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
Part 1: Basics from probability theory and stochastics
Basic Probability Theory

A **probability space** is a triple \((\Omega, E, P)\) with

- a set \(\Omega\) of elementary events (sample space),
- a family \(E\) (event space) of subsets of \(\Omega\) with \(\Omega \in E\) which is closed under \(\cap, \cup\), and - with a countable number of operands (with finite \(\Omega\) usually \(E = 2^\Omega\)), and
- a **probability measure** \(P: E \rightarrow [0,1]\) 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 $P[A \cap B] = P[A] \cdot P[B]$.

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

The conditional probability $P[A \mid B]$ of $A$ under the condition (hypothesis) $B$ is defined as:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]}$$

Event $A$ is conditionally independent of $B$ given $C$ if $P[A \mid BC] = P[A \mid C]$. 
Total Probability and Bayes’ Theorem

**Total probability theorem:**
For a partitioning of $\Omega$ into events $B_1, \ldots, B_n$:

$$P[ A] = \sum_{i=1}^{n} P[ A \mid B_i] P[ B_i]$$

**Bayes’ theorem:**

$$P[ A \mid B] = \frac{P[ A \cap B]}{P[ B]} = \frac{P[ B \mid A] P[ A]}{P[ B]}$$

$P[ A \mid B]$ is called *posterior probability*

$P[ A]$ is called *prior probability*
Example: Probabilistic Retrieval with Term Independence

Ranking Proportional to Relevance Odds

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

\[
= \frac{P[d \mid R] \times P[R]}{P[d \mid \neg R] \times P[\neg R]}
\]

\[
\sim \frac{P[d \mid R]}{P[d \mid \neg R]} = \prod_i \frac{P[X_i \mid R]}{P[X_i \mid \neg R]}
\]

\[
sim(d, q)' = \log \prod_{i \in q} \frac{P[X_i \mid R]}{P[X_i \mid \neg R]}
\]

\[
= \sum_{i \in q} \log P[X_i \mid R] - \log P[X_i \mid \neg R]
\]
Example: Naive Bayes classification with Binary Features

estimate:  \( P \left[ d \in c_k \mid d \text{ has } \vec{X} \right] = \frac{P \left[ d \text{ has } \vec{X} \mid d \in c_k \right] P \left[ d \in c_k \right]}{P \left[ d \text{ has } \vec{X} \right]} \)

\( \sim P \left[ X \mid d \in c_k \right] P \left[ d \in c_k \right] \)

\( = \prod_{i=1}^{m} P \left[ X_i \mid d \in c_k \right] P \left[ d \in c_k \right] \)  

with feature independence or linked dependence:

\( \frac{P \left[ X \mid d \in c_k \right]}{P \left[ X \mid d \notin c_k \right]} = \prod_{i} \frac{P \left[ X_i \mid d \in c_k \right]}{P \left[ X_i \mid d \notin c_k \right]} \)

\( = \prod_{i=1}^{m} p_{ik}^{X_i} (1 - p_{ik})^{1 - X_i} p_k \)  

with empirically estimated  

\( p_{ik} = P[X_i = 1 \mid c_k], \quad p_k = P[c_k] \)

\( \Rightarrow \log P[c_k \mid d] \sim \sum_{i=1}^{m} X_i \log \frac{p_{ik}}{(1 - p_{ik})} + \sum_{i=1}^{m} \log (1 - p_{ik}) + \log p_k \)

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 $X: \Omega \to M$ with $M \subseteq \mathbb{R}$ s.t. $\{e \mid X(e) \leq x\} \in E$ for all $x \in M$ ($X$ is measurable).

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

With countable set $M$ the function $f_X: M \to [0,1]$ with $f_X(x) = P[X = x]$ is called the (probability) density function (pdf) of $X$; in general $f_X(x)$ is $F'_X(x)$.

