Метод определения степени доверия автономных транспортных средств на основе социальных сил

Ю.А. Павелина

***Аннотация* — В работе описан комплексный метод оценки степени доверия в группе автономных транспортных средств (АТС) с учетом временного затухания, скользящего окна и социальных сил Хелбинга-Молнара притяжения и отталкивания. Применение экспоненциального затухания и скользящего окна, ограничивающее анализ последних взаимодействий АТС позволяет снизить влияние инерции доверия. Также для оценки степени доверия учитывается групповая репутация АТС, которая рассчитывается с помощью социальных сил, основанных на степени доверия соседних АТС и мере сходства их данных. Итоговая степень доверия формируется как взвешенная сумма личного опыта и групповой репутации, что обеспечивает баланс между индивидуальной информацией и коллективным влиянием. Метод позволяет эффективно выявлять АТС с пониженной достоверностью информации (диверсантов) и демонстрирует адаптивность к изменяющемуся поведению в динамичных системах.**

***Ключевые слова* — мультиагентные системы, временное затухание, доверие, репутация, скользящее окно.**

I. Введение

В настоящее время существует тенденция развития автоматизации систем окружающего нас мира. Одной из быстро развивающейся областей автоматизации являются транспортные системы. Группы автономных транспортных средств (АТС) являются системами с большим количеством вычислительных и исполнительных устройств, связанных между собой. Что увеличивает потребность в повышении безопасности таких систем.

В условиях неопределенности и постоянном изменении окружающей среды достоверность получаемых данных является критически важной задачей информационного взаимодействия в группах АТС. Так для повышения безопасности и эффективности работы группы оценка степени доверия источника информации является важным фактором. Стоит отметить существование угрозы фальсификации информации АТС в группе.

II. Методы расчета степеней доверия

Методы оценки степени доверия и репутации развиваются для решения задач координации в условиях неопределенности и потенциальной недобросовестности агентов.

Современные исследования разделяют модели доверия на два уровня: индивидуальный и системный. На индивидуальном уровне доминируют репутационные системы, основанные на бета-распределениях и машинном обучении. Например, [1] предлагают концепцию «актуального доверия» (actual trust), где оценка строится на прогнозе способности агента выполнять задачи, а не на исторических данных. Это согласуется с тенденцией перехода от ретроспективных к прогностическим моделям.

На системном уровне ключевым трендом стало внедрение формальных методов верификации. В работе [2] разработали фреймворк, интегрирующий логику ATLK (Alternating-time Temporal Logic with Knowledge) для количественной оценки доверия в реальном времени. Такой подход особенно важен для АТС, где задержки в оценке могут привести к катастрофическим последствиям.

Современные подходы активно интегрируют глубокое обучение для прогнозирования доверия. Распределённые графовые нейросети (DGNN) агрегируют данные от соседних агентов, учитывая семантическое сходство данных, временные задержки и топологию коммуникационных сетей. Авторы работы [3] на симуляторах АТС показали, что DGNN повышают точность обнаружения злонамеренных агентов на 22% по сравнению с традиционными методами. Однако их вычислительная сложность $O(n^2)$ ограничивает применение в реальном времени при n>100 агентов.

III. Метод расчета репутации и степеней доверия группы агентов

A. Модель группы автономных транспортных средств

Так как группа АТС является самоорганизующейся, динамической, информационно-коммуникационной системой, состоящей из интеллектуальных элементов, для описания такой системы был использован мультиагентных подход.

Рассмотрим группу автономных транспортных средств (АТС), состоящую из n интеллектуальных агентов: $\Lambda = \{A_1, \dots, A_n\}$.

Будем считать, что время такой системы дискретно $T = \{0, t_1, \dots, t_q\}$.

Каждого агента группы определяют его координаты, количество ресурсов и множество знаний о среде:

$$\mathrm{A}_i(t) = \\ = \{coord_i(t), v_i(t), R_i(t), Inf_i(t), r\_inf_i(t), r\_sens_i(t)\}$$

где:

- $coord_i(t)$ - набор пространственных характеристик агента A_i в момент времени t,

- $v_i(t)$ - скорость агента в момент времени t,
- $Res_i(t) = [res_1(t), \dots, res_h(t)]^T$ - вектор-функция, описывающая ресурсы агента в момент времени t;
- $Inf_i(t)$ - множество знаний агента в момент времени t, собранные с помощью физических устройств агента или полученные в результате информационного взаимодействия системы управления с другими агентами;
- $r\_inf_i(t)$ - радиус информационного взаимодействия агента A_i в момент времени t;
- $r\_sens_i(t)$ - доступный (возможный) радиус сбора информации физическими устройствами агента A_i в момент времени t.

