Методы анализа социальных сетей. Структурные характеристики и экспериментальные наблюдения.

Николай Ильич Базенков, к.т.н.

Институт проблем управления им. В.А. Трапезникова
Российской академии наук

23 сентября 2015 г.

Содержание курса

Лекция 1 Базовые понятия

Лекция 2 Структурные свойства социальных сетей

- Компоненты связности
- Диаметр и длина пути
- Кластеризация
- Распределение степеней
- Метрики центральности

Лекция 3 Модели формирования сетей

Лекция 4 Сообщества в социальных сетях

Плотность сети

Плотный граф. Граф $G = (N, E)$, $|N| = n$ называется плотным, если

$$|E| \sim n^2 \tag{1}$$

Разреженный граф. Граф $G = (N, E)$, $|N| = n$ называется разреженным, если

$$|E| \sim n \tag{2}$$

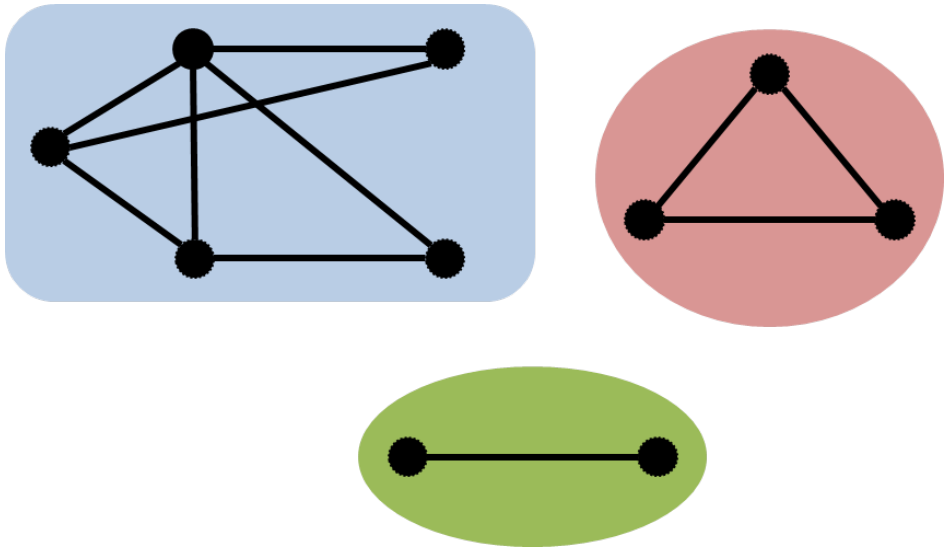
Плотность сложных сетей

Социальные, информационные, биологические и технические крупномасштабные сети, как правило, разреженные

Сеть	C.Elegans	WWW (nd.edu) 1999	WWW 2014	LiveJournal RUS 2011	Facebook 2011
Количество узлов	272	3.26×10^5	1.7×10^9	2.92×10^5	7.21×10^8
Количество ребер	4451	1.47×10^6	64×10^9	6.2×10^6	6.87×10^{10}
$\frac{2\|E\|}{n(n-1)}$	0.06	1.39×10^{-5}	2.21×10^{-8}	7.27×10^{-5}	1.32×10^{-6}

Связность

Компонентой связности неориентированного графа $G = (N, E)$ называется такой максимальный подграф, между каждой парой вершин которого существует путь



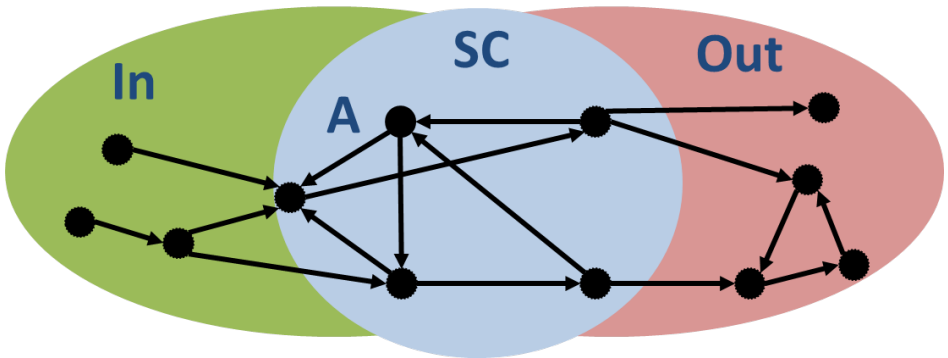
Компоненты связности ориентированных графов

