

Lecture 2: Graph Preliminaries

Mark Peot

Read: [CGH] Chapter 4.1 to 4.6.

Comments during class:

The moral graph definition for chain graphs in the notes is correct.

There was a comment on the definition of parent. The definition in the notes is correct.

There was a question on the uniqueness of paths in a tree. Given any two nodes, there is only one path from one node to another. Thus the path definition works.

1.0 Why do we care about graphs?

Graphs will be used to model the independence properties of distributions.

The idea of graph separation will be used to capture a subset of the independence properties implied by a distribution. Our general goal (approximately): graph separation will imply probabilistic independence that, in turn, will imply that the graph can be factored. If a graph can be factored, then that will imply the structure of a graph, such that graphical separation will imply independence.

The idea of graph decomposition will be used in inference. If a graph is decomposable, we can use the distributed law of multiplication to simplify the computation of marginals in the distribution represented by that graph.

A brief note on notation. I like using big roman capitals for sets and small roman capitals for the elements of sets.

Note: Much of this section is adapted from [Lauritzen].

2.0 Graphs

A *graph* is a pair $G = (X, L)$, where X is a set of *vertices* or *nodes* and L is a set of *edges* or *links*.¹ Each *edge* $(a, b) \in X \times X$ is an ordered pair of nodes².

The graphs that we will consider are *simple*: In a simple graph, each *edge* is an ordered pair of distinct vertices (no self loops are permitted).

1. In the belief net literature, both sets of terminology (nodes/links and vertices/edges) are used. However, the 'cool' people use the terms vertices and edges when referring to graphs.

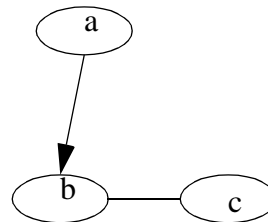
2. The book uses L_{ij} for edge (i,j) . I think that the L is redundant.

Edge (a, b) is called *undirected* if both (a, b) and (b, a) are in L . An edge is called *directed* if $(a, b) \in L$ and $(b, a) \notin L$.

$a \rightarrow b$ will be used to denote a directed edge (a, b) . $a - b$ will be used to denote an undirected edge between a and b .

Graphically, vertices are represented by ellipses, undirected edges by lines and directed edges by arrows.

The figure below represents the graph $G = (\{a, b, c\}, \{(a, b), (b, c), (c, b)\})$ or, alternatively, $G = (\{a, b, c\}, \{a \rightarrow b, b - c\})$.



2.1 Relationships in Graphs

Given a graph $G = (X, L)$:

If $a \rightarrow b$, then a is a *parent* of b and b is a *child* of a .

A *path* from a to b is an ordered sequence of two or more distinct nodes (x_0, \dots, x_n) such that $a = x_0$, $b = x_n$ and $(x_{i-1}, x_i) \in L$ for $i = \{1, \dots, n\}$. The length of the path is the number of links that it contains. If there is a path from a to b , we will write $a \Rightarrow b$.

The *adjacency set* for a is $\text{adj}(a) = \{b \in X \mid (a, b) \in L\}$, that is the set of nodes that can be reached by following all one edge paths. The *neighbors* of a are $\text{ne}(a) = \{b \in X \mid a - b \in G\}$.

A *closed path*¹ (or *cycle*) is similar to a path save that it starts and ends with the same node (a, \dots, a) . A cycle is called an *n-cycle* if it contains n edges.

A cycle is *directed* if one of the edges on the path is directed, otherwise it is *undirected* ([CGH] calls this a *loop*).

1. The book calls this a closed path, but the cool people call this a cycle.

Given graph $G = (X, L)$ and a set of nodes $A \in X$, the *induced subgraph* G_A of G is the graph $G_A = (A, L \cap A \times A)$. That is, the induced subgraph consists only of nodes in A and those links that pass between those nodes in A .

A node b is a *descendent* of a if $a \Rightarrow b$, but not $b \Rightarrow a$. Given graph $G = (X, L)$, the descendents of a are $\text{de}(a) = \{c \mid (a \Rightarrow c) \wedge \neg(c \Rightarrow a)\}$. The *nondescendents* of a , $\text{nd}(a)$, is the set $X \setminus (\text{de}(a) \cup \{a\})$.

