Anonymous Routing Using Minimum Capacity Clustering

Maike Buchin
Ruhr University Bochum, Germany

Lukas Plätz
Ruhr University Bochum, Germany

Abstract

We present a framework which allows one to use an online routing service and get live updates without revealing the sensitive starting and ending points of one's route. For that, we obfuscate the starting and ending locations in minimum capacity clusters and reveal only the route between these clusters. We compare different anonymous clustering strategies on positions in the network with efficient approximations and analyse the impact of the anonymisation on the route. We experimentally evaluate the effect of the anonymisation scheme in real-world settings.

2012 ACM Subject Classification Security and privacy → Pseudonymity, anonymity and untraceability

Keywords and phrases Anonymity, approximation Algorithms, directed Networks, minimum capacity Clustering, Privacy

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.18

Category Short Paper

Supplementary Material *Software (Code and Data)*:
`https://gitlab.ruhr-uni-bochum.de/plaetlsv/giscience23/-/tree/FrechetAbstand`

Funding *Lukas Plätz*: The work was supported by the PhD School "SecHuman – Security for Humans in Cyberspace" by the federal state of NRW.

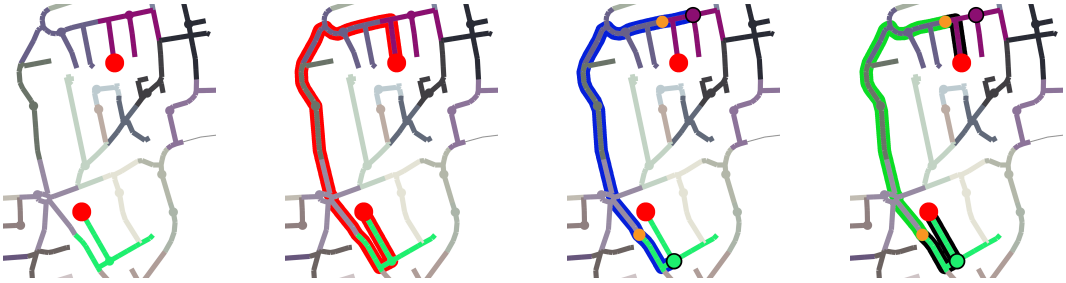
1 Introduction

Services often utilise personal routing data to offer traffic information, but it can be achieved using anonymised data. We can protect the sensitive part of our data by trading in a small amount of convenience. Anonymising the routing data enables us to use it in scientific research and redistribute it. The central idea is that the two endpoints of a route determine the shortest path. This means that the shortest paths only is helpful in re-identifying the starting and ending locations. By obfuscating these locations, we can protect privacy while sharing the remainder of the route for the public and personal benefit.

The concept of "k-anonymity" was first introduced by Sweeney [6]. It guarantees that each subject cannot be distinguished from less than $k-1$ other subjects. So finding a good k-anonymisation can be viewed as a clustering problem, with clusters requiring a minimum capacity of k. We k-anonymise locations in the network by clustering them.

To achieve k-anonymity, we adopt a concept from the routing literature introduced by Bast et al. [3]. In long-distance travel, routes around a starting location pass through a small set of nodes near the start. These nodes, known as *transit nodes*, reduce the search space and speed up the shortest path computation. We use a variation of this concept to anonymise the routing of a person. By computing transition nodes for each cluster (possibly depending on the cluster to be routed to) and routing through them, we can ensure that the path between these nodes remains the same for all starting and ending locations within the cluster.

Given clusters of at least k locations, one could find the transit nodes between two clusters. Utilising these transit nodes per cluster would keep the travel time the same. However, it would weaken anonymity, as a (specific) transit node (of several transit nodes of a cluster) may reveal in which part of the cluster the point lies. Therefore, we instead decided to actively lead the route through *check points* on the boundary of the clusters. This re-routing introduces some additional travel time, denoted as Δ. We will show that the maximum of Δ can be upper bounded with the radius of the clusters. Moreover, for a wide range of k, the mean value of Δ is insignificant in daily use. Additionally, since the check points are on the boundary of the clusters, most of the routes can be shared. See Figure 1 for an example. Routes within a cluster will not be anonymised within our framework as they are too short for gains through online services and would not use any check points.



