Combinatorics, Probability and Computing (2011) 20, 347–361. © Cambridge University Press 2011
doi:10.1017/S0963548311000095

Negative Dependence and Srinivasan's Sampling Process

JOSH BROWN KRAMER[1], JONATHAN CUTLER[2]
and A. J. RADCLIFFE[3]

[1]Department of Mathematics and Computer Science Department,
Illinois Wesleyan University, Bloomington, IL, USA
(e-mail: jbrownkr@iwu.edu)
[2]Department of Mathematical Sciences, Montclair State University, Montclair, NJ, USA
(e-mail: jonathan.cutler@montclair.edu)
[3]Department of Mathematics, University of Nebraska–Lincoln, Lincoln, NE, USA
(e-mail: aradcliffe1@math.unl.edu)

Received 3 December 2009; revised 25 January 2011; first published online 22 February 2011

Dubhashi, Jonasson and Ranjan (2007) study the negative dependence properties of Srinivasan's sampling processes (SSPs), random processes which sample sets of a fixed size with prescribed marginals. In particular they prove that linear SSPs have conditional negative association, by using the Feder–Mihail theorem and a coupling argument. We consider a broader class of SSPs that we call tournament SSPs (TSSPs). These have a tree-like structure and we prove that they have conditional negative association. Our approach is completely different from that of Dubhashi, Jonasson and Ranjan. We give an abstract characterization of TSSPs, and use this to deduce that certain conditioned TSSPs are themselves TSSPs. We show that TSSPs have negative association, and hence conditional negative association. We also give an example of an SSP that does not have negative association.

1. Introduction

The theory of negative dependence has been a subject of great interest for mathematicians of late. Intuitively, a sequence of random variables is negatively dependent if the event that some subset of them are 'large' tends to make the values of other variables 'small'. In [6], Pemantle calls for a general theory of negative dependence corresponding to that of positive dependence. One step towards this goal had already been achieved by Reimer [7] in proving the van den Berg–Kesten conjecture [10].

The examples of negative dependence properties with which this paper is concerned are that of negative association and conditional negative association (defined below). While these concepts were introduced by Joag-Dev and Proschan [5], they were studied earlier by Feder and Mihail [3] in the context of balanced matroids. Since then negative association

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

has been well studied: see, for example, [2], [4] and [8]. In the definition that follows, note that negative association is a much stronger condition than pairwise negative correlation.

Definition 1. Given a sequence $A = (A_x)_{x\in I}$ of real-valued random variables, we write A_J for the subsequence $(A_x)_{x\in J}$. We say that A is *negatively associated* if, for every pair of disjoint subsets $J, K \subseteq I$ and all non-decreasing functions f, g,

$$\mathbb{E}[f(A_J)g(A_K)] \leqslant \mathbb{E}[f(A_J)]\mathbb{E}[g(A_K)].$$

The variables $(A_x)_{x\in I}$ are *conditionally negatively associated* if, for any subset $J \subseteq I$ and sequence $a = (a_y)_{y\in J}$, the sequence $(A_{I\setminus J} \mid A_J = a)$ is negatively associated.

In this paper, we will be concerned with the negative association and conditional negative association of Srinivasan's sampling process (SSP), a method of producing a random k-subset of an n-element set [9]. A random k-subset of an n-element set can be thought of as a sequence of binary random variables $A = (A_x)_{x\in I}$, where $|I| = n$, in which $A_x = 1$ if x is in the k-set and $A_x = 0$ if not. Since we are choosing a k-subset, we have $\sum_{x\in I} A_x = k$. We freely switch between saying $x \in A$ and $A_x = 1$. One simple example of such a process is *uniform sampling of k-subsets*, where k-subsets are chosen uniformly from all subsets of the n-element set of size k. In applications, including integer linear programming, it is desirable to be able to prescribe the marginals, *i.e.*, $\mathbb{P}(A_x = 1) = p_x$ where $0 \leqslant p_x \leqslant 1$ for all $x \in I$. This rules out uniform sampling.

Srinivasan's sampling process was introduced as a method of producing random k-sets from an n-element set quickly and with given marginals. The question of whether SSPs have negative association and conditional negative association was considered by Dubhashi, Jonasson and Ranjan [1].

Before we define SSPs, we introduce some results related to negative association and conditional negative association. Feder and Mihail [3] gave a somewhat surprising sufficient condition for conditional negative association, involving the notion of variables of positive influence, defined as follows.

Definition 2. Let $A = (A_x)_{x\in I}$ be real-valued random variables and F be an A-measurable random variable (*i.e.*, some function of the A_x). Then we say A_y is a *variable of positive influence* for F if

$$\mathbb{E}[F \mid A_y = t]$$

is a non-decreasing function of t.

Feder and Mihail showed that the relatively weak property of conditional pairwise negative correlation, along with the existence of a variable of positive influence, gives conditional negative association.

Theorem 1.1 (Feder and Mihail [3]). *Let $(A_x)_{x\in I}$ be binary random variables such that, for any $J \subseteq I$, and any $a = (a_y)_{y\in J}$, the random sequence $(B_x)_{x\in I\setminus J} = (A_x \mid A_J = a)_{x\in I\setminus J}$ satisfies the following.*

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

- *Every non-decreasing B-measurable F has a variable of positive influence.*
- *The B_x are pairwise negatively correlated.*

Then the variables $(A_x)_{x\in I}$ are conditionally negatively associated.

