and relevance of our metric on four datasets for $\rho \leq 5$, with $\rho \in \mathbb{N}$. Thus, we account for all *quasi*-shortest paths for which $\delta_{i,j} \leq \delta^*_{i,j} + \{1, 2, 3, 4, 5\}$, hence computing the $\{1, 2, 3, 4, 5\}$-geodesic betweenness. The analysis guidelines and the datasets are described next.

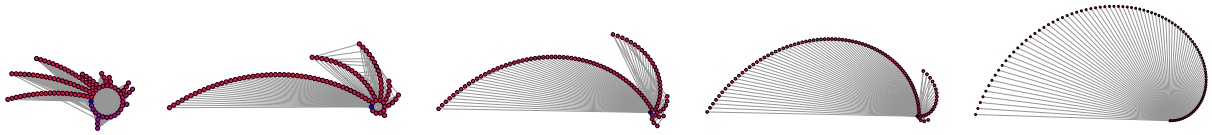
6.1 Analysis guidelines

Our analysis captures the importance of nodes according to their topological distribution in the network. We use the traditional betweenness as the baseline centrality metric to assess the characteristics of the ρ-geodesic betweenness. We begin with the (i) analysis of the correlation between the random walk, distance-scaled and ρ-geodesic betweenness with the traditional betweenness. The goal is to discover how close to the traditional betweenness they are and if they can pinpoint nodes that should be reclassified, even if strongly correlated to the traditional betweenness. Nodes can be reclassified in higher or lower positions, according to its new value of betweenness. Then, we investigate the (ii) behavior of the ranking obtained for each metric, studying the level of agreement between the metrics and the reclassification of nodes. Note that the rank position of a node depends on the value of betweenness assigned to it, such that the first node (most important) has the highest betweenness. Following, we (iii) examine how often we can prevent nodes to lose their ability to intermediate flows, and for how long they can keep the same position.

6.2 Datasets

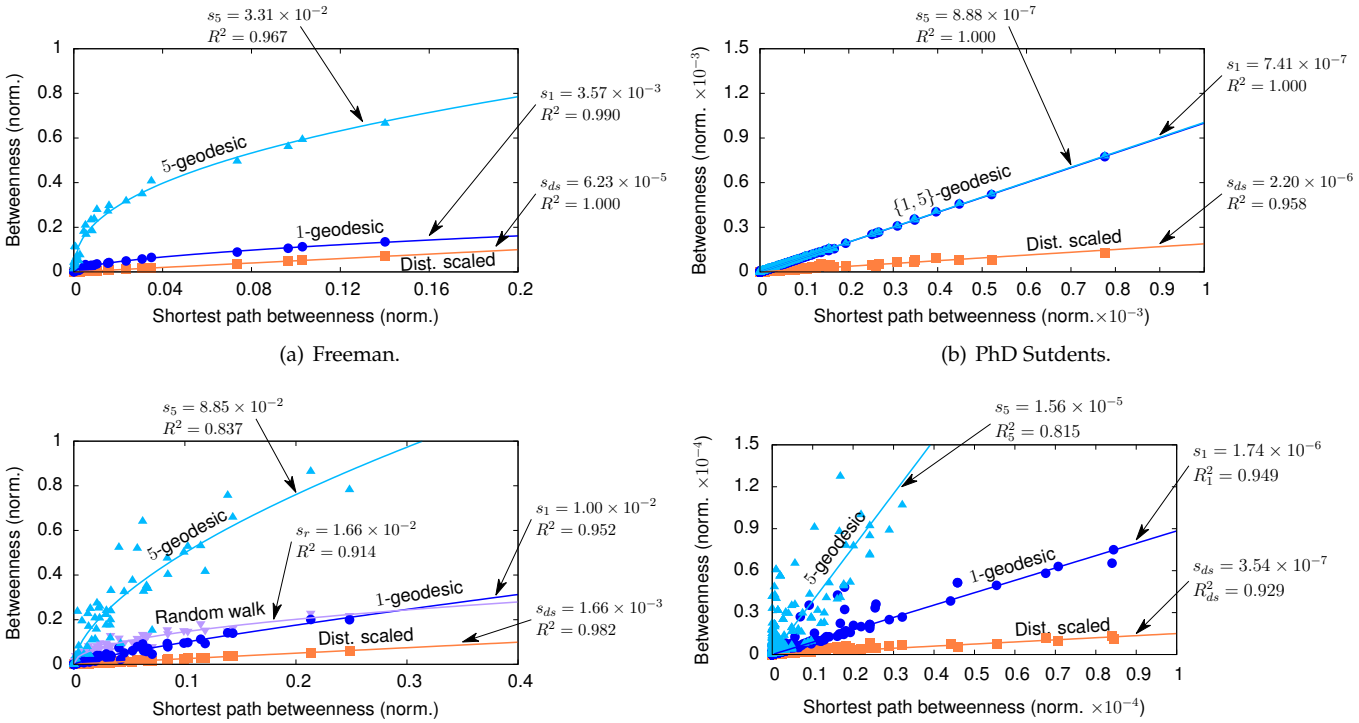
In order to maintain the generality of the metric, we use four datasets with distinct characteristics to evaluate our proposed metric. The importance of nodes is depicted according to its topological position. Smaller nodes have smaller traditional betweenness; more bluish nodes have higher degree; and more reddish, have lower degree.

