Journal of Information and Data Management, 2024, 15:1, doi: 10.5753/jidm.2024.3328

This work is licensed under a Creative Commons Attribution 4.0 International License.

Using Non-Local Connections to Augment Knowledge and Efficiency in Multiagent Reinforcement Learning: an Application to Route Choice

Ana L. C. Bazzan [**Universidade Federal do Rio Grande do Sul** | ***bazzan@inf.ufrgs.br***]
H. U. Gobbi [**Universidade Federal do Rio Grande do Sul** | ***hugobbi@inf.ufrgs.br***]
G. D. dos Santos [**Universidade Federal do Rio Grande do Sul** | ***gdsantos@inf.ufrgs.br***]

Institute of Informatics, UFRGS, Caixa Postal 15064, Porto Alegre, 90540-090, Brazil.

Received: 14 April 2023 • **Published:** 29 February 2024

Providing timely information to drivers is proving valuable in urban mobility applications. There has been several attempts to tackle this question, from transportation engineering, as well as from computer science points of view. In this paper we use reinforcement learning to let driver agents learn how to select a route. In previous works, vehicles and the road infrastructure exchange information to allow drivers to make better informed decisions. In the present paper, we provide extensions in two directions. First, we use non-local information to augment the knowledge that some elements of the infrastructure have. By non-local we mean information that are not in the immediate neighborhood. This is done by constructing a graph in which the elements of the infrastructure are connected according to a similarity measure regarding patterns. Patterns here relate to a set of different attributes: we consider not only travel time, but also include emission of gases. The second extension refers to the environment: the road network now contains signalized intersections. Our results show that using augmented information leads to more efficiency. In particular, we measure travel time and emission of CO along time, and show that the agents learn to use routes that reduce both these measures and, when non-local information is used, the learning task is accelerated.

Keywords: multiagent reinforcement learning, non-local information, urban mobility

1 Introduction

Congestion in urban traffic networks poses challenges to many developed and developing countries economies, as well as to the environment. In this context, reducing emissions and lost hours in traffic is among the top priorities of our societies. Conventional traffic management solutions may have reached their limits, as the available methods and tools are not flexible enough, or were not developed having in mind traffic patterns that are arising in mega-cities, or due to new transportation modes, mobility-on-demand, etc. Moreover, those solutions do not necessarily exploit new technologies, such as communication systems, including vehicle to infrastructure (V2I) communication. In fact, although Intelligent Transportation Systems (ITS) have received a lot of attention in the past decades, only recently has this area focused on the avenues opened by fast communication and vehicular networks. By effectively using communication-based approaches, congestion and, consequently, emissions and lost hours can be reduced.

New technologies can be employed in several ways. In this paper we concentrate on V2I as a tool to improve drivers' decisions on how to travel from A to B (pointers on other research directions appear in Section 3). Under this particular perspective, while the current pattern is that each individual driver selects a route based on his/her own experience, this is changing with the increasing penetration of new technologies that allow information exchange. Examples of these technologies are not only based on broadcast (e.g., GPS or cellphone information) but also a two-way communication channel, where drivers provide and receive traffic information.

Key here is that these technologies change the paradigm. While currently many traffic management systems are based on a central authority in charge of assigning routes for drivers, or at least providing information (e.g., Waze, Google apps, etc.) for them to decide, communication among vehicles, between vehicles and the road infrastructure, or even among elements of the infrastructure are transformative. In fact, roads are already undergoing the same changes that are seen in the economy, as well as in the society, namely, a decentralization of the decision-making process and, not least, the prominence of several players, such as IT, big tech, and, most importantly, the citizen her/himself. At the end of Section 3 we point to researchers that have surveyed this topic.

Right now we are experiencing a situation in which these technologies and platforms are trying to establish themselves, and are still focusing very much on an agenda that is decades old, namely saving travel time. However, more and more, other aspects are being considered when formulating public policies related to urban mobility. One of these aspects concerns the environment, since stop-and-go traffic may cause more emissions. Therefore, while in the past traffic engineering has focused mostly on reducing travel time, emissions are often being taken into account as well.

As discussed in the next section, there are many ways to help improve how the demand (persons, trips, goods) can efficiently use the existing supply (road infrastructure). Most of them rely on centralized approaches though. One way to mitigate this is by letting drivers experience and decide in a decentralized way by means of reinforcement learning (RL), where, given the collective nature of this process, we in fact should use multi-agent reinforcement learning (MARL), aiming at investigating how drivers (or agents) choose their preferable route based on their own learning experiences.

In a previous work Santos and Bazzan [2021]; Santos *et al.* [2021], we have connected MARL to V2I communication, in order to investigate how it could augment the information drivers use in their route choices. Later, we have considered multiobjective RL Santos and Bazzan [2022], where drivers have two objectives: reduce not only travel time, but also emission of carbon monoxide.

In a further work, we connected that line of research to the use of multiple attributes, while also considering V2I Bazzan *et al.* [2022]. In this case, a third element was added, namely, information about non-local interactions. By non-local we mean information that is gathered not in the vicinity of a given road infrastructure element like an intersection or a road segment. This is done by constructing a graph in which some elements of the infrastructure are connected according to a similarity measure regarding patterns. Patterns here relate to a set of different attributes: we consider not only travel time, but also include emission of gases. These aspects are further detailed in Section 4.

In the present paper, we extend the experiments to a scenario that also considers traffic signal controllers. This is important because such controllers are present in urban networks and thus, need to be considered. We stress that, in this work, the controllers are not learning agents.

