Rumour spreading on graphs

How does gossip spread? How do technologies diffuse within or across societies? How do fads and fashions take hold? How do people learn about job opportunities or recruit workers? All these are examples of information spread over social networks, networks of friendship, family and acquaintance. They play a role in everything from the smooth functioning of economies to the maintenance of social and cultural groupings. Those are some of the questions that motivate our study of rumour/information spread in networks. Another motivation comes from synthetic networks like the Internet or peer-to-peer networks, where it may be important to disseminate information rapidly; for example, security updates or routing tables. Rumour spreading mechanisms provide a template for many distributed algorithms on such networks. In these lectures, we shall look at some highly simplified mathematical models of information spread, with the goal of addressing precise questions about the spped of spread. But first, we familiarise ourselves with some terminology.

1 A very brief introduction to graphs

We shall use the terms graph and network interchangeably to refer to a finite set of *nodes* or *vertices*, some of which may be connected by *edges* or *arcs*. We write $G = (V, E)$ to denote a graph G with vertex set V and edge set E. An edge is a pair of vertices. Thus, E is a subset of $V \times V$, the Cartesian product of the vertex set with itself. An edge (i, j), $i, j \in V$ is called undirected if the order of the vertices doesn't matter, i.e., if (i, j) and (j, i) are the same, and is called directed otherwise. If $(i, j) \in E$ is a directed edge, we say it is directed from i to j. A graph is called directed or undirected if all its edges are directed or undirected respectively. In this course, all graphs will be either one or the other. If $(i, j) \in E$, then j is called a neighbour of i. If the graph is undirected, the degree of i is defined as the

number of neighbours it has. If it is directed, we use the terms in-degree and out-degree, with the obvious meanings (number of edges directed to i and from i respectively).

A path is a sequence $v_1, v_2, \ldots, v_k$ of nodes such that $(v_i, v_{i+1}) \in E$ or $(v_{i+1}, v_i) \in E$ for each i between 1 and $k-1$. If $v_1 = v_k$, the path is called a cycle. The path or cycle is called directed if the underlying graph is directed and $(v_i, v_{i+1}) \in E$ for each i between 1 and $k-1$. We use the term simple path and simple cycle to mean that the path (cycle) does not contain a proper subset which is itself a cycle.

A graph which contains no cycles is called acyclic. A graph is said to be connected if, for any two nodes $u, v \in V$, there is a path between u and v (i.e., $\exists n$ and $v_1, v_2, \ldots, v_n$ such that $v_1 = u$, $v_n = v$ and $(v_i, v_{i+1}) \in E$ or $(v_{i+1}, v_i) \in E$ for all i between 1 and $n-1$). A connected, acylic graph is called a tree. Observe that if a tree has k nodes, it must have exactly $k-1$ edges. (Show this by induction on k.) An undirected graph $G = (V, E)$ is called a complete graph if $E = V \times V$, i.e., there is an edge between every pair of nodes.

Let $G = (V, E)$ be a graph and let $V' \subseteq V$. The subgraph induced by the nodes in V' is defined as the graph $G' = (V', E')$ where $E' = \{(i, j) \in E : i, j \in V'\}$. In other words, it is the graph consisting of the nodes in V' as well as those edge in the original graph which connect these nodes. If a subgraph is a complete graph (on its node set), then it is called a clique. A set of nodes $V' \subseteq V$ is called an independent set if, for any two nodes in V', the edge between them is absent (whereas in a clique, for any two nodes, the edge between them is present).

2 Rumour spreading on the complete graph

Consider the following model of rumour spreading on a complete graph. There are n nodes, a single one of which initially knows the rumour. There are n independent unit rate Poisson processes, one associated with each node. At a time when there is a jump of the Poisson process $N_i(t)$ associated with node i, this node becomes active, and chooses another node j uniformly at random with which to communicate. If node i knows the rumour at this time and node j doesn't, then i informs j of the rumour; otherwise there is no change. This is called the "push" model as information is pushed from i

to j. It should be obvious what is meant by the pull and push-pull models. Start time at 0 and let T denote the first time that all nodes know the rumour. Then T is a random time and we can ask about its expected value or its distribution, and how this depends on n, the size of the graph.

