

Чем старше человек,  
тем больше ему лет

---

## Методы построения социо-демографических профилей пользователей сети Интернет

Андрей Гомзин

Институт Системного программирования РАН  
Факультет ВМК МГУ

# Содержание

Введение

Постановка задачи

Краткий обзор существующих решений

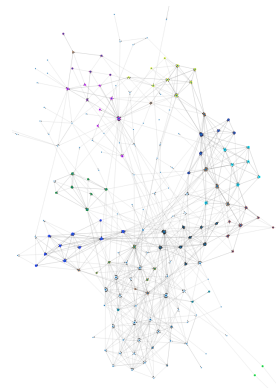
Предлагаемый подход

Эксперименты

Заключение

# Введение. Анализ социальных сетей

- ▶ Сбор пользовательских данных из социальных сервисов
- ▶ Генерация случайных графов, обладающих основными свойствами социальных сетей
- ▶ Измерение информационного влияния между пользователями в социальных сетях.
- ▶ Идентификация пользователей различных социальных сетей
- ▶ Поиск неявных сообществ пользователей на основе социальных связей между ними
- ▶ **Определение демографических атрибутов пользователей**
- ▶ **Поиск искусственно созданных аккаунтов (ботов)**



# Введение. Социо-демографический профиль

**Социо-демографические характеристики пользователей:**  
пол, возраст, семейное положение, уровень образования.

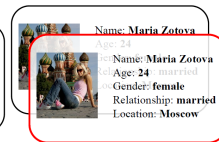
- ▶ Не все поля заполняются пользователями
- ▶ Указываются неверные значения



*атрибуты не  
указаны*



*ошибки*



*дубликаты  
профилей*

# Введение. Целевая аудитория

**Целевая аудитория** – группа людей, объединённых общими признаками, или объединённой ради какой-либо цели или задачи<sup>1</sup>

Явно указанные и извлеченные социо-демографические атрибуты используются в рекомендательных и маркетинговых системах для:

- ▶ определения целевой аудитории продукта
- ▶ поиска потенциальных потребителей

---

<sup>1</sup>Википедия: [https://ru.wikipedia.org/wiki/Целевая\\_группа](https://ru.wikipedia.org/wiki/Целевая_группа)

# Введение. Гомофилия и поведение

Для определения демографических атрибутов используются поведенческие признаки и свойство гомофилии.

**Гомофилия** – тенденция индивидов создавать и поддерживать связи с похожими на них.

- ▶ Отношение дружбы в соцсетях

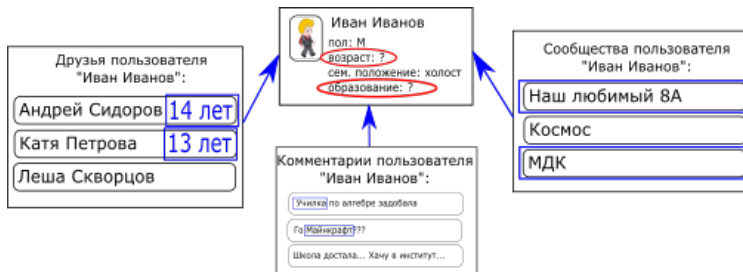
Поведенческие признаки:

- ▶ Сообщения
- ▶ Подписка на сообщества
- ▶ оценки (нравится / не нравится)

# Постановка задачи (1)

**Дано:** социальная сеть

**Найти:** неуказанные значения демографических атрибутов пользователей



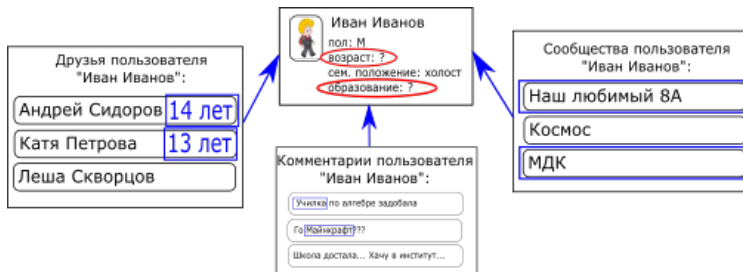
Демографические атрибуты:

- ▶ пол
- ▶ возраст
- ▶ семейное положение
- ▶ уровень образования
- ▶ ...