Section 4 details the methodology. Here, it suffices to say that we use of a relationship graph where sections of the traffic network (links) that have similar values for those attributes are connected. An exchange of information about travel time and emissions occurs, so that the infrastructure has augmented information, which is then passed to vehicles to allow them to make more informed decisions. Section 5 reports experiments that show the efficiency of our approach, where informed drivers are able to make decisions that, despite aiming at reducing their travel times, also reduce emissions. We test the approach on a network considering two cases (with and without traffic signal controllers) and also two RL algorithms (the classical QL) and our approach.

2 Background

In this section, we cover some key concepts that underlie our approach. For more details on conventional traffic assignment, we refer the reader to Chapter 10 in Ortúzar and Willumsen [2011]. For our purposes it suffices to mention that conventional approaches are centralized. Instead, this section focuses on MARL-based approaches that allow a decentralized decision-making for route choice.

Next, we give a brief introduction to RL and MARL.

Reinforcement learning (RL) is a machine learning method, in which agents learn how to map a given state to a given action, by means of a value function. RL can be modeled as a Markov decision process (MDP), where there is a set of states S, a set of actions A, a reward function $R : S \times A \rightarrow \mathbb{R}$, and a probabilistic state transition function $T(s, a, s') \rightarrow [0, 1]$, where $s \in S$ is a state the agent is currently in, $a \in A$ is the action the agent takes, and $s' \in S$ is a state the agent might end up, taking action a in state s. The tuple (s, a, s', r) represents that an agent was in state s, then took action a, ended up in state s' and received a reward r. The key idea of RL is to find an optimal policy π^*, which maps states to actions in a way that maximizes future rewards.

RL methods fall within two main categories: model-based and model-free. While in the model-based approaches the reward function and the state transition are known, in the model-free case, the agents learn R and T by interacting with an environment. One method that is frequently used in many applications is Q-Learning (QL). In QL, the agent keeps a table of Q-values that estimate how good it is for it to take an action a in state s; thus a Q-value $Q(s, a)$ holds the maximum discounted value of going from state s, taking an action a and keep going through an optimal policy. In each learning episode, the agents update their Q-values as in Equation 1, where α and γ are the learning rate and the discounting factor for future values, respectively.

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma max_a[Q(s',a') - Q(s,a)]) \quad (1)$$

In RL tasks, it is also important to define how the agent selects actions, while also exploring the environment. A common action selection strategy is the ϵ-greedy, in which the agent chooses to follow the optimal values with a probability $1 - \epsilon$, and takes a random action with a probability ϵ.

3 Related Work

Traffic assignment problem is not a new problem; there have been several works that aim at solving it. Besides conventional methods, (see Chapter 10 in Ortúzar and Willumsen [2011]), which mostly deal with planning (long term) tasks, and are centralized, RL is turning popular. In this front, methods usually fall into two categories: a traditional (state-based) RL method, and a stateless one. In the latter, each agent d actually is in only one state (its origin location), where it selects a route to travel. A route is defined as a sequence of links that take d from its origin to its destination, thus no en-route decision is necessary. Works in this category are Ramos and Grunitzki [2015]; Ramos *et al.* [2017]; Zhou *et al.* [2020]. Tumer *et al.* [2008] adds a reward shaping component (difference utilities) to QL, aiming at aligning the UE to a socially efficient solution. Multiobjective, stateless RL was employed in Huanca-Anquise [2021]; Huanca-Anquise *et al.* [2023], where agents aim at optimize travel time and a second objective, toll.

In the state-based front, each agent d makes decisions at each junction, regarding which link to select next, so that it will eventually reach its destination. In Bazzan and Grunitzki [2016] this is used to allow agents to learn how to build routes. However, they use a macroscopic perspective by means of cost functions that compute the abstract travel time.

In the present paper, the actual travel time is computed by means of a microscopic simulator (see Section 5).

As aforementioned, our approach includes V2I communication, as this kind of new technologies may lead agents to benefit from sharing their experiences, thus reducing exploration. The use of communication in transportation systems, with some sort of communication among drivers, has also been studied previously by us (Bazzan *et al.* [2006]; Grunitzki and Bazzan [2016]) as well by others (Auld *et al.* [2019]). In a different perspective, works like Yu *et al.* [2020] evaluate the impact of incomplete information sharing.

Also, in order to exploit the potential of V2I communication, in previous works Santos and Bazzan [2020, 2021]; Santos *et al.* [2021], we have connected it to MARL, in order to investigate how V2I communication could benefit drivers use in their route choices. In these works, the infrastructure is able to communicate with the vehicles, both collecting information about their most recent travel times (on given links), as well as providing them with information that was collected from other vehicles. However, links in the infrastructure only exchange information if they are connected by a junction, i.e., only local information is considered.

We have also investigated what happens when multiple objectives are considered Santos and Bazzan [2022]. In that work, drivers have two objectives: reduce travel time and emission of carbon monoxide. However, that work does not address any kind of communication. A first attempt in this direction was investigated in Bazzan *et al.* [2022], where elements of the traffic infrastructure that have similar patterns form a graph. This graph is then used to allow communication among the various elements. Patterns are formed when these elements have similar values regarding a set of attributes that include not only travel time but also emission of gases. In the present paper, we extend this approach and consider a scenario that has traffic signals.

The value of V2I communication has started to receive attention also in the traffic engineering community. The reader is referred to Mahmassani [2016] (focusing on how autonomous vehicles and connected vehicles are expected to increase the throughput of highway facilities, as well as improve the stability of the traffic stream), and Maimaris and Papageorgiou [2016] (applications).

Finally, the method we employed here grounds on graph-based methods. Since few of them do tackle communication, due to lack of space we cannot cover that literature. We refer the reader to a survey: Cui *et al.* [2022].

4 Methodology

4.1 Terminology: Road Network and Virtual Graph

We deal with two sorts of graphs. First, a road network is a (planar) graph $G = (J, L)$, where J is the set of junctions (intersections), and L is the set of links. We use the term link, since it is more commonly used in traffic engineering (and then reserve the term edge for the second graph, as described next).