(a) Locations. **(b)** Shortest Route. **(c)** Cluster Centers and check points. **(d)** Private and Public Parts.

Figure 1 shows an example of the anonymisation strategy. In the road network, streets from the same cluster have the same colour. In red are the locations and the shortest route. Next, we look at the centres shown as dots in colour for their cluster. With the shortest route in blue between them, get the check points in orange on the boundary of the clusters. Lastly, we compute the private in the black and the public in the green part of the anonymous route between the red locations.

We discuss four clustering strategies that differ in their setting and optimisation criterion while achieving a k-anonymous clustering. Later we will compare them on their impact on Δ. First is the *r-gather clustering* problem, which Aggarwal et al. [1] introduced. Here, the objective is to find clusters where each cluster contains a minimum of r points. The cluster's centre determines its radius, and the goal is to minimise the maximum radius across all clusters. They showed that this problem is NP-hard and gave a polynomial time algorithm to compute a 2-approximation, i.e. the radius is at most two times the optimal radius.

Armon [2] presents two variants of r-gathering that interest us. In the *r-gathering setting* – in contrast to r-gather clustering – the centres are chosen from a different set than the points to be clustered. The first one, called *min-max r-gathering*, minimises the maximum radius of the clusters. It is an NP-hard problem, and they presented a 3-approximation in the maximum radius in $O(n(m+r+\log n))$ time. The second strategy, *min-sum r-gathering*, minimises the sum over the distances to the centre. They showed that the problem is NP-hard and gave a $2r$ approximation in $O(n(m+r+\log n))$ time. With r-gather and min-max r-gathering, we can compute bounds to time lost by our anonymisation. However, min-sum r-gathering should lead to a better mean Δ than the other strategies.

Haunert et al. [5] introduced the *k-Anonymous Steiner Forest* for the problem of location clustering. Here they compute the optimal clustering where a cluster has to pay to the length of the street connecting them. They gave an algorithm with an approximation factor of 2 and a runtime of $O(nm)$. They applied their strategy to clustering places in a street network. However, their optimisation criterion does not align with ours, as the cost of the edge is only paid once. However, we will compare our location clustering with their result.

Brauer et al. recently presented a solution to a similar problem [4] which builds on the clustering strategy of Haunert et al. [5] to truncate trajectories. However, they only considered geometric clues[1]. With that, the trajectories leak information about the start and end points. The attacker model is heavily constrained because it cannot use the knowledge of the existing network. However, their strategy can be used to anonymise existing databases of trajectories, and they do not need the shortest paths.

We present a framework for k-anonymous online routing. We argue that under our assumptions (that all users use the shortest route), retrieving the start or ending from the route is impossible, even if the attacker knows the model of obfuscation, the clusters and the network. We bound the impact on the travel time by this framework and present a polynomial time algorithm to minimise the impacted travel time. We demonstrate the practicality of our framework with experimental results for German cities.

2 Anonymization Scheme

We will use the road network, given as a directed embedded graph $G = (V, E)$ and information on travel time $t(e)$ and population distribution $p(e)$ over the edges of the network first to compute a clustering on the edges. We then anonymise the shortest path between two points by calculating the shortest path between the cluster centres that encompass these points. We exclude the portion of the path within the clusters and replace it with the shortest route from the starting point to the path and from the path to the destination.

This setting brings two challenges with it. First, typically the vertices of a graph are clustered, whereas clustering of edges is rare. To use the point clustering techniques, we have to adapt our graph. For that, we use the directed line graph, which has a node for each edge in the directed graph and maintains the connectivity by introducing a directed edge for each pair of edges in the directed graph if the first edge ends at the start of the second edge.

Secondly, road networks are directed graphs which do not come with a canonical metric. We decided to use the length of the minimal cycle of a list of items as our distance function. We use this because it gives a symmetrical distance measure for a list of length two and satisfies the triangle inequality. Also, roundtrips are meaningful in our settings. We used items as a stand-in for vertices and edges. To make this distance measure a metric, we define cycle $[p, p]$ to have length 0. As we will primarily discuss minimal cycles, we use the list notation $[a, b, \dots]$ if we mean the shortest cycle using these objects in that order. The distance function d gives us the length of the minimal cycle.

