Technical Basics

Sergej Sizov
Information Retrieval
Summer term 2013

Technical Basics: Outline

Probability Theory & Stochastics
- Events, Probabilities, Random Variables, Distributions,
  Basics from Information Theory, Markov chains
Linear Algebra
- Vectors and matrices, eigenvectors, common decompositions
Network Analysis
- Properties of social networks

Basic Probability Theory

A probability space is a triple \((\Omega, E, P)\) with
- a set \(\Omega\) of elementary events (sample space),
- a family \(E\) of subsets of \(\Omega\) with \(\Omega \in E\) which is closed under \(\cap, \cup, -\) with a countable number of operands
  (with finite \(\Omega\) usually \(E=2^\Omega\)), and
- a probability measure \(P: E \to [0,1]\) with \(P[\Omega]=1\) and
  \(P[\bigcup_i A_i] = \sum_i P[A_i]\) for countably many, pairwise disjoint \(A_i\).

Properties of \(P\):
- \(P[A] + P[\neg A] = 1\)
- \(P[A \cup B] = P[A] + P[B] - P[A \cap B]\)
- \(P[\emptyset] = 0\) (null/impossible event)
- \(P[\Omega] = 1\) (true/certain event)

Independence and Conditional Probabilities

Two events \(A, B\) of a prob. space are independent if \(P[A \cap B] = P[A] P[B]\).

A finite set of events \(A_1, ..., A_n\) is independent if for every subset \(S \subseteq A\) the equation
\[
P[\bigcap_{i \in S} A_i] = \prod_{i \in S} P[A_i]
\]
holds.

The conditional probability \(P[A | B]\) of \(A\) under the condition (hypothesis) \(B\) is defined as:
\[
P[A | B] = \frac{P[A \cap B]}{P[B]}
\]

Event \(A\) is conditionally independent of \(B\) given \(C\) if \(P[A | BC] = P[A | C]\).

Total Probability and Bayes’ Theorem

Total probability theorem:
For a partitioning of \(\Omega\) into events \(B_1, ..., B_n:\)
\[
P[A] = \sum_{i=1}^n P[A | B_i] P[B_i]
\]

Bayes’ theorem:
\[
P[A | B] = \frac{P[A \cap B]}{P[B]} = \frac{P[B | A] P[A]}{P[B]}
\]
P\([A|B]\) is called posterior probability
P\([A]\) is called prior probability
For a random variable $X$ with distribution function $F$, the inverse function $F^{-1}(x)$ is called the quantile function (percentile) of $X$. The median of $X$ is $F^{-1}(0.5)$. If $F^{-1}(x)$ is continuous, the median is unique and $F^{-1}(0.5)$ is called the median of $X$.

**Example: Probabilistic Retrieval with Term Independence**

Ranking Proportional to Relevance Odds

$$\text{sim}(d,q) = \text{O}(R\mid d) \frac{P[f\mid R\mid d]}{P[f\mid \neg R\mid d]}$$

where $P[f\mid R\mid d]$ is odds for relevance (ratio of relevant documents) and $P[f\mid \neg R\mid d]$ is the joint distribution over the prob. space $(\Omega, E, P)$. $f$ is a function of $X$.

Bayes’ theorem

$$P[X|d\in\mathbb{E}_k] = \frac{P[d\in\mathbb{E}_k|X]P[X]}{P[d\in\mathbb{E}_k]}$$

with feature independence or linked dependence:

$$P[d\in\mathbb{E}_k|X] = \prod_{j=1}^m P[X_j|d\in\mathbb{E}_k]$$

$$P[d\in\mathbb{E}_k] = \prod_{j=1}^m P[X_j]$$

for binary classification with odds rather than probs for simplification.

**Example: Naive Bayes Classification with Binary Features**

Estimate:

$$P[f\mid d\in\mathbb{E}_k] = \frac{P[f\mid d\in\mathbb{E}_k]}{P[d\in\mathbb{E}_k]}$$

$$P[d\in\mathbb{E}_k|X_j] = \frac{P[X_j|d\in\mathbb{E}_k]P[d\in\mathbb{E}_k]}{P[d\in\mathbb{E}_k]}$$

$$P[d\in\mathbb{E}_k] = \prod_{j=1}^m P[X_j]$$

with empirically estimated

$$P[X_j] = \frac{\sum_{i=1}^n P[X_j|X_i]}{\sum_{i=1}^n P[X_i]}$$

for binary classification with odds rather than probs for simplification.

