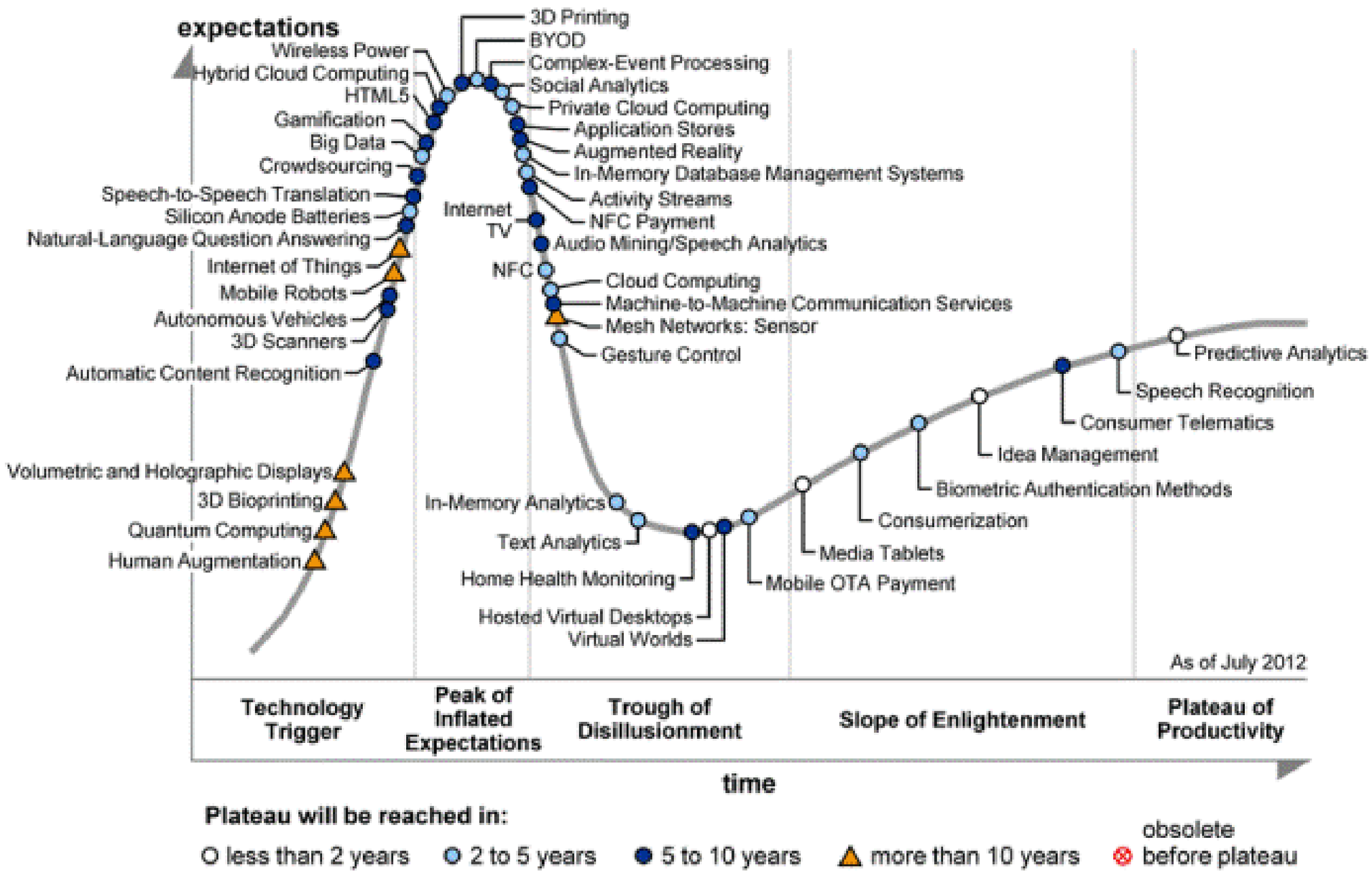
ISP RAS
Institute for System Programming
of the Russian Academy of Sciences

Современные проблемы анализа данных пользователей социальных сетей

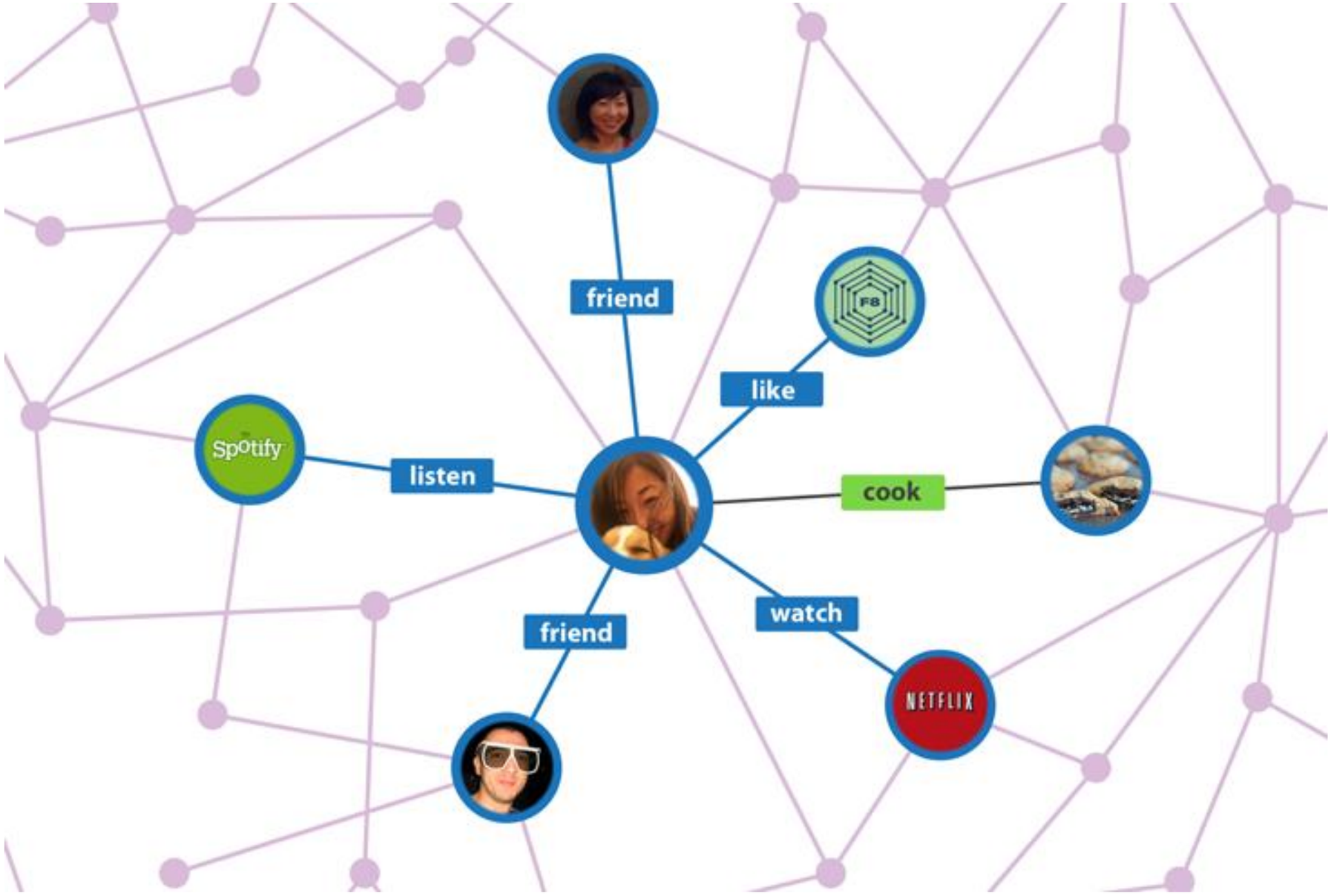
Коршунов Антон Викторович

korshunov@ispras.ru

Gartner Hype Cycle for Emerging Technologies 2012



Социальная сеть



Пользователи, объекты, связи

Социальная сеть

- С позиций data mining – разреженный граф огромной размерности. Пример - facebook
  - ✓ **1 млрд** пользователей, **100 млрд** связей
  - ✓ **2,7 млрд** новых оценок и комментариев ежедневно
  - ✓ **250 млн** новых фотографий ежедневно
- Глобально состоит из иерархии кластеров, отражающих географическую близость пользователей и похожие интересы
- Пользователь в среднем имеет аккаунты в **3 разных сетях**
- Рост числа аккаунтов с **2011** года составил **28%** (с **2,4** до **3,1 млрд**)
- Прогноз на **2015** год: **3,9 млрд** аккаунтов и **1,2 млрд** пользователей