For a random variable $X$ with distribution function $F$, the inverse function $F^{-1}(q) := \inf\{x \mid F(x) > q\}$ for $q \in [0,1]$ is called quantile function of $X$. (0.5 quantile (50th percentile) is called median)

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 \( \{1, 2, ..., m\} \):
  
  \[
P[X = k] = f_X(k) = \frac{1}{m} \quad \text{for} 1 \leq k \leq m
  \]

- **Binomial** distribution (coin toss \( n \) times repeated; \( X \): #heads):
  
  \[
P[X = k] = f_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}
  \]

- **Poisson** distribution (with rate \( \lambda \)):
  
  \[
P[X = k] = f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}
  \]

- **Geometric** distribution (#coin tosses until first head):
  
  \[
P[X = k] = f_X(k) = (1 - p)^k p
  \]
Important Continuous Distributions

• **Uniform** distribution in the interval \([a,b]\)

\[
f_X(x) = \frac{1}{b-a} \quad \text{for } a \leq x \leq b \quad \text{(0 otherwise)}
\]

• **Exponential** distribution (z.B. time until next event of a Poisson process) with rate \(\lambda = \lim_{\Delta t \to 0} \frac{\# \text{ events in } \Delta t}{\Delta t}\) :

\[
f_X(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0 \quad \text{(0 otherwise)}
\]

• **Hyperexponential** distribution: \(f_X(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}\)
Multidimensional (Multivariate) Distributions

Let $X_1, ..., X_m$ be random variables over the same prob. space with domains $\text{dom}(X_1), ..., \text{dom}(X_m)$.

The joint distribution of $X_1, ..., X_m$ has a density function

$$f_{X_1, ..., X_m}(x_1, ..., x_m)$$

with

$$\sum_{x_1 \in \text{dom}(X_1)} \cdots \sum_{x_m \in \text{dom}(X_m)} f_{X_1, ..., X_m}(x_1, ..., x_m) = 1$$

or

$$\int_{\text{dom}(X_1)} \cdots \int_{\text{dom}(X_m)} f_{X_1, ..., X_m}(x_1, ..., x_m) \, dx_m \cdots dx_1 = 1$$

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

$$\sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_m} f_{X_1, ..., X_m}(x_1, ..., x_m) \text{ or}$$

$$\int_{X_1} \cdots \int_{X_{i-1}} \int_{X_{i+1}} \cdots \int_{X_m} f_{X_1, ..., X_m}(x_1, ..., x_m) \, dx_m \, dx_{i+1} \, dx_{i-1} \cdots dx_1$$
Important Multivariate Distributions

**multinomial distribution** (n trials with m-sided dice):

\[
P [ X_1 = k_1 \land \ldots \land X_m = k_m ] = f_{X_1,\ldots,X_m}(k_1,\ldots,k_m) = \binom{n}{k_1 \ldots k_m} p_1^{k_1} \ldots p_m^{k_m}
\]

with \( \binom{n}{k_1 \ldots k_m} := \frac{n!}{k_1! \ldots k_m!} \)

**multidimensional normal distribution:**

\[
f_{X_1,\ldots,X_m}(\bar{x}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2} (\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu)}
\]

with covariance matrix \( \Sigma \) with \( \Sigma_{ij} := \text{Cov}(X_i,X_j) \)
Moments

For a discrete random variable $X$ with density $f_X$

$$E[X] = \sum_{k \in M} k f_X(k)$$  is the **expectation value (mean)** of $X$

$$E[X^i] = \sum_{k \in M} k^i f_X(k)$$  is the **$i$-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 $f_X$

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) \, dx$$  is the **expectation value** of $X$

$$E[X^i] = \int_{-\infty}^{+\infty} x^i f_X(x) \, dx$$  is the **$i$-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)
Important Properties of Expectation and Variance

\[ E[aX+b] = aE[X]+b \] for constants a, b

\[ E[X_1+X_2+...+X_n] = E[X_1] + E[X_2] + ... + E[X_n] \]
(i.e. expectation values are generally additive, but distributions are not!)