Компонента слабой связности – связна без учета ориентации ребер

Компонента сильной связности – для любой пары вершин A и B существует ориентированный путь от A к B и от B к A

Входящая компонента вершины A – все вершины, из которых достижима A

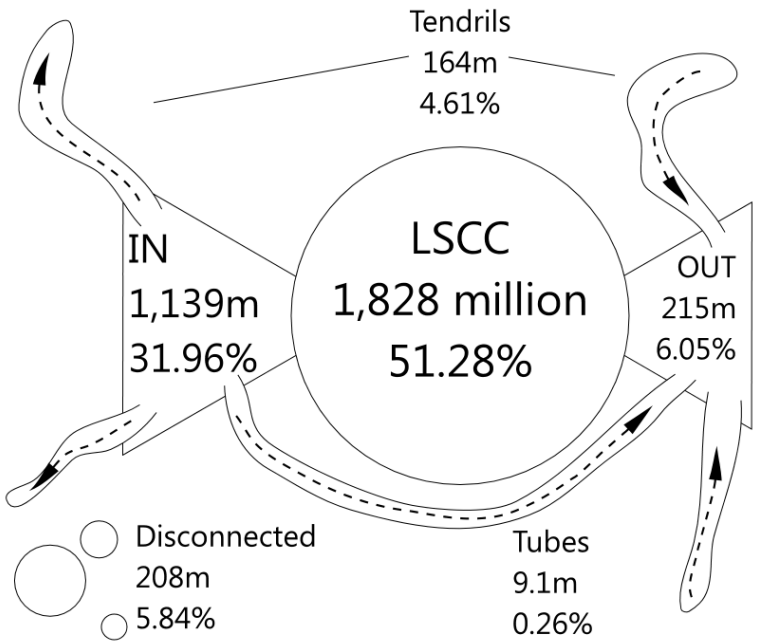
Исходящая компонента вершины A – все вершины, достижимые из A