Сети общего назначения

- Поддержание оффлайновых связей
- Новые знакомства
- Обсуждение повседневных событий
- Развлечения

facebook

ВКонтакте

одноклассники

Нишевые сети

- Поддержание профессиональных контактов
- Новые контакты с единомышленниками
- Обсуждение общих интересов

LinkedIn

Pinterest

comon
Соцсеть трейдеров

Контентные сети

- Обмен контентом
- Распространение новостей
- Образование и развитие сообществ по интересам
- Развлечения

twitter

YouTube™

last·fm

SOUNDCLOUD

Другие

- Геосоциальные
- Закладки, каталоги
- Игровые

Популярность

NORTH AMERICA	
Canada	94%
United States	98%

LATIN AMERICA	
Argentina	96%
Brazil	97%
Chile	94%
Colombia	96%
Mexico	96%
Peru	96%
Puerto Rico	90%
Venezuela	96%

EUROPE	
Austria	86%
Belgium	93%
Denmark	94%
Finland	91%
France	91%
Germany	90%
Ireland	95%
Italy	93%
Netherlands	94%
Norway	89%
Poland	95%
Portugal	96%
Spain	98%
Russia	88%
Sweden	93%
Switzerland	90%
Turkey	96%
United Kingdom	98%

MIDDLE EAST & AFRICA	
Israel	94%
South Africa	88%

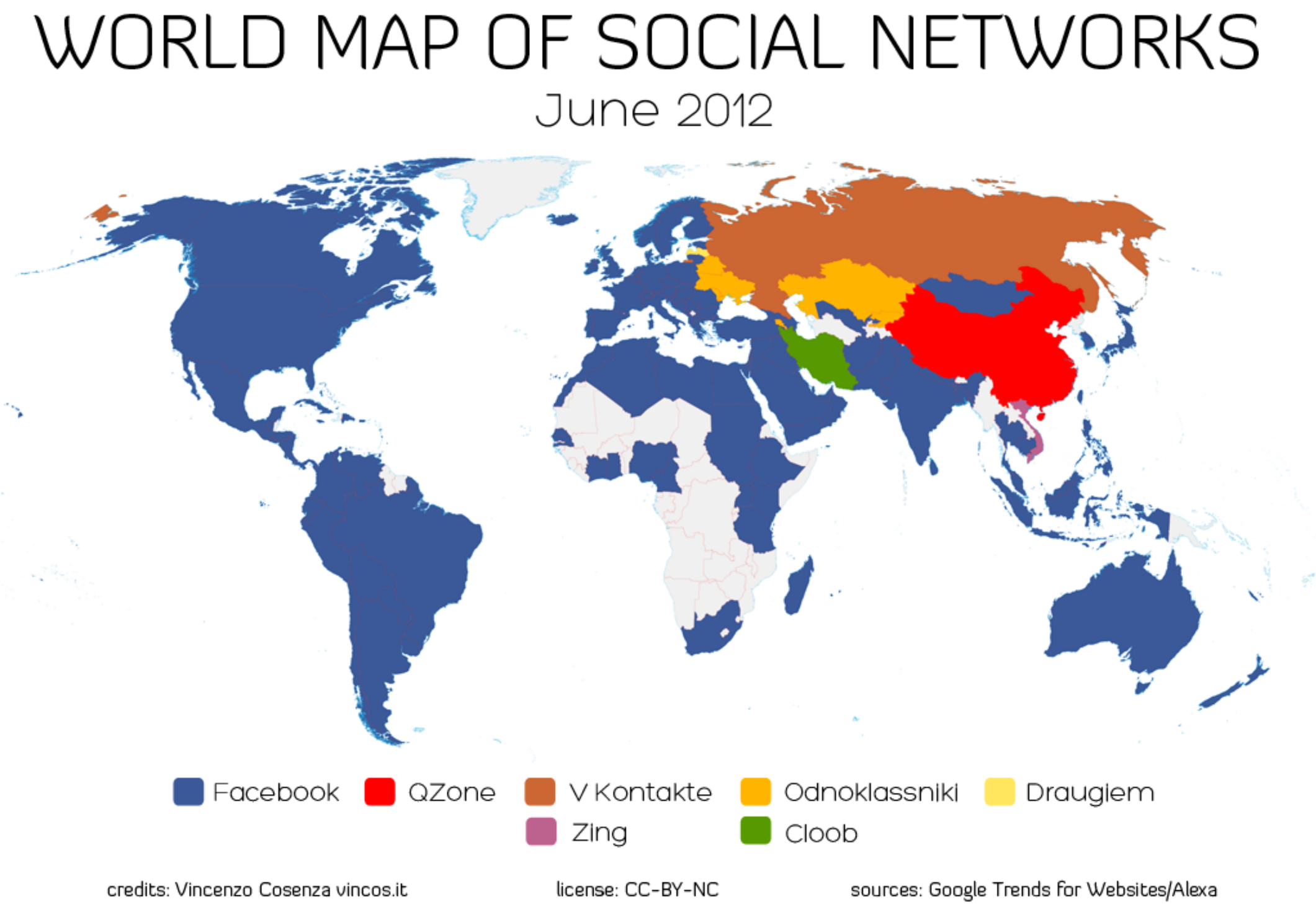
ASIA PACIFIC	
Australia	96%
China	53%
Hong Kong	93%
India	95%
Indonesia	94%
Japan	58%
Malaysia	94%
New Zealand	95%
Philippines	96%
Singapore	94%
South Korea	87%
Taiwan	94%
Vietnam	85%

Percentage of Online Population Using Social Networking around the World

Source: comScore Media Metrix, October 2011

** Data is based on the 43 countries on which comScore reports individually.*

Популярность



Facebook

Ежедневные изменения и появление новых данных

Актуальность

- Уникальный источник данных о личной жизни и интересах реальных людей: переписка, дневники, фотоальбомы
- Проникновение во все сферы деятельности: нишевые сети IT-специалистов, трейдеров, бухгалтеров
- Социализация контента: фото, видео, музыка, новости, рецепты
- Социализация сервисов: магазины, форумы, рекомендации
- Растущий спрос на социальные сервисы, упрощающие общение и обмен информацией
- Беспрецедентные возможности для таргетированного маркетинга

Анализ социальных графов

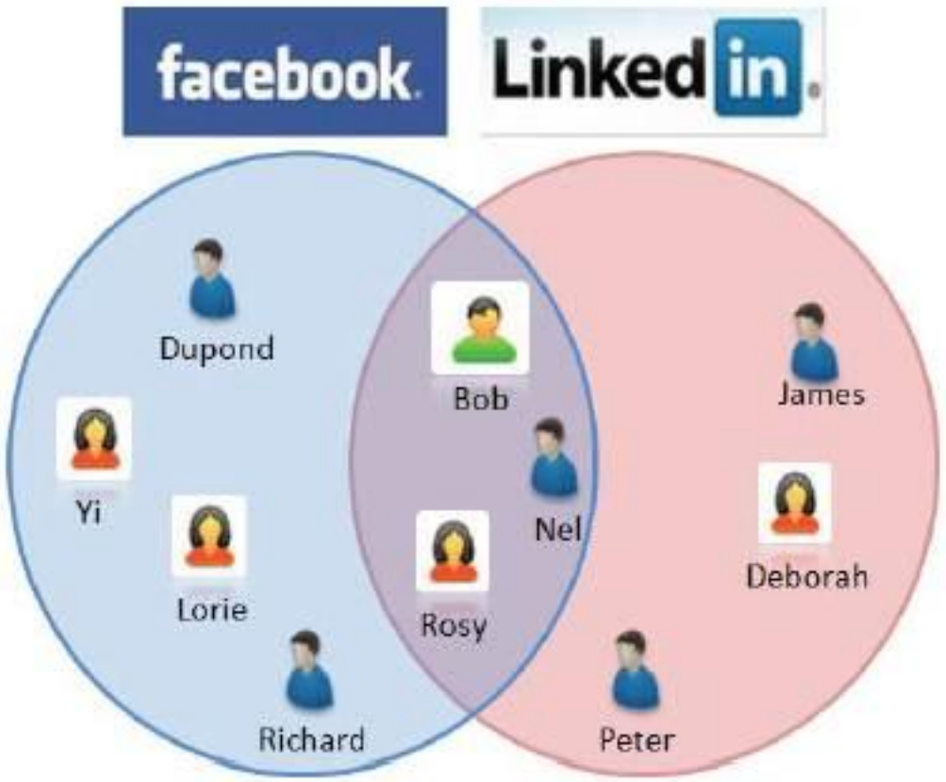
Объединение множества дисциплин

- ✓ Теория графов
- ✓ Анализ текстов на естественном языке
- ✓ Информационный поиск
- ✓ Машинное обучение и искусственный интеллект
- ✓ Социология и антропология
- ✓ Распределённые алгоритмы и системы

