PRADA: Proactive Risk Assessment and Mitigation of Misinformed Demand Attacks on Navigational Route Recommendations

Ya-Ting Yang, *Graduate Student Member, IEEE*, Haozhe Lei, and Quanyan Zhu, *Senior Member, IEEE*

***Abstract*—Leveraging recent advances in wireless communication, IoT, and AI, intelligent transportation systems (ITS) played an important role in reducing traffic congestion and enhancing user experience. Within ITS, navigational recommendation systems (NRS) are essential for helping users simplify route choices in urban environments. However, NRS are vulnerable to information-based attacks that can manipulate both the NRS and users to achieve the objectives of the malicious entities. This study aims to assess the risks of misinformed demand attacks, where attackers use techniques like Sybil-based attacks to manipulate the demands of certain origins and destinations considered by the NRS. We propose a game-theoretic framework for proactive risk assessment of demand attacks (PRADA) and treat the interaction between attackers and the NRS as a Stackelberg game. Specifically, we consider the case of local-targeted attacks, in which the attacker aims to make the NRS recommend the authentic users towards a specific road that favors certain groups. Our analysis unveils the equivalence between users' incentive compatibility and Wardrop equilibrium recommendations and shows that the NRS and its users are at high risk when encountering intelligent attackers who can significantly alter user routes by strategically fabricating non-existent demands. To mitigate these risks, we introduce a trust mechanism that leverages users' confidence in the integrity of the NRS, and show that it can effectively reduce the impact of misinformed demand attacks. Numerical experiments are used to corroborate the results and support our discussion of the Resilience Paradox, where locally targeted attacks can sometimes benefit the overall traffic conditions. Our framework not only assists risk assessment in automating the evaluation process and estimating potential impacts but also aligns with standards like ISO/IEC 27005, offering a proactive approach to managing risks in ITS.**

***Index Terms*—Information attack, risk assessment, Stackelberg game, navigational recommendation.**

I. Introduction

HARNESSING vast information available from modern wireless communication and Internet of Things (IoT) advancements [1], [2], coupled with the progress made in data science and artificial intelligence [3], [4], intelligent transportation systems (ITS) have gained substantial attention for their ability to effectively tackle traffic congestion and elevate driver experiences. Within ITS, navigational recommendation systems (NRS) such as Google Maps and Apple Maps play a vital role in complex urban environments, as users, including drivers and pedestrians, may be overwhelmed by diverse route choices [5]. Based on the given information, the NRS offers routes to simplify users' decision-making processes, aiming to reduce travel duration, elevate user experiences, and alleviate congestion [6]. However, unlike routing in computer network systems [7], [8] or routing in transportation networks for connected autonomous vehicles [9], the NRS involves human drivers who may not always adhere to the recommendations, making user compliance unguaranteed [10]. Therefore, in this work, we refer to the NRS as the platform that provides incentive-compatible path recommendations [11], [12]. This ensures that users can not be better off by unilaterally deviating from the recommended strategies, and can be interpreted as a routing game between NRS users.

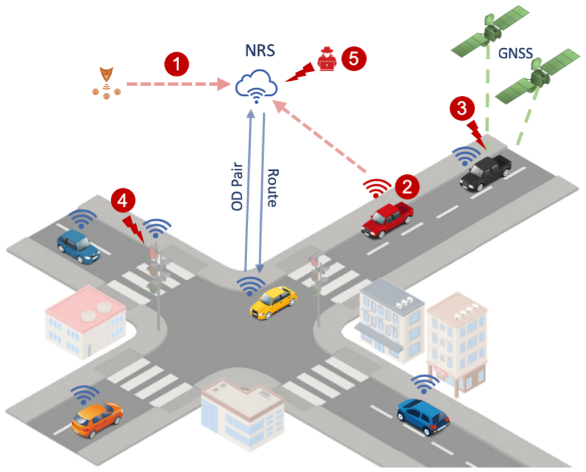


Fig. 1. The NRS receives user origin and destination (OD) requests, which are vulnerable to exploitation by malicious entities for demand attacks: (1) Sybil attack (2) Botnet (3) GNSS spoofing (4) Man-in-the-middle attack (5) Insider threats.

However, as illustrated in Fig. 1, the navigational recommendations are prone to various vulnerabilities [13] that attackers can leverage during the process to promote particular groups or businesses in a locally targeted sense or potentially exacerbate congestion levels on a broader network-wide scale. Within this context, information-based attacks emerge as a

Received 21 July 2024; revised 23 January 2025 and 27 August 2025; accepted 23 September 2025. Date of publication 29 September 2025; date of current version 16 October 2025. The associate editor coordinating the review of this article and approving it for publication was Prof. Edgar Weippl. *(Corresponding author: Ya-Ting Yang.)*

The authors are with the Department of Electrical and Computer Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: yy4348@nyu.edu; hl4155@nyu.edu; qz494@nyu.edu).

Digital Object Identifier 10.1109/TIFS.2025.3615726

critical concern, as they empower malicious entities to spread misinformation and manipulate both the NRS and its users to achieve their objectives [14]. Real-world examples include cases on the Waze platform, where residents may fabricate congestion reports to divert traffic away from their residential areas, aiming to maintain the tranquility of their surroundings [15]. Additionally, the recent study [16] demonstrates how Sybil-based attacks can effectively manipulate perceived crowdedness at places of interest and traffic congestion levels within Google Maps, which serves as an example that shows how fake users (that can lead to fake demands) can impact the NRS in real-time. These recent findings highlight that informational attacks are significant threats within NRS. Therefore, it is important to understand and estimate their potential impact, assess the associated risks, and develop proactive measures to mitigate and manage these attacks before they lead to serious consequences. Within this scope, our study is motivated by the recent research [16] and specifically focuses on the misinformed *demand attack*, which is defined as the manipulation of user demands between certain origins and destinations, and can be achieved through a variety of attack methods, including but not limited to Sybil-based attack, botnet, GNSS spoofing, man-in-the-middle attack on wireless communication, insider threat, etc.

According to ISO/IEC 27005 [17], an international standard for information security risk management, the risk assessment process includes the following key steps: (i) define the scope of the risk assessment, (ii) identify the sources of risk and how they exploit the vulnerabilities, (iii) assess the risk qualitatively or quantitatively based on the chosen methodology, (iv) design appropriate mitigation strategies for risk treatment, (v) decide whether to accept the remaining risk after treatment as well as monitor and periodically review the risk management process.

In this context, we focus on the informational demand attack within the NRS. The sources of risk are illustrated in Fig. 1 and detailed in Section III-A. Then, the risk assessment is based on event-based methodologies. One viable approach is manual evaluations by human experts. Another one is the data-driven method, which simulates user interactions within the transportation network and executes documented demand attacks manually to collect data for analysis [18]. While these approaches provide detailed insights, they tend to be task-specific and time-consuming to execute or implement. A more cost-effective alternative is the model-based approach, which offers essential estimates to guide proactive management, though it may not perfectly align with real-world observations. However, this approach is often fragmented, requiring multiple models to cover attackers, the NRS, users, and their interactions. To bridge this gap, we introduce a holistic, game-theoretic framework for Proactive Risk Assessment of Demand Attacks (PRADA). As shown in Fig. 2, PRADA integrates necessary elements into a cohesive framework, offering automated estimates and analyses for risk assessment to complement existing methods.

Specifically, the proposed PRADA framework is analyzed through three layers of games. The *inner layer* is the process for the NRS, as incentive-compatible recommendations can be interpreted as a routing game between NRS users. The equivalence between user incentive-compatible recommendations and Wardrop equilibrium recommendations aids in analyzing the *middle layer*, which captures the interaction between the threat profile of the malicious entity and the NRS as a Stackelberg game [12], [19]. In this game, the attacker acts as the leader who conveys misinformed demands, while the NRS, as the follower, responds to the provided information. The *outer layer* captures the interplay between the PRADA risk evaluator and the Stackelberg game at the middle layer. Specifically, this study focuses on the locally targeted attack for the attacker's objective, as it has rather few systematic studies. In this attack, the attacker manipulates the NRS to direct genuine users towards a specific road that benefits certain groups or businesses. From the perspective of the PRADA risk evaluator, our analysis shows that by strategically designing the misinformed demands, such as how many fake users for which OD pairs, the attacker can make the NRS redistribute the true users originally on other alternative paths towards the target road, leading to a higher risk for the NRS and its users.

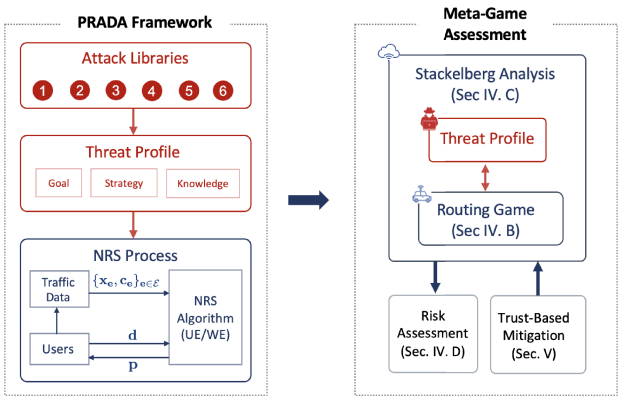


Fig. 2. The PRADA framework is analyzed through three layers of games between users, the NRS, the threat profile of the attacker, and the PRADA risk evaluator.

Building on the PRADA framework, we propose a mitigation method to address key steps (iv) and (v) of the risk assessment process. We utilize the concept of *trust score*, which measures how much users trust the integrity of the NRS and believe the recommendations are not manipulated. A higher trust score indicates greater user willingness to follow recommendations that differ from previous ones for the same origin and destination. By proposing a trust mechanism that incorporates trust score constraints into the model for NRS, we can effectively reduce the impact of demand attacks and lower the risk both locally and network-wide. Our analysis indicates that the dual variable associated with the trust constraint can be interpreted as a *trust risk factor*. It shows how sensitive the user's expected cost is to changes in the trust score.

To this end, our contribution can be summarized as follows:

- We identify vulnerabilities of the NRS and then propose a game-theoretic framework for holistic proactive risk assessment for misinformed demand attacks (PRADA).
- We employ a Stackelberg game approach to integrate the threat profile with the NRS into the PRADA framework.

Our analytical results and numerical studies demonstrate that users are at high risk when encountering intelligent attackers who target specific roads by fabricating fake demands on alternative paths.

- We introduce a trust mechanism that leverages users' confidence in the integrity of the NRS. Our findings show that the resulting trusted recommendation can effectively mitigate the impact of demand attacks, both in local-targeted and network-wide contexts.

II. Literature Review

A. Generic Attacks on ITS

Intending to enhance mobility, safety, sustainability, and traffic efficiency in urban transportation networks, modern ITS leverages a wide range of advanced technologies. These include sensors and cameras for data collection, wireless communication—particularly vehicle-to-everything (V2X) technology—for information exchange, GNSS for precise positioning, data analytics coupled with AI for processing, as well as mobile apps for distributing information. However, the extensive network of interconnected devices with vast information exchanged in ITS introduces vulnerabilities [20] that expand the potential cyber-physical attack surface (see [21] for real-world ITS attack cases). Within the domain of ITS, malicious entities or potential adversaries [13], [22] can exploit vulnerabilities within data and information infrastructure through a class of attacks known as informational attacks. These attacks, which include data falsification, data integrity breaches, and data poisoning [23], [24], are designed to divert drivers and escalate traffic congestion that leads to increased crash risks within urban transportation networks. These attacks exploit various tactics, including sensor and GNSS spoofing techniques [25], as well as employing man-in-the-middle and Sybil-based methods [16], [26].

B. Informational Attacks on NRS

We scrutinize the particular vulnerabilities inherent to the NRS, which are susceptible to a wide range of potential attacks [25]. Regarding generic attacks in the scenario of navigational guidance, attackers could compromise vehicles via wireless communication networks or manipulate real-time traffic conditions, leading to informational attacks. Such attacks can result in inaccurate traffic predictions and misguidance for drivers, contributing to network-wide traffic congestion and safety concerns [27]. For more specific examples, [28] illustrates how the availability of portable GNSS signal spoofing devices enables attackers to divert drivers from their intended destinations without their awareness. Additionally, [26] demonstrates the significant impact of a single Sybil device with limited resources on platforms like Waze, where false reports of congestion and accidents can automatically reroute user traffic. This work expands the threats discovered by recent studies [16] targeting NRS, where misinformation, such as fabricated demands, originates from Sybil-based users. Specifically, we assess the risk of locally targeted attacks that have rather few systematic studies, wherein attackers tend to strategically mislead users onto specific roads that favor certain groups.

