Моделирование коллективного поведения в социальных сетях на основе математических и нейросетевых моделей

Докладчик: *Огай Владислав Леонидович*, 324

План доклада

Цель и контекст

- Социальные сети $\Rightarrow$ сложное коллективное поведение:
  - консенсус, поляризация, эхо-камеры;
  - вирусное распространение контента;
  - «лидеры мнений» и упрямые агенты.
- Два класса моделей:
  - **математические модели** динамики мнений на графах;
  - **нейросетевые модели**: GNN и сети LLM-агентов.
- Цель: показать, как эти подходы связываются в единую формальную рамку.

Общая постановка

- Множество агентов (пользователей):

$$V = \{1, \dots, n\}, \quad |V| = n.$$

- Направленный граф социальной сети:

$$G = (V, E),$$

где $(j, i) \in E$ означает влияние j на i.

- Матрица весов влияния:

$$W = (w_{ij})_{i,j=1}^{n}, \quad w_{ij} \geq 0.$$

- Скалярное мнение агента i в момент времени t:

$$x_i(t) \in \mathbb{R}, \quad x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \in \mathbb{R}^n.$$

Статистики коллективного поведения

- Среднее мнение:

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t).$$

- Дисперсия мнений (грубая мера поляризации):

$$\mathrm{Var}(x(t)) = \frac{1}{n}\sum_{i=1}^{n}\big(x_i(t) - \bar{x}(t)\big)^2.$$

- Доли агентов по разные стороны порога θ:

$$P_+(t) = \frac{1}{n}\#\{i : x_i(t) > \theta\}, \quad P_-(t) = \frac{1}{n}\#\{i : x_i(t) < \theta\}.$$

- Векторное мнение:

$$x_i(t) \in \mathbb{R}^d, \quad x(t) \in \mathbb{R}^{n\times d}.$$

Модель Де Гроота

Динамика

$$x_i(t+1) = \sum_{j=1}^{n} w_{ij} x_j(t),$$

$$x(t+1) = Wx(t).$$

- Весовые коэффициенты:

$$w_{ij} \geq 0, \quad \sum_{j=1}^{n} w_{ij} = 1 \quad \forall i.$$

- W - стохастическая матрица по строкам
- Интуитивно: мнение агента - взвешенное среднее мнений его соседей.

Предположения:

- Граф влияния сильно связен и апериодичен.
- Матрица W *примитивна*: существует k, такое что W^k положительна.

Консенсус в модели Де Гроота

Теорема о консенсусе

Существуют стационарное распределение π:

$$\pi^\top W = \pi^\top, \quad \sum_{i=1}^{n} \pi_i = 1,$$

и предел:

$$\lim_{t\to\infty} x(t) = \mathbf{1}\,\pi^\top x(0),$$

где $\mathbf{1} = (1, \ldots, 1)^\top$.

- Итоговое мнение: взвешенный консенсус

$$x^* = \sum_{i=1}^{n} \pi_i x_i(0).$$

Сходимость

- Собственные значения матрицы W:

$$1 = \lambda_1, \quad |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|.$$

- Разложение начального вектора по собственным векторам:

$$x(0) = \sum_{k=1}^{n} \alpha_k v_k,$$

$$x(t) = W^t x(0) = \sum_{k=1}^{n} \alpha_k \lambda_k^t v_k.$$

- При $t \to \infty$ остаётся только компонент при $\lambda_1 = 1$.
- **Спектральный зазор**:

$$\gamma = 1 - |\lambda_2|,$$

определяет скорость сходимости к консенсусу.

Модель Фридкина-Джонсона: упрямые агенты

- Часть агентов сохраняет исходные установки (упрямство агентов).
- Диагональная матрица податливости:

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \quad 0 \leq \lambda_i \leq 1.$$

Динамика в модели Фридкина-Джонсона

$$x(t+1) = \Lambda W x(t) + (I - \Lambda)x(0).$$

- Если спектральный радиус $\rho(\Lambda W) < 1$, существует стационарное состояние:

$$x^* = (I - \Lambda W)^{-1}(I - \Lambda)x(0).$$

- Консенсус не обязателен: возможно устойчивое расхождение мнений.

Модель ограниченного доверия

- Каждый агент усредняет мнения только «близких» по взглядам.
- Радиус доверия:

$$\varepsilon > 0.$$

- Множество доверенных соседей:

$$\mathcal{N}_i^{\varepsilon}(x(t)) = \{j : |x_j(t) - x_i(t)| \le \varepsilon\}.$$

Динамика в модели ограниченного доверия

$$x_i(t+1) = \begin{cases} \dfrac{1}{|\mathcal{N}_i^{\varepsilon}(x(t))|} \displaystyle\sum_{j \in \mathcal{N}_i^{\varepsilon}(x(t))} x_j(t), & \mathcal{N}_i^{\varepsilon} \neq \emptyset, \\ x_i(t), & \mathcal{N}_i^{\varepsilon} = \emptyset. \end{cases}$$

- Векторно:

$$x(t+1) = W(x(t))\, x(t),$$

где $W(x(t))$ зависит от текущего распределения мнений (нелинейность).

