What is a network?

- **A collection of points joined together in pairs by lines**
  - Points that will be joined together depends on the context
    - *Points* → Vertices, nodes, actors …
    - *Lines* → Edges

- **There are many systems of interest that can be modeled as networks**
  - Individual parts linked in some way
    - Internet
      - A collection of computers linked together by data connections
    - Human societies
      - A collection of people linked by acquaintance or social interaction
Why are we interested in networks?

- While the individual components of a system (e.g., computer machines, people etc.) as well as the nature of their interaction is important, of equal importance is the pattern of connections between these components.
  - These patterns significantly affect the performance of the underlying system.
    - The patterns of connections in the Internet affect the routes packets travel through.
    - Patterns in a social network affect the way people obtain information, form opinions etc.
Why are we interested in networks?

- Scientists in a wide variety of fields have developed tools for analyzing, modeling and understanding network structures
  - Mathematical, statistical and computational tools
    - Identify the best connected node, the path between two nodes, predict how a process on a network (e.g., spread of a disease) will take place etc.

- These tools work in an abstract level – not considering specific properties of the network examined
  - General applicability to any system that can be represented as a network
The importance of structure

- The value of a system does not only depend on its individual components

- What is common in pencil and diamond?
  - They are both made of carbon
The importance of structure

- The value of a system does not only depend on its individual components.

- What is different in pencil and diamond?
  - Their structure!

Figures obtained from: [http://batteryblog.ca](http://batteryblog.ca) and [http://www.e6cvd.com](http://www.e6cvd.com)
Examples of networks

- The Internet
  - Depending on the level of granularity we examine it vertices can be network devices (host machines and routers) or autonomous systems
  - Edges are physical links between vertices
  - Studying the Internet structure can help understand and improve the performance
    - How do we route packets over the Internet?
    - How resilient is the Internet in the failure of nodes/edges?
    -Which edge’s capacity should we boost?
Examples of networks

- **The World Wide Web**
  - Often is confused with the Internet but they are different!
  - The Web is a network of information stored in webpages
    - Nodes are the webpages
    - Edges are the hyperlinks between the pages
  - The structure of this network is one of the major factors that Google exploits in its search engine
  - Directed edges

Figure obtained from: http://mathinsight.org/network_introduction
Examples of networks

- **Social networks**
  - Network of people
    - Edges can represent friendships, relative relations, co-locations etc.
  - Long tradition in network analysis
  - Traditionally social network studies were based on small scale networks
    - Online social media have provided network data on previously unreachable scale
Example of networks

- **Biological networks**
  - Food webs
    - Ecological network
      - Vertices are species in an ecosystem
      - Directed edge from A to B, iff B eats A
  - Can help study many ecological phenomena, particularly concerning energy and carbon flows in ecosystems
    - Edges typically point from the prey to the predator, indicating the direction of the flow of energy
  - Metabolic networks
    - Protein-protein interactions
Example of networks

- **Transportation networks**
  - Airline routes, road and rail networks
    - Road networks (usually)
      - Vertices: road intersections
      - Edges: roads
    - Rail networks
      - Vertices: locations
      - Edges: between locations that are connected with a single train
      - More general bipartite networks
  - Network theory can be applied to identify the “optimal” structure of a transportation network for a pre-specified objective
Example of networks

- **Delivery and distribution networks**
  - Gas pipelines, water and sewerage lines, routes used by the post office and package delivery and cargo companies
    - Gas distribution network
      - Edges: pipelines
      - Vertices: intersections of pipelines
        - Pumping, switching and storage facilities and refineries
  - River networks, blood vessels in animals and plants
Example of networks

- **Affiliation networks**
  - Two types of vertices
    - Connections allowed only among different types of vertices
  - E.g., board of directors of companies and their members, people and locations they visit etc.

![Affiliation Network Diagram]

Figure 3.2: The affiliation network of the "Southern Women Study." This network (like all affiliation networks) has two types of vertex, the open circles at the bottom representing the 18 women who were the subjects of the study and the shaded circles at the top representing the social events they attended. The edges connect each woman to the events she attended, as deduced from newspaper reports. Data courtesy of L. Freeman and originally from Davis et al. [86].