The model description above takes a node-centred perspective. But it is equivalent to a description where the clocks sit on the edges rather than the nodes. Let's consider a node i and ask what the probability is that it chooses to communicate with node j during the time interval $([t, t + dt)$. For this to happen, node i should become active during this time interval, and it should choose node j to communicate with. The first of these events corresponds to $N_i(t + dt) - N_i(t) = 1$ (the Poisson process at node i has an increment during this time period), which has probability $1 \cdot dt$ (since the Poisson process has unit rate) independent of the past at all nodes. The second event has probability $\frac{1}{n-1}$, independent of everything else, since node i chooses another node to communicate with uniformly at random. Hence, the probability that the time period $[t, t + dt)$ sees a communication event from i to j is $\frac{1}{n-1}dt$, independent of the past. In other words, the model can be recast as follows: there are $n(n-1)$ independent Poisson processes of rate $1/(n-1)$, one associated with each *directed* edge (i, j) in the complete graph on n nodes. When there is a jump in the Poisson process on edge (i, j), the rumour is pushed from node i to node j if node i is informed at this time. Otherwise, nothing happens.

Remarks

1. Note that, in both models above, the Poisson processes on all nodes and edges start at time zero. Would it not be more natural to start the process at node i only when this node first learns of the rumour? Indeed it would, but it makes no difference. Can you see why?

2. A discrete-time version of the model was studied by Boris Pittel (On spreading a rumor, *SIAM J. Appl. Math.*, 1987, pp. 213–223). In this version, time proceeds in rounds. In each round, each informed node picks another node uniformly at random and pushes the rumour to it. The models are very similar. We have chosen to work with the continuous time model because it is somewhat easier to analyse.

We now want to analyse the model described above, and find out how long it takes for all nodes to learn the rumour. Observe from the verbal description that, if we let S_t denote the set of informed nodes at time t, then $S_t, t \geq 0$

evolves as a continuous time Markov chain. The state space of this Markov chain is the set of all subsets of the node set $\{1, 2, \ldots, n\}$, which is of size 2^n. This is a rather large state space. In fact, from the symmetry in the problem, we can see that it is enough to keep track of the number of informed nodes, rather than of exactly which nodes are informed. Even for this reduced state descriptor, the evolution is Markovian. This reduces the size of the state space to n, and makes the problem more tractable.

Let T_i denote the first time that exactly i nodes are informed, so that $T_1 = 0$ and $T_n = T$. Then, $T_{i+1} - T_i$ is the random additional time it takes for the $(i+1)^{\text{th}}$ node to be informed, after the i^{th} node has been. Let S_{T_i} denote the set of nodes that are informed at time T_i. There are i nodes in this set, and $n - i$ nodes in its complement, so that there are $i(n-i)$ edges between S_{T_i} and $S_{T_i}^c$. There are independent Poisson processes of rate $1/(n-1)$ associated with each of these edges, according to which some node in S_{T_i} contacts some node in $S_{T_i}^c$ and informs it of the rumour. (Communications taking place on edges with S_{T_i} or $S_{T_i}^c$ have no effect.)

Now, using the fact that the superposition of independent Poisson processes is a Poisson process with the sum of their rates, we conclude that the time to inform a new node is the time to the first jump in a Poisson process of rate $i(n-i)/(n-1)$. In other words, $T_{i+1} - T_i$ is an $Exp(\frac{i(n-i)}{n-1})$ random variable, independent of T_i, and of the past of the rumour spreading process. Hence, recalling the formulas for the mean and variance of an exponential random variable, we have,

$$\mathbb{E}[T_{i+1} - T_i] = \frac{n-1}{i(n-i)}, \quad \text{Var}(T_{i+1} - T_i) = \Big(\frac{n-1}{i(n-i)}\Big)^2. \tag{1}$$