For a cluster C in the clustering $\mathcal{C}$, we denote its centre as c_C. For two clusters C, C', we get the check points from the minimal cycle $[c_C, c_{C'}]$. The entry check point is the first node $i_C \in C$ on the minimal cycle, coming from $c_{C'}$. The exit check point is the last node $o_C \in C$ on the minimal cycle, coming from c_C. So for the shortest cycle $[s, e]$ with $s \in S$, $e \in E$ and $S, E \in \mathcal{C}$, this leads to the anonymized cycle $\mathcal{A}(s, e) := [i_s, s, o_s, i_e, e, o_e]$.

We define the radius $\mathrm{R}(C)$ of a cluster C as $\max_{p \in C} \mathrm{d}([c_C, p])$. Furthermore, $\mathcal{R} := \max_{C \in \mathcal{C}} \mathrm{R}(C)$ denotes the maximal radius of the clusters in a clustering.

Now we can define the function $\Delta(s, e)$ from the introduction as $\mathrm{d}(\mathcal{A}(s, e)) - \mathrm{d}([s, e])$.

▶ **Lemma 1.** *Given the maximum radius $\mathcal{R}$ of a clustering, $4\mathcal{R}$ bounds the maximum extra time Δ introduced by the anonymisation scheme $\mathcal{A}$.*

[1] i.e. the closest location and a wedge in the last direction of the trajectory

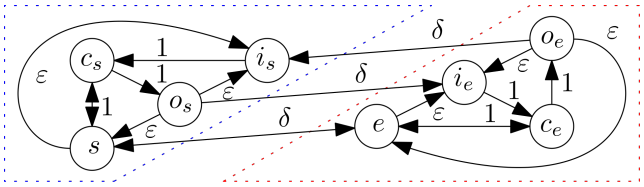


Figure 2 Example showing the tightness of the upper bound as stated in the lemma 1. If we pick c_s and c_e as the centres, we get a 4-gather with radius $2+\varepsilon$. The minimal cycle distance between s and e is 2δ. The anonymised cycle distance is $8+2\delta$. This gives us the tightness for ε approaching 0.

Proof. By definition, we have $\Delta := \mathrm{d}(\mathcal{A}(s,e)) - \mathrm{d}([s,e])$. When we insert the centres c_s and c_e into the anonymised cycle, we only make it longer but also can drop the check points as they lie on the shortest cycle between c_e and c_s.

$$\Delta \leq \mathrm{d}([i_s, c_s, s, c_s, o_s, i_e, c_e, e, c_e, o_e]) - \mathrm{d}([s,e]) = \mathrm{d}([c_s, s, c_s, c_e, e, c_e]) - \mathrm{d}([s,e])$$

If we now insert s and e between c_s and c_e we can split the long cycle into smaller ones,

$$\mathrm{d}([s, c_s, s, c_s, s, e, c_e, e, c_e, e]) = \mathrm{d}([c_s, s]) + \mathrm{d}([c_s, s]) + \mathrm{d}([s,e]) + \mathrm{d}([c_e, e]) + \mathrm{d}([c_e, e]).$$

The length of the smaller cycles within a cluster is bounded by $\mathcal{R}$. Thus, we get $\Delta \leq 4\mathcal{R}$. ◀

Remarkably, the upper bound $4\mathcal{R}$ is tight. Figure 2 shows an example for that. Also, Δ can be bounded by the actual radius of the starting and ending clusters.

We conclude that minimising the maximum radius of the clustering is a suitable proxy/substitute for anonymising with a small impact on travel time.