Проблемы

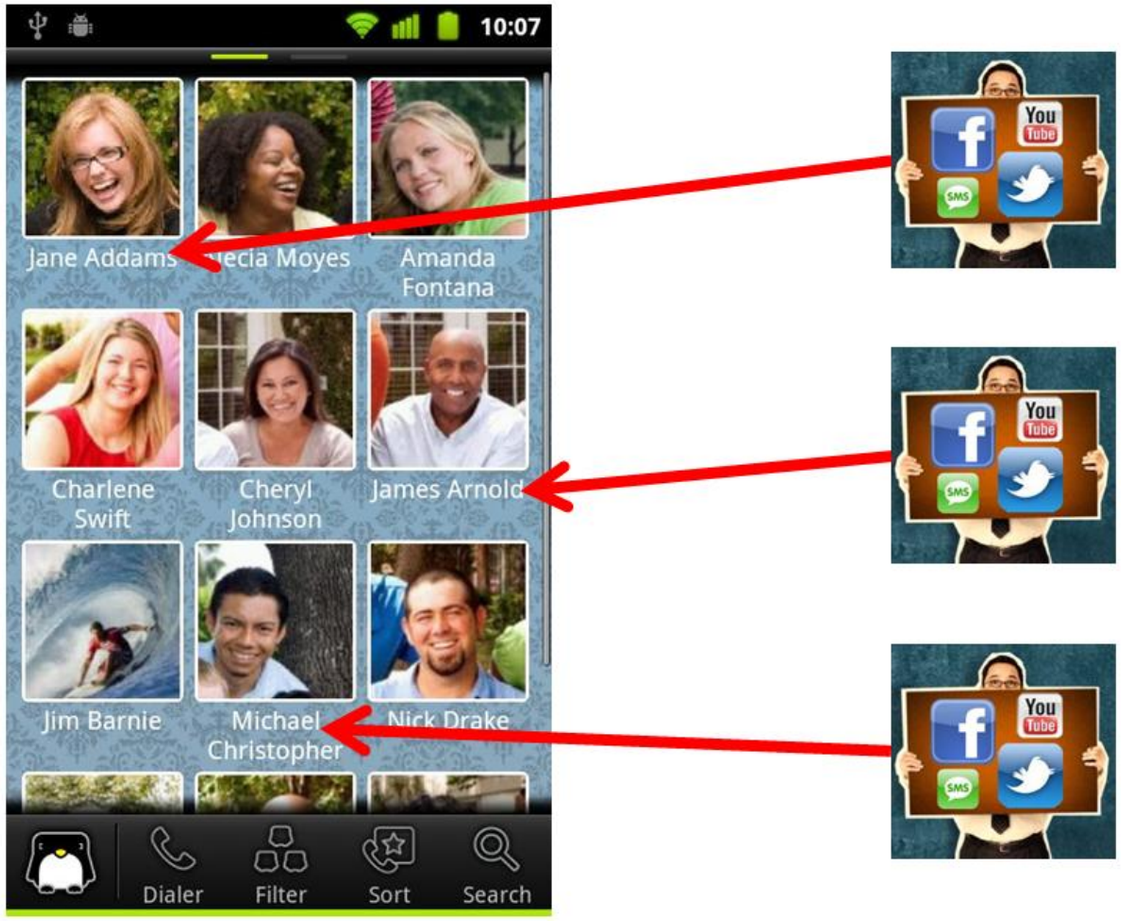
- ➢ Новые бизнес-задачи, требующие срочного решения
- ➢ Трудности работы с персональными данными
- ➢ Поиск новых областей применения известных алгоритмов
- ➢ Оптимизация распределённого хранения и обработки

Интеграция профилей пользователя



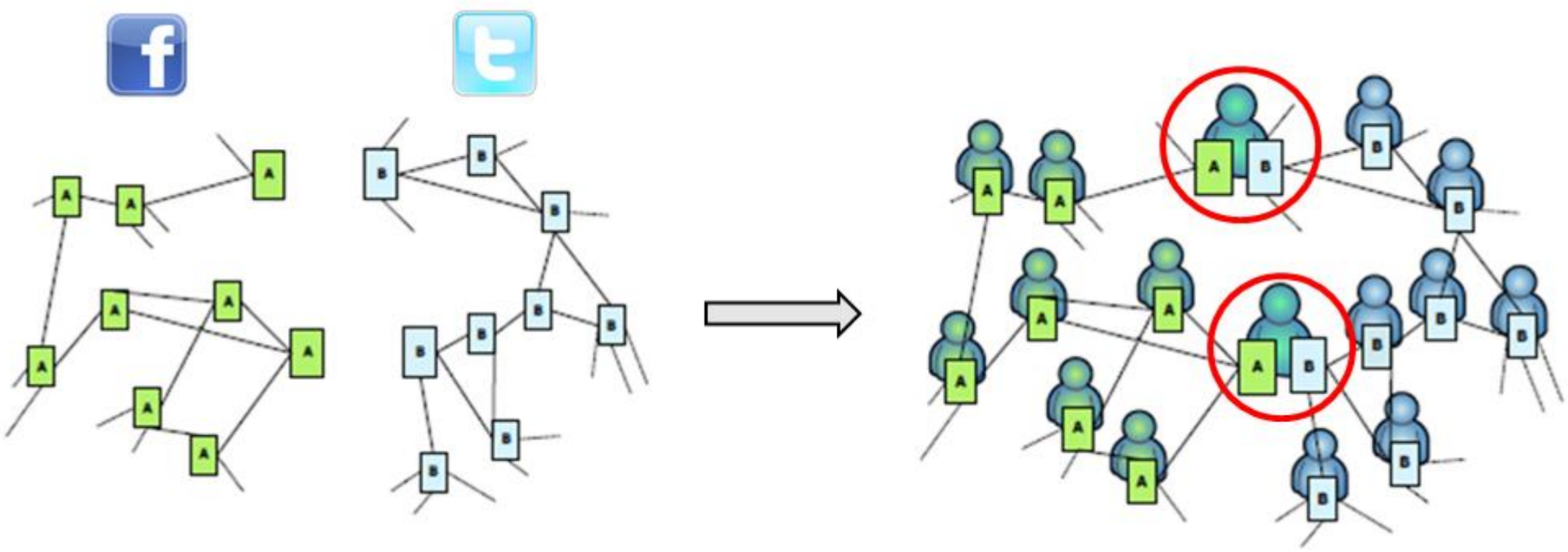
Как автоматически найти знакомых из одной сети в другой сети?

Интеграция профилей пользователя



Как автоматически объединить списки контактов?