## Постановка задачи (2)

**Дано:** социальная сеть

**Найти:** неуказанные значения демографических атрибутов пользователей



Анализируемые данные:

- ▶ отношение дружбы
- ▶ подписка на сообщества
- ▶ тексты сообщений
- ▶ значения атрибутов
- ▶ ...

# Краткий обзор существующих решений

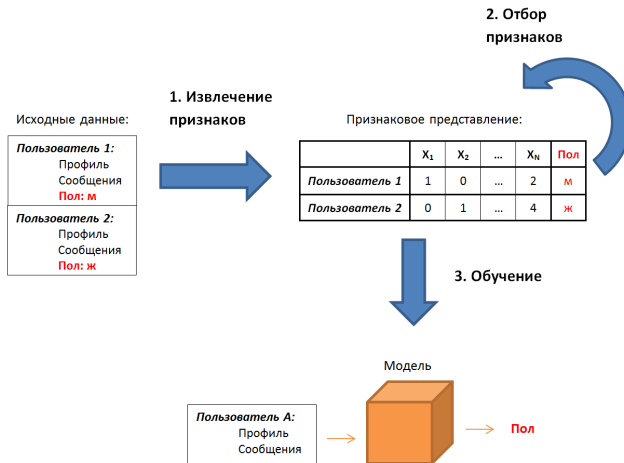
## Данные:

- ▶ Twitter
- ▶ Facebook
- ▶ Youtube
- ▶ e-mail

## Подходы:

- ▶ Эвристики (пол по имени)
- ▶ Однородные данные -> Машинное обучение с учителем
- ▶ Графы -> Кластеризация графов, векторное представление вершин
- ▶ Подходы, использующие разнородные данные

# Машинное обучение с учителем



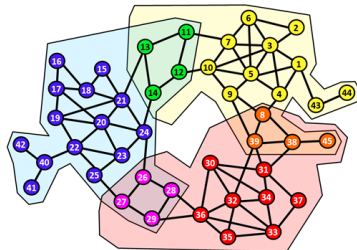
# Машинное обучение с учителем. Признаки

## Текстовые признаки:

- ▶  $n$ -граммы:  
слова, символы, части речи
- ▶ статистические
  - ▶ средняя длина сообщения
  - ▶ LIWC – Linguistic Inquiry and Word Count:
    - ▶ доля эмоций (+/-)
    - ▶ доля местоимений
    - ▶ доля длинных слов ( $|\cdot| > 6$ )
- ▶ ресурсозависимые:
  - ▶ #хештеги
  - ▶ @упоминания

## Графовые признаки:

- ▶ найденные (нечеткие) кластеры:



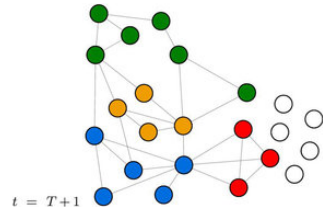
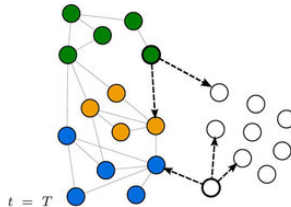
## Прочие ресурсозависимые признаки:

- ▶ Цвет фона/текста/ссылок (Twitter)

# Кластеризация социального графа

Распространение меток в графе:

1. инициализация узлов метками
2. несколько итераций распространения меток соседям:
  - ▶ Задается стратегия отправки меток
  - ▶ Задается стратегия приема меток



