# Определение места проживания пользователей социальных сетей

## на основе социального графа

Юля Трофимович, Илья Козлов, Денис Турдаков

1 декабря 2016 г.

ИСП РАН

# Table of contents

# Определение местоположения: о задаче

# Определение местоположения: о задаче

Дано:

- социальный граф (ВКонтакте), 156 057 700 вершин
- информация о месте проживания указанная пользователями (известна для 63% пользователей)

Найти:

- основное место проживания пользователей, которые не указали его в профиле

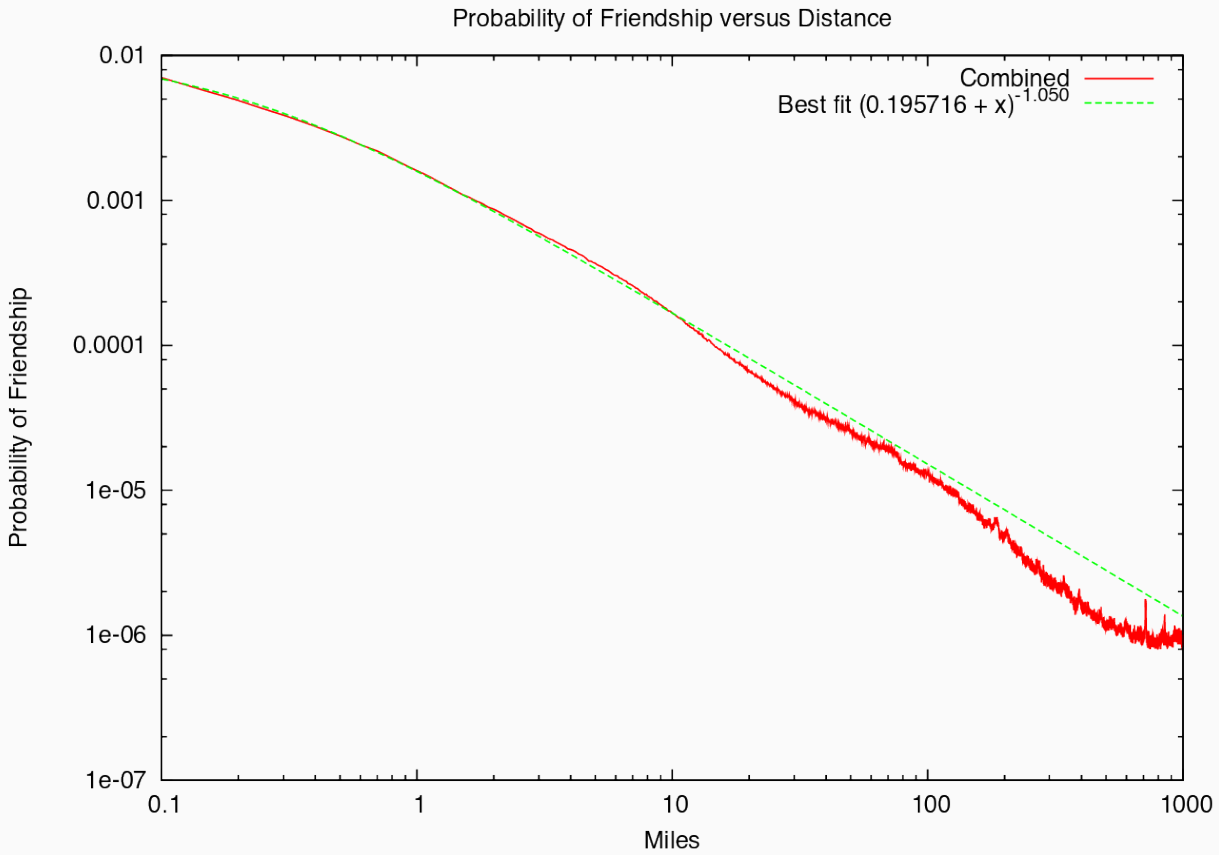
Точность:

- субъект РФ/страна

# Что было сделано до нас

# Первые попытки: вероятностные модели

Backstrom, Sun и Marlow (2010):




