Структура сложных сетей

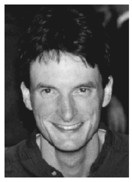
Лекция N 4 курса
"Алгоритмы для Интернета"

Юрий Лифшиц

ПОМИ РАН - СПбГУ ИТМО

Осень 2006

M.E.J. Newman



Six degrees of separation - - это гипотеза, утверждающая что от любой человек на Земле связан с любым другим по цепочке знакомств содержащей не более пяти промежуточных звеньев.

Karinthy Frigyes [1929]

Как изучать сложные сети?

Подход современной науки:

1. Собрать максимум информации о сетях реального мира
2. Разработать научный язык (терминологию) для описания свойств сетей
3. Разработать математические модели сетей имеющие сходные характеристики с реальными сетями
4. Описать процессы происходящие на сетях в математических терминах
5. Разработать алгоритмы предсказания и управления процессами на сетях

План лекции

1. Сети вокруг нас
2. Вспоминаем теорию графов
3. Математические модели сетей
   - Случайные графы
   - Модели роста сетей

А что такое сеть?

Минимальный набор:

- Вершины
- Ребра

Какие могут быть дополнительные параметры?

Дополнительные характеристики:

- Оиентированные/неориентированные ребра
- Двудольные сети
- Числовые/качественные характеристики вершин
- Числовые/качественные характеристики ребер
- Динамика сети (добавление/исчезновение новых вершин/ребер)
- Ацикличность

Часть I

Какие сети есть в реальном мире?

Как их классифицировать?

Социальные сети

Отношения между людьми:

- Сеть дружбы
- Сеть соавторства ученых
- Сеть сексуальных контактов
- Браки между кланами
- Бизнес-отношения
- Совместное появление киноактеров в фильмах
- Телефонные звонки, email'ы, сеть icq-контактов

Пример: число Эрдеша

- Поль Эрдеш имеет индекс ноль
- Его соавторы имеют индекс 1
- Соавторы соавторов имеют индекс 2, и т.д.

Гипотеза: не бывает определенного значения индекса Эрдеша больше 15.
Факт: у 98% математиков индекс Эрдеша не более 7.

Например: Paul Ërdos – Richard K. Guy – Юрий Матиясевич – Юрий Лифшиц

Посетите `http://www.oakland.edu/enp/`!

Информационные сети

Отношения между информационными объектами:

- Цитирования в научных статьях
- Граф ссылок WWW
- Цитирование в патентах
- Peer-to-peer сети
- Совместное употребление слов в текстах

Технологические сети

"Физические" связи в нашем трехмерном мире:

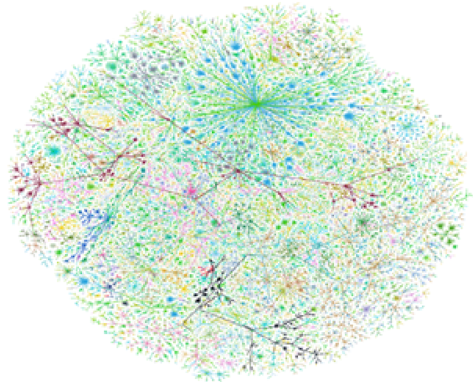
- Интернет как сеть компьютеров
- Национальные электросети
- Телефонные линии, почтовые службы доставки
- Поезда, самолеты, автобусы

Биологические сети

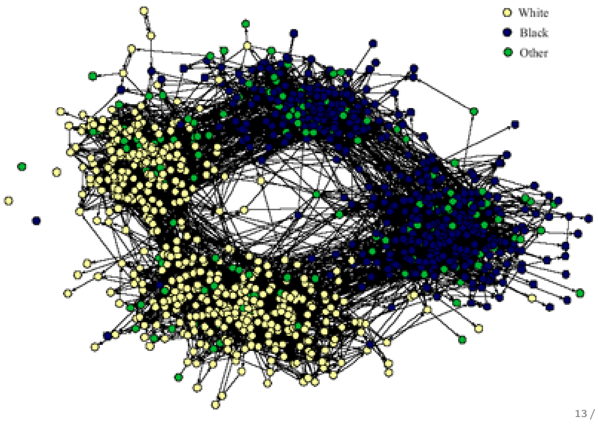
Сети внутри и между животными, растениями, людьми:

