Probability III
Graphs I

Margaret M. Fleck

11 November 2009

This lecture finishes the discussion of probability, and starts the topic of graphs (section 9.1 and some of 9.2 in Rosen).

1 Announcements

Exams should be graded by later in the week. Watch the newsgroup. We still don't know what will happen with the possible TA strike next week and we'll just have to play it by ear.

2 Conditional Probability

Suppose that we have two events E and F. The conditional probability of E given F is

$$p(E|F) = \frac{P(E \cap F)}{P(F)}$$

The idea is that we are suppose that we already know that event F is happening and, given that information, we want to figure out how likely event E is. It's the same as computing the probability of E with a reduced sample space that only contains the outcomes in F.

For example, consider bit strings of length 4. Suppose E is that the string contains two consecutive zeros and F is that the string starts with a zero. Of the 16 strings in our sample space, F contains 8 and $E \cap F$ contains five (0000, 0100, 0010, 0001, and 0011). So

$$p(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{5}{16}}{\frac{8}{16}} = \frac{5}{8}$$

Or, consider a family with two kids. Suppose E is that the family has two boys and suppose that F is that the family has at least one boy.

Then our sample space contains four outcomes (boy boy, boy girl, girl boy, girl girl). E contains one outcome (two boys) and F contains three outcomes (boy boy, boy girl, girl boy). So

$$p(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

In both cases, we're assuming that all outcomes in the sample space are equally likely When that's not true, the ideas stay the same but you need to weight everything by the probability of each outcome.

3 Independence

Suppose that E and F are events. There's two meanings of the word "independent." The stronger meaning is that there is no causal connection between E and F. If that's true, then

$$P(E \cap F) = P(E)P(F)$$

For example, if there's a 0.1 probability that I'll wear my Totoro T-shirt and a 0.3 probability that the President will wear a blue suit, there's a 0.03 probabilyt that I'll wear a Totoro T-shirt and he will wear a blue suit, because we don't influence one another's choice of clothes.

The weaker meaning of independence, which is the official definition in statistics, is that E and F are independent if and only if

$$P(E \cap F) = P(E)P(F)$$

This definition includes pairs of events that have no causal connection (as above) but also other events that might be connected to one another (or we're not sure if they are connected) but which happen to have probabilities that fit this pattern. So this is a weaker notion than lack of causal connection, but a definition that's much easier to test: compute the three probabilities and see if the equation works out.

For example, suppose that we consider bit strings of length 4, let E contain the strings starting with 1 and F contain the strings with an even number of 1's. Then $P(E) = P(F) = \frac{1}{2}$. $E \cap F$ contains exactly four strings (write out a table of all 16 and check them), so $P(E \cap F) = \frac{1}{4}$. So $P(E \cap F) = P(E)P(F)$ and therefore these two events are independent.

For some small finite problems, it's very hard to know whether two events will turn out to be independent, because this can depend on small details of the problem. For example, suppose that we have a family with two kids, E is that they have two boys and F is that they have at least one boy. Then $P(E) = \frac{1}{4}$, $P(F) = \frac{3}{4}$, and $P(E \cap F) = \frac{1}{4}$. So $P(E)P(F) = \frac{1}{4} \cdot \frac{3}{4}$ which is not equal to $P(E \cap F)$. So E and F aren't independent.

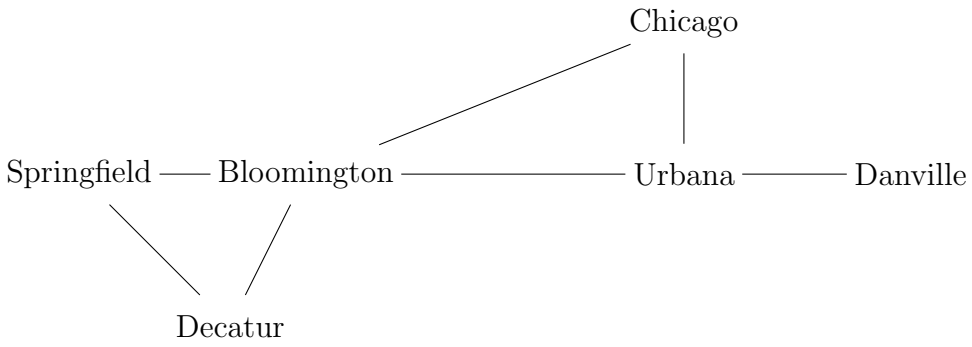
But now suppose that we have a family with three kids, E is that they have both boys and girls, F is that they have at most one boy. Then our sample space contains eight possible sequences of girls and boys. If you write them all out and check which meet the conditions for the two events, you discover that $P(E) = \frac{6}{8}$, $P(F) = \frac{4}{8}$, and $P(E \cap F) = \frac{3}{8}$. So $P(E)P(F) = \frac{6}{8} \cdot \frac{1}{2}$ which is equal to $P(E \cap F)$. So these two events are independent.

