

# Beyond the shortest path: the path length index as a distribution

Leonardo B. L. Santos<sup>a</sup>, Luiz Max Carvalho<sup>b</sup>, Giovanni G. Soares<sup>c</sup>,  
Leonardo N. Ferreira<sup>d</sup>, Igor M. Sokolov<sup>e</sup>

<sup>a</sup>*National Center for Monitoring and Early Warning of Natural Disasters (Cemaden),  
Brazil*

<sup>b</sup>*School of Applied Mathematics (EMAp), Getulio Vargas Foundation (FGV), Brazil*

<sup>c</sup>*National Institute of Space Research (INPE), Brazil*

<sup>d</sup>*Center for Humans and Machines, Max Planck Institute for Human Development,  
Germany*

<sup>e</sup>*Humboldt University of Berlin, Germany*

---

---

## 1. Motivation

Traversing graphs is a fundamental question in Graph Theory and an important consideration when dealing with complex networks. The traditional complex network approach considers only the shortest paths from one node to another [1], and does not take into account several other possible paths. This limitation is significant, for example, in urban mobility studies [2, 3], where it is important to consider alternative routes between locations.

As mentioned by Lima et al. (2016), in urban mobility settings users choose multiple routes over origin-destination pairs, and those choices often deviate from the shortest time path. Galbrun et al. (2016) highlight that chosen routes may be associated with diverse factors, for instance public safety. Tomas et al. (2022) further support these claims by showing that exceptional events, such as urban floods, may lead users to deviate from routes previously defined to risk-less (and potentially longer) ones.

Estrada and Hatano (2008) proposed the Communicability Index, a number (scalar) that takes into account not only the shortest paths but also all the walks from one node to another. Their approach was based on walks. In

---

*Email address:* santoslbl@gmail.com (Leonardo B. L. Santos)

contrast, here we are interested in paths due to the urban mobility motivation context.

On one hand, the number of walks between each pair of nodes in a simple graph is known analytically [6]. On the other hand, the analogous problem for paths is NP-hard [7]. Roberts and Kroese (2007) presented a stochastic algorithm to estimate the solution of that problem using a sequential importance sampling.

In this short report, as the first steps, we present an exhaustive approach to address the problem of finding all paths between two nodes. We show one can go beyond the shortest path but we do not need to go so far: we present an interactive procedure and an early stop possibility. We apply our ideas to the well-known Zachary’s karate club graph [8]. We do not collapse the distribution of path lengths between a pair of nodes into a scalar number; instead we look at the distribution itself - taking all paths up to a pre-defined path length (considering a truncated distribution), and show the impact of that approach on the most straightforward distance-based graph index: the walk/path length.

## 2. Preliminaries: definitions and notation

In this section, we give a few definitions and results from elementary graph theory to facilitate the understanding of the new results presented herein. Most of the following discussion is standard and can be found in [1, 6].

We start by defining a graph (Definition 2.1) and then a simple graph (Definition 2.2), which will be the main objects of interest in this paper.

**Definition 2.1 (Graph).** *A graph  $G = (V, E)$  is a set of nodes and edges, where  $V$  is the set of  $|V| = N$  nodes and  $E$  is the set of  $|E| = M$  edges.*

An edge (also called link or connection)  $(i, j), i, j \in V$  connects two nodes  $i$  and  $j$ . A self-connection or a loop is a link  $(i, i)$  that connects node  $i$  to itself. Multiple edges are two or more edges that connect the same two vertices.

A link can be undirected or directed. In an undirected graph, all edges  $(i, j)$  connect  $i$  to  $j$  and vice-versa. A directed graph has directed edges (also called arcs)  $(i, j)$ , that connect  $i$  to  $j$ , but not  $j$  to  $i$ , i.e.,  $(i, j) \neq (j, i)$ . A link can also have an associated weight, which is a numeric value.

**Definition 2.2 (Simple Graph).** A graph  $G = (V, E)$  is a simple graph if, and only if, it is undirected, there are no self-connections in  $G$ , no multiple edges or weights.

