Seed Distributions for the NCAA Men's Basketball Tournament: Why it May Not Matter Who Plays Whom*

Sheldon H. Jacobson
Department of Computer Science
University of Illinois at Urbana-Champaign
shj@illinois.edu
https://netfiles.uiuc.edu/shj/www/shj.html

* Joint work with, Alexander G. Nikolaev, Adrian, J. Lee, Douglas M. King

NCAA Men's Basketball Tournament

- National Collegiate Athletic Association (NCAA) Men's DI College Basketball Tournament (aka March Madness)
  - First held in 1939 with 8 teams
  - Since 1985, 64 teams participate annually
    - Increased to 68 teams with four play-in games (2011)

- Popularity of gambling on tournament games
  - Estimated $2.25B (US) wagered on 2007 Final Four through illegal channels alone
  - Common types of gambling: traditional (single game) and office pool (entire tournament bracket)
  - Goal: Forecast the winners of one or more tournament games

Predicting Game Winners

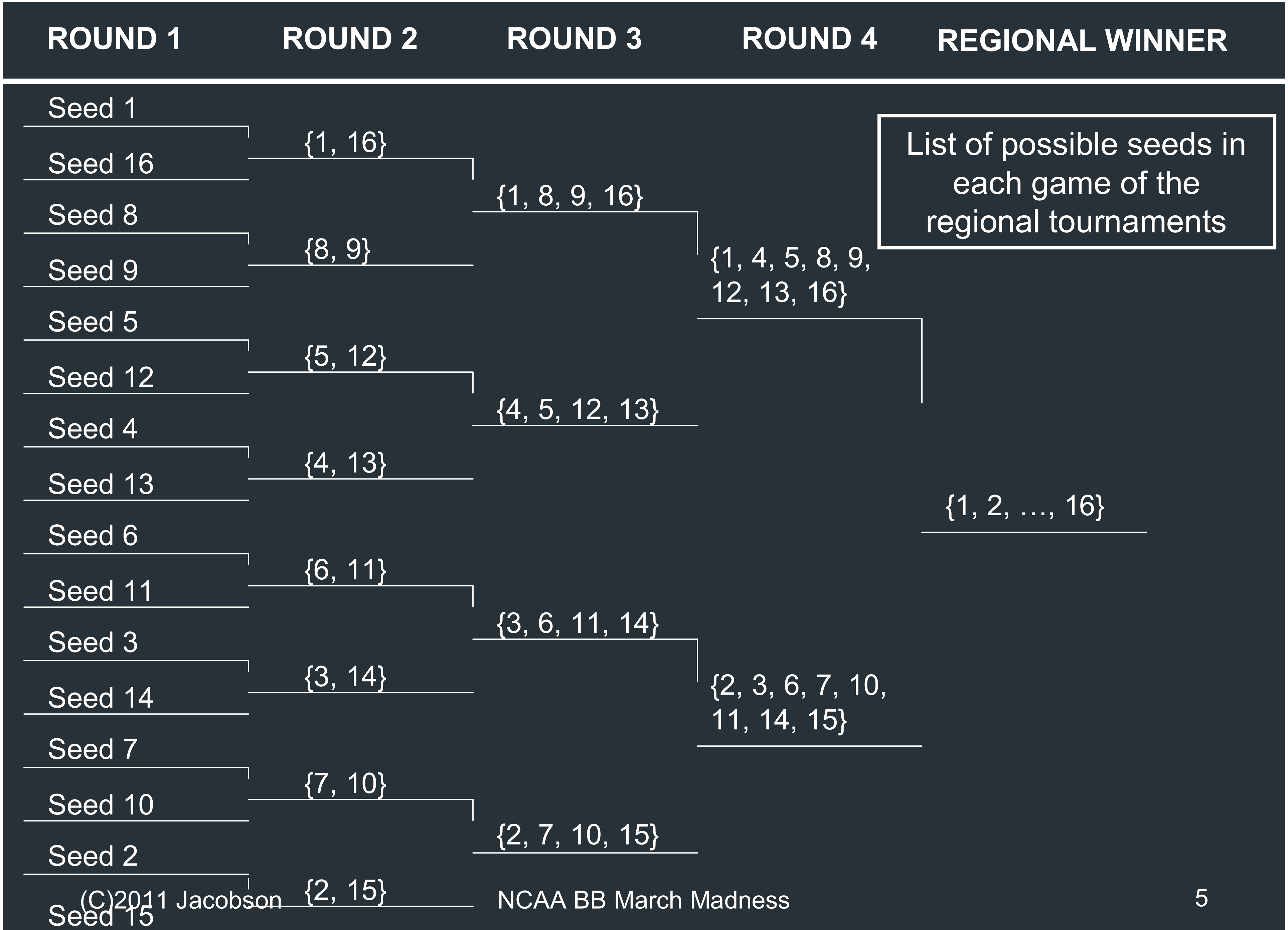
Models have been proposed to forecast game winners (e.g., binary win/lose, final score difference)

