An Enhanced Human Navigation Simulator for benchmarking Human-Aware Robot Navigation*

Noé Pérez-Higueras[1], Miguel Escudero[1], Fernando Caballero[1] and Luis Merino[1]

***Abstract*— This paper presents the new ongoing iteration of the Human Navigation Simulator (*HuNavSim*) [1], [2], an open-source ROS 2 tool for simulating diverse human-agent behaviors to facilitate the development and evaluation of human-aware robot navigation. We highlight two major advancements: the use of Behavior Trees to orchestrate complex, realistic human actions and the integration of multiple high-fidelity simulators. Recognizing the subjective nature of social navigation evaluation, we further present our ongoing work on human-aligned benchmarking [3]. By analyzing correlations between numerical metrics and human surveys, we seek to capture human perceptions such as safety or comfort through quantitative metrics.**

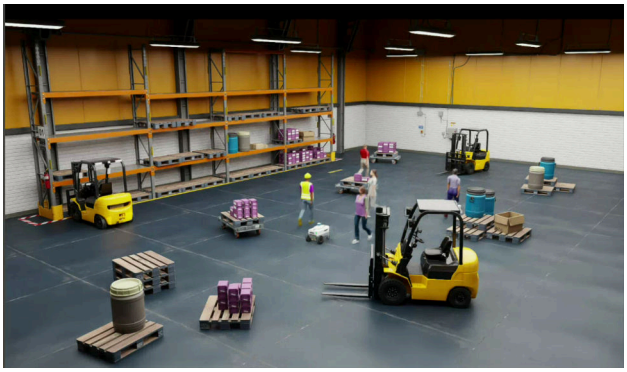

Fig. 1: Capture of HuNav agents in the NVidia Isaac Sim

I. INTRODUCTION AND RELATED WORK

As mobile robots transition from industrial settings to domestic environments shared with humans, human-aware navigation has become a pivotal research area. Developing these social robots presents two primary challenges: the high cost and difficulty of conducting repeatable real-world experiments with human participants, and the lack of consensus on standardized evaluation metrics that capture subjective qualities like comfort, legibility, and safety [4], [5].

To address the first challenge, several tools have been developed to simulate human crowds. Classical models such as the Social Force Model (SFM) utilize force-based interactions to lead agent movement [6]. Tools like PedSimROS [7] and MengeROS [8] integrated these into the ROS ecosystem, while more comprehensive frameworks like SEAN [9], [10] and CrowdBot [11] utilized the Unity engine to establish benchmarks. However, many existing simulators suffer from uniform behavior patterns that fail to capture the unpredictability of real human navigation. Furthermore, frameworks like SocNavBench [12] provide realistic environments through replayed real-world data but lack interactivity, as simulated agents do not react to the robot's actions. HuNavSim 2.0 improves upon these limitations by employing Behavior Trees (BT) to orchestrate complex and realistic human actions beyond simple goal-to-goal navigation. By adding controlled noise to SFM parameters, the simulator introduces behavioral variability closer to real-world trajectories. The tool is highly flexible, providing wrappers for multiple high-fidelity simulators including Gazebo Classic, Gazebo Fortress, NVIDIA Isaac Sim, and Webots.

The second challenge involves bridging the gap between numerical data and human perception. While robots can be evaluated using traditional performance metrics (e.g., path length, time to goal), there is no agreement on which quantitative metrics (QM) reliably capture social aspects like friendliness or unobtrusiveness. Recent efforts, such as those in SocialGym 2.0 [13] and Arena [14], [15], have focused on training social navigation via reinforcement learning, yet they often rely on fixed or limited metric sets. In HuNavSim we provide a set of 32 metrics collected from the literature. Our current work focuses on identifying human-aligned benchmarking protocols. By analyzing the correlations between the comprehensive set of quantitative metrics (QM)—including proxemics, social work, and safety measures—and qualitative human surveys (HM), we aim to identify a minimal set of representative metrics. Preliminary results suggest that while metrics like Social Work are commonly used in literature, they can be too noisy to distinguish social behavior levels as perceived by humans. Instead, metrics such as Intimate Space Occupancy and Average Minimum Distance to People show a stronger correlation with human-level assessments of safety and comfort. By integrating these findings into the HuNavSim 2.5 evaluator, we move toward an automated benchmarking system that aligns more closely with real-world human judgment.