C. Risk Assessment

Risk assessment is a systematic process for identifying, analyzing, and evaluating risks within a particular system or framework in various domains, including but not limited to energy systems [29], supply chains [30], IoT-based systems [31], autonomous vehicles [32], and transportation networks [33], [34]. It aims to understand potential adverse outcomes, enabling organizations or individuals to mitigate or manage such risks effectively. Within the field of transportation, risk assessment plays an important role, as evidenced by the substantial focus on (highway) crash risk evaluation [35], collision risk avoidance for autonomous vehicles [32], and risk-based route selection [33]. When addressing potential cyber risks in ITS, a deeper understanding of the attack model is necessary [36]. Cyber attackers are often intelligent and strategic, unlike non-strategic attackers who add disturbances uniformly or randomly. Therefore, a natural way to integrate the attack model into risk assessment and capture the interaction between the attacker and the target system is through game-theoretic approaches. These approaches are commonly employed to capture the threat posed by followers in dynamic games, such as Stackelberg games [37], [38], bargaining games [39], detection games [40], as well as in mechanism design problems involving contract designs [41] and incentive mechanisms [42], offering analytical tools and strategies for effective risk assessment and mitigation.

III. Misinformed Demand Attack

The nature of ITS, characterized by a vast network of interconnected devices and extensive data exchange, presents vulnerabilities that can be exploited by malicious entities through informational attacks, which target the system's data and information infrastructure. In this study, we specifically investigate one type of informational attack within the context of navigational recommendations, called the *demand attack*. As depicted in Fig. 1, the NRS typically receives navigational requests, including origin and desired destination (OD) pairs, from users. These requests contribute to the demand associated with each OD pair. More specifically, we denote Θ as the set of all possible OD pairs, and each OD pair $\theta \in \Theta$ is associated with a demand $d_\theta \in \mathbb{R}_{\geq 0}$ contributed from the users. Let $\mathbf{d} = \{d_\theta\}_{\theta \in \Theta}$ represent all the demands for later usage. For instance, suppose there are ten NRS users who wish to travel from the Empire State Building (origin) to Times Square (destination), which corresponds to OD pair θ, then the demand for this OD pair is $d_\theta = 10$.

Different demands generally lead to different recommendations from the NRS. For instance, with a single user, the NRS can simply suggest the shortest path. However, if many users are associated with the same OD pair, recommending the shortest path can lead to the 'flash crowd effect' [43], making it no longer the optimal choice. Malicious entities can exploit this fact to manipulate demand $\mathbf{d}$ to some other $\mathbf{d}'$, steering the NRS toward other recommendations that fulfill their own objectives. Therefore, the PRADA risk evaluator, who is responsible for conducting the risk assessment, must evaluate the risks for various libraries of attacks, consisting

of different malicious goals, types of attackers, and attack methods that can lead to demand attacks.

A. Demand Attack Methods

In this subsection, we mention some techniques indicated in Fig. 1 that the attacker with related knowledge can utilize to launch the misinformed demand attack.

1) Sybil Attacks: Attackers can generate numerous fake identities (non-existent users) as shown in Fig. 1 and then simulate these users at specific locations [26]. These non-existent users send OD pair requests to the NRS through emulators [16]. When computing recommendations, the NRS considers these fake demands alongside genuine ones, leading to recommendations that differ from those based solely on authentic demands. Consequently, the attacker can strategically redistribute legitimate users by launching Sybil-based attacks with fake demands. Furthermore, since the non-existent users do not actually drive on the roads after receiving the recommendations, the actual traffic conditions caused by legitimate users will differ from the NRS's expectations.

2) Botnet: An attacker can deploy a botnet, a network of compromised devices controlled remotely [44]. These devices can range from infected smartphones to IoT devices and computers. The attacker can command the botnet to send numerous navigational OD pair requests to the NRS, as shown in Fig. 1, simulating authentic users seeking guidance between various origins and destinations. Similar to the Sybil attack, the resulting recommendations will differ due to the fake demands (that do not exist on roads) generated by the botnet. This allows the attacker to redistribute legitimate users by utilizing the botnet to strategically create fake demands, aligning the recommendations with the attacker's objectives.

3) GNSS Spoofing: Attackers can employ GNSS spoofing techniques [25], [28] to alter the perceived location of NRS users, as illustrated in Fig. 1. This manipulation can cause users to send navigation requests with incorrect origins. For example, when a user selects 'current location' as the origin, the spoofed GNSS signal can make the NRS believe the user is in a different place. Consequently, the requested OD pair is being manipulated, affecting the overall demand considered by the NRS. This disruption can significantly impact navigational recommendations, especially if multiple NRS users are affected simultaneously, leading to incorrect navigational recommendations (that may align with the attacker's objective).

4) Wireless Communication Network: We use the man-in-the-middle attack as an illustrative example of a demand attack through wireless communication networks [45]. In this scenario, attackers attempt to position themselves between the user's device and the NRS server, allowing them to intercept communications in between (see Fig. 1). When a user sends a request for route recommendations for a specific OD pair, the attacker modifies the request before it reaches the system's servers. This modification can involve altering the origin, destination, or other parameters within the request. By manipulating multiple requests from different users, the attacker can increase or decrease the demand for specific OD pairs. As a result, the manipulated demand influences the NRS's recommendation to align with the attacker's objectives.

5) Insider Threats: An insider, such as an employee or contractor with access to the NRS infrastructure, can directly manipulate the data or algorithm within the system. This manipulation may involve altering the demand data, such as increasing or decreasing the number of requests for specific OD pairs so that the insider can bias the recommendations suggested by the NRS to align with their objectives. Since this work focuses on misinformed demand attack, insider threats here specifically correspond to manipulating demand-related data rather than arbitrarily altering the NRS algorithms.

It is worth noting that although the technical procedures of these five methods differ, they all lead to a common effect at the system level: the demand vector observed by the NRS is altered from $\mathbf{d}$ to a manipulated $\mathbf{d}'$.

B. Types of Attackers

We can categorize attackers into two main types.

1) Non-Strategic Attacker: Non-strategic attackers may lack the understanding of how the NRS generates recommendations for users, or they may not pay attention to and simply disregard this information. Instead, a non-strategic attacker often manipulates demands by uniformly or randomly increasing or decreasing the number of requests associated with some OD pairs. They then observe whether such manipulation achieves their desired outcome.

2) Strategic Attacker: Strategic attackers are often more intelligent and possess either a deep understanding of how the NRS generates recommendations for users or the ability to model the NRS. With this knowledge, they assess or observe the outcomes of the NRS when under attack. By leveraging this insight, strategic attackers can utilize efficient strategies to achieve their goals with fewer resources used and less time spent.

C. Attacker's Goals

Imagine an attacker driven by self-interest, in conflict with the overall social welfare goal of reducing congestion. This scenario can be studied at both local targeted and network-wide levels: the former pertains to specific groups or locations, while the latter considers the system-wide impact.

1) Local-Targeted Attacks: The attacker seeks to bias the system by suggesting paths that favor particular groups (e.g., higher-paying users) or businesses (e.g., those paying the attacker to ensure users see particular ads or pass by their shops). For example, a restaurant owner could pay malicious entities to ensure a certain volume of users are directed by the NRS to pass by the road where the restaurant is located, as illustrated in Fig. 3.

The impact of these attacks can be measured by the difference in traffic flow on specific roads with and without the attacks.

2) Network-Wide Attacks: The attacker aims to disrupt the system by increasing delays or congestion levels across the network, consequently raising the overall travel time cost for users. These attacks can harm the system's reputation, leading to user dissatisfaction or a loss of trust in the NRS.

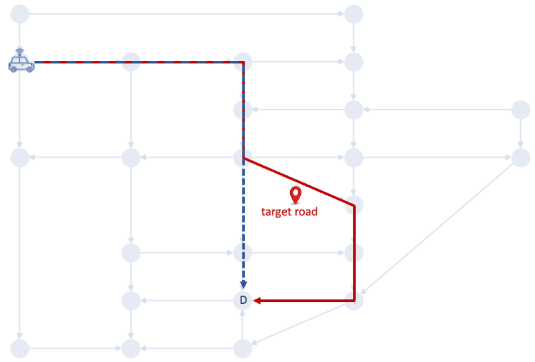


Fig. 3. Local-targeted attacks: The malicious entity manipulates the NRS to guide users through the target road. The blue dashed line represents the original recommendation from the NRS, and the red line illustrates the one under attack.

Note that in this work, we categorized attackers by types and goals for illustration of systematic risk assessment; however, real-world attackers may have non-binary types and pursue more diverse or unpredictable goals. Moreover, while this study focuses on malicious entities with local-targeted objectives, these attacks can result in both local-targeted and network-wide impacts. The metrics used to assess these risks are detailed in Section IV-D.

IV. Proactive Risk Assessment

This section aims to assess the risk caused by misinformed demand attacks on the NRS. The proposed framework for proactive risk assessment of demand attacks (PRADA) is illustrated in Fig. 2. The PRADA evaluator proactively evaluates risks by employing a library of potential and real-world documented attack scenarios. Each attack scenario is characterized by the attacker's goal (objective), attack strategy (type of attacker), and attack method (knowledge), which together form the threat profile. The resulting demand attack from the threat profile then influences the recommendations suggested by the NRS.

The framework is analyzed through meta-game, which consists of three layers of games. The inner layer focuses on the routing game between NRS users. That is, the NRS aims to provide incentive-compatible recommendations to users, as human users may not always follow recommendations if they find better alternatives that align with their pReferences. The middle layer employs a Stackelberg game approach to capture the interaction between the threat profile of the attacker and the NRS. Here, the attacker acts as the leader, manipulating demand, while the NRS, as the follower, responds to the provided information. The outer layer is a meta-game for the interplay between the PRADA risk evaluator and the middle layer. It involves assessing the impacts of different threat profiles from the attack libraries. Note that while this work focuses on incentive-compatible NRS, it is flexible enough to accommodate other types of NRS as well. In such cases, the outcomes of the inner layer will differ, and the analysis of the middle and outer layers will be based on the corresponding results from the inner layer.

We then conduct a sensitivity analysis of demand attacks and propose metrics for measuring local-targeted and network-wide impacts.

A. Settings for Urban Transportation Networks

The urban transportation network can be represented by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the set of nodes $\mathcal{V}$ corresponds to intersections and the set of edges $\mathcal{E}$ indicates the roads. Traveling along a road $e \in \mathcal{E}$ incurs a road-specific cost $c_e : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_+$ associated with the flow $x_e \in \mathbb{R}_{\geq 0}$ on road e. One usual choice for the cost $c_e(\cdot)$ is the standard Bureau of Public Roads (BPR) function

$$c_e(x_e) = t_e \left(1 + \alpha \left(\frac{x_e}{k_e}\right)^{\beta}\right)$$

for travel time costs. Here, $t_e \in \mathbb{R}_+$ represents the free-flow travel time on road e, $k_e \in \mathbb{R}_+$ signifies the capacity of road e, and $\alpha, \beta \in \mathbb{R}_{\geq 0}$ are some parameters.

B. Models for NRS

The user set of NRS is denoted as $\mathcal{U}$. Each user, $u \in \mathcal{U}$, is associated with a specific origin $O_u \in \mathcal{V}$ and destination $D_u \in \mathcal{V}$ pair. We refer to the pair as an OD pair, expressed by $\theta_u = (O_u, D_u)$, and the set of OD pairs for the NRS users is $\Theta_{\mathcal{U}} \subseteq \Theta$, with $\Theta = |\mathcal{V}| \times |\mathcal{V}|$ representing all the possible OD pairs within the network. Then, user u with OD pair θ_u has the feasible path choice set $\mathcal{S}_u = \{s_{u,1}, \ldots, s_{u,k_u}\}$, which is identical to the feasible path choice set $\mathcal{S}_\theta = \{s_{\theta,1}, \ldots, s_{\theta,k_\theta}\}$ for OD pair θ, where $\theta = \theta_u$. Each choice $s_{u,i} \in \mathcal{S}_u$ or $s_{\theta,i} \in \mathcal{S}_\theta$ provides the user u a path from the origin to the desired destination. To this end, the elements of the urban transportation network considered by the NRS can be encapsulated using the notation $\mathscr{R} = \langle \mathcal{G}, (c_e(\cdot))_{e\in\mathcal{E}}, \mathcal{U}, (\mathcal{S}_u)_{u\in\mathcal{U}} \rangle$, and we call $\mathscr{R}$ the "NRS component".