We will see some examples of these and other networks throughout this book and we will give details as needed as we go along. The rest of the present chapter is devoted to a discussion of the different empirical methods used to measure social networks. The two techniques described above, namely direct questioning of subjects and the use of archival records, are two of the most important, but there are several others that find regular use. This chapter does not give a complete review of the subject—for that we refer the reader to specialized texts such as those of Wasserman and Faust [320] and Scott [293]—but we introduce as much material as will be needed for the later chapters of the book, while at the same time, we hope, giving some flavor for the challenges of empirical study in the field of social networks.
Example of networks

- **Citation networks**
  - Vertices are papers
  - An edge exists from paper A to paper B, iff A cites B
  - Network analysis can be used to identify influential papers
    - Bibliometrics

- **Recommender networks**
  - Represent people’s preferences for things
    - E.g., preference on certain products sold by a retailer
Synthesize

- Many literature from different disciplines deals with networks
  - Sociology
  - Computer Science
  - Math
  - Statistical Physics
  - Economics
  - ...

- What have we learned?
Questions

- What can we do with network data?
- What can they tell us about the form and function of the system the network represents?
- What properties of the network systems can we measure or model?
- How are these properties related with the practical issues we care about?
Reasoning about networks

How do we reason about networks?
- **Empirical**: Study network data to find organizational principles
- **Mathematical models**: Probabilistic, graph theory
- **Algorithms**: analyzing graphs

What do we hope to achieve from studying networks?
- Patterns and statistical properties of network data
- Design principles and models
- Understand why networks are organized the way they are
  - Predict behavior of networked systems
What do we study in networks?

- **Structure and evolution**
  - What is the structure of a network?
  - Why and how did it become to have such structure?

- **Processes and dynamics**
  - How do information disseminate?
  - How do diseases spread?
Properties of networks

- Network theory has developed a large number of tools that can be used to describe and understand networks
  - Centrality: quantification of the importance of a vertex (or even an edge)
    - Various definitions that capture different aspects and can be useful in different contexts
      - Degree, eigenvector, Katz, PageRank etc.
  - Geodesic distance: minimum number of edges one would have to traverse in order to get from one vertex to the other
    - Implications on how fast things travel in the network
Properties of networks

- **Network theory includes also a number of concepts of practical importance**
  - The notion of hubs
    - A small number of vertices with extremely high degree
    - What are their implications in networks?
  - Small-world effect
    - On average geodesic distances are much smaller compared to the size of the network
    - Repercussions with regards to information diffusion
  - Communities in networks
    - The way a network breaks to communities might reveal information for the network (e.g., an organization) that are not easy to see without network data
The tale of networks

- Seven bridges of Konigsberg
  - Konigsberg was built on the banks of the river Pregel and on two islands that lie midstream
  - Seven bridges connected the land masses
  - Does there exist any walking route that crosses all seven bridges exactly once?
  - Leonhard Euler mapped the problem to a graph and proved that this is impossible
    - Foundations of graph theory
The tale of networks

- **Random networks**
  - Erdos extensively studied networks that form randomly
    - Two vertices connect uniformly at random
  - Erdos realized that if networks develop randomly, they are highly efficient
    - Even with few connections on average per link, the network can be connected with small paths
  - Erdos laid the foundations modern network theory
The tale of networks

- The Milgram Small World Experiment: Six degrees of separation
  - Randomly selected people living in Wichita, Kansas and Omaha, Nebraska
  - They were asked to get a letter to a stockbroker in Boston they had never met
    - Only the name and the occupation of the person in Boston was revealed
  - A large portion of these mails never reached the destination, but those who did reached it in a few hops
    - Funneling was also observed
    - Small paths exist, but it is hard to find them in a decentralized way
The tale of networks

- **The strength of weak ties**
  - Granovetter studied the way that people find their jobs in communities around Boston.
  - He found that in most of the cases it was not the closely related people that played a key role but rather some acquaintance.
  - Our friends have more friends than we do and hence it is most probable that they will be of help.
    - Also we are most probably exposed to the same information with our very close friends.
  - Social capital.
The tale of networks

- Syncing behavior
  - Watts and Strogatz started wondering about problems such as “How can Malaysian fireflies sync their behavior as if they were one giant organism?”
    - Is there a leader?
    - In what manner does the information travel across thousands or even millions of entities?
  - Watts strongly believed there is a strong connection with the six degrees phenomenon
    - Information seems to have an ability to travel across large populations fast
The tale of networks