- Сеть нейронов в мозге
- Реакции между протеинами
- Кровеносные сосуды
- Реки, озера, океаны
- Граф "жертва-хищник"

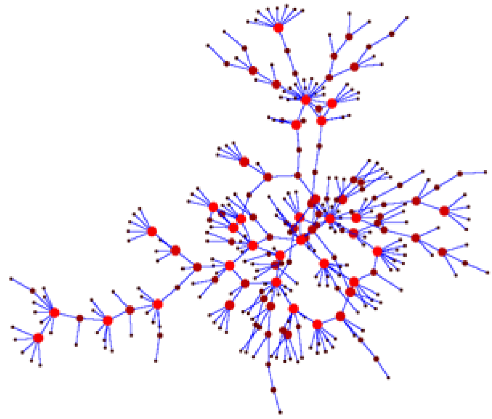
Пример: Интернет [Hal Burch, Bill Cheswick / Lumeta corp.]



Пример: дружба в школе [James Moody]



Пример: сексуальные отношения [Potterat et al.]



Часть II

Какие свойства есть у графов?

По каким формулам можно численно выразить эти свойства?

Эффект "маленького мира"

Как выразить "six degrees of separation" в числах?

Естественный ответ: вычислим среднее кратчайшее расстояние

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$$

Что делать, если больше одной компоненты связности?

Первое решение: взять среднее по всем связанным парам. Второе:

$$l^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

Транзитивность/кластеризация

Неформально, сеть имеет высокую кластеризацию, если **твои друзья дружат между собой**

Формализация через треугольники:

$$C = \frac{6 \times \text{число треугольников}}{\text{число путей длины 2}}$$

Коэффициент кластеризации отдельной вершины:

$$C_i = \frac{\text{число треугольников с вершиной} i}{\text{число пар друзей} i}$$

Тогда можно пересчитать общий коэффициент кластеризации:

$$C' = \frac{1}{n} \sum_i C_i$$

Распределение степеней

Эмпирическое распределение (для данного графа): Пусть d_i — степень i-ой вершины. Определим

$$p_k = \frac{1}{n} \#\{i | d_i = k\}.$$

Последовательность $p_1, p_2, \ldots$ называется (эмпирическим) распределением степеней.

Получилась "шумная" функция. Как ее сгладить?

Ответ: усреднять по возрастающим интервалам. Например, для каждого $2^t \le k < 2^{t+1}$ определить

$$p'_k = \frac{1}{2^t} \sum_{j=2^t}^{2^{t+1}-1} p_j$$

Распределение степеней II

Распределение Пуассона:

$$p_k = \frac{z^k e^{-z}}{k!}$$

Распределение по степенному закону (power law):

$$p_k = \frac{k^{-\alpha}}{\zeta(\alpha)}$$

Корреляции

Пусть вершины сети бывают нескольких типов. Мы будем говорить, что наблюдается эффект **assortative mixing**, если ребра "чаще" соединяют вершины внутри типа, чем между разными типами.

Пусть e_{ij} — доля ребер между типами i и j в множестве всех ребер сети. Тогда assortative mixing можно измерить таким коэффициентом Q:

$$P(j|i) = e_{ij} / \sum_k e_{ik} \qquad Q = \frac{\sum_i P(i|i) \; - \; 1}{N - 1}$$

Значения 0 и 1 как раз соответствуют полному наличию/отсутствию эффекта

Корреляции II

Возьмем частный случай assortative mixing: **зависимость расположения ребер от одинаковости степени концов**

Пишут ли авторы с большим числом соавторов статьи в основном с себе подобными или как раз наоборот?
Цитируют ли самые знаменитые сайты в основном другие известные сайты или, напротив, в основном сообщают о не столь популярных?
Тот же вопрос про поезда, актеров, нейроны и сеть хищник-жертва.

Ответ: положительная корреляция присутствует для сетей соавторства, и актеров. Отрицательная для Internet'a, поездов, web-сайтов, нейронов и сети хищник-жертва.

Корреляции III

Как измерить наличие assortativity для степеней?

Ответ 1: построить график "средняя степень друзей у вершин степени k". Посмотреть на убывание / возрастание.

