# Network Theory and Applications to Genomics

*Structural Graph Theory*

Simone Severini & Andrew Teschendorff

CoMPLEX

May 15, 2012

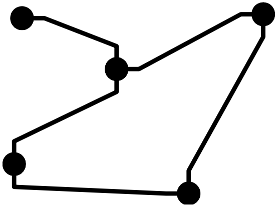
# General references


- Reinhard Diestel, *Graph Theory*, Graduate Texts in Mathematics, Springer-Verlag, Heidelberg, 2010. http://diestel-graph-theory.com/
- Béla Bollobás, *Modern graph theory*, Graduate Texts in Mathematics, Springer, New York, 1998.

# Graphs

### Definition

A **graph** $G = (V, E)$ is a pair of two sets: the elements of $V$ are called **vertices**; $E \subseteq V \times V - \{\{i, i\} : i \in V\}\}$ and its elements are called **edges**.

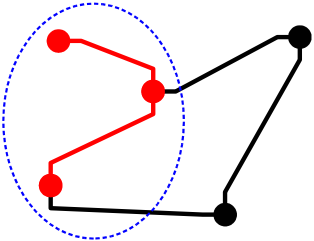


- Generalizations: **digraphs** (directed graphs); **weighted graphs**.

# Subgraphs

### Definition

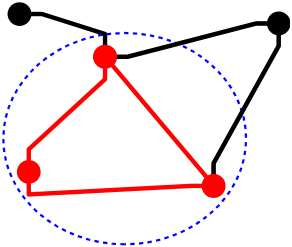
A graph $H = (W, F)$ is a **subgraph** of a graph $G = (V, E)$ if $W \subseteq V$ and $F \subseteq E$.



# Induced subgraphs

## Definition

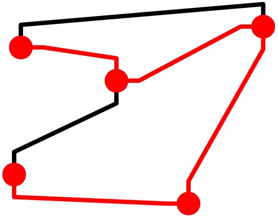
A graph $H = (W, F)$ is said to be an **induced subgraph** of a graph $G = (V, E)$ if: *(i)* $H$ is a subgraph of $G$; *(ii)* if $i, j \in V, W$ then $\{i, j\} \in F$ iff $\{i, j\} \in E$.



# Spanning subgraphs

## Definition

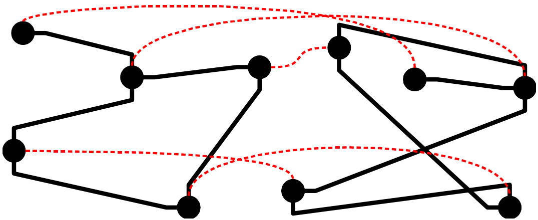
A graph $H = (W, F)$ is said to be a **spanning subgraph** of a graph $G = (V, E)$ if: *(i)* $H$ is a subgraph of $G$; *(ii)* if $W = V$.



# Isomorphism

## Definition

Two graphs $G = (V, E)$ and $H = (W, F)$ are **isomorphic** ($G \cong H$) if there is a bijection $f : V \longrightarrow W$ such that $\{i, j\} \in E$ iff $\{f(i), f(j)\} \in F$. The number of isomorphism classes on $n$ vertices is at most $(1 + o(1))\, 2^{n(n-1)/2} / n!$.



## Definition

A map on graph is called **a (complete) graph invariant** (iff) if it assigns equal values to isomorphic graphs.

# Degrees

### Definition

The **neighbourhood** of a vertex $i$ is $N(i) := \{j : \{i, j\} \in E\}$. The nonnegative integer $d_G(i) = |N(i)|$ is the **degree** of a vertex $i$. The **maximum degree** is denoted by $\Delta(G)$; the **minimum degree** by $\delta(G)$.