When studying random k-sets there is always a variable of positive influence. Indeed

$$\operatorname{cov}\left(F, \sum_{x\in I} B_x\right) = 0,$$

since $\sum_{x\in I} B_x$ is constant, so some B_x has non-negative covariance with F, and is thus a variable of positive influence. Therefore, in this case, it is enough to show conditional pairwise negative correlation in order to prove CNA. In [1], a rather complicated coupling argument is used to show this in the special case of so-called *linear SSPs*. In this paper we discuss a broader class of SSPs called *tournament SSPs*. We show that this class of SSPs can be described by a slightly different random process that we call a *tournament sample*. This allows us to give an abstract characterization of tournament SSPs, and use this characterization to show that certain conditioned tournament SSPs are, in fact, tournament SSPs themselves. This in turn allows us to deduce conditional negative association for tournament SSPs directly from the much simpler fact that they have negative association, bypassing the Feder–Mihail theorem.

Further, we disprove a conjecture of Dubhashi, Jonasson and Ranjan by exhibiting an SSP that does not even have negative association. This is in fact a counterexample to Theorem 5.1 in their paper. The proof they give is correct for linear SSPs, but does not apply, as they claim, to arbitrary SSPs.

2. Srinivasan's sampling process

Srinivasan's sampling process [9] is a probability distribution on the k-subsets of a finite set I. It is determined by a sequence $(p_x)_{x\in I}$ of probabilities summing to k, together with an ordering (which we refer to as the 'match ordering') on pairs from I. If A is a random variable with this distribution, then for every $x \in I$ we will have $\mathbb{P}(x \in A) = p_x$.

Definition 3. Given a finite set I, a vector of probabilities $p = (p_x)_{x\in I}$ satisfying $\sum_{x\in I} p_x = k$, and a total ordering $<$ on $\binom{I}{2}$, the distribution $\mathrm{SSP}(I, p, <)$ is defined as follows. The elements $x \in I$ compete in a game in which each starts with credit p_x. Matches are played according to the match ordering. After a match between x and y at least one will have credit in $\{0, 1\}$ while the sum of their credits does not change. These matches are independent, and are defined so that each player's credit evolves as a martingale. Eventually the credit of each player is either 0 or 1, and the random set A is defined to be the set of players finishing with credit 1. We call a player whose credit becomes either 0 or 1 a *loser*, in the sense that they take no further meaningful part in the game (though, of course, whether they end up in the final random set A depends on whether their final credit is 0 or 1). For a match between x and y where their current credits are w_x and w_y, this requires us to do the following.

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

(a) $0 \leqslant w_x + w_y < 1$. This is a *loser-out* match, since one player will end up with credit 0 and the other with credit $w_x + w_y$. The probability that x wins is $w_x/(w_x + w_y)$, and similarly the probability that y wins is $w_y/(w_x + w_y)$. (Even in the case where $w_x = w_y = 0$ it is useful, in order to make our proofs more consistent, to continue to think of one of x and y as the winner. We can define the probability of x winning arbitrarily.)

(b) $1 \leqslant w_x + w_y \leqslant 2$. This is a *loser-in* match, since one player will end up with credit 1 and the other with credit $w_x + w_y - 1$. We require

$$\mathbb{P}(x \text{ wins}) = \frac{1 - w_x}{2 - w_x - w_y} \quad \text{and} \quad \mathbb{P}(y \text{ wins}) = \frac{1 - w_y}{2 - w_x - w_y}.$$

In order to simplify the analysis, from here on we insist that $p_x < 1$ for each $x \in I$. In the general case one can simply start by putting any x with $p_x = 1$ directly into the random set.

The match ordering has a substantial effect on the final distribution. Note that in a given run of the generation process many matches will be irrelevant because one or other of the contestants will already have lost a prior match, *i.e.*, their credit will already be either 0 or 1.

It was conjectured in [1] that all SSPs have conditional negative association. This conjecture turns out to be false. We give here a counterexample. In fact it is a counterexample to Theorem 5.1 of [1], since it does not even have negative association.[1]

Example 1. Define $p_i = 4/7$ for $i = 1, 2, \ldots, 7$. Let $<$ be the ordering on $\binom{[7]}{2}$ given by:

$$\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{1,7\},\{3,4\},\{2,3\},\{2,4\},\{2,5\},\{2,6\},$$
$$\{2,7\},\{3,5\},\{3,6\},\{3,7\},\{4,5\},\{4,6\},\{4,7\},\{5,6\},\{5,7\},\{6,7\}.$$

Note that this ordering is the lexicographic ordering, except that match $\{3,4\}$ is played immediately after all matches involving 1. We let $A \sim \mathrm{SSP}([7], p, <)$. The probabilities of the 4-sets of [7] being chosen are displayed in Table 1. Sets that are not listed have probability 0 of being chosen. A routine calculation shows that

$$\mathbb{P}(\{2,3,4\} \subseteq A) - \mathbb{P}(2 \in A)\,\mathbb{P}(\{3,4\} \subseteq A) = \frac{2}{245},$$

and so A fails to be negatively associated since

$$\mathbb{E}(\mathbb{1}(2 \in A)\mathbb{1}(\{3,4\} \subseteq A)) > \mathbb{E}(\mathbb{1}(2 \in A))\,\mathbb{E}(\mathbb{1}(\{3,4\} \subseteq A)).$$

One can understand this by considering the question of whether 1 wins the first match (in which case 2 is definitely in A). If 1 wins then the next match is the losing match 1 versus 3, biased 4 to 1 in favour of 3. Then, assuming that 3 is not eliminated, 3 plays 4

[1] In the proof of Theorem 5.1 of [1] the authors introduce a random variable Z representing the outcome of the first match and claim that Z is independent of the random variables A_x for all x not involved in the first match. This is correct for linear SSPs, and, as we prove later, for tournament SSPs, but is not true in Example 1.

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

Table 1.