- **Freeman's EIES**: relationships in a group of 32 academics [43]. A directed edge between two nodes $[v_i, v_j]$ exists only if v_i has sent a message to v_j, totaling 460 links, with a density of 0.464.
- **Dolphins**: association relationships between 62 dolphins in Doubtful Sound, New Zeland [44]. Nodes correspond to dolphins and the interaction between them is represented by an undirected edge $\varepsilon_{i,j}$, totaling 159 links. The density of this network is 0.084.
- **PhD Students**: directed network with density 0.001 representing the relationships between 1,025 PhD



(a) $\alpha = 1.5$ (b) $\alpha = 2.5$ (c) $\alpha = 3.5$ (d) $\alpha = 4.5$ (e) Star, $|V| = 100$

Fig. 6. Comparison between a star network and sample random networks with power law degree distribution for different α. Both networks have 100 nodes and it is clear that the structure of the network changes with α, becoming more similar to a star as α increases.



(a) Freeman. (b) PhD Sutdents.

(c) Dolphins. (d) Cologne #1.

Fig. 7. The correlation with the traditional betweenness is clearly strong for all metrics, being stronger for the distance-scaled betweenness. The random walk and ρ-geodesic betweenness show more capability to identify nodes that should receive a different value for the betweenness. Note that the axis are normalized by $(|\mathcal{V}| - 1)(|\mathcal{V} - 2)$ if the graph is asymmetric, and by $0.5 \cdot (|\mathcal{V}| - 1)(|\mathcal{V} - 2)$ otherwise.

students and supervisors [45]. A directed link exists from v_i to v_j only if v_i is the supervisor of v_j, totaling 1,043 links.

- **TAPASCologne Dataset**: vehicular traffic model of Cologne, Germany [46]. We use 10 samples of the original subset, containing from 1,584 to 1,916 nodes and from 1,573 to 2,044 undirected links, depending on the snapshot sample. Each node is a vehicle and an edge exists between nodes if they are less than 50 meters away from each other. The density of all samples is 0.001.

7 Results

We explore our metric to verify the impact of the *quasi*-shortest paths on the importance of a node to the network. To this end, we divide the results in three categories, according to the guidelines presented in Section 6: (i) the ability of recognizing nodes that should be attributed another value of betweenness, (ii) the changes in the rank position of nodes, and (iii) the ability to intermediate flows.

7.1 Recognition of poorly classified nodes

It is important to know how the metrics relate to the traditional betweenness to discover how the additional requirements of each metric influence the similarity between them. Simultaneously, it is important to know if the metrics can highlight nodes that were over or underestimated, even if they are strongly correlated to the traditional betweenness. The results are shown in Figure 7, where the x-axis is the normalized traditional betweenness and each curve represents one of the other three metrics, also normalized. The normalizing factor is given by $0.5 \cdot (|\mathcal{V}| - 1) \cdot (|\mathcal{V}| - 2)$ for the undirected graphs and by $(|\mathcal{V}| - 1) \cdot (|\mathcal{V}| - 2)$ for the directed ones, as explained in Section 3. In addition, the axes in Figures 7(b) and 7(d) are scaled for better visualization. The random walk betweenness is computed only for the Dolphins dataset due to restrictions of Newman's algorithm. We only show the curves for the 1- and 5-geodesic betweenness ($\rho = \{1, 5\}$, respectively), for the sake of clearness. The curves for the other values of ρ lie between these two.

Figure 7 shows that all metrics are strongly correlated to the traditional betweenness, as their coefficient of determi-