### Random Variables

A random variable (RV) $X$ on the prob. space $(\Omega, E, P)$ is a function $X: \Omega \rightarrow M$ with $M \subseteq \mathbb{R}$ s.t. $(\omega \in \Omega, x \in M) \in E$ for all $x \in M$ (X is measurable).

$F_X: \Omega \rightarrow [0,1]$ with $F_X(x) = P[X \leq x]$ is the (cumulative) distribution function (CDF) of X.

With countable $M$ the function $F_x: M \rightarrow [0,1]$ with $F_x(x) = P[X = x]$ is called the probability density function (pdf) of X, in general $F_x(x)$ is $P_X(x)$.

For a random variable $X$ with distribution function $F$, the inverse function $F^{-1}(q)$ is called the quantile function of X. The median of $X$ is $F^{-1}(0.5)$.

Random variables with countable $M$ are called discrete, otherwise they are called continuous.

For discrete random variables the density function is also referred to as the probability mass function.

### Important Discrete Distributions

- **Bernoulli** distribution with parameter $p$: $P[X = x] = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$
- **Uniform** distribution over $[1,2, \ldots, m]$: $P[X = k] = \frac{1}{m}$ for $1 \leq k \leq m$
- **Poisson** distribution (with rate $\lambda$): $P[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}$
- **Geometric** distribution ($\#$coin tosses until first head): $P[X = k] = e^{-\lambda} \frac{\lambda^k}{k}$

### Important Continuous Distributions

- **Uniform** distribution in the interval $[a,b]$: $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$ (0 otherwise)
- **Exponential** distribution (x.B. time until next event of a Poisson process) with rate $\lambda = \lim_{\Delta t \rightarrow 0} (#\text{ events in } \Delta t) / \Delta t$:
  $$f_X(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0 (0 \text{ otherwise})$$
- **Hyperexponential distribution**: $f_X(x) = \alpha x e^{-\lambda_1 x} + (1-\alpha) \lambda_2 e^{-\lambda_2 x}$

### Multidimensional (Multivariate) Distributions

Let $X_1, \ldots, X_m$ be random variables over the same prob. space with domains $\text{dom}(X_1), \ldots, \text{dom}(X_m)$. The joint distribution of $X_1, \ldots, X_m$ has a density function

$$f_{X_1,\ldots,X_m}(x_1,\ldots,x_m)$$

with

$$\sum_{x_i \in \text{dom}(X_i)} \sum_{x_m \in \text{dom}(X_m)} f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = 1$$

or

$$\int \ldots \int f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) \, dx_m \ldots dx_1 = 1$$

The marginal distribution of $X_i$ in the joint distribution of $X_1, \ldots, X_m$ has the density function

$$\sum_{x_i} \sum_{x_m} f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) \quad \text{or}$$

$$\int \ldots \int f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) \, dx_m \ldots dx_1$$
**Important Multivariate Distributions**

- **Multinomial distribution** (n trials with m-sided dice):
  \[ P[X_1 = k_1, \ldots, X_n = k_n] = \frac{n!}{k_1! \cdots k_n!} p_1^{k_1} \cdots p_m^{k_m} \]
  with \( k_i \geq 0 \) and \( k_1 + \cdots + k_n = n \).

- **Multidimensional normal distribution**:
  \[ f_{X_1, \ldots, X_n}(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)} \]
  with covariance matrix \( \Sigma \) with \( \Sigma_{ij} = \text{Cov}(X_i, X_j) \).

