WPI-CS-TR-06-01 April 2006

Cinderella and the Big Dance

by

Craig E. Wills

Computer Science
Technical Report
Series

WORCESTER POLYTECHNIC INSTITUTE

Computer Science Department
100 Institute Road, Worcester, Massachusetts 01609-2280

Abstract

Each year a number of Cinderella stories occur in the "Big Dance," the NCAA basketball tournament, when lower seeded teams make headlines by beating higher seeded opponents. This work takes an objective view of this phenomenon by defining a new "Cinderella Index" metric to measure the degree of upsets that occur in the NCAA basketball tournament each year. It is a useful and interesting metric as it translates the yearly discussion about upsets into an objective measure that can be compared between years.

Applying the metric to the NCAA Men's tournament over the past five years shows it is dynamic as its relative values can change from round to round. Lots of upsets in one round do not necessarily correspond to upsets in the next, while lesser Cinderellas may advance far into a tournament and have a bigger impact on the Cinderella Index in later rounds. It is of particular interest that the Index for the 2006 tournament was at its highest in five years.

In applying the same metric to the NCAA Women's tournament we generally see a much smaller Index value than the Men's tournament across all rounds each year. This difference is interesting and the cause needs to be explored further.

1 Introduction

Each year at this time there is always discussion of one of the more compelling aspects of the NCAA basketball tournament—the "Cinderella" stories of lower seeded teams that rise up to defeat favorites and in the process capture the fancy of tournament followers. Rather than simply talk about these Cinderella stories, this work looks to objectively measure the occurrence of the Cinderella effect in the "Big Dance," as the tournament is known, with a measure that we term the "Cinderella Index." This work defines and measures the Cinderella Index (CI) for the current and preceding NCAA basketball tournaments. It is a useful and interesting metric because it provides an objective measure to compare the the degree of upsets that occur in different rounds and different years of the tournament.

After defining the Index, we focus on applying it to the Men's tournament. However, the CI is equally valid for the Women's tournament and we also measure it for that tournament as well. We conclude with a summary of what is learned from defining and applying the CI.

2 Cinderella Index

The CI is based on the initial seeds of each team and is a measure to determine the degree of upsets that occur in each round of the tournament. We note that the seeding itself can be a topic of great debate, but for this work the seeds are used as assigned to each team. Results of the play-in game of recent years are also ignored in the CI calculation with the winner of this game moving forward as a 16th seed.

The Cinderella Index is computed for each round of a tournament based upon all games played in that round. First, the seed value for each winning team in the round is summed together. Thus this calculation includes 32 games in the first round of the tournament, 16 games in the second round, 8 in the third round and so on. Next, the minimum possible summation of seed values is determined for each round as a baseline. In the first round, a tournament with perfect form would result in seeds 1-8 winning games in each of the four regional brackets. Taking the summation of 1 through 8 multiplied by four, results in a baseline value of 144 for round 1. Using a similar calculation, we get a baseline value of 40 (summation of 1 through 4,

four times) for round 2, a value of 12 for round 3, a value 4 for round 4 (all four #1 seeds in the Final Four), a value of 2 in round 5 and a value of 1 in round 6 (a #1 seed wins the tournament).

The Cinderella Index is simply the ratio of the accumulated seed values for a round divided by the baseline value for that round. As an example, if the Final Four (round 4) consists of a #3, #2 and two #1 seeds then the CI is $(3+2+1+1)/4 = 1.75$. The larger the value of the CI, the higher degree of Cinderella stories in that round. By definition the CI cannot be smaller than 1.0, which occurs when the highest possible seeded teams win their games in a round.

3 Cinderella Index for the Men's Tournament

The Cinderella Index not only provides a means to talk about these Cinderella stories, but it is an objective measure to examine their effect on the tournament. This measure is illustrated in Figure 1 for the six rounds of the NCAA Men's basketball tournament from the year 2001 to the present.

The figure illustrates a number of interesting aspects of Cinderella stories in the NCAA tournament. First, the figure clearly indicates that the Cinderella effect can be a dynamic one across different rounds in the tournament. For example, the relative lack of upsets in the first round of the 2004 was not repeated in the remaining rounds of that tournament. This situation occurs when modest upsets occur in the first round, such as a #9 over a #8 or a #10 over a #7, but with larger upsets in the second round when the #9 or #10 beats a #1 or #2. Conversely, the 2001 tournament had the highest degree of upsets in the first round, but a relatively small number of upsets in later rounds. This situation occurs when say a #12 Cinderella wins in round 1 only to lose to a #4 in round 2 with the expected team advancing to the third round.

As the rounds in the tournament progress, the presence of a lower seed has more of an impact. For example with Syracuse winning the tournament in 2003 as a #3 seed, the CI for the final is 3. In contrast, the 2005 tournament had the highest CI for the Final Four (round 4), yet had two #1 seeds in the final game for a baseline 1.0 CI.