A node b is an *ancestor* of a if $b \Rightarrow a$, but not $a \Rightarrow b$. The ancestors of a are $\text{an}(a) = \{c \mid (c \Rightarrow a) \wedge \neg(a \Rightarrow c)\}$.

Two nodes a and b are *connected* if $a \Rightarrow b$ and $b \Rightarrow a$. We will write this $a \Leftrightarrow b$. A *component* is a maximal set of nodes with the property that every node is connected to every other node. Let $[a]_A$ denote the set of nodes that are connected to a in induced subgraph A .

The *boundary* $\text{bd}(a)$ of a is $\text{pa}(a) \cup \text{ne}(a)$.

The expressions $\text{pa}(A)$, $\text{ch}(A)$ and $\text{bd}(A)$ refer the collections of parents of, children of, and nodes in the boundary of the nodes in A that are not themselves in A . For example,

$$\text{bd}(A) = \left(\bigcup_{a \in A} \text{bd}(a) \right) \setminus A.$$

3.0 Kinds of Graphs

An *undirected graph* is a graph that contains only undirected edges.

A *directed graph* is a graph that contains only directed edges.

A *directed acyclic graph* (or *DAG*) is a directed graph that contains no directed cycles.

A *chain graph* is a graph G with the property that the nodes X can be partitioned¹ into a sequence of numbered subsets $X = X(1) \cup \dots \cup X(n)$ such that

1. all edges in each $G_{X(i)}$ are undirected.
2. if $a \rightarrow b$, $a \in X(i)$, and $b \in X(j)$ then $i > j$.

1. A *partition* of A is a set A_1, \dots, A_n of subsets of A such that $A_i \cap A_j = \emptyset$ if $i \neq j$ and $A = A_1 \cup \dots \cup A_n$.

The subsets $X(1), \dots, X(n)$ are called the *chain components* of G .

A concise definition of *chain graph*: a graph with no directed cycles.

A directed acyclic graph is a chain graph where each chain component consists of a single node. An undirected graph is a chain graph with only one chain component.

3.1 Undirected Graphs

The undirected version G^- of graph G is the graph derived by replacing all of the edges of G with undirected edges.

A graph G is *connected* if there is a path between any two distinct nodes in its corresponding undirected graph G^- .

A *tree* is an undirected graph with the property that there is exactly one path between any two distinct nodes.

A *forest* is an undirected graph with the property that there is at most one path between any two distinct nodes.

A graph $G = (X, L)$ is *complete* if there is an edge between every distinct pair of nodes in X . A subset $A \subset X$ of the nodes of G is *complete* if the induced subgraph G_A is complete. A complete subgraph will also be called a *clique*.¹

If a clique A is not a subgraph of another clique in the graph, we will say that A is *maximal*.

The *complement* $G' = (X, L')$ of an undirected graph $G = (X, L)$ if $(a, b) \in L'$ iff $(a, b) \notin L$.

A set K is a *cover* for the edges in G if every edge in G contains a vertex in K . If no subset of K is a cover, then K is *minimal*.

Thm. 1: If K is a cover for the complement of G , then $X \setminus K$ is a clique in G . If K is minimal, $X \setminus K$ is a maximal clique.

C is an (a, b) -*separator* for G if every path from a to b includes a node in C . C is said to separate A from B if C is an (a, b) -*separator* for every $a \in A$ and $b \in B$.

1. In the first lecture, I claimed that a clique is a maximal complete subgraph. Although this is the usage in the graphical model community (in particular in the assigned texts), several people have pointed out to me that is inaccurate. We will use the term *maximal clique* to avoid this problem.

3.2 Directed Acyclic Graphs

The *family* for a node x in a directed graph is $\{x\} \cup \text{pa}(x)$.

A set of nodes S is an *ancestral set* if it contains the ancestors of all of its nodes, that is, $\text{an}(s) \in S$ for all $s \in S$.