1) User Equilibrium Recommendations (UE): We consider the scenario where the NRS recommends a mixed strategy over feasible path choices to the users. Define $\mathcal{P}_u := \Delta\mathcal{S}_u$ as the simplex of $\mathcal{S}_u$. A mixed strategy for user u is $\mathbf{p}_u \in \mathcal{P}_u$ so that $\mathbf{p}_u = \{p_{u,i}\}_{s_{u,i}\in\mathcal{S}_u}$ is a probability distribution over $\mathcal{S}_u$. Each element $p_{u,i} \in [0, 1]$ denotes the probability that the NRS recommends path $s_{u,i} \in \mathcal{S}_u$ to user u, and needs to satisfy the constraints $\sum_{i=1}^{k_u} p_{u,i} = 1, \ \forall u \in \mathcal{U}$. That is,

$$\mathcal{P}_u := \left\{ \mathbf{p}_u \in \mathbb{R}^{k_u} \mid p_{u,i} \geq 0, i = 1, \cdots, k_u, \sum_{i=1}^{k_u} p_{u,i} = 1 \right\}.$$

Then, let $\mathcal{P} := \Pi_{u\in\mathcal{U}} \mathcal{P}_u$, the recommendation suggested by the NRS to all users is $\mathbf{p} = \{\mathbf{p}_u\}_{u\in\mathcal{U}} \in \mathcal{P}$. Note that $\mathbf{p}$ can be written as $\{\mathbf{p}_u, \mathbf{p}_{-u}\}$. Here, $\mathbf{p}_u \in \mathcal{P}_u$ is the recommendation to user u, while $\mathbf{p}_{-u} \in \Pi_{u'\in\mathcal{U}\setminus u} \mathcal{P}_{u'}$ represents the recommendations to other users except user u.

In transportation, from a microscopic perspective, the probability $p_{u,i}$ can be interpreted as the expected volume generated by user u along path $s_{u,i}$. This, in turn, contributes to the overall expected road flow (load) $x_e^r : \mathcal{P} \mapsto \mathbb{R}_{\geq 0}$ on road $e \in \mathcal{E}$, which aggregates the expected volumes from users whose feasible paths include road e as below.

$$x_e^r(\mathbf{p}) = \sum_{u\in\mathcal{U}} \sum_{s_{u,i}\in\mathcal{S}_u} p_{u,i} a_{e s_{u,i}},$$

where $a_{es_{u,i}}$ is an element of the road-path incidence matrix $A_{|\mathcal{E}|\times|\Pi_{u\in\mathcal{U}}\mathcal{S}_u|} = [a_{es_{u,i}}]$, and is defined as follows.

$$a_{es_{u,i}} = \begin{cases} 1 & \text{if } e \in s_{u,i}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, from user u's perspective, recommendations to other users, $\mathbf{p}_{-u}$, can affect the expected road load on the roads within their feasible paths. In this context, $x_e^r(\mathbf{p})$ can also be written as $x_e^r(\mathbf{p}_u, \mathbf{p}_{-u})$. Accordingly, the generalized travel cost $C_{u,i} : \mathcal{P} \mapsto \mathbb{R}_+$ for user u's path $s_{u,i}$ is calculated by summing the costs of all the roads along the path, with such costs influenced by recommendations given to other users. Specifically, we can express $C_{u,i}(\mathbf{p})$ as:

$$C_{u,i}(\mathbf{p}_u, \mathbf{p}_{-u}) = \sum_{e\in s_{u,i}} c_e(x_e^r(\mathbf{p}_u, \mathbf{p}_{-u})).$$

In this context, the expected cost evaluated by user u is $F_u^r : \mathcal{P} \mapsto \mathbb{R}_{\geq 0}$, where

$$F_u^r(\mathbf{p}_u, \mathbf{p}_{-u}) = \sum_{i=1}^{k_u} p_{u,i} C_{u,i}(\mathbf{p}_u, \mathbf{p}_{-u}). \tag{1}$$

Note that a recommendation $\mathbf{p} \in \mathcal{P}$ can be interpreted as the strategy profile in a routing game between NRS users. Hence, the routing game addressed by the NRS can be defined as $\Gamma^r = \langle \mathscr{R}, \mathscr{F}^r \rangle$, where $\mathscr{F}^r = (F_u^r)_{u\in\mathcal{U}}$ represents the costs evaluated by users. However, human users may choose not to follow the NRS recommendation if they find a better alternative. Therefore, to ensure user adherence that leads to a guaranteed performance over the network, the NRS must suggest a recommendation $\mathbf{p} \in \mathcal{P}$, where $\mathbf{p}_u \in \mathcal{P}_u$ is preferred by user u given the recommendations to other users $\mathbf{p}_{-u} \in \Pi_{u'\in\mathcal{U}\setminus\{u\}}\mathcal{P}_{u'}$, for all $u \in \mathcal{U}$. That is, given the recommendations $\mathbf{p}_{-u}$ to users other than u, user u has no incentive to unilaterally deviate from the recommended $\mathbf{p}_u$. This coincides with the concept of user equilibrium (UE), which is defined as follows:

Definition 1 (User Equilibrium Recommendation): Considering a routing game addressed by the NRS defined as $\Gamma^r = \langle \mathscr{R}, \mathscr{F}^r \rangle$, a mixed strategy profile $\mathbf{p} \in \mathcal{P}$ for all the users is called a user equilibrium recommendation if it satisfies:

$$F_u^r(\mathbf{p}_u, \mathbf{p}_{-u}) - F_u^r(\mathbf{p}'_u, \mathbf{p}_{-u}) \leq 0, \quad \forall\ \mathbf{p}'_u \in \mathcal{P}_u,\ \forall u \in \mathcal{U}. \tag{2}$$

UE recommendation can be found by gradient descent-based method. Let $\text{proj}_{\mathcal{P}_u}$ represent the projection onto simplex $\mathcal{P}_u$ and $\rho \in \mathbb{R}$ denote the step size, problem (2) can be solved by finding a fixed point to:

$$\mathbf{P}_u^* = \text{proj}_{\mathcal{P}_u}\left[\mathbf{p}_u^* - \rho\nabla_u F_u^r(\mathbf{p}_u^*, \mathbf{p}_{-u}^*)\right], \quad \forall u \in \mathcal{U}. \tag{3}$$

Note that according to our work [46] [Proposition 2], we have proved that, assuming $\sum_{n=1}^{\infty}\rho_n^2 < \infty$, where ρ_n represents the step size at time step n, the update algorithm $\mathbf{P}_u^{(n+1)} = \text{proj}_{\mathcal{P}_u}\left[\mathbf{p}_u^{(n)} - \rho_n\nabla_u F_u^r(\mathbf{p}_u^{(n)}, \mathbf{p}_{-u}^{(n)})\right]$ for all user $u \in \mathcal{U}$ leads to the convergence to a UE under mild conditions.

2) Wardrop Equilibrium Recommendations (We): In this subsection, we use the concept of Wardrop equilibrium (WE) as the foundation for the WE-based recommendations.

Recall that $\mathcal{U}$ represents the user set of the NRS, with their associated set of OD pairs, denoted as $\Theta_{\mathcal{U}}$. For each OD pair $\theta \in \Theta_{\mathcal{U}}$, the demand flow aggregated from users, $d_\theta = \sum_{u\in\mathcal{U}} \mathbf{1}_{\{\theta_u=\theta\}}$, must be routed from the corresponding origin to the desired destination. As for OD pair $\theta \in \Theta \setminus \Theta_{\mathcal{U}}, d_\theta = 0$. The set of feasible paths for each OD pair θ is $\mathcal{S}_\theta = \{s_{\theta,1}, \ldots, s_{k_\theta}\}$. Then, let vector $\mathbf{y}_\theta = \{y_{\theta,i}\}_{s_{\theta,i}\in\mathcal{S}_\theta} \in \mathbb{R}^{k_\theta}$ so that each element $y_{\theta,i}$ represents the expected flow generated by the users being recommended through path $s_{\theta,i}$, and needs to satisfy the constraints: $\sum_{s_{\theta,i}\in\mathcal{S}_\theta} y_{\theta,i} = d_\theta$. That is, we can define

$$\mathcal{Y}_\theta := \left\{\mathbf{y}_\theta \in \mathbb{R}^{k_\theta} \mid y_{\theta,i} \geq 0, i = 1, \cdots, k_\theta, \sum_{i=1}^{k_\theta} y_{\theta,i} = d_\theta\right\}.$$

By denoting $\mathcal{Y} := \Pi_{\theta\in\Theta}\mathcal{Y}_\theta$, the expected flow recommended by the NRS on all the paths $s_{\theta,i} \in \mathcal{S}_\theta, \theta \in \Theta$ is $\mathbf{y} = \{\mathbf{y}_\theta\}_{\theta\in\Theta} \in \mathcal{Y}$. Similar to the UE recommendation, $\mathbf{y}$ also contributes to the expected road flow (load) $x_e^w : \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ on road $e \in \mathcal{E}$ as below.

$$x_e^w(\mathbf{y}) = \sum_{\theta\in\Theta}\sum_{s_{\theta,i}\in\mathcal{S}_\theta} y_{\theta,i} a'_{es_{\theta,i}},$$

where $a'_{es_{\theta,i}}$ is an element of the road-path incidence matrix $A'_{|\mathcal{E}|\times|\Pi_{\theta\in\Theta}\mathcal{S}_\theta|} = [a'_{es_{\theta,i}}]$, and is defined as follows.

$$a'_{es_{\theta,i}} = \begin{cases} 1 & \text{if } e \in s_{\theta,i}, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the cost evaluated by user u with OD pair $\theta = \theta_u$ for path $s_{\theta,i}$ is $F_{\theta,i}^w : \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$, where

$$F_{\theta,i}^w(\mathbf{y}) = \sum_{e\in s_{\theta,i}} c_e\left(x_e^w(\mathbf{y})\right).$$

In this context, the routing game addressed by the NRS can be defined as $\Gamma^w = \langle \mathscr{R}, \mathscr{F}^w \rangle$, where $\mathscr{F}^w = (F_{\theta,i}^w)_{s_{\theta,i}\in\mathcal{S}_\theta,\theta\in\Theta}$ represents the costs evaluated by users. Then, the WE-based recommendation is defined as the following.

Definition 2: [Wardrop Equilibrium Recommendation] Consider a routing game addressed by the NRS, defined as $\Gamma^w = \langle \mathscr{R}, \mathscr{F}^w \rangle$. A feasible path flow and road load pair $(\mathbf{y}, \mathbf{x}^w)$ with $\mathbf{y} \in \mathcal{Y}$ and $\mathbf{x}^w = \{x_e^w(\mathbf{y})\}_{e\in\mathcal{E}} \in \mathbb{R}_{\geq 0}^{|\mathcal{E}|}$ is called a Wardrop equilibrium recommendation if it satisfies:

$$\begin{aligned} F_{\theta,i}^w(\mathbf{y}) \leq F_{\theta,j}^w(\mathbf{y}) \text{ when } & y_{\theta,i} > 0, \\ & \forall s_{\theta,i}, s_{\theta,j} \in \mathcal{S}_\theta,\ \forall\theta \in \Theta. \end{aligned} \tag{4}$$

In other words, the WE-based recommendation results in the minimal prevailing costs for all used paths. Then, according to Beckmann [47], WE can be computed as the solution to the following optimization problem,

$$W(\mathbf{d}): \min_{\mathbf{y},\mathbf{x}^w} \sum_{e\in\mathcal{E}} \int_0^{x_e^w} c_e(z)dz \tag{5a}$$

$$\text{s.t.} \quad \sum_{i=1}^{k_\theta} y_{\theta,i} = d_\theta, \quad \forall\theta \in \Theta, \tag{5b}$$

$$y_{\theta,i} \geq 0, \quad \forall s_{\theta,i} \in \mathcal{S}_\theta, \ \forall \theta \in \Theta, \tag{5c}$$

$$x_e^w(\mathbf{y}) = \sum_{\theta\in\Theta} \sum_{s_{\theta,i}\in\mathcal{S}_\theta} y_{\theta,i} a'_{e s_{\theta,i}}, \quad \forall e \in \mathcal{E}, \tag{5d}$$

and the corresponding WE recommendation pair is represented as $(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}^w)$.