**Important Properties of Expectation and Variance**

**Important Multivariate Distributions**

- **Markov property**: \( P[S(t) = i \mid S(0), \ldots, S(t-1)] = P[S(t) = i \mid S(t-1)] \)
  for every choice of \( t_1, \ldots, t_n \) from the parameter space and every choice of \( x_i, \ldots, x_{i+1} \) from the state space the following holds:
  \[ P[f_{X(t+1)}] = x_1 \cdot X(t+2) \cdot x_2 \cdot \ldots \cdot X(t+1) = x_n \]
  \[ P[f_{X(t+1)}] = x_{n+1} \cdot X(t+1) = x_{n+1} \]
  A Markov process with discrete state space is called **Markov chain**. A canonical choice of the state space are the natural numbers. Notation for Markov chains with discrete parameter space: \( X_t \) rather than \( X(t) \) with \( n = 0, 1, 2, \ldots \)

**Moments**

- For a discrete random variable \( X \) with density \( f \)
  \[ E[X] = \sum_{x \in X} x f(x) \] is the expectation value (mean) of \( X \)
  \[ E[X^k] = \sum_{x \in X} x^k f(x) \] is the \( k \)-th moment of \( X \)
  \[ V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \] is the variance of \( X \)
- For a continuous random variable \( X \) with density \( f \)
  \[ E[X] = \int \limits_{-\infty}^{\infty} x f(x) dx \] is the expectation value of \( X \)
  \[ E[X^k] = \int \limits_{-\infty}^{\infty} x^k f(x) dx \] is the \( k \)-th moment of \( X \)
  \[ V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \] is the variance of \( X \)

**Theorem**: Expectation values are additive: \( E[X + Y] = E[X] + E[Y] \) (distributions are not)

**Correlation of Random Variables**

**Covariance** of random variables \( X_i \) and \( X_j \):

\[ \text{Cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])] \]

\[ \text{Var}(X_i) = E[(X_i - E[X_i])^2] \]

**Correlation coefficient** of \( X_i \) and \( X_j \):

\[ \rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)} \sqrt{\text{Var}(X_j)}} \]
The Markov chain $X_n$ with discrete parameter space is **homogeneous** if the transition probabilities $p_{ij} = P[X_{n+1} = j | X_n = i]$ are independent of $n$.

**irreducible** if every state is reachable from every other state with positive probability:
$$\sum_{n=1}^{\infty} P[X_n = j \land X_k \neq i \text{ for } k = 1, \ldots, n-1 | X_0 = i] > 0 \text{ for all } i, j$$

aperiodic if every state $i$ has period 1, where the period of $i$ is the greatest common divisor of all (recurrence) values $n$ for which
$$P[X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 | X_0 = i] > 0$$

positive recurrent if for every state $i$ the recurrence probability is 1 and the mean recurrence time is finite:
$$\sum_{n=1}^{\infty} n P[X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 | X_0 = i] = 1$$
$$\sum_{n=1}^{\infty} n P[X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 | X_0 = i] < \infty$$

ergodic if it is homogeneous, irreducible, aperiodic, and positive recurrent.

Results on Markov Chains with Discrete Parameter Space (1)

For the $n$-step transition probabilities $P_{ij}^{(n)} = P[X_n = j | X_0 = i]$ the following holds:
$$P_{ij}^{(n)} = \sum_k P_{ik}^{(n-1)} P_{kj}$$
$$\Pi^{(n)} = \Pi^{(1)} P^{(n)}$$

in matrix notation:
$$P^{(n)} = P^n$$

Results on Markov Chains with Discrete Parameter Space (2)

Every homogeneous, irreducible, aperiodic Markov chain with a finite number of states is positive recurrent and ergodic.

For every ergodic Markov chain there exist **stationary state probabilities**
$$\pi_j := \lim_{n \to \infty} P_{ij}^{(n)}$$

These are independent of $\Pi^{(0)}$, and are the solutions of the following system of linear equations:
$$\sum_j \pi_j p_{ij} = \pi_i \text{ for all } i$$

in matrix notation:
$$\Pi = \Pi P$$

Elementary Information Theory

Let $f(x)$ be the probability (or relative frequency) of the $x$-th symbol in some text $d$. The **entropy** of the text (or the underlying prob. distribution $f$) is:
$$H(d) = \sum_x f(x) \log_2 \frac{1}{f(x)}$$

$H(d)$ is a lower bound for the bits per symbol needed with optimal coding (compression).

