Анализ актуальных политических предпочтений пользователей социальной онлайн-сети Вконтакте

Над проектом работали:

Козицин Иван (МФТИ, ИПУ РАН), Норкин Дмитрий, Осипов Степан, Утешев Иван, Марченко Артемий (МФТИ), Мягков Михаил (ТГУ, university of Oregon).

При изучении динамики групп к нам подключились Вадим Сушко (МФТИ) и коллеги из ТГУ (Вячеслав Гойко, Роман Палкин, Тимур Самигулин, Эдуард).

Введение

Потенциальная научная значимость

- Онлайн социальные сети можно рассматривать, как большой социальный эксперимент.
- Извлечение мнений – важный шаг при анализе социальной динамики.

Потенциальная практическая значимость

- Проведение социальных опросов.
- Предсказание итогов выборов.
- Грамотное планирование политических кампаний.
- Обеспечение политической стабильности государства.

Введение

Предпосылки

- Переход от Web 1.0 к Web 2.0.
- Сильная персонализированность таких социальных медиа, как Facebook, Twitter и Вконтакте.
- Огромное количество данных, генерируемых пользователями, находящееся в открытом доступе.
- Возможность хранить и обрабатывать эту информацию, превращая ее в релевантное знание.

Подход к решению

Идея: политические взгляды пользователей “проецируются” на их аккаунты.

Для решения задачи необходимо найти, по каким правилам происходит отображение.

Более точно: можно ли по проекции воспроизвести сам объект.

Тип задачи: задача классификации.

Способ решения: методы машинного обучения.

Подходы к решению

Для обнаружения закономерности необходимо найти зависимость между классами пользователей и некоторыми их признаками.

Подходы к выбору признаков

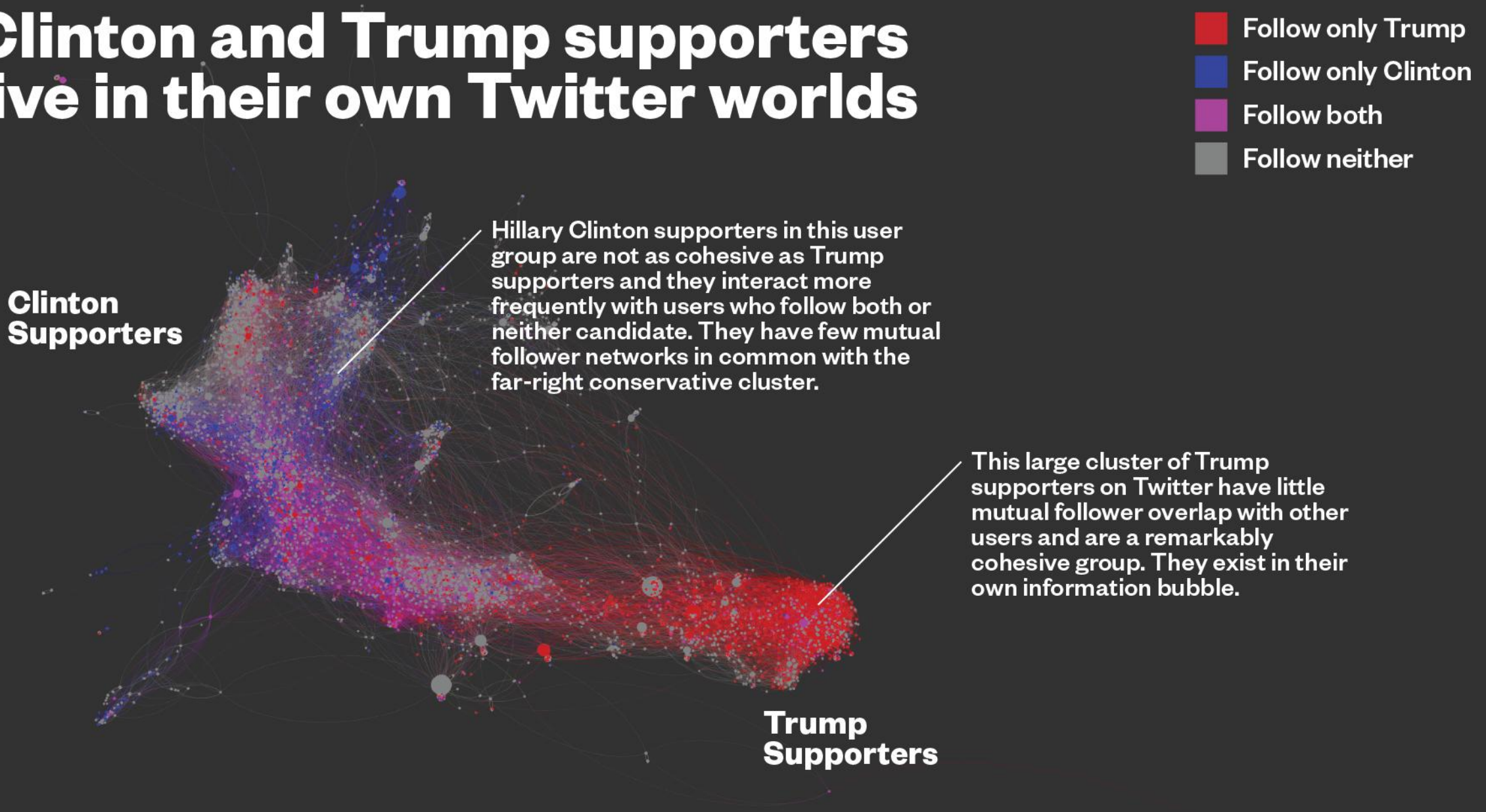
- Свойства текстов, генерируемых пользователями.
- Особенности архитектуры конкретной онлайн социальной сети (лайки, хэштеги, репосты).
- Структура социального графа и эгографов пользователей.
- “Косвенные методы”.
- Комбинации из перечисленных.

Структура социального графа и эгографов пользователей

Что под этим понимается:

- Применяем к онлайн социальной сети модель графа: аккаунты – вершины, связь “дружба” или “подписка” – ребро.
- Накладываем метаданные: политические взгляды некоторых аккаунтов – известны.
- “Близкие” аккаунты с большой долей вероятности имеют одинаковую политическую ориентацию.
- Мерой “близости” аккаунтов служит наличие ребра между соответствующими вершинами.

Clinton and Trump supporters live in their own Twitter worlds



Source: The Electome | The Laboratory for Social Machines at the MIT Media Lab

Критика

Существует ряд замечаний к рассмотренной методологии

1. Зависимости между аккаунтами пользователей и их мнениями может не существовать.
2. Нет гарантии того, что датасет отражает все существующие зависимости.
3. Пользователи могут быть размечены неверно в силу различных причин.
4. Настройки приватности.
5. Умеренные и нейтральные пользователи.
6. Этическая сторона вопроса.

Выбор признаков

Гипотеза селективной экспозиции:

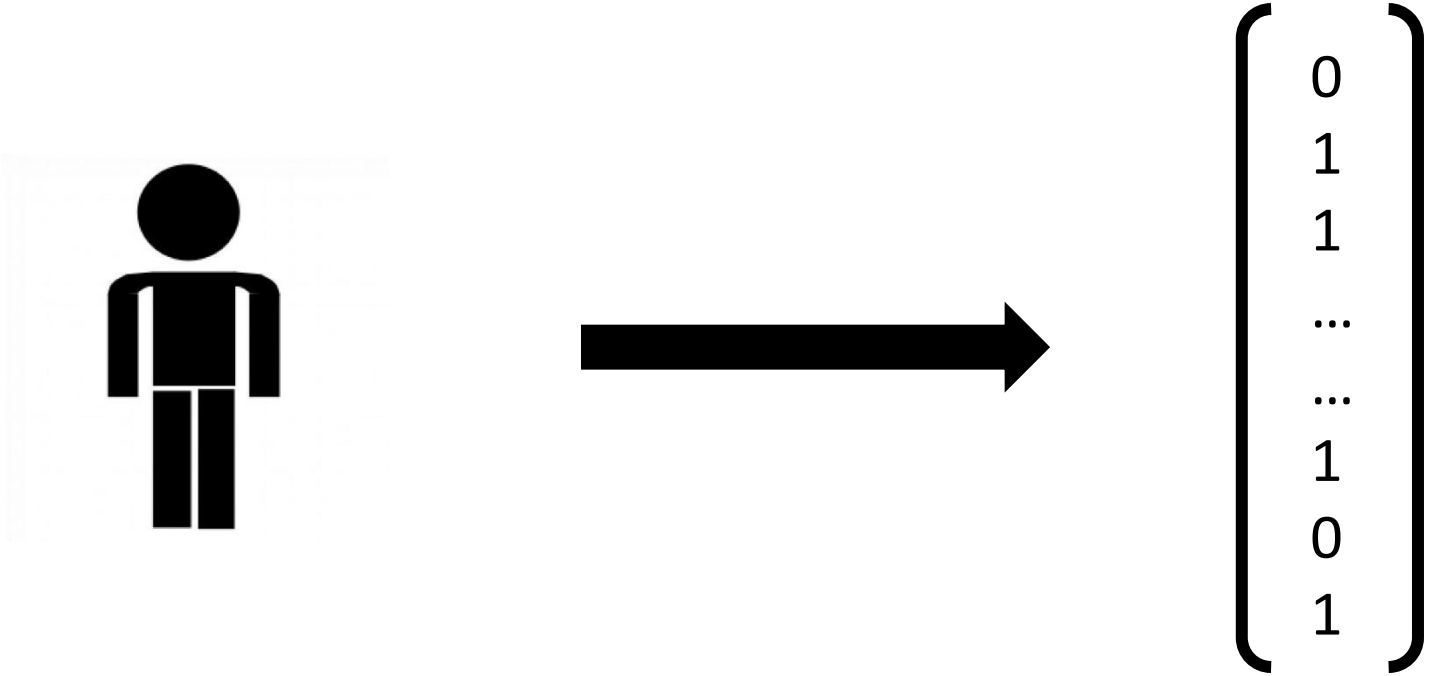
пользователь будет подписываться на источники информации согласованно со своими политическими предпочтениями.

Проще говоря:

пользователь – либерал скорее всего предпочтет новостной источник, который явно или неявно поддерживает либералов.

Выбор признаков

Произвольному пользователю ставился в соответствие вектор, каждая компонента которого соответствовала определенному паблику и принимала два значения: 0 или 1.



Выбор признаков

Преимущества такого подхода:

1. Универсальность.
2. Устойчивость к настройками приватности.
3. Требует данные, извлечение, хранение и обработка которых – относительно простой и быстрый процесс.
4. Результаты предыдущих исследований.

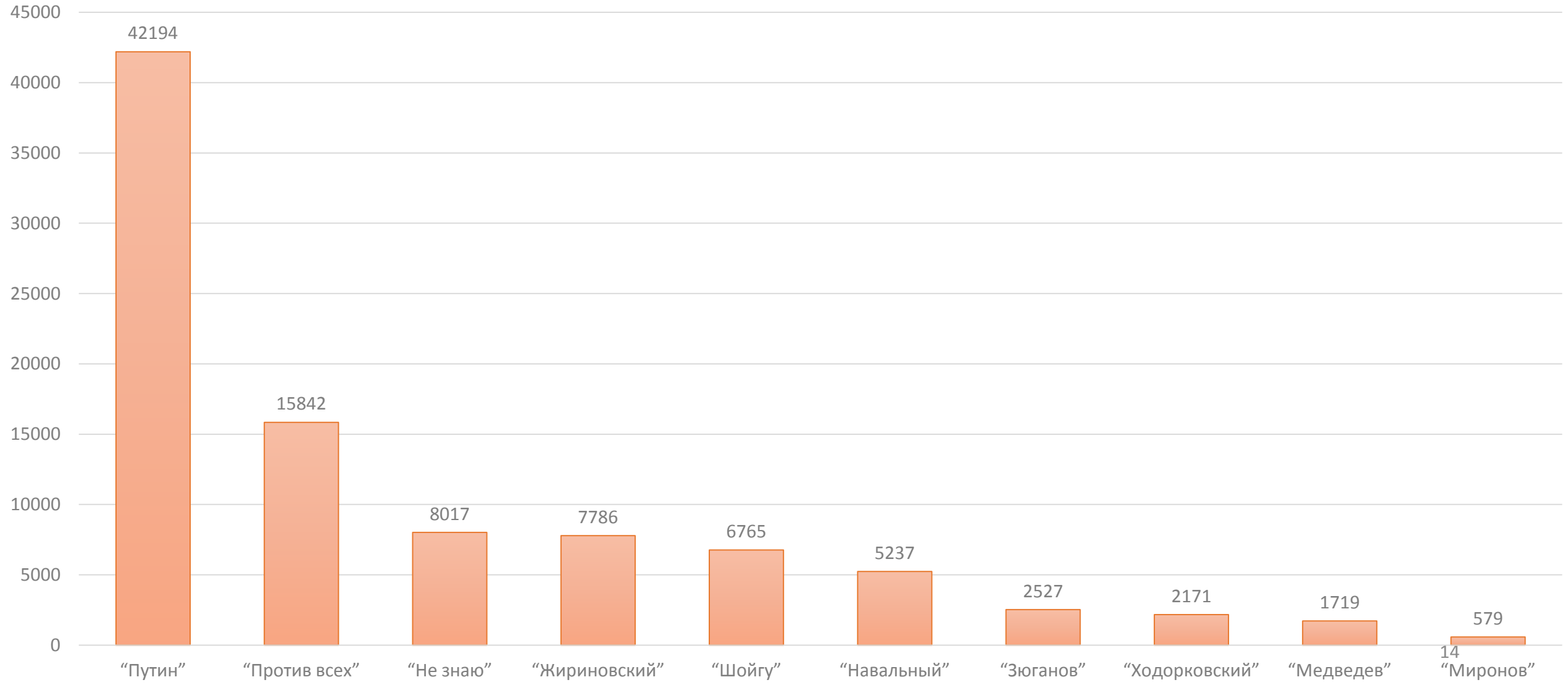
Датасет

В качестве начинки для датасета использовались результаты открытого опроса, проведенного среди пользователей Вконтакте сервисом ЦИМЕС. В нем приняли участие 92815 пользователей.

Подводные камни:

1. Открытость опроса могла повлиять на выбор пользователя.
2. Опрос был проведен в апреле 2016 года, а данные считаны в августе 2017 года.
3. В опросе могли участвовать “боты”.