3) Connection Between UE and We: Correspondence can be observed between Definition 1 and 2, which assists us in the subsequent analysis when integrating the demand attack model into our PRADA framework. That is, user u and the set of feasible paths $\mathcal{S}_u$ in UE correspond to OD pair θ and $\mathcal{S}_\theta$ in WE, the expected flow x_e^r on the road e correspond to the road load x_e^w, and the probability $p_{u,i}$ that user u be recommended on the path $s_{u,i}$ in UE corresponds to the path flow $y_{\theta,i}$ in WE. Note that letting $d_\theta = 1$ in WE leads to $\sum_{i=1}^{k_\theta} y_{\theta,i} = 1$ that corresponds to $\mathbf{p}_u \in \mathcal{P}_u$ for $\theta = \theta_u$.

Let $(\mathbf{p}, \mathbf{x}^r)$ denote the UE pair, where $\mathbf{p} \in \mathcal{P}$ and $\mathbf{x}^r = \{x_e^r(\mathbf{p})\}_{e\in\mathcal{E}} \in \mathbb{R}_{\geq 0}^{|\mathcal{E}|}$. As a result, the WE solution pair $(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}^w)$ that corresponds to the $(\mathbf{p}, \mathbf{x}^r)$ pair can be viewed as a feasible solution for the UE recommendation. More specifically, if $(\mathbf{p}, \mathbf{x}^r)$ is a WE, then for all user $u \in \mathcal{U}$ in constraint (2): Since only paths with minimum cost are utilized, all the paths used by any given user have the same cost. That is, for $p_{u,i} > 0$, the cost $C_{u,i}(\mathbf{p}_u, \mathbf{p}_{-u})$ should be the same for u. The overall expected cost $\sum_{i=1}^{k_u} p_{u,i} C_{u,i}(\mathbf{p}_u, \mathbf{p}_{-u})$ is independent of the probability $p_{u,i}$ of $C_{u,i}(\mathbf{p}_u, \mathbf{p}_{-u})$. Lastly, note that if $(\mathbf{p}, \mathbf{x}^r)$ is an equilibrium, there is no incentive for a user u to deviate to any other $\mathbf{p}'_u \in \mathcal{P}_u$.

Proposition 1: A WE solution pair $(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}^w)$ defined in Definition 2 that corresponds to the UE pair $(\mathbf{p}, \mathbf{x}^r)$ is a feasible solution for the UE recommendation defined in Definition 1.

Proof: The proof follows from the discussion above. □

Similarly, in the case where $d_\theta = \sum_{u\in\mathcal{U}} \mathbf{1}_{\{\theta_u=\theta\}} > 1$, the NRS can recommend a mixed strategy $p_{u,i}$ over the feasible path $s_{u,i} \in \mathcal{S}_u = \mathcal{S}_{\theta_u}$, with each $p_{u,i} = \widehat{y}_{\theta,i}/d_\theta$, where $\widehat{y}_{\theta,i}$ comes from the WE flow-load pair $(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}^w)$ in which only paths with minimum cost are utilized for each OD pair. Following this recommendation, the recommended path flow $\mathbf{y}^r := \{y^r_{\theta,i}\}_{\theta\in\Theta, s_{\theta,i}\in\mathcal{S}_\theta}$, with each $y^r_{\theta,i} = \sum_{u\in\mathcal{U}} p_{u,i} \mathbf{1}_{\{\theta_u=\theta\}}$, is the same as the WE path flow $\widehat{\mathbf{y}}$.

Since a feasible UE recommendation can be computed from the WE solution pair$(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}^w)$, we have the following remarks.

Remark 1: The recommended road flow load $\mathbf{x}^r$ provided by the UE recommendation coincides with $\widehat{\mathbf{x}}^w$ in WE.

It is worth noting that, according to problem (5), the WE road flow load $\widehat{\mathbf{x}}^w$ is uniquely determined if the cost function $c_e(\cdot)$ on each road $e \in \mathcal{E}$ is strictly increasing in the corresponding road flow load x_e [48].

Remark 2: Under the assumption that the cost function $c_e(\cdot)$ on each road $e \in \mathcal{E}$ is strictly increasing in the corresponding road flow load x_e, the UE road flow load $\mathbf{x}^r$ is also uniquely determined and is equivalent to $\widehat{\mathbf{x}}^w$. Hence, we can denote $\mathbf{x} \in \mathbb{R}_{\geq 0}^{|\mathcal{E}|}$ and have $\mathbf{x} = \widehat{\mathbf{x}}^w = \mathbf{x}^r$.

C. Stackelberg Game for Risk Assessment

In this work, "risk" refers to the adverse impact on the NRS and its users due to manipulated demands. To assess such risks across different attack types and goals, we adopt a Stackelberg game approach to capture the interaction between the attacker (AT) and the NRS. In this setup, the attacker acts as the leader, deciding on the misinformed demands, while the NRS, as the follower, generates recommendations based on the provided information. The risk is assessed by quantifying changes in resulting traffic flows of authentic users, measured through the local-targeted and network-wide impact metrics.

1) Strategic Attackers With Local-Targeted Objectives: The subsequent analyses focus on scenarios involving strategic attackers with local-targeted attack objectives. Specifically, the attacker intends to have a desired level of expected flow load caused by genuine NRS users on the target road $e' \in \mathcal{E}$. That is, the attacker aims to make $x^r_{e'}(\mathbf{p}) = \sum_{u\in\mathcal{U}} \sum_{s_{u,i}\in\mathcal{S}_u} p_{u,i} a_{e' s_{u,i}}$ in UE recommendation, which is equivalent to $x^w_{e'}(\widehat{\mathbf{y}}) = \sum_{\theta\in\Theta} \sum_{s_{\theta,i}\in\mathcal{S}_\theta} \widehat{y}_{\theta,i} a'_{e' s_{\theta,i}}$ in WE according to Remark 2, achieve a desired level $\gamma \in \mathbb{R}_{\geq 0}$.

Consider the situation where a Sybil-based attacker generates non-existent demands $\mathbf{d}^a \in \mathcal{D}$, where $\mathcal{D} = \mathbb{Z}_{\geq 0}^{|\Theta|}$ using Sybil (fake) users. Then, the NRS will need to consider an aggregated demand of $\mathbf{d}' = \mathbf{d} + \mathbf{d}^a$ when generating the recommendation. Note that for each OD pair θ, the demand d'_θ under attack consists of $d_\theta + d^a_\theta$. Without loss of generality, we can assume that only a proportion of $\frac{d_\theta}{d_\theta + d^a_\theta}$ of the WE expected path flow $\widehat{y}'_{\theta,i}$ with respect to $W(\mathbf{d}')$ is caused by authentic users. Hence, we denote $\widehat{y}^u_{\theta,i} = \left(\frac{d_\theta}{d_\theta + d^a_\theta}\right) \widehat{y}'_{\theta,i}$ and $\widehat{\mathbf{y}}^u = \{\widehat{y}^u_{\theta,i}\}_{\theta\in\Theta, s_{\theta,i}\in\mathcal{S}_\theta}$ for path flow generated by true users. In this case, the attacker aims to make the flow load from genuine users $x^w_{e'}(\widehat{\mathbf{y}}^u) = \sum_{\theta\in\Theta} \sum_{s_{\theta,i}\in\mathcal{S}_\theta} \widehat{y}^u_{\theta,i} a'_{e' s_{\theta,i}}$ on the targeted road e' reach the desired level γ. To this end, the Stackelberg game between the attacker (AT) and the NRS can be defined as $\Gamma^s = \langle \text{AT}, \mathcal{D}, U_{AT}, (e', \gamma), \Gamma^w \rangle$, where $U_{AT} : \mathcal{D} \mapsto \mathbb{R}_{\geq 0}$ is the attacker's cost in terms of the resources spent in fabricating fake demands. The leader-follower problem is formulated as follows, and can be solved by gradient descent-based algorithms [49].

$$\min_{\mathbf{d}^a} \ U_{AT}(\mathbf{d}^a) = \sum_{\theta\in\Theta} d^a_\theta \tag{6a}$$

$$\text{s.t.} \sum_{\theta\in\Theta} \sum_{s_{\theta,i}\in\mathcal{S}_\theta} \widehat{y}^u_{\theta,i} a'_{e' s_{\theta,i}} \geq \gamma, \tag{6b}$$

$$(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}^w) \in \arg\min W(\mathbf{d} + \mathbf{d}^a), \tag{6c}$$

$$d^a_\theta \geq 0, \forall \theta \in \Theta. \tag{6d}$$

Lastly, note that the level γ in problem (6) can not be arbitrarily large, which leads us to the following.

Remark 3: The edge load x_e^w is bounded by the total demand of the true user $d_T = \sum_{\theta\in\Theta} d_\theta$. Hence, the attacker's desired level γ is also upper-bounded by $d_T = \sum_{\theta\in\Theta} d_\theta$.

It is important to note that the proposed Stackelberg game formulation abstracts away from the specific attack method and focuses on the resulting manipulated demand perceived by the NRS. This abstraction then allows us to analyze the influence of diverse attack methods under a single mathematical framework.

2) Sensitivity Analysis for Demand Attack: Under the assumption that the cost function $c_e(\cdot)$ is continuous and increasing in x_e, a pair $(\mathbf{y}, \mathbf{x})$ is a minimizer of $W(\mathbf{d})$ if and

only if it satisfies the following Karush–Kuhn–Tucker (KKT) conditions.

$$c_e(x_e) - \lambda_e = 0, \quad \forall e \in \mathcal{E}, \tag{7a}$$

$$-\nu_\theta + \sum_{e \in \mathcal{E}} \lambda_e a'_{e s_{\theta,i}} - \mu_{\theta,i} = 0, \quad \forall s_{\theta,i} \in \mathcal{S}_\theta, \ \forall \theta \in \Theta, \tag{7b}$$

$$\mu_{\theta,i} y_{\theta,i} = 0, \quad \forall s_{\theta,i} \in \mathcal{S}_\theta, \ \forall \theta \in \Theta, \tag{7c}$$

with Lagrangian multipliers $\nu_\theta \in \mathbb{R}_{\geq 0}, \forall \theta \in \Theta$, $\lambda_e \in \mathbb{R}_{\geq 0}, \forall e \in \mathcal{E}$, and $\mu_{\theta,i} \in \mathbb{R}_{\geq 0}, \forall s_{\theta,i} \in \mathcal{S}_\theta, \forall \theta \in \Theta$. Then, a pair $(\mathbf{y}, \mathbf{x})$ satisfying the constraints with multipliers $-\boldsymbol{\nu} = -(\nu_\theta)_{\theta \in \Theta}, \boldsymbol{\lambda} = (\lambda_e)_{e \in \mathcal{E}}, \boldsymbol{\mu} = (\mu_{\theta,i})_{s_{\theta,i} \in \mathcal{S}_\theta, \theta \in \Theta}$ also satisfies

$$\nu_\theta = \sum_{e \in s_{\theta,i}} c_e(x_e) - \mu_{\theta,i} \begin{cases} = \sum_{e \in s_{\theta,i}} c_e(x_e), & y_{\theta,i} > 0, \\ \leq \sum_{e \in s_{\theta,i}} c_e(x_e), & y_{\theta,i} = 0, \end{cases}$$

which coincides with the definition of WE recommendation.

Then, with KKT conditions, we aim to examine how the WE pair $(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}^w)$ can be influenced by changes in the demand $\mathbf{d}$ according to [48].

Proposition 2: Let the pair $(\mathbf{y}, \mathbf{x})$ with the corresponding multipliers $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ described in (7) be a WE for demand $\mathbf{d}$ and $(\mathbf{y}', \mathbf{x}')$ with corresponding multipliers $\boldsymbol{\nu}'$ and $\boldsymbol{\mu}'$ be a WE for demand $\mathbf{d}'$. Then, $(\boldsymbol{\nu}' - \boldsymbol{\nu})^T (\mathbf{d}' - \mathbf{d}) \geq \boldsymbol{\mu}'^T \mathbf{y} + \boldsymbol{\mu}^T \mathbf{y}' \geq 0$.