Using a partial fraction expansion, we can rewrite the above as

$$\begin{aligned} \mathbb{E}[T_{i+1} - T_i] &= \frac{n-1}{n}\Big(\frac{1}{i} + \frac{1}{n-i}\Big), \\ \text{Var}(T_{i+1} - T_i) &= \Big(\frac{n-1}{n}\Big)^2\Big(\frac{1}{i^2} + \frac{1}{(n-i)^2} + \frac{2}{n}\big(\frac{1}{i} + \frac{1}{n-i}\big)\Big). \end{aligned} \tag{2}$$

Next, we note that the time until all nodes know the rumour is given by $T_n = \sum_{i=1}^{n-1}(T_{i+1} - T_i)$ since $T_1 = 0$. Hence, by (2) and the linearity of

expectation, we have

$$\begin{aligned}\mathbb{E}[T_n] &= \sum_{i=1}^{n-1} \mathbb{E}[T_{i+1} - T_i] = \frac{n-1}{n} \sum_{i=1}^{n-1} \Big(\frac{1}{i} + \frac{1}{n-i} \Big) \\ &= 2\frac{n-1}{n} \sum_{i=1}^{n-1} \frac{1}{i} \sim 2 \log n. \end{aligned} \tag{3}$$

Notation: For two sequences f_n and g_n, we write $f_n \sim g_n$ (read f_n is asymptotically equivalent to g_n) to mean that $\lim_{n\to\infty} f_n/g_n = 1$.

To show that $\sum_{i=1}^{n} \frac{1}{i} \sim \log n$, note that the sum is bounded below by $\int_0^n \frac{1}{x+1} dx$ and above by $1 + \int_1^n \frac{1}{x} dx$.

Thus, we have shown that the mean time needed for the rumour to spread to all nodes in a population of size n scales as $2 \log n$. We shall show that, in fact, the random time concentrates closely around this value. In order to do so, we first need to compute its variance. Recall that the random variables $T_{i+1} - T_i$ for successive i are mutually independent. Hence, $\mathrm{Var}(T_n) = \sum_{i=1}^{n-1} \mathrm{Var}(T_{i+1} - T_i)$, and we obtain using (2) that

$$\mathrm{Var}(T_n) = \Big(\frac{n-1}{n} \Big)^2 \sum_{i=1}^{n-1} \Big(\frac{1}{i^2} + \frac{1}{(n-i)^2} + \frac{2}{n} \big(\frac{1}{i} + \frac{1}{n-i} \big) \Big) \sim \frac{\pi^2}{3}. \tag{4}$$

We have used the fact that $\sum_{i=1}^{\infty} 1/i^2 = \pi^2/6$ to obtain the last equivalence. In order to use this variance estimate to show that the random variable T_n concentrates around its mean value, we will Chebyshev's inequality, which is an example of a probability inequality.

Probability Inequalities What can we say about the probability of a random variable taking values in a certain set if we only know its moments, for instance, or its generating function? It turns out that they give us some bounds on the probability of the random variable taking values in certain specific sets. We now look at some examples.

Let X be a non-negative random variable with finite mean $\mathbb{E}X$. Then, for all $c > 0$, we have

$$\textit{Markov's inequality:} \qquad \mathbb{P}(X \geq c) \leq \frac{\mathbb{E}X}{c}.$$

The proof is straightforward. Suppose X has a density, and denote it by f.

Then

$$\mathbb{E}X = \int_0^\infty x f(x) dx \geq \int_c^\infty x f(x) dx \geq \int_c^\infty c f(x) dx = c\mathbb{P}(X \geq c).$$

Re-arranging this gives us Markov's inequality. (Why does X have to be non-negative?)

Next, let X be a random-variable, not necessarily non-negative, with finite mean $\mathbb{E}X$ and finite variance $\mathrm{Var}(X)$. Then, for all $c > 0$, we have

$$\textit{Chebyshev's inequality:} \qquad \mathbb{P}(|X - \mathbb{E}X| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

The proof is an easy consequence of Markov's inequality. Note that the event $|X - \mathbb{E}X| \geq c$ is the same as the event $(X - \mathbb{E}X)^2 \geq c^2$, and apply Markov's inequality to the non-negative random variable $Y = (X - \mathbb{E}X)^2$. Note that $\mathbb{E}Y = \mathrm{Var}(X)$.