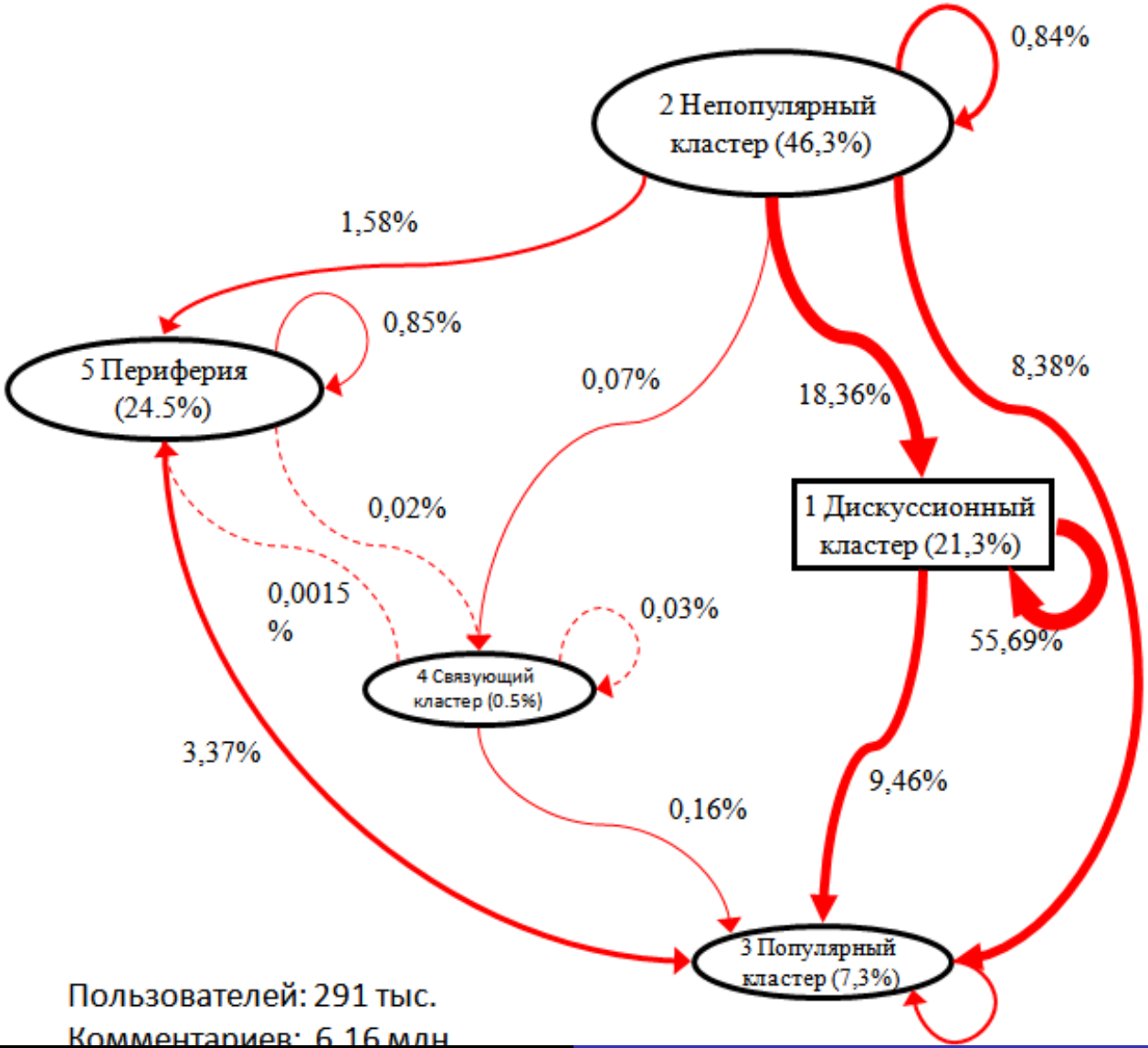
Пример – веб-граф 2014

"Бабочка" (bow-tie) – представление графа в виде наибольшей компоненты сильной связности (LSCC), входящей (IN) и исходящей (OUT)

Web Data Commons `http://webdatacommons.org/hyperlinkgraph/index.html`



Компоненты LiveJournal 2011



Пользователей: 291 тыс.
Комментариев: 6,16 млн

Диаметр графа

Расстояние между вершинами A и B – длина кратчайшего пути от A до B

Диаметр графа – наибольшее расстояние между вершинами, для которых вообще существует путь

Явление “тесного мира”

1. Диаметр и средняя длина пути в сложных сетях крайне малы
2. Считается, что $G = (N, E)$, $|N| = n$

$$diam(G) \sim \log n \tag{3}$$

3. Ориентация либо игнорируется, либо берется компонента сильной связности

Расстояния – LiveJournal 2011

1. Граф комментирования, $|N| = 292 \times 10^3$, $|E| = 6.2 \times 10^6$
2. 81% пар не связаны
3. Диаметр – 20
4. Среднее расстояние – 5.8

Кластерные коэффициенты

Локальный кластерный коэффициент – доля пар соседей узла, которые соединены ребром

$$C_i(G) = \frac{|E_i(G)|}{d_i(d_i - 1)} \tag{4}$$

1. $E_i(G)$ – множество всех ребер между соседями узла i
2. d_i – степень узла i

Средний кластерный коэффициент – усреднение по всем узлам

$$C(G) = \frac{1}{|N|} \sum_{i \in N} C_i(G) \tag{5}$$

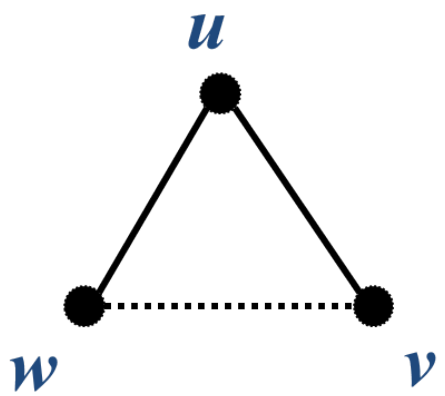
Транзитивность

Глобальный кластерный коэффициент – отношение числа треугольников к числу цепей длины 2