The complete documentation and code of *HuNavSim* is available at `https://github.com/robotics-upo/hunav_sim`.

*This work was partially funded by MCIN/AEI/10.13039/501100011033 under project PICRAH4.0 (PLEC2023-010353) and MICIU/AEI/10.13039/501100011033 under project AI-FUSE (AIA2025-163563-C33)

[1]All the authors are with Service Robotics Lab, University Pablo de Olavide, Avd. Rectora Rosario Valpuesta, 1. CP 41089. Dos Hermanas (Seville), Spain {noeperez, mescjim, fcaballero, lmercab}@upo.es

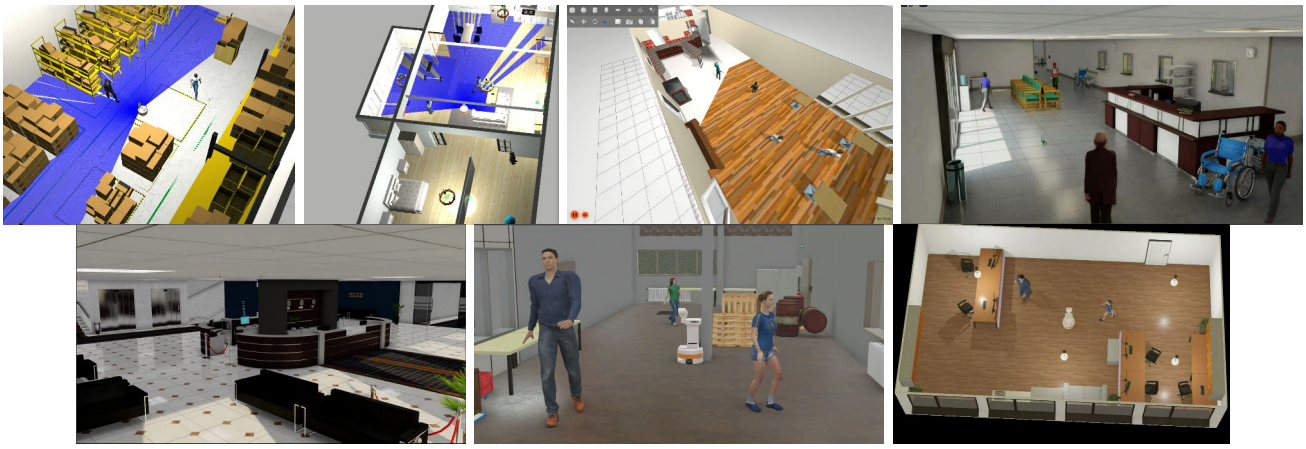

Fig. 2: Examples of simulators and some scenarios provided. From left to right and up to botton: Warehoure and home in Gazebo Classic, Cafe in Gazebo Fortress, Hospital and office hall in Isaac Sim, Industrial hall and office in Webots

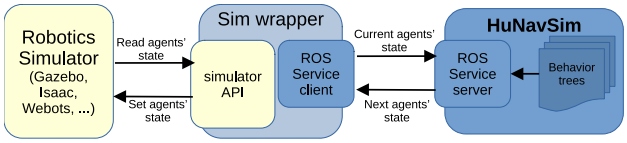

Fig. 3: General diagram of HuNavSim

II. HUNAVSIM ENHANCED SIMULATION CAPABILITIES

we briefly present the features of HuNavSim to represent human navigation behaviors in simulation.

A. System architecture and scenario generation

HuNavSim 2.0 is designed as a ROS 2-based manager that controls the poses and behaviors of human agents spawned within a base robotics simulator. The core architecture relies on communication wrappers that interface the manager with the chosen simulator, computing agent states in each simulation step based on active Behavior Trees (see Fig. 3). To streamline the development of social scenarios, the tool introduces an interactive RViz2-based workflow. Through the HuNavPanel, users can sequentially configure agent populations, set initial poses via a click-and-drag tool, and define navigation goals. This process auto-generates the necessary YAML configuration and XML BT files for each agent, which can then be graphically refined using the Groot 2 editor. This "what-you-see-is-what-you-simulate" approach allows for rapid iteration and focused ablation tests of social navigation algorithms.