- Alpha model
  - How small world networks appear?
  - What are the rules that people follow when making friends?
    - People introduce their friends to each other
      - The more common neighbors two vertices share, the more probable they are to connect

\[
y = \text{probability of connecting } u \text{ & } v
\]

\[
y \sim p + (x/N)^\alpha
\]

*default* probability \( p \)

\[x = \text{number of current common neighbors of } u \text{ & } v\]

\[N = \text{network size}\]
The tale of networks

- **Beta model**
  - Alpha model showed that small world networks are possible
  - What is the meaning of parameter $\alpha$?
  - Watts and Strogatz developed an even simpler model
    - You start from a regular lattice and you rewire edges
      - From order to randomness
      - People combine geographically contained connections with a few long-distance relations
The tale of networks

- The fitness model
  - Albert-Laszlo Barabasi and Reka Albert, realized that networks grow, and as they grow the nodes with more links get the bulk of the new links
    - Preferential attachment
    - As networks evolve hubs are formed
    - Power law
  - Later Barabasi capitalized on the similarities that these models had with the Bose-Einstein equation
    - Along with preferential attachment he introduced the attractiveness of a vertex
The future of the networks

- While most of the publicity of networks comes from social network analysis their applicability is virtually limitless.

- Network theory is applied in finance, marketing, medicine, criminology, intelligence agencies...

- Just a sample:
  
In this class we study

- Metrics for describing the structure of a network
- Random graph models
- Network formation models
- Processes in networks
- Combination of math and hands-on experience with software used for network analysis
  - Hands-on experience will be at an individual’s level effort
Course details

● **Highly technical course**
  - Linear algebra and probability knowledge is a minimum requirement

● **Some topics will be presented at a higher level**

● **Grade will be based on:**
  - Homework (30%)
  - Midterm (30%)
  - Final (30%)
  - Participation (10%)

http://www.sis.pitt.edu/~kpele/telcom2125_spring15.html
& Courseweb
Textbook

- There is one main book we will be using

- Other possible references
  - Research papers (pointers will be provided)
Part 0.1: Review on Probability Theory
Fundamentals of probabilities

- **Sample space** \( \Omega \): Set of all possible outcomes
- **Event space** \( F = 2^\Omega \) (an event is a subset of the sample space)
- **Probability measure**: function \( P : F \rightarrow \mathbb{R} \) such that:
  - \( P(A) \geq 0, \ \forall A \in F \)
  - \( P(\Omega) = 1 \)
  - For disjoint events \( A_i \), \( P(\bigcup_i A_i) = \sum_i P(A_i) \)
Example

- Consider throwing a dice twice:
  - Sample space $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$
  - Event space $F = 2^\Omega$
    - Sample events: let $A$ be the event that the sum is even and let $B$ be the event that we roll at least one 6
  - Probability measure: function $P$ is simple counting in this discrete case
    - $P(A) = 18/36=0.5$
    - $P(B) = 11/36$
For every two events A and B, the union of the two (“A or B”) is:

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

- E.g. continuing are above example

\[ P(A \cup B) = \frac{18}{36} + \frac{11}{36} - \frac{5}{36} = \frac{24}{36} = \frac{2}{3} \]
Conditional probability

Let $A$ and $B$ be two events. Then the conditional probability of $A$ given $B$ is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

“What’s the probability of $A$ once we know $B$ has happened?”

Rewriting gives us the useful product rule:

$$P(A \cap B) = P(A \mid B)P(B)$$
Independence

Two events are independent if:

\[ P(A \cap B) = P(A)P(B) \]

Equivalently: \[ P(A | B) = P(A) \quad \text{and} \quad P(B | A) = P(B) \]

Intuitively, knowing A doesn’t tell you anything about B and vice-versa.

But beware of relying on your intuition: rolling two dice (\(x_a\) and \(x_b\)), events \(x_a=2\) and \(x_a+x_b=k\) are independent if \(k=7\) and dependent otherwise!
Union bound

Recall that for any two events

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

If we are trying to upper bound the probability that A or B happens, the worst case is that A and B are disjoint (mutually exclusive) and hence, \( P(A \cap B) = 0 \)

The very useful union bound states: Let \( A_i \) be some events (not necessarily independent). Then:

\[ P\left( \bigcup_i A_i \right) \leq \sum_i P(A_i) \]
Bayes’ rule

This is the most important rule of probability!