Результаты опроса

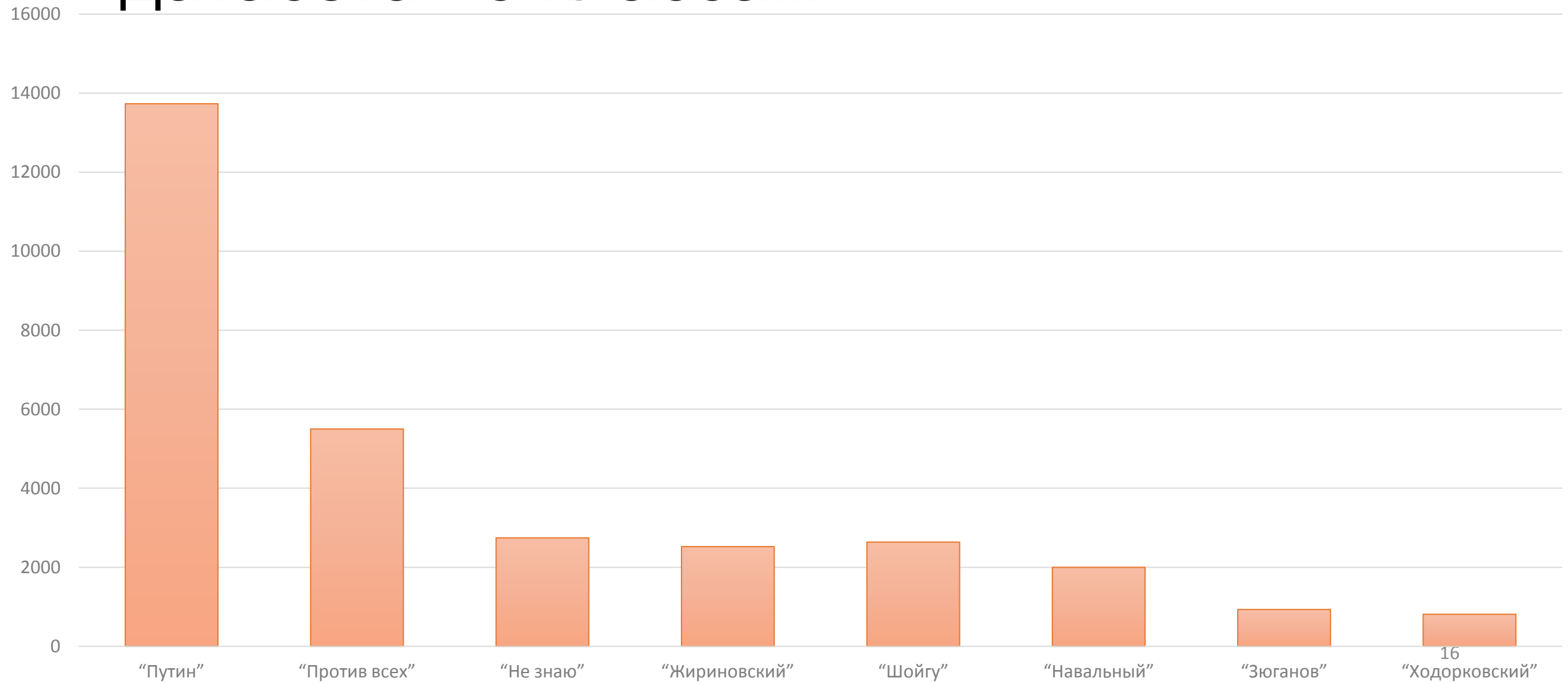


Датасет

Для создания датасета были проделаны следующие шаги:

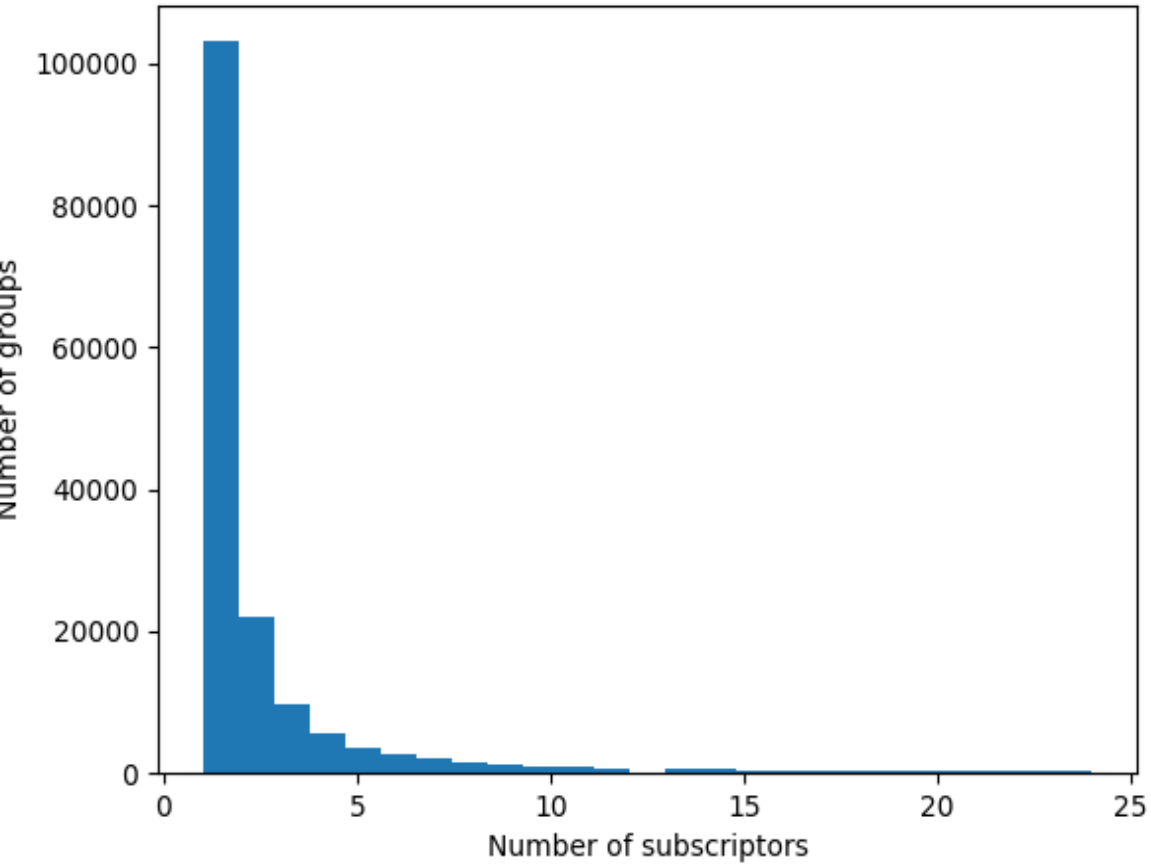
1. Списки пабликов пользователей-респондентов ***старше 18 лет*** были выкачаны с помощью методе get-subscriptions.
2. Сторонники Миронова и Медведева не учитывались ввиду их малого числа.
3. Из полученных 30904 пользователей случайно выбраны 8500.
4. Паблики, на которые подписаны не более 8000 и не менее 15 пользователей из этих 8500 служили базисом для пространства признаков.
5. Полученное пространство имело размерность 7282.
6. Данные хранились в 8 файлах, каждый из которых состоял из матрицы 7282*Nk, где Nk – число пользователей в данном классе k (k=1,…,8). Элементами матриц были нули и единицы.

Распределение пользователей датасета по классам

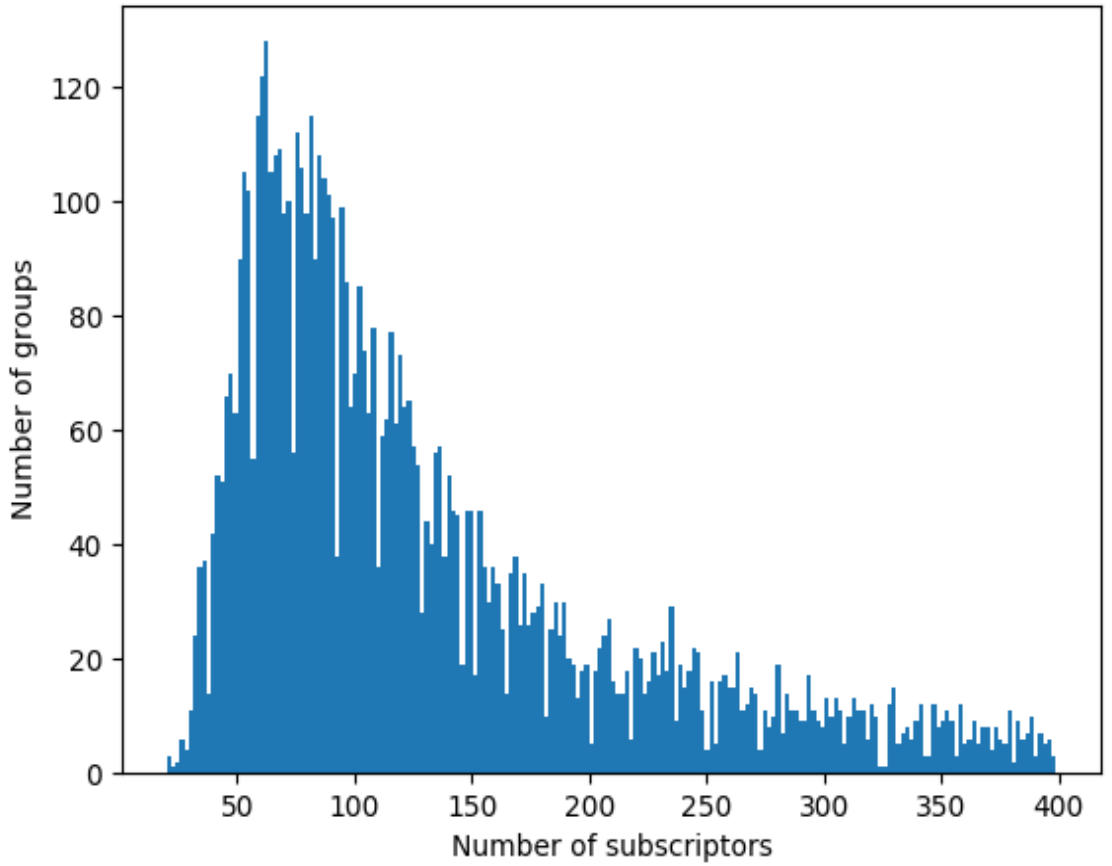


Зачем убирать малочисленные паблики?

