Organization of the NCAA Division I Baseball Tournament

J. Cole Smith
Industrial & Systems Engineering
The University of Florida

Overview

- The tournament structure
- The optimization problem
- *Brief* discussion of our algorithm
- Illustration of results

NCAA College Baseball

- There are 286 teams in Division I college baseball that balance schoolwork and athletics during a regular season of play from February to May.
  - Teams are divided into 30 conferences, with nine “independent” teams excluded.
- A playoff is held during the month of June to determine the National Champion.

Tournament Composition

- 64 of the 286 teams are selected for the tournament.
  - 30 are champions of their conference, as determined by:
    - a end-of-season tournament for 28 conferences
    - the best conference record for 2 conferences.
- The remaining 34 teams are chosen “at large” by a *selection committee*.
- The selection committee then determines the seeding and locations for the 64 teams.

Tournament Seeding

- Each team is a #1, #2, #3, or #4 seed (16 of each).
- Additionally, eight of the sixteen #1 seeds are called *national seeds*, and are individually ranked from 1 through 8.

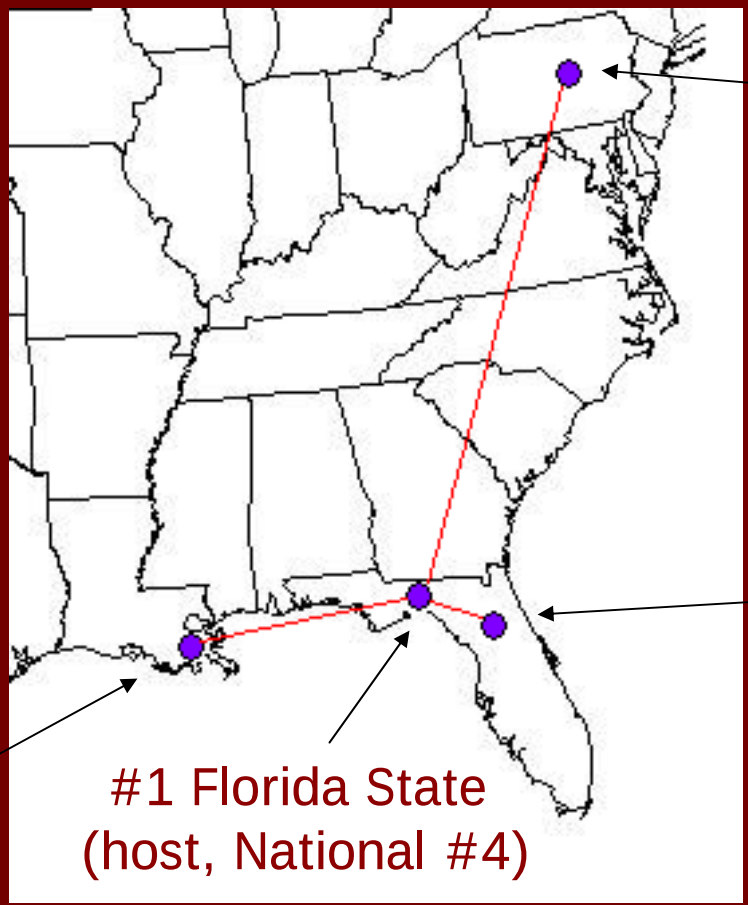
2008 Seedings

#1 seeds	#2 seeds	#3 seeds	#4 seeds
#1 Miami, FL	California, Irvine	Alabama	Bethune-Cookman
#2 North Carolina	California, Los Angeles	Arkansas	Bucknell
#3 Arizona State	Dallas Baptist	California, Berkeley	Columbia
#4 Florida State	East Carolina	Charlotte	Eastern Illinois
#5 Cal. State, Fullerton	Florida	Elon	Eastern Michigan
#6 Rice	Georgia Tech	Houston	Fresno State
#7 Louisiana State	Michigan	Kentucky	Illinois-Chicago
#8 Georgia	Missouri	Louisville	James Madison
Arizona	North Carolina – Wilmington	Mississippi	Lipscomb
Coastal Carolina	Pepperdine	New Orleans	Mt. Saint Mary's
Long Beach State	San Diego	Oklahoma	Rider
Nebraska	South Carolina	Oral Roberts	Sam Houston State
North Carolina State	Southern Mississippi	Saint John's	Stony Brook
Oklahoma State	Texas	Texas Christian	Texas Southern
Stanford	Vanderbilt	Tulane	California, Davis
Texas A&M	Wichita State	Virginia	Western Kentucky

Tournament Organization

- The tournament is organized into three phases
- Phase 1: "Regionals"
  - Sixteen regionals, each hosted by a team participating in the regional
  - Each regional is a group of four teams, each containing a #1, #2, #3, and #4 seed
  - One team advances from each regional