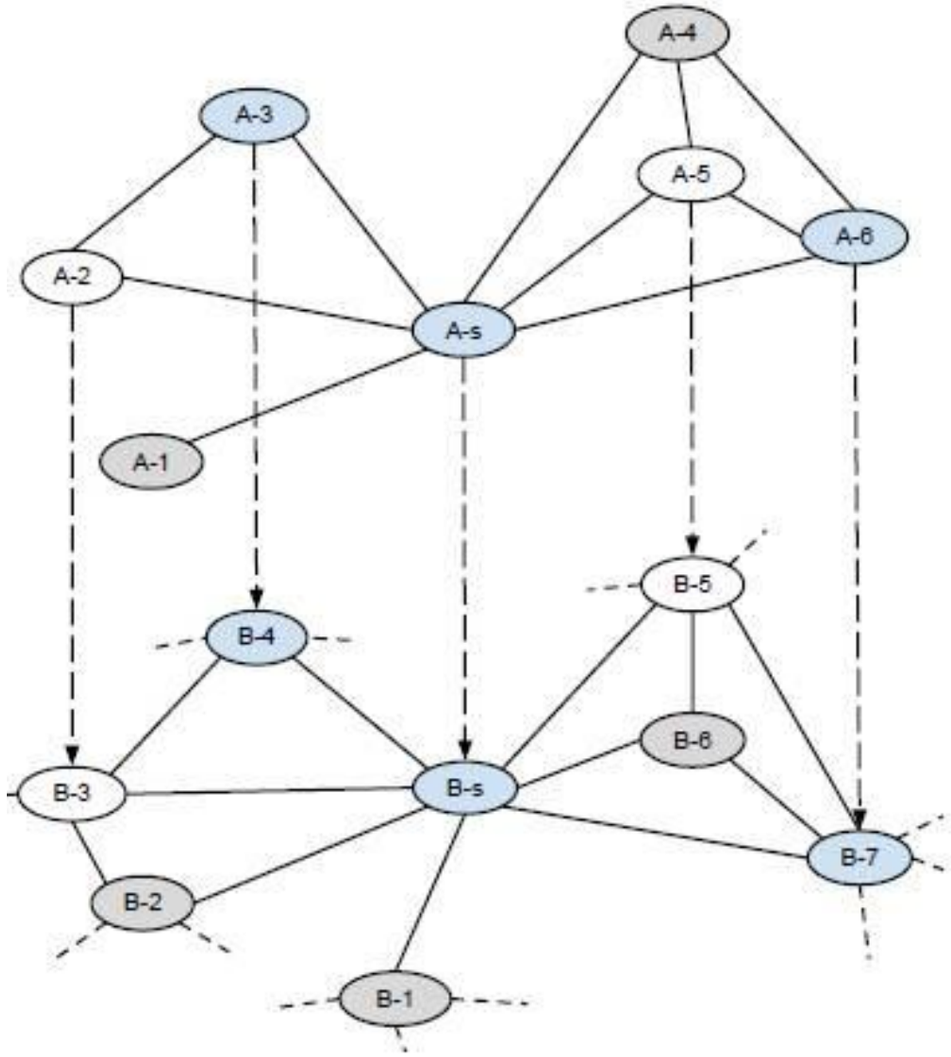
Интеграция профилей пользователя



Интеграция профилей пользователя

Решение

1. Построить модель условных случайных полей
2. Найти пользователей с точным совпадением по именам, добавить их в модель
3. Рассчитать энергии рёбер с помощью метрик строковой и графовой близости
4. Найти оптимальную конфигурацию совпадающих профилей путём оптимизации общей энергии модели
5. Отфильтровать результаты



Интеграция профилей пользователя

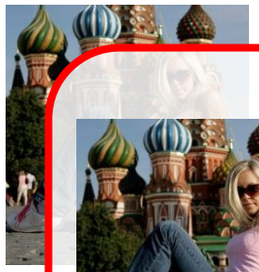
algorithm	R	P	F_1
agnostic to direction of projection			
Baseline 1 (weighted sum)	0.45	0.94	0.61
Baseline 2 (probability distance)	0.51	**1.0**	0.69
JLA, intersection, anonymized	0.6	**1.0**	0.76
JLA, intersection	0.66	0.99	0.79
Twitter $\rightarrow$ Facebook			
JLA, anonymized ($\Phi \equiv 0$)	0.62	**1.0**	0.77
JLA	0.79	**1.0**	0.89
Facebook $\rightarrow$ Twitter			
JLA, anonymized ($\Phi \equiv 0$)	0.61	**1.0**	0.76
JLA	**0.8**	**1.0**	**0.89**

Данные: 16 пар персональных сетей из Facebook и Twitter, размеченных их владельцами

Определение скрытых атрибутов

Name: **Julia Stevens**
Age: **[empty]**
Gender: **female**
Relationship: **[empty]**
Location: **France**

Name: **Rob Fee**
Age: **666**
Gender: **female**
Relationship: **single**
Location: **U.S.**



Name: **Maria Zotova**
Age: **24**
Gender: **female**
Relationship: **married**
Location: **Moscow**

Name: **Maria Zotova**
Age: **24**
Gender: **female**
Relationship: **married**
Location: **Moscow**

атрибуты не указаны

ошибки

дубликаты профилей

- Как определить пол, возраст, религию, статус отношений пользователя по текстам его сообщений?
- Кого больше среди фанатов игры/фильма/книги – мужчин или женщин, молодых или пожилых? Как использовать эти знания для рекомендации?

Определение скрытых атрибутов

Harry Bathurst
@harrybathurst
18. To ammend the words of Toy Story, you have not got a friend in! #bazinga
Brighton, UK · http://facebook.com/harry.bathurst

Harry Bathurst @harrybathurst 11 Jul
@Jedturner_1996 you see the video? youtube.com/watch?v=8sGBld...
View media

Harry Bathurst @harrybathurst 11 Jul
@Jedturner_1996 what?
View conversation

Harry Bathurst @harrybathurst 11 Jul
Scorcher merked Dappyyyy haha!
Expand

Harry Bathurst @harrybathurst 11 Jul
You know I'm hot propertyyy
Expand

Harry Bathurst @harrybathurst 11 Jul
@peterquiffs funnyyyyy
Expand

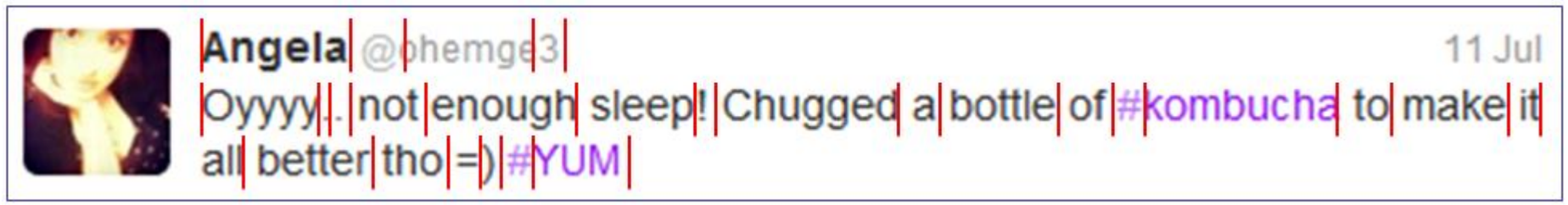
Harry Bathurst @harrybathurst 11 Jul
Fed up of going to the gym on my own! #itsboring
Expand

Basic Info

Birthday	February 1, 1994
Sex	Male
Interested In	Women
Relationship Status	In a relationship with Emily Morgan
Anniversary	April 29, 2012
Languages	English
Religious Views	Fundamental Baptist
Political Views	Care Bears Dont care

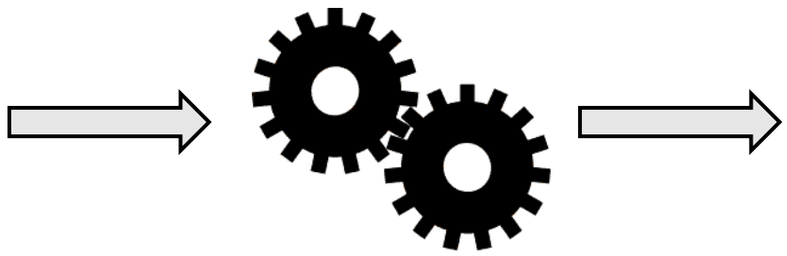
Label: Gender	Label: Age	Label: Relationship
MALE	18	NONSINGLE

Определение скрытых атрибутов



Решение

1. Извлечение бинарных признаков из твитов и профиля
2. Выбор наиболее информативных признаков для каждого класса
3. Обучение модели (алгоритм Passive-Aggressive)
4. Классификация



Gender	Age	Relationship
MALE	<20	SINGLE

Определение скрытых атрибутов

Younger (≤20 years old)	Older (>20 years old)
(('char', 6, u'revisi'), 0.15)	(('char_screen_name', 1, u'8'), 0.13)
(('token_uncased', 1, (u'exam',)), 0.14)	(('token', 2, (u'at', u'work')), 0.13)
(('token', 2, (u':', u'D')), 0.13)	(('token_uncased', 1, (u'bar',)), 0.10)
(('char_uncased', 7, u'bieber '), 0.13)	(('char', 4, u'le :'), 0.10)
(('token_uncased', 1, (u'prom',)), 0.13)	(('char', 4, u'ern '), 0.10)
(('token', 1, (u'tomorrow',)), 0.13)	(('char', 2, u':('), 0.10)
(('token_uncased', 1, (u'virgin',)), 0.12)	(('char_uncased', 7, u'hool to'), 0.09)
(('char', 3, u's :'), 0.12)	(('char', 7, u'season '), 0.09)
(('char', 4, u'nedi'), 0.12)	(('token', 1, (u'wine',)), 0.09)
(('char', 3, u' :L'), 0.12)	(('char_uncased', 7, u'hit me '), 0.09)
(('token', 2, (u'≤', u'3')), 0.12)	(('char', 3, u'g :'), 0.09)
(('char', 4, u' Yo'), 0.11)	(('char', 5, u'House'), 0.09)
(('char', 2, u'(:'), 0.11)	(('char', 6, u'ington'), 0.09)
(('char', 5, u' me? '), 0.10)	(('token', 2, (u'Nice', u'!')), 0.08)
(('char', 5, u' math'), 0.10)	(('char_uncased', 4, u' jew'), 0.08)
(('char', 3, u'KON'), 0.10)	(('char', 5, u'ompan'), 0.08)
(('char', 7, u' much i'), 0.10)	(('char', 3, u' Mo'), 0.08)
(('char_uncased', 7, u'favouri'), 0.09)	(('char_uncased', 4, u'o <3'), 0.08)
(('char', 4, u' Swa'), 0.09)	(('char', 7, u'real wo'), 0.08)
(('char_uncased', 3, u't x'), 0.09)	(('char_uncased', 4, u'. y'), 0.08)