For example, in Fig. 4 links `gneE54` and `gneE55` are both connected to the junction that appears in the center of the figure. As for the second graph, once our approach relies on non-local information, such graph is the one that connects two links $l_1 \in L$ and $l_2 \in L$ which are *not necessarily physically close* (as, e.g., `gneE49` and `gneE45` in Fig. 4), but that have similar patterns. We call this a virtual graph denoted by $VG = (L, E)$, where L is the set of links (note that now they act as vertices in VG), and E is the set of edges that connect two links that have similar patterns, as described next.

In order to define when two links are to be connected in VG, historical information is collected for a network G. This information refers to several attributes: travel time, fuel consumption and several kinds of gas emissions[1], per link, per time interval. We aggregate such information using a time window w_h and normalize the values of all attributes between zero and one. Then, the values of such attributes for each two pairs of links are compared. If two links l_1 and l_2 have the same values for all attributes (given a tolerance value, i.e. $\pm\delta_a$), then an edge connecting l_1 and l_2 is inserted in VG. Fig. 1 shows an instance of such a virtual graph, whereas Fig. 2 depicts a zoom of that graph, where some relationships among similar links can be better seen. The labels of the vertices are formed by the link ID plus the time interval in which their values were found to be similar.

4.2 How Communication Works

Next we briefly explain how the communication is performed by the elements of the road network G. We assume that every junction $j \in J$ and every link $l \in L$ is equipped with a communication device (henceforth, CommDev) that is able to send and receive messages among themselves, as well as to and from nearby vehicles (i.e., those in the link where a given CommDev is located). For instance, in Fig. 3, the red vehicle informs the corresponding CommDev about its rewards in terms of travel time and other attributes, once it has travelled that particular link. Similarly, the corresponding CommDev informs the green vehicle the expected travel time in the links ahead, so that the green vehicle is able to decide which link to take, once it reaches the next junction (i.e., the next decision state).

A junction CommDev collects information from its incoming links, defined in the physical road network G. Additionally, given that these incoming links may have virtual neighbors in the virtual graph VG, information about their virtual neighbors are also passed to each CommDev at the links and, from these to the junction CommDev. Once a CommDev at a junction j has collected such information, it updates a table in which the last 30 entries are kept in a FIFO way, for each attribute. This value was used in Santos and Bazzan [2020] for the same scenario. Moreover, in the present paper, a CommDev stores information also about travel time and CO emission.

Each CommDev then communicates to the nearby driver agents an aggregation of those values kept in the tables[2], i.e.,

[1] We collect CO, CO2, HC (hydrocarbon), PMx (particulate matter), and NOx.

[2] The information about CO is not used by the agent, given that QL only optimizes for one objective – in this case travel time – but, as discussed in

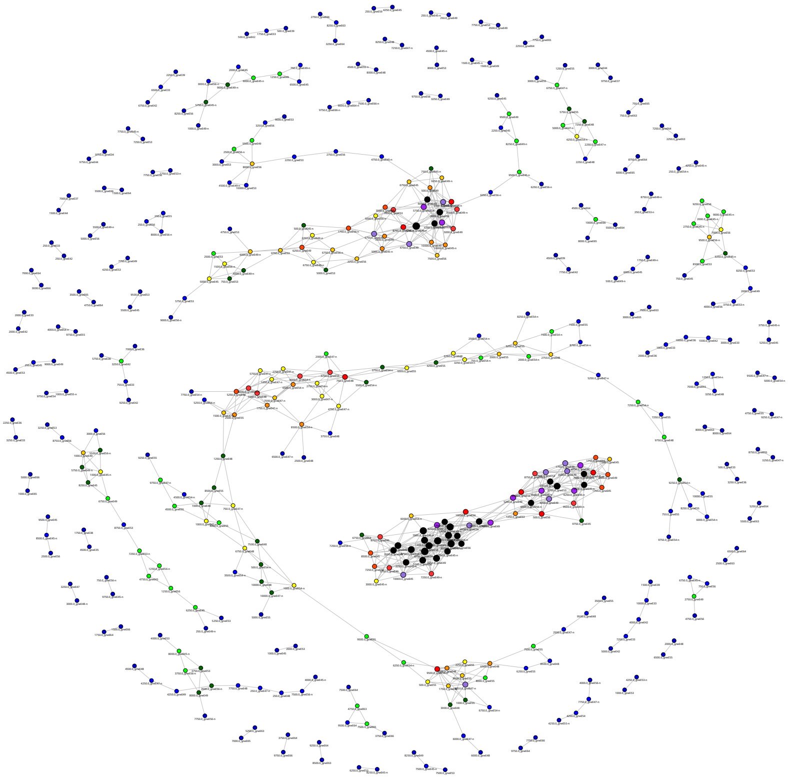

Figure 1. Instance of a virtual graph VG (Full VG).

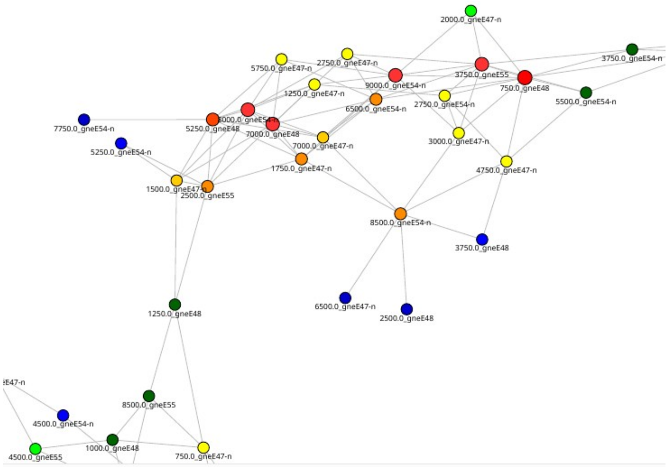

Figure 2. Zoom of a small part of the full VG.