Finally, let X be a random-variable, not necessarily non-negative, and suppose that its moment-generating function $\mathbb{E}[e^{\theta X}]$ is finite for all θ. Then, for all $c \in \mathbb{R}$, we have

$$\textit{Chernoff's inequality:} \qquad \mathbb{P}(X \geq c) \leq \inf_{\theta > 0} e^{-\theta c} \mathbb{E}[e^{\theta X}].$$

The proof follows by noting that the event $X \geq c$ is identical to the event $e^{\theta X} \geq e^{\theta c}$ for all $\theta > 0$ (the inequality gets reversed for $\theta < 0$), applying Markov's inequality to the non-negative random variable $Y = e^{\theta X}$, and taking the best bound over all possible θ.

Now, let's apply Chebyshev's inequality to the random variable T_n, the time for the rumour to reach all nodes. Using the estimates for the mean and variance of T_n in (3) and (4), we find that the statement

$$\mathbb{P}(|T_n - 2\log n| \geq c) \leq \frac{\pi^2}{3c^2}$$

is approximately true for large n. More precisely, for every $\epsilon > 0$, it holds for all n sufficiently large that

$$\mathbb{P}(|T_n - 2\log n| \geq c) \leq \frac{(1+\epsilon)\pi^2}{3c^2},$$

which tends to zero as c tends to infinity. In words, what this says is that the random variable T_n grows roughly like $2\log n$. The fluctuation around this mean value does not grow unboundedly with n, but remains bounded.

3 Rumour spreading on general graphs

In the last section, we saw that the time it takes for a rumour to spread on the complete graph (for the specific model considered) grows logarithmically in the population size. What can we say more generally? How does the shape of a graph affect the spreading time?

Let $G = (V, E)$ be a directed graph on n nodes, and let P be an $n \times n$ stochastic matrix with the property that $p_{ij} = 0$ if $(i, j) \notin E$ (i.e., the non-zero entries in P correspond to edges in the graph G). We consider the following rumour spreading model. There are n independent unit rate Poisson processes, one associated with each node. At an increment time of the Poisson process at node i, this node becomes active and picks a neighbour j with probability p_{ij}, the ij^{th} element of the matrix P. If i knows the rumour at this time and j doesn't, then i pushes the rumour to j. Otherwise, nothing happens. Note that if G is the complete graph, and P is the matrix with zero diagonal elements and all off-diagonal elements equal to $1/(n-1)$, then we recover the model of the previous section.

Again, we start with a single node s (called the source node) which is initially informed of the rumour, and are interested in the random time T until all nodes become informed. The dynamics can be modelled as a Markov process if we take the state to be the *set of informed nodes* at that time. It is not enough to keep track of the number of informed nodes, as the future dynamics depend on where in the network these nodes are located. Likewise, the distribution of the random variable T may well depend on which node we start with as the source of the rumour. As mentioned in the last section, the state space becomes very large (of size 2^n) if we have to use the subset of infected nodes as the state variable. This makes it a lot harder to obtain estimates of the mean and the variance that are sharp. What we shall do in this section instead is derive an *upper bound* on the rumour spreading time T based on some simple properties of the graph G and the matrix P that determines the choice of communication contacts.

As in the previous section, it is helpful to go from a node-centric to an edge-centric description. Pick a directed edge $(i, j) \in E$. What is the probability that i attempts to push the rumour to j in some infinitesimal time interval dt? Two things need to happen for this: node i has to be activated, and it has to choose j as its contact. The first event has probability $1 \cdot dt$ (as the Poisson process at node i is of unit rate), the second has probability p_{ij},

and they are independent of each other and of the past of all processes. In other words, the process of communications along the directed edge (i, j) is a Poisson process of rate p_{ij}, and communications along different edges are mutually independent. We shall make use of the following definition in our analysis of the rumour spreading time.

Definition. The *conductance* of the non-negative matrix P is defined as

$$\Phi(P) = \min_{S \subset V, S \neq \emptyset} \frac{\sum_{i \in S, j \in S^c} p_{ij}}{\frac{1}{n}|S| \cdot |S^c|}. \tag{5}$$

The minimum is taken over all non-empty proper subsets S of the vertex set V, S^c denotes the complement of S, and $|S|$ denotes the size of the set S.