4-set	Probability	4-set	Probability
$\{1,2,3,6\}$	1/56	$\{1,4,5,6\}$	1/14
$\{1,2,3,7\}$	1/56	$\{1,4,5,7\}$	1/14
$\{1,2,4,6\}$	9/280	$\{1,4,6,7\}$	1/28
$\{1,2,4,7\}$	9/280	$\{2,3,4,6\}$	2/35
$\{1,2,5,6\}$	3/175	$\{2,3,4,7\}$	2/35
$\{1,2,5,7\}$	3/175	$\{2,3,5,6\}$	12/175
$\{1,2,6,7\}$	3/350	$\{2,3,5,7\}$	12/175
$\{1,3,4,6\}$	1/28	$\{2,3,6,7\}$	6/175
$\{1,3,4,7\}$	1/28	$\{2,4,5,6\}$	2/35
$\{1,3,5,6\}$	1/14	$\{2,4,5,7\}$	2/35
$\{1,3,5,7\}$	1/14	$\{2,4,6,7\}$	1/35
$\{1,3,6,7\}$	1/28		

with respective probabilities 5/7 and 4/7, giving a probability of 2/7 that they are both in A. On the other hand, if 2 wins the first match then 3 and 4 play each other immediately, both with probabilities 4/7, and there is a 1/7 probability that they are both in A.

There is a natural class of SSPs whose structure is more regular. These are the linear SSPs, in which, in every match after the first, a new contender plays against the winner of the previous match.

Definition 4. Suppose that $I = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{N}$ has $x_1 < x_2 < \cdots < x_n$. Let $<_L$ be the lexicographic order on $\binom{I}{2}$ defined by $A <_L B$ if $\min(A \triangle B) \in A$. When we generate a random subset of I by the Srinivasan sampling process $\mathrm{SSP}(I, p, <_L)$, we start by playing a match between x_1 and x_2. Next we play x_3 against the winner of the first match. Then x_4 plays against the winner of the second match, and so on. This type of SSP is called a *linear SSP*.

There are several ways in which linear SSPs are easier to understand than general SSPs. The essential reason is that for a linear SSP the general outline of the process is known from the beginning. To be precise, the parameters of the jth subprocess do not depend on the prior random choices. In advance of playing match i we know whether it will be a loser-out match or a loser-in match; we know that one of the contestants is x_{i+1} and that its opponent is one of $x_1, x_2, \ldots, x_i$.

The main result of [1] is the following theorem.

Theorem 2.1 (Dubhashi, Jonasson and Ranjan [1]). *The distribution produced by a linear SSP has conditional negative association.*

3. Tournament SSPs

In this section we present the main results of the paper. We generalize Theorem 2.1 to a class of SSPs we call *tournament SSPs* or TSSPs. These are SSPs for which the match

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

schedule has a tree structure. In a TSSP the game described in Definition 3 is organized in a tournament. Leaves of the tree correspond to elements of the ground set I; internal vertices correspond to matches. In this section we give an abstract characterization of TSSPs. This allows us to deduce that certain conditioned tournament TSSPs are themselves TSSPs. This immediately implies that TSSPs have conditional negative association, since we prove that TSSPs have negative association. Our theorems for tournament SSPs apply in particular to linear SSPs.

In order to make the definition of a TSSP clear, we first need to define the notion of a tournament tree, and also the reduction of a tree (the tournament structure corresponding to the situation after one match has been played).

Definition 5. A *tournament tree* is a rooted binary tree. If T is a tournament tree with root r, and x and y are leaves of T with a common parent m, we define the *reduction of T at m*, denoted T_m, to be T with leaves x and y deleted. This is a tournament structure, which carries no information about the winner of the first match. When we need to record such information we talk about the *reduction of T in which x beats y*, denoted $T_{x/y}$, which is simply T_m with x replacing m. The root of $T_{x/y}$ is r unless $r = m$, in which case the root of $T_{x/y}$ is x. We call the non-leaf vertices of T *matches*, since they correspond to matches in the tournament. There is a natural (partial) order on the vertices of T in which $a \geqslant_T b$ if a is on the unique $b - r$ path in T. In fact $(V(T), \geqslant_T)$ is a join semi-lattice; for all $a, b \in V(T)$ there exists a unique least upper bound $a \vee b$ such that $c \geqslant a, b$ if and only if $c \geqslant a \vee b$. In particular, if $x, y \in I$ are leaves of T then $x \vee y$ is the unique match of the tournament in which x might meet y. For this reason we define $\text{match}(x, y) = x \vee y$.

Definition 6. The *TSSP associated with a tournament tree T* is any SSP in which at every stage the match being played is one of the $\geqslant_T$-minimal matches. Clearly, any such SSP has the same distribution.

There are many edge orderings that generate a TSSP with a given tournament structure. We now introduce a random process that eliminates this ambiguity, which we call a *tournament sample*. Essentially, we will indicate the winner of a given match not by the identity of the winning element, but instead by which branch of the tree they came from. This is the process we will analyse throughout the rest of the paper. It is clear that every TSSP is a tournament sample and *vice versa*.

Definition 7. Let T be a tournament tree with set of leaves I and let $p = (p_x)_{x \in I}$ be a family of probabilities with $\sum_{x \in I} p_x = k$, an integer. We start by extending p to be a function on all the vertices of T. For a match m of T define

$$k_m = \sum_{\substack{x \in I \\ x <_T m}} p_x,$$

$$p_m = k_m - \lfloor k_m \rfloor.$$

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

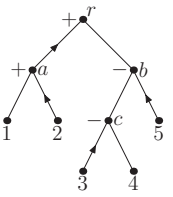


Figure 1. An example of a tournament sample; here $A = \{1, 5\}$.