Proof: Under the assumption that the cost function $c_e(\cdot)$ is continuous and increasing in x_e, which indicates that $[c_e(x'_e) - c_e(x_e)](x'_e - x_e) \geq 0, \forall e \in \mathcal{E}$, then with $x'_e = x_e^w(\mathbf{y}') = \sum_{\theta \in \Theta} \sum_{s_{\theta,i} \in \mathcal{S}_\theta} y'_{\theta,i} a'_{e s_{\theta,i}}$ and $x_e = x_e^w(\mathbf{y}) = \sum_{\theta \in \Theta} \sum_{s_{\theta,i} \in \mathcal{S}_\theta} y_{\theta,i} a'_{e s_{\theta,i}}$, we have the following:

$$\sum_{\theta \in \Theta} \sum_{s_{\theta,i} \in \mathcal{S}_\theta} \sum_{e \in \mathcal{E}} [c_e(x'_e) - c_e(x_e)] a'_{e s_{\theta,i}} (y'_{\theta,i} - y_{\theta,i}) \geq 0.$$

Note that the KKT conditions in (7a) give us $c_e(x_e) = \lambda_e, \forall e \in \mathcal{E}$. With (7b), the above inequality becomes:

$$\sum_{\theta \in \Theta} \sum_{s_{\theta,i} \in \mathcal{S}_\theta} [(\nu'_\theta + \mu'_{\theta,i}) - (\nu_\theta + \mu_{\theta,i})](y'_{\theta,i} - y_{\theta,i}) \geq 0.$$

By (7c) and $\sum_{s_{\theta,i} \in \mathcal{S}_\theta} y_{\theta,i} = d_\theta$, we have the following:

$$\sum_{\theta \in \Theta} (\nu'_\theta - \nu_\theta)(d'_\theta - d_\theta) \geq \sum_{\theta \in \Theta} \sum_{s_{\theta,i} \in \mathcal{S}_\theta} (\mu'_{\theta,i} y_{\theta,i} + \mu_{\theta,i} y'_{\theta,i}).$$

Note that $\mu'_{\theta,i}, y_{\theta,i}, \mu_{\theta,i}, y'_{\theta,i} \geq 0$, we complete the proof. □

The result of Proposition 2 can also be written as

$$\left[\sum_{e \in s_{\theta,i}} c_e(x'_e) - c_e(x_e) \right] (d'_\theta - d_\theta) \geq 0, \quad \forall s_{\theta,i} \in \mathcal{S}_\theta$$

with $y'_{\theta,i}, y_{\theta,i} > 0$. The Proposition 2 states that if one demand d_θ is increased by fake users, with other demands remaining the same, then the equilibrium cost ν_θ perceived by the NRS for the user u with OD pair $\theta_u = \theta$ is also increased.

Proposition 3: For $W(\mathbf{d})$ with demand $\mathbf{d}$, let $\mathbf{x}$ be a WE corresponds to cost $c_e(\cdot)$ and $\mathbf{x}'$ be a WE corresponds to cost $c'_e(\cdot)$, then $\left[c'_e(x_e) - c_e(x_e)\right](x'_e - x_e) \leq 0$ and $\left[c'_e(x'_e) - c_e(x'_e)\right](x'_e - x_e) \leq 0$.

Proof: Under the assumption that the cost functions $c_e(\cdot)$ and $c'_e(\cdot)$ are continuous and increasing in x_e and x'_e, respectively, with the optimality conditions for $\mathbf{x}'$ be a WE corresponds to cost function $c'_e(\cdot)$ and $\mathbf{x}$ be a WE corresponds to $c_e(\cdot)$, we have the following inequalities:

$$\left[c_e(x'_e) - c_e(x_e)\right](x'_e - x_e) \geq 0 \tag{8a}$$

$$\left[c'_e(x'_e) - c'_e(x_e)\right](x'_e - x_e) \geq 0 \tag{8b}$$

$$c_e(x_e)(x'_e - x_e) \geq 0 \tag{8c}$$

$$c'_e(x'_e)(x_e - x'_e) \geq 0 \tag{8d}$$

Then, summing up (8a), (8c), and (8d) gives us $\left[c_e(x'_e) - c'_e(x'_e)\right](x'_e - x_e) \geq 0$, while adding (8b), (8c), and (8d) leads us to $\left[c_e(x_e) - c'_e(x_e)\right](x'_e - x_e) \geq 0$. We complete the proof. □

That is, Proposition 3 demonstrates that an increasing cost on a road $e \in \mathcal{E}$ will cause the equilibrium load x_e on that road to decrease. This reduced load can be interpreted as a redistribution to alternative feasible paths. Thus, the attacker can achieve the desired flow load level γ on the target road e' by redistributing the load there. Alongside Proposition 2, this can be accomplished by strategically increasing the perceived demands (by adding non-existent ones) on certain roads, thereby raising their costs and leading to redistribution.

D. Impact Metrics for Risk Reports

In this subsection, we introduce two metrics as the outcomes of our PRADA framework. Let $\mathbf{p}$ be the recommendation to all the users without attack and $\mathbf{p}'$ is the one under attack.

1) Local-Targeted Impact: We define the *targeted impact metric (TI)* as the difference in traffic flow on each road with and without the demand attack, divided by the flow without the attack. Specifically, for each road $e \in \mathcal{E}$, the measure TI_e is given by:

$$TI_e = \frac{|x_e^r(\mathbf{p}') - x_e^r(\mathbf{p})|}{x_e^r(\mathbf{p})}, \tag{9}$$

which can assists in measuring the percentage change in traffic flow on specific or targeted road affected by the demand attack. A larger value of TI_e indicates the road e is influenced more, often implying higher risk under the attack.

2) Network-Wide Impact: Given the metrics $TI_e, \forall e \in \mathcal{E}$, we define the *network impact metric (NI)* as the mean of TI_e across all the roads/edges within the network. The measure NI is as follows:

$$NI = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} TI_e = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \frac{|x_e^r(\mathbf{p}') - x_e^r(\mathbf{p})|}{x_e^r(\mathbf{p})}. \tag{10}$$

The metric NI allows us to evaluate the percentage change in traffic flow across the entire network. It is important to note that if the demand attack primarily affects traffic flow on roads within a small area, as indicated by the TI_e values for those roads, this localized impact will be averaged out when considering the network-wide impact.

V. Risk Mitigation

In this section, we aim to explore an effective mechanism to mitigate the impact of misinformed demand attacks.

A. Mitigation Through User Trust

Denote $\mathbf{p}_u^o = \{p_{u,i}^o\}_{s_{u,i}\in\mathcal{S}_u} \in \mathcal{P}_u$ as the UE recommendation defined in Definition 1 for user u without attack (under normal traffic condition), which can be obtained from previous experience of requesting recommendation for the same OD pair. Let $T_u \in \mathbb{R}_{\geq 0}$ for all $u \in \mathcal{U}$ represent the *trust score*, which quantifies the degree to which users trust that some malicious entities do not manipulate the recommendations provided by the NRS. The trust score then leads to the following trust constraint, implying that the user trusts the recommendation $\mathbf{p}_u \in \mathcal{P}_u$ because its deviation from the user's previous experience is within an acceptable range (i.e., T_u).

Definition 3 (Trust Constraint (TC)): Consider an NRS component, denoted as $\mathscr{R}$. A recommended mixed strategy $\mathbf{p}_u \in \mathcal{P}_u$ for a user u is said to satisfy the trust constraint if the distance (in terms of Kullback–Leibler (KL) divergence) between the currently and previous recommended strategy $\mathbf{p}_u^o \in \mathcal{P}_u$ is less than the trust score $T_u \in \mathbb{R}_{\geq 0}$:

$$D(\mathbf{p}_u \| \mathbf{p}_u^o) = \sum_{i=1}^{k_u} p_{u,i} \log\left(\frac{p_{u,i}}{p_{u,i}^o}\right) \leq T_u. \tag{11}$$

A higher trust score indicates greater user trust in the current integrity of the NRS, as its associated trust constraint (11) forms a larger *trust region*. This means the user is more willing to tolerate larger differences between current and previous recommendations, believing that such variations are due to changes in traffic conditions rather than malicious demand manipulation. In contrast, $T_u = 0$ indicates a lack of trust in the current NRS, leading the user to follow only recommendations that match their past experiences.

In practice, from user u's perspective, the trust score T_u can be calculated from a weighted sum of factors contributing to the trustworthiness of the NRS. These factors may include the reputation of the NRS, which is influenced by the reliability of its data sources and algorithms, particularly in the face of potential attacks or data poisoning. The engagement level of the user's family and friends with the NRS can also enhance trust, as consistent usage by close contacts often increases confidence. In addition, historical experience is another important factor, as alignment of past recommendations with user needs and preferences also builds trust. From the NRS's perspective, the trust score T_u for user u can be determined in several ways: directly reported by the user, estimated by the NRS using the factors mentioned above, or adaptively learned by the NRS through continuous interaction and feedback from the user.

Moreover, let $\mathbf{p}_u$ be the recommendation to user u without attack and $\mathbf{p}'_u$ is the one manipulated by the attacker. If each element $p'_{u,i} = p_{u,i} + \epsilon_{u,i}$, where $\epsilon_{u,i}$ is a small perturbation due to demand attacks, we have the following sensitivity property for the manipulated recommendation using first-order Taylor expansion.

$$D(\mathbf{p}'_u \| \mathbf{p}_u^o) - D(\mathbf{p}_u \| \mathbf{p}_u^o) \approx \sum_{i=1}^{k_u} \epsilon_{u,i}\left(\log\frac{p_{u,i}}{p_{u,i}^o} + 1\right).$$

From the attacker's point of view, in order to fulfill TC, $D(\mathbf{p}'_u \| \mathbf{p}_u^o) \leq T_u$, so that user u is still willing follow the manipulated recommendation $\mathbf{p}'_u$, the perturbations $\epsilon_{u,i}, \forall s_{u,i} \in \mathcal{S}_u$ must satisfy:

$$D(\mathbf{p}_u \| \mathbf{p}_u^o) + \sum_{i=1}^{k_u} \epsilon_{u,i}\left(\log\frac{p_{u,i}}{p_{u,i}^o} + 1\right) \leq T_u,$$

which indicates that the trust score T_u bounds the total perturbation to the probabilities $p_{u,i}$ on feasible paths $s_{u,i} \in \mathcal{S}_u$.

B. Trust Mechanism for NRS

Since TC bounds and mitigates the severity of manipulation caused by demand attacks, it is reasonable to incorporate such a user trust mechanism into the navigation recommendation process. In this context, users are either learned or warned to follow only the recommendations that satisfy their TC. Then, with Definition 1 and 3, the NRS must identify feasible recommendations that can be trusted by all users, ensuring their participation and adherence to the recommended strategies. This leads to the following definition.

Definition 4 (Trusted Recommendation): Considering a routing game addressed by the NRS defined as $\Gamma^r = \langle \mathscr{R}, \mathscr{F}^r \rangle$, a trusted recommendation profile to all users $\mathbf{p} \in \mathcal{P}$ needs to satisfy:

$$F_u^r(\mathbf{p}_u, \mathbf{p}_{-u}) - F_u^r(\mathbf{p}'_u, \mathbf{p}_{-u}) \leq 0, \quad \forall\ \mathbf{p}'_u \in \mathcal{P}_u,\ \forall u \in \mathcal{U}, \tag{12a}$$

$$D(\mathbf{p}_u \| \mathbf{p}_u^o) - T_u \leq 0, \quad \forall u \in \mathcal{U}, \tag{12b}$$

Incorporating trust constraints ensures that the current recommendation provided by the NRS cannot deviate significantly from the previous one for the same OD pair, based on the assumption that traffic conditions usually evolve smoothly. Consequently, if there is a sudden change in demand caused by malicious entities, the recommendation will stay relatively aligned with past recommendations in normal circumstances.