Обозначим дистанцию необходимую для информационного взаимодействия агентов $d_{i,j}^{Inf}$.

Так для существование устойчивого информационного канала $d_{i,j}^{Inf}$ должна быть больше или равна радиусу информационного взаимодействия каждого из агентов.

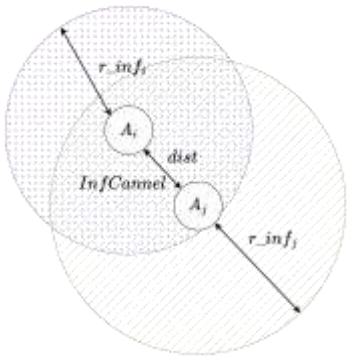

Рисунок 1 - Возможный информационный канал между агентами

B. Расчет степени доверия

1) Бета-распределение для расчета степени доверия

Бета-распределение – это вероятностное распределение, используемое для моделирования случайных величин, ограниченных интервалом [0,1].

Выбор бета-распределения обусловлен следующими факторами:

- бета-распределение является сопряженным априорным распределением для биномиального правдоподобия, что позволяет естественным образом обновлять оценки доверия при поступлении новых данных (подтверждений/опровержений сообщений) через простое инкрементное обновление параметров α и β;
- дисперсия бета-распределения естественным образом уменьшается с ростом числа наблюдений (α+β), что соответствует интуитивному представлению о снижении неопределенности при накоплении опыта взаимодействий;
- бета-распределение позволяет каждому АТС независимо поддерживать и обновлять оценки доверия к соседним агентам, используя только два параметра (α – успешные взаимодействия, β – неуспешные взаимодействия), что минимизирует вычислительные и коммуникационные издержки [4].

Для расчета степени доверия необходимо задать следующие неотрицательные параметры:

- α - число подтвержденных «подтвержденных» сообщений;
- β - число опровергнутых сообщений.

Подтвержденные сообщения – это сообщения, передаваемые агентом, которые согласуются с данными, полученными принимающим агентом.

Отвергнутые сообщения – это сообщения, информация из которых противоречит информации принимающего агента.

Тогда плотность распределения имеет следующий вид:

$$f(p,\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)},$$

где $B(\alpha,\beta)$ - бета-распределение:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

где Γ - гамма распределение.

Степень доверия есть математическое ожидание бета-распределения:

$$D = \frac{\alpha}{\alpha+\beta}.$$

Неопределенность степени доверия оценивается как дисперсия бета-распределения:

$$Var(D) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

откуда следует, что чем больше $\alpha+\beta$, тем ниже неопределенность оценки степени доверия.

2) Временное затухание степени доверия

При поступлении нового взаимодействия параметры бета-распределения обновляются следующим образом:

- $\delta_\alpha = 1,\ \delta_\beta = 0$ – при подтверждении достоверности сообщения;
- $\delta_\alpha = 0,\ \ \delta_\beta = 1$ – при обнаружении нарушения достоверности сообщения.

Для снижения влияния устаревания репутации необходимо ввести коэффициент временного затухания степени доверия $\lambda_{\Delta t}$: $\lambda_{\Delta t} \in (0;1]$.

Тогда обновление параметров степени доверия с учетом временного затухания выглядит следующим образом:

$$\alpha_{t+1} = \lambda_{\Delta t}\alpha_t + (1-\lambda_{\Delta t})\delta_\alpha,$$
$$\beta_{t+1} = \lambda_{\Delta t}\beta_t + (1-\lambda_{\Delta t})\delta_\beta.$$
$$D^{time} = \frac{\alpha_{t+1}}{\alpha_{t+1}+\beta_{t+1}}$$

3) Скользящее окно

Для снижения эффекта инерции доверия необходимо использовать метод скользящего окна. Основной идеей скользящего окна является использование не всей истории взаимодействия, а только последних N событий:

$$\alpha_{t+1}^{window} = 1 + \sum_{k=0}^{w-1}\delta_{\alpha,t-k},$$
$$\beta_{t+1}^{window} = 1 + \sum_{k=0}^{w-1}\delta_{\beta,t-k}.$$
$$D^{window} = \frac{\alpha_{t+1}^{window}}{\alpha_{t+1}^{window} + \beta_{t+1}^{window}}$$

4) Комбинация методов

Для оценки степени доверия агента необходимо учитывать временное затухание, а также скользящее окно. Тогда степень доверия агента A_i есть:

$$D_i^{result} = \mu D_i^{time} + (1-\mu)D_i^{window},$$

где μ – вес временного затухания.