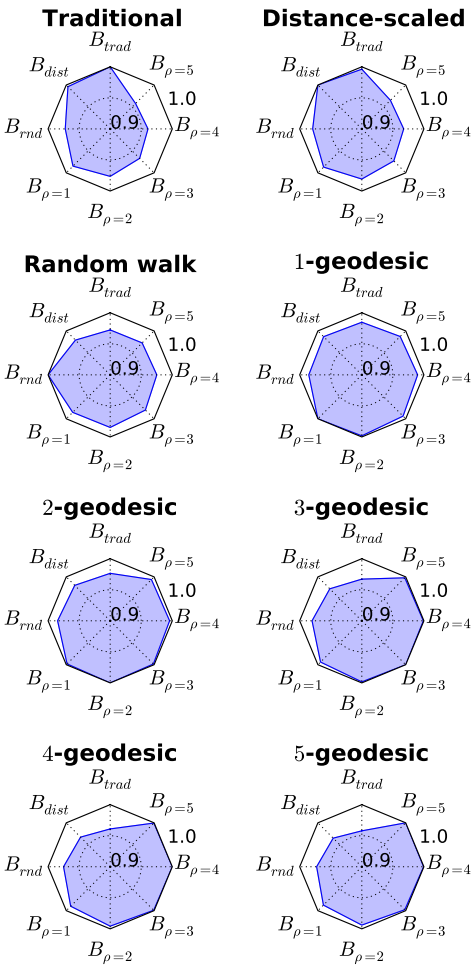


Fig. 8. The Kendall's W coefficients plotted to each pairwise combination of betweenness centrality metrics show a high level of agreement between them. The lowest concordance happens between the traditional and the ρ-geodesic betweenness for $\rho = 5$. This is due to the potentially numerous *quasi*-shortest paths considered on the metric computation, more significantly varying the importance of nodes and, consequently, their rank positions.

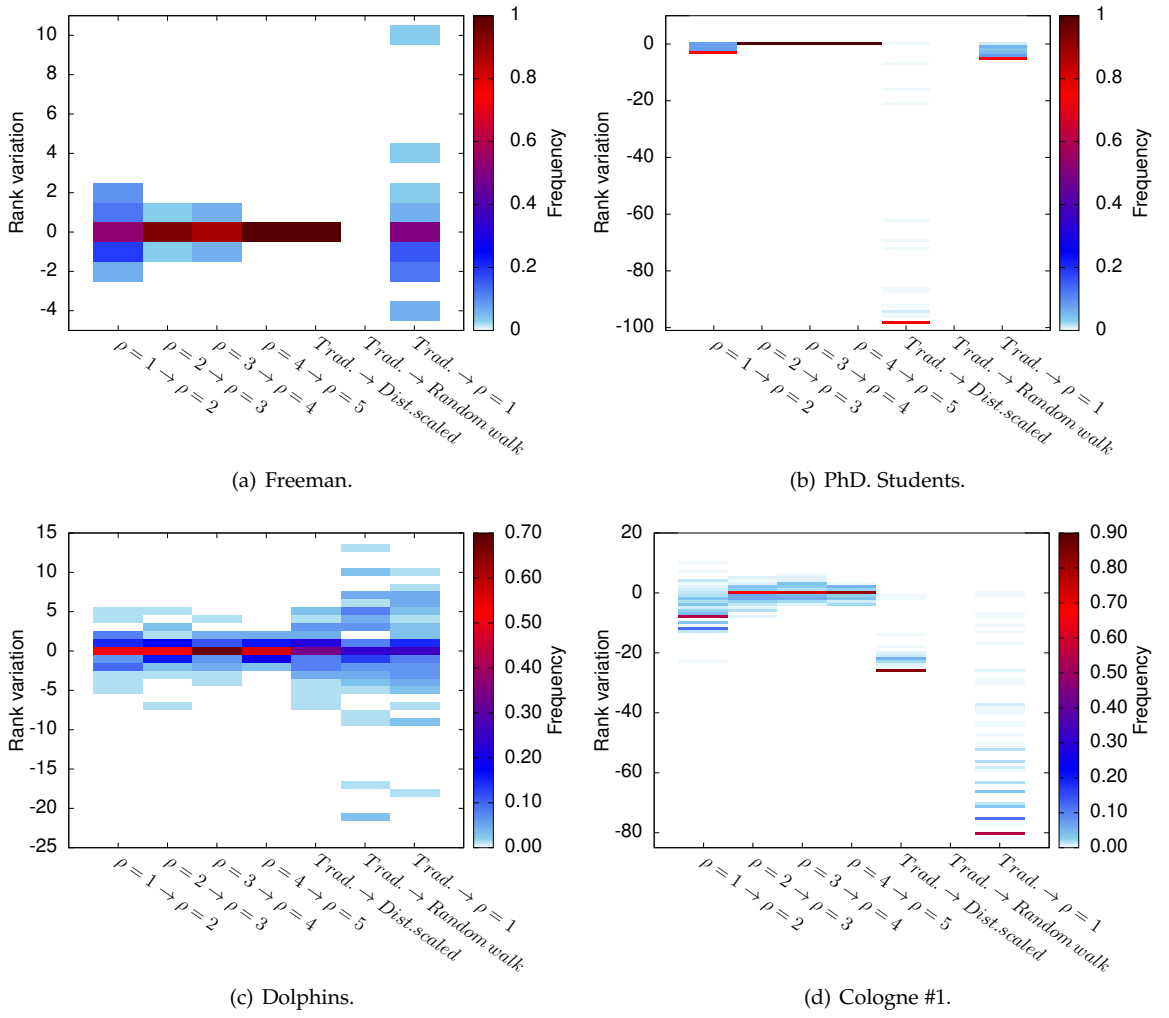
nation (R^2) is high. The strongest correlation is found for the distance-scaled and 1-geodesic betweenness. As ρ increases the correlation decreases, since the participation of nodes in *quasi*-shortest paths increases, pinpointing more nodes that should be reclassified. Nonetheless, $\rho = 1$ is already enough to pinpoint some nodes. This is shown by the dispersion of points around the curve or, mathematically, by the standard deviation of each fitting, combined with the R^2 value. By comparing these parameters for each curve, we find that the higher the standard deviation and the lower the R^2, the more misclassified nodes we can identify. Note, that R^2 cannot be very small (< 0.35), as it would give us only a moderate correlation, meaning that the metrics are almost completely different, and that is not of our interest.

