Теоретические основы деанонимизации пользователей при помощи теории графов

1. Основные понятия теории графов

1.1. Определение графа

Граф $G(V, E)$ - упорядоченное множество, состоящее из узлов V (вершин) и рёбер E (связей между вершинами).

Применение в работе:
- Социальная сеть представляется как граф, где вершины - пользователи, рёбра - дружеские связи.
- Анонимизация заключается в удалении идентификаторов вершин, оставляя только структуру графа.

---

2. Метрики центральности (для анализа уязвимости вершин)

2.1. Степенная центральность (Degree Centrality)

Формула:
$C_D(v) = \frac{deg(v)}{n-1}$

Где:
- $deg(v)$ - степень вершины v (количество связей),
- n - общее количество вершин в графе.

Применение в работе:
- Вершины с высокой степенной центральностью имеют много связей и образуют “хабы” в графе.
- В социальных сетях это популярные пользователи, чья структура окружения уникальна и уязвима для деанонимизации.
- Источник: [1, с. 6] - “Транспортная карта Цюриха. Явно выделяются центральные узлы, которые выполняют функцию хабов”.

2.2. Центральность по посредничеству (Betweenness Centrality)

Формула:
$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Где:
- σ_{st} - количество кратчайших путей между вершинами s и t,
- $\sigma_{st}(v)$ - количество кратчайших путей между s и t, проходящих через вершину v.

Применение в работе:
- Вершины с высокой центральностью по посредничеству соединяют разные части графа.
- В социальных сетях это пользователи, объединяющие разные сообщества (например, коллеги + одноклассники).
- Их деанонимизация позволяет "вскрыть" структуру всего графа.
- Источник: [2, с. 10] - "Betweenness - алгоритм на основе коэффициента 'центральности по посредничеству'".

2.3. Собственная центральность (Eigenvector Centrality)

Формула:
$$Ax = \lambda x$$

Где:
- A - матрица смежности графа,
- λ - собственное значение,
- x - собственный вектор (вектор центральности).

Применение в работе:
- Учитывает не только количество связей, но и "вес" соседних вершин.
- Вершины с высокой собственной центральностью находятся в окружении влиятельных пользователей.
- Это повышает их уникальность и уязвимость к деанонимизации.
- Источник: [3, с. 14] - "Eigenvector Centrality".

2.4. PageRank

Формула:

$$PR(u) = (1 - d) + d \cdot \sum_{v \in M(u)} \frac{PR(v)}{C(v)}$$

Где:
- d - коэффициент затухания (обычно 0.85),
- $M(u)$ - множество вершин, ссылающихся на вершину u,
- $C(v)$ - количество исходящих связей из вершины v.

Применение в работе:
- PageRank оценивает влияние и уязвимость вершин в графе.
- Вершины с высоким PageRank имеют уникальные структурные позиции.
- В работе используется для ранжирования узлов по уязвимости к деанонимизации.
- Источник: [4, с. 32] - “Коэффициент затухания (Damping factor) = 0.85, как в классическом PageRank”.

---

3. Свойства социальных графов

3.1. Коэффициент кластеризации

Формула для локального коэффициента:
$$C_i = \frac{\text{число треугольников с вершиной } i}{\text{число "вилок", центром которых является вершина } i}$$

Глобальный коэффициент кластеризации:
$$C = \frac{3 \times \text{число треугольников в графе}}{\text{число "вилок" в графе}}$$

Применение в работе:
- Высокий коэффициент кластеризации указывает на наличие плотных сообществ.
- Уникальные комбинации треугольников и “вилок” служат “отпечатками” для деанонимизации.
- Источник: [5, с. 6] - “Существует другой подход к вычислению коэффициента C. Посчитаем количество треугольников…”.

3.2. Модулярность графа

Формула:
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Где:

- A_{ij} - элемент матрицы смежности,
- k_i, k_j - степени вершин i и j,
- m - общее количество рёбер в графе,
- $\delta(c_i, c_j)$ - дельта-функция (1 если вершины в одном сообществе, иначе 0).

Применение в работе:
- Модулярность помогает выявить сообщества в графе.
- Высокая модулярность указывает на наличие четко выраженных сообществ, которые могут быть использованы как “отпечатки” для деанонимизации.
- Источник: [2, с. 8] - “Модулярность графа (Q)”.
- Источник: [6, с. 41] - “В этой формуле просчитывается количество связей, находящихся внутри одного кластера, и потом они складываются”.

3.3. Степенное распределение

Формула:
$P(k) \sim k^{-\gamma}$

Где:
- $P(k)$ - вероятность того, что вершина имеет степень k,
- γ - показатель степени (обычно между 2 и 3 для социальных сетей).