### Lemma (Degree sum formula)

*In a graph $G$, $\sum_{i \in V} d_G(i) = 2|E|$.*

### Lemma

*In any graph, there are two vertices of equal degree.*

### Lemma (Handshaking lemma)

*The number of vertices of odd degree in a graph is even.*

### Definition

The **average degree** of a graph $G$ on $n$ vertices is $\langle d_G \rangle = 2|E|/n$.

# Degree distributions

### Definition

The **degree sequence** of a graph $G$ is the list of its degrees $\{d_G(i) : i \in V\}$.

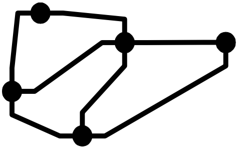
### Definition

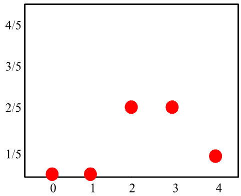
The **degree distribution** of a graph $G$ on $n$ vertices is the list $\{P_G(k) : k = 0, 1, ..., n-1\}$, where $P_G(k) = n_k / n$ and $n_k = |\{i : d_G(i) = k\}|$.

### Definition

A **unigraph** is a graph characterized by its degree sequence.

# Degree distributions: example

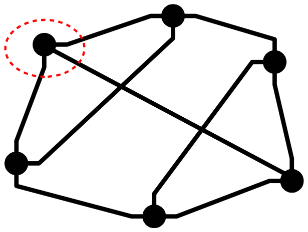




# Regular graphs

## Definition

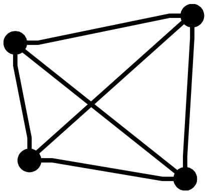
A graph $G$ is **k-regular** if $d_G(i) = k$ for every vertex $i$.



# Complete graphs

## Definition

A **complete graph** $K_n$ is an $(n-1)$-regular graph on $n$ vertices.



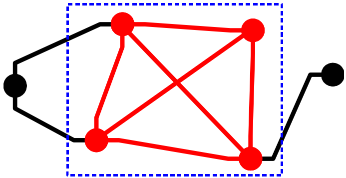
# Cliques

### Definition

A **clique** is a complete induced subgraph.

### Definition

The **clique number** $\omega(G)$ of a graph $G$ is the order of its largest clique.



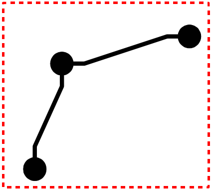
# Paths and cycles

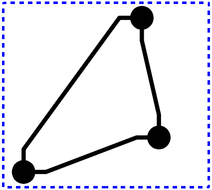
## Definition

A **path** $P_n = (V, E)$ of **length** $n - 1$ is a graph of the form $V = \{1, 2, ..., n\}$ and $E = \{\{1, 2\}, \{2, 3\}, ..., \{n - 1, n\}\}$.

## Definition

A **cycle** $C_n = (V, E)$ of **length** $n$ is a graph of the form $V = \{1, 2, ..., n\}$ and $E = \{\{1, 2\}, \{2, 3\}, ..., \{n - 1, n\}, \{1, n\}\}$.

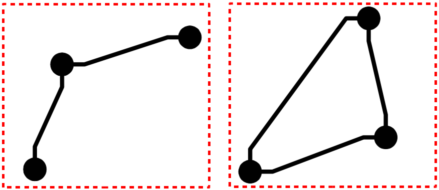




# Connectedness

## Definition
A graph is **connected** if there is a path between any two of its vertices.

## Definition
A maximal connected subgraph of a graph $G$ is said to be a **connected component**.



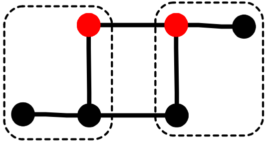
# Domination number

### Definition

A set $S \subseteq V$ is a **dominating set** in a graph $G = (V, E)$ if $N(S) \cup S = V$.

### Definition

The **domination number**, $\gamma(G)$, is the size of the smallest dominating set.