$$T(G) = \frac{3K_3(G)}{P_2(G)} \tag{6}$$

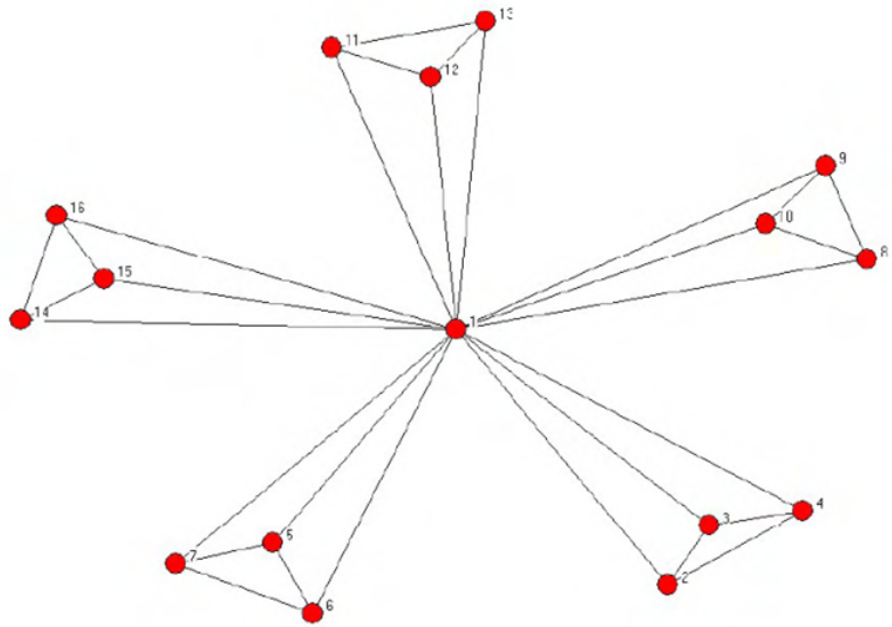
Альтернативная формула

$$T(G) = \frac{\sum_{i \in N} \binom{d_i}{2} C_i(G)}{\sum_{i \in N} \binom{d_i}{2}} \tag{7}$$