Применение в работе:
- Степенное распределение означает, что большинство вершин имеют небольшую степень, но есть небольшое количество вершин с очень высокой степенью (хабы).
- Эти хабы играют ключевую роль в структуре графа и могут быть использованы для деанонимизации.
- Источник: [5, с. 7] - “Степенные распределения в сетях”.

3.4. Свойство малого мира

Формула для средней длины пути:
$L = \frac{1}{n(n-1)} \sum_{i \neq j} d(i, j)$

Где:
- $d(i, j)$ - длина кратчайшего пути между вершинами i и j.

Применение в работе:

- Свойство малого мира (небольшая средняя длина пути) означает, что даже удаленные узлы связаны короткими путями.
- Это упрощает распространение информации о структуре графа и делает деанонимизацию более эффективной.
- Источник: [1, с. 12] - "В этой сумме самое большое слагаемое – z^l, т.е. приближенно $z^l = N$. Тогда для $N \approx 6,7$ млн и $z = 50$ (друзей) получаем $L \approx 5,8$, т.е. подтверждается идея о 6 рукопожатиях".

---

4. Методы выделения сообществ

4.1. Алгоритм Walktrap

Формула расстояния между вершинами:
$d(u, v) = \sqrt{\sum_{i=1}^{n}(p_i(u) - p_i(v))^2}$

Где:
- $p_i(u)$ - вероятность того, что случайное блуждание длины t из вершины u закончится в вершине i.

Применение в работе:
- Алгоритм Walktrap показывает высокую точность в выделении сообществ.
- В работе используется для выявления структурных особенностей графа, используемых для деанонимизации.
- Источник: [2, с. 14] - "§5.5 Walktrap".

4.2. Split-Join distance

Формула:
$sjd(C_1, C_2) = \frac{1}{n}(|C_1 - C_2| + |C_2 - C_1|)$

Где:
- C_1, C_2 - два разбиения графа на сообщества,
- $|C_1 - C_2|$ - минимальное количество операций, необходимых для преобразования C_1 в C_2.

Применение в работе:
- Split-Join distance используется для оценки качества выделения сообществ.
- Это критично для оценки точности деанонимизации.

- Источник: [2, с. 9] - “Редакторское расстояние для разбиений (split-join distance)”.

4.3. Нормализованная взаимная информация (NMI)

Формула:
$I_{norm}(X, Y) = \frac{2I(X,Y)}{H(X)+H(Y)}$

Где:
- $I(X, Y)$ - взаимная информация между разбиениями X и Y,
- $H(X)$, $H(Y)$ - энтропии разбиений.

Применение в работе:
- NMI используется для сравнения разных разбиений графа на сообщества.
- Помогает оценить стабильность структуры графа и его уязвимость к деанонимизации.
- Источник: [2, с. 9] - “Нормализованная взаимная информация”.

---

5. Методы деанонимизации

5.1. JLA-модель (Joint Link-Attribute)

Формула гибридной близости:
$\text{total\_score} = \alpha \cdot \text{struct\_sim} + (1 - \alpha) \cdot \text{attr\_sim}$

Где:
- α - вес структурной близости (обычно 0.7),
- struct_sim - структурная близость,
- attr_sim - близость атрибутов (цифровых отпечатков).

Структурная близость через коэффициент Жаккара:
$\text{struct\_sim}(v, u) = \frac{|N(v) \cap N(u)|}{|N(v) \cup N(u)|}$

Где:
- $N(v)$, $N(u)$ - окрестности вершин v и u.

Близость атрибутов:
$\text{attr\_sim}(v, u) = \sum_i w_i \cdot \delta(a_i(v), a_i(u))$

Где:

- w_i - вес атрибута i,
- $\delta(a_i(v), a_i(u))$ - 1 если атрибуты совпадают, иначе 0.

Применение в работе:
- JLA-модель объединяет структурную информацию и атрибуты профилей для повышения точности деанонимизации.
- Источник: [7, с. 8] - "Рис. 2: Структура 'JLA-модели'".

5.2. Сетевой коэффициент Дайса

Формула:
$\text{network-distance}(v, u) = 1 - 2 \cdot \frac{w(L_v \cap L_u)}{w(L_v) + w(L_u)}$

Где:
- L_v, L_u - множества вершин, связанных с v и u соответственно,
- $w(L) = |L|$ - вес множества.

Применение в работе:
- Коэффициент Дайса используется для измерения структурного сходства между вершинами.
- Критично для сопоставления графов в процессе деанонимизации.
- Источник: [7, с. 7] - "В данной работе в качестве функции расстояния в графе B используется коэффициент Дайса".

1 2073 (1).pdf
2 2015_417_SlavnovKA.pdf
3 Python for Graph and Network Analysis
4 podhody-k-matematicheskoy-otsenke...
5 04ianote.pdf
6 Благов А.В.
7 identifikatciia_polzovatelei_sotcialnykh_setei.pdf
8 isp_26_2014_1_439.pdf