The NRS's problem of finding trusted recommended mixed strategies for all users can also be interpreted using a non-cooperative game, defined as $\tilde{\Gamma}^r := (\mathscr{R}, \mathscr{F}^r, (T_u)_{u\in\mathcal{U}})$, where $\mathscr{F}^r = (F_u^r)_{u\in\mathcal{U}}$ and F_u^r is expressed in (1). Each user $u \in \mathcal{U}$ of the NRS is a player of the game $\tilde{\Gamma}^r$. User u aims to minimize his/her own expected cost F_u^r by deciding a mixed strategy $\mathbf{p}_u \in \mathcal{P}_u$ over feasible path choice set $\mathcal{S}_u$ given other users' strategies $\mathbf{p}_{-u}$, under the trust constraint that $\mathbf{p}_u$ cannot deviate too much from previous experience $\mathbf{p}_u^o$. That is, for all user $u \in \mathcal{U}$ in $\tilde{\Gamma}^r$, given other users' strategies $\mathbf{p}_{-u}$,

$$\begin{aligned} \mathrm{OP}_u:\ \min_{\mathbf{p}_u\in\mathcal{P}_u}\quad & F_u^r(\mathbf{p}_u, \mathbf{p}_{-u}) \\ \text{s.t.}\quad & D(\mathbf{p}_u \| \mathbf{p}_u^o)) - T_u \leq 0. \end{aligned} \tag{13}$$

Then, by denoting $C'_{u,i}(\mathbf{p}) = \sum_{e\in s_{u,i}} t_e\left[1 + \alpha\left(\frac{x_e^r(\mathbf{p})}{k_e}\right)^\beta + \beta x_e^u(\mathbf{p}_u)\frac{\alpha}{k_e}\left(\frac{x_e^r(\mathbf{p})}{k_e}\right)^{\beta-1}\right]$, we have the following proposition.

Proposition 4: Consider the problem defined in (12). Under the conditions that for all $u \in \mathcal{U}$, the expected cost F_u is continuously differentiable in $\mathbf{p} \in \mathcal{P}$ and convex in $\mathbf{p}_u \in \mathcal{P}_u$, and that the trust constraint is active, the trusted recommendation $\mathbf{p}^*$ is as follows: $\forall u \in \mathcal{U}$,

$$p_{u,i}^* = \frac{p_{u,i}^o \exp\left(\frac{-C'_{u,i}(\mathbf{p})}{\lambda_u} - 1\right)}{\mu'_{u,i}}, \forall s_{u,i} \in \mathcal{S}_u, \tag{14}$$

where $\mu'_u = \sum_{i=1}^{k_u} p^*_{u,i}$ is the normalization term and $\lambda_u \in \mathbb{R}_+$ is the Lagrange multiplier for the trust constraint.

Proof: Let $\mathscr{L}_u(\mathbf{p}_u, \lambda_u, \mu_u)$ as follows denote the Lagrangian of user u's optimization problem $\mathrm{OP_u}$:

$$\begin{aligned}\mathscr{L}_u(\mathbf{p}_u, \lambda_u, \mu_u) = &\sum_{i=1}^{k_u} p_{u,i} C_{u,i}(\mathbf{p}_u, \mathbf{p}_{-u}) \\ &+ \lambda_u \left[\sum_{i=1}^{k_u} p_{u,i} \log\left(\frac{p_{u,i}}{p^o_{u,i}}\right) - T_u\right] \\ &- \mu_u \sum_{i=1}^{k_u} p_{u,i},\end{aligned}$$

where $\lambda_u \in \mathbb{R}_+$, $\mu_u \in \mathbb{R}_+$ are the Lagrange multipliers. We consider $C_{u,i}(\mathbf{p}_u, \mathbf{p}_{-u}) = \sum_{e \in s_{u,i}} t_e \left(1 + \alpha \left(\frac{x^r_e(\mathbf{p})}{k_e}\right)^\beta\right)$, and the first-order condition $\partial \mathscr{L}_u / \partial p_{u,i} = 0$ for each $p_{u,i}$ becomes:

$$\begin{aligned}\sum_{e \in s_{u,i}} t_e \left[1 + \alpha \left(\frac{x^r_e(\mathbf{p})}{k_e}\right)^\beta + \beta x^u_e(\mathbf{p}_u) \frac{\alpha}{k_e} \left(\frac{x^r_e(\mathbf{p})}{k_e}\right)^{\beta-1}\right] \\ + \lambda_u \left[\log\left(\frac{p_{u,i}}{p^o_{u,i}}\right) + 1\right] - \mu_u = 0.\end{aligned}$$

By letting $\log(\mu'_u) = -\mu_u / \lambda_u$, then

$$\frac{C'_{u,i}(\mathbf{p})}{\lambda_u} + \log\left(\frac{p_{u,i}}{p^o_{u,i}}\right) + 1 + \log(\mu'_u) = 0.$$

Therefore, for all $u \in \mathcal{U}$, $s_{u,i} \in \mathcal{S}_u$, we have each

$$p^*_{u,i} = \frac{p^o_{u,i} \exp\left(\frac{-C'_{u,i}(\mathbf{p})}{\lambda_u} - 1\right)}{\mu'_u},$$

where $\mu_u, \forall u \in \mathcal{U}$ are normalizations ensuring $\mathbf{p}^* \in \mathcal{P}$. □

Under stable conditions, for each $s_{u,i} \in \mathcal{S}_u$, $u \in \mathcal{U}$ being used, $C'_{u,i}(\mathbf{p})$ remains identical, and $\mathbf{p}$ remains the same as $\mathbf{p}^o$. However, if there is a demand attack, each $C'_{u,i}(\mathbf{p})$ perceived by the NRS will differ, causing $\mathbf{p}$ to deviate from $\mathbf{p}^o$ and tilt towards paths with lower perceived costs (which may not be the true costs) caused by misinformed demand. Additionally, the extent of deviation from previous $\mathbf{p}^o$ to current $\mathbf{p}$ depends on the multipliers λ_u, which are associated with the trust score T_u for each $u \in \mathcal{U}$. Let $\mathbf{p}^*_u$ denote the trusted recommendation for user u in Proposition 4, the optimal λ^*_u can be found by numerically evaluating the dual function defined below.

$$\begin{aligned}\mathscr{G}_u(\lambda_u) &:= \min_{\mathbf{p}_u \in \mathcal{P}_u} F^r_u(\mathbf{p}_u, \mathbf{p}_{-u}) + \lambda_u \left[D(\mathbf{p}_u \| \mathbf{p}^o_u) - T_u\right] \\ &= \sum_{i=1}^{k_u} p^*_{u,i} C_{u,i}(\mathbf{p}^*_u, \mathbf{p}_{-u}) \\ &\quad + \lambda_u \left[\sum_{i=1}^{k_u} p^*_{u,i} \log\left(\frac{p^*_{u,i}}{p^o_{u,i}}\right) - T_u\right] \qquad (15)\end{aligned}$$

which leads us to the dual problem of $\mathrm{OP_u}$ as follows:

$$\mathrm{DOP}_u : \max_{\lambda_u \in \mathbb{R}_+} \mathscr{G}_u(\lambda_u) \qquad (16)$$

Note that Proposition 4 considers the situation that the trust score T_u is carefully determined so that TC is active with $\lambda_u > 0$. When $\lambda_u = 0$, which suggests that TC is non-binding, potentially due to the UE recommendation $\mathbf{p}_u$ defined in Definition 1 under current traffic condition is close to the previous experienced $\mathbf{p}^o_u$, or because the user has high confidence in the current recommendation (i.e., T_u is large), the NRS can then recommend the user with $\mathbf{p}^*_u \in \arg\min_{\mathbf{p}_u \in \mathcal{P}_u} F^r_u(\mathbf{p}_u, \mathbf{p}_{-u})$.

Algorithm 1 Trust Mechanism

1: **Input** NRS component $\mathscr{R} = \langle \mathcal{G}, (c_e(\cdot))_{e \in \mathcal{E}}, \mathcal{U}, (\mathcal{S}_u)_{u \in \mathcal{U}} \rangle$
2: **Collect** trust scores T_u from all the users $u \in \mathcal{U}$
3: **Obtain** $\mathbf{p}^o_u, \forall u \in \mathcal{U}$ from historical data
4: **Initialize** recommendation $\mathbf{p}$ based on (3)
5: **for** $u \in \mathcal{U}$ **do**
6: **if** TC for u non-binding **then**
7: $\mathbf{p}^*_u = \mathbf{p}_u$
8: **else**
9: $p^*_{u,i} = \frac{p^o_{u,i} \exp\left((-C'_{u,i}(\mathbf{p})/\lambda_u) - 1\right)}{\mu'_{u,i}}, \forall s_{u,i} \in \mathcal{S}_u$
10: $\lambda^*_u \in \arg\max_{\lambda_u \in \mathbb{R}_+} \mathscr{G}_u(\lambda_u)$
11: **end if**
12: **end for**
13: **Return** $\mathbf{p}^*$ to users and PRADA risk evaluator

To this end, the proposed trust mechanism can be summarized by the following Algorithm 1.

In practice, recognizing the vulnerabilities illustrated in Fig. 1, an NRS can consult the PRADA risk evaluator to assess risks for threat profiles from attack libraries. If the risk metrics TI and NI in the reports surpass the company's standards, one approach for the NRS to mitigate these risks is to collect user trust scores and implement a trust mechanism. The PRADA risk evaluator can then use the trust recommendation $\mathbf{p}^*$ from Algorithm 1 to reassess the risks and ensure they align with the company's standards.

C. Sensitivity Analysis for Trust Mitigation

In this subsection, we aim to examine the relationship between the multiplier λ_u, the optimal value for problem (12), and the user's trust score T_u. Suppose T_u is changed to T'_u (due to positive or negative news related to the NRS), where $T'_u = T_u + \eta_u$, the TC then becomes

$$D(\mathbf{p}_u \| \mathbf{p}^o_u) - T_u \le \eta_u. \qquad (17)$$

Proposition 5: Consider the optimization problem $\mathrm{OP_u}$. Under the assumptions that F_u is convex in $\mathbf{p}_u$ and T_u, T'_u are chosen so that the Slater's condition holds, let $v^*_{u,0}$ and v^*_{u,η_u} denote the optimal value for $\mathrm{OP_u}$ associated with T_u, and T'_u, respectively, and let λ^*_u represent the optimal dual variable for $\mathrm{OP_u}$ associated with T_u, then

$$v^*_{u,t_u} \ge v^*_{u,0} - \lambda^*_u \eta_u. \qquad (18)$$

Proof: Suppose that $\mathbf{p}_u \in \mathcal{P}_u$ is any feasible point for $\mathrm{OP_u}$ associated with T'_u, then by strong duality

$$\begin{aligned}v^*_{u,0} = \mathscr{G}_u(\lambda^*_u) &\le F^r_u(\mathbf{p}_u, \mathbf{p}_{-u}) + \lambda^*_u \left[D(\mathbf{p}_u \| \mathbf{p}^o_u) - T_u\right] \\ &\le v^*_{u,\eta_u} + \lambda^*_u \eta_u,\end{aligned}$$

which completes the proof. □

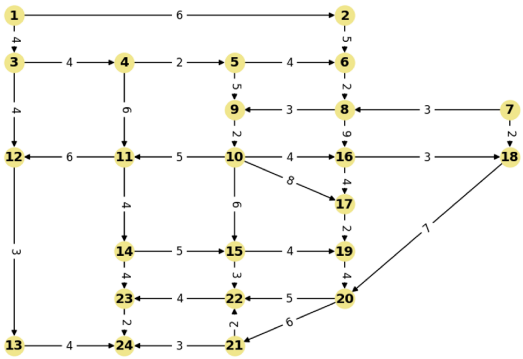


Fig. 4. An example network (based on the network structure of Sioux Falls) for our case study. The value on each edge denotes the free-flow road travel time.

TABLE I
CASE STUDY SETUP

OD pair	User #	Feasible paths
2-17	1 to 20	Path 1: 2-6-8-9-10-16-17 Path 2: 2-6-8-9-10-17 Path 3: 2-6-8-16-17
9-19	21 to 40	Path 1: 9-10-11-14-15-19 Path 2: 9-10-15-19 Path 3: 9-10-16-17-19 Path 4: 9-10-17-19

BPR cost with $\alpha = 0.4$, $\beta = 2$, $k_e = 50, \forall e \in \mathcal{E}$.