The last set of results in Figure 1 shows the cumulative results for all rounds and shows that of the last five tournaments, the 2001 tournament

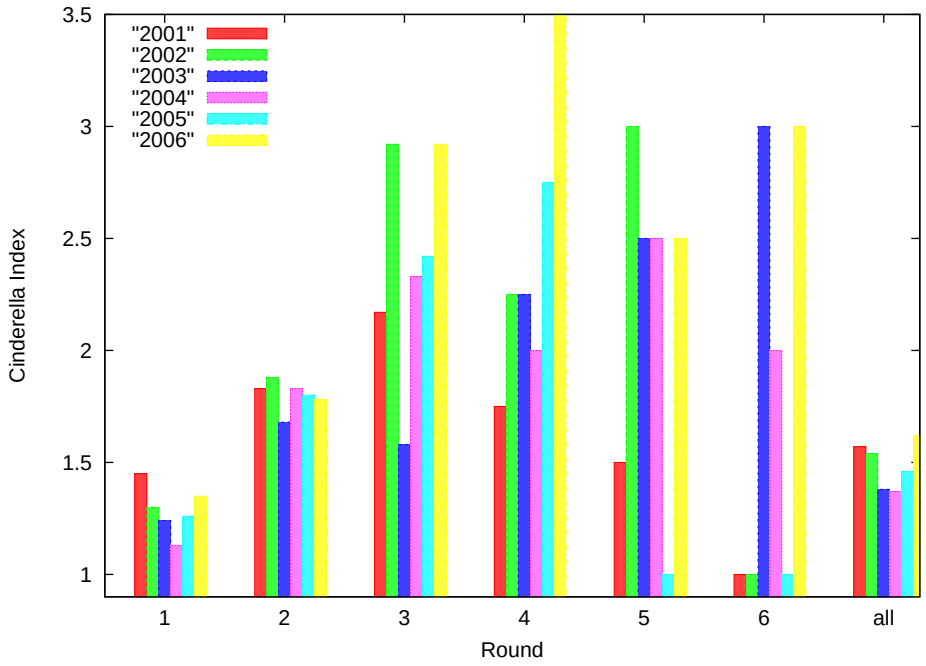

Figure 1: Cinderella Index for the NCAA Men's Basketball Tournament

had the highest Cinderella Index. It is interesting to note that the 2006 tournament has the highest CI in the first round and the highest cumulative CI of any tournament since 2001. The fourth round in 2006 was actually 5.0.

4 Cinderella Index for the Women's Tournament

Figure 2 shows the Cinderella Index values for the NCAA Women's tournaments from 2003 (oldest year for which data was easily available) to the current year. The CI values for the Women's tournament are notable for their low values. For example, in 2003 the CI for round 1 was 1.04 meaning that virtually all higher seeds won the 32 first round games. The CI values for the first round of all years are smaller than the smallest first round CI value in the Men's tournament. This result of relatively small CI values generally holds true for all years except for round 4 of the 2004 when a #1, #2, #4 and #7 seed advanced to the Final Four for a 3.5 CI. The cumulative CI for all rounds of each Women's tournament is also smaller than any cumulative CI for all rounds of any Men's tournament.

The sharp contrast in CI results between the Men's and Women's tournaments suggests that the Women's tournament is much more likely to hold form with "Cinderellas" less welcome in this Dance. One possible cause for this difference is a sharper distinction between the top and bottom seeds in the Women's tournament compared to the Men's. An obvious point of interest is to determine the Cinderella Index for a longer period of time.

5 Summary

This work defines a new "Cinderella Index" metric to measure the degree of upsets that occur in the NCAA basketball tournament each year. It is a useful and interesting metric as it translates the yearly discussion about upsets into an objective measure that can be compared between years.

Applying the metric to the NCAA Men's tournament over the past five years shows it is dynamic as its relative values can change from round to round. Lots of upsets in one round do not necessarily correspond to upsets in the next, while lesser Cinderellas may advance far into a tournament

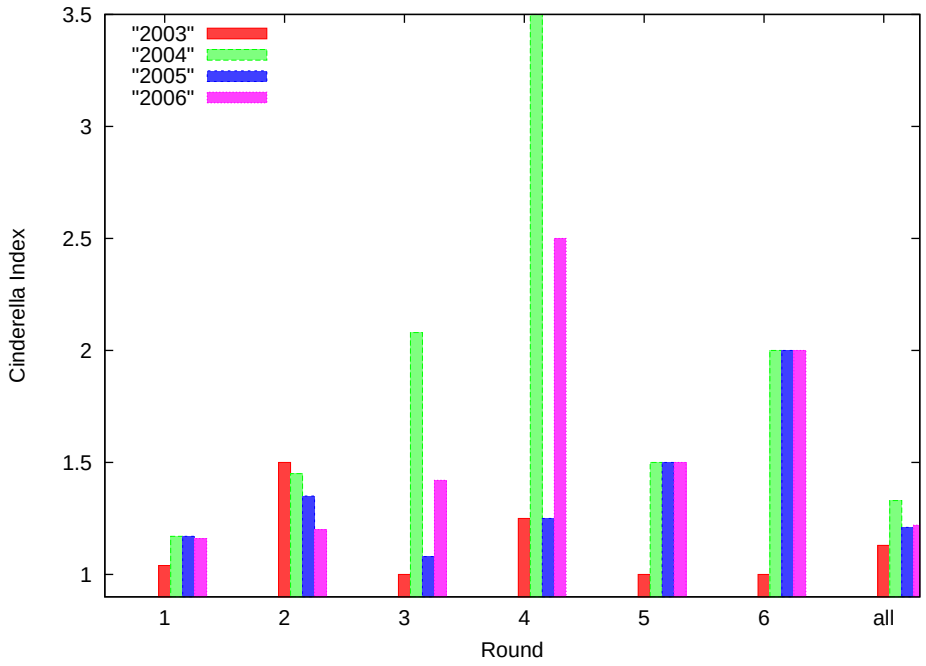

Figure 2: Cinderella Index for the NCAA Women's Basketball Tournament

and have a bigger impact on the Cinderella Index in later rounds. It is of particular interest that the Index for the first two rounds of the 2006 is at its highest in five years.

In applying the same metric to the NCAA Women's tournament we generally see a much smaller Index value than the Men's tournament across all rounds each year. This difference is interesting and the cause needs to be explored further.