Examples like this are fragile, in that it's very hard to tell whether they will turn out to be independent or not. You basically have to work out the numbers and check.

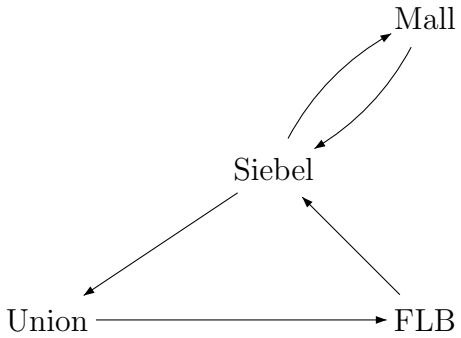
4 Graphs

Graphs are a very general class of objects, used to represent a wide variety of relationships and complex objects in computer science. A simple example of

a graph is the map of roads connecting cities shown below. In this example, each edge can be traversed in both directions. This is called an undirected graph.



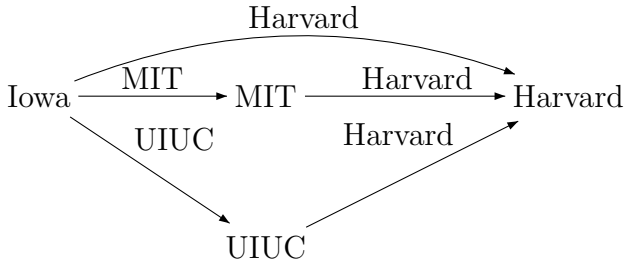
Sometimes transportation along certain routes is only one-way. For example, the examples below shows a possible (but non-existent) circular bus route serving three buildings on our campus. This is a directed graph. When a link works in both directions, you have to show this explicitly, as in the case of the busses to/from the Mall.



Graphs can also represent more abstraction relationships among objects. For example, web pages contain hyperlinks to other web pages. Each of these links has some anchor text, which labels the page at the other end of the link.

Web search systems such as Google analyze the structure of these labelled links, as well as the text on each page, to determine which pages are the best ones to return for a given search keyword.

These hyperlink relationships are typically not symmetrical: a link in one direction doesn't imply a link in the opposite direction. For example, my web page might point to Obama's web page, but the reverse link is unlikely. In fact, analysis of web pages suggests there are certain classes of pages that have mostly incoming links or mostly outgoing links. A running joke at most places other than Harvard is that other people compare to Harvard but Harvard ignores everyone else. So we might have a hyperlink graph like



In this example, notice that I've put labels on the graph edges, showing the anchor text for each link. In this class, we won't make much use of labelled edges, but you'll see them a lot in future classes e.g. automata theory.

5 Some definitions

Formally, a (finite) graph consists of a finite set of vertices V and a finite set of edges E. There are infinite graphs, but we won't see any in this class.

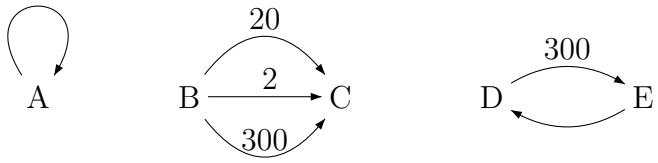
For a directed graph, E is a set of ordered pairs of edges, i.e. a subset of $V \times V$. There's several ways to formalize the edges for an undirected graph. The approach chosen by Rosen is to make each edge a set $\{v_1, v_2\}$ containing

two vertices. Alternatively, we could use ordered pairs but have a convention that (v_1, v_2) is in E if and only if (v_2, v_1) is in E. We won't need to worry about this level of detail this term.

There are many terms for different types of graphs, not entirely standard across authors. Rather than memorizing a large set of terms, it's best to concentrate on our design options:

- Is the graph directed or undirected?
- Can it contain self-loops, i.e. edges that go from a node back to the same node?
- Can it contain multi-edges, i.e. two different edges with the same starting and ending nodes?

A self loop and a multi-edge are illustrated below. The labels on the multi-edge might be, for example, the costs for various ways to get between two cities B and C (e.g. car, airplane, hitchhiking). Notice that the edges connecting D and E are not multi-edges, because they go in opposite directions.



Which of these options are allowed depends on the needs of the application. Self-loops are often allowed in directed graphs, and often forbidden in undirected graphs. Multi-edges are usually allowed only in applications where labels on the edges make the different edges differ from one another in interesting ways.