

## Анализ гиперссылок в сети Веб: модели, подходы и алгоритмы

### Лекция 1. Моделирование сети WWW

RUSSIR-2007. Сычев А.В.

## Моделирование сети WWW

- WWW = *World Wide Web*.
- К какому классу сетей отнести?
  - ☑ WWW – социальная сеть.
  - ☑ WWW – информационная сеть (сеть знаний).
- Для последнего десятилетия характерно смещение внимания исследователей от анализа небольших графов и свойств отдельных вершин или ребер к масштабным исследованиям статистических свойств графов.
- Это изменение обусловило переход от визуальных методов исследования (при небольших размерах графов) к альтернативным методам, заменяющим человеческий глаз как аналитический инструмент, оказавшийся неэффективным для графов размером свыше миллионов вершин.

RUSSIR-2007. Сычев А.В.

## WWW как социальная сеть

↪ *Социальная сеть* – множество людей или групп людей с заданным шаблоном взаимодействия между собой.

- В данном качестве исследуются такие свойства этой сети как *централизация* (какие индивидуумы наиболее тесно связаны с другими или имеют наибольшее влияние) и *связность* (в какой степени и как связаны между собой индивидуумы через сеть).

RUSSIR-2007. Сычев А.В.

## WWW как информационная сеть

- ☑ *WWW* – сеть из веб-страниц, содержащих информацию и связанных между собой гиперссылками (не путать с Интернетом!).
- ☑ Данная сеть активно исследуется, начиная с момента ее формирования в начале 90-х годов.

RUSSIR-2007. Сычев А.В.

## WWW как веб-граф

- Сеть гипертекстовых документов может быть представлена в виде ориентированного графа  $G(V, E)$ , содержащего узлы (веб-страницы)  $V$ , которые связаны между собой направленными ребрами (гиперссылками)  $E$ .

RUSSIR-2007. Сычев А.В.

## Терминология

- *Вершина* – фундаментальная единица сети. Это может быть сайт, узел, агент и т.п.
- *Ребро* – линия соединяющая две вершины.
- *Направленный/ненаправленный*. Ребро считается направленным, если оно отображает связь между вершинами только в одном направлении. Если же ребро показывает двухстороннюю связь между вершинами, то оно является ненаправленным.
- *Степень* – количество ребер, связанных с вершиной.

RUSSIR-2007. Сычев А.В.

## Терминология

- *Компонента* – множество вершин достижимых между собой по путям вдоль ребер графа.
- *Геодезический путь* – кратчайшее расстояние в сети между двумя вершинами.
- *Диаметр* – длина (число ребер) наибольшего геодезического пути между двумя вершинами.

RUSSIR-2007. Сычев А.В.

## Феномен малого мира

- Большинство пар вершин в большинстве сетей оказываются связанными посредством коротких путей.
- Если количество вершин, находящихся на расстоянии  $r$  от заданной растет экспоненциально от  $r$  (что справедливо для большинства сетей), то значение среднего геодезического (фактически кратчайшего) расстояния  $l$  между парами вершин в сети растет всего лишь как  $\log n$  ( $n$  – число вершин в графе)

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

RUSSIR-2007. Сычев А.В.

## Феномен малого мира

- Более точное определение дается следующим образом:

*в сети наблюдается данный феномен, если значение  $l$  растет логарифмически (или более медленно) в зависимости от размера сети.*

RUSSIR-2007. Сычев А.В.

## Эксперимент Стэнли Милгрэма (1967)

- Цель эксперимента – проверить насколько “мир тесен”.
- Схема эксперимента:
  - ☞ участник эксперимента должен отправить письмо определенному биржевому брокеру в Бостоне. Если участник с ним не знаком, то
  - ☞ письмо должно быть отправлено приятелю участника, который по его мнению может быть знаком с этим брокером. В письме необходимо указать свои имя и адрес для отслеживания цепочки, а также для предотвращения заикливания.
  - ☞ На отдельной карточке участник должен написать свое имя, а также имена тех, от кого он получил письмо и кому он отправил (для отслеживания пути писем, не дошедших до адресата).
- Участвовало 150 добровольцев.
- Длина цепочки между первым отправителем и бостонским брокером варьировалась от 2 до 10 человек, в среднем – 5.

RUSSIR-2007. Сычев А.В.

## Эксперимент Стэнли Милгрэма

- Уоттсом и Строгачем был повторен эксперимент с участием 61168 добровольцев из 166 стран.
- Средняя длина цепочки совпала с той, что была получена Милгрэмом.
- Но из 24 тысяч писем дошли до адресата только 384.

RUSSIR-2007. Сычев А.В.

## Диаметр веб-графа

- Исследование, проведенное в 1999 году [2], показало, что среднее расстояние между вершинами веб-графа, т.е. диаметр  $d$ , при размере  $N = 8 \times 10^8$  получилось равным 19.
- Была получена зависимость:  $d = 0.35 + 2.06 * \log(N)$ .

RUSSIR-2007. Сычев А.В.

