Sic transit, et non

Marco LiCalzi

1 Introduction

People are formidable seekers of patterns. This propensity may bring an evolutionary advantage. A caveman able to guess that the moving grass is hiding a predator in ambush will survive to tell his tale, and may pass this life-saving ability to his progeny. The counterpart of a propensity to discern patterns is the tendency to spot them even in random events. A wind breeze may be exchanged for a hiding tiger. Coincidences may easily turn into conspiracies. But, until false positives tend to carry less dramatic consequences than false negatives (as they say, "better safe than sorry"), pattern-seeking will not disappear.

This paper recounts the story for a pattern that mathematicians call transitivity. There are myriads of situations where pairwise comparisons can be efficiently communicated and understood as transitive relations. People have become attuned to expecting transitivity, and may be surprised (or misled) when their expectations are not met. Let us enter a corner of the transitive jungle, with an eye to telling tigers apart from breezes [1, 2].

Social networks are a good place to start. Currently, Bob and Carol are friends with Ann. We show friendship as a link between adjoining nodes, on the left side of Fig. 1. A common principle in sociology states that Bob and Carol are likely to become friends themselves in the future. This new relationship would form a triangle, as shown on the right side of Fig. 1 where the third link B-C closes the triangle left open by A-B and A-C. Sociologists calls this a *triadic closure*.

There are three reasons to expect triadic closure. One is based on opportunity: if Ann enjoys her time with both Bob and Carol, the latter ones will get a chance to meet and become friends. A second facilitating factor is trust: when Bob and Carol know that each of them is one of Ann's friends, they share a stronger motive for developing a mutual trust. The third reason is that Ann may actively encourage Bob

Marco LiCalzi
Department of Management, Università Ca' Foscari Venezia, e-mail: licalzi@unive.it

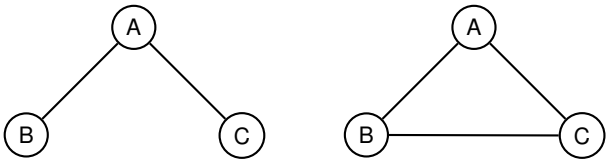


Fig. 1 Triadic closure

and Carol to become friends and reduce the latent stress arising from their being unconnected.

Clearly, triadic closure is not a necessary consequence, but simply an intuitively natural regularity. Our own experience may suggest positive and negative examples of triadic closure. What is of interest to us is that the pattern underlying a triadic closure is the same as the mathematical notion of transitivity.

A binary relation $\bowtie$ on a set S describes which pairs from S share a link. We write $A \bowtie B$ when A links to B and put an arrow from A to B. Note that the arrow is directional: $B \bowtie A$ stands for an arrow that goes from B to A. Assuming that friendship is bilateral, our former example combines two arrows in a straight link.

Transitivity states that if $B \bowtie A$ and $A \bowtie C$, then it must be true that $B \bowtie C$. Fig. 2 shows the premise on the left, and the conclusion on the right, using dashed arrows. As in a triadic closure, the presence of a relation between two pairs entails a relation

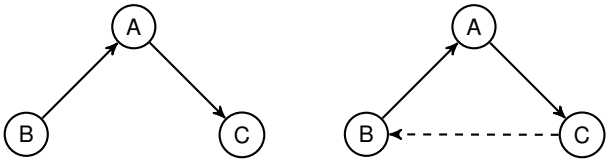


Fig. 2 Transitivity

for the third pair. There is a hint of magic here: the two arrows going through A seem to force the third arrow to appear from nowhere. (It was there to start with, indeed: transitivity makes you sure to find it.) The *Sic transit* in the title is Latin for "Thus passes". Transitivity is the property that whatever passes between B and A, and between A and C, it also goes between B and C.

Transitivity is the common thread for some of the most important families of binary relations. Mathematicians subsume them under the name of weak orders. The leading examples are two. One is equality: replace $\bowtie$ with $=$, and you get that $A = B$ and $B = C$ must imply $A = C$. The other is superiority: replace $\bowtie$ with $\geq$, and you see that $A \geq B$ and $B \geq C$ entails $A \geq C$. (No generality is lost using $\geq$ instead of $\leq$.)