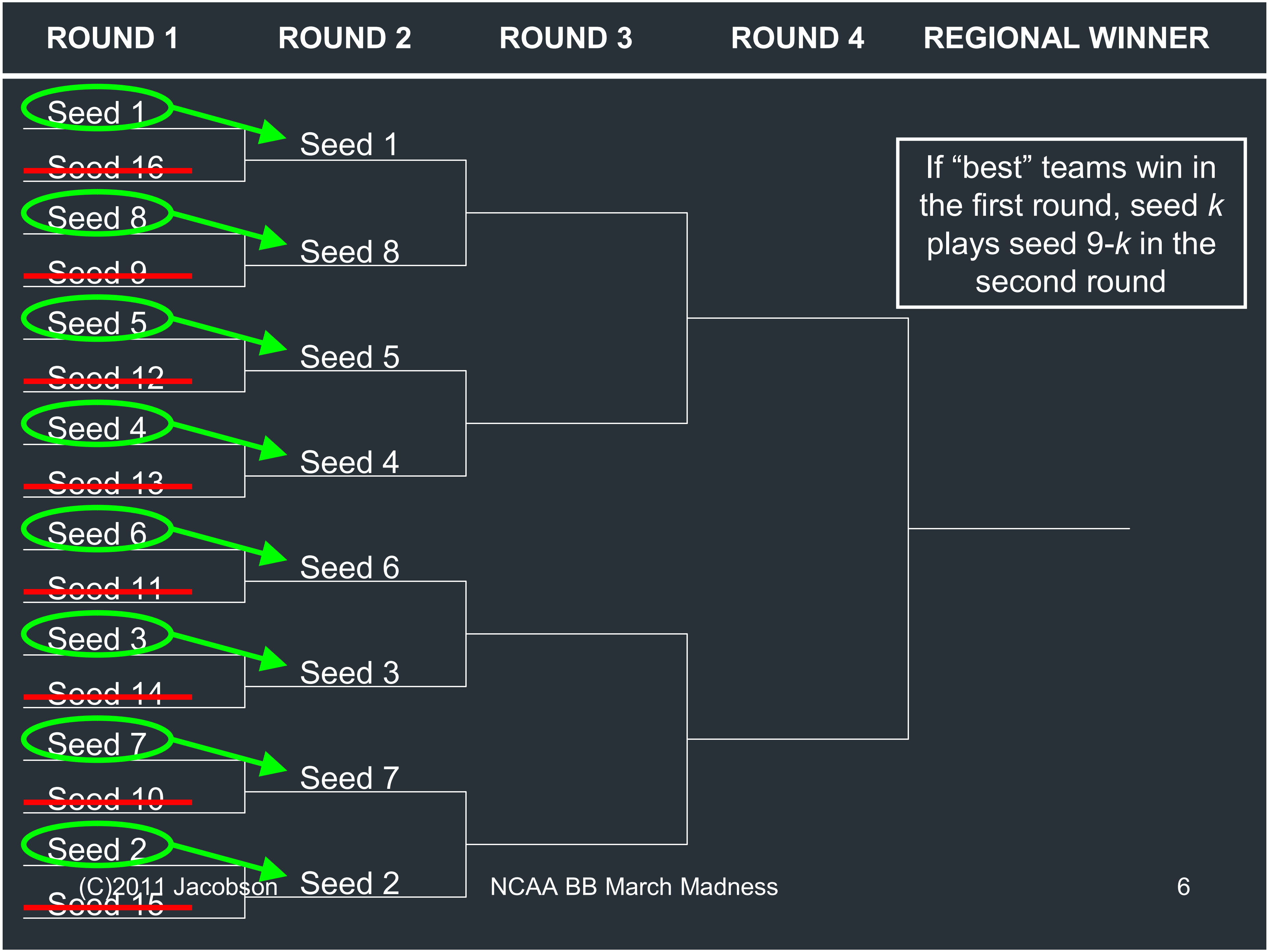
- Predictors:
  - Outcomes of season games (winner, score)
  - Las Vegas odds
  - Other rankings (RPI, Sagarin, Massey, Pomeroy)