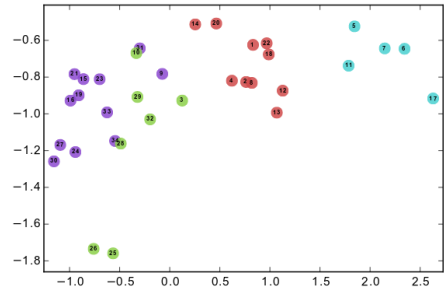
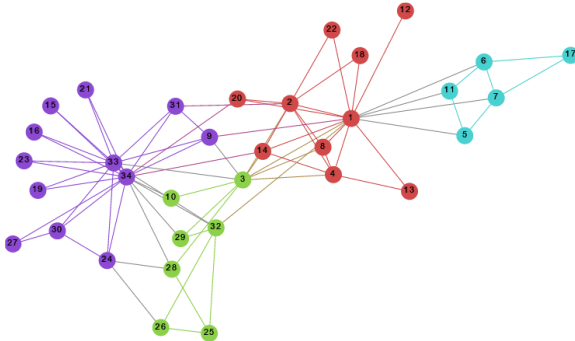
Определение демографических атрибутов с помощью кластеризации:

**Кластер объединяет в себя пользователей с одинаковым набором атрибутов.**

Для этого инициализируем метки согласно известным значениям атрибута (например, "М" и "Ж")

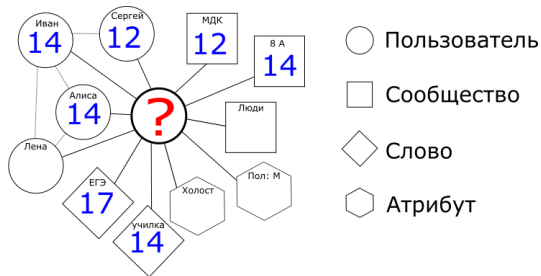
## Векторное представление вершин графа

- ▶ Каждая вершина  $v$  графа  $G(V, E)$  отображается в  $R^n$
- ▶ Близость между узлами графа сохраняется в  $R^n$ .



## Предлагаемый подход

- ▶ Построение социо-лингвистического графа
- ▶ Распространение меток



# Социо-лингвистический граф

## Узлы:

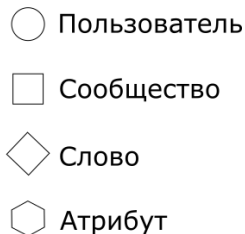
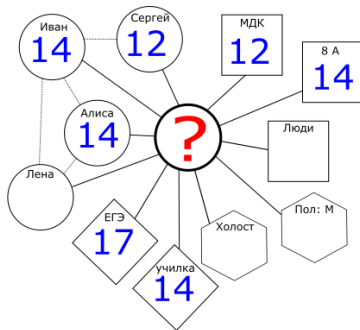
- ▶ пользователи
- ▶ сообщества
- ▶ текстовые признаки (слова)
- ▶ значения атрибутов

## Ребра:

- ▶ отношение дружбы
- ▶ подписка на сообщества
- ▶ употребление слов
- ▶ явно указанные значения

## Метки:

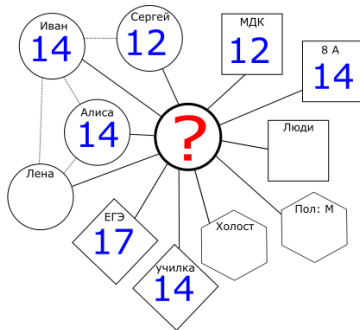
- ▶ значения атрибута



# Алгоритм определения значений атрибутов

Схема алгоритма:

1. Инициализация (разметка)
2. Распространение меток от узлов-пользователей к другим узлам (сообщества, слова, атрибуты)
3. Распространение меток от всех узлов к узлам-пользователям



- Пользователь
- Сообщество
- ◇ Слово
- ⬡ Атрибут

# Разметка (Вконтакте). Возраст, семейное положение

## ▶ Возраст

- ▶ Возраст пользователя извлекается из даты его рождения. Поле «дата рождения» может быть представлено в трех вариантах:

1. **DD-MM-YYYY** - доступна полная дата
2. **YYYY** - доступен год рождения
3. **DD-MM** - доступна дата без года

## ▶ Семейное положение

- ▶ Поле “семейное положение” пользователя Вконтакте принимает следующие значения:

- ▶ **1 — не женат/не замужем; => «не в официальном браке»**
- ▶ **2 — есть друг/есть подруга; => «не в официальном браке»**
- ▶ 3 — помолвлен/помолвлена;
- ▶ **4 — женат/замужем; => «в официальном браке»**
- ▶ 5 — всё сложно;
- ▶ **6 — в активном поиске; => «не в официальном браке»**
- ▶ 7 — влюблён/влюблена;
- ▶ 0 — не указано.