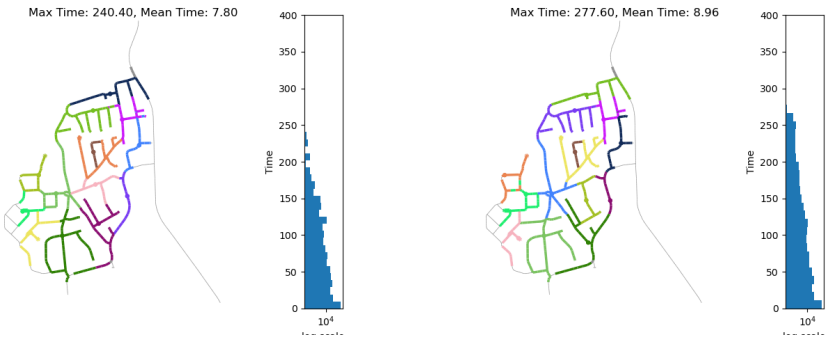
3 Experimental Results

We tested our anonymisation scheme in several cities in Germany. We used the data from OpenStreetMap for the network and the German census data to estimate the number of people living next to the streets. We import the street network with the travel time for the edges from OpenStreetMaps. The German census [7] from 2011 provided a 100 m times 100 m square grid of people living in each cell.[2] We distributed each square's population evenly on every curve in that square for a realistic distribution.[3]

To use the r-gather clustering, we used a line digraph of the network to switch the roles of edges and vertices. Because all clustering strategies assume that each point has equal weight, we used a multigraph with as many edges for a street as people.

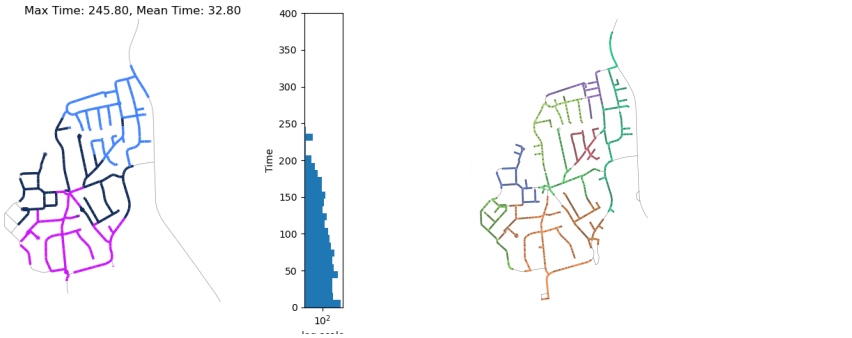
Our primary focus was to analyse Δ for the different strategies and r, as this bounds the detour induced by the anonymisation scheme. For that, we computed the shortest and anonymised path for every pair of vertices. With that, we calculated the values of Δ. We empirically found that the maximum Δ is often close to two times the maximum radius of the clusters. This could be explained by the fact that most edges of a street network are undirected, and in that case, two times the maximum radius is the upper bound. Nevertheless, there are instances where it gets close to the upper bound of 4 times the radius. However, we also see that the mean of the distribution is much closer to 0 than the maximum. Factors that play a role in this are the directedness and density of the network.

[2] The data was anonymised, so no individual or pair of people could be identified.
[3] We used networkx for processing, geopandas for geocoding, matplotlib for plotting, osmnx for import, and scipy to compute the distance matrix.



(a) min-max r-gathering.

(b) min-sum r-gathering.



(c) r-gather.

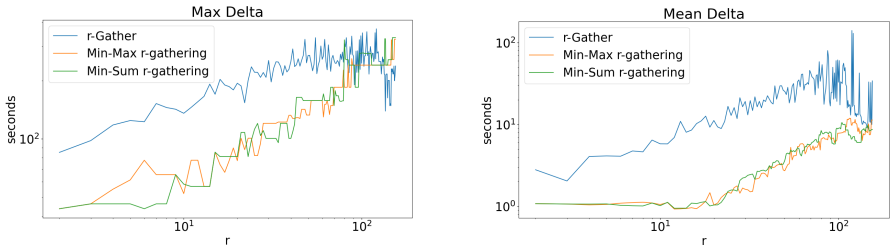
(d) Anoynmous Steiner Forest by Haunert et al. [5].