- Useful to the general public?
  - Difficult to gather relevant predictor data and implement the model
  - Simple alternatives are attractive

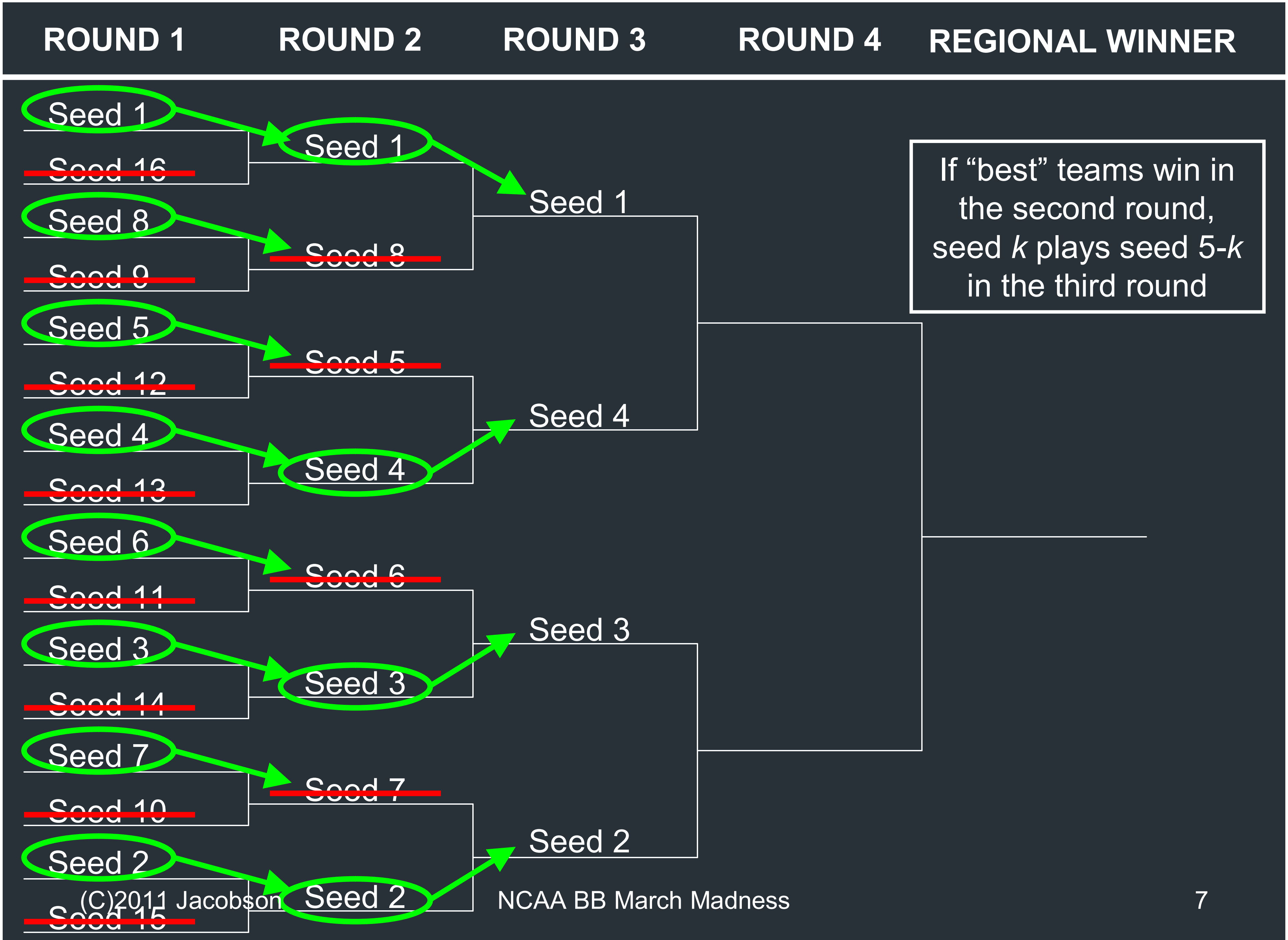
Tournament Structure

- Selection committee
  - Chooses 37 “at large” participants (31 conference champions)
  - Creates 4 regions of 16 teams each (plus 4 play-in game teams)
  - Assigns an integer seed to each team in each region, with values from 1 (best) to 16 (worst)
  - Several issues unrelated to team skill are considered (geography, conference affiliation) when placing teams in regions
- Format of the bracket in each region
  - Single elimination
  - First round: seed k plays seed $17-k$
  - Later rounds: opponents determined by results of earlier rounds