Suppose now that m is a match with children a and b. (The children of m might of course be either leaves or matches.) We say that m is a *loser-out match* if $p_a + p_b < 1$; otherwise we say it is a *loser-in* match. Let Z_m be a two-valued random variable whose possible values are a and b. The Z_m are chosen to be independent, and to satisfy

$$\mathbb{P}(Z_m = a) = \begin{cases} p_a/(p_a + p_b) & m \text{ is a loser-out match,} \\ (1 - p_a)/(2 - p_a - p_b) & m \text{ is a loser-in match,} \end{cases}$$

$$\mathbb{P}(Z_m = b) = \begin{cases} p_b/(p_a + p_b) & m \text{ is a loser-out match,} \\ (1 - p_b)/(2 - p_a - p_b) & m \text{ is a loser-in match.} \end{cases}$$

For $x \in I$ let $a_0, a_1, a_2, \ldots, a_q$ be the unique $x - r$ path in T (where $a_0 = x$ and $a_q = r$). Define

$$\ell(x) = \begin{cases} \min\{i \geqslant 1 \ : \ Z_{a_i} \neq a_{i-1}\} & \text{if this set is non-empty,} \\ q + 1 & \text{otherwise,} \end{cases}$$

and

$$m(x) = \begin{cases} a_{\ell(x)} & \ell(x) \leqslant q, \\ \infty & \text{otherwise.} \end{cases}$$

Thus $m(x)$ is the first match lost by x. (The symbol ∞ represents winning the final – the match r.) Finally we define the random set S by

$$S = \{x \in I \ : \ m(x) \text{ is a loser-in match}\}.$$

This random variable S is the *tournament sample with parameters T and p*. We denote its distribution by $\mathrm{Tourn}(T, p)$. If we wish also to specify the Z_m we write $S = \mathrm{Tourn}(T, p, Z)$.

Example 2. Consider the tournament structure shown in Figure 1, with $p_1 = p_2 = 2/3$, $p_3 = p_4 = 1/6$ and $p_5 = 1/3$. We have $k = 2$ and loser-in matches are marked with a $+$, loser-out matches with a $-$. The values of the Z_a are indicated by the arrows; there is an arrow from x to m if $Z_m = x$. In this case $A = \{1, 5\}$.

We first prove some basic facts about tournament samples. Following that we prove an abstract characterization of tournament samples that will allow us to deduce rather quickly that certain conditioned tournament samples are tournament samples, and hence that they have conditional negative association.

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

Lemma 3.1. *Suppose T is a tournament tree with set of leaves I and $p = (p_x)_{x\in I}$ is as in Definition 7. Let m be a match in T with children x, y that are leaves, and set $S = \mathrm{Tourn}(T, p, Z)$. We define S' to be the tournament sample $S' = \mathrm{Tourn}(T_m, p, Z)$. (Here p_m is defined, as in Definition 7, to be $p_x + p_y - \lfloor p_x + p_y \rfloor$ and we drop Z_m from the list of match results.)*

(a) *S is given in terms of S' by*

$$S = \begin{cases} S' & m \text{ is a loser-out match, } m \notin S', \\ S' \cup \{Z_m\} \setminus \{m\} & m \text{ is a loser-out match, } m \in S', \\ S' \cup \{x, y\} \setminus \{m\} & m \text{ is a loser-in match, } m \in S', \\ S' \cup \{x, y\} \setminus \{Z_m\} & m \text{ is a loser-in match, } m \notin S'. \end{cases}$$

(b) *For any $x \in I$ we have $\mathbb{P}(x \in S) = p_x$. Moreover, if some match m' has children a, b with $p_a + p_b = 0$, then the distribution of $Z_{m'}$ does not affect the distribution of S.*

(c) *S has size $k = \sum_{x\in I} p_x$.*

Proof. Straightforward. □

Our goal is to prove that tournament samples have conditional negative association. We start by proving that they have negative association, closely following the proof of Theorem 5.1 in [1].

Theorem 3.2. *If T is a tournament structure with set of leaves I and $p = (p_x)_{x\in I}$ is a family of probabilities with $\sum_x p_x = k \in \mathbb{N}$ and $S \sim \mathrm{Tourn}(T, p)$, then S has negative association.*

Proof. We may suppose that $S = \mathrm{Tourn}(T, p, Z)$. Suppose that J, K are disjoint subsets of I and that f, g are increasing functions on $\{0,1\}^J$ and $\{0,1\}^K$ respectively. We need to show that

$$\mathbb{E}[f(S_J)g(S_K)] \leqslant \mathbb{E}[f(S_J)]\mathbb{E}[g(S_K)].$$

Pick a leaf match m (*i.e.*, a match both of whose children, x and y say, are leaves). Then we have, by the conditional covariance formula, that

$$\mathrm{cov}(f(S_J), g(S_K)) = \mathbb{E}[\mathrm{cov}(f(S_J), g(S_K) \mid Z_m)] + \mathrm{cov}(\mathbb{E}[f(S_J) \mid Z_m], \mathbb{E}[g(S_K) \mid Z_m]).$$

For the first term, note that both $(S \setminus \{y\} \mid Z_m = x)$ and $(S \setminus \{x\} \mid Z_m = y)$ are tournament samples (on $T_{x/y}$ and $T_{y/x}$ respectively). Thus, by induction,

$$\mathrm{cov}(f(S_J), g(S_K) \mid Z_m) \leqslant 0,$$

hence

$$\mathbb{E}[\mathrm{cov}(f(S_J), g(S_K) \mid Z_m)] \leqslant 0.$$

As far as the second term is concerned, note that $S_{I\setminus\{x,y\}}$ is independent of Z_m, so the second term is 0 unless one of x, y belongs to J and the other to K. (Since J and K are disjoint this is the only possibility.) Without loss of generality, let us suppose that $x \in J$

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

and $y \in K$. Suppose, firstly, that m is a loser-out match. Let $Y = \mathbb{1}(Z_m = x)$. Note that, as in Lemma 3.1, if we set $S' = \text{Tourn}(T_m, p, Z)$ then S' is independent of Z_m and

$$S = \begin{cases} S' & m \notin S', \\ S' \cup \{Z_m\} \setminus \{m\} & m \in S'. \end{cases}$$