Figure 7(b) shows a singular behavior. The ρ-geodesic betweenness is quite identical to the traditional betweenness in the PhD. Students network, independently of ρ. This happens because the relationships between nodes in this network have strong socialization tendencies, which turns out to produce few multiple paths. The correlation between the random walk and traditional betweenness can be observed in Figure 7(c). We note that, in this scenario, this metric is similar to the 1-geodesic betweenness ($\rho = 1$), and both can almost equally identify that some nodes should be reclassified.

7.2 Impact on node classification

Knowing that the ρ-geodesic betweenness can identify nodes that should be reclassified, we further investigate how it performs this task and how the value of ρ influences the ranking. Such rank is established using the betweenness of the nodes, such that the most important node has the highest betweenness and is the first in the rank, while the node with the lowest betweenness is the less important and, thus, the last in the rank. We use the node ranking for the Dolphins network to analyze the level of agreement between the metrics. To this end we compute the Kendall's W coefficient for each pair combination of the metrics. The more close to the border is the blue octagon in Figure 8, the higher is the level of agreement between the metric and all the others. The ranking provided by the distance-scaled betweenness, for instance, is almost in perfect agreement with the one for the traditional betweenness. The disagreement between the random walk and the traditional betweenness is higher than the one between the 1-geodesic betweenness and the traditional betweenness. As ρ increases, in turn, the disagreement with the traditional betweenness also increases, because the *quasi*-shortest paths accounted become significant. This also happens if we compare the concordance between the random walk betweenness and our metric. This discussion does not reflect, however, the rate with which nodes are reclassified. Although the concordance between the metrics is high, the reclassification rate is also high. For instance, we found that compared to the traditional betweenness, several nodes are reclassified independently of the metric we use. We have a reclassification rate of 66.1% using the distance-scaled betweenness, 75.8% for the random walk betweenness, and for the ρ-geodesic betweenness we have 74.2%, 77.4%, 79.0%, 75.8%, and 77.4% for $\rho \in \{1, 2, 3, 4, 5\}$, respectively. This happens because, contrary to Kendall's W coefficient, the reclassification rate does not account whether the change in the rank position is significant.

In order to investigate the intensity of the reclassification, we analyze in Figure 9 how the rank varies according to the metrics we use. The x-axis represents the transition between the metrics, while the y-axis shows the number of positions that a node gained or lost when we change from one metric to the other. The color grid shows how frequently the nodes gain or lose y positions. Figure 9(a) illustrates the results for the Freeman dataset. We observe that at least half of the nodes keep the same position when we change from the traditional to the 1-geodesic betweenness ($\rho = 1$), as shown by the purplish color for $y = 0$. Note that we can find nodes that gain up to 10 positions if we use the proposed metric. In turn, if we use the distance-scaled betweenness, 100% of nodes stay in the same position. We also observe that increasing ρ affects the ranking with nodes losing or gaining up to 2 positions. The variation stops at $\rho = 4$, as for $\rho = 5$



(a) Freeman. (b) PhD. Students.

(c) Dolphins. (d) Cologne #1.

Fig. 9. The distance-scaled, random walk and ρ-geodesic betweenness are able to redistribute the node ranking to different extents, compared to the traditional betweenness.

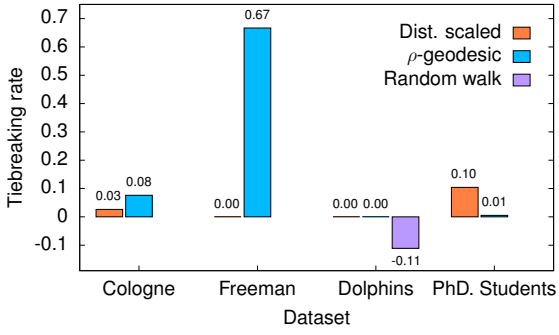
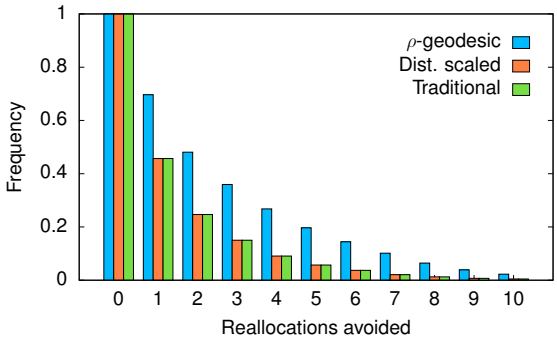


Fig. 10. Compared to the traditional betweenness, the distance-scaled and ρ-geodesic betweenness are able to spread the classification rank, giving room to more positions. Hence, we find less nodes tied in the same position. The random walk betweenness surprisingly increases the number of tied nodes in the Dolphins dataset.