До чистки:



После чистки:



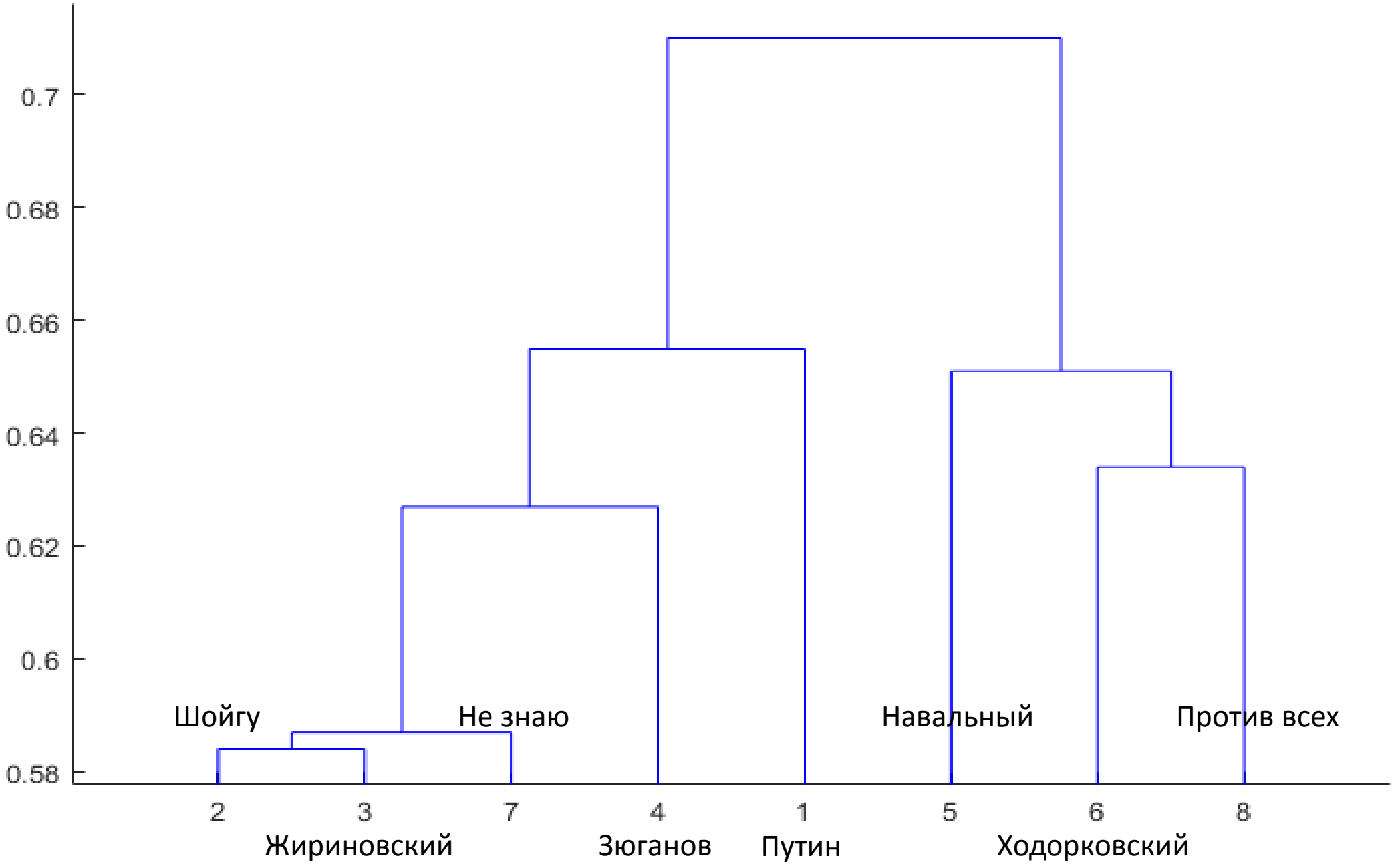
Понижение размерности

- Были использованы различные методы понижения размерности
- Лучше всего зарекомендовал себя ***метод главных компонент***
- Понижение размерности не улучшало точность классификации
- Понижение размерности до 10 – 20 признаков практически не ухудшало ее