\[ \text{Var}[aX+b] = a^2 \text{Var}[X] \] for constants a, b

\[ \text{Var}[X_1+X_2+...+X_n] = \text{Var}[X_1] + \text{Var}[X_2] + ... + \text{Var}[X_n] \]
if \( X_1, X_2, \ldots, X_n \) are independent RVs
Correlation of Random Variables

**Covariance** of random variables $X_i$ and $X_j$:

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

$$\text{Var}(X_i) = \text{Cov}(X_i, X_i) = E[X^2] - E[X]^2$$

**Correlation coefficient** of $X_i$ and $X_j$

$$\rho(X_i, X_j) := \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)} \sqrt{\text{Var}(X_j)}}$$
Markov Chains

- State set: finite or infinite
- Time: discrete or continuous
- State transition probabilities: $p_{ij}$
- State probabilities in step $t$: $p_i(t) = P[S(t)=i]$
- Markov property: $P[S(t)=i | S(0), ..., S(t-1)] = P[S(t)=i | S(t-1)]$

Interested in stationary state probabilities:

$$p_j := \lim_{t \to \infty} p^{(t)}_j = \lim_{t \to \infty} \sum_k p^{(t-1)}_k p_{kj}$$

$$p_j = \sum_k p_k p_{kj}$$

$$\sum_j p_j = 1$$

Guaranteed to exist for irreducible, aperiodic, finite Markov chains

Markov Chain Diagram:

- States: 0: sunny, 1: cloudy, 2: rainy
- Transition probabilities:
  - $P[0 \rightarrow 0] = 0.8$
  - $P[0 \rightarrow 1] = 0.2$
  - $P[1 \rightarrow 1] = 0.5$
  - $P[1 \rightarrow 2] = 0.3$
  - $P[2 \rightarrow 2] = 0.3$
  - $P[2 \rightarrow 0] = 0.4$

State Probabilities:

- $p_0 = 0.8 p_0 + 0.5 p_1 + 0.4 p_2$
- $p_1 = 0.2 p_0 + 0.3 p_2$
- $p_2 = 0.5 p_1 + 0.3 p_2$
- $p_0 + p_1 + p_2 = 1$

⇒ $p_0 \approx 0.657$, $p_1 = 0.2$, $p_2 \approx 0.143$
Markov Chains: Formal Definition

A **stochastic process** is a family of random variables \( \{X(t) \mid t \in T\} \). T is called parameter space, and the domain M of X(t) is called state space. T and M can be discrete or continuous.

A stochastic process is called **Markov process** if for every choice of \( t_1, \ldots, t_{n+1} \) from the parameter space and every choice of \( x_1, \ldots, x_{n+1} \) from the state space the following holds:

\[
P \left[ X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1 \land X(t_2) = x_2 \land \ldots \land X(t_n) = x_n \right] = P \left[ X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n \right]
\]

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: \( X_n \) rather than \( X(t_n) \) with \( n = 0, 1, 2, \ldots \).
The Markov chain $X_n$ with discrete parameter space is

**homogeneous** if the transition probabilities $p_{ij} := P[X_{n+1} = j \mid X_n=i]$ are independent of $n$

**irreducible** if every state is reachable from every other state with positive probability:

$$
\sum_{n=1}^{\infty} P[X_n = j \mid X_0 = i] > 0 \quad \text{for all } i, j
$$

**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

$$
P[X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 \mid X_0 = i] > 0
$$
The Markov chain $X_n$ with discrete parameter space is

**positive recurrent** if for every state $i$ the recurrence probability is 1 and the mean recurrence time is finite:

$$\sum_{n=1}^{\infty} P[ X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 \mid X_0 = i ] = 1$$

$$\sum_{n=1}^{\infty} n P[ X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 \mid X_0 = i ] < \infty$$