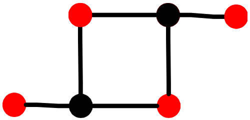
# Independence number

**Definition**

A set $S \subseteq V$ is an **independent set** in a graph $G = (V, E)$ if $\{i, j\} \notin E$ for every $i, j \in S$.

**Definition**

The **independence number**, $\alpha(G)$, is the size of the largest independent set.



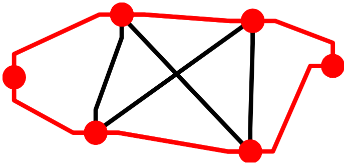
# Hamiltonicity

## Definition

A **Hamilton cycle** in a graph $G$ is a cycle containing all vertices of $G$. A graph is said to be **Hamiltonian** if it has an Hamilton cycle.

## Theorem (Dirac, 1952)

*Every graph on $n \geq 3$ vertices and minimum degree at least $\lfloor n/2 \rfloor$ is Hamiltonian.*



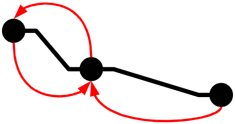
# Walks and tours

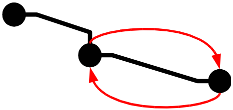
## Definition

A **walk** in a graph is an alternating sequences of vertices and edges such that every two (possibly identical) edges adjacent in the sequence are incident with the same vertex.

## Definition

A **tour** in a graph is walk in which the initial and final vertex coincide.





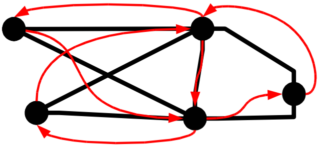
# Eulerian graphs

## Definition

An **Euler tour** in a graph is a tour that includes every edge exactly once.
An **Eulerian graph** is a graph with an Euler tour.

## Theorem (Euler, 1736)

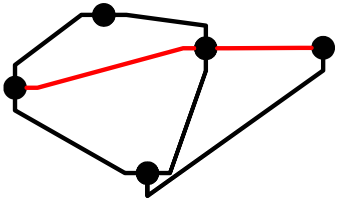
*A connected graph is* **Eulerian** *if and only if the degree of every vertex is even.*



# Diameter

## Definition
The **distance** $d_G(i,j)$ in $G$ of two vertices $i,j$ is the length of a shortest path between $i$ and $j$ in $G$.

## Definition
The greatest distance, $diam(G)$, between any two vertices in $G$ is the **diameter** of $G$.



# Clustering coefficient

## Definition

The **neighbourhood** of a vertex $i$ is the set $N_G(i) = \{j : \{i,j\} \in V\}$.
The neighbourhood of a set $S$ is $N_G(S) = \{j : \{i,j\} \in V \text{ and } i \in S\}$.

## Definition

The **clustering coefficient** of a vertex $i$ in a graph $G$ is

$$C_G(i) = \frac{2\left|\{\{k,l\} \in E(G) : k,l \in N_G(i)\}\right|}{d_G(i)^2 - d_G(i)}.$$

## Definition

The **average clustering coefficient** of a graph $G$ is

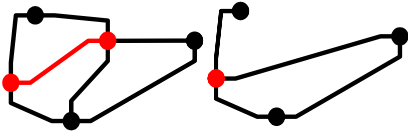
$$\langle C_G \rangle = \frac{1}{n} \sum_{i=1}^{n} C_G(i).$$

- D. J. Watts, S. H. Strogatz, *Nature* **393** (1998) 440.

# Minors

## Definition

An **edge contraction** is an operation which removes an edge while simultaneously merging together its two vertices.



## Definition

A graph $H$ is called a **minor** of the graph $G$ if $H$ is isomorphic to a graph that can be obtained by zero or more edge contractions on a subgraph of G.

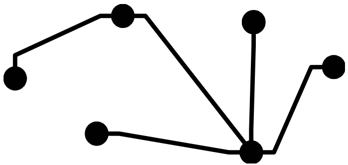
# Trees

## Definition

A **tree** is a graph without cycles as subgraphs. A tree has $n$ vertices and $n - 1$ edges.