The difference between equality and superiority is akin to that one between requited and unrequited love: it goes both ways, or it does not. In visual terms, equality

can be represented with plain links, as in Fig. 1; superiority calls for arrows, as in Fig. 2.

Intuitively, when facing a relation reminiscent of equality or superiority, we tend to expect transitivity. For instance, suppose that $\bowtie$ stands for "faster than": if a cheetah is faster than a giraffe, and a giraffe is faster than a tortoise, it is no surprise to anybody that a cheetah is faster than a tortoise. Or, replace $\bowtie$ with "beats": in the game of poker, a full house beats a flush and a flush beats a two pair, so a full house should also beat a two pair. The analogical mapping with $\geq$ for these two relations comes natural to almost everyone.

However, language is slippery, and thus we may be surprised or misled. The philosopher Eubulides of Miletus (4th century BCE) is famous for a few paradoxes that build on the ambiguities underlying some (alleged) forms of transitivity. The paradox of the bald man argues that a man with a full head of hair is not bald. Let x the number of hair on his head. Removing a single hair will not make it bald: hence, $x \approx (x-1)$ and $(x-1) \approx (x-2)$ imply $x \approx (x-2)$. A man with two less hair is still not bald. But taking away enough hair eventually results in baldness. Hence, the long chain $x \approx (x-1) \approx \ldots \approx 1 \approx 0$ must break somewhere, and it is unclear where. The sorites (in Ancient Greek: heap) paradox is of the same nature: a single grain of sand does not form a heap, and the addition of a single grain cannot turn a non-heap into a heap; however, adding a sufficient number of grains will eventually create a heap. How does transitivity falls apart?

2 Circular ambiguities

When the binary relation is about superiority, its directionality reinforces the expectations underlying transitivity. Here is an example. Consider a finite set S endowed with a binary relation $\succ$ (to be read "beats"): for any x and y in S, it holds that either $x \succ y$ or $y \succ x$. It only seems natural to expect that some element in S should emerge as the champion beating every other element. Yet, this is not always so, as the game Rock-paper-scissors (RPS) demonstrates.

RPS is the best known representative of Sansukumi-ken, a category of East Asian hand games based on three hand gestures. The rules for *jan-ken* or RPS state: paper beats rock; rock beats scissors; and scissors beats paper. When we put these on a triangle, we find the cycle shown on the left of Fig. 3, and transitivity fails. We call

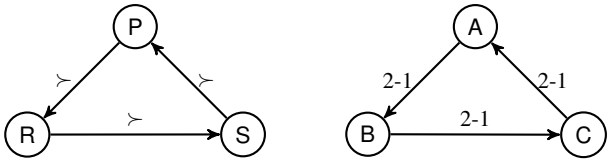


Fig. 3 Circular ambiguities

this case a (ternary) cycle, but it is known under several names. One is so delightfully evocative that "circular ambiguities" was chosen as the title of this section. The non-transitive situations discussed in the rest of this section are instances of circular ambiguities.

A famous example is the Condorcet paradox in voting theory. In its simplest form, the paradox requires three candidates (Ann, Bob, Carol) and three electors (Dan, Ken, Paula). Each elector has a transitive preference over candidates. Dan ranks them A $\succ$ B $\succ$ C; Ken orders them B $\succ$ C $\succ$ A; and Paula's ranking is C $\succ$ A $\succ$ B. If we ask electors to compare two candidates by voting over which of the two is better, A beats B for 2 votes against 1, B beats C (2 to 1), and C beats A (2 to 1); see the right-hand side of Fig. 3.

Each elector has an individual transitive preference and his own most preferred candidate, but the electorate as a whole is trapped in a cycle and cannot decide a winner. If the election is based on successive pairwise ballots, their arrangement affects which candidate may be declared a winner. For instance, if Ann is to be favoured, it suffices to pit Bob against Carol at first, let Bob emerge as the first-round winner, and then let him advance to a second and final ballot against Ann.

A variant of the Condorcet paradox arises when a single person evaluates (at least) three alternatives with respect to three different criteria. A classic example is the time-honoured fable of the princess who is mulling over three possible suitors. They fare differently over three criteria of equal importance to her. For brawn, Bob beats Dan who beats Ken; for riches, Dan beats Ken who beats Bob; for brain, Ken beats Bob who beats Dan. After a pairwise comparisons of the candidates, the princess finds herself trapped in a circular ambiguity, because each candidate loses to another who scores better in two counts out of three.