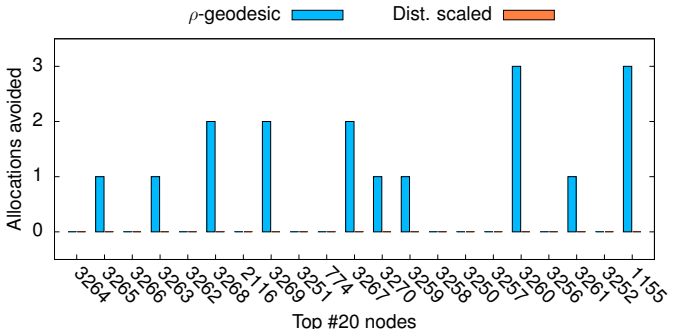
the influence of *quasi*-shortest paths ends. We highlight that, in all scenarios, for $\rho > 2$ most nodes keep their positions unchanged, as shown by the reddish rectangles for $y = 0$.

Figure 9(b) shows the result for the PhD. Students dataset. Some nodes change their position when we use the ρ-geodesic betweenness, but the distance-scaled betweenness has the most significant influence on the ranking for this scenario, as it spreads the classification. This corroborates the correlation results found for this dataset. We observe in Figures 9(c) and 9(d) that all metrics change significantly the node ranking. For the Dolphins dataset the random walk and ρ-geodesic betweenness redistribute several nodes in the rank. The variation on the positions of the rank is representative, with a range that lies, approximately, between $[-21, 13]$ for the random betweenness and $[-18, 10]$ for our metric. In the Cologne network, the ρ-geodesic betweenness has more power of modification compared with the other metrics, ranging from -80 to 28 positions. Note that the number of nodes in each dataset is very different and changing a certain number of positions in each one has a different impact. For instance, if we disregard ties, i.e., each position can be occupied by one node only, a node in the Freeman dataset promoted by 10 positions improves its importance by 31.25%. In contrast, a node in the sample 1 of the TAPASCologne dataset, in the same condition, would improve its importance by only 0.64%.

We further investigate how the centrality metrics evaluated herein behave in the presence of ties. The idea is to find how the metrics assess the importance of nodes once tied in the same rank position by the traditional between-



(a) Prevention of loss of intermediation ability.



(b) Number of times the loss of intermediation ability was prevented for the top 20 nodes.

Fig. 11. The proposed ρ-geodesic betweenness is able to reduce the number of times that nodes lose their ability to intermediate flows in the network compared to the other metrics, even for the most important nodes. The ρ-geodesic betweenness can prevent up to 3 losses more compared to the traditional and distance-scaled betweenness.

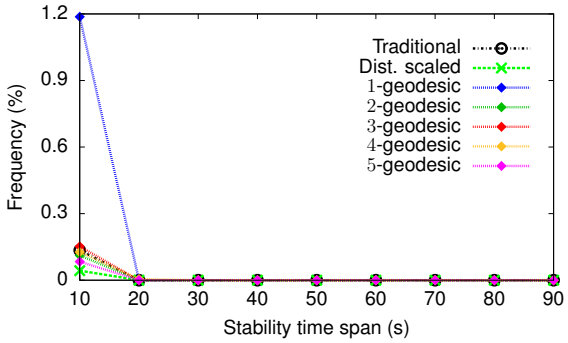


Fig. 12. In the highly dynamic scenario of the TAPASCologne vehicular network, the ρ-geodesic betweenness is the metric that changes less frequently the ranking of nodes. Of all nodes in the network, 1.2% remain in the same rank position for up to 10 seconds.

ness. The result are shown in Figure 10 for all datasets, considering $1 \leq \rho \leq 5$. We define the rate of broken ties as $1 - \#\ tied\ nodes\ other\ metrics / \#\ tied\ nodes\ trad\ vision$. We observe that the tiebreaking rate for the ρ-geodesic betweenness is usually equal or greater than the other metrics. The only exception is the PhD. Students scenario, singular in its construction, which does not provide many multiple paths that could benefit the ρ-geodesic betweenness. Note that, in this figure, a negative tiebreaking rate means that the number of tied nodes increased.

