Кластеризация

Теория кластеризации в социальных сетях

1. Коэффициент кластеризации

Коэффициент кластеризации - ключевая метрика, показывающая, насколько плотно связаны друзья пользователя между собой.

Формула для узла v:
$C(v) = \frac{(2\times\text{количество треугольников, содержащих v})}{(k_v\times(k_v-1))}$

Где:
- k_v - степень узла v (количество друзей)
- Числитель - удвоенное количество треугольников (т.к. каждый треугольник учитывается для двух ребер)
- Знаменатель - максимально возможное количество связей между друзьями

Интерпретация:
- $C(v) = 1$ - все друзья пользователя дружат между собой (полный граф)
- $C(v) = 0$ - ни один друг пользователя не дружит с другими друзьями

Источник: [04ianote.pdf, с. 6]

2. Обнаружение сообществ

Социальные сети естественным образом разделяются на плотно связанные подграфы (сообщества), такие как:
- Школьные/университетские друзья
- Рабочие коллеги
- Друзья по хобби

Алгоритмы обнаружения сообществ:
1. Метод Лувена (Louvain) - оптимизирует модулярность графа
2. Girvan-Newman - удаляет ребра с высокой центральностью по посредничеству
3. Label Propagation - быстрый алгоритм на основе распространения меток

Модулярность Q:
$Q = \Sigma[e_i i - (a_i)^2]$

Где:
- $e_i i$ - доля ребер внутри сообщества i
- a_i - доля ребер, инцидентных узлам сообщества i

3. Теоретическая основа для деанонимизации через кластеризацию

Почему кластеризация помогает деанонимизировать?
1. Уникальность комбинации сообществ:
- Пользователь принадлежит к нескольким сообществам (работа, школа, спорт)
- Комбинация этих принадлежностей уникальна, как “структурный отпечаток”

1. Мостики между сообществами:
    - Пользователи с низкой кластеризацией часто являются “мостами” между сообществами
    - Эти узлы особенно информативны для идентификации
2. Структурная неоднородность:
    - Согласно [isp_26_2014_1_439.pdf, с. 39] “...поиск сообществ пользователей и измерение информационного влияния между пользователями.”

Насчёт алгоритмов

Представим, что у нас есть большая компания людей на вечеринке. Все между собой как-то знакомы, но внутри есть небольшие группки, которые чаще общаются друг с другом. Алгоритмы обнаружения сообществ как раз ищут эти группки. Вот как работают три основных метода:

1. Метод Лувена (Louvain)

Как будто: Ты предлагаешь людям объединиться в команды так, чтобы внутри команды все активно общались, а между командами - редко.

Как работает:
- Сначала каждый человек - отдельная “команда”
- Потом люди по одному переходят в соседние команды, если это улучшает общение внутри (становится больше разговоров внутри команды и меньше снаружи)

- Процесс повторяется, пока не найдётся оптимальное разделение

Пример из жизни: Как в школе - сначала ты объединяешься с теми, с кем сидишь за одной партой, потом эти парты объединяются в классы, классы - в параллели.

Плюс: Очень быстрый и эффективный для больших сетей (как Facebook).

2. Girvan-Newman

Как будто: Ты ищешь “мостики” между группами и перерезаешь их, пока группки не отделятся друг от друга.

Как работает:
- Считает, какие связи (рукопожатия между людьми) являются “мостами” между группами
- Удаляет самые важные мостики (связи, через которые проходит больше всего “сообщений” между людьми)
- Повторяет, пока не останутся отдельные группки

Пример из жизни: В большом городе есть районы, соединённые мостами. Чтобы разделить город на независимые районы, сначала закрывают самые загруженные мосты.

Плюс: Точно находит границы между сообществами.
Минус: Медленный для больших сетей (как раз потому, что считает все “мостики”).

3. Label Propagation (Распространение меток)

Как будто: Каждый человек сначала придумывает себе имя группы, а потом начинает копировать имена у большинства своих друзей.

Как работает:
- Каждый человек получает уникальную “метку” (например, свой номер)
- На каждом шаге человек меняет свою метку на ту, которая чаще встречается у его друзей
- Через несколько шагов все в одной группе имеют одинаковую метку

Пример из жизни: На вечеринке сначала все говорят разные сленговые слова, но со временем в каждой группе устанавливается свой сленг, и все в группе начинают говорить одинаково.

Плюс: Очень быстрый - как слухи, распространяющиеся по компании.

Минус: Иногда получается неоднозначно (может зависеть от порядка обработки людей).

Какой когда использовать?

- Лувен - когда нужно быстро обработать большую сеть (например, Facebook с миллиардами пользователей)
- Girvan-Newman - когда важна точность, а сеть небольшая (например, анализ школьного класса)
- Label Propagation - когда нужно очень быстро получить приблизительный результат (например, для предварительного анализа)