B. Enhanced human behavior and evaluation

1) Behavioral Variability and Complex Interactions: A significant advancement in version 2.0 is the introduction of behavioral variability and complex social tasks. While traditional models often produce uniform agent responses, HuNavSim 2.0 adds controlled noise to the Social Force Model (SFM) parameters, resulting in trajectories that more closely mimic the unpredictability of real human movement. Beyond simple goal-directed walking, the system leverages Behavior Trees to orchestrate intricate social interactions. We have implemented a comprehensive set of actions and conditions that allow agents to:

- Perform environmental tasks (e.g., workers checking material in a warehouse).
- Initiate and participate in conversation groups.
- React dynamically to the robot or other agents, such as following a specific person or stopping to assess a robot's presence.

A key advancement in HuNavSim 2.0 is the introduction of behavioral variability to overcome the uniformity of traditional crowd models. To better reflect real human navigation, controlled noise is added to the Social Force Model (SFM) parameters, based on sensitivity analysis identifying realistic ranges. Users can choose fixed parameters for repeatability, custom settings for specific needs, or stochastic sampling from normal distributions within these ranges to generate natural variability.

Beyond local varied movement, the system utilizes Behavior Trees (BT) to orchestrate intricate social tasks that far exceed simple goal-directed walking. While the first version of the tool was restricted to six basic reactions to a robot, HuNavSim 2.0 provides a modular library of actions and conditions that can be combined to model complex human behavior. Key capabilities enabled by these new nodes include:

- **Task-Oriented Actions**: Agents can perform environmental tasks, such as workers in a warehouse, navigating (*GoTo*), inspecting items (*LookingAtPoint*), and pausing for specific durations (*StopAndWaitTimer*).
- **Social Group Dynamics**: The system supports the dynamic creation of conversation groups using the ConversationFormation action. These can be triggered by sophisticated social conditions, such as checking if another agent is paying attention (*IsLookingAtMe*) or currently talking (*isSpeaking*).
- **Human-Robot Interaction**: Agents can exhibit ad-

vanced reactive behaviors, such as following a robot (*FollowAgent*) or actively approaching it to simulate curiosity (*ApproachRobot*).

This BT-based architecture allows researchers to build highly realistic and interactive scenarios, ensuring that simulated humans react dynamically both to each other and to the robot's presence. By providing an interactive what-you-see-is-what-you-simulate workflow in RViz 2, these complex trees can be auto-generated or graphically refined using the Groot 2 editor, facilitating rapid scenario iteration and focused benchmarking.

2) Multi-Simulator Support and Benchmarking: HuNavSim 2.0 has expanded its compatibility to support four major robotics simulators: Gazebo Classic, Gazebo Fortress, NVIDIA Isaac Sim, and Webots. This allows researchers to evaluate navigation across diverse high-fidelity environments, from hospitals to industrial halls (see Fig. 2). For evaluation, the tool includes an Evaluator module that logs experimental data and computes a comprehensive set of 32 social navigation metrics. This set includes traditional efficiency measures (e.g., time to goal, path length) and advanced social indicators such as Proxemics, Social Work, and the newly integrated "danger and surprise" metrics. Users can select which metrics to compute and trigger when to start the calculations via ROS 2 services to define specific evaluation windows within a simulation. To enhance the benchmarking capabilities of the next version of HuNavSim (2.5), we are currently working on a protocol to identify human-aligned metrics. While the simulator can compute a comprehensive set of 32 metrics, there is a significant lack of consensus on which numerical values truly reflect subjective human experiences such as comfort and safety.

3) Use of LLMs and VLMs for scenario generation: In HuNavSim, we also explore the use of Large Language Models (LLMs) and Vision–Language Models (VLMs) to support two tasks: the creation of complex physical environments and the generation of behavior trees (BTs) for human navigation.

The generation of human behaviors through BTs shows promising results. However, it still lacks the required stability and robustness, occasionally producing incomplete BTs. This functionality is already integrated into HuNavSim, allowing users to decide whether to use it.