## Lemma

*There is a unique path between any two vertices of a tree.*



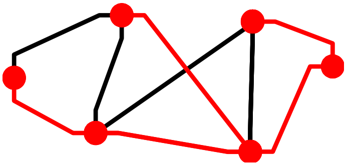
# Spanning trees

## Definition

A **spanning tree** is a spanning subgraph which is also a tree.

## Lemma

*Every graph contains a spanning tree.*



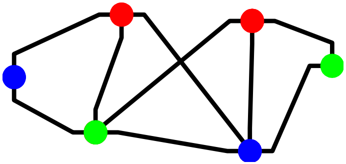
# Chromatic number

## Definition

The **chromatic number** $\chi(G)$ of a graph $G$ is the minimum number of colours for which every two adjacent vertices have a different colour.

## Lemma

$\omega(G) \leq \chi(G) \leq \Delta(G+1)$.



# Girth

## Definition

The **girth** of a graph is the length of the shortest cycle.

## Theorem (Erdős, 1959)

*For any $k$ and $l$, there is a graph $G$ such that $\chi(G) > k$ and girth $> l$.*

- The **probabilistic method** is a collection of techniques used to prove the existence of a prescribed mathematical object. It works by showing that if we choose an object from a specified class, there is a nonzero probability that the result is the desired one.

# Bipartite graphs

## Definition

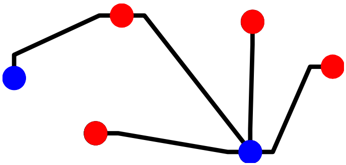
A graph $G = (V, E)$ is **bipartite** if $\chi(G) = 2$.

## Lemma

*A graph is bipartite iff it does not contain cycles of odd length.*

## Lemma

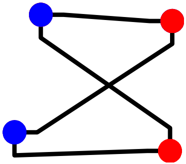
*A bipartite graph on $n$ vertices can have at most $\frac{1}{4}n^2$ edges.*



# Complete bipartite graphs

## Definition

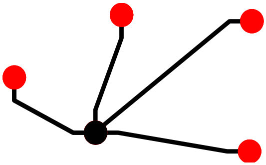
A graph $K_{n,m} = (V, E)$ is **complete bipartite** if: *(i)* $V = A \cup B$, with $|A| = n$ and $|B| = m$; *(ii)* $\{i, j\} \in E$ for every $i \in A$ and $j \in B$.



# Star graphs

## Definition

A **star (graph)** on $n$ vertices is a graph $K_{1,n-1}$.



# Perfect graphs

## Definition

A graph $G$ is **perfect** if $\omega(H) = \chi(H)$, for every induced subgraph $H$ of $G$.

## Definition

The **complement** of a graph $G = (V, E)$ is the graph $\overline{G} = (W, F)$, where *(i)* $W = V$ and *(ii)* for every $i, j \in W$, $\{i, j\} \in F$ iff $\{i, j\} \notin E$.

## Theorem (Chudnovsky-Robertson-Seymour-Thomas, 2002)

*A graph $G$ is perfect iff neither $G$ nor $\overline{G}$ contains a cycle of odd length at least 5 as an induced subgraph.*

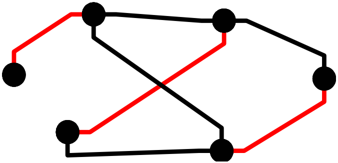
# Matchings

## Definition

A **matching** in a graph is set of disjoint edges. A matching is **perfect** if it is spanning.

## Theorem (Hall, 1935)

*A bipartite graph* $G = (V = A \cup B, E)$ *has a matching of* $A$ *iff* $|N_G(S)| \geq |S|$ *for every* $S \subseteq A$*.*



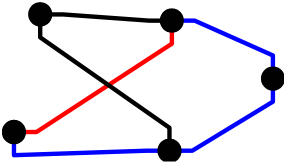
# Edge cuts

## Definition

An **edge cut** of two vertices $i$ and $j$ in a connected graph is the minimum number of edges who removal gives two connected components including $i$ and $j$, respectively.