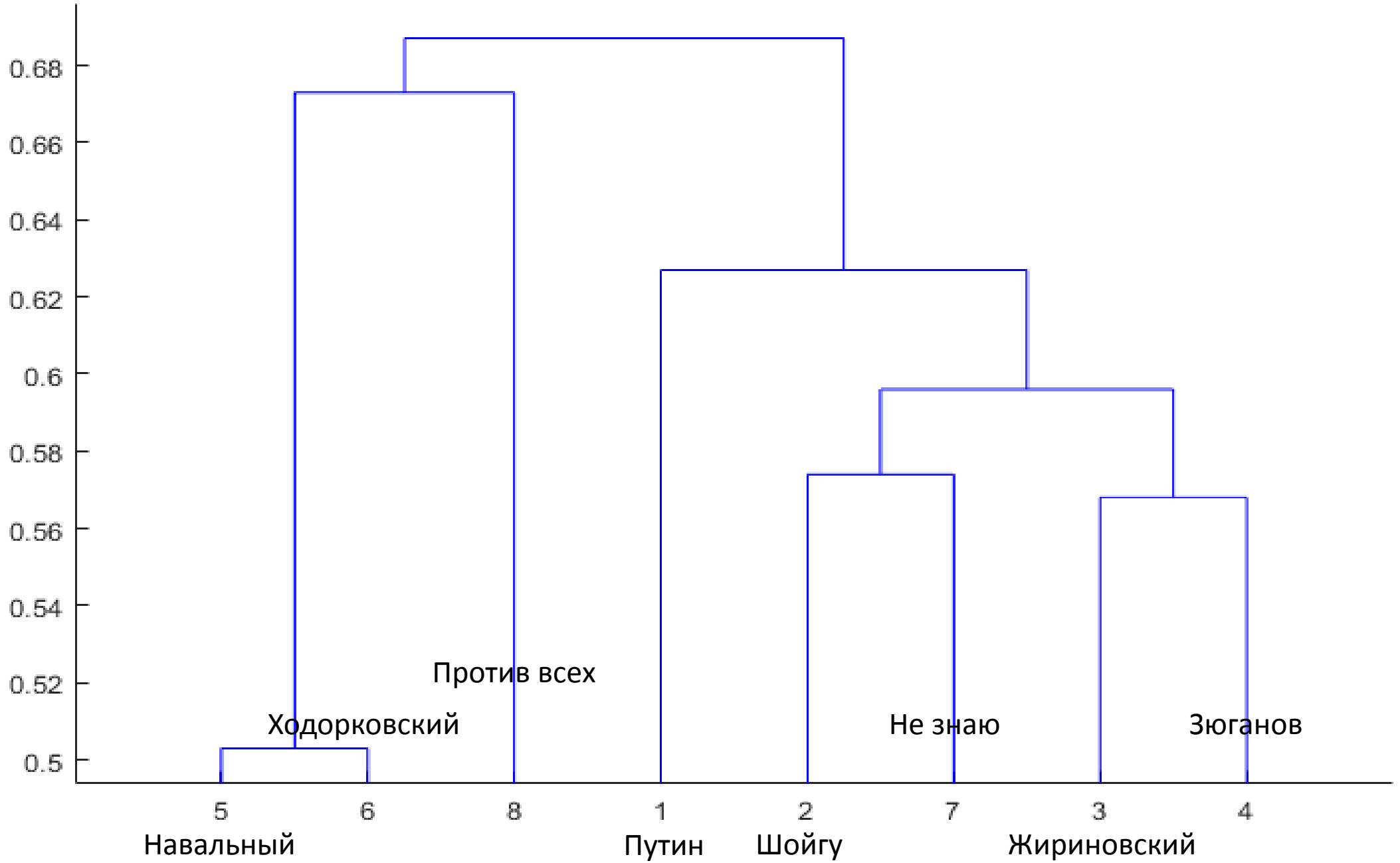
Дизайн экспериментов

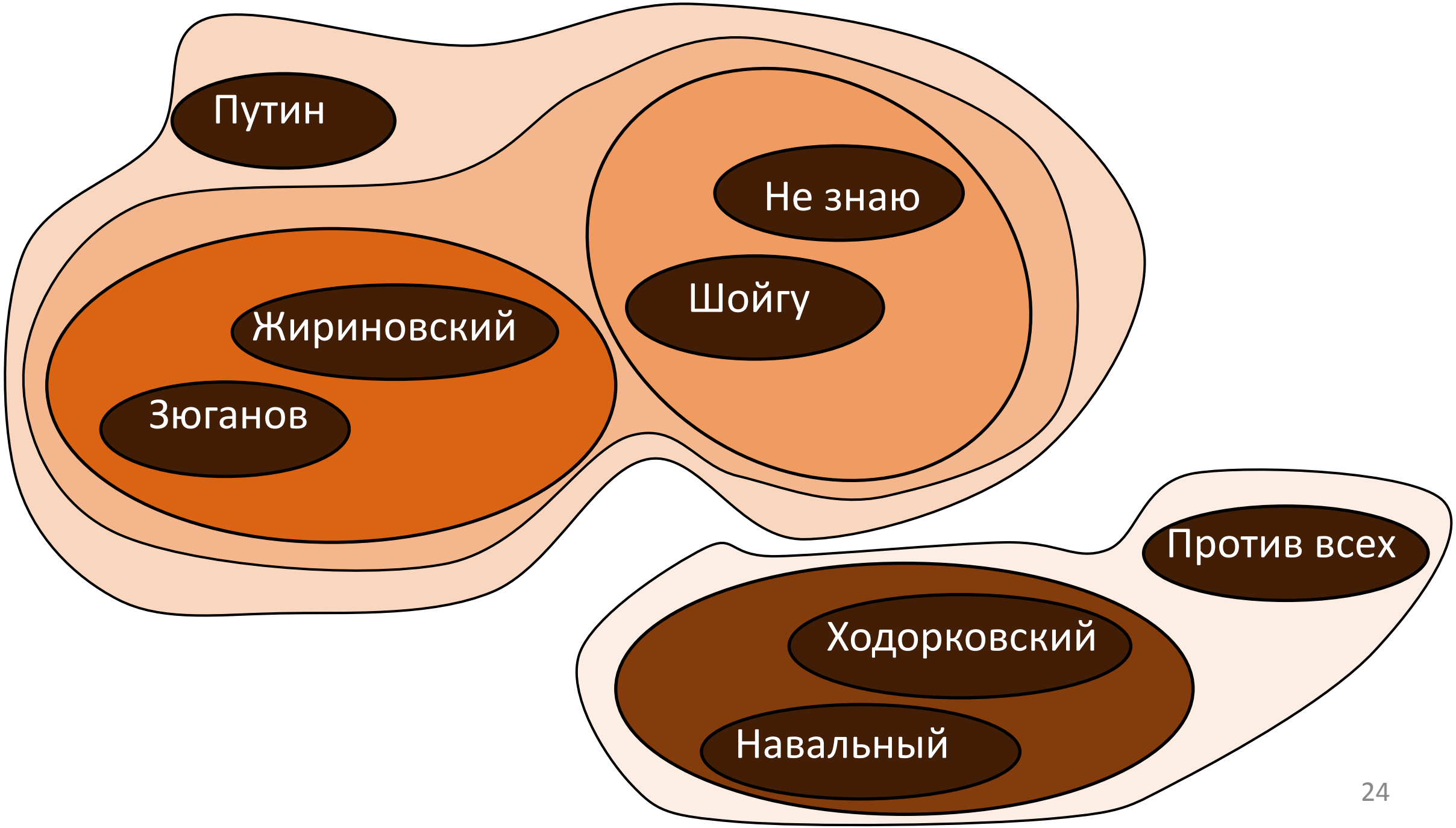
- Мы провели все возможные 28 бинарных классификаций – 28 ***задач***
- Каждая задача подразумевала ***3 – х фолдовую кросс – валидацию***.
- Из каждого участвующего в задании класса случайно выбирались примерно по тысяче пользователей.

	Путин	Шойгу	Жириновский	Зюганов	Навальный	Ходорковский	Не Знаю	Против Всех
Путин		.698	.695	.685	.835	.753	.655	.725
Шойгу	.698		.584	.627	.841	.758	.587	.771
Жириновский	.695	.584		.627	.81	.764	.595	.743
Зюганов	.685	.627	.627		.817	.747	.634	.747
Навальный	.835	.841	.81	.817		.66	.788	.651
Ходорковский	.753	.758	.764	.747	.66		.72	.634
Не Знаю	.655	.587	.595	.634	.788	.72		.71
Против Всех	.725	.771	.743	.747	.651	.634	.71	