The ρ-geodesic betweenness is strongly correlated to the traditional betweenness but it moves away from the latter as ρ increases. As such, we can identify nodes that should be given more importance. Even for $\rho = 1$ we can find poorly classified nodes, but to a lesser proportion. Following, we investigate how our metric can influence nodes intermediation ability, overtime, supposing that flows follow the shortest-path rule. In this case, not being part of any shortest path implies that the node cannot intermediate communications. It is expected that the number of nodes that participate in shortest path influences how many nodes can be elected to be part of a shortest path. Generalizing, we claim that the more nodes have null betweenness, the less nodes can be elected to participate in a data path or to store any resource. This is harmful especially in dynamic networks or in the presence of node failures, because having less options of nodes to play the role of a data path component could break the network into disconnected components or stop main functionalities if a principal node fails and there is no good option to replace it. Note that resources in this context can have multiple meanings, from information itself to physical or virtual machines that play important roles in the network. To analyze this aspect, we took 10 samples from the TAPASCologne dataset, each 10 seconds.

Figure 11(a) is a cumulative distribution that shows how often we can prevent the loss of ability to intermediate flows. The x-axis is the number of times that we could avoid to lose this ability, while the y-axis represents how often at least x losses were potentially avoided during the period in analysis, i.e., $P(X \leq x)$. While $x = 0$ means that the loss of the intermediation ability was never prevented, $x = 10$ means that nodes never lost this ability. Clearly, the ρ-geodesic betweenness is capable of always avoid the loss of the intermediation ability more than the other metrics. Of course, th0is result is of poor use if the nodes that keep the ability are the ones that are never used because their betweenness is very small, ergo the last nodes in the rank. Therefore, we analyze the top initial 20 nodes in the network to further verify if they also benefit from this behavior.

Figure 11(b) shows how many times more we can avoid nodes to lose their intermediation ability when we change from the traditional to the distance-scaled and ρ-geodesic betweenness, for the initially top-ranked nodes in the network. The x-axis is the node label in ascending order of importance. The y-axis indicates the difference between the number of times the loss was avoided by the distance-scaled and ρ-geodesic betweenness compared to the traditional betweenness, in absolute values. We observe that for the top 20 nodes, the distance-scaled betweenness never avoids more losses than the traditional betweenness. On the other hand, the ρ-geodesic betweenness is able to avoid up to 3 losses more than the traditional betweenness for almost half of these nodes.

For last, we analyze how long nodes can remain in the same rank position using each metric. Figure 12 shows how frequently we can find nodes that can keep the same position during the 90 seconds time interval analyzed for the TAPASCologne dataset. The majority of nodes in this interval frequently jumps between rank positions and none of them is able to maintain the same position for more than 20 seconds. Hence, in Figure 12, we only show the results for $10 \leq x \leq 30$. We observe that few nodes remain in the same

rank position and they do so for approximately 10 seconds maximum. Although this is valid for the minority of nodes in this network, we can quickly note that the ρ-geodesic betweenness is the metric that achieves the highest number of nodes that can keep the same rank position, reaching 1.2% of nodes for $\rho = 1$, which is 6 times higher than the traditional and ρ-geodesic betweenness for $\rho = \{3, 4\}$. Therefore, we claim that the ρ-geodesic betweenness is the metric that can better keep the ranking unchanged for the highly dynamic scenario provided by the Cologne network.

8 Conclusion

We proposed the ρ-geodesic betweenness centrality, a variation of the traditional betweenness that uses both shortest and *quasi*-shortest paths. The idea is to increase the importance of nodes that do not necessarily fall on shortest paths, but can still be considered critical to the network operation, reducing the reorganization and costs of the network upon failure of a central node. The random walk betweenness also follows this idea, but it considers that information travels at random using all existing paths. This is not the case in some situations, such as in the majority of computer and transport networks, and even in some social networks. In addition, the complexity of this metric is higher than the one of our metric. Further, although similar in concept, the ρ-geodesic betweenness is quite different from the random walk betweenness in practice, specially for networks that follow a power law degree distribution with $2 \leq \alpha \leq 3$. We verified the impact of the ρ-geodesic betweenness through the analysis of four datasets with distinct characteristics, for which we also computed the traditional and distance-scaled betweenness. We additionally computed the random walk betweenness for the dataset that represents a completely connected and undirected graph. Our results show that our metric is able to reclassify nodes, promoting those that participate in many paths. Also, the ρ-geodesic betweenness spreads the classification rank, giving room to break ties between nodes, as the number of *quasi*-shortest paths they fall on can be different. The random walk betweenness also present these characteristics, but depending on the dataset, it can increase the number of nodes tied in the same position. We also observed that the ρ-geodesic betweenness has the potential to avoid the loss of the ability to intermediate flows in networks that use shortest path based rules to distribute resources, which can range from information flow to real or virtual machines, lowering associated costs. This is true even for the most important nodes in the network. As a consequence, if we use rules based on the ρ-geodesic betweenness we can potentially reduce the waste of resources. Some of these networks can quickly change their topology and in the vehicular network scenario that we analyzed, we found that the ρ-geodesic betweenness can provide the longest rank stability to the larger number of nodes compared to the other metrics. As future work, we plan to extend our algorithm to compute the metric for weighted networks and we intend to investigate the performance of the network running under rules based on the ρ-geodesic betweenness. Also, we will face it to real networks (by running experiments on realistic platforms) and study its relevance in different use cases.