C. Определение групповой репутации агента

1) Адаптация модели социальных сил

Модель Дирка Хелбинга и Питера Молнара [5] описывает взаимодействие агентов через "силы". Модель Хелбинга-Молнара, разработанная для пешеходной динамики, описывает движение как результат

социальных сил: притяжения к цели, отталкивания от препятствий и случайных возмущений.

В отличие от традиционных графовых моделей, где связи статичны, при использовании социальных сил репутация динамически пересчитывается через силы, зависящие от:

- пространственной близости,
- семантического сходства данных,
- исторической надежности.

В контексте расчета степеней доверия эти силы можно интерпретировать как влияние подтвержденных и опровергнутых данных, а также пространственные отношения агентов.

Для каждого агента A_i в радиусе $dist^{Inf}$ вычисляется влияние соседних агентов A_j через социальные силы Хелбинга и Молнара.

Сила доверия (притяжения):

$$F_{i,j}^{trust} = \Omega \cdot \frac{D_j \cdot sim(i,j)}{d_{i,j}^2+\epsilon} g(d_{i,j}),$$

где:

- $d_{i,j}$ – метрика расстояния между агентами;
- $sim(i,j)$ – мера сходства между данными от A_i и A_j;
- $e_{i,j}$ – единичный вектор направления от агента A_i к агенту A_j;
- D_j – степень доверия агента A_i к агенту A_j;
- Ω – положительный коэффициент, определяющий амплитуду силы доверия;
- ϵ – малая константа;
- $g(d_{i,j})$ – параметр, характеризующий возможность построения информационного канала между агентами:

$$g(d_{i,j}) = \{^{1, \text{если } g(d_{i,j}) \le d_{i,j}^{Inf}}_{0, \text{если } g(d_{i,j}) > d_{i,j}^{Inf}}.$$

Мера сходства данных вычисляется как степень согласованности информации, передаваемой агентами A_i и A_j. Формально:

$$sim(i,j) = 1 - \frac{\|mes_i - mes_j\|}{\|mes_i\| + \|mes_j\|},$$

где mes_i и mes_j – векторы признаков сообщений агентов A_i и A_j, например, координаты обнаруженных объектов, типы событий.

Аналогично рассчитывается сила конфликта (отталкивания) агентов A_i и A_j, находящиеся на расстоянии $d_{i,j}^{Inf}$:

$$F_{i,j}^{conflict} = -\Phi \cdot \frac{D_j \cdot (1 - sim(i,j))}{d_{i,j}^2+\epsilon} g(d_{i,j}),$$

где Φ – положительный коэффициент, определяющий амплитуду силы конфликта.

2) Определение групповой репутации агента на основе социальных сил

Групповая репутация агента A_i определяется как взвешенная сумма, учитывающая оба значения социальных сил агентов, находящихся на расстоянии $d_{i,j}^{Inf}$:

$$R_i = \frac{\sum_{j \neq i} (F_{i,j}^{trust} + F_{i,j}^{conflict}) \cdot D_{j,i}}{\sum_{j \neq i} |F_{i,j}^{trust} + F_{i,j}^{conflict}|},$$

где $D_{j,i}$ – степень доверия агента A_j к A_i.

Как видно из формулы групповой репутации агента силы отталкивания уменьшают влияние агентов с несовпадающими данными.

D. Определение итоговой степени доверия

Итоговая степень доверия определяется на основе личного опыта (п. *B*) и групповой репутации (п. *C*), т.е.

$$D_i^{result} = \gamma D_i + (1-\gamma) R_i$$

IV. Вычислительный эксперимент

A. Описание эксперимента

Для оценки эффективности разработанного метода расчета степеней доверия был разработан программный симулятор, имитирующий определение степени доверия агентов в группе.

Для реализации метода применялись следующие ключевые программные средства и библиотеки:

- NumPy – позволяет работать с многомерными массивами и числовыми операциями, что позволило моделировать параметры бета-распределения, экспоненциальное затухание и скользящее окно.
- NetworkX – для построения и визуализации графов доверия между агентами, что обеспечило наглядное представление групповой репутации и взаимодействий на основе модели социальных сил Хелбинга-Молнара.
- Matplotlib – для построения графиков динамики средней степени доверия и цветовой визуализации итогового графа доверия с градиентным отображением степени надежности агентов.