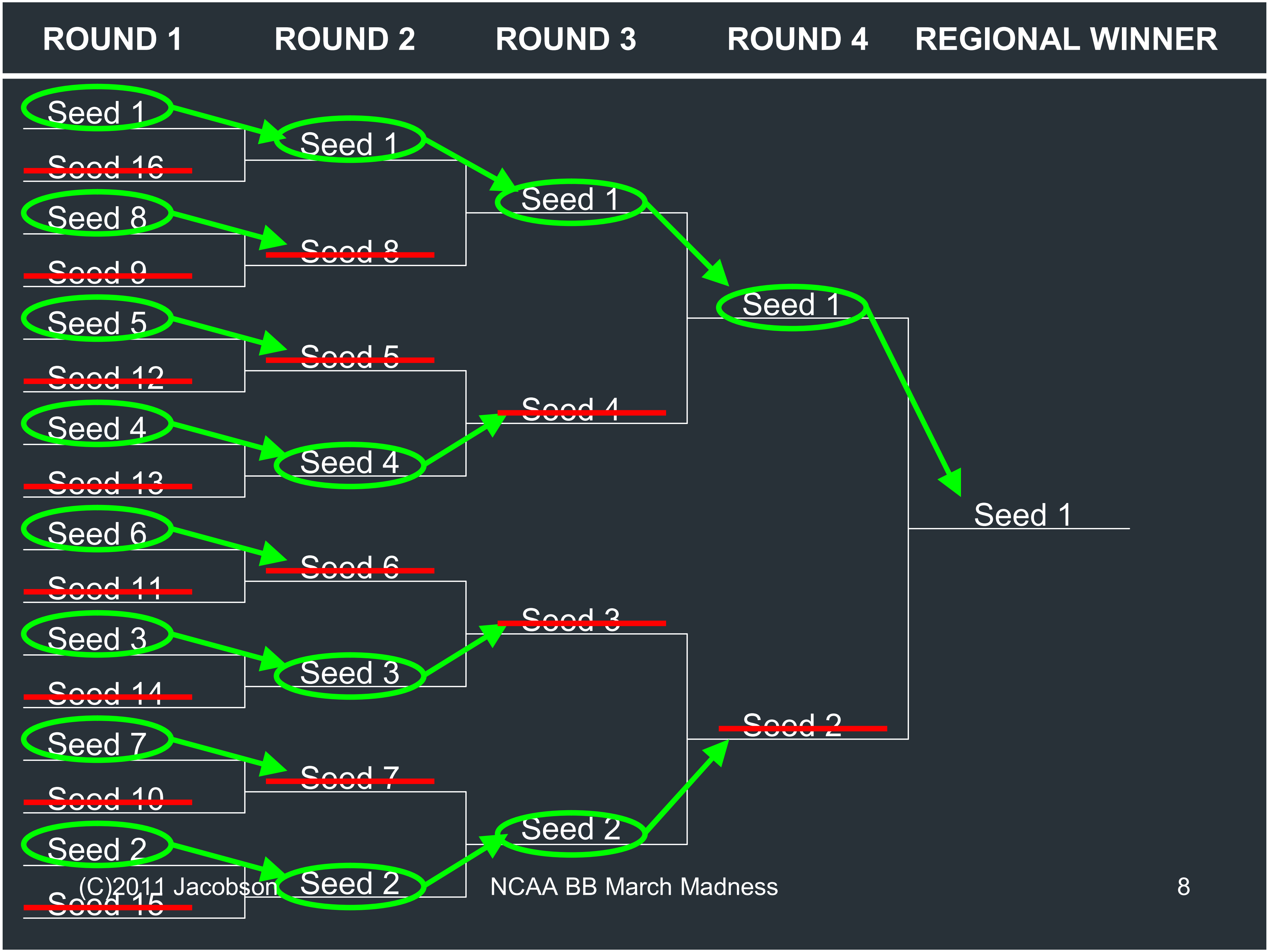

ROUND 1
ROUND 2
ROUND 3
ROUND 4
REGIONAL WINNER
List of possible seeds in each game of the regional tournaments
Seed 1
Seed 16
Seed 8
Seed 9
Seed 5
Seed 12
Seed 4
Seed 13
Seed 6
Seed 11
Seed 3
Seed 14
Seed 7
Seed 10
Seed 2
Seed 15
{1, 16}
{8, 9}
{5, 12}
{4, 13}
{6, 11}
{3, 14}
{7, 10}
{2, 15}
{1, 8, 9, 16}
{4, 5, 12, 13}
{3, 6, 11, 14}
{2, 7, 10, 15}
{1, 4, 5, 8, 9, 12, 13, 16}
{2, 3, 6, 7, 10, 11, 14, 15}
{1, 2, …, 16}



If “best” teams win in the first round, seed *k* plays seed 9-*k* in the second round