## Theorem (Menger, 1927)

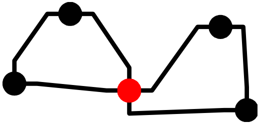
*The size of the edge cut of two vertices $i$ and $j$ is equal to the maximum number of pairwise edge-disjoint paths between $i$ to $j$.*



# Vertex/edge connectivity

## Definition

The **vertex (edge) connectivity** of a graph $G$, $\nu(G)$ $(\eta(G))$, is the number of vertices (resp. edge) that we need to delete in order to increase the number of connected components of $G$.



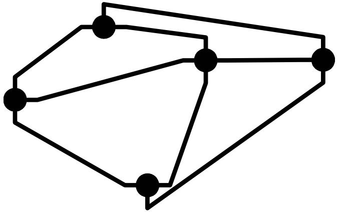
# Genus

## Definition

The **genus** of a graph is the smallest integer $g$ such that the graph can be drawn on a sphere with $g$ handles without edge intersections.

## Definition

A graph is **planar** if its genus is zero.



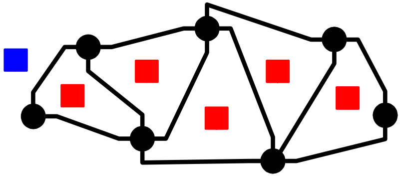
# Euler's formula

## Definition

A **face** of a planar graph is a region bounded by edges, including the outer one.

## Theorem

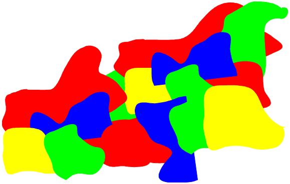
*In a planar graph $G$, we have $V - E + f = 2$, where $f$ is the number of faces of $G$.*



# Four colour theorem

Theorem (Appel-Haken, 1976)

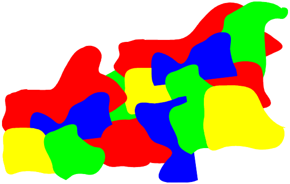
*If a graph $G$ is planar then $\chi(G) \leq 4$.*



# Thickness

## Definition

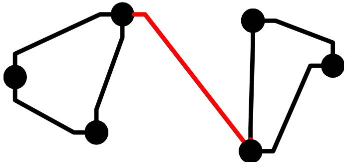
The **thickness** $t(G)$ of a graph $G$ is the minimum number of planar subgraphs $P_i$ in $G$ such that $\bigcup P_i = G$.



# Bridges

## Definition

A **bridge** (or **bottleneck**) in a graph is an edge whose removal increases the numbe of connected component.A graph is **bridgeless** if it does not contain a bridge.



# Graph products

## Definition

A **graph product** is an operation that takes two graphs $G_1$ and $G_2$ and produces a graph $H$ with the following properties: (i) $V(H) = V(G_1) \times V(G_2)$; (ii) $\{\{u_1, u_2\}, \{v_1, v_2\}\} \in E(H)$ if some conditions are satisfied.

## Definition

**Cartesian product**: $u_1 = v_1$ and $\{u_2, v_2\} \in E(G_2)$ or $\{u_1, v_1\} \in E(G_1)$ and $u_2 = v_2$.

## Definition

**Tensor product**: $\{u_1, v_1\} \in E(G_1)$ and $\{u_2, v_2\} \in E(G_2)$.

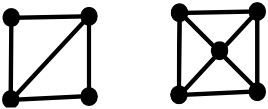
## Definition

**Strong product**: $u_1 = v_1$ and $\{u_2, v_2\} \in E(G_2)$ or $\{u_1, v_1\} \in E(G_1)$ and $u_2 = v_2$ or $\{u_1, v_1\} \in E(G_1)$ and $\{u_2, v_2\} \in E(G_2)$.