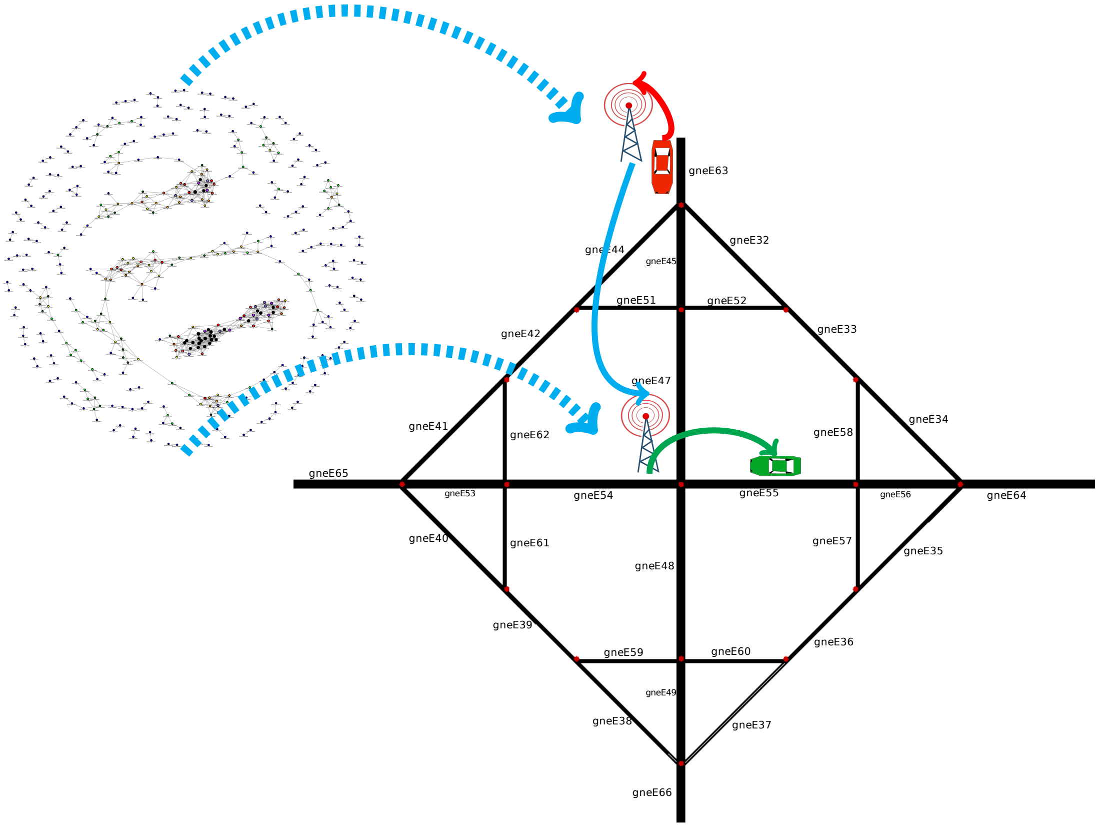

Figure 3. Road network is superimposed with a scheme of the communication among various elements of the road infrastructure, as well as the VG shown in Fig. 1.

potential rewards that the agent may obtain if selecting each action in that particular state. The agent then perceives this information as expected rewards for the actions available to it.

4.3 MDP Formulation

As mentioned in Section 2, a RL learning task is formulated by an MDP. In our case, given a network G, the set of states is defined by J. Two particular states are the origin and the destination of an agent. They define the so-called origin-destination (OD) matrix or set of OD pairs, which basically shows how many trips start and end in each location of the network. A_d^j denotes the set of actions available to agent d at j, which are the links that leave j. Since we deal with a maximization task, each reward is given by the negative of the travel time experienced by d at link l. This value is provided by the microscopic simulator we use (see next section).

Note that in the standard QL algorithm, the agents update their Q-values based on the feedback from the action they have just taken. However, in our case agents also update their Q-values based on the expected rewards received by each CommDev. This means that every time they reach a junction, they also update their Q-values with the information provided by the CommDevs.

We also remind that we deal with a commuting scenario, where each agent performs day-to-day experiments in order to learn how to travel in the network G to go from its origin to its destination.

5 Experiments, Results, and Analysis

5.1 Scenario

Simulations were performed using a microscopic simulator called SUMO Lopez *et al.* [2018], whose API was used to allow vehicle agents to interact with the simulator during simulation time.

The network used is shown in Fig. 4, where the main links are two-way. Note that this figure depicts only six traffic signal controllers, for sake of clarity. However, the network is basically symmetric so that the other three quadrants have further signalized junctions. Fig. 5 then shows a zoom of the network, showing details of two signalized junctions (at the left side of the central horizontal arterial).

We employ this network in two variants. The first one is as in Bazzan *et al.* [2022], i.e., without the traffic signal controllers. This means that we use SUMO's allway-stop rule to decide which vehicles have right to go. This leads to all vehicles having to stop before crossing the junction. In the second variant, this rule is replaced by a signalized junction, meaning that the right to cross is decided by a signal; we use SUMO's actuated controller that gives priority to the lanes according to an actuated mechanism[3].

Section 6, this will be addressed in a future work, in a similar way as in Santos and Bazzan [2022].