Probability of friendship as a function of distance

# Первые попытки: вероятностные модели

Backstrom, Sun и Marlow (2010):

$$p(|l_u - l_v|) = 0.0019(|l_u - l_v| + 0.196)^{-1.05} \tag{1}$$

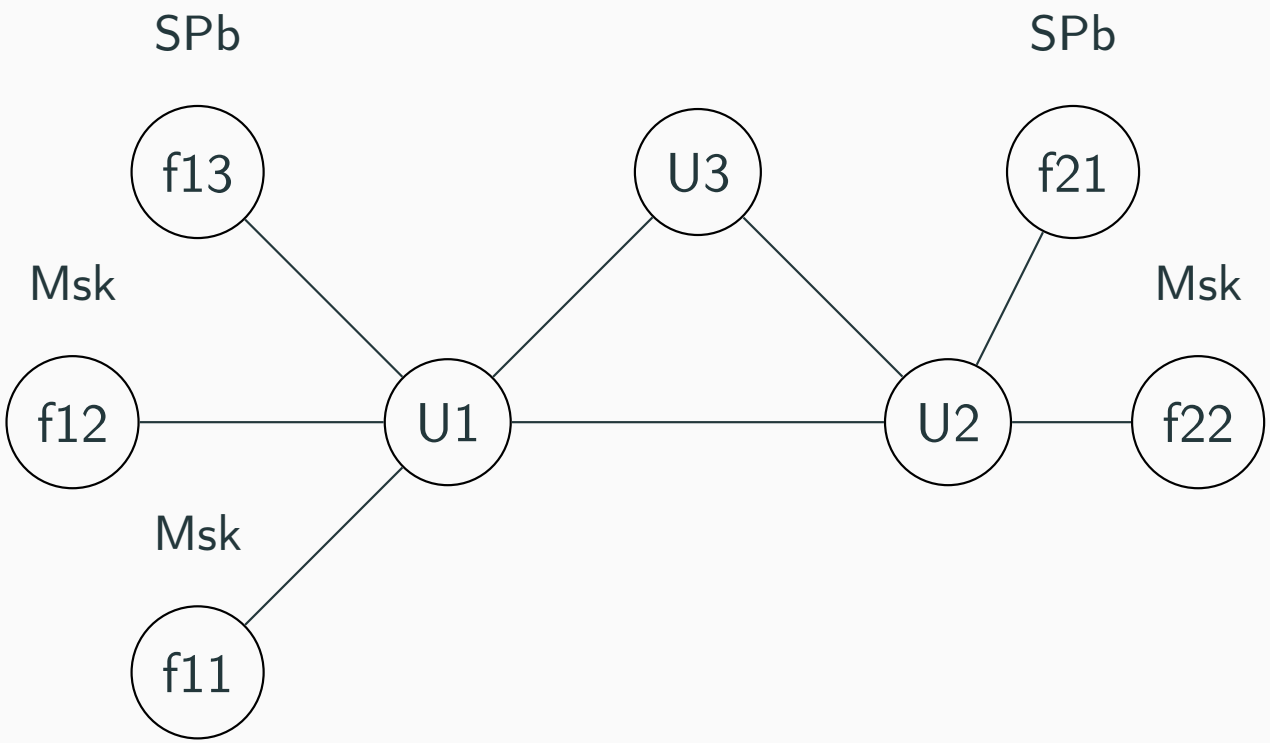
$$\prod_{(u,v)\in E} p(|l_u - l_v|) \prod_{(u,v)\notin E} 1 - p(|l_u - l_v|) \tag{2}$$

# Label propagation для определения местоположения

Jurgens (2013) “That’s What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships.”:

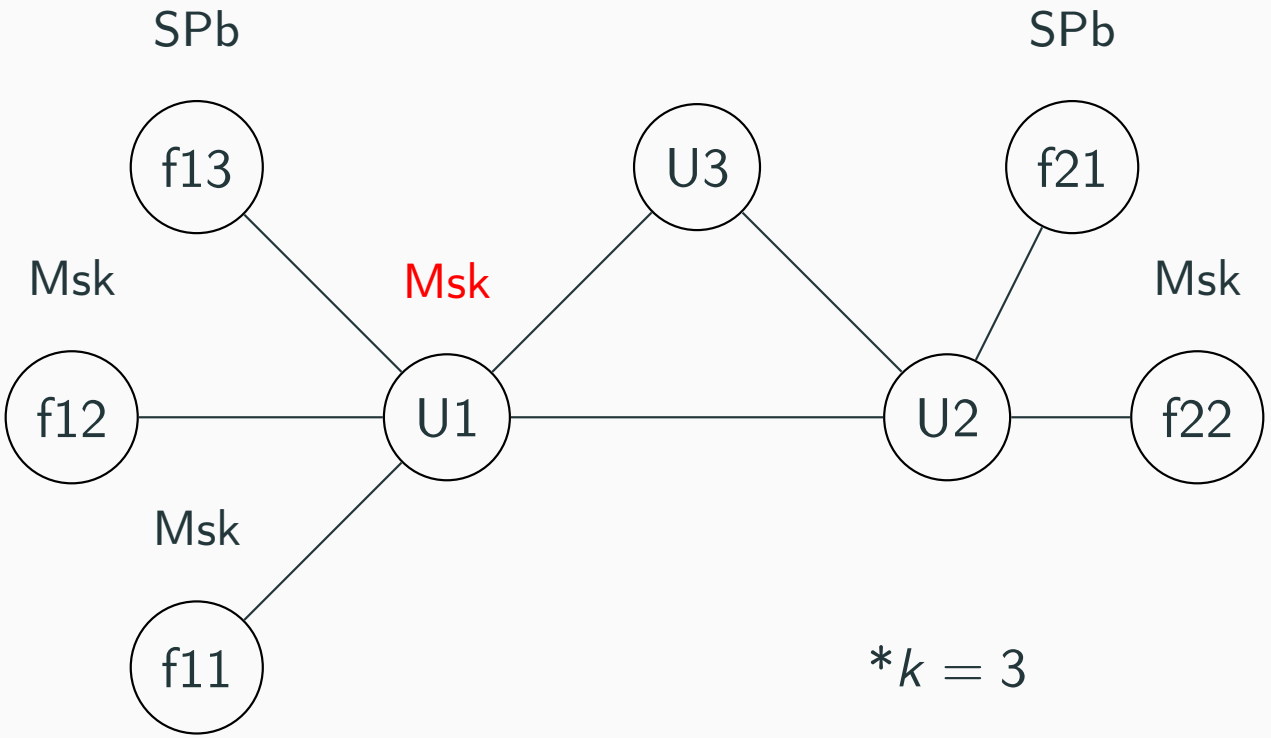
- Инициализация: присвоение меток вершинам
- Итеративно:
  - Каждая вершина отсылает свою метку соседним вершинам
  - Каждая вершина обрабатывает полученные метки, и, при необходимости, обновляет свою
- Окончание работы: если в ходе итерации ни одна вершина не обновила свою метку, работа алгоритма заканчивается