# Line graphs

## Definition

Given a graph $G$, its **line graph** $L(G)$ is a graph such that (i) each vertex of $L(G)$ represents an edge of $G$; and (ii) two vertices of $L(G)$ are adjacent if and only if their corresponding edges share a common vertex in $G$.



# Betweeness centrality

## Definition

The **betweeness centrality** of a vertex $i$ is

$$g(i) = \sum_{j \neq k} \sigma_{j,k}(i) / \sigma_{j,k},$$

where $\sigma_{j,k}$ is the total number of shortest paths between $j$ and $k$ and $\sigma_{j,k}(i)$ is the total number of shortest paths between $j$ and $k$ through $i$.

- L. C. Freeman, *Sociometry* **40**, 35 (1977).

# Assortativity

## Definition

The **assortativity coefficient** of a graph $G$ with $m$ edges is defined as

$$r(G) = \frac{m^{-1}\sum_{\{i,j\}\in E} d(i)\, d(j) - a}{m^{-1}\sum_{\{i,j\}\in E} \frac{1}{2}\left(d^2(i) + d^2(j)\right) - a},$$

where

$$a := \left[m^{-1}\sum_{\{i,j\}\in E} \frac{1}{2}\left(d(i) + d(j)\right)\right]^2 .$$

- The assortativity coefficient is a correlation coefficient between the degrees of all adjacent vertices. A positive assortativity coefficient indicates that vertices tend to be adjacent to other vertices with the same or similar degree.
- M. E. J. Newman, *Phys. Rev. Lett.* **89**, 208701 (2002).

# Regularity Lemma (1)

## Definition

Given a graph $G = (V, E)$, let $X, Y \subseteq V$ with $X \cap Y = \emptyset$. Let $\|X, Y\| = |\{\{i,j\} \in E : i \in X, j \in Y\}|$. The **density** of the pair $(X, Y)$ is defined as $d(X, Y) = \frac{\|X,Y\|}{|X||Y|}$.

## Definition

Given a graph $G = (V, E)$, let $A, B \subseteq V$ with $A \cap B = \emptyset$. Then $(A, B)$ is an $\epsilon$-**regular pair** if for all $X \subseteq A$ and $Y \subseteq B$ with $|X| \geq \epsilon |A|$ and $|Y| \geq \epsilon |B|$ we have $|d(X, Y) - d(A, B)| \leq \epsilon$.

## Definition

Let $\{V_0, V_1, ..., V_k\}$ be a partition of $V$. The set $V_0$ (possibly empty) is called an **exceptional set**. The partion is called an $\epsilon$-**regular partition** if: *(i)* $|V_0| \leq \epsilon |V|$; *(ii)* $|V_1| = ... = |V_k|$; *(iii)* all but at most $\epsilon k^2$ of the pairs $(V_i, V_j)$ with $1 \leq i < j \leq k$ are $\epsilon$-regular.

# Regularity Lemma (2)

## Theorem (Szemerédi, 1978)

*For every* $\epsilon > 0$ *and every integer* $n \geq 1$ *there exists an integer* $N$ *such that every graph on at least* $m$ *vertices has an* $\epsilon$*-regular partition* $\{V_0, V_1, ..., V_k\}$ *with* $n \leq k \leq N$*.*

- The intuition behind Szemerédi regularity lemma is that every large enough graph can be divided into subsets of about the same size so that the edges between different subsets behave almost randomly.
- see J. Komlós, A. Shokoufandeh, M. Simonovits, E. Szemerédi, *LNCS*, **2292** (2002), 84–112.

# Ramsey numbers

- **Ramsey Theory** studies conditions when a combinatorial object contains necessarily some smaller given objects.

## Definition

The **Ramsey number** $R(k;l)$, where $k$ and $l$ are arbitrary integers, is defined as the smallest $n$ such that for a graph $G$ with $n$ or more vertices $\omega(G) = k$ or $\alpha(G) = l$. (Ramsey's theorem guarantees this.)