[3] See: `https://sumo.dlr.de/docs/Simulation/Traffic_`

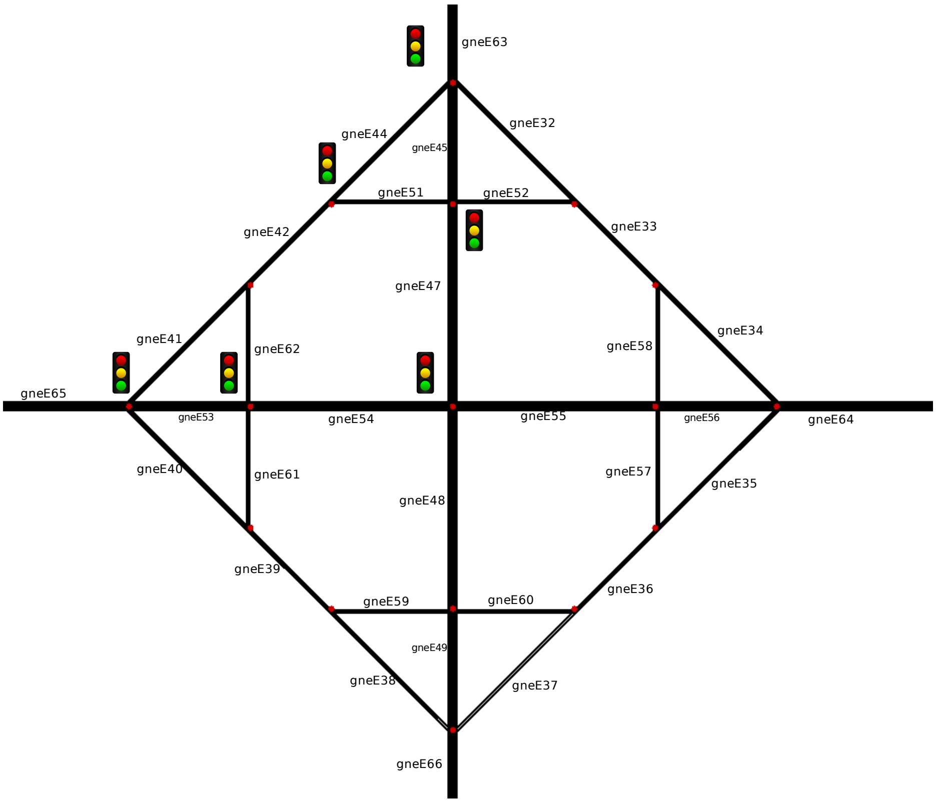

Figure 4. Road network used in the experiments (labels for some links that run in an opposite direction are omitted; they are similarly labelled: `gneE66-n` for example.). This figure also depicts some of the junctions that are signalized (for sake of clarity, only top left quadrant is depicted; the others are similar).

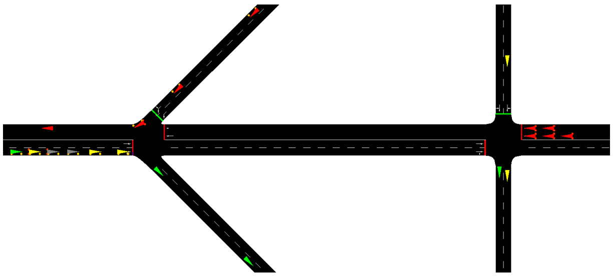

Figure 5. Zoom of the road network showing only two junctions (central horizontal arterial at the left side); the colors of the vehicles refer to their destinations.

We now explain the demand side, i.e., the trips that use the network (with or withouth traffic signals). Trips originate in each of the four most external links (`gneE63`, `gneE64`, `gneE65`, and `gneE66`), and have the other three of these links as destination (as, e.g., `gneE63` to `gneE64`, `gneE65`, and `gneE66`), thus defining 12 OD pairs, with 400 trips each. This demand was set to maintain the network populated at around 30% of its maximum capacity, (given that a vehicle occupies $5m$), which is considered a high occupation.

5.2 Model Parameters

For the various parameters of the model, we have used the same values as in Santos *et al.* [2021]. This way, we have set learning rate $\alpha = 0.5$, the discount factor $\gamma = 0.9$, and $\epsilon = 0.05$. These values guarantee that the future rewards have a considerable amount of influence in the agent's current choice, since γ has a high value. Other parameters take these values: $w_d = 250$ and $\delta_a = 0.005$.

5.3 Results and Discussion

To measure the performance, we collect travel time and CO emission, over all links $l \in L$ of the network. Given the probabilistic nature of the process, 30 runs were performed.

Plots ahead, in which shadows account for the deviations over the runs, show a comparison between the cases we deal with. First, we show the results when no learning is used, both for travel time (Fig. 6) and CO emission (Fig. 10). Figures 7 and 11 then also show these two quantities, now for the case when QL is used. The case in which QL is combined with the virtual graph VG has performance as in Fig. 8 (travel time) and Fig. 12 (CO). Finally, the case in which the network with signalized junctions is used as infrastructure is shown in figures 9 and 13.

In all cases, we note that it takes some time for all vehicles to be loaded, hence the initial increase that appears in all plots. As aforementioned, oscillations are shown in all plots as shadows. They happen either due to agents exploring their route choices, or due to SUMO's route assignment (see ahead).

We start by discussing the plots that refer to how travel time changes along time: Fig. 6 – Fig. 9. In the former, agents do not learn. However, SUMO has a mechanism that computes a route for each vehicle that is departing, based on the current occupation of the links in the network. This can be seen as a kind of optimization, but one that is performed

`Lights.html#type_actuated`

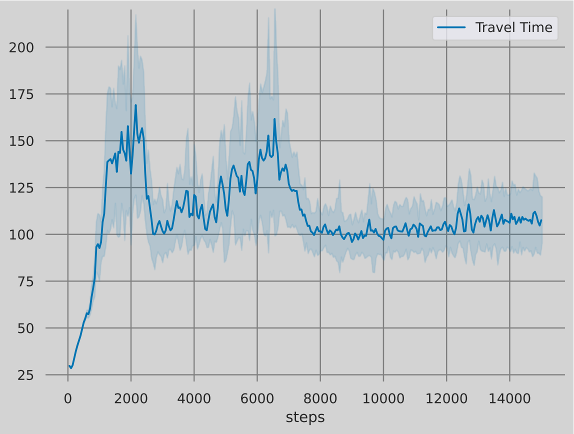

Figure 6. No learning: Travel time (average over all links).

centrally by the simulator. The travel time starts to stabilize around step 8,000, roughly at 100 steps or seconds (per link).