For two events $A$ and $B$ (with $P(B)>0$):

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Often used to update beliefs:

posterior = “support B provides for A” x “prior”
Example

If a person has malaria, there is a 90% chance that his/her test results are positive. However the test result are not very correct; there is a chance for 1% false positive. Also only 1% of the total population gets affected by Malaria. Now one person's test result came out as Positive. What’s the odds that he will actually have Malaria?

M: event that person has malaria
B: event that blood test is positive

\[ P(B|M) = 0.9 \]
\[ P(B|\sim M) = 0.01 \]
\[ P(M) = 0.01 \]
Example

Bayes’ rule:

\[ P(M \mid B) = \frac{P(B \mid M)P(M)}{P(B)} \]

What about \( P(B) \)? **Law of total probability**

\[ P(B) = P(B \mid M)P(M) + P(B \mid \sim M)P(\sim M) \]

Substituting we finally get: \( P(M \mid B) = 0.476 \)

How does this result change if you use a superior blood test that has only 0.1% false positives?
Random variables

- A random variable is technically a real-valued function defined on the sample space $X : \Omega \to \mathbb{R}$
- Probabilities of random variables come from underlying $P$ function: $P(X = k) = P(\{\omega \in \Omega \mid X(\omega) = k\})$
  - It is called random variable because it is a variable that does not take on a single – deterministic – value, but it can take on a set of different values, each with an associated probability
  - Let $X$ be a random variable (r.v.) that counts the number of 6’s we roll in a 2 dice rolls
    - $P(X=2)=P(\{6,6\}) = 1/36$
    - $P(X=1)=P(\{1,6\})+P(\{2,6\})+P(\{3,6\})+\ldots+P(\{5,6\})+P(\{6,1\})+\ldots=10/36$
    - $P(X=0)=?$
A probability mass function (pmf) assigns a probability to each possible value of a random variable (in the discrete case).

Example: “funny dice”
Distributions

Distribution over sum of two dice rolls

$p(S)$

$S$

\[\begin{array}{cccccccccc}
2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\frac{1}{36} & \frac{1}{18} & \frac{1}{12} & \frac{1}{9} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\
\end{array}\]
Probability density functions

The PDF of a continuous random variable $X$ describes the relative likelihood for $X$ to take on a given value:

$$ P(a \leq X \leq b) = \int_a^b f(x) \, dx $$
The CDF of a random variable \( X \) is:

\[
F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y) \, dy
\]
Properties of distribution functions

CDF
\[ 0 \leq F_X(x) \leq 1 \]

\( F_X \) monotone increasing, with \( \lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1 \)

PMF
\[ 0 \leq p_X(x) \leq 1 \]
\[ \sum_x p_X(x) = 1 \]
\[ \sum_{x \in A} p_X(x) = p_X(A) \]

PDF
\[ f_X(x) \geq 0 \]
\[ \int_{-\infty}^{\infty} f_X(x) \, dx = 1 \]
\[ \int_{x \in A} f_X(x) \, dx = P(X \in A) \]
Some common random variables

\(\sim \text{ Bernoulli}(p) \ (0 \leq p \leq 1): \ p_X(x) = \begin{cases} 
  p & x = 1, \\
  1 - p & x = 0.
\end{cases} \)

\(\sim \text{ Geometric}(p) \ (0 \leq p \leq 1): \ p_X(x) = p(1 - p)^{x-1} \)

\(\sim \text{ Uniform}(a, b) \ (a < b): \ f_X(x) = \begin{cases} 
  \frac{1}{b-a} & a \leq x \leq b, \\
  0 & \text{otherwise}.
\end{cases} \)

\(\sim \text{ Normal}(\mu, \sigma^2): \ f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \)
**Expectation and variance**

- If the discrete random variable $X$ has pmf $p(x)$, then the expected value of $X$ is: $E[X] = \sum_x xp(x)$
- For a continuous r.v. we have a similar result: $E[X] = \int_{-\infty}^{\infty} xf(x) \, dx$
- Expectation is linear:
  \[ \forall a \in \mathbb{R} \Rightarrow E[a] = a \]
  \[ E[ag(X) + bh(X)] = aE[g(X)] + bE[h(X)] \]
- $Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$
  - Variance is not linear

What is the expectation of a rolling dice?
Indicator variables

- An indicator variable just indicates whether an event occurs or not

\[ I_A = \begin{cases} 
1, & \text{if } A \text{ occurs} \\
0, & \text{otherwise} 
\end{cases} \]