Эффекты модели ограниченного доверия

- В зависимости от ε:
  - **консенсус**, если ε достаточно велико;
  - **кластеризация** мнений (несколько устойчивых «островков»);
  - **поляризация** (два крупных кластера по разные стороны).
- Эти эффекты хорошо согласуются с наблюдаемыми эхо-камерами в соцсетях.

Модель Деффюанта-Вайсбуха (парами)

- Выбирается случайная пара соседей (i, j).
- Если мнения не слишком различаются:

$$|x_i(t) - x_j(t)| \leq \varepsilon,$$

происходит взаимный сдвиг (компромисс).

Правило обновления

$$x_i(t+1) = x_i(t) + \mu\big(x_j(t) - x_i(t)\big),$$
$$x_j(t+1) = x_j(t) + \mu\big(x_i(t) - x_j(t)\big),$$

где $0 < \mu \leq \frac{1}{2}$.

- Стохастическая динамика, приводящая к формированию кластеров мнений.

Подписанные графы и модель Алтафини

- Влияние может быть положительным (доверие) и отрицательным (вражда).
- Подписанная матрица смежности:

$$A = (a_{ij}), \quad a_{ij} \in \mathbb{R}.$$

$$a_{ij} > 0 \Rightarrow \text{дружба}, \qquad a_{ij} < 0 \Rightarrow \text{вражда}.$$

Подписанный лапласиан

$$D_{ii} = \sum_{j=1}^{n} |a_{ij}|, \quad L_s = D - A.$$

Непрерывная динамика:

$$\dot{x}(t) = -L_s x(t).$$

Структурный баланс и биполярный консенсус

- Граф структурно сбалансирован, если вершины можно разбить на две группы

$$V = V_1 \cup V_2,$$

  такие что:
  - внутри каждой группы рёбра преимущественно положительные;
  - между группами --- отрицательные.
- Тогда асимптотическое поведение:
  - либо консенсус: $x_i(t) \to c$ для всех i;
  - либо **биполярный консенсус**:

$$x_i(t) \to c \quad (i \in V_1), \qquad x_i(t) \to -c \quad (i \in V_2),$$

    для некоторого $c \neq 0$.
- Формализует феномен «двух враждебных лагерей» в обществе.

Идея графовых нейросетей

- Для каждого агента i вводится эмбеддинг:

$$h_i^{(0)} = \phi(x_i(0), f_i),$$

где f_i - признаки пользователя.

- На каждом слое l выполняется обмен сообщениями:

$$m_i^{(l)} = \mathrm{AGGREGATE}^{(l)}\big(\{h_j^{(l-1)} : j \in \mathcal{N}(i)\}\big),$$

$$h_i^{(l)} = \mathrm{UPDATE}^{(l)}\big(h_i^{(l-1)}, m_i^{(l)}\big).$$

Спектральная GCN-аппроксимация

- Матрица смежности с петлями:

$$\tilde{A} = A + I, \quad \tilde{D}_{ii} = \sum_j \tilde{A}_{ij}.$$

- На слое l:

$$H^{(l)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\,\tilde{D}^{-1/2}H^{(l-1)}W^{(l)}\right),$$

где:

  - $H^{(l)}$ - матрица эмбеддингов,
  - $W^{(l)}$ - обучаемые веса,
  - $\sigma(\cdot)$ - нелинейность (ReLU, tanh).

- Похоже на модель Де Гроота, но с обучаемыми и нелинейными операторами.

Графовые сети внимания (GAT)

- Веса влияния обучаются через механизм внимания.
- Коэффициенты внимания:

$$e_{ij}^{(l)} = a\big(Wh_i^{(l-1)}, Wh_j^{(l-1)}\big),$$

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum\limits_{k\in\mathcal{N}(i)} \exp(e_{ik}^{(l)})}.$$

- Обновление вектора:

$$h_i^{(l)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} W h_j^{(l-1)}\right).$$

- $\alpha_{ij}^{(l)}$ можно интерпретировать как *обучаемые коэффициенты влияния*.

Прогноз реакций и диффузии

- Пусть y_i - бинарная реакция (лайк/репост) на контент:

$$y_i \in \{0, 1\}.$$

- GNN даёт оценку:

$$\hat{y}_i = \sigma(u^\top h_i^{(L)}),$$

где u - обучаемый вектор.

- Функция потерь (кросс-энтропия):

$$\mathcal{L}(\theta) = -\sum_i \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right].$$

- После обучения можно:
  - оценивать «влияние» узлов;
  - моделировать диффузию контента и сценарии вмешательства.

LLM-агент как стохастическая политика

- Состояние агента i в момент t:

$$s_i(t) = \big(\mathcal{H}_i(t), \mathcal{F}_i(t), p_i\big),$$

где:

  - $\mathcal{H}_i(t)$ - история сообщений;
  - $\mathcal{F}_i(t)$ - лента;
  - p_i - параметры (роль, идеология).