A helpful object for characterizing a graph is its adjacency matrix, whose definition is given in Definition 2.3.

**Definition 2.3 (Adjacency matrix).** The adjacency matrix  $\mathbf{A}$  of a graph  $G = (V, E)$  is the  $N \times N$  matrix whose entries  $A_{ij}$  are given by

$$A_{ij} = \begin{cases} 1, & \text{if } i, j \in V \text{ share an edge;} \\ 0, & \text{otherwise.} \end{cases}$$

In this paper, we are concerned with traversing the graph, i.e., starting from a source node  $i \in V$ , visiting a collection of nodes, and arrive a target node  $j \in V$ , where  $i = j$  is a possibility. Here we distinguish between trajectories that allow multiple visits to the same node (and associated) edges, called **walks** (Definition 2.4); and trajectories where each node and vertex can only be visited once, called **paths**, presented in Definition 2.5.

We start our discussion with trajectories that can visit the same node multiple times, called **walks**:

**Definition 2.4 (Walk).** Consider a simple graph  $G = (V, E)$  and a pair of nodes  $i, j$  in  $V$ . A walk  $w$  in  $G$  from  $i$  to  $j$  is an alternating sequence of edges and nodes from  $i$  (node of origin/source) to  $j$  (node of destination/target).

With this definition in hand, we are prepared to state Theorem 2.1, which tells us that the number of walks of a given (finite) length is finite so long as  $|V|$  is finite.

**Theorem 2.1 (Finite number of walks).** Consider a simple graph  $G = (V, E)$ . Take  $i, j \in V$ , the number of walks of length  $l$  between  $i$  and  $j$  is given by

$$f_W^{ij}(l) = (A^l)_{ij},$$

where  $A_{ij}$  is the corresponding entry in the adjacency matrix of  $G$  – see Definition 2.3.

*Proof.* This is a well-known result. See Lemma 2.5 in [6]. □

Now, consider trajectories in a graph without ever visiting any node twice. Such a trajectory is called a **path**:

**Definition 2.5 (Path).** Consider a simple graph  $G = (V, E)$  and a pair of nodes  $i, j$  in  $|V|$ . A path  $p_{ij}$  in  $G$  from  $i$  to  $j$  is an open ( $i \neq j$ ) walk from  $i$  to  $j$ , and with no repeated edges or nodes.

As Definition 2.5 makes clear, paths are specializations (restrictions) of walks. This might prompt the reader to think that one can study paths by considering restrictions to results about walks. As we will show later on, this is not always the case. Our approach is somehow similar to Self Avoiding Walks (SAW) [9], but we fix not only the source but also the target for each path.

In this paper, we will devote attention to connected graphs (Definition 2.6), that is, graphs for which there exists at least one path for every pair of vertices  $i, j \in V$ .

**Definition 2.6 (Connected Graph).** A simple graph  $G = (V, E)$  is connected if for every pair of vertices  $i, j$  one can construct a subset  $C_{ij} \subseteq V$ , with  $|C_{ij}| = K$  where the vertices  $c_1, \dots, c_K \in C_{ij}$  are such that  $i$  and  $c_1$  share an edge as do  $j$  and  $c_K$  and also  $c_k$  and  $c_{k+1}$  share an edge, for  $2 \leq k \leq K - 1$ . In other words,  $G$  is connected if and only if one can always construct at least one path between  $i, j \in V$ , for every such pair.

The number of vertices visited in the path  $p_{ij}$  is the path length (Definition 2.7), and the shortest such path (Definition 2.8) is usually of great interest as it is related to many optimization problems, such as the traveling salesman problem. We now state a few more definitions related to traversal of graphs, which will be useful in the remainder of the paper.

**Definition 2.7 (Path Length).** Consider a simple graph  $G = (V, E)$ . The number of edges on the path from  $i$  to  $j$  is the path length ( $l'$ ) of that path.