- They have a very useful property:

\[ E[I_A] = 1P(I_A = 1) + 0P(I_A = 0) = P(I_A = 1) = P(A) \]
Method of indicators

- Suppose $X$ is the number of events that occur among a collection of events $A_1, A_2, \ldots, A_n$. Let $X_1, X_2, \ldots, X_n$ be indicator variables for these events. Then,

$$X = \sum_{i=1}^{n} X_i \implies E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P(A_i)$$

- When to use: Useful when we need to find $E[X]$ for a counting variable $X$, especially when we can break it down into counting the occurrences of events in a collection $A_1, A_2, \ldots, A_n$, where the probability $P(A_i)$ is easy to compute.
  - Note: The events $A_i$ do not need to be independent!
Example: N professors are at a dinner and take a random coat when they leave. What is the expected number of professors with the right coat?

Let $G$ be the number of professors that get the right coat, and let $G_i$ be an indicator function for the event that professor $i$ gets his own coat. Then: $G=G_1+G_2+\ldots+G_n$

$$E[G] = E[G_1 + G_2 + \ldots + G_n] =$$

$$= E[G_1] + E[G_2] + \ldots + E[G_n] =$$

$$= \frac{1}{n} + \frac{1}{n} + \ldots + \frac{1}{n} = 1$$

Linearity of expectations does not assume independence!
Some useful inequalities

- **Markov’s inequality**
  - $X$ r.v. and $a > 0$
    $$P(X \geq a) \leq \frac{E[X]}{a}$$
  - Continuing our previous example we can see that the probability of at least 5 professors get the right coats is no higher than 20%
    - Regardless of $N$

- **Chebyshev’s inequality**
  - $X$ r.v. with finite, non-zero variance $\sigma^2$. For any real $k > 0$
    $$P(|X - E[X]| \geq k\sigma) \leq \frac{1}{k^2}$$
Chernoff bound

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli with $P(X_i=1)=p_i$

- With $\mu = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} p_i$

$P\left(\sum_{i=1}^{n} X_i \geq (1+\delta)\mu\right) \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu$, $\forall \delta$

- Multiple variants of Chernoff bounds exist, which can be useful in different settings
Parameter estimation: Maximum likelihood

- Assume we have a parametrized distribution $f_X(x; \theta)$ and we do not know parameter $\theta$ (could be a vector as well)
- We observe IID samples $x_1, x_2, \ldots, x_n$
- Goal: Estimate $\theta$
- The maximum likelihood estimator (MLE) is the value $\hat{\theta}$ that maximizes the likelihood of observing the data/samples you observed
MLE example

You flip a coin with unknown bias $p$ of landing heads $n$ times and get $n_H$ heads and $n_T$ tails. What is the MLE estimate for the coin’s bias?

The likelihood of observing the data given a particular $\theta$ is:

$$P(D \mid \theta) = \theta^{n_H} (1 - \theta)^{n_T}$$

Usually we work with the log-likelihood. That is,

$$\log(P(D \mid \theta)) = n_H \log \theta + n_T \log(1 - \theta)$$
MLE example

• We want to maximize this last expression (which also maximizes the original likelihood – why?)

We take the derivative of the log-likelihood with respect to $\theta$ and set it equal to zero

$$\frac{d}{d\theta} \log(P(D | \theta)) = 0 \Rightarrow$$

$$\frac{d}{d\theta} \left[ n_H \log \theta + n_T \log(1 - \theta) \right] = 0 \Rightarrow$$

$$\frac{n_H}{\theta} - \frac{n_T}{1 - \theta} = 0 \Rightarrow$$

$$\hat{\theta} = \frac{n_H}{n_H + n_T}$$

Closed form solutions are not always possible!
Part 0.2: Review on Linear Algebra
Matrices and vectors

- **Matrix** is a rectangular array of numbers, e.g., \( A \in \mathbb{R}^{m \times n} \)

\[
A = \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{pmatrix}
\]

- **Vector**: a matrix consisting of only one column, e.g., \( x \in \mathbb{R}^n \)

\[
x = \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}
\]
Matrix multiplication

- If $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}, C = AB$ then $C \in \mathbb{R}^{m \times p}$

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- Special cases
  - Matrix – vector multiplication
  - Inner produce of two vectors
    - E.g., $x, y \in \mathbb{R}^{n}$

$$x^T y = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}$$
Matrix multiplication
Properties of matrix multiplication & Operators