**ergodic** if it is homogeneous, irreducible, aperiodic, and positive recurrent.
For the **n-step transition probabilities**

\[ p_{ij}^{(n)} := P \left[ X_n = j \mid X_0 = i \right] \]

the following holds:

\[ p_{ij}^{(n)} = \sum_k p_{ik}^{(n-1)} p_{kj} \text{ with } p_{ij}^{(1)} := p_{ik} \]

\[ = \sum_k p_{ik}^{(n-l)} p_{kj}^{(l)} \text{ for } 1 \leq l \leq n - 1 \]

in matrix notation:

\[ P^{(n)} = P^n \]

For the **state probabilities after n steps**

\[ \pi_j^{(n)} := P \left[ X_n = j \right] \]

the following holds:

\[ \pi_j^{(n)} = \sum_i \pi_i^{(0)} p_{ij}^{(n)} \text{ with initial state probabilities } \pi_i^{(0)} \]

in matrix notation:

\[ \Pi^{(n)} = \Pi^{(0)} P^{(n)} \]  

*(Chapman-Kolmogorov equation)*
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

$$\pi_j := \lim_{n \to \infty} \pi_j^{(n)}$$

These are independent of $\Pi^{(0)}$ and are the solutions of the following system of linear equations:

$$\pi_j = \sum_i \pi_i \ p_{ij} \quad \text{for all } j$$

$$\sum_j \pi_j = 1$$

In matrix notation:

$$\Pi = \Pi \ P$$

(with 1×n row vector $\Pi$)

$$\Pi \vec{1} = 1$$
Part 2: Basics of Information Theory
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

$$D(f \parallel g) := \sum_x f(x) \log \frac{f(x)}{g(x)}$$

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:

$$H(f, g) := H(f) + D(f \parallel g) = -\sum_x f(x) \log g(x)$$
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:

\[
H(d) = \sum_i p_i \log_2 \frac{1}{p_i}
\]

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

Flickr - Tags vs Photo Descriptions
Vocabulary Size and Tag/Word Assignments

Weeks since January 2004

- Tags
- Words
- Words + NLP
- Dist. TAS
- Dist. WAS
The Entropy of Flickr (2)

Flickr - Tags vs Photo Descriptions

Entropy

Weeks since January 2004

Entropy

Tags
Dist. Tags
Description
Dist. Description
Description + NLP
Dist. Description + NLP
Part 3: Using Vectors and Matrices
IR and Linear Algebra

Vector space model revisited..

- 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
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 $x, y \in S$:

$$\|x\|_2 := \sqrt{\sum_{i=1}^{n} X_i^2} = 1 = \|y\|_2 \quad \text{and} \quad x \cdot y = 0$$
Eigenvalues and Eigenvectors

Let $A$ be a real-valued $n \times n$ matrix, $x$ a real-valued $n \times 1$ vector, and $\lambda$ a real-valued scalar. Solutions $x \neq 0$ and $\lambda$ of the equation $Ax = \lambda x$ are called an **Eigenvector** and **Eigenvalue** 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 polynom $f(\lambda)$ of $A$:

$$f(\lambda) = |A - \lambda I| = 0$$

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

$$|A| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} A^{(ij)}$$

where matrix $A^{(ij)}$ is derived from $A$ by removing the i-th row and the j-th column.

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

If $A$ is symmetric then all Eigenvalues of $A$ are real.

If $A$ is symmetric and positive definite then all Eigenvalues are positive.
Singular Value Decomposition (SVD)

Theorem:
Each real-valued $m \times n$ matrix $A$ with rank $r$ can be decomposed into the form $A = U \times \Delta \times V^T$ with an $m \times r$ matrix $U$ with orthonormal column vectors, an $r \times r$ diagonal matrix $\Delta$, and an $n \times r$ matrix $V$ with orthonormal column vectors. This decomposition is called singular value decomposition and is unique when the elements of $\Delta$ are sorted.