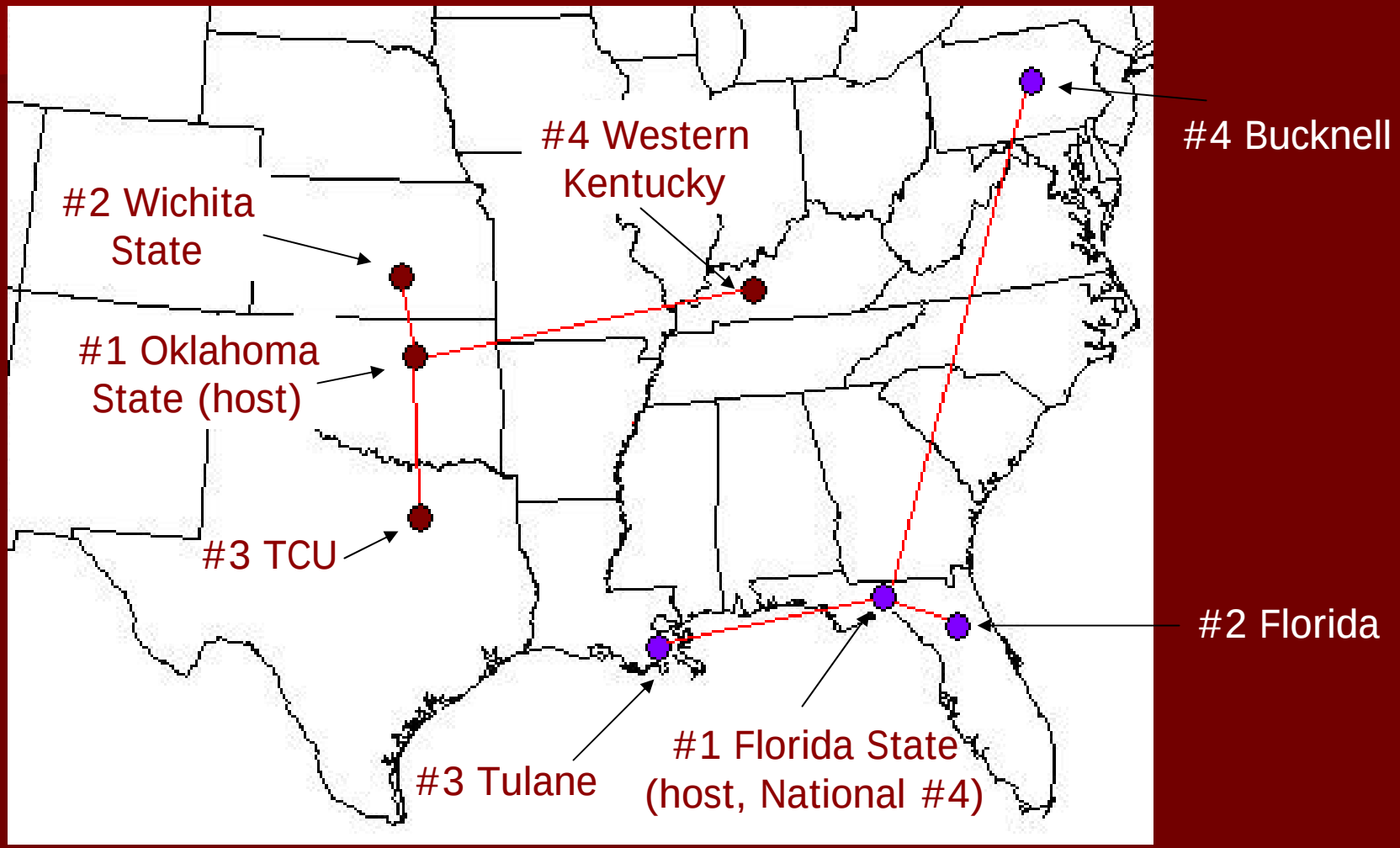
Regional Illustration



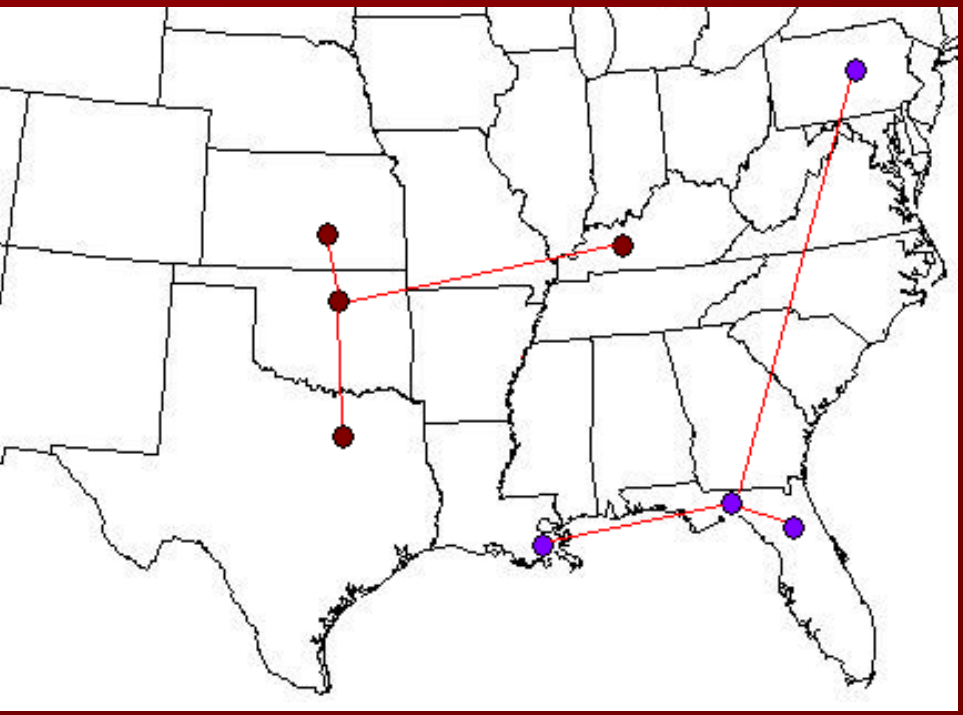
Tournament Organization

- Phase 2: "Super Regionals"
  - **Ahead of time**, the 16 regionals are paired with one another
  - Winners of these regionals compete against one another in the second weekend
  - Super regionals are best-of-three; winner goes to the College World Series

Super Regional Illustration



Super Regional Illustration



The Stillwater, OK regional was paired with the Tallahassee, FL regional.

Champions meet in the super regional

Anticipated matchup:
#1 Oklahoma St at #1 Florida St (N4)

Another possible matchup:
#2 Wichita St at #1 Florida St (N4)