The creation of physical environments for the Gazebo simulator yields satisfactory results, significantly reducing the time required to generate environments. This feature has been delivered as an optional component of HuNavSim: during installation, users are prompted to choose whether to include it. Scenario generation for other base simulators, such as NVIDIA Isaac Sim or Webots, is planned as future work.

III. TOWARDS HUMAN-ALIGNED BENCHMARKING

Our current research addresses the gap between easy-to-compute quantitative metrics (QM) and costly, non-reproducible human-level assessments (HM). By establishing which numerical measures correlate with human judgment, we can provide standardized, automated benchmarks that serve as reliable preliminary screening tools.

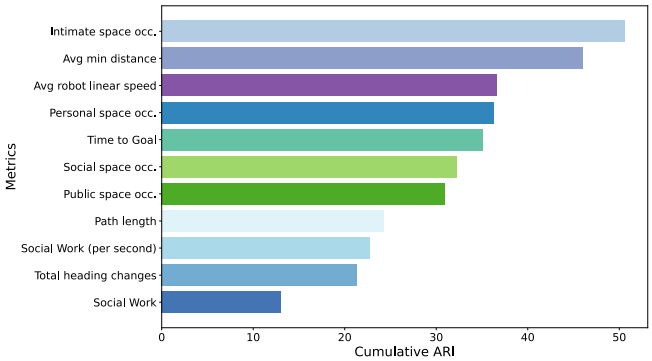

Fig. 4: Histogram of the cumulative ARI. The QM metrics with the highest cumulative ARI result are more relevant for clustering experiments as done using the HM feature set.

A. data collection and experimental setup

We developed a framework to identify an optimal subset of representative metrics (QM*) through a joint analysis of real-world trajectories and human surveys. We conducted 24 social navigation experiments using a Jackal robot across eight distinct scenarios, including Passing, Crossing, and complex Mixed interactions.

A population of 70 participants evaluated these experiments using a 5-point Likert scale across four human-centric criteria: unobtrusiveness, friendliness, smoothness, and avoidance foresight. To process this data, we employed an unsupervised K-means clustering approach to see if quantitative data naturally aligns with human categorical perceptions, alongside statistical tests like Spearman's Rho and Kendall's Tau tests.

B. Unsupervised approach

While human labels were available, we specifically chose an unsupervised K-means clustering approach to determine if the inherent, high-dimensional structure of the quantitative data naturally aligns with human categorical perceptions without the bias of supervised training. This allows for a robust assessment of whether numerical metrics can autonomously capture the underlying patterns of social interaction observed by humans.

We employed the K-means clustering algorithm to group the 24 experimental samples using the QM and HM feature spaces independently. An internal composition analysis using the silhouette score indicated that a division into K=2 clusters was the most distinguishable configuration, achieving a score of 0.548. To measure the matching accuracy between the resulting clusters, we utilized the Adjusted Rand Index (ARI) [16], treating the HM-based clusters as the ground truth. We computed the ARI for all 2,047 possible combinations of quantitative metrics. As shown in Figure 4 (Histogram of the cumulative ARI), we calculated a cumulative ARI score for each metric to represent its relevance in achieving human-like groupings. The five most prevalent metrics identified

through this process were intimate space occupancy, average minimum distance to the closest person, average robot linear velocity, personal space occupancy, and time to goal.

C. Statistical analysis

To investigate the specific links between individual metrics, we conducted a correlation study using two non-parametric methods: Spearman's Rho, and Kendall's Tau. These methods are particularly appropriate for the ordinal 5-point Likert-scale evaluations used in our human survey. Spearman and Kendall coefficients were utilized to evaluate monotonic relationships, with significance thresholds set at $|\rho > 0.4|$ and $|\tau > 0.25|$ (and $p < 0.05$ in both cases). Figure 5 presents the average absolute strength of correlations that met these criteria. The results demonstrate that metrics such as average minimum distance to person and intimate space occupancy correlate strongly with human perceptions of unobtrusiveness and friendliness. Conversely, Social Work (SW), despite its popularity in literature, was found to be too "noisy" to reliably distinguish between different social behavior levels as perceived by humans.