- **Associative**: \((AB)C = A(BC)\)
- **Distributive**: \(A(B+C) = AB + AC\)
- **Non commutative**: \(AB \neq BA\)

- **Transpose**: \(A \in \mathbb{R}^{m \times n}\text{ then } A^T \in \mathbb{R}^{n \times m}: (A^T)_{ij} = A_{ji}\)

- **Properties**
  - \((A^T)^T = A\)
  - \((AB)^T = B^T A^T\)
  - \((A+B)^T = A^T + B^T\)
Identity matrix

- **Identity matrix:** \( I_n \in \mathbb{R}^{n \times n} \)

\[
I_n = \begin{cases} 
  1, & i = j \\
  0, & \text{otherwise}
\end{cases}
\]

\[
\forall A \in \mathbb{R}^{m \times n} : AI_n = I_i A = A
\]

\[
I_1 = [1], \quad I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \cdots, \quad I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}
\]
Diagonal matrix

Diagonal matrix: \( D = \text{diag}(d_1, d_2, \ldots, d_n) \)

\[
D_{ij} = \begin{cases} 
  d_i, & i = j \\
  0, & \text{otherwise}
\end{cases}
\]

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 4 & 0 \\
0 & 0 & -3
\end{bmatrix}
\]
Other special matrices

- Symmetric matrices: a square matrix $A$ is symmetric if $A = A^T$

- Orthogonal matrix: a square matrix $U$ is orthogonal if $UU^T = U^T U = I$
Linear independence and rank

- A set of vectors \( \{x_1, x_2, \ldots, x_n\} \) is linearly independent if
  \[
  \#\{\alpha_1, \ldots, \alpha_n\} : \sum_{i=1}^{n} \alpha_i x_i = 0
  \]

- Rank of an \( mxn \) matrix \( A \) is the maximum number of linearly independent columns (or equivalently, rows)

- Properties:
  - \( \text{rank}(A) \leq \min(m,n) \)
  - \( \text{rank}(A) = \text{rank}(A^T) \)
  - \( \text{rank}(AB) \leq \min\{\text{rank}(A),\text{rank}(B)\} \)
  - \( \text{rank}(A+B) \leq \text{rank}(A) + \text{rank}(B) \)
Matrix inversion

- If $A$ is an $n \times n$ matrix with $\text{rank}(A) = n$, then the inverse of $A$, denoted with $A^{-1}$ is the matrix that: $AA^{-1} = A^{-1}A = I$

- Properties
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^T = (A^T)^{-1}$

- The inverse of an orthogonal matrix is its transpose
Eigenvalues and eigenvectors

- Consider a real matrix $n \times n$. $\lambda \in C$ is an eigenvalue of $A$ with corresponding eigenvector $x \in C^n (x \neq 0)$ if

$$Ax = \lambda x$$

- Eigenvalues: the $n$ possibly complex roots of the (characteristic) polynomial equation $\det(A-\lambda I)=0$

$$Ax = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 3 + 1 \cdot (-3) \\ 1 \cdot 3 + 2 \cdot (-3) \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1 \cdot \begin{bmatrix} 3 \\ -3 \end{bmatrix}.$$
Eigenvalue and eigenvector properties

- Usually eigenvectors are normalized to unit length
- If $A$ is symmetric, then all the eigenvalues are real and the eigenvectors are orthogonal to each other
  - Their inner product is zero
- $tr(A) = \sum_{i=1}^{n} \lambda_i$
- $\det(A) = \prod_{i=1}^{n} \lambda_i$
- $rank(A) = \left| \{1 \leq i \leq n \mid \lambda_i \neq 0 \} \right|$
Matrix eigendecomposition

- Consider an $k \times k$ matrix $A$, with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ and eigenvectors $x_1, x_2, \ldots, x_k$
- Also $P = [x_1 | x_2 | \ldots | x_k]$ and $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$. Then:

$$AP = A \begin{bmatrix} X_1 & X_2 & \cdots & X_k \end{bmatrix} = \begin{bmatrix} AX_1 & AX_2 & \cdots & AX_k \end{bmatrix} = \begin{bmatrix} \lambda_1 X_1 & \lambda_2 X_2 & \cdots & \lambda_k X_k \end{bmatrix} = \begin{bmatrix} \lambda_1 x_{11} & \lambda_2 x_{21} & \cdots & \lambda_k x_{k1} \\ \lambda_1 x_{12} & \lambda_2 x_{22} & \cdots & \lambda_k x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1 x_{1k} & \lambda_2 x_{2k} & \cdots & \lambda_k x_{kk} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{ kk} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} = PD,$$
Matrix eigendecomposition