Another possible matchup:
#2 Florida at #1 Oklahoma St

And this is technically possible:
#4 Bucknell at #4 Western Kentucky

Note on hosting

- The first two phases are held on-campus. Who gets to host?
  - Determined by the selection committee.
  - Usually a #1 seed, but not always.
  - Their problem, not ours.

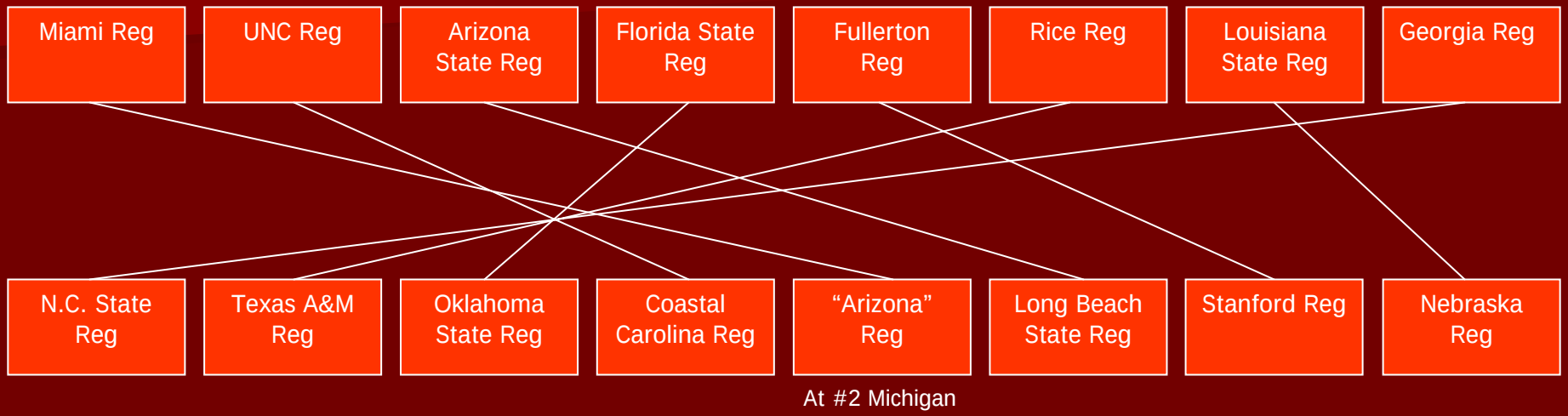
College World Series

- Phase 3: College World Series
  - Eight super regional winners advance to Omaha, Nebraska
  - (Format is actually “two regionals followed by a super regional.”)
  - Must **not** optimize this!
    - To minimize travel, we’d rig regionals for teams like Nebraska, Creighton, etc.!