This makes it clear that $\mathbb{E}[f(S_J) \mid Y]$ is increasing in Y since possible values of S_J are either equal for $Y = 0$ and $Y = 1$, or equal apart from having $S_x = Y$. Similarly $\mathbb{E}[g(S_K) \mid Y]$ is decreasing in Y, from which it follows immediately that

$$\text{cov}(\mathbb{E}[f(S_J) \mid Z_m], \mathbb{E}[g(S_K) \mid Z_m]) \leqslant 0.$$

The case where m is a loser-in match is similar. This time we set $Y = \mathbb{1}(Z_m = y)$. We have

$$S = \begin{cases} S' \cup \{x, y\} \setminus \{m\} & m \in S', \\ S' \cup \{x, y\} \setminus \{Z_m\} & m \notin S'. \end{cases}$$

Again the possible values of S_J for a given S' are either equal or differ in that $S_x = Y$. Again we also have that $\mathbb{E}[g(S_K) \mid Y]$ is decreasing in Y, and

$$\text{cov}(\mathbb{E}[f(S_J) \mid Z_m], \mathbb{E}[g(S_K) \mid Z_m]) \leqslant 0.$$

Having shown that both terms in the conditional covariance formula are non-positive we have $\text{cov}(f(S_J), g(S_K)) \leqslant 0$, hence S has negative association. □

We now state some not-quite-so-basic properties of tournament samples, properties that, it turns out, characterize them. For a tournament sample $\text{Tourn}(T, p, Z)$, recall that Z_m specifies the branch of T from which the winner of m comes. It is convenient to be able to refer to the contestant that wins m. We write W_m for the (random) winner of match m.

Theorem 3.3. *Suppose that $S = \text{Tourn}(T, p, Z)$ is a tournament sample. For m a match of T define*

$$\begin{aligned} D(m) &= \{z \in I \ : \ z <_T m\}, \\ S^+(m) &= \begin{cases} S \setminus D(m) & W_m \notin S, \\ (S \cup \{m\}) \setminus D(m) & W_m \in S, \end{cases} \\ S^-(m) &= S_{D(m)} \setminus \{W_m\}, \\ N_m &= |S \cap D(m)|, \end{aligned}$$

and recall that

$$\begin{aligned} k_m &= \sum_{\substack{x \in I \\ x <_T m}} p_x, \\ p_m &= k_m - \lfloor k_m \rfloor. \end{aligned}$$

Then, for all matches m in T we have the following.

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

(a) $S^+(m) = \mathrm{Tourn}(T_m, p, Z)$, *where we abuse notation and write* p *and* Z *for the restrictions of* p *and* Z *to the leaves and matches of* Z_m *respectively.*

(b) *If* m*'s children are both leaves,* x *and* y *say, then* S *can be reconstructed from* $S^+(m)$ *and* Z_m *as*

$$S = \begin{cases} S^+(m) & m \text{ is a loser-out match and } m \notin S^+(m), \\ S^+(m) \cup \{Z_m\} \setminus \{m\} & m \text{ is a loser-out match and } m \in S^+(m), \\ S^+(m) \cup \{x, y\} \setminus \{Z_m\} & m \text{ is a loser-in match and } m \notin S^+(m), \\ S^+(m) \cup \{x, y\} \setminus \{m\} & m \text{ is a loser-in match and } m \in S^+(m). \end{cases} \tag{3.1}$$

(c) $\mathbb{E}(N_m) = k_m$.

(d) *The random variables* N_m *satisfy*

$$\lfloor k_m \rfloor = \lfloor \mathbb{E} N_m \rfloor \leqslant N_m \leqslant \lceil \mathbb{E} N_m \rceil = \lceil k_m \rceil.$$

(e) $S_{D(m)}$ *and* $S_{I \setminus D(m)}$ *are conditionally independent given* N_m.

Proof. Parts (a)–(d) are straightforward. We briefly sketch the proof of (e). From part (c) we know that $N_m - \mathbb{1}(W_m \in S)$ is a constant, so we wish to show that $S_{D(m)}$ and $S_{I \setminus D(m)}$ are conditionally independent given the random variable $\mathbb{1}(W_m \in S)$. Given subsets A, B of $D(m)$ and $I \setminus D(m)$, respectively, we have

$$\begin{aligned} &\mathbb{P}(W_m \in S)\mathbb{P}(S_{D(m)} = A, S_{I \setminus D(m)} = B, W_m \in S) \\ &\quad = \mathbb{P}(S^+(m) = B \cup \{m\}) \sum_{x \in A} \mathbb{P}(W_m \in S)\mathbb{P}(S^-(m) = A \setminus \{x\}, W_m = x) \\ &\quad = \mathbb{P}(S_{I \setminus D(m)} = B, W_m \in S)\mathbb{P}(S_{D(m)} = A, W_m \in S). \end{aligned}$$

A similar calculation establishes that

$$\begin{aligned} &\mathbb{P}(W_m \notin S)\mathbb{P}(S_{D(m)} = A, S_{I \setminus D(m)} = B, W_m \notin S) \\ &\qquad = \mathbb{P}(S_{D(m)} = A, W_m \notin S)\mathbb{P}(S_{I \setminus D(m)} = B, W_m \notin S), \end{aligned}$$

completing the proof. □

We are now almost ready to state a theorem that characterizes those random processes that are tournament samples. Before we can do that, however, we need to clarify the relationship between the random variables S_x and the various Z_m, where $S = \mathrm{Tourn}(T, p, Z)$ is a tournament sample. Note that there is more information in the random variables Z than in the sample S. Some match results may be rendered irrelevant by the results of later matches. In the next lemma we prove that this 'irrelevant' information can, in a certain sense, be taken from an arbitrary source. To be more precise we would like to show that we can choose the Z_m so that they are $(\tilde{Z}, S)$-measurable, where $\tilde{Z}$ is any 'suitable' source of randomness. The following lemma provides the details.