В ходе эксперимента была смоделирована группа из 30 агентов, среди которых 30% выступали в роли диверсантов с пониженной достоверностью сообщений после 40-го временного шага. Целью диверсантов было введение в заблуждение остальных агентов группы

Метод оценивал степень доверия каждого агента с учетом личного опыта, обновляемого с помощью экспоненциального затухания и скользящего окна, а также групповой репутации, вычисляемой через социальные силы.

Все агенты $\Lambda = \{A_1, .., A_n\}$ в симуляторе объединены в систему, где каждый элемент может взаимодействовать друг с другом на некотором радиусе взаимодействия для расчета степеней доверия.

Входные параметры моделирования следующие:

- количество агентов: 30, выбрано как репрезентативный размер для плотного городского трафика или колонны АТС, обеспечивающий достаточное число взаимодействий для статистически значимых выводов;
- 30% агентов случайно выбираются как диверсанты – их поведение после определённого момента времени меняется;
- каждый агент имеет начальную степень доверия 0,5, что соответствует нейтральной позиции (максимальная неопределенность в бета-распределении с α=β=1), отражающая отсутствие предварительной информации об агентах;
- радиус взаимодействия агентов: 0,5 нормализованное расстояние, соответствующее типичной дальности связи V2V (Vehicle-to-Vehicle) 200-300 м, в симуляторе пространство нормализовано к единичному квадрату;
- порог степени доверия для выбывания агентов из взаимодействия: 0,3, агенты с доверием ниже этого значения считаются ненадежными и исключаются из

обмена информацией; выбран эмпирически как баланс между чувствительностью и устойчивостью к шуму;

- коэффициент забывания: 0,6, определяет скорость затухания старых данных: значение означает, что вклад прошлых взаимодействий уменьшается на 40% на каждом шаге, что соответствует быстрой адаптации в динамичной среде;
- глубина скользящего окна: 15, что при частоте обмена данными 10 Гц соответствует примерно 1,5 секундам истории;
- вес временного компонента личной степени доверия: 0,4;
- вес личного опыта: 0,7;
- коэффициент силы доверия: 2,0, что
- коэффициент силы конфликта: 4,0 (значение для силы конфликта в 2 раза больше, чтобы система быстрее реагировала на расхождения в данных).

Агенты в симуляторе ведут себя по следующему алгоритму:

1. в начальный момент времени агенты инициализируются в случайном месте, при этом агенты не могут появляться в одном и том же секторе;
2. далее на каждом шаге агент генерирует событие (достоверное/недостоверное) в зависимости от своего типа и времени,
3. далее агенты обновляют свою степень доверия с учетом временного затухания и скользящего окна;
4. рассчитываются силы доверия и конфликта на основе радиуса взаимодействия агентов на основе «схожести» переданной информации;
5. рассчитывается групповая репутация агента;
6. рассчитывается итоговое значение степени доверия для каждого агента.

Динамика поведения диверсантов меняется на 40-ом временном шаге, после чего они увеличивают число недостоверных сообщений. Их доверие падает из-за накопления негативного опыта у других агентов, и они исключаются из дальнейшего взаимодействия с группой.

Основной целью моделирования была проверка возможности предложенного метода обнаруживать агентов-диверсантов при информационном взаимодействии.

На рисунке 2 представлены результаты эксперимента, где красный график показывает среднюю степень доверия к агентам-диверсантам, синий - к обычным, а зеленый - среднюю степень доверия к группе.

Рисунок 2 – График динамики средней степени доверия группы агентов

Как видно из рисунка 2, разработанный метод позволяет выявлять и исключать агентов-диверсантов из информационного взаимодействия почти сразу после начала изменения их поведения. Так первый агент-диверсант был обнаружен уже на 11 шаге после изменения поведения. Последний агент был обнаружен на 14 шаге.

При частоте обновления 10 Гц (типичной для систем V2V) 11-14 шагов соответствуют 1,1-1,4 секундам. Для автономных транспортных средств, движущихся со скоростью 60 км/ч (≈16,7 м/с), это означает прохождение дистанции 18-23 метра с момента начала враждебного поведения до изоляции диверсанта. Это время реакции, учитывая необходимость накопления статистически значимого числа несоответствий для уверенного обнаружения.

На рисунке 3 показана пространственная структура степени доверия на последнем шаге симуляции. Линиями между агентами обозначены произошедшие процессы информационного взаимодействия между агентами.