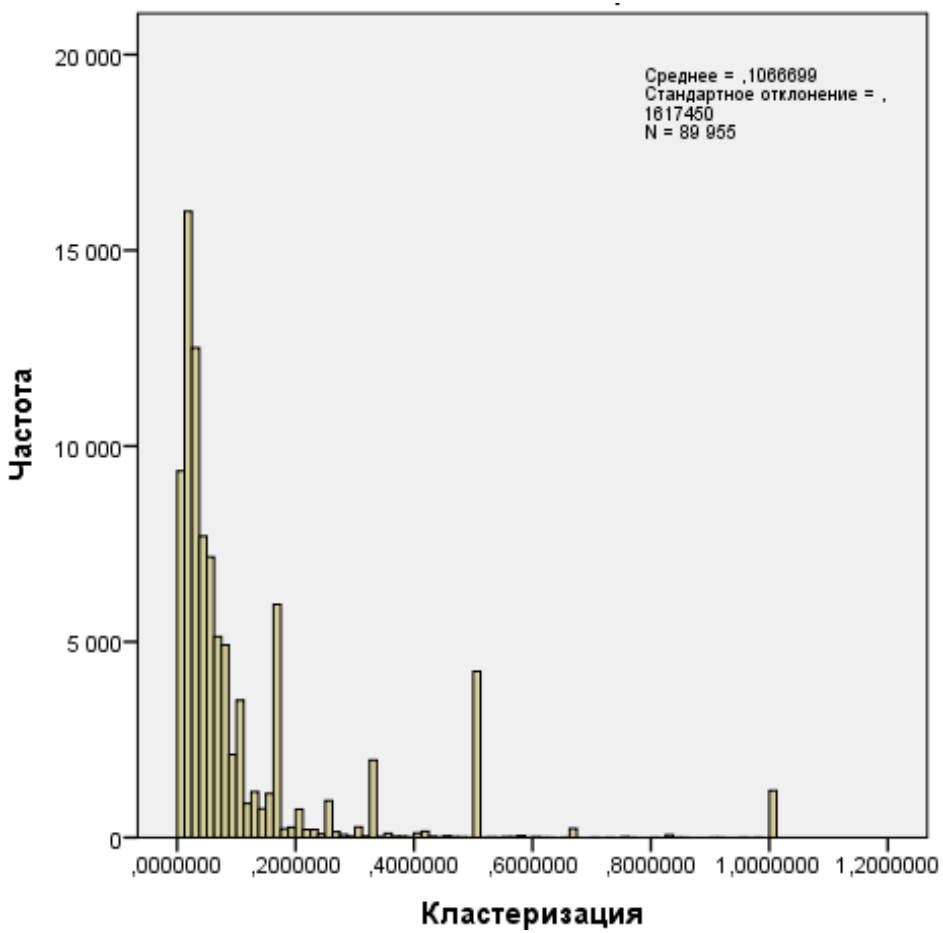


Разница ср. класт. и транзитивности

Пример из Jackson – $C(G) \to 1$, а $T(G) \to 0$ при $n \to \infty$



Кластеризация – LiveJournal 2011



1. Граф комментирования, $|N| = 292 \times 10^3$, $|E| = 6.2 \times 10^6$
2. Средний коэфф класт. – 0.1

Распределение степеней

Неориентированный граф $G = (N, E)$, N – множество вершин, E – множество ребер

Степень – число ребер, связанных с вершиной i

$$d_i(G) = |\{(i,j) : (i,j) \in E\}| \tag{8}$$

Распределение степеней (degree distribution) определяет долю вершин с заданной степенью k

$$P(k) = \frac{|\{i : d_i(G) = k, i \in N\}|}{|N|} \tag{9}$$

Степенной закон (power law)

1. В сложных сетях распределение степеней (degrees) очень часто подчиняется степенному закону (power law):

$$p(k) = ck^{-\gamma} \tag{10}$$

$c > 0$, $\gamma > 1$

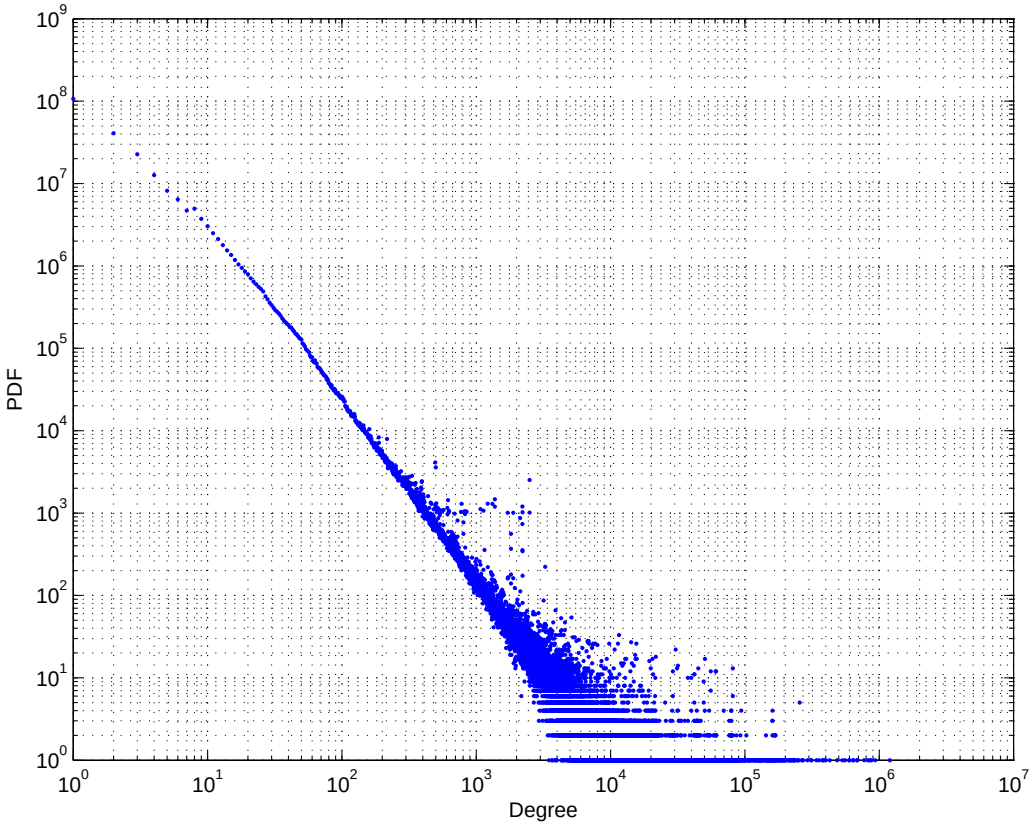
2. Распределение (12) часто называют безмасштабным