- LLM задаёт распределение действий:

$$\pi_\theta\big(a_i(t) \mid s_i(t)\big) = \mathbb{P}_\theta\big(a_i(t) \text{ --- текст / действие} \mid \text{prompt}(s_i(t))\big).$$

- Состояние сети:

$$s(t+1) = F\big(s(t), a(t)\big), \quad a(t) = \{a_i(t)\}_{i=1}^{n}.$$

Извлечение числовых мнений из текста

- Определим отображение (классификатор)

$$g : \text{текст} \times \text{контекст} \to [-1, 1],$$

- Тогда:

$$x_i(t) = g\big(a_i(t),\, \mathcal{H}_i(t)\big) \in [-1, 1].$$

- Получаем траектории мнений LLM-агентов:

$$x_{\text{LLM}}(0), x_{\text{LLM}}(1), \ldots, x_{\text{LLM}}(T).$$

- Можно попытаться аппроксимировать их более простой моделью динамики.

Калибровка к линейной модели мнений

- Рассмотрим линейную модель Фридкина-Джонсона:

$$x_{\text{lin}}(t+1) = \Lambda W x_{\text{lin}}(t) + (I - \Lambda) x_{\text{lin}}(0).$$

- Подбираем параметры (W, Λ), минимизируя расхождение:

$$\min_{W,\Lambda} \sum_{t=0}^{T} \left\| x_{\text{LLM}}(t) - x_{\text{lin}}(t) \right\|_2^2.$$

- Итог:
  - получаем *эффективную матрицу влияния* W LLM-агентов;
  - можем применить классические результаты о консенсусе / поляризации.

Генеративное социальное моделирование соцсетей

- Лайки/репосты/подписки моделируются как действия LLM-агентов.
- Матрица подписок в момент t:

$$A(t) = (a_{ij}(t)), \quad a_{ij}(t) \in \{0, 1\}.$$

- Лента агента i:

$$\mathcal{F}_i(t) = \{a_j(\tau) : j \in \mathcal{N}_A(i),\ \tau \leq t\},$$

  с ранжированием по скору $R(a_j(\tau), i)$.
- Обновление подписок:

$$A_{ij}(t+1) = G_{ij}\big(A(t), s_i(t+1), a_i(t)\big).$$

- Возникает совместная динамика мнений $x(t)$ и структуры сети $A(t)$.

Гибридная макро-микро модель

- Микро-уровень: LLM-агенты

$$a(t) \sim \pi_\theta(\cdot \mid s(t)), \quad s(t+1) = F(s(t), a(t)).$$

- Макро-уровень: динамика мнений

$$x(t+1) = \mathcal{M}\big(x(t); \Theta_{\text{macro}}(a(0{:}t))\big),$$

где $\mathcal{M}$ - выбранная модель (Де Гроот, Алтафини и т.д.).

- Параметры Θ_{macro} (например, W, Λ, ε) оцениваются по данным симуляции.

Интервенции и про-социальные цели

- Пусть $J(x(T))$ - функционал качества:
  - поляризация:
$$J_{\text{pol}}(x) = \mathrm{Var}(x);$$
  - токсичность / радикализация и т.п.
- Интервенция u (изменение ранжирования, добавление «мостов» и т.д.) влияет на динамику:
$$x^{(u)}(T), \quad x^{(\emptyset)}(T).$$
- Эффект интервенции:
$$\Delta J(u) = J\big(x^{(u)}(T)\big) - J\big(x^{(\emptyset)}(T)\big).$$
- Задача: подобрать u так, чтобы
$$\Delta J(u) \ll 0,$$
при ограничениях на вмешательство.

Выводы

- Классические математические модели (DeGroot, Friedkin--Johnsen, HK, подписанные графы) дают:
  - строгие условия консенсуса, поляризации, кластеризации;
  - интерпретацию через спектр матриц влияния и структуру графа.
- Нейросетевые модели (GNN, LLM-агенты):
  - позволяют работать с реальными, сложными данными соцсетей;
  - дают реалистичную микродинамику поведения пользователей.
- Гибридные модели:
  - комбинируют интерпретируемость и выразительную мощность;
  - задают основу для оценки про-социальных интервенций.

Список литературы

- А. В. Проскурников, Р. Темпо. *A tutorial on modeling and analysis of dynamic social networks. Part I*, Annual Reviews in Control, Vol. 43 (Supplement C), 2017, с. 65–79.
- А. В. Проскурников, Р. Темпо. *A tutorial on modeling and analysis of dynamic social networks. Part II*, Annual Reviews in Control, Vol. 46 (Supplement), 2018
- Ч.-Ю. Чэнь, С. Саха, М. Бансал., *ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs*, arXiv:2309.13007 [cs.CL], 2023 (rev. 2024).
- М. Ларой, П. Тёрнберг., *Can We Fix Social Media? Testing Prosocial Interventions using Generative Social Simulation*, arXiv:2508.03385 [cs.SI], 2025.

Спасибо за внимание!