## Example

In any group of six people, it is possible to find three persons who all know each other or three persons who do not know each, *i.e.*, $R(3,3) = 6$; $R(k,k) \leq \binom{2k-2}{k-1} \leq \frac{4^k}{\sqrt{k}}$.

- F. P. Ramsey, *Proceedings of the London Mathematical Society*, **30** (1930), pp. 264–286.

# Max-Flow Min-Cut

## Definition

In a digraph $D = (V, E)$, a vertex of in-degree 0 (resp. out-degree 0) is called a **source** (resp. **sink**). Each arc $(i, j)$ has a **capacity** $c_{i,j} \geq 0$. A **flow** is a map $f : E \longrightarrow \mathbb{R}^{\geq 0}$ such that: *(i)* $f(i, j) \leq c_{i,j}$; *(ii)* $\sum_j f(i, j) = \sum_j f(j, i)$, for every $i \in V$.

## Problem

*Maximize* $f_{tot} = \sum_j f(1, j) = \sum_j f(j, n)$, *where* $1$ *is the source and* $n$ *is the sink.*

## Definition

A **cut** in a digraph is a set of arcs such that if they are deleted, there is not path from the source to the sink (see edge cut). The **capacity** of a cut is defined to be the sum of the capacities of every arc in the cut.

## Theorem (Max-Flow Min-Cut)

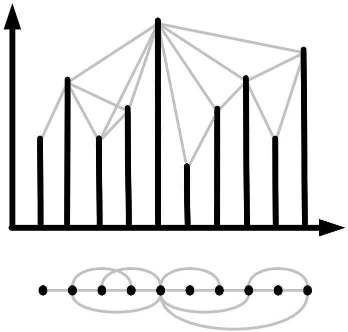
*In every digraph, the maximum flow equals the minimum capacity of a cut.*

- Ford–Fulkerson, Elias–Feinstein–Shannon, 1956.

# Visibility graphs

## Definition

Let $X = (x_1, x_2, ..., x_n)$ be a sequence with $x_i \in \mathbb{R}^{\geq 0}$ for $i = 1, 2, ..., n$. Each element $x_i \in X$ is associated to a line segment in $\mathbb{R}^2$ with end points $(i, 0)$ and $(i, x_i)$. The *visibility graph* of the sequence $X$, denoted by $G(X)$, is the graph with set of vertices $\{1, 2, ..., n\}$ and an edge $i \sim j$ if, for any $i < k < j$, we have $x_k < x_j + (x_i - x_j)\frac{j-k}{j-i}$.

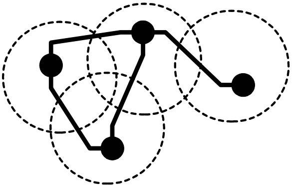


# Intersection graphs

## Definition

The **intersection graph** of a family of sets is a graph whose vertices are the sets; two vertices are adjacent if the sets have nonempty intersection.

- Every graph is an intersection graph.
- Interval graphs.
- Unit disk graphs



# Graph widths

- Most NP-hard problems are efficiently solvable for graphs with **bounded widths**. Treewidth quantifies how "close" a graph is to being a tree.

## Definition

A **tree decomposition** $(T, X)$ of a graph $G = (V, E)$ consists of a tree $T = (I, F)$ and a function $X : I \longrightarrow 2^V$ such that: (i) $\bigcup_{i \in I} X(i) = V$; (ii) for every $\{i, j\} \in E$ there is $u \in I$ such that $\{i, j\} \in X(u)$; (iii) $X(i) \cap X(k) \subseteq X(j)$, for every $j$ on the path between $i$ and $k$. The **width** of a tree decomposition is $\max_i |X(i)| - 1$. The **treewidth**, $tw(G)$, is the minimum width over all tree decompositions.