The emergence of a cycle upsetting the transitivity of "beats" is not limited to games or fables. In the coast range of California, the reproductive success of three morphs of side-blotched lizards follows a non-transitive pattern, where each morph invades another morph when rare, and is in turn invaded by another morph when common [3]. The changes in morph fitnesses match the cycle associated with RPS.

If a binary relation exhibits a cycle, a circular ambiguity arises and the binary relation is not transitive. A cycle need not be ternary, as it may take more than three elements to form one. Moreover, there may be cycles of different length. In the realm of gesture-based games, S. Kass and K. Bryla have proposed a five-gesture variant of RPS called "Rock-paper-scissors-Spock-Lizard" (RPSSL), that has made a guest appearance in Episode 8 of the second season of *The Big Bang Theory*; see `http://www.samkass.com/theories/RPSSL.html`. For lovers of extreme sports, D.C. Lovelace has assembled a version called RPS101 with 101 different gestures; see `http://www.umop.com/rps101.htm`.

We are ready to explore a few surprising results, amenable to an (unexpected) lack of transitivity. Consider three teams formed by three tennis players. The teams are named A, B, C. Players are ranked. A higher ranked player always beats someone with a lower rank. The matrix below lists the ranks for the three players in each team. Each pair of teams plays a round-robin competition, based on $3 \times 3 = 9$ matches. When A plays against B, A wins 5 matches out of 9 because A_1 (rank 8)

	1	2	3
A	8	1	6
B	3	5	7
C	4	9	2

Fig. 4 A magic square

wins all his matches and A_3 (rank 6) beats B_1 (rank 3) and B_2 (rank 5). At the end of the tournament, each team beats a second one 5-4 and loses to the third team 4-5: there is a circular ambiguity, yielding the picture on the left of Fig. 3 with the score 2-1 replaced by 5-4. We are unable to conclude which team is strongest, in spite of knowing players' ranks.

The arrangement of ranks for the players forms a magic square, where the numbers add up to the same value in each row, column, and diagonal. In particular, our 3×3 square carries a sum of 15, and is equivalent to a pattern known in the Chinese literature as the Lo Shu square since 650 BCE. Western Occultism refers to it as the Magic Square of Saturn. More mundane gamblers rely on it for constructing a biased bet.

Replace A, B, C with the three suits ♠, ♣, ♡ and players' ranks with corresponding cards. The gambler asks his victim to choose a suit, and afterwards picks one of the other two. Each player uncovers a random card from his suit, and the highest rank receives €1 from the loser. Since each pair of cards is equally likely, choosing the right suit in the cycle gives the gambler odds of 5-4 and an (expected) margin of 11% on each bet. This is a clever application of the Steinhaus-Trybula paradox [4], by which we can construct three random variables X, Y, and Z such that $P(X > Y)$, $P(Y > Z)$, and $P(Z > X)$ all exceed $1/2$.

It is reasonable for a gambler to propose bets biased in his favour, and one should be wary of accepting bets that look too good to be true. Consider the sparring over a bet between two famous (and arguably smart) celebrities such as Bill Gates and Warren Buffett. Here is what Gates narrates on the Jan.-Feb. 1996 issue of the *Harvard Business Review*:

> One area in which we do joust now and then is mathematics. Once Warren presented me with four unusual dice, each with a unique combination of numbers (from 0 to 12) on its sides. He proposed that we each choose one of the dice, discard the third and fourth, and wager on who would roll the highest number most often. He graciously offered to let me choose my die first.
>
> "Okay," Warren said, "because you get to pick first, what kind of odds will you give me?"
>
> I knew something was up. "Let me look at those dice," I said.
>
> After studying the numbers on their faces for a moment, I said, "This is a losing proposition. You choose first."
>
> Once he chose a die, it took me a couple of minutes to figure out which of the three remaining dice to choose in response. Because of the careful selection of the numbers on each die, they were non-transitive. Each of the four dice could be beaten by one of the others: die A would tend to beat die B, die B would tend to beat die C, die C would tend to beat die D, and die D would tend to beat die A. This meant that there was no winning first

> choice of a die, only a winning second choice. It was counterintuitive, like a lot of things in the business world.