For two prob. distributions $f(x)$ and $g(x)$ the **relative entropy** (Kullback-Leibler divergence) of $f$ to $g$ is
$$D(f \parallel g) := \sum_x f(x) \log_2 \frac{f(x)}{g(x)}$$

Relative entropy is a measure for the (dis-)similarity of two probability or frequency distributions.

It corresponds to the average number of additional bits needed for coding information (events) with distribution $f$ when using an optimal code for distribution $g$.

The **cross entropy** of $f(x)$ to $g(x)$ is:
$$H(f,g) := H(f) + D(f \parallel g) = -\sum_x f(x) \log_2 g(x)$$
**Compression**

- Text is sequence of symbols (with specific frequencies)
- Symbols can be:
  - letters or other characters from some alphabet $\Sigma$
  - strings of fixed length (e.g., trigrams)
  - or words, bits, syllables, phrases, etc.

**Limits of compression:**

Let $p_i$ be the probability (or relative frequency) of the $i$-th symbol in text $d$.

Then the **entropy** of the text $H(d) = \sum p_i \log_2 \frac{1}{p_i}$ is **a lower bound** for the average number of bits per symbol in any compression (e.g., Huffman codes).

**Note:**

Compression schemes such as Ziv-Lempel (used in zip) are better because they consider context beyond single symbols; with appropriately generalized notions of entropy, the lower-bound theorem does still hold.

---

**The Entropy of Flickr**

The Entropy of Flickr (2)

---

**Part 3: Using Vectors and Matrices**

**IR and Linear Algebra**

- Terms characterize documents (term-document matrix)
- But also: documents characterize terms...
- User/resource/tag relationships in folksonomies
- Alternate document/term representation forms also possible:
  - e.g. concept based (using thesaurus like WordNet)
  - e.g. with rich background document corpus (using Wikipedia in ESA)

But matrix representation is also frequently used for:

- Representing connections in graph models (adjacency matrix)
- Representing transitions in stochastic models (probabilistic matrix)

... and many other IR relevant applications and methods

---

**Foundations from Linear Algebra**

A set $S$ of vectors is called **linearly independent** if no $x \in S$ can be written as a linear combination of other vectors in $S$.

The **rank** of matrix $A$ is the maximal number of linearly independent row or column vectors.

A **basis** of an $n \times n$ matrix $A$ is a set $S$ of linearly independent row or column vectors such that all rows or columns are linear combinations of vectors from $S$.

A set $S$ of $n \times 1$ vectors is an **orthonormal basis** if for all $x, y \in S$:

$$||X||_2 = \sqrt{\sum_{i=1}^n x_i^2} = 1 = ||y||_2 \text{ and } x \cdot y = 0$$
**Eigenvalues and Eigenvectors**

Let A be a real-valued non-negative matrix, x a real-valued n×1 vector, and λ a real-valued scalar. Solutions x ≠ 0 and λ of the equation Ax = λx are called an **Eigenvalue** and **Eigenvector** of A.

Eigenvectors of A are vectors whose direction is preserved by the linear transformation described by A.

The **Eigenvalues** of A are the roots (Nullstellen) of the characteristic polynomial f(λ) of A:  
\[ f(λ) = |A - λI| = 0 \]

with the determinant (developing the i-th row):

\[ |A - λI| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} f^{(j)}(λ) \]

where matrix A(λ) is derived from A by replacing the i-th row and the j-th column.

The real-valued non-negative matrix A is symmetric if \( a_{ij} = a_{ji} \) for all i, j. A is **positive definite** if for all n×1 vectors \( x \neq 0 \): \( x^T A x > 0 \).

If A is symmetric and positive definite then all Eigenvalues of A are real.

**Singular Value Decomposition (SVD)**

**Theorem:**
Each real-valued m×n matrix A with rank r can be decomposed into the form \( A = U \Sigma V^T \) with an m×m matrix U with orthonormal column vectors, an r×r diagonal matrix \( \Sigma \), and an n×r matrix V with orthonormal column vectors.

