МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра теоретических основ компьютерной безопасности и криптографии

Графы социальных сетей

АВТОРЕФЕРАТ
дипломной работы

студента 6 курса 631 группы
специальности 10.05.01 Компьютерная безопасность
факультета компьютерных наук и информационных технологий
Заварина Максима Сергеевича

Научный руководитель
д. ф.-м. н., доцент ____________ М. Б. Абросимов
22.01.2022 г.

Заведующий кафедрой
д. ф.-м. н., доцент ____________ М. Б. Абросимов
22.01.2022 г.

Саратов 2022

ВВЕДЕНИЕ

В нашем мире существует огромное количество информации, которую можно представить в виде графов. Например: генеалогические деревья людей, схема метро, протяженность дорог, схема транспорта, молекулярная структура вещества, электрические цепи электроприборов, сайты, научные статьи и т.д.

Рассмотрим социальные сети. Их тоже можно представить в виде графа. Они стали частью нашей жизни и важной задачей становится их анализ. Этот анализ представляет собой исследование, рассматривающее социальные отношения в терминах теории сетей. Через анализ социальных сетей, можно получить разные взаимосвязи. Исследование графов социальных сетей широко применяется в социологии, экономике, медицине.

Социальные графы – очень мощный аналитический инструмент. Благодаря им можно получить уникальную информацию, например, о неформальной структуре организации.

Бывают разные социальные графы: графы публикаций научных статей, графы взаимосвязи сотрудников в компании, графы интересов и многие другие. Они обладают некоторым набором специфических свойств, характерных только для такого типа графов. Для изучения схожих свойств используют случайные графы.

Целью данной работы будет исследование и анализ социальных графов, а также исследование алгоритмов для генерации случайных графов, приближенных к графам социальных сетей. Для анализа был выбран граф друзей из сети ВКонтакте и граф актеров по версии IMDb.

Дипломная работа состоит из введения, 8 разделов, заключения, списка использованных источников и 4 приложений. Общий объем работы – 76 страниц, из них 40 страниц – основное содержание, включая 31 рисунок и список использованных источников из 14 наименований.

КРАТКОЕ СОДЕРЖАНИЕ

1 Основные понятия теории графов

В данном разделе вводятся основные понятия теории графов, такие как: неориентированный граф, путь, длина пути, цикл, цепь, диаметр, изоморфизм, компонента связности, ориентированный граф, клика, мост, точка сочленения, граф маленького мира, решетка.

2 Социальный граф и его применение

Социальный граф – это граф, вершины которого являются социальными объектами, например, профили каких-либо пользователей, а ребра – социальные связи между ними.

Задачи, которые решает социальный граф:

1) Идентификация пользователей. Обнаружение профилей, принадлежащих конкретному человеку, в нескольких социальных сетях. Решение этой задачи позволяет получить более полный социальный граф, что очень полезно во многих задачах, таких как:

- Социальный поиск;
- Поиск социальных объектов (пользователей, их данных и т. д.), основанный на анализе набора связей, в которых находятся искомые объекты;
- Генерация рекомендаций.

2) Выявление «настоящих» связей. При применении такого подхода из открытых источников, выявляются настоящие друзья, родственники и т.д.

3 Метрики для социальных графов

Существует множество разных метрик для социальных графов. Приведем только те, для которых будет проведено исследование. Будем формулировать определения в терминах социальных сетей.

1) Центральность – степень, которая показывает «важность» определенного пользователя (кластера пользователей) внутри графа.

2) Треугольник – эта метрика показывает количество треугольников у каждого пользователя.

3) Клика – группа, в которой все пользователи имеют «прямые» связи (вершины соединены ребром) друг с другом.

4) Коэффициент кластеризации – вероятность того, что два разных пользователя, связанные с общим для них человеком, тоже будут связаны. Формула для вычисления:

$$c_x = \frac{2T(x)}{d(x)(d(x)-1)}$$

5) Граф принадлежит к маленькому миру если у него небольшая средняя длина кратчайшего пути и большой коэффициент кластеризации. В контексте социальной сети это значит, что двух незнакомых пользователей связывает небольшое количество промежуточных знакомых. Для определения принадлежности к маленькому миру используют коэффициенты ω и σ.

Коэффициент ω находится в диапазоне от -1 до 1. Значения, близкие к 0, означают, что G имеет характеристики маленького мира. Значения, близкие к -1, означают, что G имеет форму решетки, тогда как значения, близкие к 1, означают, что G является случайным графом. Формула для нахождения коэффициента ω:

$$\omega = \frac{L_l}{L} - \frac{C}{C_l}$$

L_l – длина пути для эквивалентной случайной решетчатой сети, C_l – коэффициент кластеризации для эквивалентной случайной решетчатой сети.