**Definition 2.8 (Shortest Path Length).** Consider a simple graph  $G = (V, E)$ . The number of edges on the shortest path from  $i$  to  $j$  is the shortest path length ( $s'$ ) of that path. The  $s'$  is a number associated with the pair  $i-j$ : for each pair  $i-j$  there is one and only one  $s'$ :  $s'(i,j)$ .

**Remark 2.1 (Shortest Path).** Consider a simple graph  $G = (V, E)$ . For any two vertices  $i, j \in V$  there is at least one path from  $i$  to  $j$  which the path length is the shortest possible.

### 3. Main problem: computing the frequency and length of walks and paths

First, let us take a look at the number of walks: in a simple graph  $G = (V, E)$ , for any two vertices  $i, j \in V$ , there is an infinite number of walks from  $i$  to  $j$ , which holds even for finite graphs. However, if we take a finite path length, the number of walks with that path length is finite, and it is given by  $f(l_{ij}) = (A^n)_{ij}$ , with  $n = l_{ij}$ . One might now ask what the expected value for  $l_{ij}$  is. As we can always get a walk longer than any other, we cannot define a normalized probability measure; thus, this expectation does not exist.

Let us define the shortest walk length from  $i$  to  $j$ ,  $s_{ij}$ , as the minimum value of  $l_{ij}$  - obviously that the walk associated with this length is a path. Then, once any (finite)  $l_{ij}$  can be expressed as  $l_{ij} = s_{ij} + k$ ,  $k \geq 0$ ,  $k \in N$ , and, therefore, a “truncated” expected value, under a  $k$ -th order approximation, is:

$$\mathbb{E}[l] = \frac{\sum_{n=s_{ij}}^{n=s_{ij}+k} n(A^n)_{ij}}{\sum_{n=s_{ij}}^{n=s_{ij}+k} (A^n)_{ij}}. \quad (1)$$

Now, let us move from walks to paths. Between any pair of nodes  $i$ - $j$  in  $G$ , there is at least one path  $p$ , from  $i$  to  $j$  - we are considering a single connected component in  $G$ . While  $\# p$  is finite, the problem of counting the number of s-t (source-destination) paths in a graph is NP-complete [7].

Here we propose a  $k$ -th order approximation for the case of paths. The length of  $p_{ij}$  is  $l'_{ij}$ , and it is between 1 and  $N - 1$ . Let us define the shortest path length from  $i$  to  $j$ ,  $s'_{ij}$ , as the minimum value of  $l'_{ij}$ . Any  $l'_{ij}$ , therefore, can be expressed as  $l'_{ij} = s'_{ij} + k$ , for a finite value of  $k \leq N - 2$ . Finally, the expected value, under the  $k$ -th order approximation, is:

$$\mathbb{E}[l'] = \frac{\sum_{n=s'_{ij}}^{n=s'_{ij}+k} n f(n)}{\sum_{n=s'_{ij}}^{n=s'_{ij}+k} f(n)}, \quad (2)$$

where  $f(n)$  is the frequency of a  $l'_{ij}=n$ .

There is no analytical expression for  $f(n)$  in the literature. Finding all paths in a graph can be very computationally expensive -  $O(N!)$  in the worst case: a complete graph with order  $N$ . Here we perform a depth-limited

search (DLS) in order to find  $f(n)$ , which can found here.<sup>12</sup>.

It is worth highlighting we do an exhaustive search - finding all possible paths from a node to another. However, the main insight is that we do not need to go so far beyond the shortest paths - in order words: we do not use a so much high value of  $k$  in the  $k$ -th order approximation.

## 4. Results

In this section, we present an analytical result considering complete graphs (4.1), and, based on a depth-limited Search, results for the Zachary’s Karate Club graph (4.2).

### 4.1. Complete graphs

In a complete graph all nodes are directly connected to all others ( $s'_{ij}=1$ ,  $\forall i,j$ ). The number of paths between any pair of nodes is a combinatorial result based on the arrange of  $N-2$  nodes in a path of length  $l'$ .