Acknowledgment

This work has been partially supported by CAPES, CNPq and Faperj, by the CAPES-COFFECUB CROMO project and CPER/FEDER DATA. We would like to thank Prof. Dr. Jefferson Elbert Simões for the helpful discussion and insights.

References

[1] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant MANETs," in *ACM MobiHoc*, 2007, pp. 32–40.
[2] A. Arulselvan, C. W. Commander, L. Elefteriadou, and P. M. Pardalos, "Detecting critical nodes in sparse graphs," *Computers & Operations Research*, vol. 36, no. 7, pp. 2193–2200, 2009.
[3] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science*, vol. 323, no. 5916, pp. 892–895, 2009.
[4] P. Pantazopoulos, M. Karaliopoulos, and I. Stavrakakis, "On the local approximations of node centrality in internet router-level topologies," in *IWSOS*, 2013, pp. 115–126.
[5] A. M. A. H. Seyed Amir Ali Ghafourian Ghahramani and K. Kavousi, "A network model for vehicular ad hoc networks: An introduction to obligatory attachment rule," *IEEE Transactions on Network Science and Engineering*, vol. 3, no. 2, pp. 82–94, 2016.
[6] C. Kiss and M. Bichler, "Identification of influencers - measuring influence in customer networks," *Decision Support Systems*, vol. 46, no. 1, pp. 233–253, 2008.
[7] K. Thilakarathna, A. C. Viana, A. Seneviratne, and H. Petander, "Mobile social networking through friend-to-friend opportunistic content dissemination," in *ACM MobiHoc*, 2013, pp. 263–266.
[8] M. Bouet, J. Leguay, T. Combe, and V. Conan, "Cost-based placement of vDPI functions in NFV infrastructures," *International Journal of Network Management*, vol. 25, no. 6, pp. 490–506, 2015.
[9] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.
[10] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: Social-based forwarding in delay tolerant networks," in *ACM Mobihoc*, 2008, pp. 241–250.
[11] K. Wehmuth and A. Ziviani, "DACCER: Distributed assessment of the closeness centrality ranking in complex networks," *Computer Networks*, vol. 57, no. 13, pp. 2536–2548, 2013.
[12] S. Dolev, Y. Elovici, and R. Puzis, "Routing betweenness centrality," *JACM*, vol. 57, no. 4, pp. 25:1–25:27, Apr. 2010.
[13] A. P. Giles, O. Georgiou, and C. P. Dettmann, "Betweenness centrality in dense random geometric networks," in *ICC*, 2015, pp. 6450–6455.
[14] J. Yim, H. Ahn, and Y.-B. Ko, "The betweenness centrality based geographic routing protocol for unmanned ground systems," in *IMCOM*, 2016, pp. 74:1–74:4.
[15] A. Jain, "Betweenness centrality based connectivity aware routing algorithm for prolonging network lifetime in wireless sensor networks," *Wireless Networks*, vol. 22, no. 5, pp. 1605–1624, 2016.
[16] L. C. Freeman, S. P. Borgatti, and D. R. White, "Centrality in valued graphs: A measure of betweenness based on network flow," *Social Networks*, vol. 13, no. 2, pp. 141–154, 1991.
[17] U. Brandes and D. Fleischer, "Centrality measures based on current flow," in *STACS*, 2005, pp. 533–544.
[18] M. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, no. 1, pp. 39 – 54, 2005.
[19] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social Networks*, vol. 28, no. 4, pp. 466 – 484, 2006.
[20] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, 2010.
[21] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
[22] A. Bavelas, "A mathematical model for group structures," *Human Organization*, vol. 7, no. 3, pp. 16–30, 1948.
[23] A. Shimbel, "Structural parameters of communication networks," *Bulletin of Mathematical Biophysics*, vol. 15, no. 4, pp. 501–507, 1953.
[24] M. E. Shaw, "Group structure and the behavior of individuals in small groups," *The Journal of Psychology*, vol. 38, no. 1, pp. 139–149, 1954.