Эквивалентная случайная решетчатая сеть – граф построенный по модели Ваттса-Строгаца, с параметрами: n = количеству вершин в исследуемом графе, p = 0.5, k = средней степени вершин в исследуемом графе.

Если коэффициент σ больше 1, то граф классифицируется как маленький мир. Формула для нахождения коэффициента σ:

$$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}$$

L_r – длина пути для эквивалентного случайного графа, C_r – коэффициент кластеризации для эквивалентного случайного графа.

Эквивалентный случайный граф – граф построенный по модели Эрдеша-Реньи, с параметрами: n = количеству вершин в исследуемом графе, p = 0.5.

4 Граф интересов и его применение

Граф интересов – это представление интересов конкретного человека, полученное на основе анализа его активности в социальных сетях. Вершинами такого графа являются увлечения этого человека. Также вершиной может быть и профиль человека в социальной сети. Ребра графа показывают взаимоотношения между вершинами этого графа.

5 Модели построения случайных графов

В математике случайный граф – это общий термин для обозначения вероятностного распределения графов. Случайные графы описываются как распределение вероятности.

Существует большое число различных моделей, генерирующих случайные графы близкие по свойствам к реальным сетям. Их можно разделить по генерируемым ими графам на несколько основных классов:

- Модели случайных графов (модель Эрдеша-Реньи).
- Простейшие модели безмасштабных сетей (модель Барабаши-Альберта, модель Боллобаша-Риордана, модель копирования и др.).
- Более гибкие модели безмасштабных сетей (модель Чунг-Лу, модель Янсона-Лучака).

5.1 Модель Эрдеша-Реньи

Зафиксируем число $p \in [0,1]$, которое равно вероятности каждого отдельного ребра независимо от остальных ребер в случайном графе. Это схема

Бернулли, в которой проводятся C_n^2 испытаний. При успехе в каком-то из испытаний, соответствующее ребро добавляется в случайный граф – иначе не добавляется.

5.2 Модель Барабаши-Альберта

Барабаши и Альберт предложили свой взгляд на процесс формирования Интернета. Пусть в каждый момент времени появляется какой-то новый сайт, и он ставит фиксированное количество ссылок на своих предшественников. Этот сайт будет ссылаться на тех, кто популярен. Предполагаем, что вероятность, с которой новый сайт поставит ссылку на один из прежних сайтов, пропорциональна числу уже имевшихся на тот сайт ссылок.

Модели случайных графов, основанные на этой идее, называются моделями предпочтительного присоединения.

5.3 Расширенная модель Барабаши-Альберта

Расширенная модель Барабаши-Альберта немного отличается. Начнем с m_0 изолированных вершин, и на каждом временном шаге мы выполняем одно из следующих трех действий:

1) С вероятностью p к графу добавляются m ($m \le m_0$) новых дуг. Для этого мы случайным образом выбираем вершину в качестве начальной точки. Новая дуга описывает, например, что к сайту добавляется новая ссылка. Другой конец дуги выбирается с вероятностью

$$\Pi(k_x) = \frac{k_x+1}{\sum_y(k_y+1)} \quad (1)$$

Новые дуги преимущественно указывают на популярные вершины. Этот процесс повторяется m раз.

2) С вероятностью q, переопределяем m существующих дуг. Для этого случайным образом выбираем вершину x и связанную с ним дугу l_{xy}. Затем мы удаляем эту дугу и заменяем ее новой дугой $l_{xy'}$ которая соединяет x с вершиной y', с вероятностью $\Pi(k_x')$ по формуле (1). Этот процесс повторяется m раз.

3) С вероятностью $(1-p-q)$ к графу добавляются новые вершины. Новая вершина имеет m новых дуг, которые с вероятностью $\Pi(k_x)$ присоединены к вершинам x, которые уже есть в графе.

Если $p = q = 0$, модель ведет себя так же, как модель Барабаши-Альберта.

5.4 Модель Ваттса-Строгаца

Модель Ваттса-Строгаца – это модель генерации случайного графа, который имеет высокий коэффициент кластеризации вершин и довольно небольшую среднюю длину пути.

Эта модель, что-то среднее между регулярной кольцевой цепочкой и графом, построенным по модели Эрдеша-Реньи. Модель Ваттса-Строгаца позволяет создавать сети с большим числом функциональных структур, в отличие от модели Эрдеша-Реньи.