**Theorem 4.1 (Number of paths in a complete graph).** *Let  $G$  be a complete simple graph. Then the number of paths of length  $k + 1$  is*

$$f(k + 1) = \begin{cases} 1, & k = 0, \\ \prod_{r=2}^{r=k+1} N - r, & 0 < k < N - 1, \\ 0, & k \geq N - 1 \end{cases}$$

It is possible to note that, in a simple but complete graph, between any pair of nodes:

- There is only one walk (and path) of length 1;
- The number of walks grows exponentially with the number of nodes;
- The number of paths grows with the number of nodes, but at a rate inversely related to the number of nodes;

---

<sup>1</sup>[https://github.com/gioguarnieri/all\\_paths](https://github.com/gioguarnieri/all_paths)

<sup>2</sup>We discussed “to go beyond the shortest path” in 2018 and implemented the first complete version of this code in August 2019. The COVID-19 pandemic has changed research agendas worldwide. We resume this paper in 2022.

- The most frequent path length are the longest ones (lengths  $N-1$  and  $N-2$ );
- It is always possible to get a walk longer than a previous one;
- There is no path of length longer than  $N-1$ .

Figure 1 illustrates this result for the a complete graph with 10 nodes (C-10).

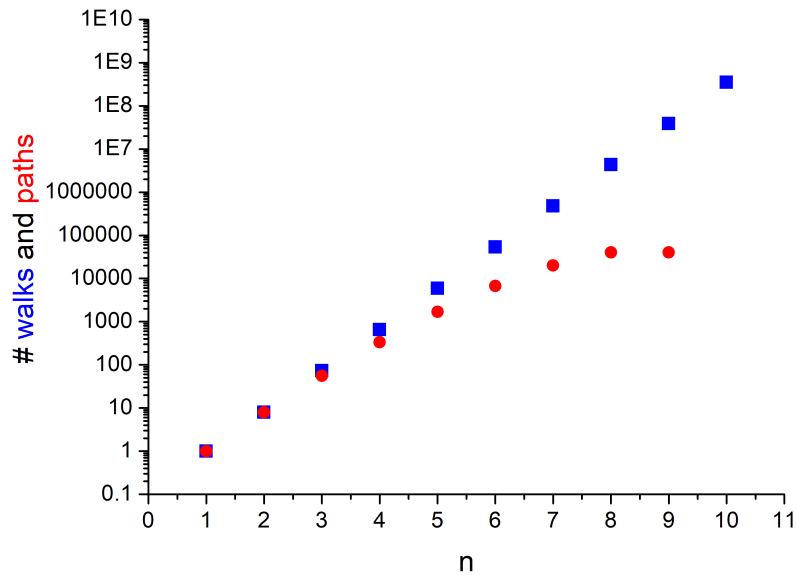


Figure 1: Number of walks (blue squares) and paths (red circles) from 0 to 1, in the C10 graph, for each path length.

#### 4.2. Zachary's Karate Club

Zachary's Karate Club graph is a well known graph [8, 10], with  $N = 34$ ,  $M = 78$ , 1 connected component. Figure 2 shows the Zachary's Karate Club graph, with numerated nodes.