$$\frac{p(20)}{p(10)} = \frac{p(200)}{p(100)} = \left(\frac{2}{1}\right)^{-\gamma} \tag{11}$$

3. В логарифмическом масштабе – прямая:

$$\log(p(k)) = \log(c) - \gamma \log(k) \tag{12}$$

Степенной закон в сложных сетях



1. Обычно степенной закон выполняется для $k \geq k^{min} \geq 1$
2. Правый хвост распределения зашумлен

Кумулятивное распределение степеней

Кумулятивная функция распределения (CDF) – меньше подвержена шуму, чем частотная (PDF)

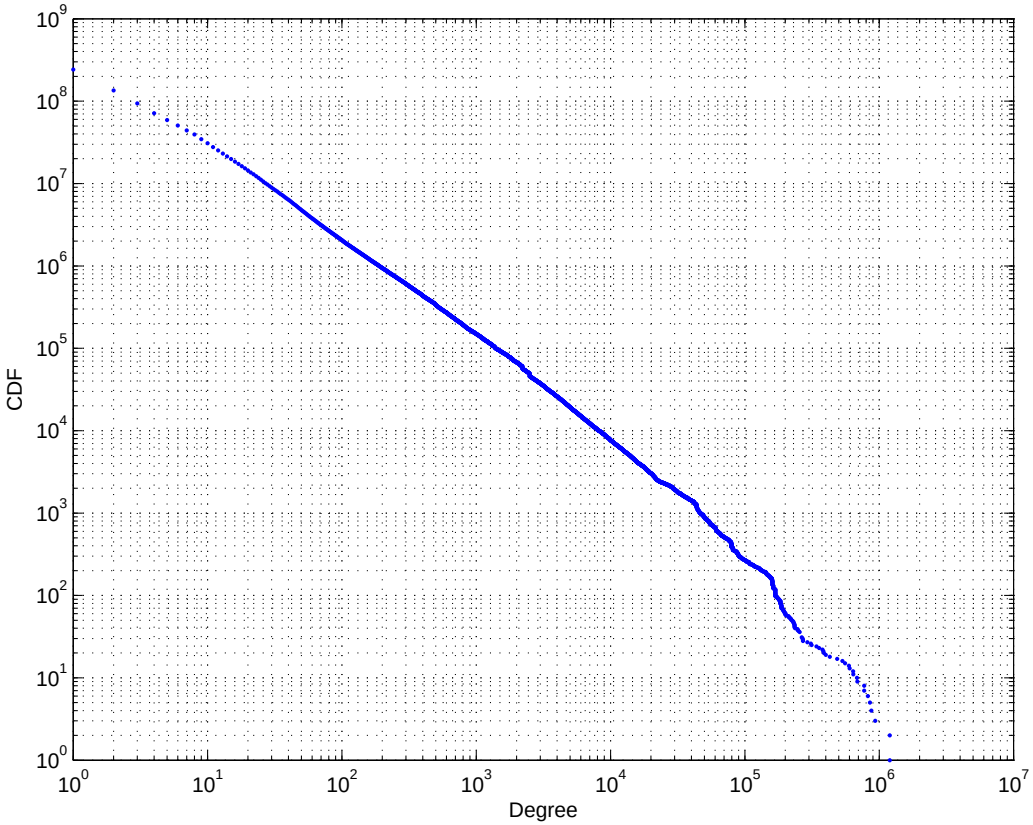
$$P(k) = \sum_{t \geq k}^{\infty} p(t) \tag{13}$$