Next we compare: (i) travel time when agents do not learn versus when they use QL (Fig. 7 is lower than that in Fig. 6); (ii) when QL is used with and without the virtual graph (Fig. 6 Fig. 8).

In the former case, when agents use QL, they have to experiment in the beginning, until they converge to better decisions (about choice of links). This leads to lesser travel times: at the end of the simulation the travel time shown in Fig. 7 is lower than that in Fig. 6 by 25%.

When the virtual graph is used (Fig. 8), the convergence to a lower travel time happens earlier in comparison to the case in which only QL is used (Fig. 6). This happen due to the fact that links with similar patterns exchange information that helps each CommDev to better inform drivers about which links to select. Also, note that there are less deviations (blue shadow). In short, although drivers do converge to similar travel times ($\approx$ 705), this time is reached earlier when the virtual graph is used.

In the case with traffic signal controllers, Fig. 9, the pattern is different from the aforementioned cases. First, the use of traffic signal also leads to a stop-and-go behavior, due to vehicles having to stop for some seconds in each signalized junction; this pattern corresponds to the behavior each driver experiences in the real-world. Second, the initial behavior is similar as the drivers are being loaded. Then, driver agents do experimentation until roughly time step 4,000 when the use of the approach leads them to find their best routes. Note that the maximum value for travel time is now reduced to less than 100 steps. Further, the choices of routes converge to travel times in the same order as when the VG is used, i.e., $\approx$ 70. Also noticeable, the agents seem to find their ways without changing their choices as much as in the previous cases; this can be seen by the fact that there are less sharp peaks in this plot. We credit this behavior to the fact that the use of traffic signals heavily determine the choices drivers can make; this will be tested in further experiments.

As for the emission of CO, we recall that, when using QL, drivers only optimize for travel time, as QL does not handle more than one reward value, except if they are somehow combined in a function. Despite this, the use of the virtual graph (that corresponds to similarities among links using several attributes, including CO), also leads to reduction of CO emission. This can be seen by comparing Fig. 10, Fig. 11,

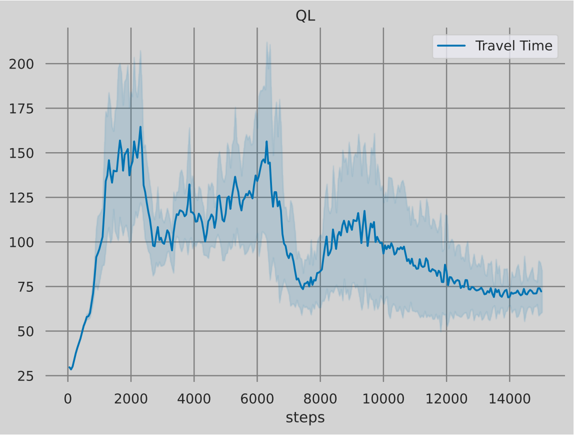

Figure 7. QL, no virtual graph: Travel time (average over all links).

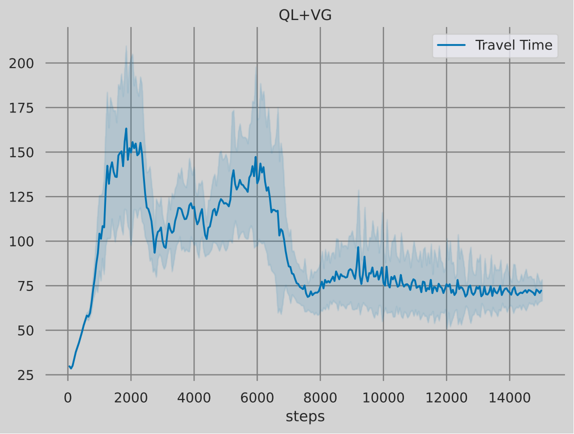

Figure 8. QL plus virtual graph: Travel time (average over all links).

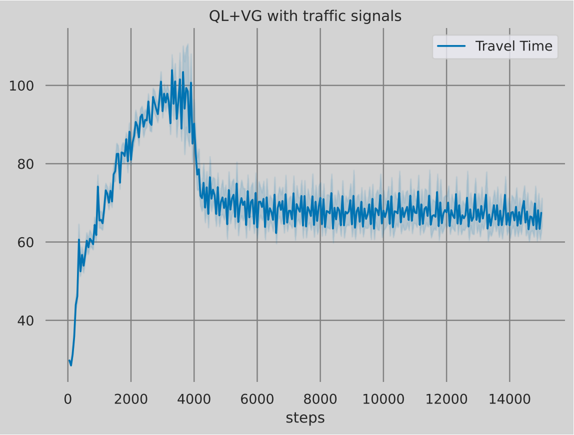

Figure 9. QL plus virtual graph with traffic signals: Travel time (average over all links).

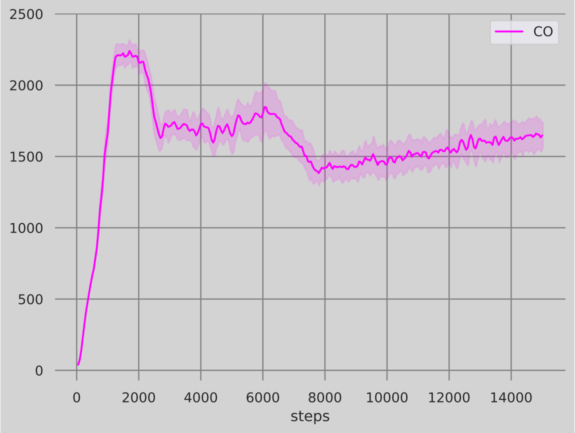

Figure 10. No learning: CO emission in mg/s (average over all links).