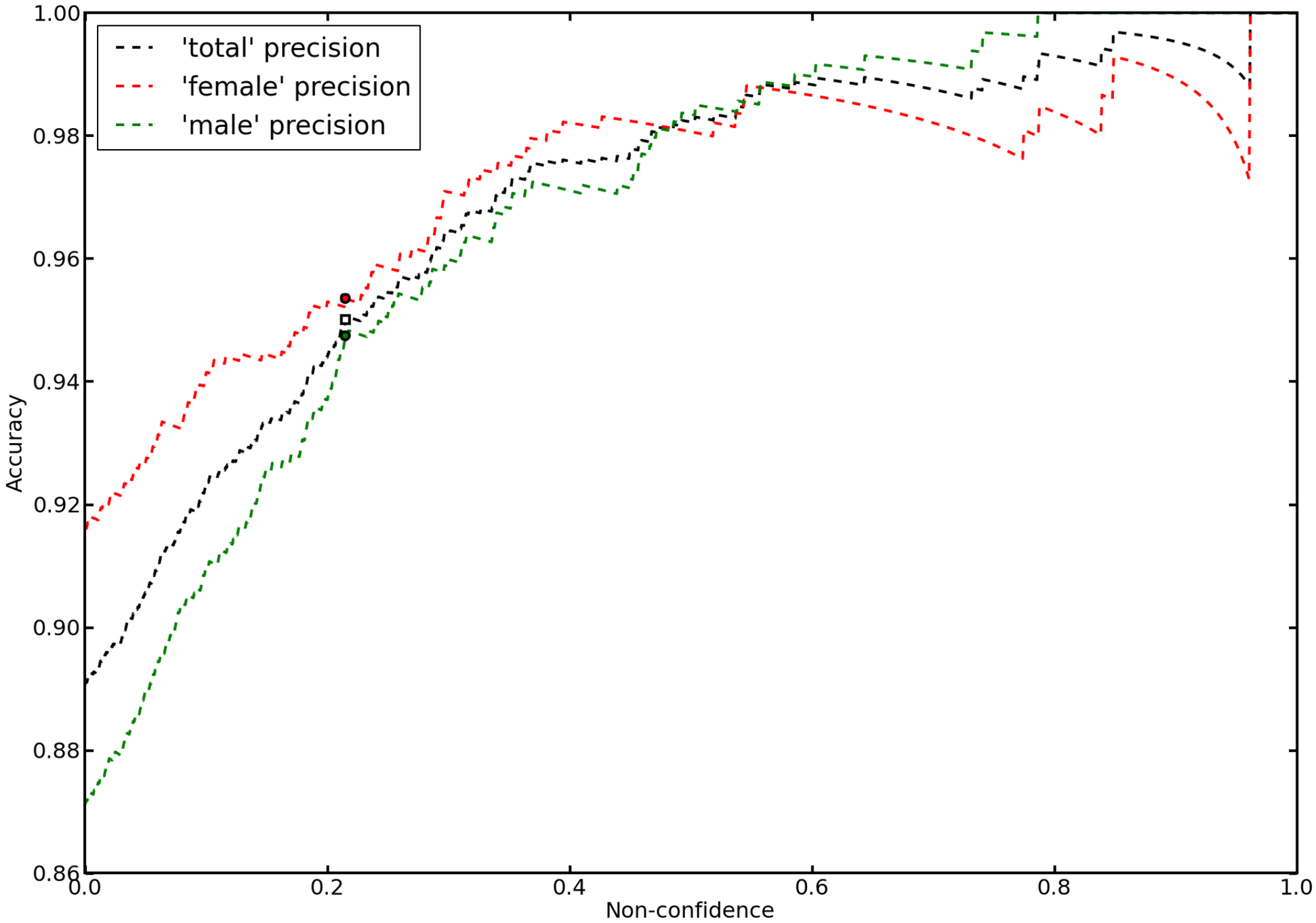
Определение скрытых атрибутов

Male	Female
(('token_name', 1, (u'alex',)), 0.17)	(('char_name', 2, u'a_'), 0.561)
(('char_name_uncased', 3, u'an_'), 0.16)	(('char_name_uncased', 2, u'i_'), 0.22)
(('char_name_uncased', 2, u'k_'), 0.16)	(('char_name_uncased', 2, u'e_'), 0.21)
(('char_name', 1, u'_'), 0.15)	(('char_name', 3, u'ie_'), 0.18)
(('char_name', 2, u'd_'), 0.14)	(('char_name_uncased', 2, u'y_'), 0.17)
(('char_name', 2, u't_'), 0.14)	(('char_name', 1, u'Y'), 0.16)
(('char_name', 1, u'k'), 0.14)	(('char_name_uncased', 3, u'rac'), 0.16)
(('char_name', 2, u'ev'), 0.14)	(('char_name_uncased', 3, u'jen'), 0.14)
(('char_name_uncased', 3, u'ust'), 0.13)	(('char_name_uncased', 3, u'ash'), 0.13)
(('char_name', 4, u'icha'), 0.13)	(('char_name', 2, u'hl'), 0.13)
(('char_name', 1, u'J'), 0.13)	(('char_name_uncased', 4, u'laur'), 0.12)
(('char_name', 3, u'ew_'), 0.12)	(('char_name_uncased', 2, u'ka'), 0.12)
(('char_name', 2, u'm_'), 0.12)	(('char_name_uncased', 3, u'ara'), 0.12)
(('char_name_uncased', 3, u'is_'), 0.12)	(('char_name', 2, u'nn'), 0.12)
(('char_name', 2, u'am'), 0.11)	(('char_name_uncased', 3, u'ine'), 0.12)
(('char_name_uncased', 2, u'ry'), 0.11)	(('char_name', 2, u'Su'), 0.11)
(('token_name', 1, (u'chris',)), 0.11)	(('char', 3, u'epy'), 0.11)
(('char_name', 3, u'ke_'), 0.11)	(('char_name', 3, u'gan'), 0.11)
(('char_name_uncased', 3, u'wil'), 0.11)	(('char_name', 2, u'Am'), 0.11)
(('char_name', 3, u'And'), 0.11)	(('char_name_uncased', 2, u'sa'), 0.11)

Определение скрытых атрибутов

Single	Non-single
(('char', 7, u'_Avenge'), 0.21) (('char', 4, u'bums'), 0.19) (('char', 5, u'd_rel'), 0.19) (('char', 5, u'ys_lo'), 0.17) (('char_uncased', 4, u'neat'), 0.15) (('char_uncased', 3, u'd_z'), 0.15) (('char_uncased', 3, u'aja'), 0.15) (('char_uncased', 6, u'e,_can'), 0.15) (('token', 2, (u':', u'Thank')), 0.15) (('char', 7, u'golden_'), 0.14) (('char_uncased', 4, u'tip_'), 0.14) (('char', 3, u'RAG'), 0.14) (('char', 2, u'$5'), 0.14) (('char', 6, u'haha_w'), 0.14) (('char_uncased', 3, u'kur'), 0.14) (('char', 6, u'_locke'), 0.14) (('token_uncased', 3, (u'in', u'the', u'future')), 0.14) (('char_uncased', 7, u'dancers'), 0.14) (('token_uncased', 3, (u'heading', u'back', u'to')), 0.14) (('char', 5, u'ine..'), 0.14)	(('char', 2, u'zH'), 0.29) (('token', 2, (u'my', u'girlfriend')), 0.28) (('char_uncased', 3, u'bew'), 0.23) (('char', 6, u'y_boyf'), 0.22) (('token', 3, (u'!', u'!', u'X')), 0.22) (('char', 5, u'cool!'), 0.22) (('char_name', 2, u'el'), 0.20) (('char_uncased', 7, u'_burger'), 0.20) (('char_uncased', 4, u'poop'), 0.20) (('char_uncased', 6, u'done._'), 0.19) (('token', 2, (u'mate', u'ㄴ')), 0.19) (('char_uncased', 5, u'roche'), 0.19) (('char_uncased', 4, u'e?_t'), 0.18) (('char_uncased', 5, u'_baby'), 0.18) (('char_uncased', 2, u'8z'), 0.18) (('char_uncased', 7, u'_tell_s'), 0.18) (('char', 5, u'_#ilo'), 0.17) (('char', 6, u'e_you!'), 0.17) (('token', 1, (u'shift',)), 0.17) (('char_uncased', 5, u'neil_'), 0.17)

Определение скрытых атрибутов



Открытые проблемы

Где взять данные?