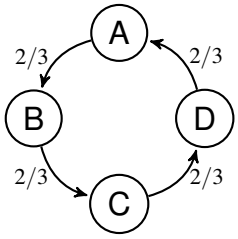
Buffett confirms the story in a interview published on *Time* in June 2001:

> I showed him a set of four dice with numbers arranged in a complex way so that any one of them would on average beat one of the others. He was one of three people I ever showed them to who figured this out and saw the way to win was to make me choose first which one I'd roll.

We are not told which set of dice was in the billionaires' hands, but they would not work differently from a set named after his inventor — Bradley Efron, best known for the bootstrap resampling technique. The values on the faces of the four Efron's dice are as follows.

0	0	4	3	3	3	2	2	2	1	1	1
4	4	4	3	3	3	2	6	6	5	5	5
A			B			C			D		

Each die beats the following one in the list, with a probability of 2/3, creating the 4-wise cycle shown below.



Varieties of non-transitive dice abound, but space prevents a thorough report. However, we mention two different kinds of dice to entice the reader's curiosity about these fascinating cubes that epitomise randomness. Sicherman's dice are the only pair of 6-sided dice different from the standard ones and carrying positive integers on their faces that exhibit the same probability distribution for their sum as generated by two standard dice. If you only bet on the sum of two dice, standard or Sicherman's makes no difference.

Lake Wobegon Dice [5] are named after the "News from Lake Wobegon" of the radio show *A Prairie Home Companion*, "where all the women are strong, all the men are good looking, and all the children are above average." A set of Lake Wobegon Dice has the property that, after a simultaneous toss, each die has a higher probability of scoring above the resulting average than below. Roughly speaking, each die tends to be above average. The simplest example is a set of three 3-sided dice with values $(1,2,2)$. After a simultaneous toss, each die has probability $10/27$ to beat the average, $8/27$ to be beaten, and $9/27$ to match it.

Our last foray among circular ambiguities shift the focus from dice to coins. Penney's game [6] is played as follows. Ann and Bob toss a fair coin until a sequence of

fixed length appears. Ann chooses a specific sequence of heads and tails. Then Bob picks another sequence of the same length, and bets that it will appear before Ann's. This game is non-transitive: for any sequence of length at least three, there is another sequence with a higher probability of occurring first. To fix ideas, consider any of the eight sequences of length three (HHH, HHT, …, TTT): the optimal choice for Bob is the sequence that starts with the opposite of Ann's middle choice, followed by the first two items in her choice. For instance, if Ann picks HTT, Bob should choose HHT. This rule gives Bob a substantial edge: for any of Ann's choices, his optimal choice ensures a winning probability of at least $2/3$.

3 Many ratings, one ranking

Up to minor quibbles, the signature of a transitive relation of the superiority kind is the ranking associated with it. In tournaments based on pairwise matches (or comparisons), including the case of three suitors chasing the princess' hand, we may find a cycle $i \succ j \succ k \succ i$. This circular ambiguity prevents the emergence of a ranking.

To circumvent these difficulties, a common approach is to introduce ratings and bypass pairwise comparisons. A rating attributes a score to a candidate. Comparisons among scores naturally respect transitivity, and candidates are accordingly ranked. Well-known examples are the practice to assess students by grading a final exam or to rank runners by clocking their time to cover a distance. Whenever a candidate can be scored on a single dimension, transitivity is ensured and a ranking may follow.

It may sometimes be difficult (if not impossible) to interpret a rating as the score measuring a performance. However, as far as any candidate can be consistently associated with a rating, a ranking still follows. For instance, a vast (but relatively unknown) family of spectral methods rate each candidate by a distinct number extracted from the eigenvector of a matrix [7]. The most celebrated case is the PageRank algorithm originally used by Google to answer a search query: the technique (and its subsequent variants) rates millions of webpages, and the ratings are used to return a list of links ranked by their relevance for the submitted query [8]. Spectral methods can be given elegant alternative interpretations [9].

We focus on a different issue: how to combine multiple ratings into a single ranking. This problem arises when candidates are scored over multiple dimensions, or evaluated by different raters. For instance, returning to Condorcet's paradox, suppose that each of three electors scores the three candidates over the scale 3-2-1. Dan rates A=3, B=2, C=1; Ken gives scores B=3, C=2, A=1; and Paula assigns C=3, A=2, B=1. The vectors of scores for each candidates, up to permutations, are the same: A=(3,2,1); B=(2,1,3); C=(1,3,2). How shall we aggregate the individual ratings and produce the single final rating used to rank them?

