О задаче классификации пользователей социальной сети Одноклассники на основе анализа графа друзей

Алексей Зиновьев, аспирант ОмГУ

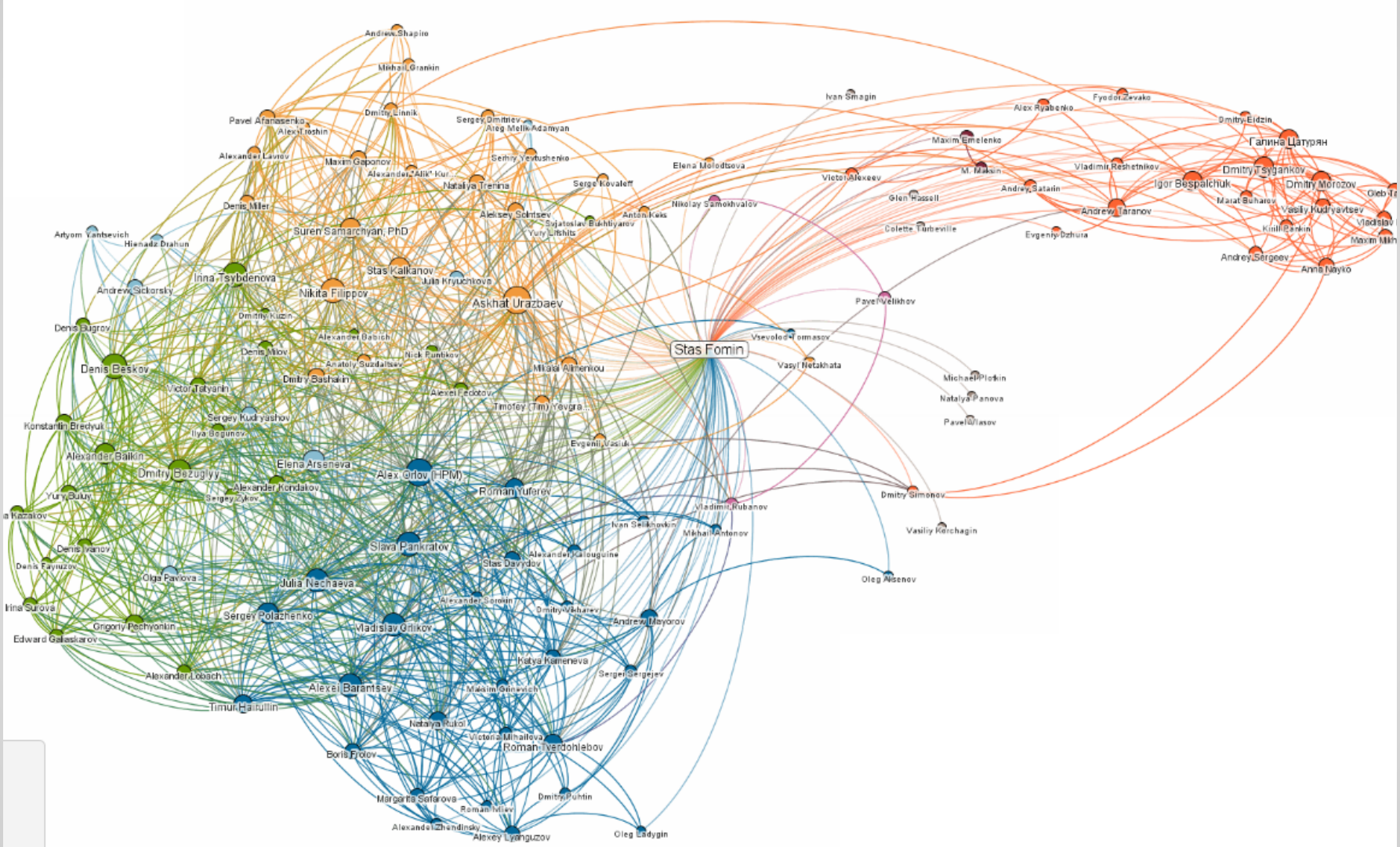
Параметры соц.сети

В целом:

- 200 000 000 пользователей
- 8 500 000 сообществ

В сутки:

- 40 000 000 пользователей
- 250 000 000 сообщений
- 8 000 000 постов
- 12 000 000 фотографий
- 7 000 000 дружб


Andrew Shapiro
Mikhail Grankin
Pavel Afanasenko
Alexei Toshin
Dmitry Linnik
Sergey Dmitriev
Aleg Melik-Adamyan
Alexander Lavov
Maxim Gaponov
Alexander "Alik" Kur
Nataliya Trenina
Semyr Yevtushenko
Serge Kovaleff
Denis Miler
Aleksey Solntsev
Sviatoslav Bukhtiyarov
Yuriy Litshits
Anton Keks
Suren Samarchyan, PhD
Artyom Yantsevich
Hienadz Drahun
Inna Tsybdenova
Stas Kalkanov
Julia Kryuchkova
Nikita Filippov
Askhat Urazbaev
Andrew Sickorsky
Dmitry Kuzin
Denis Bugrov
Alexander Babich
Denis Milov
Nick Puntikov
Anatoly Suzdaltsev
Denis Beskov
Dmitry Bashkirin
Victor Tsiryanin
Alexei Fedotov
Mikaela Almeniou
Timotey (Tim) Yevgra...
Sergey Kudryashov
Konstantin Bredyuk
Iliya Begunov
Alexander Balkin
Elena Arseneva
Dmitry Bezuglyy
Alex Orlov (HPM)
Alexander Kondakov
Sergey Zykov
Yuriy Buluy
Roman Yuferev
Evgeniy Vasiuk
Ivan Selikhovkin
Mikhail Antonov
Vladimir Rubanov
Slava Pankratov
Alexander Kalouguine
Stas Davydov
Denis Ivanov
Denis Fayruzov
Olga Pavlova
Julia Nechaeva
Irina Surova
Sergey Polazhenko
Grigoriy Pechyonkin
Edward Galaskarov
Vladislav Grikov
Alexander Sorokin
Dmitry Vikharev
Andrey Mayorov
Katya Kameneva
Sergei Sergejev
Alexander Lobach
Alexei Baramtsev
Maxim Grinevich
Timur Haitullin
Natalya Kulot
Victoria Mihailova
Roman Tverdohlebov
Boris Frolov
Margarita Seferova
Dmitry Puhtin
Roman Iviev
Alexander Zhendinsky
Alexey Lyanguzov
Oleg Ladygin
Ivan Smagin
Fyodor Zevako
Alex Ryabenko
Dmitry Eidzin
Maxim Emelenko
Гаянна Цатурян
Elena Molodtsova
Victor Alexeev
M. Maksin
Vladimir Reshetnikov
Dmitry Tsygankov
Igor Bespalchuk
Dmitry Morozov
Andrey Satarin
Glen Hassell
Marat Buharov
Andrew Taranov
Vasiliy Kudryavtsev
Colette Turbeville
Evgeniy Dzhura
Kirill Rankin
Andrey Sergeev
Anna Nayko
Pavel Velikhov
Stas Fomin
Vsevolod Tomasov
Vasyl Netakhata
Michael Plotkin
Natalya Panova
Pavel Nasov
Dmitry Simonov
Vasiliy Korchagin
Oleg Alsenov

Вредоносная деятельность

- Нарушение законов этики, морали и статей УК РФ
- Создание “теневой” подсети спам-аккаунтов
- Взлом профилей реальных пользователей
- Рассылка спама со своих и взломанных профилей
- Привлечение внимание при помощи появления в списке посетивших страницу пользователя