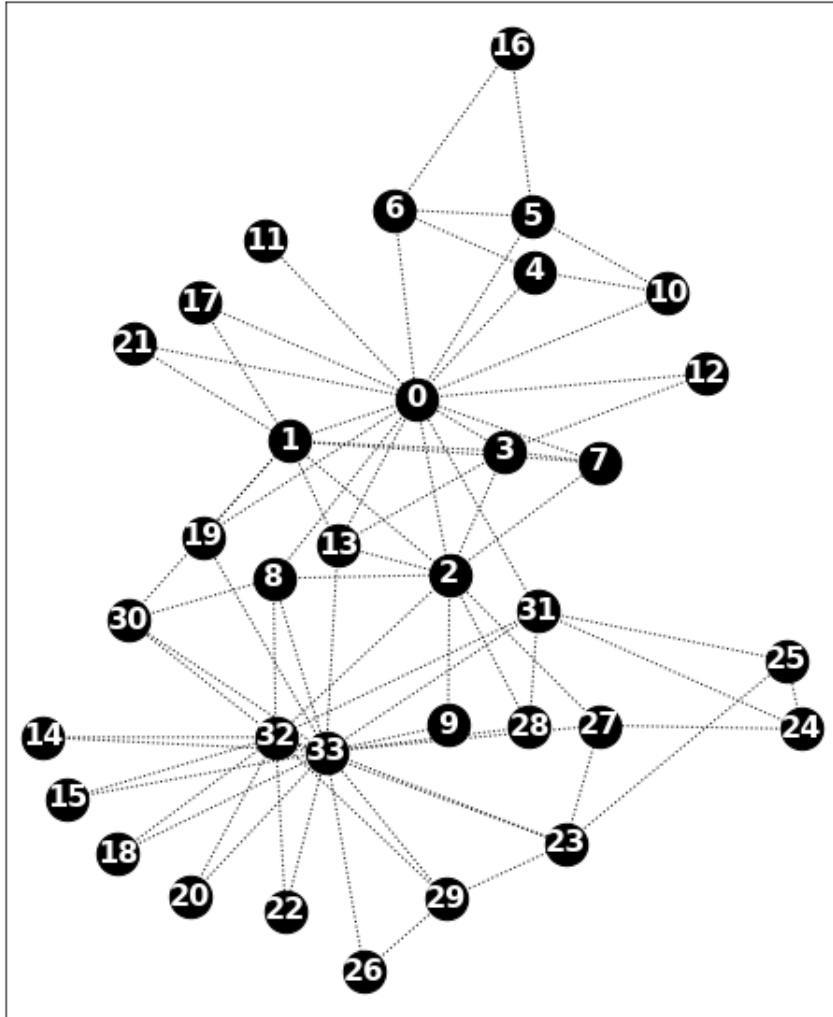


Figure 2: Zachary's Karate Club graph [8] has with  $N = 34$  vertices and  $M = 78$  edges and a single connected component.

Figure 3 shows the distribution of the number of walks and paths with specific lengths from node 0 to node 1 in the Zachary's Karate Club graph. Considering walks and paths between nodes 0 and 1, it is possible to note that:

- There is an edge connecting nodes 0 and 1, so,  $A_{01} = 1$ .
- As  $(A^1)_{01} = A_{01} = 1$ , there is only 1 walk from 0 to 1 with length 1.



- That walk is a path as well.
- The shortest path length between nodes 0 and 1 is 1:  $s'_{01} = 1$ .
- As  $(A^2)_{01} = 7$ , there are 7 walks from 0 to 1 with length 2.
- All those walks are paths as well.
- As  $(A^3)_{01} = 37$ , there are 37 walks from 0 to 1 with length 3.
- However, only 13 of those 37 walks are paths as well.
- The length of the longest path between nodes 0 and 1 is 18.

The number of paths is calculated using the Depth-limited search (DLS), setting the path length as the limit of the DLS. The longer the path length, the more significant the difference between the number of walks and paths between a pair of nodes. There are infinite walks between nodes 0 and 1, but precisely 8.854.467.719.776.520.000 ( $\approx 8E18$ ) walks with lengths up to 18. The number of paths between nodes 0 and 1 is 80.137 ( $\approx 8E4$ ).

Going beyond the shortest path, let us calculate the expected value for  $l$ , under the k-order approximation:

$$\mathbb{E}[l_{ij}] = \frac{\sum_{n=s}^{s+k} n(A^n)_{ij}}{\sum_{n=s}^{s+k} (A^n)_{ij}}. \quad (3)$$

The expected value for  $w_{01}$ , under the k=17-order approximation is, thus:

$$\mathbb{E}[l_{01}] = \frac{\sum_{n=1}^{18} n(A^n)_{01}}{\sum_{n=1}^{18} (A^n)_{01}}. \quad (4)$$

On the other hand, the expected value for  $l'$ , under the k-order approximation is:

$$\mathbb{E}[l'_{ij}] = \frac{\sum_{n=1}^{s'_{ij}+k} n f_P^{(ij)}(n)}{\sum_{n=1}^{s'_{ij}+k} f_P^{(ij)}(n)} \quad (5)$$