Figure 3 Shown are different clusterings with the minimum capacity of r=100 of Bonn Ückesdorf in Germany. In general, the mean Δ is far from the maximum and clumped around 0 (blue histography on the right side of each subfigure with a log scale). 3a shows the min-max r-gathering with the best maximum cluster size and mean time in seconds. 3b shows the min sum r-gathering. 3c shows the r-gather, which has a higher meantime because it only finds large clusters. 3d shows the clustering from Haunert et al. [5]. Here, the cluster of locations are retrieved from the buildings in OpenStreetMaps. As they grow trees until they are big enough, these clusters are connected.

We show the example of Bonn Ückesdorf, a small suburb. This allows us to compare our clusterings with the clustering of Haunert et al. [5]. Figure 3 shows the four clustering strategies. The clusters are randomly coloured, and we depict streets without inhabitants as thin grey lines. On the right of each subfigure are the histograms of the Δ of each route between different clusters. The r-gathering approaches lead to a significantly smaller mean Δ. The approximation algorithm for r-gather produces equal-sized clusters. The clusters in r-gather have the problem that they are not connected. The disconnection comes from the flow problem satisfying the minimum capacity. Here the edges can be arbitrarily distributed between the clusters when they have enough edges and their influence radius overlap.

In Figure 4, we compare Δ for different r and strategies. In this setting, r-gather gives bad results for small r but catches up for larger r. It also seems to be less stable as the other. Surprisingly, the r-gatherings stay close to each other in max and mean Δ.

All clustering strategies had runtimes from a few seconds to minutes on a city scale. The computations were done with a regular desktop pc and programmed in Python. Anonymising a route does not require much more time than a normal routing query. We need to look up the centres of the cluster of our endpoints and query the route between the centres. Finding the boundary point of the cluster on the route is straightforward, and routing to these



(a) Max Δ. **(b)** Mean Δ.

Figure 4 The graph on the left depicts the max Δ and on the right mean Δ for different r for Bonn Ückesdorf. Both axes use a log scale.

checkpoints only needs a short-distance route query. Furthermore, routing between different cities can be done by clustering every city individually. Therefore, it is only necessary that the different clustering do not overlap, as that would break the k-anonymity.

4 Conclusion

We have developed a framework for anonymous routing that has minimal impact on travel time. We explored four minimum capacity clustering strategies and their effects on travel time. Our analysis revealed that the r-gather and min-max r-gathering strategies provided upper bounds for the maximum extra travel time. We also examined the min-sum r-gathering strategy and found that both r-gathering cluster strategies resulted in shorter mean extra travel times. In the future, we plan to use a weighted version of the clustering strategies to aggregate edges in the graph, which might lead to faster computations. Additionally, we believe that a finer subdivision of the streets could reduce extra travel time even further.

References

1 Gagan Aggarwal, Rina Panigrahy, Tomás Feder, Dilys Thomas, Krishnaram Kenthapadi, Samir Khuller, and An Zhu. Achieving anonymity via clustering. *ACM Trans. Algorithms*, 6(3), 2010. doi:10.1145/1798596.1798602.

2 Amitai Armon. On min–max r-gatherings. *Theoretical Computer Science*, 412(7):573–582, 2011. Selected papers from WAOA 2007: Fifth Workshop on Approximation and Online Algorithms. doi:10.1016/j.tcs.2010.04.040.

3 Holger Bast, Stefan Funke, Peter Sanders, and Dominik Schultes. Fast routing in road networks with transit nodes. *Science*, 316(5824):566–566, 2007. doi:10.1126/science.1137521.

4 Anna Brauer, Ville Mäkinen, Axel Forsch, Juha Oksanen, and Jan-Henrik Haunert. My home is my secret: concealing sensitive locations by context-aware trajectory truncation. *International Journal of Geographical Information Science*, 36(12):2496–2524, 2022. doi:10.1080/13658816.2022.2081694.

5 Jan-Henrik Haunert, Daniel Schmidt, and Melanie Schmidt. Anonymization via clustering of locations in road networks. In *GIScience 2021 Short Paper Proceedings*. UC Santa Barbara: Center for Spatial Studies., 2021. doi:10.25436/E2CC7P.

6 Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002. doi:10.1142/S0218488502001648.

7 Statistische Ämter des Bundes und der Länder. Bevölkerung im 100 meter-gitter, 2018. URL: https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html?nn=559100.