Выгода социальной сети

- Предотвращаем распространение “эпидемий” по взлому профилей и утечку персональных данных
- Предотвращаем спам до его появления
- Уменьшаем количество жалоб
- Снижаем нагрузку на модераторов
- Сокращаем штат модераторов

Данные

- Граф связей (~ 9 * 10^6, 39 Гб)
- Демография
- “Лайки/Классы” пользователей
- История “логинов” (включая неудачные попытки авторизации) (~ 3,2 * 10^8, 12 Гб)
- Посты в группах
- Жалобы на спам

Набор инструментов для анализа

- R 3.0.3 (прототипирование моделей)
- python 23 + scypi + numpy + pandas (моделирование)
- Hadoop 2.6 (кластерная инфраструктура)
- Pig 14 (вычисление неграфовых признаков)
- Giraph 1.1 (вычисление графовых признаков)

Производственная задача

Необходимо на основании графа друзей, истории заходов в соц.сеть, а также анализа иной деятельности предложить мат.модель, позволяющую с высокой достоверностью (ошибка первого рода не более 1% и ошибка второго рода не более 10%) определять, что пользователь является злоумышленником.

Существующее на данный момент решение имеет ошибку первого рода более 10% и ошибку второго рода более 20%.

Задача

Множество объектов - пользователи соц.сети.

Каждый из объектов необходимо отнести к одному из двух следующих классов: User, Spamer.

Учебная выборка составлена на основе жалоб пользователей на реальных злоумышленников.

Основные признаки

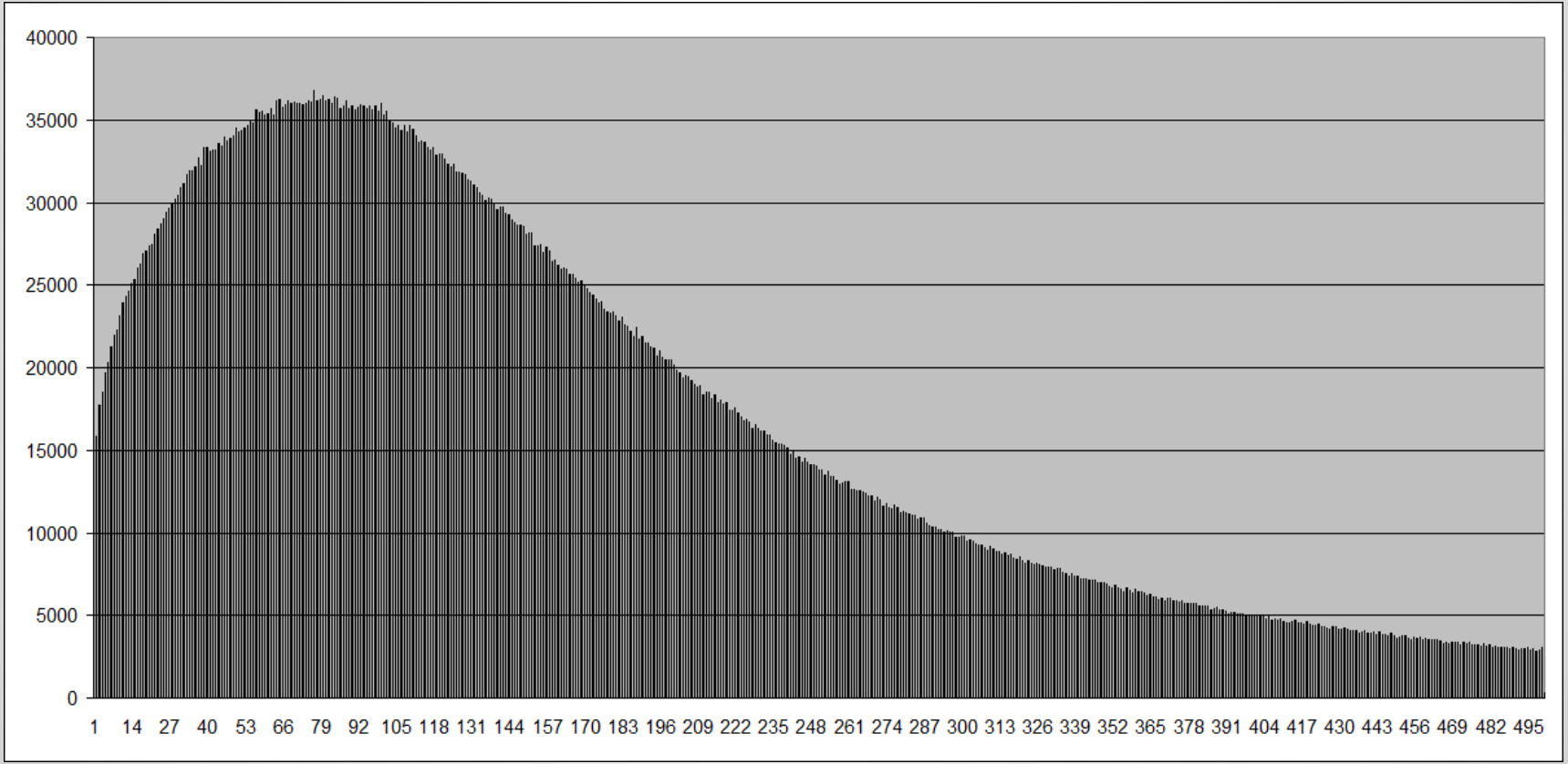
- Local feature: степень вершины
- Global feature: значение PageRank для каждой вершины
- Global-local feature: значение локального коэффициента кластеризации (LCC)
- Количество попыток аутентификации
- Анкетные данные (пол, возраст, регион проживания)
- Удаленность от столицы и от государственных границ