Theorem:
In the singular value decomposition $A = U \times \Delta \times V^T$ 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 $A^T \times A$,
- the columns of $U$ are the Eigenvectors of $A \times A^T$,
- the columns of $V$ are the Eigenvectors of $A^T \times A$. 
SVD for Regression

Theorem:
Let $A$ be an $m \times n$ matrix with rank $r$, and let $A_k = U_k \times \Delta_k \times V_k^T$, where the $k \times k$ diagonal matrix $\Delta_k$ contains the $k$ largest singular values of $A$ and the $m \times k$ matrix $U_k$ and the $n \times k$ matrix $V_k$ contain the corresponding Eigenvectors from the SVD of $A$.

Among all $m \times n$ 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:

![Term-document matrix diagram](image)

**Observation**: dot product $t_i^T \cdot t_j$ gives the “similarity” between $t_i$, $t_j$ over the documents. The matrix product $A \times A^T$ contains all these dot products. Element $(x,y)$ (= element $(y,x)$) contains the dot product $t_x^T \cdot t_y$.

The matrix $A^T \times A$ contains the dot products between document vectors, and gives their “similarity” over terms: element $(m,n)$ (=element $(n,m)$) = $d_m^T \cdot d_n$

**SVD Theorem**: In the singular value decomposition $A = U \times \Delta \times V^T$ 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 $A^T \times A$,
- the columns of $U$ are the Eigenvectors of $A \times A^T$,
- the columns of $V$ are the Eigenvectors of $A^T \times A$. 
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 polysems that exhibit different term correlations for different meanings
  (e.g. „Java“ with „Library“ vs. „Java“ with „Kona Blend“ vs. „Java“ with „Borneo“)

**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 \( m \times n \) term-document matrix. Then:

- \( U \) and \( U_k \) are the \( m \times r \) and \( m \times k \) term-topic similarity matrices,
- \( V \) and \( V_k \) are the \( n \times r \) and \( n \times k \) 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 of \( m \times 1 \) vectors into latent-topic space:

\[
d_j a U_k^T \times d_j =: d_j' \\
qu a U_k^T \times q =: q'
\]

Scalar-product similarity in latent-topic space:

\[
d_j'^T \times q' = ((\Delta_k V_k^T)_{*j})^T \times q'
\]
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 Erdös 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>2.537527</td>
<td>112</td>
<td>2562</td>
</tr>
<tr>
<td>2</td>
<td>Donald Pleasence</td>
<td>2.542376</td>
<td>180</td>
<td>2874</td>
</tr>
<tr>
<td>3</td>
<td>Martin Sheen</td>
<td>2.551210</td>
<td>136</td>
<td>3501</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>2.566284</td>
<td>104</td>
<td>2552</td>
</tr>
<tr>
<td>7</td>
<td>Eddie Albert</td>
<td>2.567036</td>
<td>112</td>
<td>3333</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>2.577880</td>
<td>107</td>
<td>2865</td>
</tr>
<tr>
<td>10</td>
<td>John Gielgud</td>
<td>2.578980</td>
<td>122</td>
<td>2942</td>
</tr>
<tr>
<td>11</td>
<td>Anthony Quinn</td>
<td>2.579750</td>
<td>146</td>
<td>2978</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/

876 Kevin Bacon 2.786981 46 1811
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 $E=\{\langle 1,2 \rangle, \langle 2,1 \rangle, \langle 1,3 \rangle, \langle 3,2 \rangle, \langle 3,4 \rangle, \langle 4,5 \rangle \}$
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 $d_{\text{in}}(i)$ of node $i$
  – number of edges pointing to node $i$

out-degree $d_{\text{out}}(i)$ 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

$f_k = \text{fraction of nodes with degree } k = \text{probability of a randomly selected node to have degree } k$
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

