Московский физико-технический институт
(государственный университет)

Факультет управления и прикладной математики

Кафедра системного программирования

---

Ефремова Мария Александровна

# Использование распределения подграфов в графе для определения демографических атрибутов пользователей сети Интернет

Выпускная квалификационная работа бакалавра

---

***Научные руководители:*** к.ф.-м.н. Турдаков Денис Юрьевич
Дробышевский Михаил Дмитриевич

Москва 2018

# Содержание

# 1 Введение

Пользователи сети Интернет принимают активное участие в создании контента, например: оставляют комментарии в социальных сетях, пишут отзывы на товары интернет-магазинов, ведут блоги и общаются на форумах. Однако многие ресурсы дают возможность не только делиться текстовой информацией, но и оставлять персональные данные - заполнять профиль. Как правило, к таким данным относятся имя, возраст, пол, контактная информация, интересы и прочее.

Информация в профиле, может оказаться неполной, например некоторые поля могут быть необязательными для заполнения, часто их оставляют пустыми. Более того, некоторые пользователи намеренно оставляют неверные данные. Возникает задача автоматического предсказывания недостающих демографических атрибутов, так как, к примеру, для проведения социологических исследований, а также для работы гибридных и основанных на знаниях рекомендательных систем и таргетированной рекламы необходим наиболее полный набор характеристик [1, 2].

Во многих работах используются методы машинного обучения для решения этой задачи. Сначала осуществляется сбор данных для построения модели, после чего производится обучение модели, а затем установление неизвестных атрибутов с помощью полученной модели и оценка её качества [3].

В данной работе рассматривается задача предсказания демографических атрибутов пользователей социальной сети ВКонтакте; для признакового описания объектов вводятся признаки, использующие распределения подграфов в графах.

# 2 Постановка задачи

Цель исследования - ввести распределения подграфов в качестве нового признака для описания объектов (пользователей) и проверить гипотезу о том, что с их помощью можно предсказывать демографические атрибуты, а также сравнить с другими методами - например такими, как использование распределения атрибутов соседей и распространение меток - и выяснить, как введённый признак влияет на качество предсказания.

Для достижения поставленной цели необходимо решить следующие подзадачи:

- Выбрать пользователей с полным набором известных атрибутов (пол, возраст, семейное положение, образование) для обучающей выборки;
- Для пользователей из обучающей выборки постороить графы социальных связей до второй окрестности. Каждому пользователю ставится в соответствие граф $G(V, E)$, в котором $v \in V$ - это либо сам исследуемый пользователь, либо пользователь из списка его друзей и друзей друзей, а ребро $(u, v) \in E$ - дружественная связь между пользователями $u$ и $v$;
- Построить вектора-распределения подграфов в полученных графах;
- Построить классификационные модели, где в качестве признакового описания объектов рассматриваются признаки, использующие посчитанные ранее распределения подграфов;
- Произвести сравнение качества предсказания других методов решения задачи и данного.