So, the expected value for  $l'_{01}$ , under the k=17-order approximation:

$$\mathbb{E}[l'_{01}] = \frac{\sum_{n=1}^{18} n f_P^{(ij)}(n)}{\sum_{n=1}^{18} f_P^{(ij)}(n)} \quad (6)$$

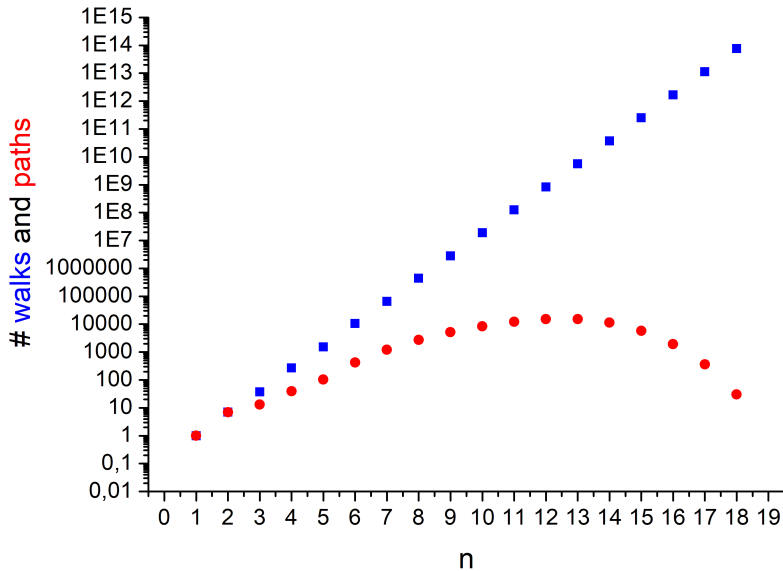


Figure 3: Number of walks (blue squares) and paths (red circles) from node 0 to node 1, in the Zachary Karate Club graph, for each path length.

Figure 4 shows our “delta measure”:  $\mathbb{E}[w_{01}] - sw_{01}$  (for walks),  $\mathbb{E}[l'_{01}] - sl'_{01}$  (for paths), for each value of  $k$ .

It is essential to highlight that difference, for the walks-case, grows indefinitely. However, in the case of the paths, it converges: it happens because when we allow longer paths, although the numerator increases (path length), that increment decreases - once, in a non-complete graph, the number of paths with a length close to the longest possible is smaller than the number of paths with intermediate lengths.

Considering all origins and destinations (all the nodes), we notice that the largest number of paths are between nodes 16 to 25, with 4319868 paths, going from length 4 to 23. This is interesting since node 16 appears in every case where the shortest path length is equal to the diameter of the network, is on the top of the mean length of paths, and has the highest mode. Node 16 seems the most unapproachable from the network by looking at these data.

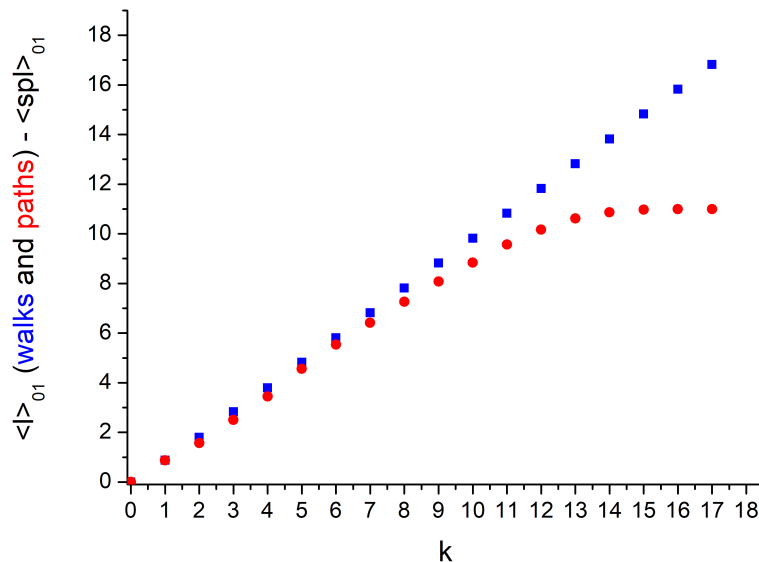


Figure 4:  $\mathbb{E}[w_{01}] - sw_{01}$  (blue squares) and  $\mathbb{E}[l'_{01}] - sl'_{01}$  (red circles), in the Zachary karate club graph, for different k-order approximations