Данные для обучения

Для 10000 пользователей были вычислены следующие признаки:

- age, is_male, is_female
- degree, lcc, page_rank, geo_lcc
- good_auth_per_week, bad_auth_per_week
- dist_from_Moscow, dist_from_borders

Распределение числа друзей



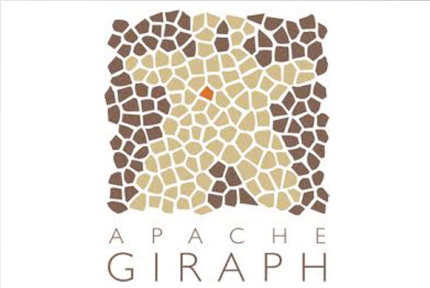
Вычисление признаков

4 сервера, каждый из которых нес 8 ядер и 30 Гб RAM, запущенные в Google Compute Engine.

Hadoop Cluster + Pig для вычисления неграфовых фич и degree.

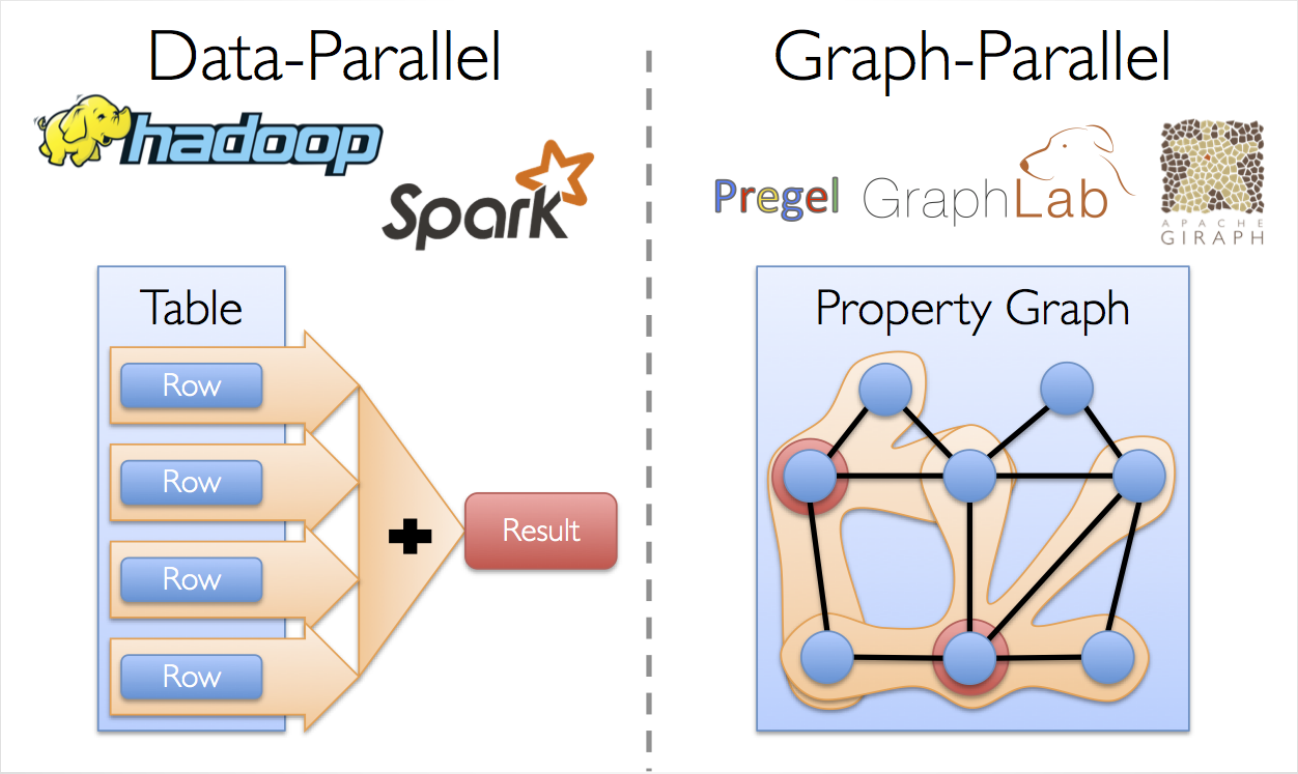
Giraph, запущенный на том же кластере, для вычисления PageRank и lcc.