In a well-known dispute with Condorcet over voting methods, Jean-Charles de Borda (1733-1799) proposed to sum up the individual ratings. Following Borda's

count, each of the three candidates obtains 6 as final score. Leaving open the practical question (who is the winner of the election?), this approach solves the circular ambiguity by declaring a tie. However, different rules may yield different answers. For a trivial example, suppose that Dan is given the scale 4-2-1; applying the Borda count makes A a single winner with a final score of 7. Similarly, a matchmaker could sway the princess' heart towards a candidate by modifying the scoring scale for one of the three attributes she is considering.

There is no established consensus on how to aggregate multiple ratings. Often, people follow long-established conventions that are perceived as valid rules. Their intrinsic merits are rarely questioned, and there is little awareness of the perils involved. Sometimes, a novel approach is touted as the panacea by extolling its advantages and keeping mum on its weak points. Creative solutions are encouraged by the seeming facility associated with processing scores, and by the huge variety of details that differentiate situations where multiple ratings are aggregated.

The only principle that goes unchallenged is *dominance*: if each score for A is greater than the corresponding score for B, then A should be ranked higher than B. If Ann has higher grades than Bob in every subject at school, dominance expects Ann to rank above Bob. Dominance is transitive: if A dominates B, and B dominates C, then A dominates C. If dominance is present, circular ambiguities can be easily avoided. However, dominance is rare: the case where A has higher scores than B in some subjects, and lower in others, is much more common. When dominance fails, we can usually come up with seemingly natural aggregation methods that lead to opposing conclusions: A may be ranked above, equal to, or below B. Absent dominance, the choice of the method may upset the final ranking of two candidates.

A well-known example is the (unsporting) controversy between U.S.A. and China over the 2008 Summer Olympics. The U.S. press ranked countries by tallying all medals (equivalent to mapping gold, silver, and bronze on a scale 1-1-1). The Chinese opted for counting only the gold medals (1-0-0). Each country championed the aggregation method that gave them the top position in a global country ranking, in spite of the Olympic Charter stating that the games "are competitions between athletes in individual or team events and not between countries". Sensible suggestions based on more balanced scales (e.g., 4-2-1 or 5-3-1) have not attracted many followers.

Among the aggregation methods that respect dominance, scoring systems are commonly used in sport competitions. In many football leagues, the champion is chosen by having each team play against every other team in a round-robin tournament. Each match delivers a score on a scale 3-1-0 (for win-tie-lose), and scores are aggregated by taking their sum. At the end of a season, the highest ranking team is declared the champion. It is not difficult to find examples where a different set of scores W>T>L would select a different winner, while respecting dominance. FIFA ruled out these alternatives, by formally adopting the 3-1-0 scale in 1995.

In a round-robin tournament, the basic scheme for a scoring system has an individual current rating r for each player (or team). Following a new match, the player obtains a score s and the rating is updated to $r' = r + s$. The score s obtained after a match does not depend on the ranking or strength of the opponent. Beating the

current leader or the bottom-ranked one yields the same increment s to a player's current rating. Another restriction is that the system requires a round-robin tournament to ensure consistency over the final ratings.

Arpad Elo (1903-1992) introduced to chess a system that takes into account the relative quality of the opponents, and applies even if participants play a different number of matches. Roughly speaking, the Elo system computes an expected score e for the match between i and the other player, based on their current individual scores. Given the actual score s obtained in the match, the individual rating for i is updated by adding the difference between his current and expected score: $r_i' = r_i + (s - e)$. The most important consequence of this approach is that the unusual outcome of a low-ranked player beating a top-ranked one yields substantially bigger variations in their ratings. The Elo method is adopted by FIDE and other organisations, with some controversy over the calibration for the formula computing the expected score.

FIFA requires a ranking for national teams to arrange the draw for the final tournament in the World Cup. This ranking, however, cannot be based on the same system as the football leagues because national teams do not play round-robin tournaments. This bears similarity with the situation in world chess. FIFA has devised its own procedure for ranking men's national teams, based on the average score of recent matches and on their importance, as well as on the strength of the teams and of the confederation they belong to. The formula used by FIFA inevitably attracts its fair share of criticisms, that may partly reflect vested economic interests associated with football. (Interestingly, FIFA's women's ratings are based on a different method, closer to Elo's.)