Proposition 5 indicates that if λ_u^* is large and user u shrinks his/her trust region (i.e., $\eta_u < 0$), then the optimal value of user u's expected cost becomes much higher. Conversely, if λ_u^* is small, even if user u expands his/her trust region (i.e., $\eta_u > 0$), the optimal value of user u's expected cost does not decrease substantially. That is, when λ_u^* is large (small), the optimal value of user u's expected cost is more (less) sensitive to changes in the trust region. Thus, λ_u^* can be interpreted as a *risk factor*. A larger λ_u^* indicates a higher risk for user u in modifying the trust region, as it leads to more significant changes in the user's optimal expected cost.

It is essential for users to understand that the trust score must be carefully determined due to the trade-offs between low and high values. For instance, if a user has low confidence in the NRS, resulting in a very small T_u, the recommendation will closely follow the user's past experiences. While this minimizes the impact of potential demand attacks, it also means the user may lose the chance to adapt to gradually changing traffic conditions if no attack occurs. On the other hand, a high trust score allows users to receive recommendations that reflect the latest traffic conditions, optimizing their travel time when there is no attack. However, this high trust also increases susceptibility to demand attacks, potentially leading to more significant manipulation of their recommendations.

VI. DISCUSSION THROUGH CASE STUDY

We use the traffic network abstracted in Fig. 4 as a case study of our PRADA framework, where we adopt the structure from the Sioux Falls network [50], and utilize the BPR function for the cost $c_e(\cdot)$ on each road $e \in \mathcal{E}$ with parameters $\alpha = 0.4$, $\beta = 2$, and $k_e = 50$ for simplicity. The number displayed on each edge represents the free-flow time cost t_e. Then, we focus on the case where 20 users seeking to travel from node 2 to 17 (OD 2-17) and other 20 users from node 9 to 19 (OD 9-19). The feasible path set for users is specified in Table I.

A. Risks Under Different Attacker Models

In the context of a misinformed demand attack, attack methods (1)–(5) in subsection III-A lead to fabricated user demands for a set $\mathcal{K} \subset \mathcal{V} \times \mathcal{V}$ of distinct OD pairs. We consider the case where the attacker has a local-targeted objective, and aims to make NRS recommend a level of $\gamma = 20$ flow load from authentic users passing $(10, 17)$, the target road. (The flow load without attack is 12, originally.) We compare the risk in terms of TI and NI of the following types of attackers in Fig. 5. This risk report provides the PRADA risk evaluator with a holistic overview, highlighting the attacks that require the most attention and urgent mitigation.

1) Strategic Attacker: A strategic attacker who has the knowledge of NRS can identify the desired fake demand levels by solving the leader-follower problem in section IV-C. The expected flow on the target road $(10, 17)$ caused by authentic users meets the desired level $\gamma \geq 20$ by generating a total of less than 35 non-existent demands within the traffic network.

2) Non-Strategic Attacker: A non-strategic attacker may not know how NRS generates the recommendation for users. Hence, the uniform attacker evenly distributes the total demand across all OD pairs near the target road, while the random attacker distributes the demand randomly among the OD pairs. Note that for comparison, the non-strategic attackers are restricted to allocating the same amount of demands, totaling 35, as the optimal strategic attacker.

Fig. 5 shows the risk report in terms of network-wide impact (NI) and local-targeted impact (TI) on roads/edges along users' feasible paths. First, we can observe that none of the attack scenarios affect roads $(2, 6)$ and $(6, 8)$. Since true users with the OD pair 2-17 must travel through these two roads to reach their destination (as all three feasible paths include these roads), the attacker cannot impact these roads by redistributing users through fake demands. This suggests that the road users must go through is at lower risk of demand attack, as it is hard for malicious entities to influence the flow load by fabricating non-existent demands. Moreover, the alternative roads to the target road $(10, 17)$, including $(10, 15)$, $(10, 16)$, and $(16, 17)$, are at higher risk. This is because the attacker achieves their goal by manipulating the NRS to redistribute users originally passing through these roads to the target road, which also illustrates the analysis in Section IV-C2.

However, we can observe that the local-targeted impacts (TI) on roads $(8, 9)$ and $(8, 16)$ are lower under strategic attacks compared with non-strategic attacks. This is because non-strategic attacks may accidentally cause a greater increase (or decrease) in traffic on the road $(8, 9)$ and a more significant decrease (or increase) in traffic on the road $(8, 16)$ than a

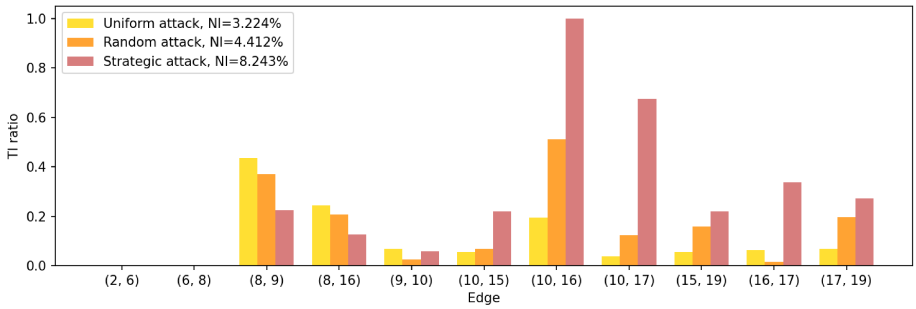


Fig. 5. Risk report in terms of TI (local-targeted impact on roads along users' feasible paths) and NI (network-wide impact) when encountering non-strategic (random, uniform) and strategic attackers.

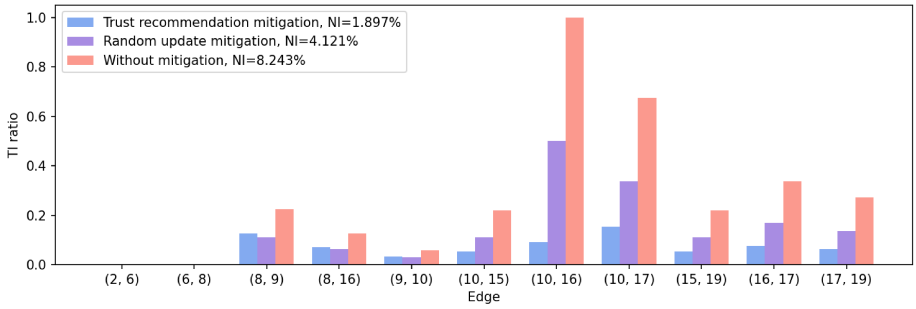


Fig. 6. Risk report in terms of TI and NI when adopting mitigation methods (random update and trusted recommendation) under strategic attacks. The trusted recommendation mitigation can effectively reduce both the severe local-targeted and network-wide impacts compared to the random update one.

strategic attack, potentially leading to higher risks. In contrast, the risks associated with the target road $(10, 17)$ from a non-strategic attack may not be as high as those from a strategic attack, as the traffic changes from OD pair 9-19 can accidentally offset those from OD pair 2-17 under a non-strategic attack, reducing the overall impact on the road $(10, 17)$. Lastly, the network-wide impact (NI) indicates that the risks posed by strategic attackers are higher compared to non-strategic ones. This heightened risk in the risk report emphasizes the urgent need for mitigation strategies against intelligent attackers.

B. Potential Mitigation of the Risk

In this subsection, we aim to assess the risk when the NRS adopts the trusted recommendation described in Definition 4. To evaluate whether such a trust mechanism can effectively mitigate the impact of misinformed demand attacks, we compare it with a straightforward random update mitigation method and a scenario without any mitigation. The risk report associated with these mitigation methods is shown in Fig. 6. This report aids the PRADA risk evaluator in determining the most efficient mitigation mechanism.

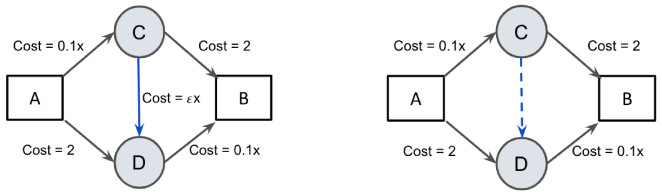
1) Trusted Recommendation Mitigation: In practice, users may not adhere to recommendations that differ significantly from previously received recommendations for the same OD pair. Therefore, to ensure user compliance, the NRS incorporates user trust constraints into its recommendations, called trusted recommendations. Such a trust mechanism ensures that the current recommendation does not deviate significantly from the previous one for the same OD pair. The degree of deviation allowed in the current recommendation depends on the user's trust score T_u. A higher trust score indicates greater user trust in the integrity of NRS, with the user interpreting deviations as responses to sudden changes in traffic conditions rather than malicious demand attacks. In this context, the current recommendation for all users is given based on Algorithm 1.

2) Random Update Mitigation: In practice, not all users receive the updated recommendations simultaneously. Therefore, we consider the random update algorithm that may assist in mitigating the risk of sudden changes in demands perceived by the NRS caused by demand attacks. In this context, each user gets his/her current recommendation $\mathbf{p}_u$ with a predefined probability $0 < \pi_u < 1$; otherwise remains $\mathbf{p}_u^o$. That is,

$$\mathbf{p}_u = \begin{cases} \mathbf{p}_u \text{ satisfying (1)}, & \text{w.p. } \pi_u, \\ \mathbf{P}_u^o, & \text{w.p. } 1-\pi_u. \end{cases} \tag{19}$$

The probability of update π_u may vary based on the user's driving habits or the capabilities of V2X technologies within the region containing the user's origin and destination. Here, we consider $\pi_u = 0.5, \forall u \in \mathcal{U}$ for simplicity.

Fig. 6 shows the risk report in terms of NI and TI on roads/edges along users' feasible paths when the NRS adopts mitigation methods. With $\pi_u = 0.5$ for all users $u \in \mathcal{U}$, the random update mitigation method reduces the risk from strategic attacks by half. It is important to note that a lower π_u could potentially decrease risks further, but it may also result in users losing access to the most recent recommendations based on current traffic conditions. Noticing this, we can consider a more complex scenario where the probability of receiving updated recommendation π_u for all $u \in \mathcal{U}$ can be



(a) Without attack (b) Under attack

Fig. 7. A carefully crafted example for the discussion on the "Resilience Paradox".

dynamically adjusted (updated), which provides more flexibility in balancing the trade-off between the risk mitigation and the most recent recommendations. For instance, π_u can be increased during peak hours, in areas with higher driving risks, or in response to sudden changes in traffic patterns. Besides, π_u can also be adjusted based on user feedback regarding the timeliness of updates. If a user expresses dissatisfaction with delayed recommendations, π_u for that user can be increased to improve their experience. In addition to adjusting π_u, synchronization methods can assist in meeting real-time needs. One viable approach is to synchronize the updates within specific time windows, determined by factors like geographical areas or traffic conditions. For example, during high-impact events like accidents or road closures, shorter synchronization windows may be necessary. Another method is setting thresholds for critical traffic conditions that trigger immediate updates, ensuring users receive timely recommendations in urgent situations.

As for the proposed trusted recommendation in Fig. 6, it can effectively mitigate risk by constraining flow load changes based on user trust scores. Additionally, comparing the road $(10, 16)$ with the road $(8, 16)$, we can observe that the trusted recommendation mitigation is more obvious when the roads are originally facing higher risks of demand attack.

C. Discussion on the Resilience Paradox

To this end, a natural question is: *Can the locally targeted attack lead to a better overall outcome (total travel time costs for users) in some situations?* We begin with a carefully crafted example using the classical Braess' network [51], which illustrates Braess's Paradox, a well-documented phenomenon where adding roads to a network can sometimes degrade overall traffic performance. Conversely, removing roads from a network can, in certain cases, improve performance.

Within the transportation network shown in Fig. 7, there are 30 users aiming to go from node A to node B, and the ϵ is small enough so that the cost on C-D is close to 0 even though all 30 users are passing through. Before the attack (illustrated in Fig. 7a), the RS will recommend a mixed strategy $(1/3, 1/3, 1/3)$ on path A-C-B, A-C-D-B, and A-D-B, respectively. The overall costs on these three paths are all 4, which leads to a total travel time cost of 120 for users. Suppose the attacker wants more "users" to pass D-B by fabricating a large demand on C-D to make C-D seem congested to the RS, as in Fig. 7b. The RS will recommend a strategy $(0.5, 0, 0.5)$ on paths A-C-B, A-C-D-B, and A-D-B, respectively. The overall costs on A-C-B and A-D-B are both 3.5, which leads to the total travel time cost for users becoming 105. Therefore, we can conclude that the cost under attack is better than the performance without attack in this carefully crafted example. This points out that the local-targeted attack is a potential aspect worth further investigation.