## Структура веб-графа – “бабочка”

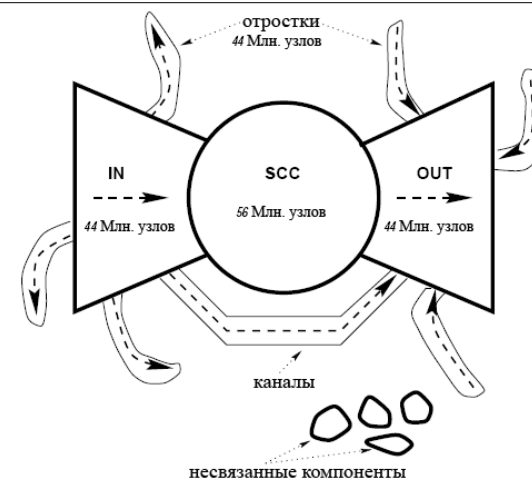
• Исследование, проведенное в 1999 году на веб-графе, содержащем 200 миллионов страниц и 1.5 миллиарда гиперссылок между ними [3], выявило более сложную структуру этого графа:

*большинство пар вершин не имеют связывающих их цепочек вообще, также имеется значительно число пар вершин, между которыми может быть установлена связь посредством цепочек, содержащих сотни промежуточных звеньев.*

• было обнаружено центральное сильной связанное ядро (SCC), подграф, содержащий только направленные ссылки на ядро (IN), подграф, содержащий только направленные ссылки из ядра (OUT), относительно изолированные “отростки”, связанные с одной из трех крупных компонент, названных выше. Имелись также полностью изолированные компоненты, не имевшие связей с названными выше компонентами.

RUSSIR-2007. Сычев А.В.

## Структура веб-графа – “бабочка”



RUSSIR-2007. Сычев А.В.

## Структура веб-графа – “бабочка”

- Было также установлено, что диаметр компоненты SCC оказался равен, по меньшей мере, 28, в то время как диаметр всего графа превысил 500.
- Для двух случайно выбранных из веб-графа страниц вероятность существования пути между ними составила всего 24%. При этом если направленный путь существует, то его длина составляет примерно 16 гиперссылок. Для ненаправленных путей средняя длина составила 6 ребер.

RUSSIR-2007. Сычев А.В.

## Степенные распределения

- Обозначим долю вершин в сети, имеющих степень  $k$ , как  $p_k$ .  
Эту величину можно рассматривать как вероятность того, что случайно выбранная вершина в сети будет иметь степень  $k$ .
- Также можно ввести кумулятивную функцию распределения степеней вершин:

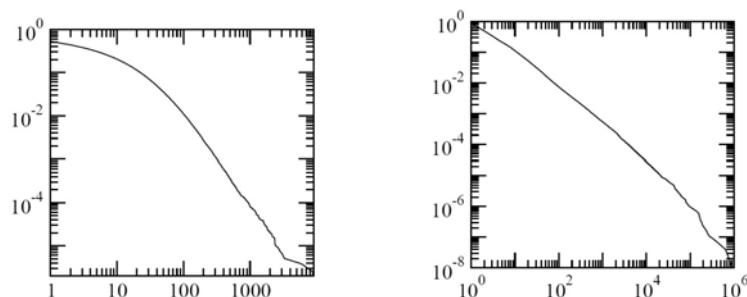
$$P_k = \sum_{k'=k}^{\infty} p_{k'}$$

которая фактически является вероятностью того, что степень вершины больше или равна  $k$

RUSSIR-2007. Сычев А.В.

## Степенные распределения

Ниже приведены графики зависимости  $P_k$  от  $k$  для сетей цитирования (слева) и для WWW (справа):



RUSSIR-2007. Сычев А.В.

## Степенные распределения

- Исследования показали, что гиперссылки в Веб не подчиняются модели независимой случайной генерации. В первом приближении вероятность появления новой ссылки у страницы подчиняется степенному закону:

$$\Pr(ucx = k) \propto \frac{1}{k^{a_{ucx}}} \quad \Pr(vx = k) \propto \frac{1}{k^{a_{vx}}}$$

$$a_{ucx} = 2.45, a_{vx} = 2.1.$$

То есть:  $p_k \sim k^{-\alpha}$

- Отсюда получаем:  $P_k \sim \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-(\alpha-1)}$

RUSSIR-2007. Сычев А.В.

## Степенные распределения

- Сети, в которых наблюдается подобная зависимость обычно называют *scale-free сетями* (безмасштабными сетями), т.е. для них выполняется следующее условие:

$$f(a \cdot x) = b \cdot f(x)$$

- Важной особенностью таких сетей, является то, что хотя степенное распределение наблюдается во всей сети в целом, подсети данной сети могут иметь другую форму распределения. Например, в сети WWW в целом наблюдается степенное распределение, однако для доменов внутри этой сети характерно унимодальное распределение.