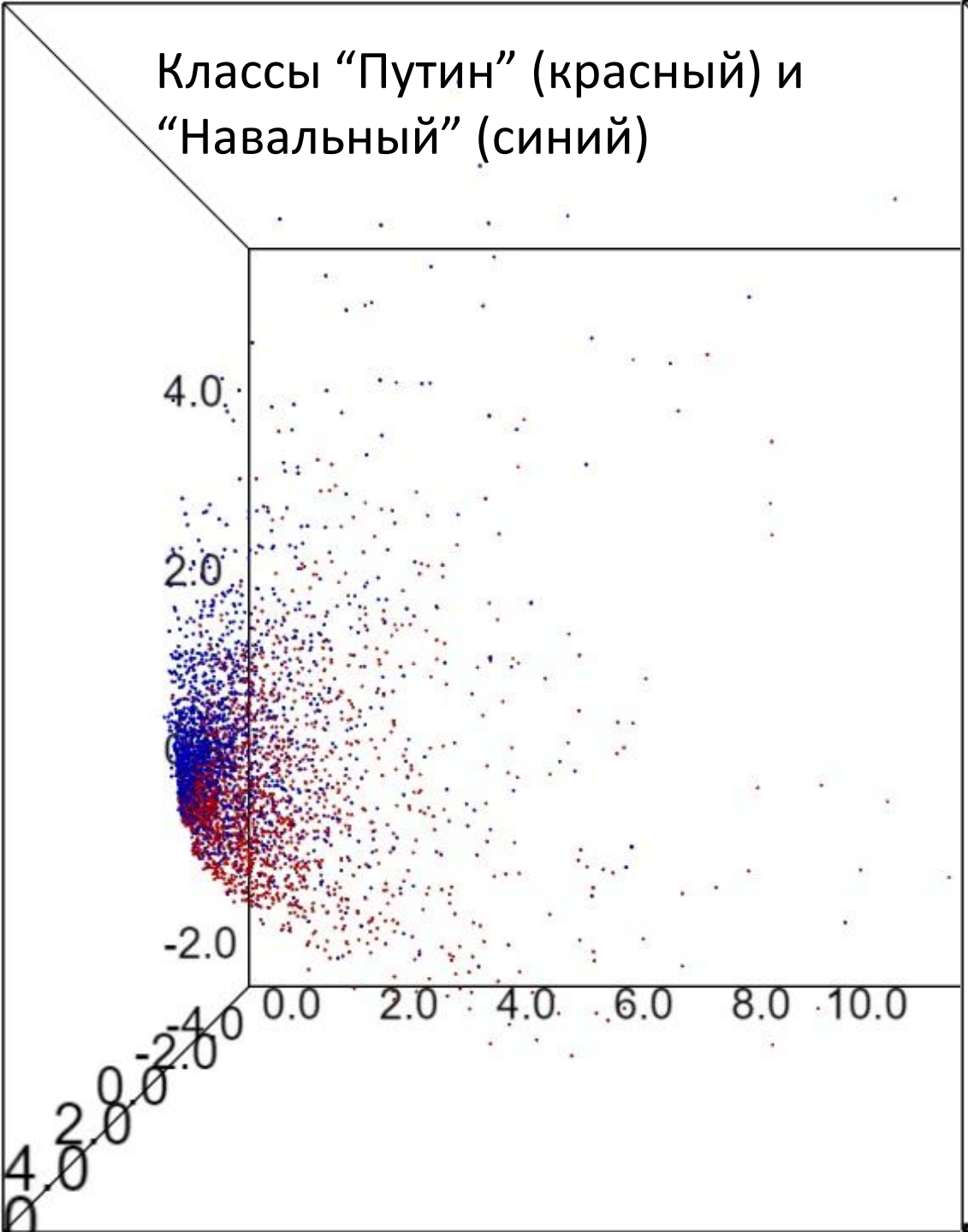

0.7
0.68
0.66
0.64
0.62
0.6
0.58
Шойгу
Не знаю
Навальный
Против всех
2
3
7
4
1
5
6
8
Жириновский
Зюганов
Путин
Ходорковский

	Путин	Шойгу	Жириновский	Зюганов	Навальный	Ходорковский	Не Знаю	Против Всех
Путин		.678	.714	.726	.846	.687	.627	.713
Шойгу	.678		.596	.655	.838	.787	.574	.784
Жириновский	.714	.596		.568	.81	.724	.618	.757
Зюганов	.726	.655	.568		.826	.71	.687	.771
Навальный	.846	.838	.81	.826		.503	.786	.673
Ходорковский	.687	.787	.724	.71	.503		.763	.717
Не Знаю	.627	.574	.618	.687	.786	.763		.73
Против Всех	.713	.784	.757	.771	.673	.717	.73	

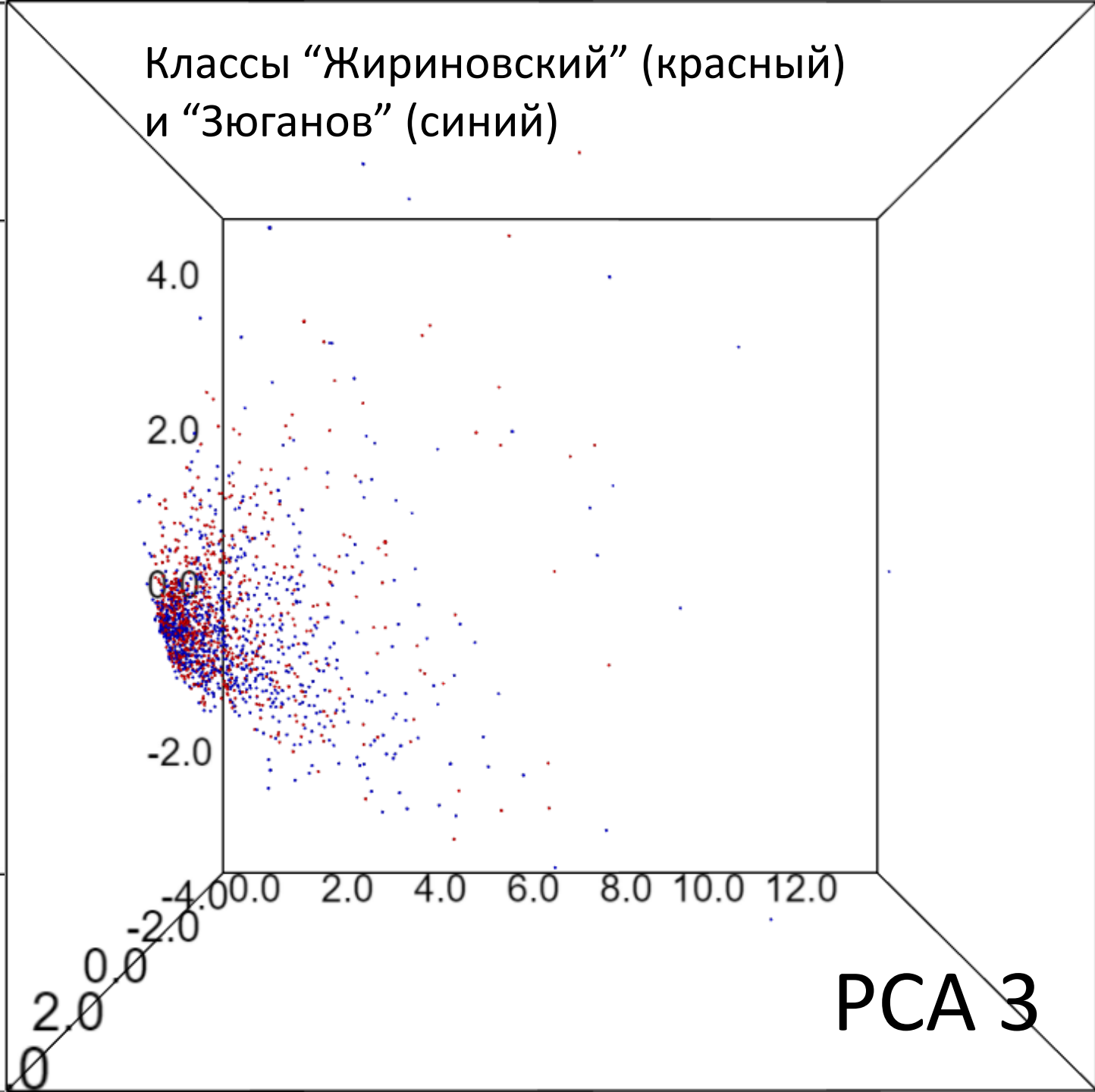

0.68
0.66
0.64
0.62
0.6
0.58
0.56
0.54
0.52
0.5
5
6
8
1
2
7
3
4
Навальный
Ходорковский
Против всех
Путин
Шойгу
Не знаю
Жириновский
Зюганов


Путин
Не знаю
Шойгу
Жириновский
Зюганов
Против всех
Ходорковский
Навальный

Классы “Путин” (красный) и “Навальный” (синий)



Классы “Жириновский” (красный) и “Зюганов” (синий)



PCA 3

Вывод

1. Информационные источники, на которые подписаны пользователи позволяют неплохо отличать сторонников кандидатов “от власти” и сторонников представителей внесистемной оппозиции.
2. Точность классификаций растет с ростом разногласий относительно оценки действующей власти в стране

Еще немного – про динамику онлайн групп

- ***Онлайн – группы*** (в т.ч. паблики) – аккаунты, управляемые некоторыми пользователями (администраторами), которые посвящены чему – либо (в т.ч. пропаганде политических взглядов).
- Полученные выше результаты свидетельствуют о том, что подписка на эти онлайн – группы ***очень сильно коррелирует с политической идеологией***.

Динамика онлайн - групп

Таким образом, динамика онлайн – групп тесно связана с динамикой мнений пользователей, по крайней мере, их политических взглядов.