# Label propagation для определения местоположения



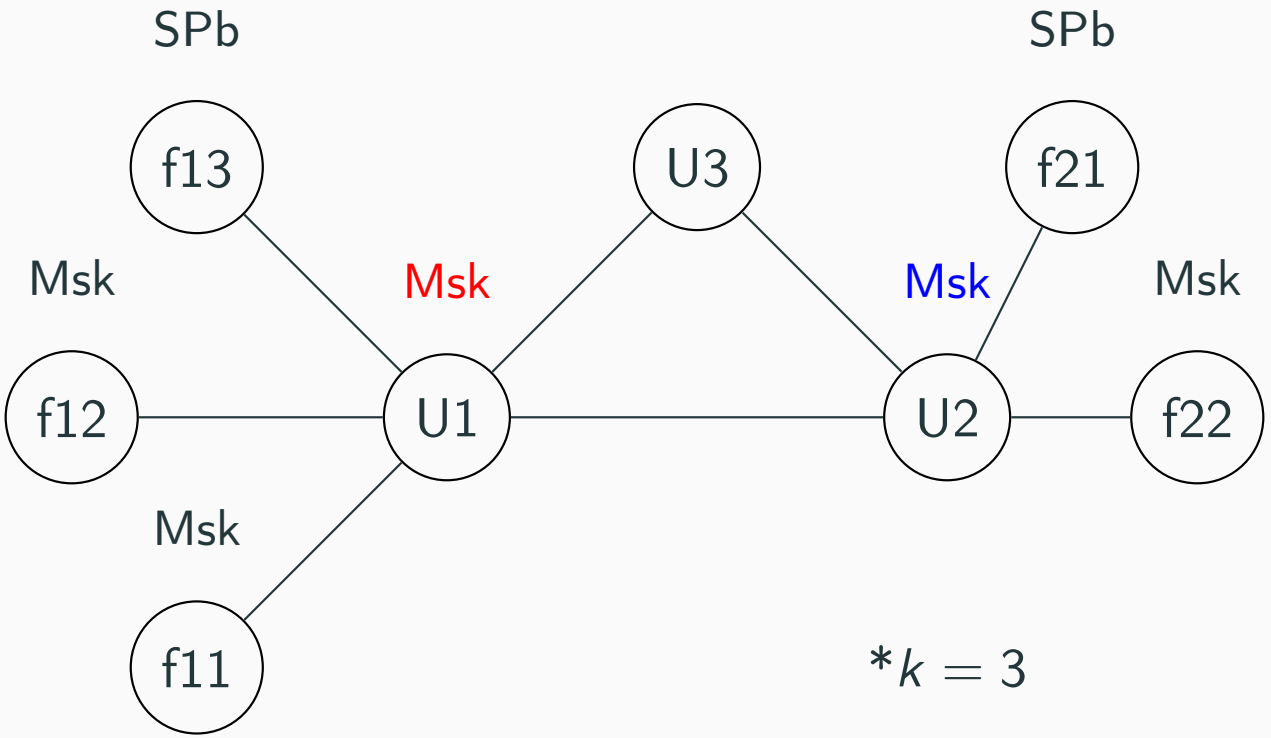

Пример социального графа

# Label propagation для определения местоположения



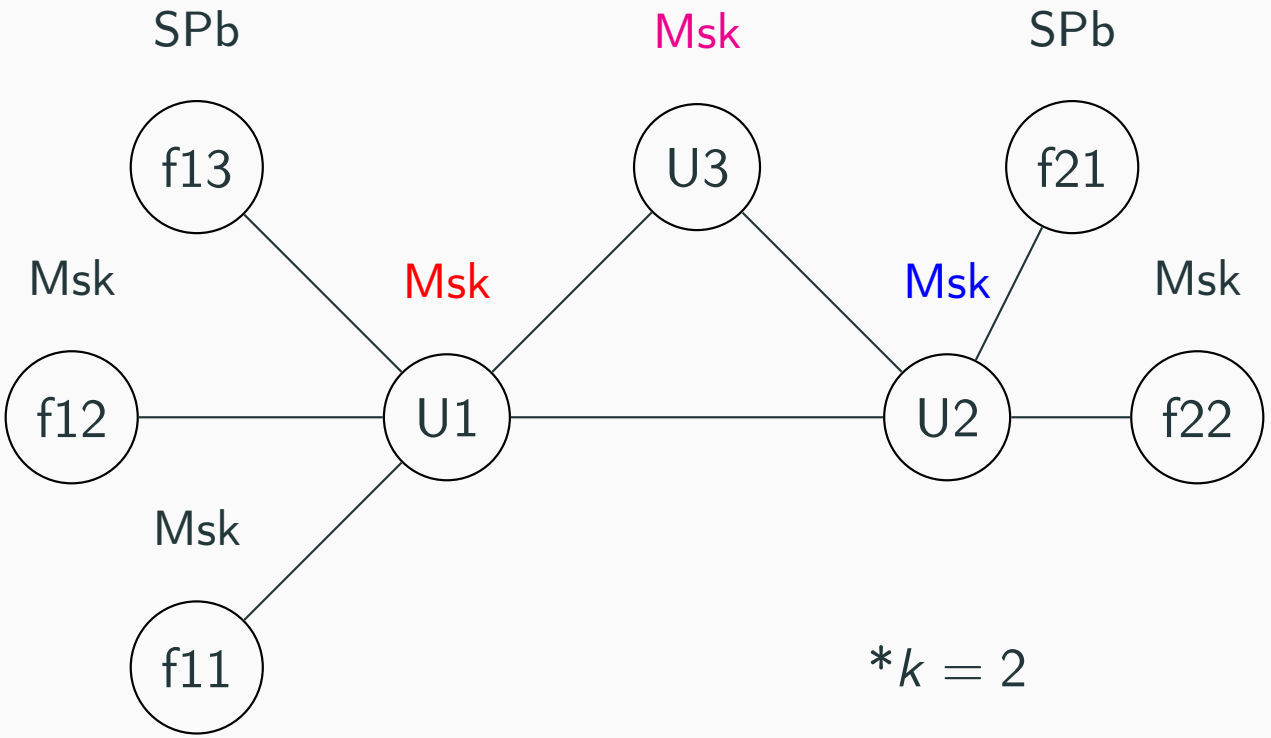

Социальный граф после первой итерации подхода на основе распространения меток

# Label propagation для определения местоположения




Социальный граф после второй (и третьей тоже) итерации подхода на основе распространения меток

# Label propagation для определения местоположения




Социальный граф после четвертой итерации подхода на основе распространения меток