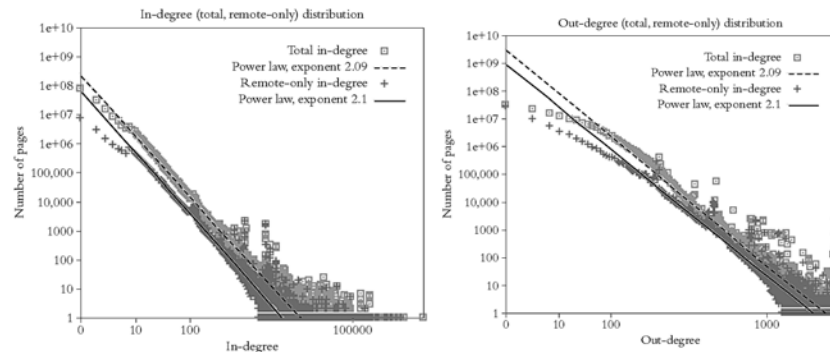
RUSSIR-2007. Сычев А.В.

## Модель предпочтительного прикрепления

- Вновь возникающий узел веб-графа устанавливает соединения с уже существующими узлами не равновероятно, но с большей вероятностью с узлами, имеющими большое количество связей. Количество таких соединений является константой.
- “Победителям достается все”.

RUSSIR-2007. Сычев А.В.

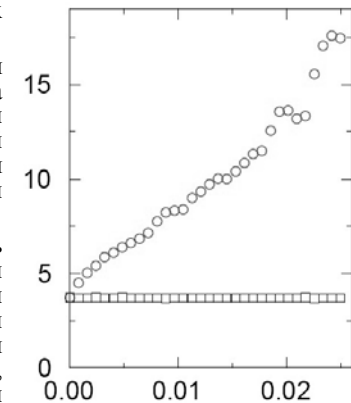
## Модель предпочтительного прикрепления



RUSSIR-2007. Сычев А.В.

## Стойкость сети

- Со степенными распределениями связана стойкость сети к удалению отдельных узлов.
- Если узлы удаляются из сети, то средняя длина пути между вершинами графа увеличивается. В пределе получается множество не связанных между собой узлов. При этом различные сети отличаются между собой по степени устойчивости к удалению узлов.
- На рисунке представлена зависимость среднего расстояния между вершинами графа от доли удаляемых вершин (для сети Интернет). Квадратными маркерами обозначены точки со случайным выбором вершин для удаления, круглыми — в порядке убывания степеней вершин (начиная с самой высокой).



RUSSIR-2007. Сычев А.В.

## Динамика веб-графа во времени

В работе [4] была рассмотрена эволюция графов, представляющих реальные сети, во времени. При этом были обнаружены поразительные на первый взгляд феномены:

- большинство из наблюдавшихся графов уплотнялись со временем, причем скорость роста ребер превышала более чем линейно скорость роста узлов (порядка  $O(\log n)$  или  $O(\log(\log n))$ ).
- среднее расстояние между узлами со временем укорачивается, что противоречит устоявшемуся ранее мнению, что данный параметр должен увеличиваться со временем как  $O(\log n)$  или  $O(\log(\log n))$ .

RUSSIR-2007. Сычев А.В.

## Динамика веб-графа во времени

Ранее рассматривавшиеся модели исходили из следующих двух допущений:

- *Допущение о постоянстве степени. вершин.* Средняя степень вершины в сети остается неизменной во времени (или, другими словами, количество ребер зависит линейно от количества вершин).
- *Допущение о медленном росте диаметра.* Диаметр является медленно растущей функцией от размера сети (графы “малого мира”).

RUSSIR-2007. Сычев А.В.

## Динамика веб-графа во времени

Авторами статьи было предложено пересмотреть эти умозрительные принципы следующим образом:

- *Эмпирическое наблюдение: степенные законы уплотнения*: сети со временем уплотняются, что дает увеличение средней степени вершин по степенному закону.
- *Эмпирическое наблюдение: укорачивание диаметров*: эффективный диаметр во многих случаях уменьшается по мере роста сети.

RUSSIR-2007. Сычев А.В.

## Динамика веб-графа во времени

В процессе эволюции графа наблюдается следующее соотношение:

$$e(t) \propto n(t)^a$$

где  $e(t)$  и  $n(t)$  описывают закон изменения количества ребер и вершин в графе во времени, константа  $a$  принимает значения из интервала от 1 до 2.

RUSSIR-2007. Сычев А.В.

## Литература

1. M.E.J. Newman “*The structure and function of complex networks*” // SIAM Review, 45: 167-256, 2003.
2. R.Albert, H.Jeong, A.-L. Barabasi “*Diameter of the world-wide web*”. Nature 401, 130-131 (1999).
3. A.Broder, R.Kumar, F.Maghoul, P.Raghavan, S.Rajagopalan, R.Stata, A.Tomkins and J.Wiener ”*Graph structure in the web*”. Computer Networks 33, 309-320 (2000).
4. J.Leskovec, J.Kleinberg, C.Faloutsos “*Graphs over time: densification laws, shrinking diameters and possible explanations*” // KDD’05 August 21-24, 2005. Chicago, Illinois, USA.

RUSSIR-2007. Сычев А.В.