D. Identification of Representative Metrics

The preliminary analysis, presented in the previous sections, revealed that a minimal set of five metrics could be most effective for achieving a human-like evaluation trend:

1) Intimate Space Occupancy: Time spent within the closest proximity to humans.
2) Average Minimum Distance (AMD): The average of the closest points reached during interaction.
3) Personal Space Occupancy: Time spent within the personal boundary.
4) Average Robot Linear Velocity: Correlated strongly with human perceptions of avoidance foresight.
5) Time to Goal: A key indicator of navigation efficiency that also impacts perceived social quality.

Notably, the results indicated that Social Work (SW), though a popular academic metric derived from the Social Force Model, is often too "noisy" to reliably distinguish between different social behavior levels as perceived by humans. Similarly, path length showed poor discriminatory power compared to human feedback. These results require further investigation and analysis with larger datasets and a wider variety of social scenarios.

E. Integration with HuNavSim

By identifying this QM* set, we could refine the evaluator module of HuNavSim described in Section II-B.2. While the simulator would retain the flexibility to compute the full suite of 32 metrics, our goal is to prioritize these human-aligned measures to provide researchers with a benchmarking protocol that more closely mirrors real-world human judgment. This automated approach would allow for rapid iteration during the development of social navigation algorithms before proceeding to expensive final validation with human participants.

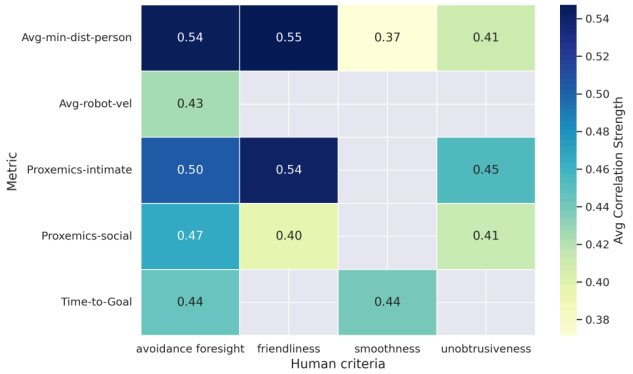

Fig. 5: Heatmap of consistent correlations between single QM and HM metrics, according to both Kendall and Spearman analysis.

IV. CONCLUSIONS AND FUTURE WORK

This paper has presented the ongoing work on HuNavSim 2.5, an enhanced open-source ROS 2 library designed to simulate diverse and realistic human navigation behaviors in shared environments. By leveraging Behavior Trees, the tool offers a modular and extensible architecture to orchestrate complex social tasks, while the introduction of controlled noise into the Social Force Model (SFM) ensures behavioral variability that closely mirrors real-world human trajectories. A significant contribution of our current work is the study of a human-aligned benchmarking protocol. Our correlation analysis between quantitative metrics and human surveys has identified a representative subset of quantitative metrics (QM*) that captures subjective human perceptions of safety and comfort. By prioritizing these metrics, HuNavSim 2.5 can provide an automated screening tool that allows researchers to iterate rapidly on social navigation algorithms before proceeding to costly real-world validation. Despite these advancements, we acknowledge that human surveys remain the gold standard for final system validation, as current numerical metrics still struggle to fully capture subtle aspects like trajectory smoothness. Future work will focus on three key areas:

- Advanced Prediction Models: Integrating more sophisticated human motion models beyond the classical SFM to further improve simulation realism.
- Extend the study on the human-aligned metrics by using larger datasets, a wider set of metrics and different tools for analysis.

Through these developments, we aim to provide the robotics community with a standardized and robust framework for the development and evaluation of human-aware navigation systems.

ACKNOWLEDGMENT

We declare the use of the AI assistant NotebookLM in the writing of this paper.

REFERENCES

[1] M. Escudero-Jiménez, N. Pérez-Higueras, A. Martínez-Silva, F. Caballero, and L. Merino, "Hunavsim 2.0: An enhanced human navigation simulator for human-aware robot navigation," 2025. [Online]. Available: https://arxiv.org/abs/2507.17317

[2] N. Pérez-Higueras, R. Otero, F. Caballero, and L. Merino, "Hunavsim: A ros 2 human navigation simulator for benchmarking human-aware robot navigation," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7130–7137, 2023.