As in the analysis for the complete graph, let T_k denote the first time that exactly k nodes are informed of the rumour. Thus, $T_1 = 0$ and T_n is the time that all nodes are informed. Let S_k denote the (random) subset of nodes that are informed at time T_k. What can we say about $T_{k+1} - T_k$? For each node $i \in S_k$ and $j \in S_k^c$, the events of i contacting j occur according to a Poisson process of rate p_{ij}. Moreover, these are mutually independent for distinct ordered pairs of nodes. Thus, using the fact that the superposition of independent Poisson processes is a Poisson process with the sum of their rates, we see that the total rate at which an informed node contacts an uninformed node and informs it of the rumour is given by $\sum_{i \in S_k, j \in S_k^c} p_{ij}$. (Contacts between nodes within S_k, or within S_k^c, result in no change of the system state, so we can ignore them.) Hence, conditional on S_k, the time $T_{k+1} - T_k$ until an additional node becomes informed is exponentially distributed with parameter $\sum_{i \in S_k, j \in S_k^c} p_{ij}$. Consequently,

$$\mathbb{E}[T_{k+1} - T_k | S_k] = \frac{1}{\sum_{i \in S_k, j \in S_k^c} p_{ij}}. \tag{6}$$

In order to compute $\mathbb{E}[T_n]$ exactly, we would have to consider every possible sequence of intermediate sets along which the system can go from just the source being informed to all nodes being informed, computing the probability of the sequence and using the conditional expectation estimate above. This is impractical for most large networks. Instead, we shall use the conductance to bound the conditional expectation of $T_{k+1} - T_k$. Observe from (5) and (6) that

$$\mathbb{E}[T_{k+1} - T_k | S_k] \leq \frac{1}{\Phi(P)} \frac{n}{k(n-k)} = \frac{1}{\Phi(P)} \Big(\frac{1}{k} + \frac{1}{n-k} \Big). \tag{7}$$

We have used the fact that $|S_k| = k$ by definition, and so $|S_k^c| = n - k$. Note that while the exact conditional expectation of $T_{k+1} - T_k$ depends on the actual set S_k of informed nodes at time T_k, the bound does not; it only depends on k, the number of informed nodes at this time.

We can use this bound to easily obtain a bound on the expected time to inform all nodes, following the same steps as in the analysis of the complete graph. First, write $T_n = \sum_{i=1}^{n-1} T_{i+1} - T_i$ since $T_1 = 0$. Next, use the linearity of expectation, and the bound in (7), to get

$$\mathbb{E}[T_n] \leq \sum_{k=1}^{n-1} \frac{1}{\Phi(P)} \Big(\frac{1}{k} + \frac{1}{n-k} \Big) = \frac{2}{\Phi(P)} \sum_{k=1}^{n-1} \frac{1}{k} \sim \frac{2 \log n}{\Phi(P)}. \tag{8}$$

The above expression is a bound on the expected value of the random variable T_n. Can we also something about the distribution of the random variable. Using Markov's inequality, we obtain

$$\mathbb{P}\Big(T_n > \frac{c \log n}{\Phi(P)} \Big) \leq \frac{\Phi(P) \mathbb{E}[T_n]}{c \log n} \leq \frac{2}{c},$$

which tends to zero as c tends to infinity. In other words, the random variable T_n is of order $\log n / \Phi(P)$ in probability.