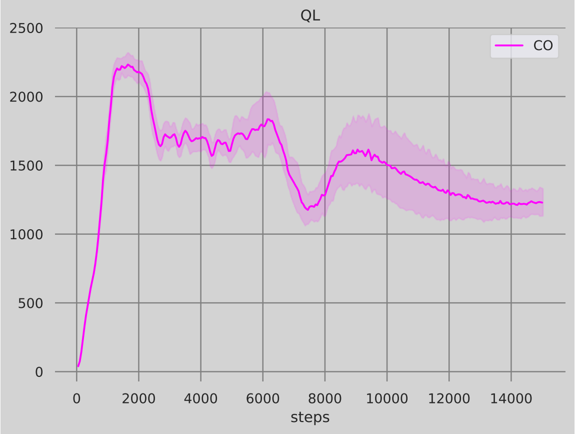

Figure 11. QL, no virtual graph: CO emission in mg/s (average over all links).

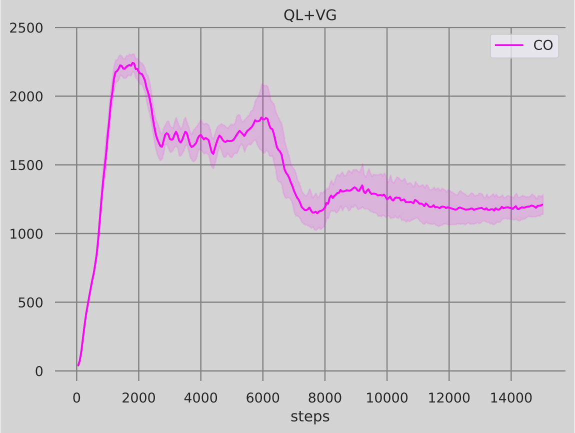

Figure 12. QL plus virtual graph: CO emission in mg/s (average over all links).

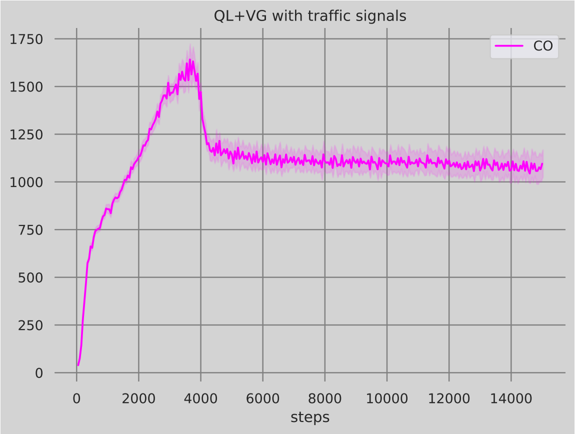

Figure 13. QL plus virtual graph with traffic signals: CO emission in mg/s (average over all links).

and Fig. 12. When learning is not used, the emission of CO converges to above 1500 mg/s. This is reduced to ≈ 1250 when QL is used. Comparing QL with and without a virtual graph, as with travel time, the convergence is achieved earlier in the latter (around time step 10,000).

The case depicted in Fig. 13 can be explained by the same behavior discussed previously in what regards travel time when signalized junctions are part of the scenario, namely, that they seem to influence the drivers choices heavily, thus leading to less experimentation and a more stable pattern. Nevertheless, as for travel time, the choice converge to emissions in the same order as when the virtual graph is used.

6 Concluding Remarks

The use of new communication technologies in urban mobility is turning more and more important. MARL is an attractive method for route choice, as it mimics the way drivers perform experimentation in their daily commuting.

The present paper discussed experiments using a method that combines MARL with V2I communication to allow the road infrastructure to collect and use non-local information, and form a virtual neighborhood, where links that have similar patterns regarding attributes such as travel time and emission of gases are virtual neighbors. Such augmented vision is then passed to vehicles for their decision-making about which link to follow next.

Specifically, experiments considered a network without signalized junctions, as in Bazzan *et al*. [2022], as well as a case with signal controllers. The results for the former showed that the use of a virtual graph improves the efficiency of the learning process. The case of signalized junctions has shown a different pattern in regard to the metrics used (travel time and emission of CO). It seems that the time stopped at the junctions has an influence in the route choices, leading drivers to converge to routes with less experimentation. This point remains to be checked. For this, we propose a change in the way the trips use the network. Recall that the trips originate in each of the four most external links and have the other three of these links as destination. This causes the trips to be short. Thus, to better test the aforementioned hypothesis, we plan to change the demand so that trips take longer so that the effect of the signals can be reduced. Another line of investigation refers to the use of a multiobjective RL approach. Recall that the reward of the drivers only considers travel time (even if, as stressed, the virtual graph is constructed using more attributes). Thus, we plan to reformulate the problem, so that rewards are expressed as a vector, in which not only travel time is used, but also further attributes, as in Santos and Bazzan [2022]. Note that this is different from the formulation in which reward is expressed as some kind of linear or non-linear function, in which case classical QL could be used without modification. The authors in that paper have employed an extension of QL, namely Pareto-QL Van Moffaert and Nowé [2014].

Acknowledgements

The authors are grateful for the suggestions made by the reviewers of KDMiLe 2022.

Funding

This research was funded by the following funding agencies: FAPESP and MCTI/CGI (grant number 2020/05165-1); CNPq (grant no. 304932/2021-3 for Ana Bazzan); and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil) - Finance Code 001.