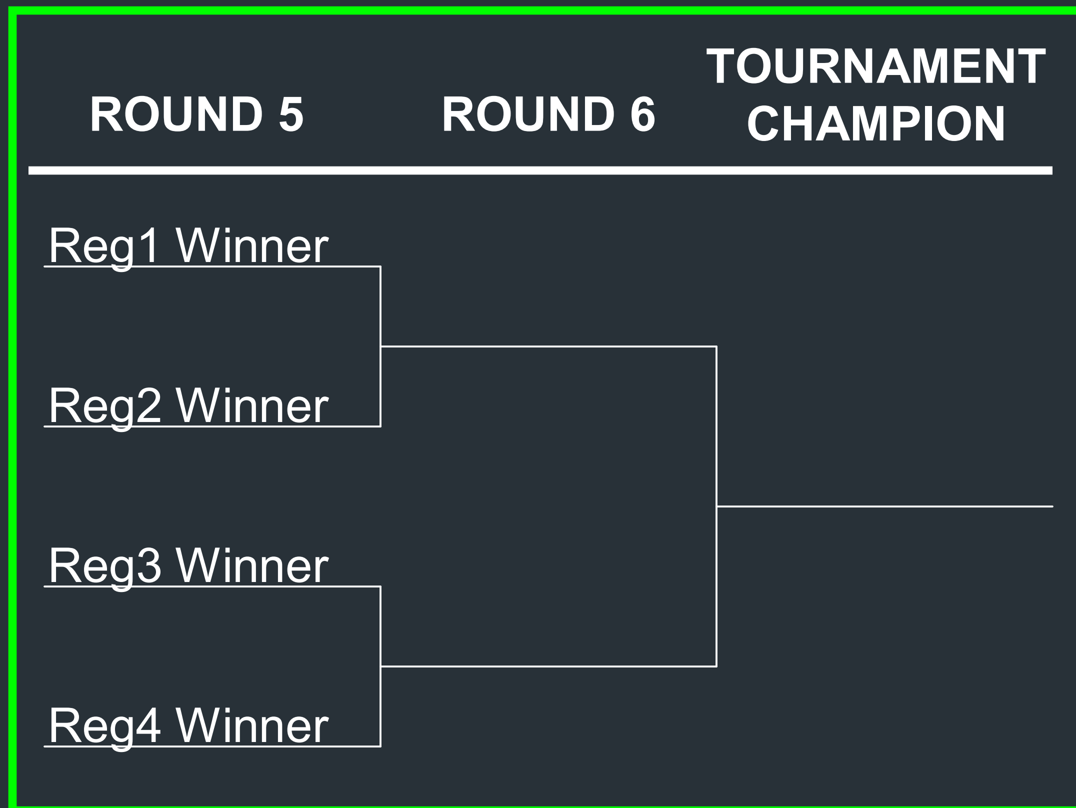


If “best” teams win in the second round, seed *k* plays seed 5-*k* in the third round


ROUND 1
ROUND 2
ROUND 3
ROUND 4
REGIONAL WINNER
Seed 1
Seed 16
Seed 8
Seed 9
Seed 5
Seed 12
Seed 4
Seed 13
Seed 6
Seed 11
Seed 3
Seed 14
Seed 7
Seed 10
Seed 2
Seed 15
Seed 1
Seed 8
Seed 5
Seed 4
Seed 6
Seed 3
Seed 7
Seed 2
Seed 1
Seed 4
Seed 3
Seed 2
Seed 1
Seed 2
Seed 1

The Final Four

- Four regional winners meet in two more rounds
- Two identical seeds can play in a single game
- Any seed can play against any seed
(in theory)



Is It Best To Pick the Better Seed?

- One way to forecast winners**: Pick the better seed**
  - Simplicity of this method makes it attractive
  - Does it provide good predictions?
- Selection committee tends to assign better seeds to better teams
- When seed differences are large, games tend to be more predictable (and hence, fewer upsets)

Predictions by Round

- As the tournament progresses, seed differences tend to be smaller
  - 70% in round 4 (Elite Eight) have been seeded No. 3 or better
  - 76% in round 5 (National Semi-final) have been seeded No. 3 or better
  - 83% in round 6 (National Final) have been seeded No. 3 or better
  - 89% of tournament champions have been seeded No. 3 or better

- Other indicators of success?
  - To appear in the r^{th} round, a team must have won its preceding r-1 games
  - Teams with worse seeds tend to face more skilled competition earlier in the tournament

- Are seed less informative as tournament progresses?
  - Jacobson and King (2009) focus on the top three seeds.

Goals of the Study