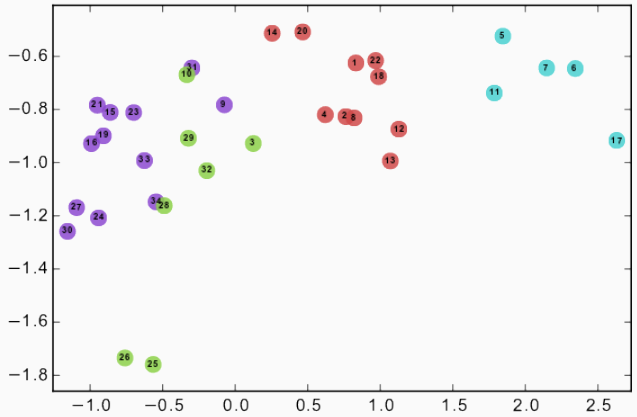
# Machine learning для определения местоположения

- SVM
  - Rout и др. (2013): City size, City population bins, Triads
- PGM
  - Chen, Liu и Zou (2016): Tie strength (Random Walk with Restart)
  - Jia и др. (2016): Influential friends (Sequential Random Walk with Restart)

# Наш подход

# Graph node embeddings

(a) Input: Karate Graph

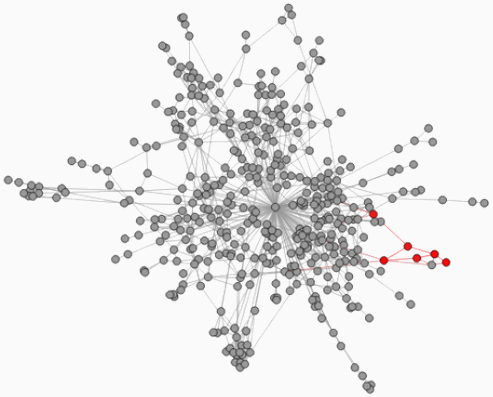



(b) Output: Representation

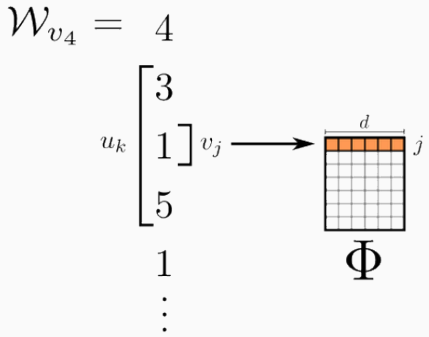
Вложение Zachary's Karate Club в плоскость