To access the table containing the statistics of all paths click [here](#)<sup>3</sup>.

## 5. Conclusions

This short report presents an original idea about going “beyond the shortest path”. After presenting some fundamental concepts in graph theory, we presented an analytical solution for the problem of counting the number of possible paths between two nodes in complete graphs, and a depth-limited approach to get all possible paths between each pair of nodes in a general graph. Using the simple and well-known Zachary’s karate club graph, we showed the distribution of walks and path lengths.

The most important result is that we can go beyond the shortest path (facing an NP-hard problem), but (fortunately) we do not need to go so far: there is a convergence/saturation value for the path-length expected value -

<sup>3</sup>[https://github.com/gioguarnieri/Pesquisa\\_Doutorado/blob/master/all\\_paths\\_data.md](https://github.com/gioguarnieri/Pesquisa_Doutorado/blob/master/all_paths_data.md)

once, in a non-complete graph, the number of paths with a length close to the longest possible is smaller than the number of paths with intermediate lengths. The value of that control parameter ( $k$  - number of edges beyond the shortest path length) can be even smaller when considering penalties for longer paths.

In future work we plan to apply those ideas to a real-world problem, addressing urban mobility-related problems.

## References

- [1] A.-L. Barabási, M. Pósfai, Network science, Cambridge University Press, Cambridge, 2016. URL: <http://barabasi.com/networksciencebook/>.
- [2] A. Lima, R. Stanojevic, D. Papagiannaki, P. Rodriguez, M. C. González, Understanding individual routing behaviour, J. R. Soc. Interface 13 (2016) 20160021. URL: <https://royalsocietypublishing.org/doi/10.1098/rsif.2016.0021>.
- [3] E. Galbrun, K. Pelechrinis, E. Terzi, Urban navigation beyond shortest route: The case of safe paths, Information Systems 57 (2016) 160–171. URL: <https://www.sciencedirect.com/science/article/pii/S0306437915001854>. doi:<https://doi.org/10.1016/j.is.2015.10.005>.
- [4] L. R. Tomas, G. G. Soares, A. A. S. Jorge, J. F. Mendes, V. L. S. Freitas, L. B. L. Santos, Flood risk map from hydrological and mobility data: a case study in são paulo (brazil), Transactons in GIS - accepted for publication (2022). URL: <https://onlinelibrary.wiley.com/doi/10.1111/tgis.12962>.
- [5] E. Estrada, N. Hatano, Communicability in complex networks, Physical review. E, Statistical, nonlinear, and soft matter physics 77 (2008) 036111. URL: <https://arxiv.org/abs/0707.0756>.
- [6] N. Biggs, Algebraic graph theory, Cambridge university press, 1974.
- [7] B. Roberts, D. P. Kroese, Estimating the number of s-t paths in a graph, Journal of Graph Algorithms and Applications 11 (2007) 195–214. URL: <https://jgaa.info/accepted/2007/RobertsKroese2007.11.1.pdf>.

- [8] W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* 33 (1977) 452–473. URL: <https://www.jstor.org/stable/3629752>.
- [9] C. P. Herrero, Self-avoiding walks on scale-free networks, *Phys. Rev. E* 71 (2005) 016103. URL: <https://link.aps.org/doi/10.1103/PhysRevE.71.016103>. doi:10.1103/PhysRevE.71.016103.
- [10] M. Girvan, M. E. Newman, Community structure in social and biological networks, *Proceedings of the national academy of sciences* 99 (2002) 7821–7826.