Lemma 3.4. *Let* $S = \mathrm{Tourn}(T, p, Z')$ *and let* $\tilde{Z} = (\tilde{Z}_m)$ *be any family of independent two-valued random variables with* $\mathbb{P}(\tilde{Z}_m = a) = \mathbb{P}(Z'_m = a)$ *for all matches* m *and all vertices* a,

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

and in addition $\tilde{Z}$ is independent of Z'. Define, for $a \in V(T)$, the random variable $X_a = N_a - \lfloor k_a \rfloor$. If we let, for every match m with children a, b,

$$Z_m = \begin{cases} \tilde{Z}_m & X_a = X_b \\ a & X_a = 1, X_b = 0, m \text{ is a loser-out match}, \\ b & X_a = 0, X_b = 1, m \text{ is a loser-out match}, \\ b & X_a = 1, X_b = 0, m \text{ is a loser-in match}, \\ a & X_a = 0, X_b = 1, m \text{ is a loser-in match}, \end{cases}$$

then the (Z_m) are independent, $(\tilde{Z}, S)$-measurable, and $S = \mathrm{Tourn}(T, p, Z)$.

Proof. By construction, $\mathrm{Tourn}(T, p, Z) = \mathrm{Tourn}(T, p, Z')$, since we have only referred to values of $\tilde{Z}$ to decide matches that were irrelevant. Moreover, the (Z_m) are obviously $(S, \tilde{Z})$-measurable since the X_as are S-measurable. Now we need to verify that the Z_m are independent. Suppose then that T has at least two matches, and that $\hat{m}$ is a leaf match of T. It is easy to check, by induction, that the $(Z_m)_{m \neq \hat{m}}$ are independent. Let the children of $\hat{m}$ be x and y. Let ζ be a sequence with $\mathbb{P}((Z_m)_{m \neq \hat{m}} = \zeta) > 0$. Now there are two cases, depending on whether $X_x = X_y$, *which is determined by* ζ. Firstly, if $X_x = X_y$ then

$$\begin{aligned} \mathbb{P}(Z_{\hat{m}} = x, (Z_m)_{m \neq \hat{m}} = \zeta) &= \mathbb{P}(\tilde{Z}_{\hat{m}} = x, (Z_m)_{m \neq \hat{m}} = \zeta) \\ &= \mathbb{P}(\tilde{Z}_{\hat{m}} = x)\mathbb{P}((Z_m)_{m \neq \hat{m}} = \zeta), \end{aligned}$$

since $\tilde{Z}_m$ is independent of S and $(Z_m)_{m \neq \hat{m}}$ by hypothesis. The other case is that $X_x \neq X_y$. If $\hat{m}$ is a loser-out match then

$$\begin{aligned} &\mathbb{P}(Z_{\hat{m}} = x, (Z_m)_{m \neq \hat{m}} = \zeta) \\ &\quad = \mathbb{P}(x \in S, (Z_m)_{m \neq \hat{m}} = \zeta) \\ &\quad = \mathbb{P}(x \in S, (Z_m)_{m \neq \hat{m}} = \zeta, N_{\hat{m}} = 1) \\ &\quad = \mathbb{P}(x \in S, (Z_m)_{m \neq \hat{m}} = \zeta \mid N_{\hat{m}} = 1)\mathbb{P}(N_{\hat{m}} = 1) \\ &\quad = \mathbb{P}(x \in S \mid N_{\hat{m}} = 1)\mathbb{P}((Z_m)_{m \neq \hat{m}} = \zeta \mid N_{\hat{m}} = 1)\mathbb{P}(N_{\hat{m}} = 1) \qquad (*) \\ &\quad = \frac{p_x}{p_x + p_y}\mathbb{P}((Z_m)_{m \neq \hat{m}} = \zeta, N_{\hat{m}} = 1) \\ &\quad = \mathbb{P}(Z_{\hat{m}} = x)\mathbb{P}((Z_m)_{m \neq \hat{m}} = \zeta). \end{aligned}$$

The equality $(*)$ follows from the facts that S_x is conditionally independent of $S_{I \setminus D(\hat{m})}$ given $N_{\hat{m}}$, and that $(Z_m)_{m \neq \hat{m}}$ is $((\tilde{Z}_m)_{m \neq \hat{m}}, S_{I \setminus D(\hat{m})})$-measurable by induction. The case where $X_x \neq X_y$ and $\hat{m}$ is a loser-in match is similar. □

We are now ready to prove our characterization of tournament SSPs. Essentially the proof is a straightforward induction, but we need to be careful in moving up from a smaller case that we have enough independence between our inductively established tournament sample and the behaviour of our SSP. This is where Lemma 3.4 comes in.

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

Theorem 3.5. *Let T be a tournament structure with set of leaves I and let A be a random variable with values in $\{0,1\}^I$. Define $p_i = \mathbb{P}(A_i = 1)$, $p = (p_i)_{i\in I}$ and, for m any match in T, set $N_m = \sum_{i<_T m} A_i$. Then $A \sim \mathrm{Tourn}(T,p)$ if and only if the following hold.*

(1) *For all matches m of T,*

$$\lfloor \mathbb{E}(N_m) \rfloor \leqslant N_m \leqslant \lceil \mathbb{E}(N_m) \rceil.$$

(2) *For any match m, variables $A_{D(m)}$ and $A_{I\setminus D(m)}$ are conditionally independent given N_m.*

Proof. The implication in the forward direction follows from Theorem 3.3(d) and (e).