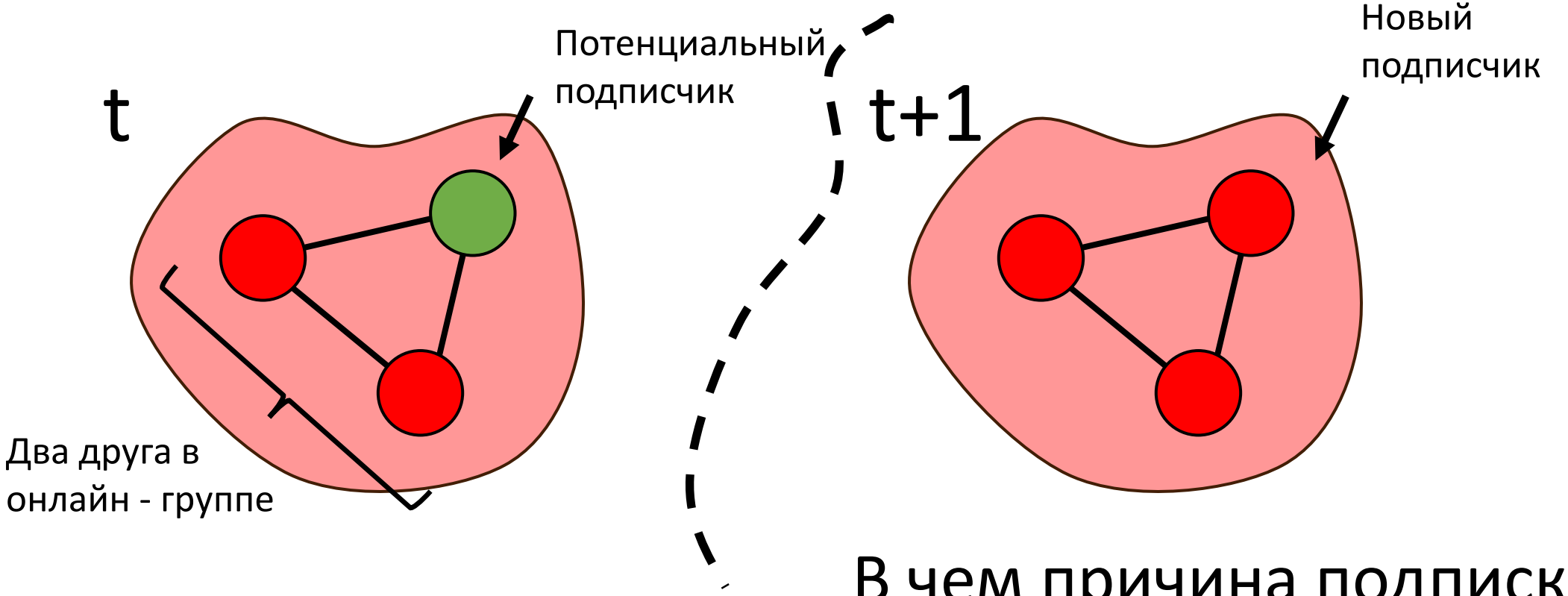
Пользуясь этим, можно попытаться уловить ***межличностное социальное влияние***, которое может проявляться между пользователями.

Возникает вопрос:

“Как друзья пользователя, подписанные на конкретную онлайн - группу влияют на решение пользователя подписаться на нее?”

Динамика онлайн - групп

Нужно быть осторожным: очень трудно отличить явление ***социального влияния*** и эффект ***ассортативности*** со стороны.



В чем причина подписки?

Динамика онлайн - групп

Существуют два объяснения:

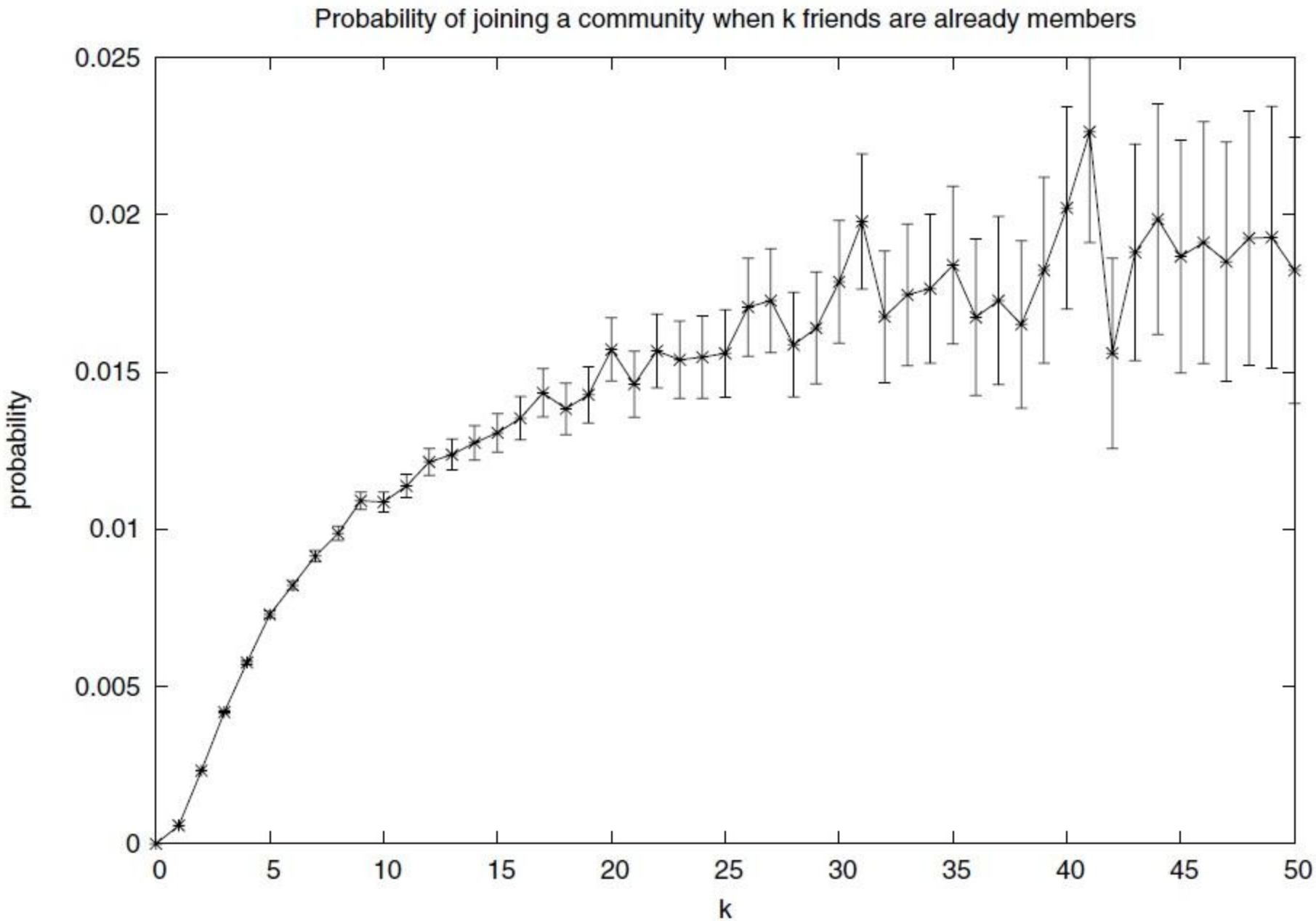
1. Гипотеза об ***ассортативности***: “дружба” (в том числе и в Вконтакте) образуется между “похожими” людьми, которые, вероятно, ведут себя одинаково в одних и тех же ситуациях (“похожие” -> “дружба”).
2. Эффект ***социального влияния***: есть “дружба”, которая и является каналом распространения поведения (“дружба” -> “похожие”).