The *moral graph* G^m for G contains an edge $a - b$ iff

- $a \rightarrow b$ or $b \rightarrow a$ is contained in G , or
- there exists a node c in G , such that $a \rightarrow c$ and $b \rightarrow c$.

The moral graph is obtained by ‘marrying’ the parents of each node and then dropping the directions of the edges.

The *undirected version* of a graph is the graph obtained by replacing every directed edge with an undirected edge.

A directed graph is a *polytree* if its undirected version is a tree.

3.3 Chain Graphs

The *moral graph* G^m for chain graph G contains an edge $a - b$ iff

- $a \rightarrow b$ or $b \rightarrow a$ is contained in G ,
- $a - b$ is contained in G , or
- there exists a nodes c_1 and c_2 in the same chain component in G , such that $a \rightarrow c_1$ and $b \rightarrow c_2$.

3.4 Triangulation, Perfect Numbering, and Decomposition

3.4.1 Decomposition

Disjoint subsets A , B , and C are a (weak) *decomposition* of an undirected graph, denoted (A, B, C) if

- $V = A \cup B \cup C$,
- C separates A from B , and
- C is a complete subgraph.

Decomposition (A, B, C) is *proper* if A and B are nonempty.

C decomposes G into induced subgraphs $G_{A \cup C}$ and $G_{B \cup C}$.

G is *decomposable* if G is complete or there exists a proper decomposition (A, B, C) into decomposable subgraphs $G_{A \cup C}$ and $G_{B \cup C}$.

3.4.2 Triangulation

A cycle in an undirected graph has a *chord* if two non-consecutive vertices in that path are neighbors.

A *triangulated* graph is an undirected graph with the property that every cycle of length greater than 4 possesses a chord.

Thm. 2: If G is triangulated then every induced subgraph is triangulated.

Thm. 3: [Lauritzen] The following conditions are equivalent for undirected graph G

- (i) G is decomposable,
- (ii) G is triangulated, and
- (iii) every minimal (a, b) -separator is complete.

Proof: By induction. The three conditions are true for any graph with three vertices. Assume that the result holds for all graphs $G = (X, L)$ with $|X| \leq n$. Now consider a graph G with $n + 1$ vertices.

(i) implies (ii): Say that G is decomposable. If G is complete, then it is triangulated. If G is not complete, then it has a proper decomposition into $G_{A \cup C}$ and $G_{B \cup C}$. The inductive assumption states that both of these graphs are triangulated. The only possibility for a chordless cycle is one that includes a node in A and a node in B . This cycle intersects C twice since C separates A from B . Since C is complete, this cycle must include a chord.

(ii) implies (iii): Let C be a minimal (a, b) -separator. If C has only one vertex, it is complete. Otherwise, there are at least two nodes in C , c_1 and c_2 . Since C is a minimal separator, there will be a path from a to b via both c_1 and c_2 . By concatenating these paths (flipping one), we find a “cycle” that can contain repeated points, $(a, \dots, c_1, \dots, b, \dots, c_2, \dots, a)$. Call A the connected component of a in $G_{X \setminus C}$. Call B the connected component of b in $G_{X \setminus C}$. Since $G_{X \setminus C}$ is triangulated, repeated points along with chords other than a link between c_1 and c_2 can be used to shorten the cycle until it consists of one node in A , one node in B , c_1 and c_2 . This produces a cycle of size 4.

Since C is a minimal separator, then $c_1 - c_2$. Repeating for all pairs of nodes in C yields that C is complete.

(iii) implies (i): Either G is complete or there are two nonadjacent nodes, a and b . Let C be a minimal (a, b) -separator and partition the nodes X into $[a]_{X \setminus C}$, $[b]_{X \setminus C}$, C and the remaining nodes D . Since C is complete, the triple (A, B, C) , where $A = [a]_{X \setminus C} \cup D$ and $B = [b]_{X \setminus C}$, form a decomposition of G . Each of the subgraphs $G_{A \cup C}$ and $G_{B \cup C}$ must also be decomposable, because if C_1 is a minimal (a_1, b_1) separator in $G_{A \cup C}$, it is contained in the minimal (a_1, b_1) separator in G , which is complete by assumption, therefore C_1 is complete. The inductive assumption implies then that $G_{A \cup C}$ and $G_{B \cup C}$ are complete. \square

The smallest graph that is not decomposable is a 4-cycle.