This decomposition is called **singular value decomposition** and is unique when the elements of \( \Sigma \) are sorted.

**Theorem:**
In the singular value decomposition \( A = U \Sigma V^T \) of matrix A the matrices U, \( \Sigma \), and V can be derived as follows:

- \( \Sigma \) consists of the singular values of A, i.e. the positive roots of the Eigenvalues of \( A^T A \),
- the columns of U are the Eigenvectors of \( A A^T \),
- the columns of V are the Eigenvectors of \( A^T A \).

**SVD for Regression**

**Theorem:**
Let A be an m×n matrix with rank r, and let \( A_k = U_k \Delta_k V_k^T \),

where the k×k diagonal matrix \( \Delta_k \) contains the k largest singular values of A and the m×k matrix \( U_k \) and the n×k matrix \( V_k \) contain the corresponding Eigenvectors from the SVD of A.

Among all m×n matrices C with rank at most k, \( A_k \) is the matrix that minimizes the Frobenius norm

\[ \|A - C\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (A_{ij} - C_{ij})^2 \]

**SVD and Vector Space Model**

**Term-document matrix:**

\[
\begin{pmatrix}
\vdots \\
A_{ij} \\
\vdots \\
\end{pmatrix}
\]

Observation: dot product \( f'_t \) gives the ‘similarity’ between \( f_t , f'_t \) over the documents. The matrix product \( A x A^T \) contains all these dot products. Element \((x,y) \in \text{content} \ (y,x) \in \text{context} \) contains the dot product \( f_t f'_y \).

The matrix \( A^T A \) contains the dot products between document vectors, and gives their ‘similarity’ over terms: element \((m,n) \in \text{element} \ (n,m) \in \text{content} \) \( d_m d_n \).

**SVD Theorem:** In the singular value decomposition \( A = U \Delta V^T \) of matrix A the matrices U, \( \Delta \), and V can be derived as follows:

- \( \Delta \) consists of the singular values of A, i.e. the positive roots of the Eigenvalues of \( A^T A \),
- the columns of U are the Eigenvectors of \( A A^T \),
- the columns of V are the Eigenvectors of \( A^T A \).

**Key Idea of Latent Concept Models**

**Objective:**
Transformation of document vectors from high-dimensional term vector space into lower-dimensional **topic vector space** with:

- exploitation of term correlations (e.g. “Web” and “Internet” frequently occur in together)
- implicit differentiation of polysemes that exhibit different term correlations for different meanings (e.g. “Java” with “Library” vs. “Java” with “Kona Blend” vs. “Java” with “Ionomix”)

**Mathematically:**
given: m terms, n docs (usually n > m) and a m×n term-document similarity matrix A, needed: largely similarity-preserving mapping of column vectors of A into k-dimensional vector space (k << m) for given k

**Latent Semantic Indexing (LSI)**

[Deerwester et al. 1990]

A is the m×n term-document matrix. Then:

- U and \( U_k \) are the m×m and m×k term-topic similarity matrices,
- V and \( V_k \) are the n×n and n×k document-topic similarity matrices,
- \( A A^T \) and \( A_k A_k^T \) are the term-term similarity matrices,
- \( A^T A \) and \( A_k^T A_k \) are the document-document similarity matrices

**Mapping:**
mapping of m×1 vectors into latent-topic space:

\[
d_j^T \cdot a_i = d_j^T U_i \Rightarrow d_j = U_i \cdot a_i
\]

**Scalar-product similarity in latent-topic space:**

\[
d_j^T q = (A_k V_k^T a_j)^T q
\]
Part 4: Understanding Social Networks

Small World Experiment (Six degrees of separation)

Stanley Milgram: 1967: Letters were handed out to people in Nebraska to be sent to a target (stock broker) in Boston
- People were instructed to pass on the letters to someone they knew on first-name basis
- The letters that reached the destination followed paths of avg length 5.2 (i.e. around 6)

Duncan Watts: 2001: Milgram's experiment recreated on the internet
- using an e-mail message as the "package" that needed to be delivered, with 48,000 senders and 19 targets (in 157 countries).
- the avg number of intermediaries was also around 6.