Example. Let $G = (V, E)$ be the complete graph, and suppose $p_{ij} = 1/(n-1)$ for every ordered pair (i, j), $i \neq j$. This is exactly the model of rumour spreading on a complete graph that we first analysed. What does the bound tell us in this case? To answer that, we need to compute the conductance $\Phi(P)$ for this example. Fix a subset S of the node set consisting of k nodes, where k is not equal to zero or n. For each node in this set, there are $n - k$ edges to nodes in S^c. The communication rate on each of these edges is $1/(n-1)$. Hence, we get

$$\sum_{i \in S, j \in S^c} p_{ij} = \frac{k(n-k)}{n-1},$$

and so

$$\frac{\sum_{i \in S, j \in S^c} p_{ij}}{\frac{1}{n} |S| \cdot |S^c|} = \frac{n}{n-1},$$

irrespective of the choice of S. Hence, taking the minimum over S gives us $\Phi(P) = \frac{n}{n-1}$. Substituting this in (8), we obtain that $\mathbb{E}[T_n]$ is bounded by a quantity that is asymptotic to $2 \log n$, which is precisely the same as the

exact analysis gave us. Thus, the bound is tight in this example. In general, of course, it won't be tight, but in many examples, it may be good enough to be useful, and yield at least the right scaling in n of the rumour spreading time, though not the exact constants. In your homework problems, you'll see examples both of when the bound is good and when it is not.

4 Stochastic ordering

In the last section, we observed that $T_{k+1} - T_k$ is exponentially distributed with parameter $\lambda_{S_k} = \sum_{i \in S_k, j \in S_k^c} p_{ij}$, which depends on the random set S_k reached at the first time that k nodes are informed. We noted that the parameter of this exponential is bounded from below by $k(n-k)\Phi(P)/n$, and hence that its mean is bounded from below by the reciprocal of this quantity. Here $\Phi(P)$ denotes the conductance of the matrix P. While it sufficed for our analysis in the last section to bound the mean of $T_{k+1} - T_k$, and thereby the mean time to inform all nodes of the rumour, we might want to be able to say more. We might want to bound the random variable T_n in some probabilistic sense. Stochastic ordering is a way of comparing random variables.

Definition. We say that X is stochastically dominated by Y, denoted $X \preceq Y$, if $F_X(a) \geq F_Y(a)$ for all $a \in \mathbb{R}$.

A relation R on a set S is called a partial order if it is reflexive (xRx for all $x \in S$) and transitive (xRy and yRz implies xRz), and if xRy and yRx together imply that x equals y. It is a total order if any two elements of the set are comparable, i.e., for all x and y in S, either xRy or yRx (or both). The set S together with the relation R is called a partially ordered set or (totally) ordered set accordingly. For example, the real numbers with the usual relation $\leq$ is an ordered set, while $\mathbb{R}^2$ with the relation $(x_1, y_1) \leq (x_2, y_2)$ if and only if $x_1 \leq x_2$ and $y_1 \leq y_2$ is a partially ordered set.

It is easy to see that the stochastic order relation defines a partial order on the set of probability distributions on the real numbers. With some abuse of terminology, we will also call it a partial order on random variables. (The distinction is that, while $\preceq$ is reflexive and transitive on the set of random variables, the relations $X \preceq Y$ and $Y \preceq X$ together only imply that X and Y have the same distribution. They may not be the same random variable,

or even defined on the same sample space. (Recall that a random variable is a function from a sample space Ω to the real numbers. As an example, let $\Omega = \{1, 2\}$, with $P(1) = 1/2$ and $P(2) = 1/2$, define the random variable X by $X(1) = 0$, $X(2) = 1$ and the random variable Y by $Y(1) = 1$, $Y(2) = 0$. These are different random variables because they are different functions, but they have the same probability distribution.)) The following theorem gives another description of the stochastic order relation.

Theorem 1 *The following statements are equivalent:*

1. *The random variable X is stochastically dominated by the random variable Y, i.e., $X \preceq Y$.*
2. *$\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$ for every monotone non-decreasing function $f : \mathbb{R} \to \mathbb{R}$ for which the corresponding expectations are defined (possibly infinite).*

We will not prove this theorem. Taking f to be the identity function $f : x \mapsto x$, an immediate corollary of the above theorem is that $\mathbb{E}X \leq \mathbb{E}Y$ whenever $X \preceq Y$ (provided the random variables X and Y have expectations).

Example. Let X and Y be exponential random variables with parameters λ and μ respectively, and suppose $\lambda \geq \mu$. Then,

$$F_X(a) = 1 - e^{-\lambda a} \geq 1 - e^{-\mu a} = F_Y(a)$$

for all $a \geq 0$, while $F_X(a) = F_Y(a) = 0$ for $a < 0$. Hence, $X \preceq Y$.

