What is a Semigraphoid?

Probability is ubiquitous in our everyday lives – we use probabilistic models to predict future behavior. For example, we use probability to predict weather patterns, to determine which treatment will be the most effective for a sick patient, or to determine the best way to invest our money. Aside from these tangible applications, there are also many indirect uses of probability as well. Many common algorithms, such as quicksort or primality testing, use "random choices" to increase performence by the time spent making decisions. While counterintuitive, this often works because we can show that "randomly" is "not too bad." Whatever the particular details, the common foundation for these applications is that we develop a model for a complex system we cannot understand completely, and then use probabilistic techniques to study the behavior of these systems.

As often happens in math, we will try to make an abstract model for these more concrete objects, and this is where *semigraphoids* come in.

Conditional Independence

Suppose you want to model some real world situation (for example the interactions of different components in a biological system). One approach to this involves modelling components of the system using *random variables*, which can roughly be thought of as some model (e.g. a probability distribution) for possible outcomes of an observation. Of course, in a complicated system, the individual components do not act independently, and there is some dependence of one component on others. To capture these interactions, we often specify a *joint probability distribution*, which characterizes the probability of making a set of observations.

One useful tool for studying interactions among random variables is *conditional independence*. Intuitively, two random variables X and Y are independent conditioned upon a collection of random variables $Z = \{Z_1, Z_2, \ldots, Z_n\}$ when knowing X, or knowing Y, when you already know Z does not affect the outcome of the other, and we write this as $X \perp\!\!\!\perp Y \mid Z$.

For example, suppose you have a well-trained dog that only barks either when there is someone at the door or there is an earthquake. In this situation, our random variables might be: E (whether an earthquake is happening), D (whether there is someone at the door), and B (whether your dog is barking).

In this case, there are 6 possible independences: $E \perp\!\!\!\perp D$, $E \perp\!\!\!\perp B$, $D \perp\!\!\!\perp B$, $E \perp\!\!\!\perp D \mid B$, $E \perp\!\!\!\perp B \mid D$, and $D \perp\!\!\!\perp B \mid E$. Some of these are incompatible with our situation. For example, E and B are not independent, because knowing E gives us information about B. However, E and D are independent because knowing E does not give us information about D. Now something a little strange can occur: once we condition on B, it's no longer the case that E and D are independent. If you know that your dog is barking and there is no one at the door, then you can conclude that there is an earthquake. In general, we obtain this information from a joint probability distribution, and conditional independence gives qualitative information about a collection of random variables.

Semigraphoids

One way to understand conditional independence more deeply is to abstract it. Suppose someone gives you n random variables $X_1, X_2, \ldots, X_n$, and a set of conditional independence relations of the form $X_i \perp\!\!\!\perp X_j \mid \{X_{k_1}, X_{k_2}, \ldots, X_{k_l}\}$. Ideally, you'd like to understand whether or not these conditional independence relations could have been the conditional independence relations of an actual probability distribution. (Remember that in the example with E, D, B, there were some conditional independence statements that were incompatible with the situation.)

Unfortunately, the answer is that you cannot do this: there is not finite list of rules you can check that will give you a test to determine whether an arbitrary set of conditional independence relations come from an actual probability distribution...naturally, we won't let this stop us. Our next best bet is to write down some very agreeable rules which must be satisfied. (This means that we can decide when something does not come from a probability distribution, but not when it does.) Here is the list:

Symmetry $X \perp\!\!\!\perp Y \implies Y \perp\!\!\!\perp X$

Decomposition $X \perp\!\!\!\perp A, B \implies X \perp\!\!\!\perp A$ and $X \perp\!\!\!\perp B$

Weak Union $X \perp\!\!\!\perp A, B \implies X \perp\!\!\!\perp A \mid B$ and $X \perp\!\!\!\perp B \mid A$

Contraction $X \perp\!\!\!\perp A \mid B$ and $X \perp\!\!\!\perp B \implies X \perp\!\!\!\perp A, B$

Intersection $X \perp\!\!\!\perp A \mid B, C$ and $X \perp\!\!\!\perp B \mid A, C \implies X \perp\!\!\!\perp A, B \mid C$

For a collection of random variables $X_1, \ldots, X_n$, and a set S of conditional independence statements (statements of the form $X_i \perp\!\!\!\perp X_j \mid \{X_{k_1}, \ldots, X_{k_n}\}$, we say that X is independent of Y given Z if $X \perp\!\!\!\perp Y \mid Z \in S$ (here X, Y, Z are sets of random variables). We call sets S satisfying the first four rules *semigraphoids*. Semigraphoids additionally satisfying the fifth rule are called *graphoids*.

It takes some work, but you can prove that any probability distribution satisfies these rules. However, some of them are fairly intuitive: for example, if X and Y are independent, then the order I declare them in does not matter. Or if X is independent of multiple things, then it should be independent of those things individually.

In the most general case, it is difficult to use this abstraction to obtain something useful. However, an active field of current research aims to study families of semigraphoids and understand their properties using tools from various other fields of math.

Why do I care?

While occasionally very elegant, abstraction for the sake of abstraction is something that is often somewhat difficult to appreciate concretely. In this case, however, there are many tangible outcomes. One nice application of these is to the problem of *causal inference*. n Suppose you have a complicated system, and you want to understand how the individual components interact. You're able to make some observations – maybe you decide that some

components are independent, while others seem to depend on each other, etc. In this way, you write down a set of conditional independence statements you suspect (or know) are true. However, completely determining the system would require too many experiments or too many resources to perform, and you can only gather limited information. (For example, this scenario comes up frequently in the sciences, especially in modelling biological or chemical networks.)

With your data, you would still like to try to determine how your system is put together, and which pieces interact with what other pieces. This can guide future experiments, and suggest previously unidentified connections.

First, you can test whether your data is correct: you should check that the semigraphoid rules are not violated. If they are, then your data could not have come from a probability distribution, and thus probably not from real life data. Next, you can try to determine which semigraphoid best completess your data. This involves determining some metric by which to determine how "good" a fit a semigraphoid is, and there are some general heuristics for how to do this.

Essentially, semigraphoids provide a formal framework within which to analyze your data, and they allow you to apply many mathematical tools to various problems without needing to understand the technical details of these methods.