For the backward direction, the proof proceeds by induction on the number of leaves of T. The result is trivial if T has only one vertex. Suppose then that T has at least two leaves, and let $\hat{m}$ be a match both of whose children, x and y say, are leaves. We define a random subset B of the leaves of $T_{\hat{m}}$ as follows. We will describe the $\{0,1\}$-valued variables B_z that specify whether each leaf z of $T_{\hat{m}}$ is in or out. For leaves of the original tree we just set $B_z = A_z$. For the special leaf $\hat{m}$ we compare $|A \cap \{x,y\}|$ with its expectation and include $\hat{m}$ if we have exceeded the expected value. That is,

$$B_z = \begin{cases} N_{\hat{m}} - \lfloor \mathbb{E}(N_{\hat{m}}) \rfloor & z = \hat{m}, \\ A_z & \text{otherwise.} \end{cases}$$

Note first that since $\lfloor \mathbb{E}(N_{\hat{m}}) \rfloor \leqslant N_{\hat{m}} \leqslant \lceil \mathbb{E}(N_{\hat{m}}) \rceil$ we have $B_{\hat{m}} \in \{0,1\}$. We will show that B and $T_{\hat{m}}$ satisfy the hypotheses of the theorem. To see this, let us write, for m a match in $T_{\hat{m}}$,

$$M_m = \sum_{z<_{T_{\hat{m}}} m} B_z.$$

Note that all matches m of $T_{\hat{m}}$ are also matches of T, and M_m is either N_m or $N_m - 1$ (the second case occurs when $\hat{m}$ is a loser-in match and m is an ancestor of $\hat{m}$). Thus (1) is trivially satisfied. The required conditional independence follows from that of the A_z. Thus, by induction, $Y \sim \mathrm{Tourn}(T_m, q)$ where $q_{\hat{m}} = p_x + p_y - \lfloor p_x + p_y \rfloor$ and otherwise $q_z = p_z$. Thus there exist random variables Z_m for each match of $T_{\hat{m}}$ such that $B = \mathrm{Tourn}(T_{\hat{m}}, q, Z)$. By Lemma 3.4 we can assume that there are random variables $\tilde{Z}$ independent of A such that Z is $(\tilde{Z}, B)$-measurable. We will now make a careful choice of a Bernoulli random variable $Z_{\hat{m}}$ such that $A = \mathrm{Tourn}(T, p, Z)$. We start by letting Z_0 be a random variable, independent of $\tilde{Z}$ and B, distributed as

$$Z_0 \sim \begin{cases} \mathrm{Bernoulli}\left(\frac{p_x}{p_x+p_y}\right) & p_x + p_y < 1, \\ \mathrm{Bernoulli}\left(\frac{1-p_x}{2-p_x-p_y}\right) & p_x + p_y \geqslant 1. \end{cases}$$

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

Then we set

$$Z_{\widehat{m}} = \begin{cases} Z_0 & \text{if } N_{\widehat{m}} = 0, \\ x & \text{if } A_x = 1,\ A_y = 0 \text{ and } \widehat{m} \text{ is a loser-out match}, \\ y & \text{if } A_x = 0,\ A_y = 1 \text{ and } \widehat{m} \text{ is a loser-out match}, \\ y & \text{if } A_x = 1,\ A_y = 0 \text{ and } \widehat{m} \text{ is a loser-in match}, \\ x & \text{if } A_x = 0,\ A_y = 1 \text{ and } \widehat{m} \text{ is a loser-in match}. \end{cases}$$

For this $Z_{\widehat{m}}$ it is straightforward to check, from (3.1), that $A = \text{Tourn}(T, p, Z)$; in the cases where $Z_{\widehat{m}}$ is defined to be Z_0 its value is ignored in the calculation of $\text{Tourn}(T, p, Z)$, and in the other cases it is chosen to have the correct value to ensure that $A = \text{Tourn}(T, p, Z)$.

It remains of course to prove that $Z_{\widehat{m}}$ is independent of the other Z_m and has the right distribution. For any fixed sequence $\zeta = (\zeta_m)_{m \neq \widehat{m}}$ we wish to show that $\mathbb{P}(Z_{\widehat{m}} = x \mid (Z_m)_{m \neq \widehat{m}} = \zeta)$ has the appropriate value. We split into two cases according to whether $\widehat{m}$ is a loser-in or a loser-out match. If it is a loser-out match then we have

$$\begin{aligned} \mathbb{P}(Z_{\widehat{m}} = x \mid (Z_m)_{m \neq \widehat{m}} = \zeta) &= \mathbb{P}(Z_{\widehat{m}} = x \mid (Z_m)_{m \neq \widehat{m}} = \zeta, N_{\widehat{m}} = 1)\mathbb{P}(N_{\widehat{m}} = 1) + \\ &\qquad \mathbb{P}(Z_{\widehat{m}} = x \mid (Z_m)_{m \neq \widehat{m}} = \zeta, N_{\widehat{m}} = 0)\mathbb{P}(N_{\widehat{m}} = 0) \\ &= \mathbb{P}(A_x = 1 \mid (Z_m)_{m \neq \widehat{m}} = \zeta, N_{\widehat{m}} = 1)\mathbb{P}(N_{\widehat{m}} = 1) + \\ &\qquad \mathbb{P}(Z_0 = 1)\mathbb{P}(N_{\widehat{m}} = 0) \\ &= \mathbb{P}(A_x = 1 \mid N_{\widehat{m}} = 1)\mathbb{P}(N_{\widehat{m}} = 1) + \mathbb{P}(Z_0 = 1)\mathbb{P}(N_{\widehat{m}} = 0) \\ &= \frac{p_x}{p_x + p_y}(\mathbb{P}(N_{\widehat{m}} = 1) + \mathbb{P}(N_{\widehat{m}} = 0)) \\ &= \frac{p_x}{p_x + p_y}. \end{aligned}$$