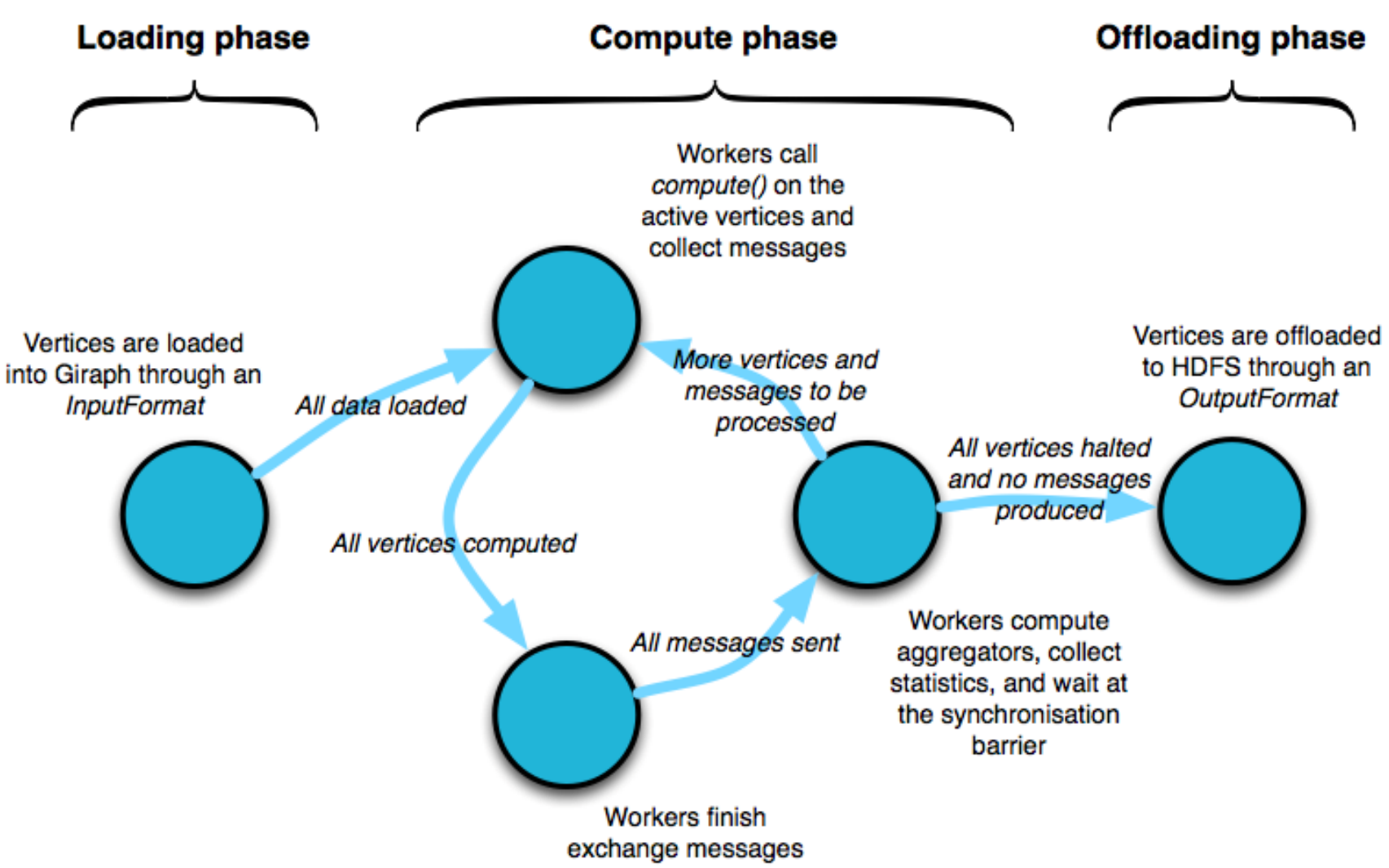
Почему Giraph?



- Open-source воплощение Pregel
- Запускалось на уже имеющейся инфраструктуре Hadoop
- Вычисления происходят в памяти (а ее было достаточно)
- Простая возможность для проведения итерационных вычислений (важно при вычислении

Думай вершинами, а не строками...




Loading phase
Compute phase
Offloading phase
Workers call compute() on the active vertices and collect messages
Vertices are loaded into Giraph through an InputFormat
All data loaded
More vertices and messages to be processed
Vertices are offloaded to HDFS through an OutputFormat
All vertices halted and no messages produced
All vertices computed
All messages sent
Workers compute aggregators, collect statistics, and wait at the synchronisation barrier
Workers finish exchange messages

Время расчета

Итерационный расчет PageRank, написанный на Pig прошел за 25 итераций, 123 минуты (~ 5 минут на итерацию)

Аналогичная реализация для Giraph заняла 45 итераций и 25 минут (~ 35 секунд на итерацию) при условии запуска 1 worker per 1 core

Построение модели

Для построения модели использовались kNN, полиномиальная регрессия и решающие деревья (Random Forest, C4.5).