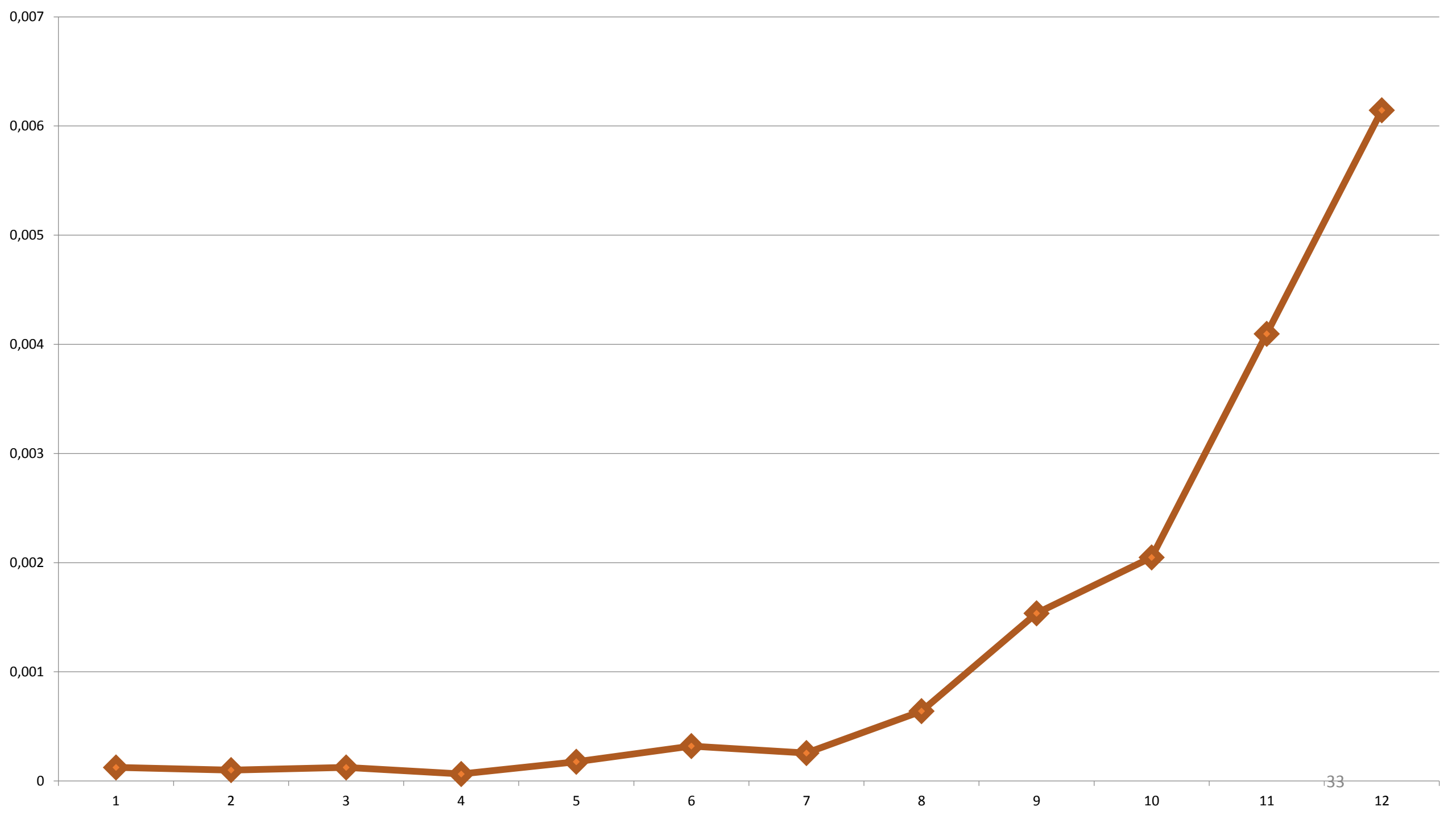
Динамика групп

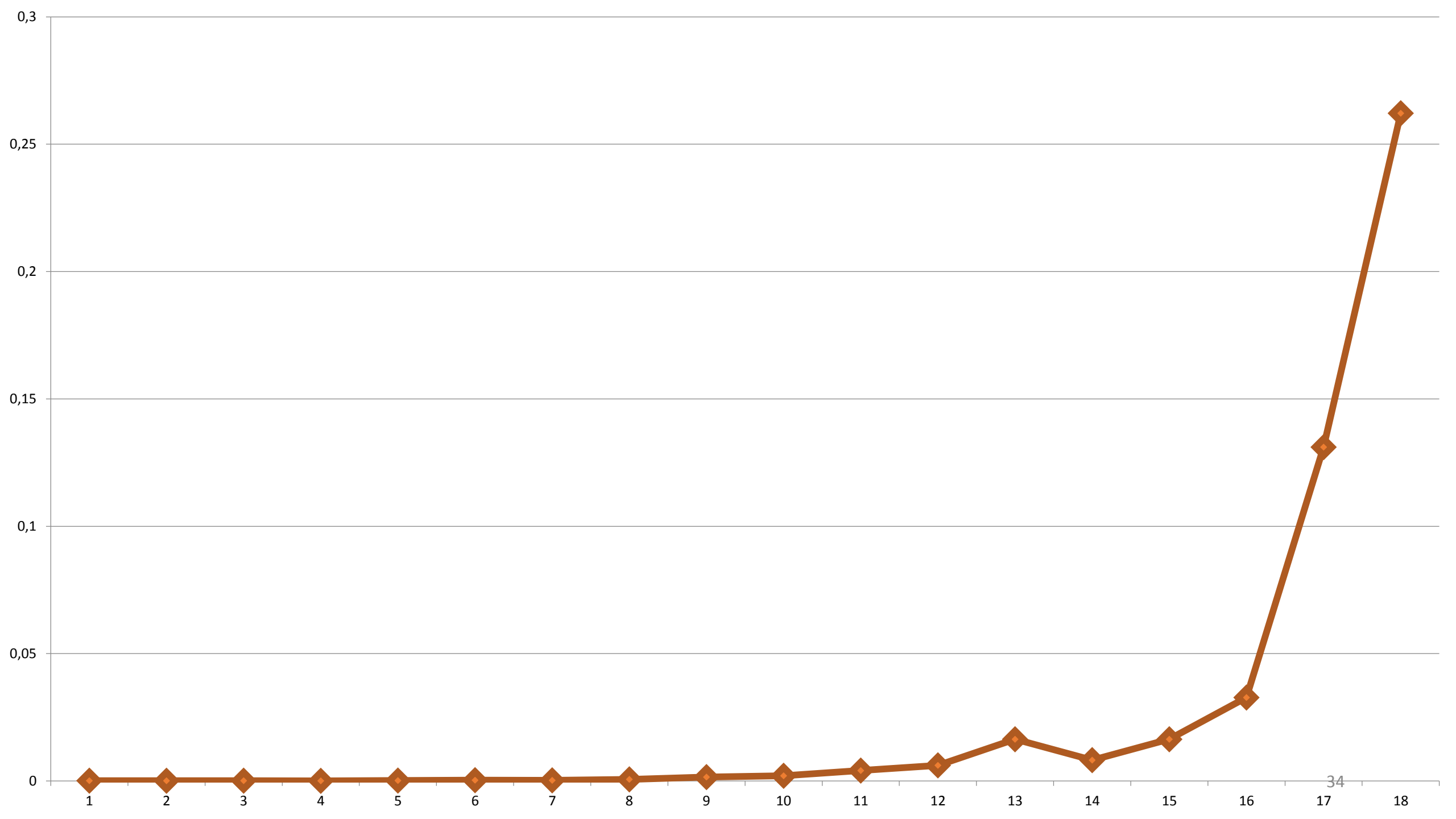
Один из способов оценить эти явления (не разделяя их) – зафиксировать онлайн группу, а потом посмотреть за ее динамикой.

Важно то, что состав участников групп находится в открытом доступе.

Таким образом, можно построить зависимость вероятности вступления пользователя в группу в зависимости от числа его друзей, уже находящихся в этой группе.


Probability of joining a community when k friends are already members
probability
k
0.025
0.02
0.015
0.01
0.005
0
0
5
10
15
20
25
30
35
40
45
50


0,007
0,006
0,005
0,004
0,003
0,002
0,001
0
1
2
3
4
5
6
7
8
9
10
11
12


0,3
0,25
0,2
0,15
0,1
0,05
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Спасибо за внимание!