(a) collaborations in mathematics

(b) citations

(c) World Wide Web

(d) Internet

(e) protein interactions

Taken from [Newman 2003]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>(\ell)</th>
<th>(\alpha)</th>
<th>(C^{(1)})</th>
<th>(C^{(2)})</th>
<th>(r)</th>
<th>Ref(s).</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20,416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7,673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105,323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107,182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311,313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,592,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311,313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
<td>8,9</td>
<td></td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td></td>
<td></td>
<td>0.16</td>
<td>136</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td></td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265,266</td>
<td></td>
</tr>
<tr>
<td>www nd. edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14,34</td>
</tr>
<tr>
<td>WWW AltaVista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2</td>
<td></td>
<td></td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>3.0/–</td>
<td></td>
<td></td>
<td></td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>2.7</td>
<td></td>
<td>0.44</td>
<td>119,157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.189</td>
<td>86,148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>–0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>–0.033</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>–0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>–0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,097</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>–0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.366</td>
<td>6,354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>–0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>–0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>–</td>
<td>0.16</td>
<td>0.23</td>
<td>–0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>–</td>
<td>0.20</td>
<td>0.087</td>
<td>–0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>–</td>
<td>0.18</td>
<td>0.28</td>
<td>–0.226</td>
<td>416,421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices \(n\); total number of edges \(m\); mean degree \(z\); mean vertex-vertex distance \(\ell\); exponent \(\alpha\) of degree distribution if the distribution follows a power law (or “–” if not; in/out-degree exponents are given for directed graphs); clustering coefficient \(C^{(1)}\) from Eq. (3); clustering coefficient \(C^{(2)}\) from Eq. (6); and degree correlation coefficient \(r\), Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Web Structure: Power-Law Degrees

Study of Web Graph (Broder et al. 2000)

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

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

$$C^{(2)} = \frac{1}{n} C_i$$
Example

The two clustering coefficients give different measures
$C^{(2)}$ increases with nodes with low degree

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

\[ C^{(1)} = \frac{3}{8} \]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s).</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>76,735</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,900</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td>15/20</td>
<td>0.16</td>
<td>0.16</td>
<td>136</td>
<td>8, 9</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>136</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td></td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>-0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265, 266</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>-0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW AltaVista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td></td>
<td></td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td></td>
<td>3.0/5</td>
<td></td>
<td></td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td></td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td></td>
<td>2.7</td>
<td>0.44</td>
<td></td>
<td>119, 157</td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>-0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td></td>
<td>0.10</td>
<td>0.080</td>
<td>-0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td></td>
<td>0.69</td>
<td>-0.033</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>-0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td></td>
<td>0.033</td>
<td>0.012</td>
<td>-0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,997</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>-0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>-0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>-0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>-0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td></td>
<td>0.16</td>
<td>0.23</td>
<td>-0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td></td>
<td>0.20</td>
<td>0.087</td>
<td>-0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td></td>
<td>0.18</td>
<td>0.28</td>
<td>-0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

**TABLE II** Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex-vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law ($\propto n^{-\alpha}$) if not; in/out-degree exponents are given for directed graphs; clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
The probability of two of your neighbors also being neighbors is $p$, independent of local structure

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

Table 1: Clustering coefficients, $C$, for a number of different networks; $n$ is the number of node, $z$ is the mean degree. Taken from [146].