3.4.3 Perfect Numbering

Given a set of nodes $X = \{x_1, \dots, x_n\}$, a *numbering*, α , is a bijection that assigns each number in $\{1, \dots, n\}$ to a unique node:

$$\alpha : \{1, \dots, n\} \rightarrow X.$$

A numbering α the nodes X in G is called a *perfect numbering* if the subset of nodes $\text{bd}(\alpha(i)) \cap \{\alpha(1), \dots, \alpha(i-1)\}$ is complete for $i = 2, \dots, n$.

Thm. 4: (Dirac's Lemma [Lauritzen]) Let G be a triangulated graph with at least two nodes. Then G has at least two nodes with complete boundaries. If G is not complete, these nodes can be chosen to be non-adjacent.

Proof: By induction on $|X|$. If $|X| = 2$, then the lemma is true. Assume that the lemma holds for graphs with $|X| \leq n$ and let $|X| = n + 1$. If G is complete, the lemma is true. If G is not complete, then there is a proper decomposition (A, B, C) for G into subgraphs $G_{A \cup C}$ and $G_{B \cup C}$. The induction assumption used on $G_{A \cup C}$ yields a pair (a_1, a_2) of nodes that have complete boundaries. If $G_{A \cup C}$ is complete, we can choose one of these nodes to be in A . If $G_{A \cup C}$ is not complete, one of the nodes (a_1, a_2) must be in B because C is complete. Thus, there is a node with a complete boundary in A . By symmetry, there is also a vertex with a complete boundary in B . Since C separates A and B , there must be a nonadjacent pair of nodes with complete boundaries in G . \square

Thm. 5: A graph has a perfect numbering if and only if it is decomposable.

Proof: (decomposable implies perfect numbering) By induction: Say that the theorem is true for graphs with less than k nodes. If G is decomposable, then it has two nodes with perfect boundaries. Pick one and label it $\alpha(k)$. By hypothesis, $G_{X \setminus \alpha(k)}$ has a perfect numbering, $\alpha(1), \dots, \alpha(k-1)$. Therefore, the numbering $\alpha(1), \dots, \alpha(k)$ is perfect.

(perfect numbering implies decomposition). By induction: Say that the theorem is true for graphs with less than k nodes. Let $A = \{\alpha(k)\}$, let $B = X \setminus (\{\alpha(k)\} \cup \text{bd}(\alpha(k)))$ and let $C = \text{bd}(\alpha(k))$. If a graph has a perfect numbering, then it has the proper decomposition (A, B, C) . Since the boundary of $\alpha(k)$ is complete, $G_{A \cup C}$ is complete. By induction hypothesis, $G_{B \cup C}$ is decomposable. \square

3.4.4 Methods for triangulating graphs.

Untriangulated graphs can be triangulated by adding new edges, called *fill edges* to the graph.

One method (Maximum Cardinality Search) is discussed in [CGH], Chapter 4.

Another is called *node elimination*.

The algorithm:

1. Select any node x in G .
2. Add fill edges to the boundary of x until $\text{bd}(x)$ is complete.
3. Eliminate x and the edges that contain x from G .
4. Repeat until all nodes have been eliminated from G .

Add the fill edges added in step 2 to G . The resulting graph is triangulated.

An *elimination order* is the sequence of nodes removed during the algorithm.

The elimination order is *perfect* if no fill edges are added during step 2.

Thm. 6: The elimination order for a graph is perfect iff the graph is triangulated.

3.5 Basic Complexity Results

The following undirected graph problems are *NP*-complete:

1. What is the largest clique in an undirected graph G ? (Reduction to k -cover: Is there a cover of the complement of G that has k nodes or less?)
2. What is the minimum number of fill edges that I have to add to triangulate a graph?