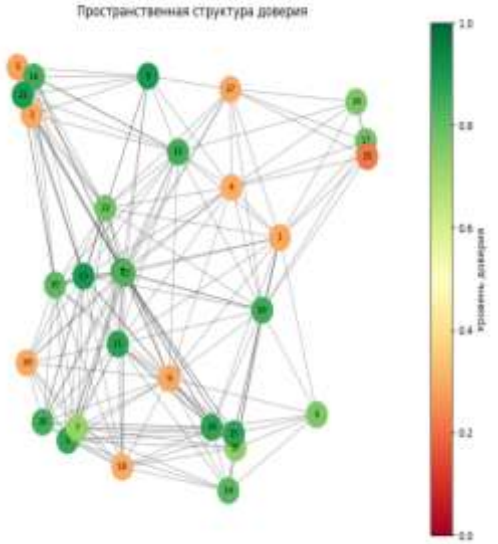

Рисунок 3 – Пространственная структура степеней доверия

V. Заключение

Обеспечение достоверности информации в группах автономных транспортных средств является сложной и важной задачей, для достижения допустимого уровня транспортной безопасности.

Предлагаемый метод к оценке степени доверия на основе временного затухания, скользящего окна и социальных сил учитывает возможные временные изменения в поведении агента, а также схожесть информации, поступающей от него и других АТС группы. Интеграция предлагаемого метода в процесс функционирования и взаимодействия автономных транспортных средств позволяет повысить безопасность такой системы за счет снижения уровня влияния недостоверной информации.

Эффективность методы была проверена с использованием программного моделирования, результаты которого показывают возможность применения метода на практике. В частности, метод оценки степени доверия демонстрирует, как динамическое изменение доверия и социальные взаимодействия влияют на выявление и изоляцию диверсантов в группе.

В качестве плана дальнейших исследований определено многофакторное слияние и сравнение данных с объектами инфраструктуры Интернета транспортных средств (Internet of Vehicles, IoV).

Благодарности

Работа выполнена в Университете ИТМО при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках проекта № 70-2024-001354 «Разработка технологий и демонстратора комплексной системы группового управления, взаимодействия и организации поведения группы БВС при выполнении целевых задач».

Библиография

[1] Akintunde M. et al. Actual Trust in Multiagent Systems. – 2024.
[2] Cheng M. et al. A general trust framework for multi-agent systems //Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. – 2021. – С. 332-340.
[3] Jin W. et al. A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives //arXiv preprint arXiv:2503.13415. – 2025.
[4] Jøsang A., Ismail R., Boyd C. A survey of trust and reputation systems for online service provision //Decision support systems. – 2007. – Т. 43. – №. 2. – С. 618-644.
[5] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. Institute of Theoretical Physics, University of Stuttgart, 70550 Stuttgart, Germany.

Статья получена 17 ноября 2025.
Павелина Юлия Александровна, аспирант факультета Безопасности Информационных Технологий, Национальный исследовательский университет ИТМО (email: lyakhovenko.kam@gmail.com).

A method for determining the trustworthiness of autonomous vehicles based on social forces

J.A. Pavelina

***Abstract* — This paper describes a comprehensive method for estimating trust in a group of autonomous vehicles (AVs) using time decay, a sliding window, and Helbing-Molnar social forces of attraction and repulsion. Using exponential decay and a sliding window to limit the analysis of recent AV interactions reduces the impact of trust inertia. Trust is also assessed using the AVs' group reputation, calculated using social forces based on the trustworthiness of neighboring AVs and the similarity of their data. The final trust score is formed as a weighted sum of personal experience and group reputation, which ensures a balance between individual information and collective influence. The method effectively identifies AVs with reduced information reliability (saboteurs) and demonstrates adaptability to changing behavior in dynamic systems.**

***Keywords* — multi-agent systems, time decay, trust, reputation, sliding window.**

REFERENCES

[1] Akintunde M. et al. Actual Trust in Multiagent Systems. – 2024.
[2] Cheng M. et al. A general trust framework for multi-agent systems //Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. – 2021. – C. 332-340.
[3] Jin W. et al. A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives //arXiv preprint arXiv:2503.13415. – 2025.
[4] Jøsang A., Ismail R., Boyd C. A survey of trust and reputation systems for online service provision //Decision support systems. – 2007. – T. 43. – №. 2. – C. 618-644.
[5] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. Institute of Theoretical Physics, University of Stuttgart, 70550 Stuttgart, Germany.