In the last section, we showed that, conditional on the set of informed nodes S_k reached at time T_k, the random variable $T_{k+1} - T_k$ is exponentially distributed with parameter $\sum_{i \in S_k, j \in S_k^c} p_{ij}$. Moreover, we showed that this parameter is bounded from below by $k(n-k)\Phi(P)/n$. Hence, by the example above, the random variable $T_{k+1} - T_k$ is stochastically dominated by an exponential with parameter $k(n-k)\Phi(P)/n$. We want to go on from this to claim that T_n is stochastically bounded by a sum of independent exponential random variables with parameters $k(n-k)\Phi(P)/n$, $k = 1, 2, \ldots, n-1$. To argue this, we want to be able to reason as follows: if Y_1 and Y_2 are independent random variables, if $X_1 \preceq Y_1$ and $X_2 | X_1 \preceq Y_2$ (for almost every X_1), then $X_1 + X_2 \preceq Y_1 + Y_2$. Such reasoning would be justified if we could replace the order relation $X \preceq Y$ which relates distributions, with an order

relation $X \leq Y$, which deterministically relates the corresponding random variables. Recalling that random variables are functions on a sample space, what $X \leq Y$ says is that the function X is dominated by the function Y pointwise, i.e., $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. The following theorem says that we can indeed think like this.

Theorem 2 (Strassen's theorem) *Let X and Y be random variables, possibly on different probability spaces, and suppose $X \preceq Y$. Then, there exist random variables $\tilde{X}$ and $\tilde{Y}$ defined on the same probability space such that $\tilde{X}$ has the same distribution as X, $\tilde{Y}$ has the same distribution as Y, and $\tilde{X} \leq \tilde{Y}$ almost surely (with probability 1).*

Proof. Let F_X and F_Y denote the cdfs of the random variables X and Y respectively, and recall that these are monotone increasing functions, though not necessarily continuous. For such a function $F : \mathbb{R} \to [0, 1]$, we define the generalised inverse

$$F^-(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}, \quad y \in (0, 1).$$

Note that, if F is strictly increasing and continuous, then it is invertible, and the generalised inverse defined above coincides with the usual inverse, where $F^{-1}(y)$ is defined as the unique x solving $F(x) = y$. We have been careful to only define F^- on $(0, 1)$; it is well-defined on $[0, 1]$ if $F(x_1) = 0$ and $F(x_2) = 1$ for some$x_1, x_2 \in \mathbb{R}$, but will not be defined if $F(x)$ only tends to 0 and 1 as x tends to $-\infty$ and $+\infty$ respectively.

Let U be a random variable uniformly distributed on $(0, 1)$, and define $\tilde{X} = F_X^-(U)$, $\tilde{Y} = F_Y^-(U)$. Then, $\tilde{X}$ and $\tilde{Y}$ are defined on the same probability space, namely the one on which U is defined.

Now, as $X \preceq Y$, we have $F_X(a) \geq F_Y(a)$ for all $a \in \mathbb{R}$, by definition. Consequently, $\{a : F_Y(a) \geq b\} \subseteq \{a : F_X(a) \geq b\}$, and hence $F_Y^-(b) \geq F_X^-(b)$, for all $b \in (0, 1)$. In particular, $\tilde{Y} \geq \tilde{X}$, for any realisation of the random variable U. This completes the proof of the theorem. □

The theorem extends to random vectors (joint distributions) as well. The notion of stochastic ordering for random vectors is similar to that for random variables: we say $(X_1, \ldots, X_n) \preceq (Y_1, \ldots, Y_n)$ if $\mathbb{E}[f(\mathbf{X})] \leq \mathbb{E}[f(\mathbf{Y})]$ for all non-decreasing functions. Here, a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be non-decreasing if $f(\mathbf{x}) \leq f(\mathbf{y})$ whenever $\mathbf{x} \leq \mathbf{y}$ in the usual partial order on $\mathbb{R}^n$. Equivalently, f is called non-decreasing if it is non-decreasing in each of its n variables.