VII. Conclusion

This paper assesses the risk of potential informational attacks on navigational recommendation systems (NRS). We introduce the attack methods and identify vulnerabilities that attackers can exploit to launch demand attacks, achieving locally targeted goals that benefit certain groups or businesses. Then, we propose a holistic framework for proactive risk assessment of demand attacks (PRADA) that integrates necessary elements. Given that modern attackers are often intelligent, our focus, from the perspective of the PRADA risk evaluator, is on strategic attacks. We analyze the interaction between the attacker and the incentive-compatible NRS through a Stackelberg game. Our study indicates that users are at high risk when facing strategic attacks that target specific roads by creating non-existent demands for OD pairs with alternative path options. To mitigate these risks, we introduce a trust mechanism, and our investigation shows that it is a viable approach to reducing the risk posed by misinformed demand attacks in both local-targeted and network-wide senses.

While the proposed PRADA framework captures the core interactions between the attacker and the NRS, it relies on modeling assumptions in order to facilitate tractable analysis. Hence, it may not fully reflect real-world complexities such as user bounded rationality or attackers with different preferences. Extending the framework to incorporate these broader behavioral models can be one important direction for future work. In addition, according to the case study, the trusted recommendation mechanism can improve resilience against demand attacks. However, the mechanism itself may also be targeted by sophisticated adversaries. For example, attackers could poison the trust score estimation or reduce user confidence through misinformation. Therefore, exploring attack-aware trust mechanisms for risk mitigation can be another future direction. One of the other possible directions will be investigating different scenarios, such as impacts of misinformed traffic conditions (costs) attacks on NRS.

References

[1] Z. Lv, R. Lou, and A. K. Singh, "AI empowered communication systems for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4579–4587, Jul. 2021.

[2] F. Zantalis, G. Koulouras, S. Karabetsos, and D. Kandris, "A review of machine learning and IoT in smart transportation," *Future Internet*, vol. 11, no. 4, p. 94, Apr. 2019.

[3] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3152–3168, Aug. 2020.

[4] A. Haydari and Y. Yilmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 11–32, Jan. 2022.

[5] M. van Essen, T. Thomas, E. van Berkum, and C. Chorus, "From user equilibrium to system optimum: A literature review on the role of travel information, bounded rationality and non-selfish behaviour at the network and individual levels," *Transp. Rev.*, vol. 36, no. 4, pp. 527–548, Jul. 2016.

[6] S. Das, E. Kamenica, and R. Mirka, "Reducing congestion through information design," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 1279–1284.

[7] D. K. Goldenberg, L. Qiuy, H. Xie, Y. R. Yang, and Y. Zhang, "Optimizing cost and performance for multihoming," in *Proc. Conf. Appl., Technol., architectures, protocols Comput. Commun.*, Aug. 2004, pp. 79–92, doi: 10.1145/1015467.1015478.

[8] Z. Zhang, M. Zhang, A. Greenberg, Y. C. Hu, R. Mahajan, and B. Christian, "Optimizing cost and performance in online service provider networks," in *Proc. NSDI*, 2010, p. 3.

[9] F. Rossi, R. Zhang, Y. Hindy, and M. Pavone, "Routing autonomous vehicles in congested transportation networks: Structural properties and coordination algorithms," *Auto. Robots*, vol. 42, no. 7, pp. 1427–1442, Oct. 2018.

[10] K. Kollias, A. Chandrashekharapuram, L. Fawcett, S. Gollapudi, and A. K. Sinop, "Weighted Stackelberg algorithms for road traffic optimization," in *Proc. 29th Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2021, pp. 57–68, doi: 10.1145/3474717.3483652.

[11] Y. Ning and L. Du, "Robust and resilient equilibrium routing mechanism for traffic congestion mitigation built upon correlated equilibrium and distributed optimization," *Transp. Res. B, Methodol.*, vol. 168, pp. 170–205, Feb. 2023.

[12] Y.-T. Yang, H. Lei, and Q. Zhu, "Strategic information attacks on incentive-compatible navigational recommendations in intelligent transportation systems," 2023, *arXiv:2310.01646*.

[13] T. Mecheva and N. Kakanakov, "Cybersecurity in intelligent transportation systems," *Computers*, vol. 9, no. 4, p. 83, Oct. 2020.

[14] M. Waniek, G. Raman, B. AlShebli, J. C.-H. Peng, and T. Rahwan, "Traffic networks are vulnerable to disinformation attacks," *Sci. Rep.*, vol. 11, no. 1, p. 5329, Mar. 2021.

[15] (2016). *Waze To Go: Residents Fight Off Crowdsourced Traffic for a While*. [Online]. Available: https://nakedsecurity.sophos.com/2016/06/07/waze-to-go-residents-fight-off-crowdsourced-traffic-for-a-while/

[16] C. Eryonucu and P. Papadimitratos, "Sybil-based attacks on Google maps or how to forge the image of city life," in *Proc. 15th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, May 2022, pp. 73–84.

[17] (2022). *Information Security, Cybersecurity and Privacy Protection—Guidance on Managing Information Security Risks*. [Online]. Available: https://www.iso.org/standard/80585.html

[18] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: An overview from machine learning perspective," *J. Big Data*, vol. 7, no. 1, pp. 1–29, Dec. 2020.

[19] S. Ramasubramanian, "Introduction to game theory," *Game Theory Mach. Learn. for Cyber Secur.*, vol. 9, no. 3, p. 81, 2004.

[20] D. Hahn, A. Munir, and V. Behzadan, "Security and privacy issues in intelligent transportation systems: Classification and challenges," *IEEE Intell. Transp. Syst. Mag.*, vol. 13, no. 1, pp. 181–196, Spring. 2021.

[21] N. Huq, R. Vosseler, and M. Swimmer, "Cyberattacks against intelligent transportation systems," Trend Micro, Tech. Rep., 2017.

[22] D. Fletcher and P. Bye, "Cybersecurity in transit systems," Nat. Academies Press, Washington, DC, USA, Tech. Rep. TCRP Project J-07, Topic SA-50, 2022, doi: 10.17226/26475.

[23] S. A. Almalki and J. Song, "A review on data falsification-based attacks in cooperative intelligent transportation systems," *Int. J. Comput. Sci. Secur.*, vol. 14, p. 22, Mar. 2020.

[24] Y. Pan and Q. Zhu, "On poisoned Wardrop equilibrium in congestion games," in *Proc. Int. Conf. Decis. Game Theory Secur.*, 2023, pp. 191–211.

[25] Y. Cao, Q. Luo, and J. Liu, "Road navigation system attacks: A case on GPS navigation map," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jan. 2019, pp. 1–5.

[26] G. Wang, B. Wang, T. Wang, A. Nika, H. Zheng, and B. Y. Zhao, "Ghost riders: Sybil attacks on crowdsourced mobile mapping services," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1123–1136, Jun. 2018.

[27] J. Lin, W. Yu, N. Zhang, X. Yang, and L. Ge, "Data integrity attacks against dynamic route guidance in transportation-based cyber-physical systems: Modeling, analysis, and defense," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8738–8753, Sep. 2018.

[28] K. C. Zeng, Y. Shu, S. Liu, Y. Dou, and Y. Yang, "A practical GPS location spoofing attack in road navigation scenario," in *Proc. 18th Int. Workshop Mobile Comput. Syst. Appl.*, Feb. 2017, pp. 85–90, doi: 10.1145/3032970.3032983.

[29] A. Chehri, I. Fofana, and X. Yang, "Security risk modeling in smart grid critical infrastructures in the era of big data and artificial intelligence," *Sustainability*, vol. 13, no. 6, p. 3196, Mar. 2021.

[30] M. Pournader, A. Kach, and S. Talluri, "A review of the existing and emerging topics in the supply chain risk management literature," *Decis. Sci.*, vol. 51, no. 4, pp. 867–919, Aug. 2020.

[31] K. Ntafloukas, D. P. McCrum, and L. Pasquale, "A cyber-physical risk assessment approach for Internet of Things enabled transportation infrastructure," *Appl. Sci.*, vol. 12, no. 18, p. 9241, Sep. 2022.

[32] G. Li et al., "Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios," *Transp. Res. C, Emerg. Technol.*, vol. 122, Jan. 2021, Art. no. 102820.

[33] N. Koohathongsumrit and W. Meethom, "An integrated approach of fuzzy risk assessment model and data envelopment analysis for route selection in multimodal transportation networks," *Expert Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114342.

[34] F. Luo, Y. Jiang, Z. Zhang, Y. Ren, and S. Hou, "Threat analysis and risk assessment for connected vehicles: A survey," *Secur. Commun. Netw.*, vol. 2021, Sep. 2021, Art. no. 1263820.

[35] Q. Liu, C. Li, H. Jiang, S. Nie, and L. Chen, "Transfer learning-based highway crash risk evaluation considering manifold characteristics of traffic flow," *Accident Anal. Prevention*, vol. 168, Apr. 2022, Art. no. 106598.

[36] M. Kalinin, V. Krundyshev, and P. Zegzhda, "Cybersecurity risk assessment in smart city infrastructures," *Machines*, vol. 9, no. 4, p. 78, Apr. 2021.

[37] C. Casorrán, B. Fortz, M. Labbé, and F. Ordóñez, "A study of general and security Stackelberg game formulations," *Eur. J. Oper. Res.*, vol. 278, no. 3, pp. 855–868, Nov. 2019.

[38] Y. Yang, T. Zhang, and Q. Zhu, "A game-theoretic analysis of auditing differentially private algorithms with epistemically disparate herd," in *Proc. Int. Conf. Decis. Game Theory Secur.*, 2023, pp. 349–368.

[39] D. Guerrero, A. A. Carsteanu, and J. B. Clempner, "Solving Stackelberg security Markov games employing the bargaining Nash approach: Convergence analysis," *Comput. Secur.*, vol. 74, pp. 240–257, May 2018.

[40] L. Xiao, D. Xu, N. B. Mandayam, and H. V. Poor, "Attacker-centric view of a detection game against advanced persistent threats," *IEEE Trans. Mobile Comput.*, vol. 17, no. 11, pp. 2512–2523, Nov. 2018.

[41] R. Zhang and Q. Zhu, "`FlipIn`: A game-theoretic cyber insurance framework for incentive-compatible cyber risk management of Internet of Things," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2026–2041, 2020.

[42] Q. Zhu, C. Fung, R. Boutaba, and T. Basar, "GUIDEX: A game-theoretic incentive-based mechanism for intrusion detection networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 11, pp. 2220–2230, Dec. 2012.

[43] A. Grzybek, G. Danoy, P. Bouvry, and M. Seredynski, "Mitigating flash crowd effect using connected vehicle technology," *Veh. Commun.*, vol. 2, no. 4, pp. 238–250, Oct. 2015.

[44] J. Ashraf et al., "IoTBoT-IDS: A novel statistical learning-enabled botnet detection framework for protecting networks of smart cities," *Sustain. Cities Soc.*, vol. 72, Sep. 2021, Art. no. 103041.

[45] M. A. Al-Shareeda, M. Anbar, S. Manickam, and I. H. Hasbullah, "Review of prevention schemes for man-in-the-middle (MITM) attack in vehicular ad hoc networks," *Int. J. Eng. Manage. Res.*, vol. 10, no. 3, pp. 153–158, Jun. 2020.

[46] Y.-T. Yang, H. Lei, and Q. Zhu, "Adaptive incentive-compatible navigational route recommendations in urban transportation networks," 2024, *arXiv:2409.00236*.

[47] M. Beckmann, C. B. McGuire, and C. B. Winsten, *Studies in the Economics of Transportation*. Connecticut, USA: Yale Univ. Press, 1956. [Online]. Available: http://cowles.yale.edu/sites/default/files/files/pub/misc/specpub-beckmann-mcguire-winsten.pdf

[48] G. Still, "Lectures on parametric optimization: An introduction," Optim. Online, 2018.

[49] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[50] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla, "An efficient approach to solving the road network equilibrium traffic assignment problem," *Transp. Res.*, vol. 9, no. 5, pp. 309–318, Oct. 1975.

[51] X. Di, X. He, X. Guo, and H. X. Liu, "Braess paradox under the boundedly rational user equilibria," *Transp. Res. B, Methodol.*, vol. 67, pp. 86–108, Sep. 2014.