See also:
- The Kevin Bacon game
- The Erdős number
- etc.

Kevin Bacon Experiment

Craig Fass, Brian Turtle and Mike Ginelli: 1994: motivated by Bacon's most recent movie „The Air Up There“ and his career discussion

Vertices: actors and actresses

Edge between u and v if they appeared in a movie together

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Average distance</th>
<th># of movies</th>
<th># of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rod Steiger</td>
<td>2.037527</td>
<td>112</td>
<td>2562</td>
</tr>
<tr>
<td>2</td>
<td>Donald Pleasence</td>
<td>2.542730</td>
<td>140</td>
<td>2674</td>
</tr>
<tr>
<td>3</td>
<td>Martin Sheen</td>
<td>2.551210</td>
<td>136</td>
<td>2593</td>
</tr>
<tr>
<td>4</td>
<td>Christopher Lee</td>
<td>2.552497</td>
<td>201</td>
<td>2993</td>
</tr>
<tr>
<td>5</td>
<td>Robert Mitchum</td>
<td>2.557181</td>
<td>136</td>
<td>2905</td>
</tr>
<tr>
<td>6</td>
<td>Charlton Heston</td>
<td>2.562044</td>
<td>194</td>
<td>3552</td>
</tr>
<tr>
<td>7</td>
<td>Christopher Lee</td>
<td>2.567030</td>
<td>112</td>
<td>3533</td>
</tr>
<tr>
<td>8</td>
<td>Robert Vaughn</td>
<td>2.570193</td>
<td>126</td>
<td>2761</td>
</tr>
<tr>
<td>9</td>
<td>Donald Sutherland</td>
<td>2.571889</td>
<td>107</td>
<td>2665</td>
</tr>
<tr>
<td>10</td>
<td>John Gielgud</td>
<td>2.579902</td>
<td>122</td>
<td>2942</td>
</tr>
<tr>
<td>11</td>
<td>Anthony Quinn</td>
<td>2.573573</td>
<td>146</td>
<td>2878</td>
</tr>
<tr>
<td>12</td>
<td>James Earl Jones</td>
<td>2.584440</td>
<td>112</td>
<td>3787</td>
</tr>
</tbody>
</table>

Kevin Bacon

No. of movies: 46
No. of actors: 1811
Average separation: 2.79

Is Kevin Bacon the most connected actor?

See also:
- http://oracleofbacon.org/

Quantities of Interest

Connected components:
- how many, and how large?

Network diameter:
- maximum (worst-case) or average?
- exclude infinite distances? (disconnected components)
- the small-world phenomenon

Clustering:
- to what extent that links tend to cluster "locally"?
- what is the balance between local and long-distance connections?
- what roles do the two types of links play?

Degree distribution:
- what is the typical degree in the network?
- what is the overall distribution?

Graph theory: undirected graph notation

Graph $G=(V,E)$
- $V =$ set of vertices
- $E =$ set of edges

undirected graph $E=((1,2),(1,3),(2,3),(3,4),(4,5))$

Graph theory: directed graph notation

Graph $G=(V,E)$
- $V =$ set of vertices
- $E =$ set of edges

directed graph $E=((1,2),(2,1),(1,3),(3,2),(3,4),(4,5))$
undirected graph: degree distribution

degree $d(i)$ of node $i$
- number of edges incident on node $i$
degree sequence
- $[d(1),d(2),d(3),d(4),d(5)]$
- $[2,2,3,2,1]$
degree distribution
- $[(1,1),(2,3),(3,1)]$

Directed graph: in/outdegrees

in-degree $d_{in}(i)$ of node $i$
- number of edges pointing to node $i$
out-degree $d_{out}(i)$ of node $i$
- number of edges leaving node $i$
in-degree sequence
- $[1,2,1,1,1]$
out-degree sequence
- $[2,1,2,1,0]$

Degree distributions

Problem: find the probability distribution that best fits the observed data

Power-law distributions

The degree distributions of most real-life networks follow a power law

\[ p(k) = Ck^{-\alpha} \]