CDF степенного закона получается следующим образом

$$P(k) = C \sum_{t \geq k}^{\infty} t^{-\gamma} \approx C \int_{k}^{\infty} t^{-\gamma} dt \tag{14}$$

$$= \frac{C}{\gamma - 1} k^{-(\gamma - 1)} \tag{15}$$

Кумулятивное распределение степеней



Кумулятивная функция распределения (CDF):

1. Тоже степенной закон
2. Менее подвержена шуму
3. Нет потери информации как при разбиении на слоты

Оценка параметров степенного закона

Нормализация – выбор константы C

Показатель степени – подбор γ

1. Аппроксимация прямой линией – МНК
2. Метод максимального правдоподобия

Нормализация

Условие нормализации поскольку $p(k)$ – частота

$$\sum_{k=0}^{\infty} p(k) = 1 \tag{16}$$

Начальная степень $k = 0$ не согласуется со степенным законом, поэтому рассматриваем $k \geq k_{min}$

Константа нормализации:

$$C = \frac{1}{\sum_{k=k_{min}}^{\infty} k^{-\gamma}} \approx \frac{1}{\int_{k_{min}}^{\infty} x^{-\gamma} dx} \tag{17}$$

$$= (\gamma - 1) k_{min}^{\gamma - 1}, \tag{18}$$

$$\gamma > 1 \tag{19}$$

Подробнее см. Newman M.E.J. Networks: An Introduction, section 8.4

Линейная регрессия

Шум: – при переходе к логарифмической шкале шум уже не Гауссов. В случае CDF данные уже зависимы

Высокий r^2: нельзя отличить степенное распределение от другого с высоким r^2 (например, лог-нормального)

Нарушается нормализация: для линии регрессии не выполняется условие $\sum_{k \geq k_{min}} k = 1$

Вывод: линейная регрессия – некорректный метод оценки показателя γ (см. Clauset, Shalizi, Newman 2009)

Метод максимального правдоподобия

Считаем, что степенной закон выполняется, начиная с $k_i \geq k_{min}$, всего N точек

Показатель степени:

$$\gamma \approx 1 + N \left[\sum_{i:k_i \geq k_{min}} \ln \frac{k_i}{k_{min} - \frac{1}{2}} \right]^{-1} \tag{20}$$

Ошибка:

$$\sigma = \frac{\gamma - 1}{\sqrt{N}} \tag{21}$$

Точность Формула (22) дает хорошее приближение для $k_{min} \geq 6$. Подробнее см. Clauset, Shalizi, Newman 2009

Код и данные `http://tuvalu.santafe.edu/~aaronc/powerlaws/`

Моменты степенного закона

Общая формула m-й момент распределения:

$$\langle k^m \rangle = \sum_{k=0}^{\infty} k^m p(k) = \sum_{k=0}^{k_{min}-1} k^m p(k) + C \sum_{k=k_{min}}^{\infty} k^{m-\gamma} \tag{22}$$

$$\approx C_0(k_{min}) + C \int_{k_{min}}^{\infty} k^{m-\gamma} dk \tag{23}$$

Аппроксимация m-го момента степенного распределения

$$\langle k^m \rangle \approx C_0(k_{min}) + \frac{C}{m-\gamma+1} \left[k^{m-\gamma+1} \right]_{k_{min}}^{\infty} \tag{24}$$