# Graph node embeddings: DeepWalk

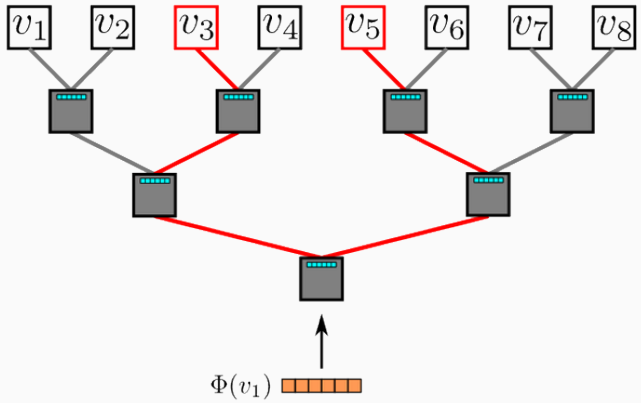
Perozzi, Al-Rfou и Skiena (2014): **DeepWalk**



(a) Random walk generation.




(b) Representation mapping.




(c) Hierarchical Softmax.

Overview of DeepWalk

# Graph node embeddings: BLM

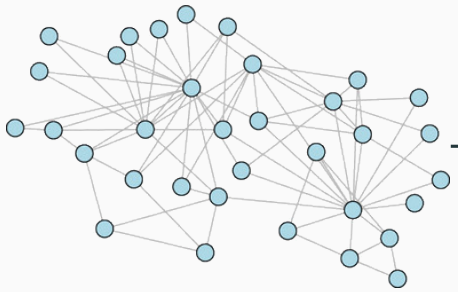
Ivanov и Bartunov (2015): **Bilinear Link Model**

$$\vec{In} \in \mathbb{R}^D, \vec{Out} \in \mathbb{R}^D \tag{3}$$

Вероятность ребра между вершинами $v \to u$

$$p(v|u) = \frac{\exp(In_u^T Out_v)}{\sum_{w \in V} \exp(In_u^T Out_w)} \tag{4}$$

# Подход на основе graph node embeddings



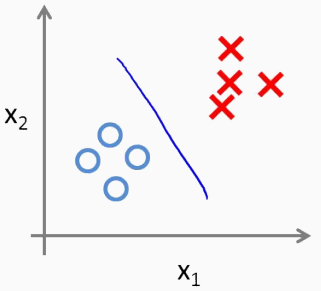
Social graph

BLM

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{d1} & x_{d2} & \dots & x_{dn} \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_d \end{bmatrix}$$

Graph embeddings

ML




Classification model*

* Мы использовали нейронную сеть - многослойный персептрон

# Данные и эксперимент

- Размер датасета: 99 055 808 пользователей
- Обучающая выборка: 15 003 815 ($\sim$15%) выбранных случайным образом пользователей
- Тестовые выборки:
  (a) 84 051 993 ($\sim$85%) - пользователи не вошедшие в обучающую выборку
  (b) 4 588 844 ($\sim$4.6%) - пользователи не вошедшие в обучающую выборку и оставившие хотя бы одно сообщение в топ-миллионе активных групп
- Количество классов: 277

# Результаты

(a) Пользователи **вне зависимости от наличия сообщений**

| Метрика | | Подход | |
|---|---|---|---|
| | | LP | GE |
| Accuracy | | 0.463 | **0.516** |
| Macro | Precision | **0.522** | 0.308 |
| | Recall | 0.150 | **0.167** |
| | F1-score | 0.192 | **0.203** |
| Weighted | Precision | **0.760** | 0.729 |
| | Recall | 0.463 | **0.516** |
| | F1-score | 0.555 | **0.595** |

Обозначения: LP - подход на основе распространения меток, GE - подход на основе векторного представления

# Результаты

(b) Только пользователи, оставившие **хотя бы одно сообщение**

| Метрика | | Подход | |
|---|---|---|---|
| | | LP | GE |
| Accuracy | | 0.786 | **0.839** |
| Macro | Precision | **0.374** | 0.302 |
| | Recall | 0.241 | **0.268** |
| | F1-score | 0.255 | **0.271** |
| Weighted | Precision | 0.819 | **0.823** |
| | Recall | 0.786 | **0.840** |
| | F1-score | 0.788 | **0.827** |

Обозначения: LP - подход на основе распространения меток, GE - подход на основе векторного представления

# Спасибо за внимание ^_^

**The End**