Right-skewed/Heavy-tail distribution
- there is a non-negligible fraction of nodes that has very high degree (hubs)
- scale-free: no characteristic scale, average is not informative
  - highly concentrated around the mean
  - the probability of very high degree nodes is exponentially small

Power-law signature

\[ p(k) = Ck^{-\alpha} \]

Power-law distribution gives a line in the log-log plot

\[ \log p(k) = -\alpha \log k + \log C \]

\[ \alpha : \text{power-law exponent (typically } 2 \leq \alpha \leq 3) \]

Power-law: Examples

Taken from [Newman 2003]

<table>
<thead>
<tr>
<th>Network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>Clustering coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>lattice</td>
<td>regular</td>
<td>2000</td>
<td>1000</td>
<td>0.032</td>
</tr>
<tr>
<td>lattice</td>
<td>random</td>
<td>2000</td>
<td>1000</td>
<td>0.002</td>
</tr>
<tr>
<td>regular</td>
<td>1000</td>
<td>500</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>random</td>
<td>1000</td>
<td>500</td>
<td>0.005</td>
<td></td>
</tr>
</tbody>
</table>

Web Structure: Power-Law Degrees

Study of Web Graph (Broder et al. 2000)

- power-law distributed degrees: \( P[\text{degree}=k] \sim (1/k)^\alpha \)
  - \( \alpha \approx 2.1 \) for indegrees
  - \( \alpha \approx 2.7 \) for outdegrees

Clustering (Transitivity) coefficient

Measures the density of triangles (local clusters) in the graph.

Two different ways to measure it:

The ratio of the means:

\[
C^{(1)} = \frac{\sum \text{triangles centered at node } i}{\sum \text{triples centered at node } i}
\]

Example

\[
C^{(1)} = \frac{3}{1+1+6} = \frac{3}{8}
\]

Clustering coefficient for node \( i \)

\[
C_i = \frac{\text{triangles centered at node } i}{\text{triples centered at node } i}
\]

The mean of the ratios:

\[
C^{(2)} = \frac{1}{n} \sum C_i
\]

Example

\[
C^{(2)} = \frac{1}{5} (1+1+1+6) = \frac{13}{30}
\]

\[
C^{(1)} = \frac{3}{8}
\]

The two clustering coefficients give different measures:

- \( C^{(2)} \) increases with nodes with low degree.

<table>
<thead>
<tr>
<th>Network</th>
<th>n</th>
<th>dmax</th>
<th>dmean</th>
<th>dharmonic</th>
<th>dmax/dmean</th>
<th>dmax/dharmonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme</td>
<td>10</td>
<td>1</td>
<td>0.5</td>
<td>1.5</td>
<td>4</td>
<td>1.67</td>
</tr>
<tr>
<td>Protein</td>
<td>20</td>
<td>2</td>
<td>1</td>
<td>1.5</td>
<td>1.33</td>
<td>1.5</td>
</tr>
<tr>
<td>Yeast</td>
<td>15</td>
<td>2.5</td>
<td>1.5</td>
<td>2</td>
<td>1.67</td>
<td>1.25</td>
</tr>
<tr>
<td>E. coli</td>
<td>20</td>
<td>3.5</td>
<td>2.5</td>
<td>3</td>
<td>2.22</td>
<td>1.5</td>
</tr>
<tr>
<td>Page Rank</td>
<td>100</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>2.5</td>
<td>5</td>
</tr>
<tr>
<td>Internet</td>
<td>10,000</td>
<td>1000</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Power grid</td>
<td>3000</td>
<td>1000</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>ridge</td>
<td>1000</td>
<td>50</td>
<td>25</td>
<td>25</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>social network</td>
<td>1000</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>model network</td>
<td>1000</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>food web</td>
<td>3000</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

Clustering coefficient for random graphs

The probability of two of your neighbors also being neighbors is p, independent of local structure

- clustering coefficient \( C = p \)
- when \( z \) is fixed \( C = z/n = O(1/n) \)