Authors' Contributions

Ana Bazzan contributed to the main idea (on the virtual network), to the conception of this study and the writting of the manuscript. Henrique U. Gobbi has modified the existing code (partially developed by Guilherme D. dos Santos in a previous work), performed the experiments, analyzed the results (together with Ana Bazzan). All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

Auld, J., Verbas, O., and Stinson, M. (2019). Agent-based dynamic traffic assignment with information mixing. *Procedia Computer Science*, 151:864–869.

Bazzan, A. L., Gobbi, H. U., and dos Santos, G. D. (2022). More knowledge, more efficiency: Using non-local information on multiple traffic attributes. In *Proceedings of the KDMiLe 2022*, Campinas. SBC.

Bazzan, A. L. C., Fehler, M., and Klügl, F. (2006). Learning to coordinate in a network of social drivers: The role of information. In Tuyls, K., Hoen, P. J., Verbeeck, K., and Sen, S., editors, *Proceedings of the International Workshop on Learning and Adaptation in MAS (LAMAS 2005)*, number 3898 in Lecture Notes in Artificial Intelligence, pages 115–128.

Bazzan, A. L. C. and Grunitzki, R. (2016). A multiagent reinforcement learning approach to en-route trip building. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 5288–5295. DOI: 10.1109/IJCNN.2016.7727899.

Cui, K., Tahir, A., Ekinci, G., Elshamanhory, A., Eich, Y., Li, M., and Koeppl, H. (2022). A survey on large-population systems and scalable multi-agent reinforcement learning.

Grunitzki, R. and Bazzan, A. L. C. (2016). Combining car-to-infrastructure communication and multi-agent reinforcement learning in route choice. In Bazzan, A. L. C., Klügl, F., Ossowski, S., and Vizzari, G., editors, *Proceedings of the Ninth Workshop on Agents in Traffic and Transportation (ATT-2016)*, volume 1678 of *CEUR Workshop Proceedings*, New York. CEUR-WS.org.

Huanca-Anquise, C. A. (2021). Multi-objective reinforcement learning methods for action selection: dealing with multiple objectives and non-stationarity. Master's thesis, Instituto de Informática, UFRGS, Porto Alegre, Brazil.

Huanca-Anquise, C. A., Bazzan, A. L., and Tavares, A. R. (2023). Multi-objective, multi-armed bandits: Algorithms for repeated games and application to route choice. *Revista de Informática Teórica e Aplicada*. To appear.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., and Wießner, E. (2018). Microscopic traffic simulation using SUMO. In *The 21st IEEE International Conference on Intelligent Transportation Systems*.

Mahmassani, H. S. (2016). Autonomous vehicles and connected vehicle systems: Flow and operations considerations. *Transp. Sci.*, 50(4):1140–1162. DOI: 10.1287/trsc.2016.0712.

Maimaris, A. and Papageorgiou, G. (2016). A review of intelligent transportation systems from a communications technology perspective. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 54–59. DOI: 10.1109/ITSC.2016.7795531.

Ortúzar, J. d. D. and Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons, Chichester, UK, 4 edition.

Ramos, G. de O., da Silva, B. C., and Bazzan, A. L. C. (2017). Learning to minimise regret in route choice. In Das, S., Durfee, E., Larson, K., and Winikoff, M., editors, *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, pages 846–855, São Paulo. IFAAMAS.

Ramos, G. de O. and Grunitzki, R. (2015). An improved learning automata approach for the route choice problem. In Koch, F., Meneguzzi, F., and Lakkaraju, K., editors, *Agent Technology for Intelligent Mobile Services and Smart Societies*, volume 498 of *Communications in Computer and Information Science*, pages 56–67. Springer. DOI: 10.1007/978-3-662-46241-6_6.

Santos, G. D. dos and Bazzan, A. L. C. (2020). Accelerating learning of route choices with C2I: A preliminary investigation. In *Proc. of the VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 41–48. SBC. DOI: 10.5753/kdmile.2020.11957.

Santos, G. D. dos and Bazzan, A. L. C. (2021). Sharing diverse information gets driver agents to learn faster: an application in en route trip building. *PeerJ Computer Science*, 7:e428. DOI: 10.7717/peerj-cs.428.

Santos, G. D. dos and Bazzan, A. L. C. (2022). A multiobjective reinforcement learning approach to trip building. In Bazzan, A. L., Dusparic, I., Lujak, M., and Vizzari, G., editors, *Proc. of the 12th International Workshop on Agents in Traffic and Transportation (ATT 2022)*, volume 3173, pages 160–174. CEUR-WS.org.

Santos, G. D. dos, Bazzan, A. L. C., and Baumgardt, A. P. (2021). Using car to infrastructure communication to accelerate learning in route choice. *Journal of Information and Data Management*, 12(2).

Tumer, K., Welch, Z. T., and Agogino, A. (2008). Aligning social welfare and agent preferences to alleviate traffic congestion. In Padgham, L., Parkes, D., Müller, J., and Parsons, S., editors, *Proceedings of the 7th Int. Conference on Autonomous Agents and Multiagent Systems*, pages 655–662, Estoril. IFAAMAS.

Van Moffaert, K. and Nowé, A. (2014). Multi-objective reinforcement learning using sets of Pareto dominating policies. *J. Mach. Learn. Res.*, 15(1):3483–3512.

Yu, Y., Han, K., and Ochieng, W. (2020). Day-to-day dynamic traffic assignment with imperfect information, bounded rationality and information sharing. *Transportation Research Part C: Emerging Technologies*, 114:59–83.

Zhou, B., Song, Q., Zhao, Z., and Liu, T. (2020). A reinforcement learning scheme for the equilibrium of the in-vehicle route choice problem based on congestion game. *Applied Mathematics and Computation*, 371:124895.