Условие существования m-го момента:

$$\gamma > m+1 \tag{25}$$

Первый и второй моменты

Мат. ожидание существует при $\gamma > 2$

$$\langle k \rangle = \frac{\gamma - 1}{\gamma - 2} k_{min} \tag{26}$$

Второй центр. момент существует при $\gamma > 3$

$$\langle k^2 \rangle = \frac{\gamma - 1}{\gamma - 3} (k_{min})^2 \tag{27}$$

Заключение

1. Сложные сети разной природы обладают сходными структурными характеристиками
2. Малый диаметр
3. Высокая кластеризация
4. Степенной закон

Структурные оценки влиятельности узлов

Задача – по структуре связей оценить характеристики узлов:

1. Популярность
2. Влиятельность
3. Важность для сети
4. и т.д.

Центральность по степени

1. Просто нормализованная степень:

$$C_i^{deg}(G) = \frac{d_i(G)}{n-1} \tag{28}$$

2. Оценивает популярность, известность, авторитетность
3. Легко поддается манипуляциям
4. Количество друзей в Facebook и LJ, подписчики в Twitter, кол-во ссылок на сайт и т.д.

Близость (closeness)

1. Близость, обратная длине пути

$$Clos_i(G) = \frac{n-1}{\sum_{j \neq i} l(i,j)} \tag{29}$$

$l(i,j)$ – длина кратчайшего пути от i к j

2. Близость с затуханием (Decay centrality)

$$Dec_i(G) = \sum_{j \neq i} \delta^{l(i,j)} \tag{30}$$

$0 < \delta < 1$

3. При $\delta \to 1$ $Dec_i(G) =$ размер компоненты узла
4. При $\delta \ll 1$ $Dec_i(G)$ пропорциональна степени

Промежуточность (betweenness)

1. $P(kj)$ – количество кратчайших путей от k к j
2. $P_i(kj)$ – количество кратчайших путей от k к j, проходящих через i
3. Оценивает значимость узла для распространения информации по сети

$$Btw_i(G) = \sum_{k \neq j \neq i} \frac{P_i(kj)/P(kj)}{(n-1)(n-2)/2} \tag{31}$$

Собственные векторы и центральность

1. Идея – влиятельность узла связана с влиятельностью его соседей, их соседей и т.д.
2. Реализации:
   1. Престиж Каца (Katz prestige)
   2. Центральность собственного вектора (eigenvector or Bonacich centrality)
   3. Google PageRank

Центральность собственного вектора

1. A – матрица связности графа, $a_{ij} = 1$, если есть ребро (i, j)
2. Центральность узла i задается как

$$x_i = \sum_{j \in N} a_{ij} x_j \tag{32}$$

3. Центральность x – решение системы

$$Ax = \lambda x \tag{33}$$

4. x – собственный вектор A. Считаем, что λ – наибольшее собственное число A

Ориентированные графы

1. Матрица несимметрична, $a_{ij} = 1$, если есть ребро от j к i
2. Входящие ребра
$$x_i = \sum_{j \in N} a_{ij} x_j \tag{34}$$
3. x – правый собственный вектор A
4. Вершины, не принадлежащие компоненте сильной связности, получают $x_i = 0$

PageRank

1. Матрица несимметрична, $a_{ij} = 1$, если есть ребро от j к i
2. PageRank
$$Pr_i = \alpha \sum_{j \in N} a_{ij} \frac{Pr_j}{d_j^{out}} + (1 - \alpha)\frac{1}{n} \quad (35)$$
3. $0 < \alpha < 1$, $1 - \alpha$ – коэффициент телепортации
4. $\alpha = 0.85$ в оригинальном PageRank

Заключение

1. Все структурные метрики имеют свои недостатки
2. Выбор конкретной метрики зависит от задач
3. Любой структурной метрикой можно манипулировать
4. Оценки влиятельности пользователей, основанные только на структуре сети, часто не адекватны