- Compare historical performance of the seed distributions in each round.

- Model the seed distributions in each round

- Comparisons model with statistical hypothesis testing
  - X^2 Goodness-of-fit

- Data Sources
  - NCAA: Historical tournament results (1985 – 2010)

Statistical Hypothesis Testing Requirements

- A sufficient number of samples
  - 1,638 total games (63 games over 26 years)
    - Play-in and First Four games not included
  - When subsets are taken based on seeds and rounds, sample sizes drop dramatically
- A random sample. To this effect, assume:
  - Historical data are a representative sample of each seed's performance
  - Each seed has a constant probability of winning against any other seed in a specified round

The Math
Behind The Numbers

Geometric Distribution

- Common (nonnegative) discrete random variable.
- Defined as the number of independent and identically distributed Bernoulli random variables (with probability p) until the first success occurs.
- If Y is distributed geometric with probability p, then

$$P\{Y=k\} = (1-p)^{k-1}p,\ k=1,2,\ldots.$$

Key Theorem*

Let X_1, X_2, … be an arbitrary sequence of Bernoulli trials. Let Z be the number of these Bernoulli trials until the first success. Then Z is a geometric random variable with probability p iff

$$P\{X_i = 1 \mid \Sigma_{h=1,2,\ldots,i-1} X_h = 0\} = p \text{ for all } i = 1,2,\ldots.$$

Implication: Provides a N&S condition for a geometric RV.

Intuition: If the first i-1 seed positions have not advanced to the next round (i.e., won), then the probability that the ith seed position advances is p, the same value for all seed positions i.