Как скачать *Facebook*?

- Через API невозможно, нужно качать и парсить HTML
- Граф из **10^9** пользователей, **~160** друзей в среднем
- 10^9 * 160 * 8 (int64) = **1192 Гб** - только на хранение связей
- 10^9 * 100 Кб (средний вес HTML-страницы) = **93 Тб** – трафик
- **Вывод**: качать весь граф долго, дорого и непрактично
- **Решение**: взять подграф, повторяющий свойства всей сети

Другие проблемы

- Эмуляция пользовательской сессии
- Блокировки по IP-адресу и аккаунту
- Работа в распределённом режиме

Где взять данные?

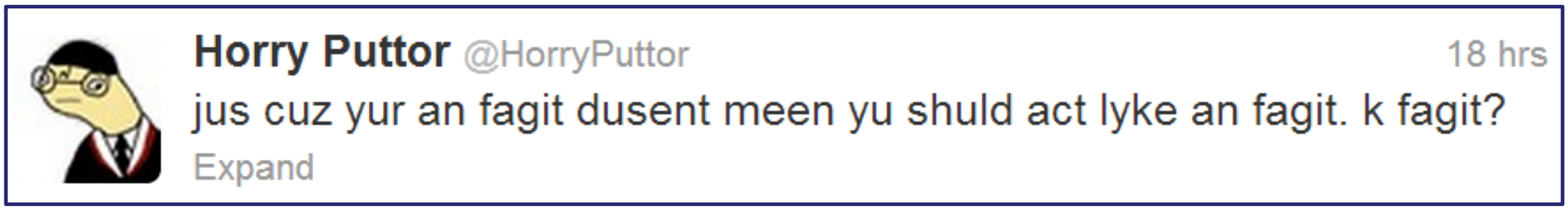
Как сгенерировать граф, похожий на *Facebook*?

- Изучить свойства реальной сети
- Разработать математическую модель случайного графа
- Написать распределённый алгоритм генерации по модели
- Сравнить свойства реального и сгенерированного графов

Открытые вопросы

- Как смоделировать направленный граф контентной сети?
- Как сгенерировать граф с известной структурой сообществ?
- Как сгенерировать профили пользователей с учётом свойства гомофилии?
- Как оценить сходство синтетического и реального графов?

Нормализация текста

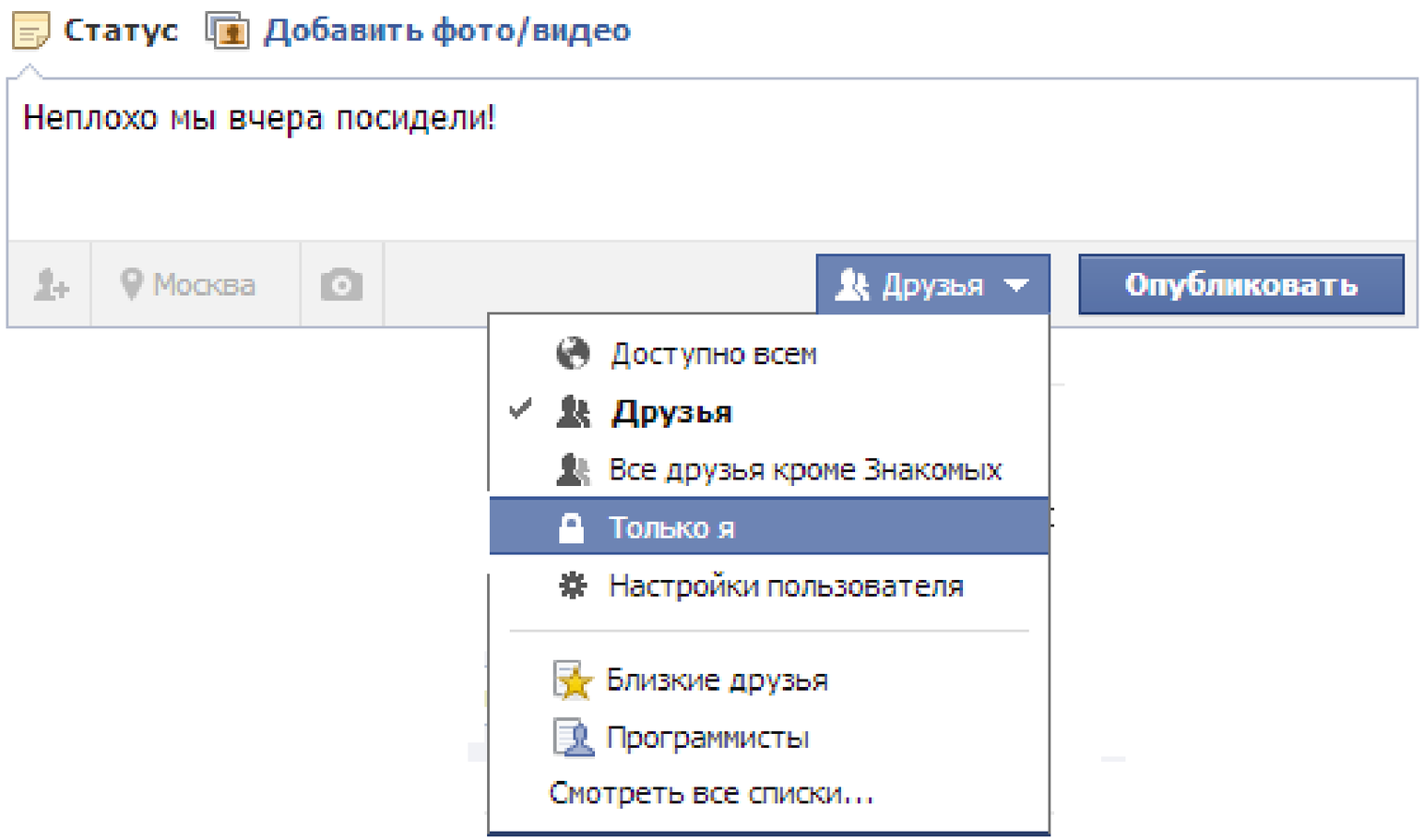


Проблемы:

- Сокращения
- Повторяющиеся буквы/слова
- Опечатки
- Слэнг и неологизмы

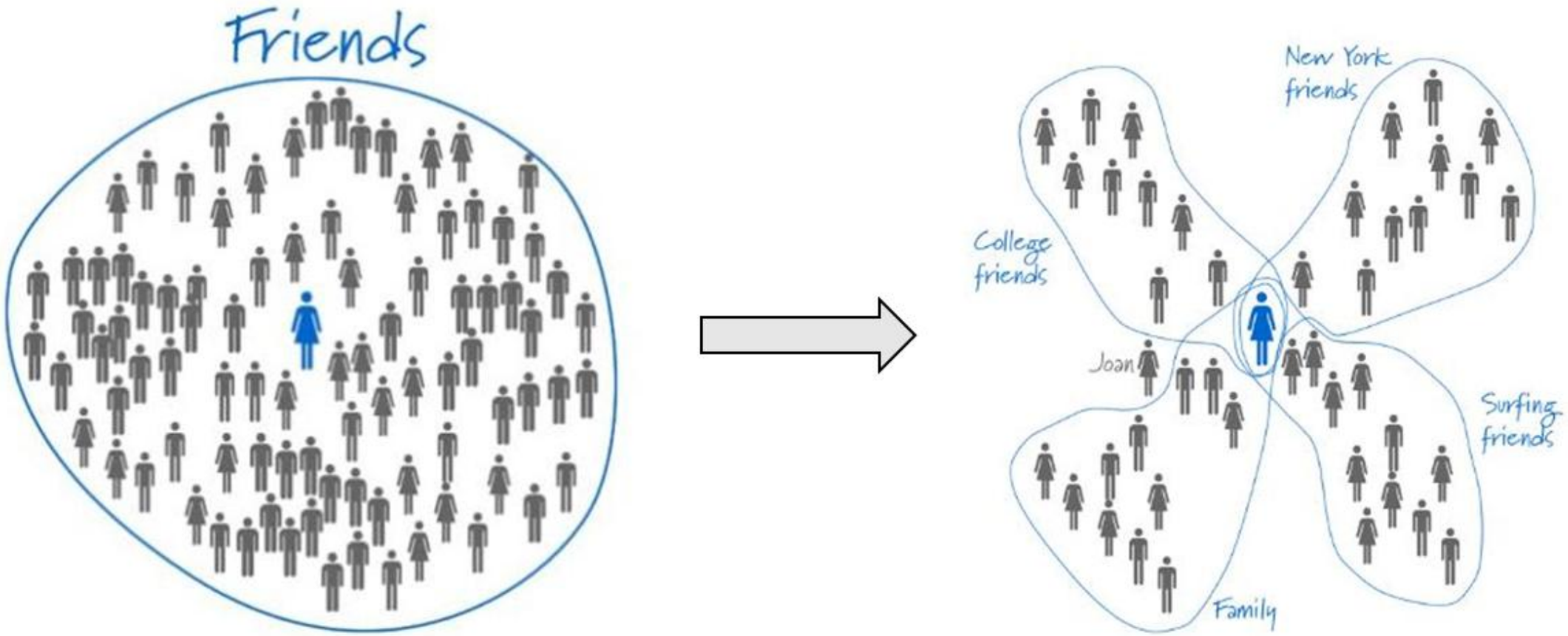
Как сделать возможной обработку текстов сообщений с помощью традиционных методов?