The third equality follows from the conditional independence of A_x and A_y from $A_z, z \notin \{x, y\}$, together with the independence of $\tilde{Z}$ from A and the $(\tilde{Z}, (A_z)_{z \neq x,y})$-measurability of Z. The fourth is a simple calculation: we have

$$\mathbb{P}(A_x = 1 \wedge N_{\widehat{m}} = 1) = \mathbb{P}(A_x = 1) = p_x,$$

and, since $N_{\widehat{m}}$ is a Bernoulli random variable,

$$\mathbb{P}(N_{\widehat{m}} = 1) = \mathbb{E}(N_{\widehat{m}}) = \mathbb{E}(A_x) + \mathbb{E}(A_y) = p_x + p_y.$$

The other case is when $\widehat{m}$ is a loser-in match; this case is proved entirely analogously. □

Corollary 3.6. *If $S \sim \text{Tourn}(T, p)$, J is a subset of I, and $\zeta = (\zeta_j)_{j \in J}$ then the random variable $S' = (S \mid S_J = \zeta)$ is distributed as $\text{Tourn}(T, q)$ where $q_x = \mathbb{P}(i \in S \mid S_J = \zeta)$.*

Proof. We need only check the conditions of Theorem 3.5. For a match m in T we write $M_m = \sum_{x <_T m} S'_x$. Since S is a tournament sample M_m takes at most two adjacent values, so certainly S' satisfies (1). Now we need to prove that for m a match of T we have $S'_{D(m)}$ and $S'_{I \setminus D(m)}$ conditionally independent given M_m. Of the 'fixed' values $S_J = \zeta$ some are in $D(m)$ and some are not. Let us write $J_1 = D(m) \cap J$, $J_2 = J \setminus D(m)$ and ζ_1, ζ_2 for the restriction of ζ to J_1, J_2 respectively. Suppose then that ξ, ψ are possible values for

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

$S'_{D(m)}$ and $S'_{I\setminus D(m)}$ respectively. (In particular, we know that ξ restricted to J_1 equals ζ_1, and similarly ψ restricted to J_2 equals ζ_2.) Then

$$\begin{aligned}\mathbb{P}(S'_{D(m)} = \xi \mid S'_{I\setminus D(m)} = \psi, M_m = k) &= \frac{\mathbb{P}(S'_{D(m)} = \xi, S'_{I\setminus D(m)} = \psi, M_m = k)}{\mathbb{P}(S'_{I\setminus D(m)} = \psi, M_m = k)} \\ &= \frac{\mathbb{P}(S_{D(m)} = \xi, S_{I\setminus D(m)} = \psi, N_m = k)/\mathbb{P}(S_J = \zeta)}{\mathbb{P}(S_{J_1} = \zeta_1, S_{I\setminus D(m)} = \psi, N_m = k)/\mathbb{P}(S_J = \zeta)} \\ &= \frac{\mathbb{P}(S_{D(m)} = \xi, N_m = k)}{\mathbb{P}(S_{J_1} = \zeta_1, N_m = k)} \\ &= \mathbb{P}(S'_{D(m)} = \xi \mid M_m = k).\end{aligned}$$

The third equality follows from the conditional independence of $S_{D(m)}$ and $S_{I\setminus D(m)}$ given $N_m = k$. This proves the conditional independence of $S'_{D(m)}$ and $S'_{I\setminus D(m)}$ given M_m, which, by Theorem 3.5, proves the corollary. □

It is now immediate to deduce the main result of our paper, that tournament SSPs have conditional negative association.

Theorem 3.7. *TSSPs have conditional negative association.*

Proof. By Corollary 3.6 we know that conditioned TSSPs are themselves TSSPs. These have negative association by Theorem 3.2. Thus TSSPs have conditional negative association. □

4. Further directions

There are still a large number of natural open questions in this area. In general it would certainly be interesting to know whether there are other classes of SSPs that have conditional negative association, or indeed negative association. However, the most interesting question concerns a related process called a *random ordering SSP*. In this process we start by picking an ordering from some distribution on the set of all orderings on $\binom{I}{2}$, and then we run an SSP using this ordering. The techniques in this paper seem to offer very little traction in this more general setting. Even proving that a random ordering SSP that starts by picking a random linear ordering on I has negative association seems a difficult task.

Acknowledgements

The authors would very much like to thank the referees, who substantially improved the clarity of our presentation.

References

[1] Dubhashi, D., Jonasson, J. and Ranjan, D. (2007) Positive influence and negative dependence. *Combin. Probab. Comput.* **16** 29–41.

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press

[2] Dubhashi, D. and Ranjan, D. (1998) Balls and bins: A study in negative dependence. *Random Struct. Alg.* **13** 99–124.
[3] Feder, T. and Mihail, M. (1992) Balanced matroids. In *STOC '92: Proc. 24th Annual ACM Symposium on Theory of Computing* (New York), ACM, pp. 26–38.
[4] Grimmett, G. R. and Winkler, S. N. (2004) Negative association in uniform forests and connected graphs. *Random Struct. Alg.* **24** 444–460.
[5] Joag-Dev, K. and Proschan, F. (1983) Negative association of random variables, with applications. *Ann. Statist.* **11** 286–295.
[6] Pemantle, R. (2000) Towards a theory of negative dependence. *J. Math. Phys.* **41** 1371–1390.
[7] Reimer, D. (2000) Proof of the van den Berg–Kesten conjecture. *Combin. Probab. Comput.* **9** 27–32.
[8] Semple, C. and Welsh, D. (2008) Negative correlation in graphs and matroids. *Combin. Probab. Comput.* **17** 423–435.
[9] Srinivasan, A. (2001) Distributions on level-sets with applications to approximation algorithms. In *42nd IEEE Symposium on Foundations of Computer Science* (Las Vegas 2001), IEEE Computer Society Press, pp. 588–597.
[10] van den Berg, J. and Kesten, H. (1985) Inequalities with applications to percolation and reliability. *J. Appl. Probab.* **22** 556–569.

https://doi.org/10.1017/S0963548311000095 Published online by Cambridge University Press