<table>
<thead>
<tr>
<th>Network</th>
<th>$n$</th>
<th>$z$</th>
<th>$C$ measured</th>
<th>$C'$ for random graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet [153]</td>
<td>6,374</td>
<td>3.8</td>
<td>0.24</td>
<td>0.00060</td>
</tr>
<tr>
<td>World Wide Web (sites) [2]</td>
<td>153,127</td>
<td>35.2</td>
<td>0.11</td>
<td>0.00023</td>
</tr>
<tr>
<td>power grid [192]</td>
<td>4,941</td>
<td>2.7</td>
<td>0.080</td>
<td>0.00054</td>
</tr>
<tr>
<td>biology collaborations [140]</td>
<td>1,520,251</td>
<td>15.5</td>
<td>0.081</td>
<td>0.000010</td>
</tr>
<tr>
<td>mathematics collaborations [141]</td>
<td>253,339</td>
<td>3.9</td>
<td>0.15</td>
<td>0.000015</td>
</tr>
<tr>
<td>film actor collaborations [149]</td>
<td>449,913</td>
<td>113.4</td>
<td>0.20</td>
<td>0.00025</td>
</tr>
<tr>
<td>company directors [149]</td>
<td>7,673</td>
<td>14.4</td>
<td>0.59</td>
<td>0.0019</td>
</tr>
<tr>
<td>word co-occurrence [90]</td>
<td>460,902</td>
<td>70.1</td>
<td>0.44</td>
<td>0.00015</td>
</tr>
<tr>
<td>neural network [192]</td>
<td>282</td>
<td>14.0</td>
<td>0.28</td>
<td>0.049</td>
</tr>
<tr>
<td>metabolic network [69]</td>
<td>315</td>
<td>28.3</td>
<td>0.59</td>
<td>0.090</td>
</tr>
<tr>
<td>food web [138]</td>
<td>134</td>
<td>8.7</td>
<td>0.22</td>
<td>0.065</td>
</tr>
</tbody>
</table>
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_{ij} = \text{shortest path between } i \text{ and } j

Diameter:
\[ d = \max_{i,j} d_{ij} \]

Characteristic path length:
\[ \ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij} \]

Harmonic mean
\[ \ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1} \]
<table>
<thead>
<tr>
<th>social</th>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>ℓ</th>
<th>α</th>
<th>C(1)</th>
<th>C(2)</th>
<th>r</th>
<th>Ref[s]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td></td>
<td>company directors</td>
<td>undirected</td>
<td>7,673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td></td>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td></td>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td></td>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td></td>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td>–</td>
<td>0.006</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>4.4</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>–</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td></td>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td></td>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.006</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265, 266</td>
</tr>
<tr>
<td></td>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td></td>
<td>WWW AltaVista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>–</td>
<td></td>
<td>74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td></td>
<td>3.0/–</td>
<td></td>
<td></td>
<td>351</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td></td>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td></td>
<td>2.7</td>
<td>0.44</td>
<td></td>
<td>119, 157</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td></td>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>–0.003</td>
<td>416</td>
</tr>
<tr>
<td></td>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>–0.033</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td></td>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>–0.016</td>
<td>318</td>
</tr>
<tr>
<td></td>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>–0.119</td>
<td>395</td>
</tr>
<tr>
<td></td>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,997</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>–0.154</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>biological</td>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>–0.240</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>–0.156</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>–</td>
<td>0.16</td>
<td>0.23</td>
<td>–0.263</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>–</td>
<td>0.20</td>
<td>0.087</td>
<td>–0.326</td>
<td>272</td>
</tr>
<tr>
<td></td>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>–</td>
<td>0.18</td>
<td>0.28</td>
<td>–0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex-vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or “–” if not; in/out-degree exponents are given for directed graphs); clustering coefficient C(1) from Eq. (3); clustering coefficient C(2) from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
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 $i$

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

\[
r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \bigg/ \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \quad \quad \quad \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>directed</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td>–</td>
<td>0.8</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td>265, 266</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>3.0/–</td>
<td></td>
<td>0.351</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>2.7</td>
<td></td>
<td>0.44</td>
<td>119, 157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>–0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>–0.033</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>–0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>–0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,097</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>–0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>–0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>–0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>–</td>
<td>0.16</td>
<td>0.23</td>
<td>–0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>–</td>
<td>0.20</td>
<td>0.087</td>
<td>–0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>–</td>
<td>0.18</td>
<td>0.28</td>
<td>–0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices $n$; total number of edges $m$; mean degree $z$; mean vertex-vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law (or “–” if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
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)

SCC = strongly connected component

..strongly connected core tends to have small diameter

Source: A.Z. Broder et al., WWW 2000
Summary: real network properties

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?