Поиск сообществ



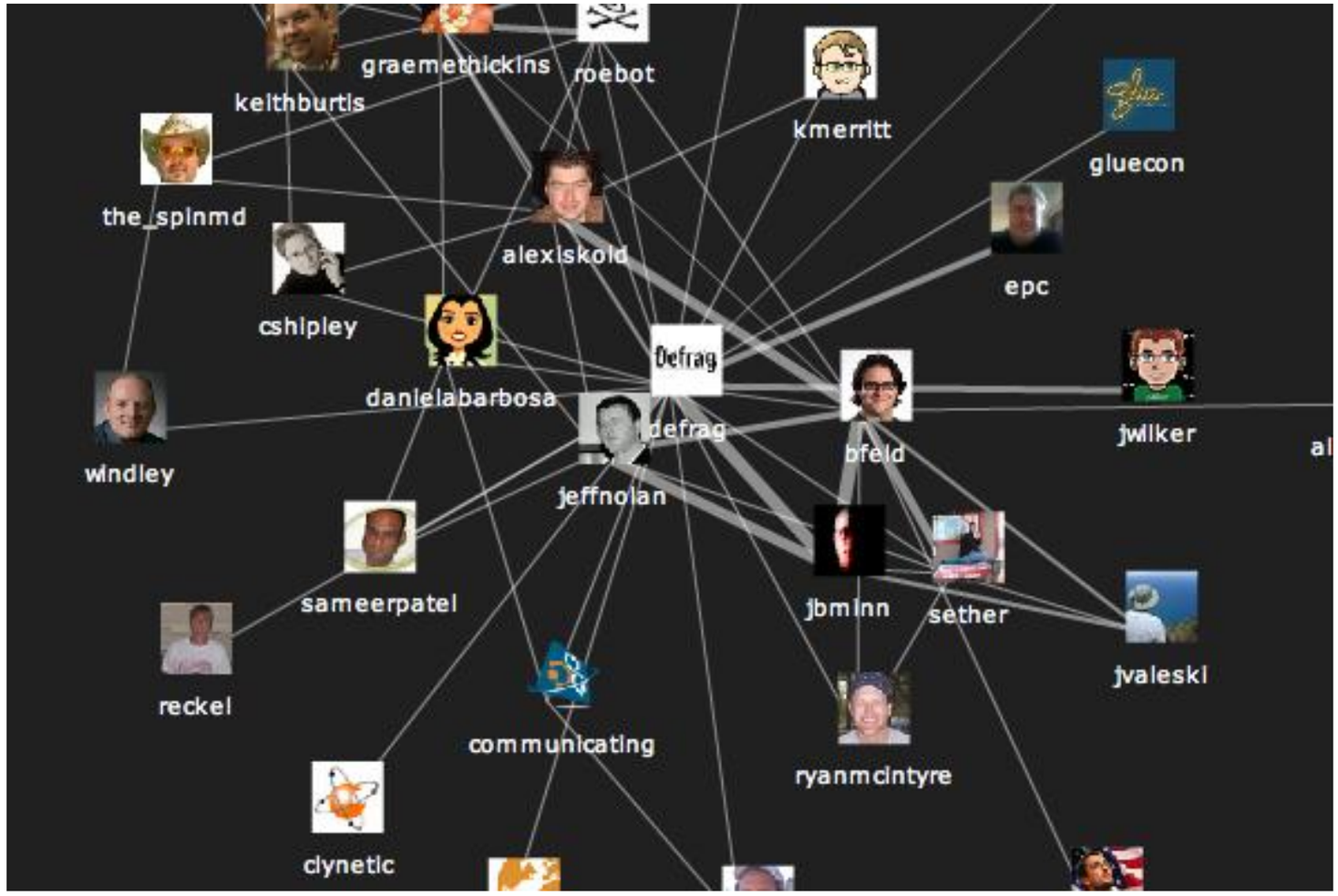
- Иметь возможность выбирать группу получателей - удобно
- Создавать списки друзей вручную - неудобно

Поиск сообществ



- Как автоматически распределить своих друзей по кругам общения? - локальная перспектива
- Как выявить и визуализировать модульную структуру социального графа? – глобальная перспектива

Поиск сообществ



- Кто находится в одном тематическом сообществе с выбранным пользователем?
- Какие существуют тематические сообщества по данному запросу?
- Как использовать эти знания для рекомендации друзей и контента?

Поиск сообществ

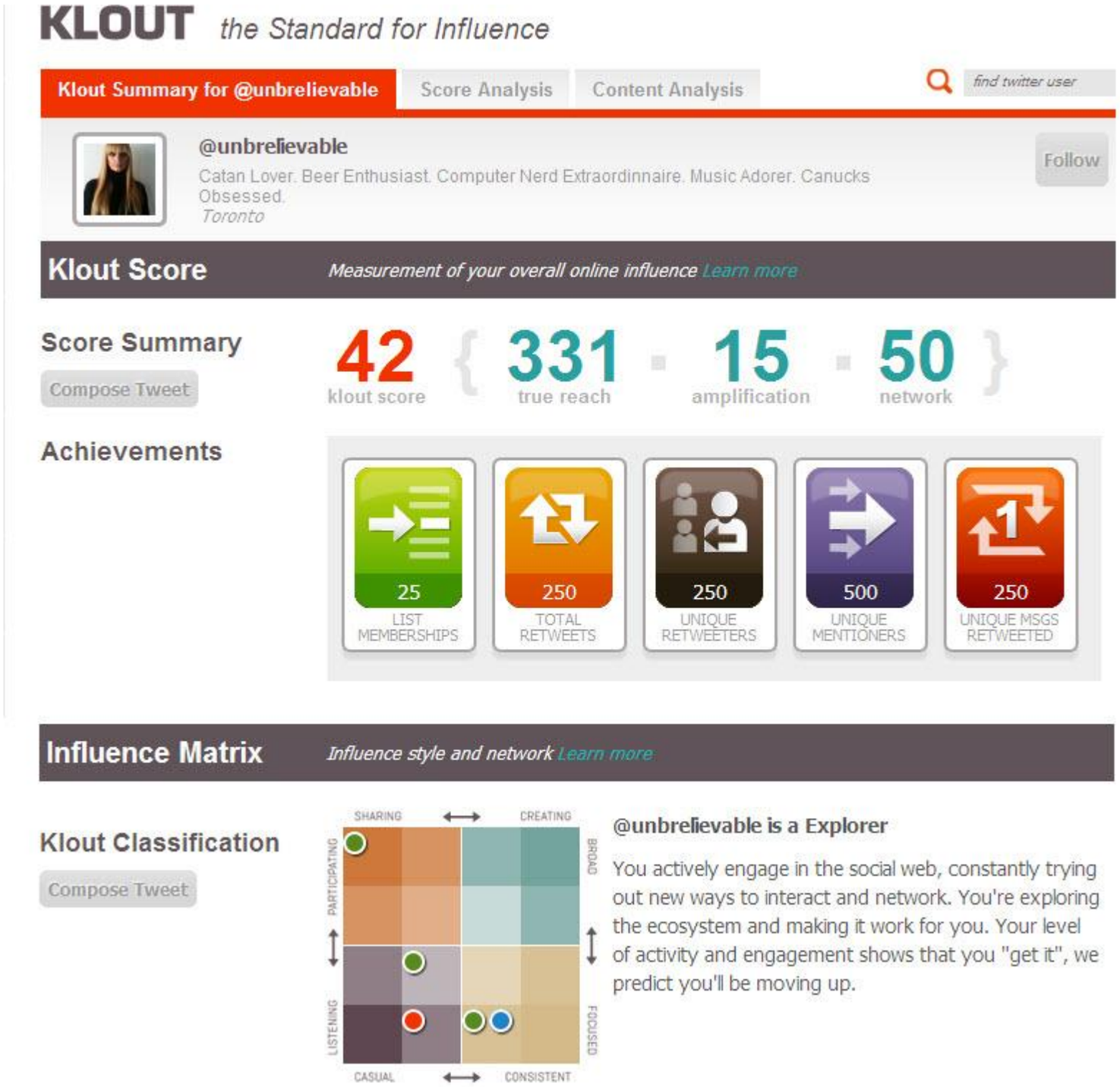
Открытые вопросы:

- Как производить эффективный и точный поиск тематических сообществ?
  - Вероятностные тематические модели
- Как тестировать гипотезы о способах возникновения и развития сообществ?
  - Генеративные модели + байесовский вывод
- Как искать различные типы сообществ, используя только графовую связность?
  - Оптимизация различных графовых статистик, характерных для конкретного типа сообщества
- Как автоматически выявлять иерархию сообществ?
- Как улучшить качество результатов для сильно пересекающихся сообществ?
- Как искать сообщества на графах с миллиардами вершин?