5.5 Модель Холма и Кима

Одна из моделей, была изучена П. Холмом и Б. Дж. Кимом в 2002 г. Она основана на модели Барабаши-Альберта, но включает «этап формирования треугольников». Развитие сети в самой модели выполняется итеративно по следующим правилам:

1) (Рост) На каждой итерации t добавляется одна новая вершина t;

2) На каждой итерации к сети добавляется m ребер, и каждый новая вершина t соединяется с m существующим вершинам по следующим правилам:

a) (Предпочтительное присоединение) Новая вершина связывается с одной из существующих вершин i сети с вероятностью, пропорциональной k_i.

b) Каждое из оставшихся $m-1$ ребер новой вершины t присоединяется следующим образом:

- (b1) (Формирование треугольника) С вероятностью p ребро соединяется с произвольной соседней вершиной для вершины i, выбранного на шаге 2(a).

- (b2) С вероятностью $1-p$ ребро привязано к одной из вершин s сети (не обязательно рядом с вершиной i) с вероятностью, пропорциональной степени этой вершины k_s .

6 Инструменты визуализации графов социальных сетей

6.1 Приложение «Интерактивный граф друзей»

Приложение «Интерактивный граф друзей» в сети ВКонтакте, которое дает нам возможность просмотреть граф друзей. Этот граф можно масштабировать, также можно менять размер самого окна приложения. Особенностью является то, что можно таким же образом построить граф любого из друзей пользователя.

6.2 MultiNet

MultiNet – это программа, которая позволяет анализировать исследовательские данные, представленные в виде социальных или иных сетей. Программа позволяет получить как численный анализ исходных данных, представляющих некий граф, так и визуальное представление сети.

6.3 Визуализация друзей ВКонтакте YASIV

«Визуализация друзей ВКонтакте YASIV» – это веб-приложение, которое строит граф друзей пользователя в социальной сети ВКонтакте. В основе приложения лежит JavaScript-библиотека для укладки графов, сами графы отображаются с помощью WebGL.

7 Программа для генерации случайных графов

Программа написана на языке Python с использованием библиотек:

- Networkx – библиотека, для анализа характеристик исследуемого графа (количество треугольников и т.д.).
- Matplotlib – библиотека, для визуализации графа.

Модель Эрдеша-Реньи. Первый пример: n = 110 и p = 0.2. Второй пример: n = 230 и p = 0.1.

Модель Барабаши-Альберта. Первый пример: n = 110, m = 6. Второй пример: n = 230, m = 8.

Расширенная модель Барабаши-Альберта. Первый пример: n = 110, m = 6, p = 0.3, q = 0.4. Второй пример: n = 230, m = 6, p = 0.2, q = 0.5.

Модель Ваттса-Строгаца. Первый пример: n = 110, k = 20, p = 0.6. Второй пример: n = 230, k = 22, p = 0.7.

Модель Холма и Кима. Первый пример: n = 110, m = 12, p = 0.6. Второй пример: n = 230, m = 12, p = 0.8.

8 Программы для анализа социальных графов

Программа для анализа графа друзей из сети ВКонтакте написана на языке Python с использованием библиотек: Vk_api, Networkx, Matplotlib.

Первый пример работы для графа с количеством вершин = 110, а другой пример для графа с количеством вершин = 318.

Программа для анализа графа актеров написана на языке Python с использованием библиотек, IMDbPy, Networkx, Matplotlib.

Построим граф первых 50 актеров (с индексами 0000001 – 0000050) по версии IMDb. Для чистоты эксперимента так же, построим другой граф актеров (с индексами 0001000 – 0001050).

После анализа получившихся графов можно заметить, что случайные графы, построенные по расширенной модели Барабаши-Альберта и модели Холма и Кима имеют похожие характеристики с этими графами

Можно сделать вывод, что эти две модели, более правильно отражают поведение социальных графов. Используя их можно искать какие-то закономерности в социальных графах.

ЗАКЛЮЧЕНИЕ

Несомненно, сейчас трудно представить современный мир без социальных сетей. Анализировать социальные сети с каждым годом становится все легче, так как появляются новые инструменты для исследования.

В данной работе были рассмотрены важные характеристики для социальных графов. Так же было разработано приложение, написанное на языке Python, в среде разработки IDLE, для визуализации и исследования случайных графов. После этого было проведено сравнение разных моделей случайных графов с реальными социальными графами. Для анализа социальных графов было разработано 2 приложения. Первое для анализа графа друзей из сети ВКонтакте. Второе для анализа графа актеров по версии IMDb.

В результате можно сделать следующие выводы. Случайные графы, построенные по моделям: Эрдеша-Реньи, Ваттса-Строгаца, Барабаши-Альберта, оказались недостаточно близкими к реальным социальным графам. Расширенная модель Барабаши-Альберта и модель Холма и Кима показали очень хорошие результаты. Их характеристики очень близки к реальным социальным графам. Эти модели можно использовать для анализа каких-либо закономерностей в реальных социальных графах.