[25] B. S. Cohn and M. Marriott, "Networks and centres of integration in Indian civilization," *Journal of Social Research*, vol. 1, pp. 1–9, 1958.
[26] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Social Networks*, vol. 11, no. 1, pp. 1–37, 1989.
[27] G. C. Ketan Savla and M. A. Dahleh, "Robust network routing under cascading failures," *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 1, pp. 53–66, 2014.
[28] U. Brandes and T. Erlebach, *Network Analysis: Methodological Foundations*, 1st ed. Springer, 2005.
[29] R. Geisberger, P. Sanders, and D. Schultes, "Better approximation of betweenness centrality," in *ALENEX*, 2008, pp. 90–100.
[30] D. S. V. Medeiros, M. E. M. Campista, N. Mitton, M. D. Amorim, and G. Pujolle, "Weighted betweenness for multipath networks," in *GIIS*, 2016, to appear.
[31] A.-M. Kermarrec, E. Le Merrer, B. Sericola, and G. Trédan, "Second order centrality: Distributed assessment of nodes criticity in complex networks," *Computer Communications*, vol. 34, no. 5, pp. 619–628, 2011.
[32] A. Lulli, L. Ricci, E. Carlini, and P. Dazzi, "Distributed current flow betweenness centrality," in *SASO*, 2015, pp. 71–80.
[33] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, no. 2, pp. 136–145, 2008.
[34] D. A. Bader and K. Madduri, "Parallel algorithms for evaluating centrality indices in real-world networks," in *ICPP*, 2006, pp. 539–550.
[35] ——, "Designing multithreaded algorithms for breadth-first search and st-connectivity on the cray mta-2," in *ICPP*, 2006, pp. 523–530.
[36] A. E. Sariyüce, K. Kaya, E. Saule, and U. V. Çatalyürek, "Betweenness centrality on gpus and heterogeneous architectures," in *GPGPU*, 2013, pp. 76–85.
[37] M. Bernaschi, G. Carbone, and F. Vella, "Scalable betweenness centrality on multi-gpu systems," in *CF*, 2016, pp. 29–36.
[38] K. Lehmann and M. Kaufmann, "Decentralized algorithms for evaluating centrality in complex networks," Wilhelm Schickard Institut, Tech. Rep., 2003.
[39] V. Havel, "A remark on the existence of finite graphs," *Časopis pro pěstování matematiky*, vol. 80, no. 4, pp. 477–480, 1955.
[40] S. L. Hakimi, "On realizability of a set of integers as degrees of the vertices of a linear graph. i," *Journal of the Society for Industrial and Applied Mathematics*, vol. 10, no. 3, pp. 496–506, 1962.
[41] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, no. 5, pp. 50–59, 2003.
[42] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks," *Advances in physics*, vol. 51, no. 4, pp. 1079–1187, 2002.
[43] S. Freeman and L. Freeman, *The Networkers Network: A Study of the Impact of a New Communications Medium on Sociometric Structure*, ser. Social sciences research reports. School of Social Sciences, University of California, 1979.
[44] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, and S. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.
[45] D. S. Johnson, "The genealogy of theoretical computer science," *SIGACT News*, vol. 16, no. 2, pp. 36–44, Jul 1984.
[46] S. Uppoor and M. Fiore, "Large-scale urban vehicular mobility for networking research," in *IEEE VNC '11*, pp. 62–69.

Dianne Scherly Varela de Medeiros received her bachelor and M.Sc degrees on Telecommunications Engineering from Universidade Federal Fluminense (UFF), Rio de Janeiro, Brazil, in 2011 and 2013, respectively. Currently she is a PhD. student at Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, in exchange at Université Pierre et Marie Curie, Paris, France. Her major research interests are in wireless and vehicular networks, data analysis and graph theory.

Miguel Elias M. Campista (S'06-M'10) is an associate professor with Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, since 2010. He received his D.Sc. degree in Electrical Engineering from UFRJ in 2008. In 2012, Miguel has spent one year with LIP6 lab at Université Pierre et Marie Curie, Paris, France, as invited professor. He heads the Electronic area of his department at PEE/COPPE/UFRJ. His major research interests are in wireless networks.

Nathalie Mitton Nathalie Mitton received the MSc and PhD. degrees in Computer Science from INSA Lyon in 2003 and 2006 respectively. She received her Habilitation à Diriger des Recherches(HDR) in 2011 from Université Lille 1. She is currently an Inria full researcher since 2006 and from 2012, she is the scientific head of the Inria FUN team which is focused on small computing devices like electronic tags and sensor networks. Her research interests are mainly focused on self-organization, self-stabilization, energy efficient routing and neighbor discovery algorithms for wireless sensor networks as well as RFID middleware. She is involved in the set up of the FIT platforms (http://fit-equipex.fr/), the FP7 Aspire or VITAL projects and in several program and organization committees such as AdHocNow 2016&2015, VTC 2016, InterIoT 2016 & 2015, MobiCom 2015, AdHocNets 2015 & 2014, HPCC 2014, WiMob 2013, MASS 2012 & 2011, etc. She also supervises several PhD students and engineers.

Marcelo Dias de Amorim received the B.Sc. and M.Sc. degrees in electronic engineering from Universidade Federal do Rio de Janeiro (UFRJ), Brazil, and the Ph.D. degree from Université de Versailles, France. He is currently a CNRS Research Director at the LIP6 Laboratory, Université Pierre et Marie Curie, Paris, France, where he heads the Networks and Performance Analysis team. His research interests include the design and evaluation of mobile networked systems.

Guy Pujolle is a Professor at Université Pierre et Marie Curie, Paris, France. Guy Pujolle is a pioneer in high-speed networking having led the development of the first Gbit/s network to be tested in 1980. He was at the origin of several inventions and important patents like DPI, Wi-Fi controller, virtual networks, metamorphic networks, and green networks.