# Разметка (Вконтакте). Уровень образования

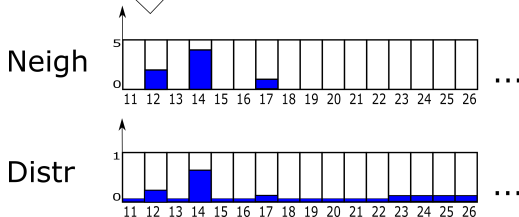
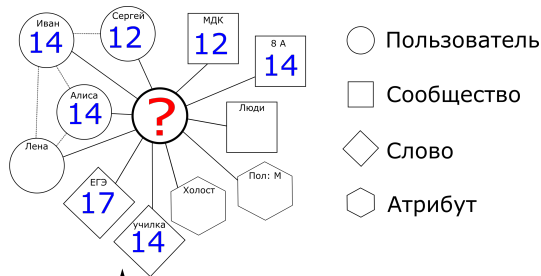
## ► Уровень образования

- Максимальный год окончания школы => среднее образование
- Минимальный год окончания университета => высшее образование
- Атрибут «Образование» по таблице:

|                         |             | окончание вуза |             |            |
|-------------------------|-------------|----------------|-------------|------------|
|                         |             | неизвестно     | отсутствует | имеется    |
| окончание средней школы | неизвестно  | неизвестно     | среднее     | высшее     |
|                         | отсутствует | отсутствует    | отсутствует | неизвестно |
|                         | имеется     | среднее        | среднее     | высшее     |

## Стратегия приема меток. Идея

1. Вычисляем распределение значений атрибута соседей.
2. Зашумляем его. Это фильтрует узлы-признаки с малым количеством соседей (например, редкие слова, которые используются 2-3 пользователями)
3. Вычисляем, насколько в распределении одно из значений преобладает; если сильно преобладает, то присваиваем метку узлу, если нет — то узел остается без метки.



## Стратегия приема меток (1)

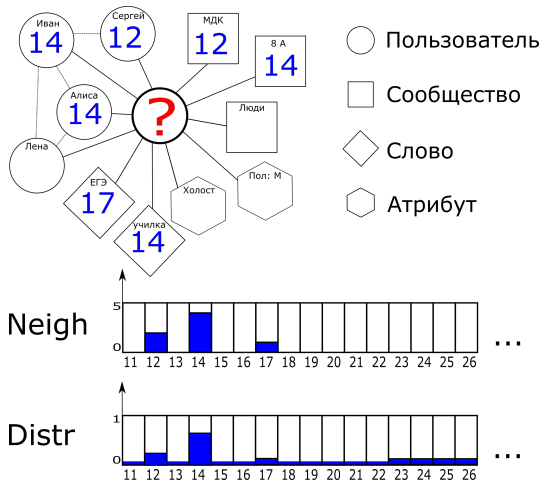
$Gen$  – распределение значений атрибута на всех данных.  $\|Gen\| = 1$

1.  $Neigh$  – распределение значений атрибута соседей.  $Neigh(v)$  – количество соседей
- 2.

$$Distr(v) = \frac{Neigh(v) + w \times Gen(v)}{N}$$

где  $N$  подбирается таким образом, что:

$$|Distr| = \sqrt{\sum_{v \in V} Distr^2(v)} = 1$$



## Стратегия приема меток (2)

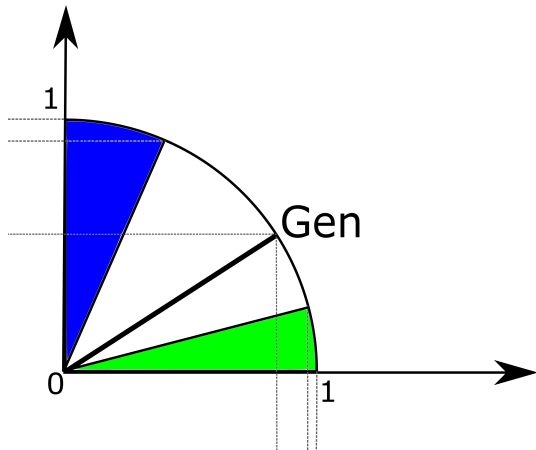
$Gen$  – распределение значений атрибута на всех данных.  $\|Gen\| = 1$