Оценка информационного влияния

- Как понять, чья активность в сети влияет на действия и предпочтения конкретного пользователя?
- Как использовать эти знания для рекомендации друзей и контента?
- Как использовать авторитетных пользователей в маркетинговых целях?

Оценка информационного влияния



Оценка информационного влияния

Открытые вопросы:

- Как построить модель, учитывающую все способы распространения влияния?
  - Графические вероятностные модели
- Как оценить качество результатов?
  - Предсказание репостов интересных материалов
- Как оценить влияние на графах с миллиардами вершин?

Обработка «больших данных»

Большие данные

- ✓ **10^9** пользователей
- ✓ Связи в виде графа, средняя степень ~**160**
- ✓ Профили, текстовые сообщения, мультимедийный контент
- ✓ Объектная графовая модель: «*всё связано со всем*»
- ✓ Данные постоянно обновляются, поступают в потоке

Большие проблемы

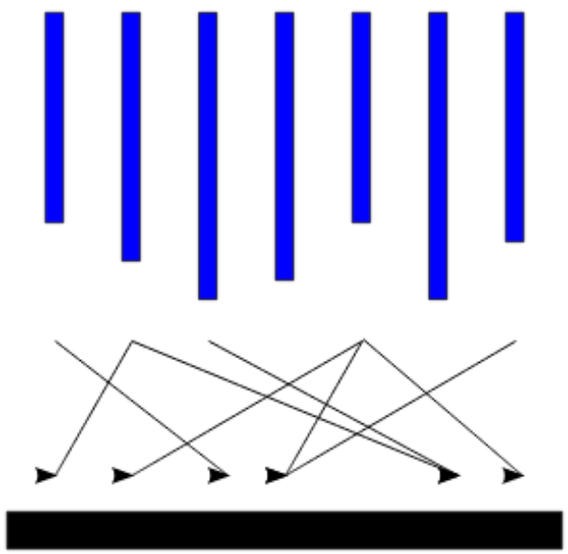
- ➢ Разработка параллельных версий алгоритмов для обработки данных в пакетном и потоковом режимах
- ➢ Поиск компромиссов между качеством результатов, производительностью и масштабируемостью
- ➢ Осваивание и развитие стремительно растущего стека технологий

Модель Bulk Synchronous Parallel

Вход Список вершин с атрибутами + список рёбер с атрибутами

Алгоритм Последовательность супершагов (итераций)

Выход Изменённый граф или рассчитанные статистики



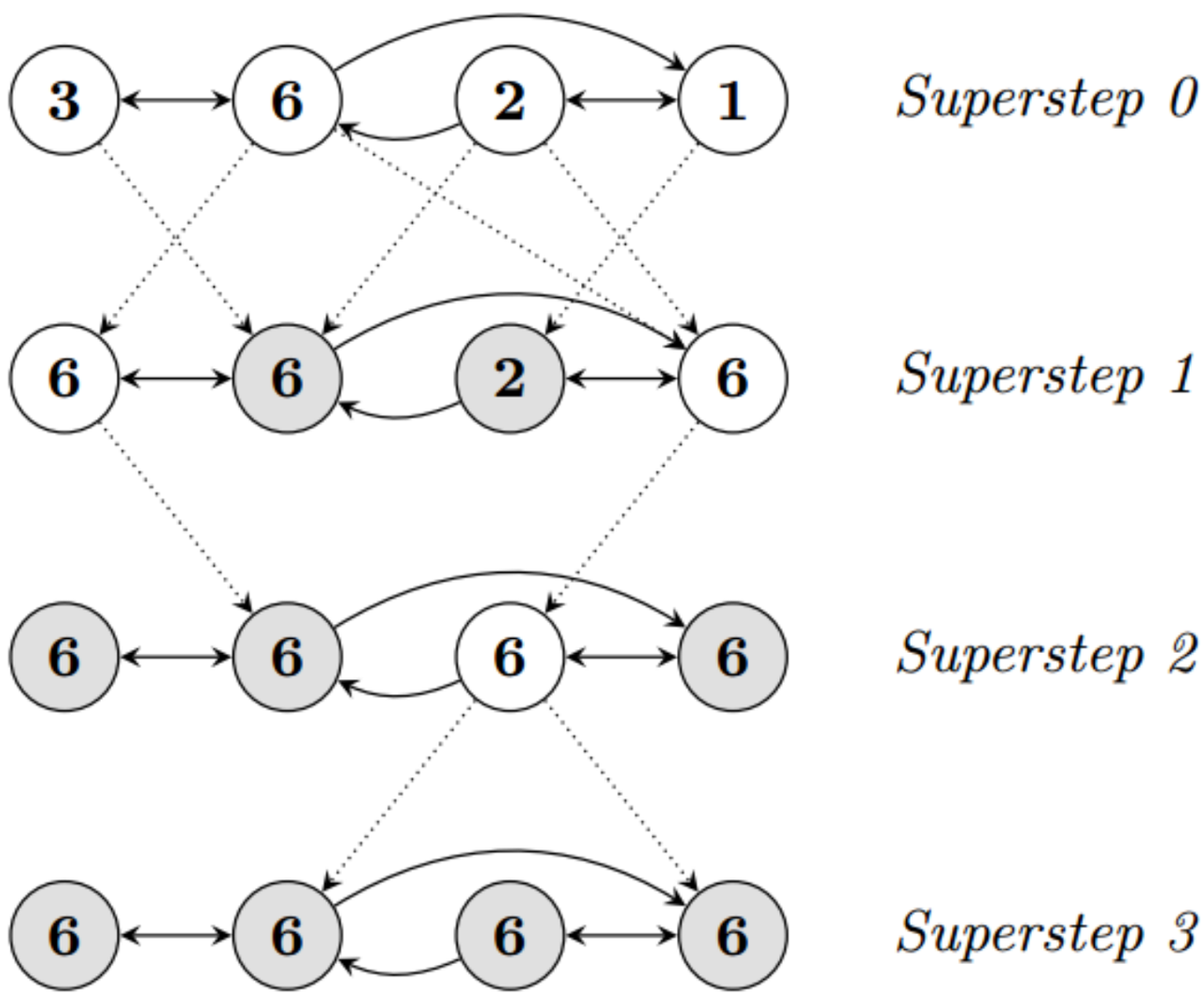
Супершаг - предопределённая функция, исполняемая всеми вершинами:

- ✓ Несколько параллельных независимых процессов (1 процесс = 1 вершина)
- ✓ Вершины могут обмениваться сообщениями
- ✓ Структура графа и состояние вершин может меняться
- ✓ Барьерная синхронизация в конце каждого супершага

Пример: Распределенные графовые фреймворки **Spark.Bagel, Giraph, GraphLab**

Модель Bulk Synchronous Parallel

Пример: поиск максимального значения



Новые модели и алгоритмы

- Изучение закономерностей общения и обмена информацией
- Разнородные данные: граф, тексты сообщений, фото, видео, музыка, логи
- Потребность в новых математических моделях для тестирования гипотез
- Потребность в эффективных алгоритмах для решения новых задач
- Потребность в новых сервисах для облегчения коммуникации

Вопросы?

Social Media Explained
I'm eating bacon
I like bacon
I have skills including eating bacon
This is where I eat bacon
Watch me eat my bacon
Here's a vintage photo of my bacon
Here's a recipe with bacon
I work for Google and eat bacon
I'm listening to music about bacon
last.fm
©COREY SMITH
coreysmith.ws