* Shishebor and Towhidi (2004)

Sets of Seeds in Each Round

- Possible seeds defined by *sets of seeds* in each round
  – First round: Seed No. n plays Seed No. 17-n, n = 1,2,…,8

- Rounds r = 1,2,3:
  - 2^{4-r} non-overlapping sets of 2^r possible winners
    - r = 1: {1,16} {2,15}, {3,14}, {4,13}, {5,12}, {6,11}, {7,10}, {8,9}
    - r = 2: {1,8,9,16}, {2,7,10,15}, {3,6,11,14}, {4,5,12,13}
    - r = 3: {1, 4,5, 8,9, 12,13,16}, {2,3,6,7,10,11,14,15}

- Rounds r = 4,5,6:
  - One set of 16 possible winners

Define $Z_{j,r}$ as the j^{th} set in the r^{th} round
Define $t_{i,j,r}$ as the i^{th} element in set $Z_{j,r}$

Truncated Geometric Distribution

Truncate the geometric distribution (finite number of seeds)

– Ensure that discrete probabilities sum to one

For set j in round r, $P\{Z_{j,r} = t_{i,j,r}\} = \kappa_{j,r}\, p_{j,r}\, (1-p_{j,r})^{i-1}$

- $i = 1,2,\ldots,\min\{2^r,16\}$ (position in set)
- $j = 1,2,\ldots,\max\{2^{4-r},1\}$ (set in round)
- $r = 1,2,\ldots,6$ (round in tournament)

– Coefficients:

- $\kappa_{j,r} = 1/(1-(1-p_{j,r})^{2\wedge r})$ for set $j = 1,2,\ldots, 2^{4-r}$ in round $r = 1,2,3$
- $\kappa_r = 1/(1-(1-p_{r,1})^{16})$ for round $r = 4,5,6$ (only one set $j = 1$).

Important Note:

$p_{j,r}$ must be estimated for *each position in each set in each round*

Geometric Distribution Validation: Values for $p_{j,r}$

Round r	Set j	Position i	$p_{j,r}$
1	1,2,3,4,5,6,7,8	1	(1.00, .961, .846, .786, .663, .683, .596, .461)
2	1,2,3,4	1	(.875, .644, .510, .423)
		2	(.692, .486, .725, .633)
3	1,2	1	(.721, .462)
		2	(.483, .464)
		3	(.467, .433)
		4	(.750, .353)
4	1	1	(.433)
		2	(.390)
		3	(.361)
		4	(.391)
		5	(.429)
		6	(.375)
5	1	1	(.481)
		2	(.407)
		3	(.500)
6	1	1	(.615)
		2	(.400)
		3	(.500)

Probability of Seed Combinations

$R(r) = 2^{6-r}$ = number of teams that win in round r = 1,2,…,6.

- Teams that advance to the next round

Given that there are four nonoverlapping regions, there are

- four independent geometric rv's for each set in round r = 1,2,3,4,
- two independent geometric rv's for r = 5,
- one geometric rv's for r = 6

Probability of seed combinations in a round are computed by taking the product of

- Probabilities of each seed appearing in that round
- Number of distinct permutations that the four seeds can assume in set j in round r across the four regions

Estimates for $p_{j,r}$

- Estimates for $p_{j,r}$ computed by method of moments
- Y(n,p) truncated geometric with parameter p and n

$$E(Y(n,p)) = (1/p) – n(1-p)^n/(1-(1-p)^n)$$

- Iterative bisection algorithm used to solve for an estimate of $p_{j,r}$ using the average seed position over the past 26 tournaments in each set (j) within each round (r)

Round r	Set j	$p_{j,r}$
3 (Elite Eight)	1,2	(.684, .455)
4 (Final Four)	1	(.400)
5 (National Finals)	1	(.456)
6 (National Champion)	1	(.510)

The Final Four

Seed Frequency in Final Four

Seed n	No. Times Actually Appeared	Expected No. Times Should Appear	δ_n
1	45	41.6	0.28
2	23	25.0	0.15
3	13	15.0	0.26
4	9	9.0	0.00
5	6	5.4	0.07
6	3	3.2	0.02
7	0	1.9	1.94
8	3	1.2	2.89
9	0	0.7	0.70
10	0	0.4	0.42
11	2	0.3	12.15
12	0	0.2	0.15
13	0	0.2	0.09
14	0	0.1	0.05
15	0	0.0	0.03
16	0	0.0	0.02

$$\delta_1 = \frac{(45-41.6)^2}{41.6} = 0.28$$

Standardized Measure

Final Four Seed Combinations

- Compute probability of Final Four seed combinations
- Reciprocal is expected frequency between occurrences