- Thus, $A = PDP^{-1}$

- Furthermore, by induction we can show that: $A^n = PD^nP^{-1}$

- The above matrix diagonalization is a special case of Singular Value Decomposition
Singular Value Decomposition

- Singular Value Decomposition (SVD) is typically used for dimensionality reduction.
- For instance let’s consider the following matrix that represents some dataset:

\[
A = \begin{pmatrix}
1 & 2 & 1 \\
-2 & -3 & 1 \\
3 & 5 & 0
\end{pmatrix}
\]

- The rank of this matrix is 2 (why?)
- Why is “low” rank interesting?
  - Each row of A is expressed as a 3-dimensional vector with the standard base vectors [1 0 0], [0 1 0] and [0 0 1]
  - However, the two linearly independent rows of A can form a new basis (i.e., [1 2 1] and [-2 -3 1]
  - Now each row has new coordinates [1 0], [0 1] and [1 -1] of lower dimensionality

\[\text{Rank} = \text{“dimensionality”}\]
Singular Value Decomposition

\[ A_{m \times n} = U_{m \times r} \Sigma_{r \times r} (V_{n \times r})^T \]

- **A**: Input (data) matrix
  - m x n matrix (e.g., m users, n products)
- **U**: Left singular vectors
  - m x r matrix (e.g., m users, n ‘concepts’)
- **\( \Sigma \)**: Singular values
  - r x r diagonal matrix (strength of each ‘concept’)
  - (r: rank of matrix A)
- **V**: Right singular vectors
  - n x r matrix (n products, r ‘concepts’)

Singular Value Decomposition

\[ A \approx U\Sigma V^T = \sum_i \sigma_i \vec{u}_i \circ \vec{v}_i^T \]
Singular Value Decomposition

\[ A \approx U \Sigma V^T = \sum_i \sigma_i \tilde{u}_i \cdot \tilde{v}_i^T \]

- \( \sigma_i \): scalar
- \( \tilde{u}_i \): vector
- \( \tilde{v}_i \): vector
SVD Theorem

- It is **always** possible to decompose a real matrix $A$ into $A = U\Sigma V^T$, where:
  - $U, \Sigma, V$: unique
  - $U, V$: column orthonormal
    - $U^TU = I$ and $V^TV = I$
  - $\Sigma$: diagonal
    - Entries are positive and sorted in decreasing order ($\sigma_{11} \geq \sigma_{22} \geq \ldots \geq 0$)
Best low rank approximation

How do we do dimensionality reduction?
- By setting the $k$ smallest singular values to 0
  - This essentially removes the contribution of the corresponding left and right singular vectors as well

$A = U \Sigma V^T$

$B$ is the best approximation of $A$ (in the sense of the Frobenius norm)
Theorem:
Let $A = U\Sigma V^T$ where $\Sigma = \sigma_1 \geq \sigma_2 \geq \ldots$, and rank($A$)=r. Then $B=USV^T$ is the best rank-$k$ approximation to $A$ where: $S$ diagonal $r \times r$ matrix where $s_i = \sigma_i$ ($i=1,\ldots,k$) else $s_i = 0$

What does “best” mean?
- $B$ is a solution to: $\min_B \|A - B\|_F$ where rank($B$) = $k$

How many singular values to keep?
- Heuristic: keep 80%-90% of the energy ($= \sum \sigma_i^2$)
Relationship between SVD & eigendecomposition

- SVD gives: $A = U\Sigma V^T$
- Eigendecomposition gives: $A = PDP^{-1}$
  - $A$ needs to be symmetric for eigendecomposition
- $U$, $V$, $P$ are orthonormal
- $D$ and $\Sigma$ are diagonal

$$AA^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma V^T (V\Sigma^T U^T) = U\Sigma \Sigma^T U^T$$
$$A^T A = V\Sigma^T U^T (U\Sigma V^T) = V\Sigma \Sigma^T V^T$$

The above equations show how to compute the SVD of a matrix through eigendecomposition

$$D = \Sigma \Sigma^T, \text{ U is the eigenvectors of } AA^T$$
$$\text{ and V is the eigenvectors of } A^T A$$