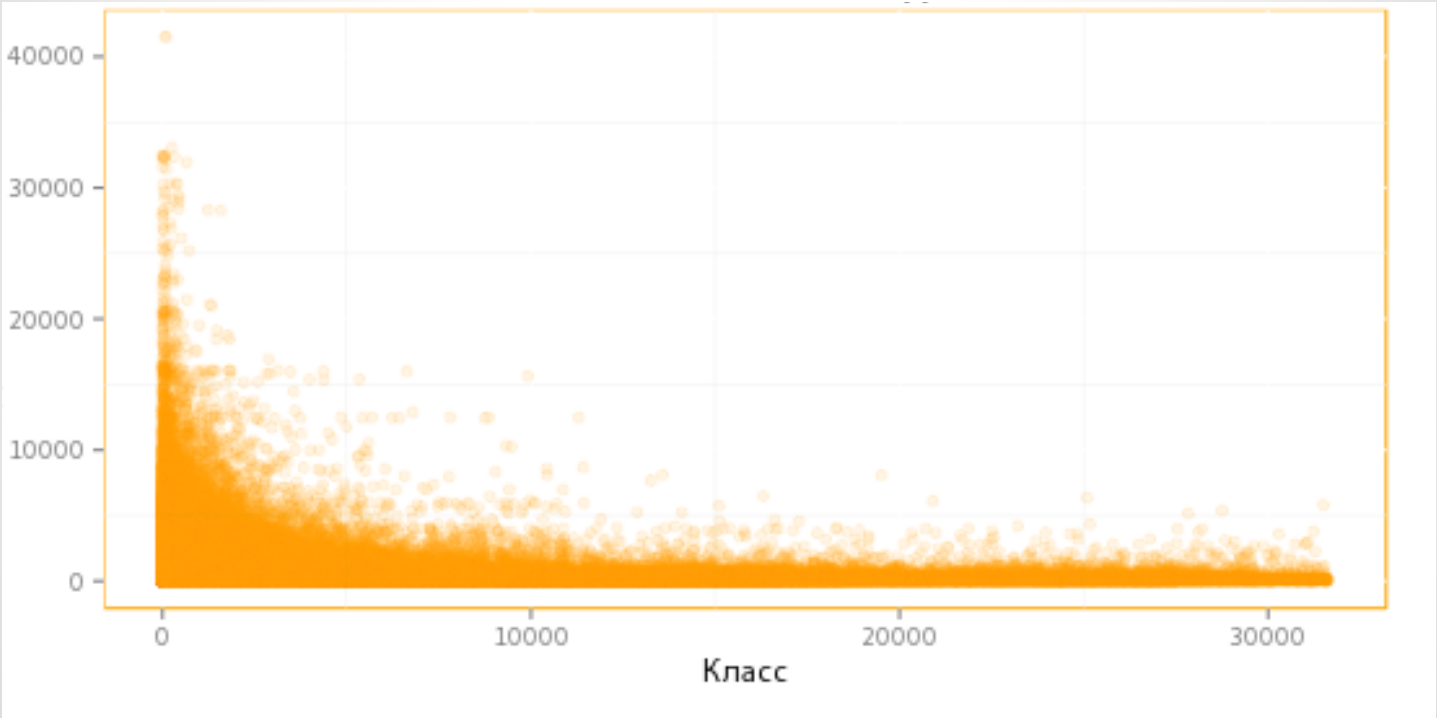
Самими лучшими оказались kNN (n = 7) и C4.5 с ошибкой первого рода в 5% и 3% соотв., ошибкой второго рода 12% и 19% соответственно.

Важность факторов

Самыми важными факторами оказались geo_lcc и degree, за ними с большим отрывом следуют lcc, dist_from_Moscow, good_auth_per_week и page_rank.

А вот те данные, которые указываются в анкетах играли большую роль в модели, построенной kNN, и незначительную в построении решающих деревьев.

Заключение

- Расчет признаков, основанных на связях пользователей друг с другом сложен и требует использования Giraph
- Признаки, полученные на основе анализа структуры связей в социальной сети являются значимыми в решении задачи поиска спам - аккаунтов.
- Hadoop + Pig + Giraph в Google Compute Engine - легко масштабируемое решение, которое подходит для решения различных задач в SNA

Haec habui, quae dixi