[3] S. Trepella, M. Martini, N. Pérez-Higueras, A. Ostuni, F. Caballero, L. Merino, and M. Chiaberge, "Metrics vs surveys: Can quantitative measures replace human surveys in social robot navigation? a correlation analysis," 2025. [Online]. Available: https://arxiv.org/abs/2510.02941

[4] A. Francis, C. Pérez-D'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra, H.-T. L. Chiang, M. Everett, S. Ha, J. Hart, J. P. How, H. Karnan, T.-W. E. Lee, L. J. Manso, R. Mirsky, S. Pirk, P. T. Singamaneni, P. Stone, A. V. Taylor, P. Trautman, N. Tsoi, M. Vázquez, X. Xiao, P. Xu, N. Yokoyama, A. Toshev, and R. Martín-Martín, "Principles and guidelines for evaluating social robot navigation algorithms," *J. Hum.-Robot Interact.*, vol. 14, no. 2, Feb. 2025. [Online]. Available: https://doi.org/10.1145/3700599

[5] P. T. Singamaneni, P. Bachiller-Burgos, L. J. Manso, A. Garrell, A. Sanfeliu, A. Spalanzani, and R. Alami, "A survey on socially aware robot navigation: Taxonomy and future challenges," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1533–1572, 2024. [Online]. Available: https://doi.org/10.1177/02783649241230562

[6] Helbing and Molnár, "Social force model for pedestrian dynamics." *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, vol. 51 5, pp. 4282–4286, 1995.

[7] T. L. B. Okal, "Ros pedsim," 2017, https://github.com/srl-freiburg/pedsim_ros.

[8] A. Aroor, S. L. Epstein, and R. Korpan, "Mengeros: a crowd simulation tool for autonomous robot navigation," in *AAAI 2017 Fall Symposium on Artificial Intelligence for Human-Robot Interaction*, 2017, pp. 123–125.

[9] N. Tsoi, M. Hussein, J. Espinoza, X. Ruiz, and M. Vázquez, "Sean: Social environment for autonomous navigation," in *the 8th International Conference on Human-Agent Interaction (HAI '20)*, 11 2020, pp. 281–283.

[10] N. Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, and M. Vázquez, "Sean 2.0: Formalizing and generating social situations for robot navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 047–11 054, 2022.

[11] F. Grzeskowiak, D. Gonon, D. Dugas, D. Paez-Granados, J. J. Chung, J. Nieto, R. Siegwart, A. Billard, M. Babel, and J. Pettré, "Crowd against the machine: A simulation-based benchmark tool to evaluate and compare robot capabilities to navigate a human crowd," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 3879–3885.

[12] A. Biswas, A. Wang, G. Silvera, A. Steinfeld, and H. Admoni, "Socnavbench: A grounded simulation testing framework for evaluating social navigation," *ACM Transactions on Human-Robot Interaction*, jul 2022. [Online]. Available: https://doi.org/10.1145/3476413

[13] Z. Sprague, R. Chandra, J. Holtz, and J. Biswas, "Socialgym 2.0: Simulator for multi-agent social robot navigation in shared human spaces," 2023. [Online]. Available: https://arxiv.org/abs/2303.05584

[14] L. Kästner, V. Shcherbyna, H. Zeng, T. A. Le, M. H.-K. Schreff, H. Osmaev, N. T. Tran, D. Diaz, J. Golebiowski, H. Soh, and J. Lambrecht, "Demonstrating Arena 3.0: Advancing Social Navigation in Collaborative and Highly Dynamic Environments," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[15] V. Shcherbyna1, L. Kästner, D. Diaz, H. G. Nguyen, M. H.-K. Schreff, T. Lenz, J. Kreutz, A. Martban, H. Zeng, and H. Soh, "Arena 4.0: A comprehensive ros2 development and benchmarking platform for human-centric navigation using generative-model-based environment generation," 2024. [Online]. Available: https://arxiv.org/abs/2409.12471

[16] D. Steinley, "Properties of the hubert-arable adjusted rand index." *Psychological methods*, vol. 9, no. 3, p. 386, 2004.