It is possible to evaluate the effectiveness of FIFA's system against Elo's method by comparing the official ranking against the (unofficial) Elo's ranking at `http://www.eloratings.net`. The table below lists the top ten football teams in the Elo ranking as of 5 November 2015, along with the corresponding FIFA ranks. The third column reports the Elo points; to appreciate the order of magnitude, it may be helpful to know that, when Germany (then rated 2098) beat Brazil (then rated 2100) in the semifinal of the 2014 World Cup by 7 goals to 1, 82 points were transferred from the latter to the former.

Elo Rank	Team	Elo Points	FIFA Rank	Difference
1	Germany	2046	2	-1
2	Brazil	2012	8	-6
3	Argentina	2006	3	=
4	Spain	1982	6	-2
5	Chile	1954	5	=
6	Colombia	1935	7	-1
7	England	1934	9	-2
8	France	1931	24	-16
9	Uruguay	1927	12	-3
10	Belgium	1903	1	+9

The scoring system for the Formula One World Championship works on a different premise. Each race involves several drivers simultaneously. Since 2010, the first ten classified drivers in each race receives a score out of a fixed scale (with a few minor exceptions). The first three positions are rated 25-18-15; hence, one first and

one third place give a total score of 40 and generate a higher rank than the 36 points from two second places. A similar principle rules the 1968 Olympic scoring system for sailing competitions.

4 Some pitfalls

Ratings usually reflect a varying degree of subjectivity. This is substantial in sports such as skating or gymnastics, where judges have a lot of latitude; however, as many fans will attest, referees' decisions may influence the outcome of a match or a competition. Subjectivity lurks behind common practices as grading exams, or rating wines, books, hotels, and so on. It is probably impossibile to ensure that different raters yield perfectly comparable scores. Even in highly standardised situations, there will be some variability. Several academic studies confirm that grading is disputable. The 2004 Guide to the new SAT essay states that the College Entrance Examination Board expects that more than 92% of all scored essays will receive ratings within ± 1 point of each other on the 6-point SAT essay scale.

A ranking obtained by aggregating ratings from different people appeal to our instincts because it is transitive. However, the variability in subjective evaluations should make us wary. This section takes no stance on the psychological issues, preferring to point out how common techniques mar the (pseudo-)objective validity of these rankings. The fine print behind an aggregation method matters almost as much as its input.

We begin with a paradox for grading systems [10]. Assume that raters use a scale with more than two grades. Let the average-grade (AG) winner be the candidate with the highest average grade, and the superior-grade (SG) winner the candidate who receives more superior grades in pairwise comparisons. The paradox states that the AG winner may be different from the SG winner and, moreover, every rater except one may grade the SG winner higher than the AG winner. For instance, suppose that Ann, Bob, and Carol are students in the same class. Dan rates A=4, B=1, C=1 in Physical Education; Ken gives grades A=3, B=4, C=4 in History; Paula gives A=1, B=2, C=2 in Mathematics. The average grade is $8/3$ for Ann, and $7/3$ for Bob and Carol; so Ann is the AG winner who (in many schools) would be considered the best students of the three. However, Bob and Carol receive higher grades than Ann in two subjects out of three: they are the SG winners, and might stake a reasonable claim for being considered better students than Ann.

Subjective ratings that are over-rated (pun intended!) concern wines. The evidence about the consistency of wine critics under blind testing is disappointing. In 1978, Robert M. Parker Jr. began to rate wines on a 50-point scale along with the usual review. The numbers gave consumers a tool to quickly gauge the (alleged) quality of their purchases, and spawned a revolution in the industry of wine reporting. Nowadays, similar adornments are published for a variety of goods and experiences. While a single critic only has to worry about the consistency of his

own judgments, modern aggregators let users provide their ratings and attempt to combine these into a single ranking.