Scenario	Probabilty	Expected # Occurrences	# Actual Occurrences	Expected Frequency (years)
Zero No. 1 Seeds	0.130	3.4	1	7.70
One No. 1 Seed	0.346	9.0	10	2.89
Two No. 1 Seeds	0.346	9.0	11	2.89
Three No. 1 Seeds	0.154	4.0	3	6.49
Four No. 1 Seeds	0.026	0.7	1	38.46

Most Likely Final Four Seed Combinations

Seeds	Actual Occurrences (Tournament Year)	Probability	Expected Frequency (in Years)
1,1,2,3	1991, 2001, 2009	0.066	15
1,1,1,2	1993	0.062	16
1,1,2,2	2007	0.055	18
1,2,2,3	1994, 2004	0.040	25
1,1,1,1	2008	0.026	39
1,2,3,3	1989, 1998, 2003	0.024	42
…		…	…
1,5,8,8	2000	0.0000312	32015

* Compiled based on data from 1985-2010 tournaments

Final Four Seed Combination Odds

Seed Description	Probability	Expected Frequency (years)
One or More 16	0.000756	1307
One or More 15 or 16	0.002037	491
One or More 14, 15, or 16	0.004152	241
One or More 13, 14, 15, or 16	0.007665	130
One or More 12, 13, 14, 15, or 16	0.013493	74
One or More 11, 12, 13, 14, 15, or 16	0.023137	43
All 16's	1.34 E-15	747 Trillion
No teams 1, 2, or 3	0.00220	454
No teams 1 or 2	0.016927	59

* Compiled based on data from 1985-2010 tournaments

2011 Final Four

Odds against any 3,4,8,11 seeds in the Final Four:
    121,000 to 1

Odds against UConn, UKentucky, Butler, VCU in the FF:
    2.9 Million to 1

Probability of UConn (#3) winning the NC: .0306

Number of ESPN Brackets: 5.9 Million

Number who chose UConn: 279,308

Expected number picking UConn, assuming all No. 3 seeds
    are equally likely: 181,000

2011 Final Four

Probability of UKentucky (#4) winning the NC: .0150
Number of ESPN Brackets that chose UKentucky: 107,249
Expected number picking UKentucky, assuming all No. 4 seeds are equally likely: 89,000

Probability of Butler (#8) winning the NC: .00347
Number of ESPN brackets that chose Butler: 4,325
Expected number picking Butler, assuming all No. 8 seeds are equally likely: 5,100

Probability of VCU (#8) winning the NC: .000102
Number of ESPN brackets that chose VCU: 1,023
Expected number picking VCU, assuming all No. 11 seeds are equally likely: 600

Conclusions and Limitations

- Truncated geometric distribution used to compute probability of seed combinations in each round
  - Distribution fits closest (via X^2 goodness of fit test) in later rounds of tournament (Elite Eight and onwards)
- Rule changes may impact seed winning probabilities over time
  - Introduction of 35 second clock
  - Expansion of three point arc
  - Selection committee criteria changes
- Distribution parameters, $p_{j,r}$, must be updated annually following each year's tournament

March Madness

Let the games begin!

http://bracketodds.cs.illinois.edu

Website Developers: Ammar Rizwan and Emon Dai
(Students, Department of Computer Science,
University of Illinois at Urbana-Champaign)

Website Functionality

Uses model to odds against seed combinations in

- **Elite Eight**
- **Final Four**
- **National Finals**
- **National Championship**

Allows one to

- **Compare the relative likelihood of seed combinations**
- **Compute conditional probabilities of seed combinations in the final two rounds.**

Note: Model can do much more than the web site functionality.

Thank you

BracketOdds

http://bracketodds.cs.illinois.edu

Sheldon H. Jacobson, Ph.D.
https://netfiles.uiuc.edu/shj/www/shj.html
(217) 244-7275
Skype: sheldon.jacobson1
shj@illinois.edu