3. Вычисляем метку  $val$  и вес  $S$ :

$$s = \max_v \frac{Distr(v) - Gen(v)}{1 - Gen(v)}$$

$$val = \arg \max_v \frac{Distr(v) - Gen(v)}{1 - Gen(v)}$$

4. Узел принимает соответствующую метку, если  $s > t$ .



# Эксперименты. Данные

Социальная сеть: Вконтакте.

Данные:

- ▶ 50M профилей
- ▶ 50M списков друзей
- ▶ 1M списков подписчиков сообществ
- ▶ тексты комментариев за 0.5 года из 1M активных сообществ

## Эксперименты. Гомофилия vs Поведение

- ▶ Измерялись значения достоверности (ассурасу) и F1-меры
- ▶ Точность для атрибута "Возраст". Возраст определен точно, если:

$$|act - pred| < 0.15act$$

- ▶ Эксперименты:
  - ▶ Только узлы-пользователи:  
(>20 друзей, хотя бы у одного указан атрибут)
  - ▶ Только узлы-сообщества:  
(>20 сообществ)
  - ▶ Только узлы-слова:  
(>20 слов)
  - ▶ Пользователи, сообщества, слова:  
(пересечение всех предыдущих)
- ▶ Параметры:
  - ▶ Распространение от пользователей к другим узлам:  $T = 0.8$ ,  $W = 2.0$
  - ▶ Распространение к пользователям (предсказание):  $T = 0.0$ ,  $W = 0.0$

# Эксперименты. Гомофилия vs Поведение. Возраст

|                    | Пользователи | Сообщества | Слова | Все  |
|--------------------|--------------|------------|-------|------|
| MAE                | 1.66         | 1.13       | 5.44  | 1.76 |
| Асс (15%)          | 0.91         | 0.91       | 0.53  | 0.90 |
| Доля пользователей | 0.99         | 0.02       | <0.01 | 0.99 |

## Эксперименты. Гомофилия vs Поведение. Семейное положение

|                    | Пользователи | Сообщества | Слова | Все  |
|--------------------|--------------|------------|-------|------|
| Acc                | 0.70         | 0.72       | 0.64  | 0.71 |
| F1macro            | 0.65         | 0.67       | 0.63  | 0.66 |
| Доля пользователей | 0.98         | 0.81       | 0.83  | 0.98 |

# Эксперименты. Гомофилия vs Поведение. Образование

|                    | Пользователи | Сообщества | Слова | Все  |
|--------------------|--------------|------------|-------|------|
| Acc                | 0.59         | 0.63       | 0.47  | 0.58 |
| F1macro            | 0.47         | 0.44       | 0.37  | 0.48 |
| Доля пользователей | 0.99         | 0.35       | 0.46  | 0.98 |

# Планы

- ▶ **Сравнение с методами, использующими разнородные данные**
- ▶ Национальность, доход, дети, языки
- ▶ Лайки, репосты
- ▶ Стены пользователей и комментарии (вместо комментариев в группах)
- ▶ Определение неверно указанных значений

# Заклучение

- ▶ Гомофилия, поведение, социо-демографические профили пользователей
- ▶ Метод построения социо-демографического профиля
- ▶ Выводы:
  - ▶ Гомофилия (друзья) хорошо работает для определения возраста
  - ▶ Нужно улучшать текстовые признаки
  - ▶ Параметрами алгоритма можно регулировать точность/покрытие
  - ▶ Нужно фильтровать данные

# Спасибо за внимание