Graphs: paths

Path from node i to node j: a sequence of edges (directed or undirected from node i to node j)
- path length: number of edges on the path
- nodes i and j are connected
- cycle: a path that starts and ends at the same node

Graphs: shortest paths

Shortest Path from node i to node j
- also known as BFS path, or geodesic path

Measuring the small world phenomenon

\[ d_i = \text{shortest path between } i \text{ and } j \]
\[ \text{Diameter: } d = \max d_i \]

Characteristic path length:

\[ \ell = \frac{1}{n(n-1)/2} \sum d_i \]

Harmonic mean

\[ \ell^{-1} = \frac{1}{n(n-1)/2} \sum d_i^{-1} \]


<table>
<thead>
<tr>
<th>Network</th>
<th>n</th>
<th>dmax</th>
<th>dmean</th>
<th>dharmonic</th>
<th>dmax/dmean</th>
<th>dmax/dharmonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme</td>
<td>10</td>
<td>1</td>
<td>0.5</td>
<td>1.5</td>
<td>4</td>
<td>1.67</td>
</tr>
<tr>
<td>Protein</td>
<td>20</td>
<td>2</td>
<td>1</td>
<td>1.5</td>
<td>1.33</td>
<td>1.5</td>
</tr>
<tr>
<td>Yeast</td>
<td>15</td>
<td>2.5</td>
<td>1.5</td>
<td>2</td>
<td>1.67</td>
<td>1.25</td>
</tr>
<tr>
<td>E. coli</td>
<td>20</td>
<td>3.5</td>
<td>2.5</td>
<td>3</td>
<td>2.22</td>
<td>1.5</td>
</tr>
<tr>
<td>Page Rank</td>
<td>100</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>2.5</td>
<td>5</td>
</tr>
<tr>
<td>Internet</td>
<td>10,000</td>
<td>1000</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Power grid</td>
<td>3000</td>
<td>1000</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>ridge</td>
<td>1000</td>
<td>50</td>
<td>25</td>
<td>25</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>social network</td>
<td>1000</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>model network</td>
<td>1000</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>food web</td>
<td>3000</td>
<td>500</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>
Degree correlations

Do high degree nodes tend to link to high degree nodes?

Newman

- compute the correlation coefficient of the degrees $x_i$, $y_i$ of the two endpoints of an edge $i$

\[
K_{or}(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}
\]

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \\
\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]


<table>
<thead>
<tr>
<th>Community</th>
<th>Size</th>
<th>Average Degree</th>
<th>Density</th>
<th>Average Path Length</th>
<th>Diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>1000</td>
<td>2.4</td>
<td>0.008</td>
<td>4.2</td>
<td>9</td>
</tr>
<tr>
<td>WWW</td>
<td>5000</td>
<td>6.0</td>
<td>0.001</td>
<td>8.5</td>
<td>12</td>
</tr>
<tr>
<td>Small World</td>
<td>100</td>
<td>5.0</td>
<td>0.010</td>
<td>3.5</td>
<td>4</td>
</tr>
</tbody>
</table>

Graphs: fully connected cliques

Clique $K_n$

A graph that has all possible $n(n-1)/2$ edges

Directed Graph

**Strongly connected** graph: there exists a path from every $i$ to every $j$

**Weakly connected** graph: If edges are made to be undirected the graph is connected

Web Structure: Connected Components

Study of Web Graph (Broder et al. 2000)

Most nodes have only a small number of neighbors (degree), but there are some nodes with very high degree (power-law degree distribution)

- scale-free networks

If a node $x$ is connected to $y$ and $z$, then $y$ and $z$ are likely to be connected

- high clustering coefficient

Most nodes are just a few edges away on average.

- small world networks
A "Canonical" Natural Network has...

Few connected components:
- often only 1 or a small number, indep. of network size

Small diameter:
- often a constant independent of network size (like 6)
- or perhaps growing only logarithmically with network size or even shrink?
- typically exclude infinite distances

A high degree of clustering:
- considerably more so than for a random network
- in tension with small diameter

A heavy-tailed degree distribution:
- a small but reliable number of high-degree vertices
- often of power law form

Is it possible that there is a unifying underlying generative process?