A favorite example of mine is the Paris Wine Tasting Competition held in 1976. Skipping many juicy details for lack of space, the core of the story is that nine French judges rated ten quality red wines from France and California under blind tasting. Every judge independently rated each wine over a 20-point scale. The final score for each wine was the average of the ratings it received. To the surprise of many, including the organiser himself, the top ranked selection was a California wine. The final scores for the top three wines were 14.17 for the 1973 vintage of Stag's Leap Wine Cellars, 14.00 for the 1970 Château Mouton-Rothschild, and 13.94 for the 1970 Château Haut-Brion.

Looking at the data suggests some of the typical issues that usually go unnoticed by the rankings' users. The paradox of grading systems applies: the AG winner was judged worse than the third wine by five critics out of nine. A majority of experts agrees that a wine is better than another, but the average score favours this latter. In simple words, a minority of strong negative opinions may overtake a majority of milder positive opinions. There is great variability in the critics' ratings: the average score per critic ranges from 9.2 to 13.5. Shall we correct for differences in critics' standards? Note that a 1-point difference between two ratings from the same critic is worth more than 0.11 points in the final average, so two per-critic points are enough to switch the first and second positions. How significant is the conclusion? If one of the nine critics is not invited to participate, the final score would be different: if we pick one of the them randomly, the probability that the winner would be another wine is $2/3$. One of the critics gave scores using half-points, thereby implicitly doubling (and changing) the official scale. His scores were accepted, but should the other judges have been given a chance to revise their grades using half-points?

Our last example deals with the attempt to take care of differences in scoring scales by normalising the values [11]. Suppose that Ann, Bob, and Carol are rated by three different criteria, and are given the scores listed in the magic square of Fig. 4. Then the average score for each candidate is 5, and they are all tied. Someone points out that the range for each criterion is different ($8-3=5$ for 1, $9-1=8$ for 2, and $7-2=5$ for 3) and suggests to normalise scores by a linear transformation mapping each of them onto $[0,1]$ before taking the average.

The normalisation is easily accomplished by computing $x' = \frac{x-x_{min}}{x_{max}-x_{min}}$ for each column. This gives the table on the left of Fig. 5, with average scores A$= 4/7$, B$= 1/2$, and C$= 3/7$. The final rank is A$\succ$B$\succ$C. Now, suppose that the score for

	1	2	3
A	1	0	5/7
B	0	1/2	1
C	2/7	1	0

	1	2	3
A	1	0	5/7
B	0	4/5	1
C	2/7	1	0

Fig. 5 Normalised scores for the magic square

C in the second column goes down from 9 to 6. Since this pejorative change affects

only C that is ranked last, it is natural to expect no changes between the ranks of A and B. However, after normalisation, the second column changes as shown on the right of Fig. 5. The new average scores are A$=4/7$, B$=3/5$, and C$=3/7$ giving a new rank B$\succ$A$\succ$C. A worse score for C bumps B up from second to first place!

References

1. Gardner, M.: The paradox of the nontransitive dice and the elusive principle of indifference. Sci. Am. **223**), 110–114, 1970.
2. Gardner, M.: On the paradoxical situations that arise from nontransitive relations. Sci. Am. **231** 120–125, 1974.
3. Sinervo, B., Lively, C.M.: The rock–paper–scissors game and the evolution of alternative male strategies. Nature **340**. 240–243, 1996.
4. Steinhaus, H., Trybula, S.: On a paradox in applied probabilities. B. Acad. Pol. Sci. Smap. **7**, 67–69, 1959.
5. Moraleda, J., Stork, D.G.: Lake Wobegon Dice. Coll. Math. J. **43**, 153–160, 2012.
6. Penney, W.: Problem 95: Penney Ante. J. of Recreat. Math., **2**, 241, 1969.
7. Langley, A.N., Meyer, C.D.: A survey of eigenvector methods for web information retrieval. Siam Rev. **47**, 135–161, 2005.
8. Vigna, S.: Spectral ranking. `arXiv:0912.0238`, v14, September 2015.
9. Callaghan, T., Mucha, P.J., Porter, M.A.: Random walker ranking for NCAA Division I-A football. Amer. Math. Monthly. **114**, 761–777, 2007.
10. Brams, S., Potthoff, R.: The paradox of grading systems. `http://www.politics.as.nyu.edu/docs/IO/2578/GradingParadox.pdf`, March 2015.
11. D'Agostino, M., Dardanoni, V., Ghiselli Ricci, R.: How to standardize (if you must). Mimeo, May 2015.