Ответ 2: ходить на лекции по статистике. Сосчитать коэффициент Пирсона корреляции степеней на концах всех ребер в сети.

Много других свойств

- Betweenness — как часто данная вершина лежит на кратчайших путях между другими вершинами
- Повторения небольших подграфов — motif analysis
- Уязвимость связности при выкидывании части вершин
- Размер наибольшей компоненты

Часть III

Как породить графы, похожие по своим характеристикам на сети из реальной жизни?

Какие параметры могут быть у этих моделей?

Как подобрать параметры модели, чтобы она стала максимально похожа на реальную сеть?

Случайные пуассоновские графы

Первая модель случайного графа, которая приходит в голову?

Solomonoff/Rapoport and Erdös/Rényi:

- Фиксируем параметр p
- Независимо для каждой пары вершин с вероятностью p проводим ребро

Erdös/Rényi:

- Фиксируем общее число ребер m
- Берем случайный граф с n вершинами и m ребрами

Конфигурационная модель

- Фиксируем распределение степеней D
- Выбираем n чисел согласно распределению D
- Рисуем "хвосты ребер" в соответствии с полученными числами $d_1, \ldots, d_n$
- Случайным образом разбиваем хвосты на пары и соединяем их в ребра

Улучшения конфигурационной модели

Можем внести следующие изменения в модель:

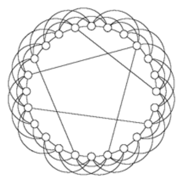
- С помощью распределений входящих и исходящих степеней построить ориентированный граф
- Обобщение на двудольные графы
- Марковские случайные графы: блуждание по "графу графов" с вероятностями, пропорциональными желаемым свойствам.

Модель "маленького мира"

Строим "географическую" сеть:

- Располагаем n вершин равномерно по кругу
- Соединяем каждую вершину со всеми, находящимися на расстоянии не более k
- Для каждого ребра с вероятностью p один из его концов заменяем на случайную вершину

Картинка из работ D.J.Watts:



Модель Прайса

"Богатые становятся богаче!"

- Начинаем с одной вершины
- Параметр m задает среднее число исходящих ребер
- При рождении новой вершины выбираем ее исходящую степень и для каждого ребра его конец определяются с **вероятностью, пропорциональной входящим степеням старых вершин**.

Мотивирующий пример: граф цитирований в научных статьях.

Модель Прайса II

- Параметр m задает среднее число исходящих ребер
- При рождении новой вершины выбираем ее исходящую степень и для каждого ребра его конец определяются с **вероятностью, пропорциональной входящим степеням старых вершин**.

Выразим вероятность стать адресатом нового ребра. **Фиксированная** вершина степени k будет процитирована с вероятностью:

$$\frac{1}{N}\frac{k+1}{\sum_i (i+1)p_i} = \frac{1}{N}\frac{k+1}{m+1}$$

Расширения базовых моделей роста

В модель Прайса можно вносить следующие изменения:

- Строить неориентированный граф, вероятность стать адресатом будет пропорциональна полной степени
- Изменить добавочный коэффициент, сделать вероятности пропорциональными $k + k_0$
- При рождении ввести случайный коэффициент привлекательности η_i, вероятности проведения ребра будет пропорциональны $\eta_i k_i$ или $\eta_i + k_i$
- Рассмотреть нелинейные вероятности, пропорциональные k^γ

Задачи

Найти знак корреляции степеней в модели Прайса. Там assortativity или dissortativity?

Найти знак корреляции числа френдов в Livejournal. Там assortativity или dissortativity?

Найти номер Эрдеша для Анатолия Абрамовича Шалыто

Главные моменты

Сегодня мы узнали:

- Реальные сети: биологические, социальные, информационные, технологические
- Свойства сетей: связность, кластеризация, распределение степеней, корреляции базовых свойств между собой
- Модели: Пуассона, конфигурационная, Прайса, Small World

Вопросы?

Источники

Страница курса http://logic.pdmi.ras.ru/~yura/internet.html

Использованные материалы:

- M.E.J. Newman
  The Structure and Function of Complex Networks
  http://www.santafe.edu/files/gems/paleofoodwebs/Newman2003SIAM.pdf
- Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener
  Graph structure of the Web
  http://www.people.cornell.edu/pages/dc288/Paper1.pdf