Tournament Organization

- None of the sixteen regionals can contain teams from the same conference.
- The sixteen regionals are paired with one another *a priori*. (That's what makes this hard!)
- National seeds cannot be paired with other national seeds: Each pairing is between a #1 seed that has a national seed and a #1 seed that does not.
- #1 seeds that are paired with one another cannot belong to the same conference

Illustration



Example matching:

#1 Georgia (#8 nat'l)	#1 N. C. State
#2 Georgia Tech	#2 East Carolina
#3 Louisville	#3 St. John's
#4 Mt. Saint Mary's	#4 Stony Brook

Teams belong to the the same conference

Optimization Problem

- Input from the committee:
  - 64 tournament teams
  - All seedings
  - Hosts for each regional
- Organize assignment of teams to regional (a host is automatically assigned to its own regional).
- Give “super regional” pairings of regionals.
- Minimize **regional travel cost** plus **expected super regional travel cost**.

Illustration of Travel Cost

Example matching:

#1 Florida State (N4)	#1 Oklahoma State
#2 Florida	#2 Wichita State
#3 Tulane	#3 Texas Christian
#4 Bucknell	#4 Western Kentucky

First-phase cost:

distance(Florida – Florida State) + distance(Wichita State – Oklahoma State) +
distance(Tulane – Florida State) + distance(Texas Christian – Oklahoma State) +
distance(Bucknell – Florida State) + distance(Western Kentucky – Oklahoma State)

Second-phase cost (16 terms):

distance(Florida St – Oklahoma St) × prob(Florida St and Oklahoma St advance) +
distance(Florida St – Wichita St) × prob(Florida St and Wichita St advance) +
distance(Florida St – Western Ky) × prob(Florida St and Western Ky advance) +
distance(Florida – Oklahoma St) × prob(Florida and Oklahoma St advance) +
distance(Florida – Wichita St) × prob(Florida and Wichita St advance) +
… +
distance(Bucknell – Western Ky) × prob(Bucknell and Western Ky advance)

Probabilities

- Need to create advancing probabilities: depends on who else is present in your regional!
- Give every team a power rating
  - Ex: Florida State = 91, Florida = 47/2, Tulane = 20/3, Bucknell = 6/4.
  - Need more details? Off-line, it's boring.
- Probability of advancing:
  - (Team's power)/(total regional power).
  - FSU: 74%, Florida: 19%, Tulane: 6%, Bucknell: 1%
  - These are pretty close to historical success rates, but they're very much back-of-the-envelope...

One more consideration

- We have a basketball tournament paper too
- There, “all 1-seeds are created equally” (and 2-seeds, and 3-seeds, etc.)
- Here, that’s not true.
  - Committee would never pair the strongest #1 with the strongest #2 in basketball, much less baseball
  - Total power in each regional should be balanced
  - We enforce that they have to be within 10% of the average regional power

Model Idea

- Variables $x_{tr} = 1$ if team $t = 1,\ldots,64$ is assigned to regional $r = 1,\ldots,16$, and 0 otherwise.
  - (Fix all hosts to their regions.)
- Variables $y_{uv} = 1$ if regional u is matched with regional v, and 0 otherwise.
- You can imagine the assignment and regional balancing constraints...

The objective

- First-stage assignment plus expected second-stage assignment.

$$\min \sum_{t=1}^{64}\sum_{r=1}^{16} d(t,r)x_{tr} + \sum_{t=1}^{64}\sum_{t'=1}^{64}\sum_{u=1}^{8}\sum_{v=9}^{16} d(t,t')y_{uv}\left(\frac{p_t x_{tu}}{\sum_{a=1}^{64} p_a x_{au}}\right)\left(\frac{p_{t'} x_{t'v}}{\sum_{b=1}^{64} p_b x_{bv}}\right)$$

It's integer, it's nonlinear (nonconvex), can't be linearized, the nonlinear terms are fifth-order polynomial terms, and it's large-scale.

And it's ugly too.

Solution Method Approach

- Make an optimistic guess
- Add cutting planes from nonlinear and assignment analyses
- Refine the guess
- This will eventually yield an optimal solution, but its real purpose is to get us close and prove how close we are

Solutions

Year	Actual Expected Travel	Optimized Expected Travel	Percentage Improvement
2006	52,202km	45,786km	12.3%
2007	49,797km	38,472km	22.7%
2008	51,850km	40,855km	21.2%

Long Trips: 2006 and 2007

- Question: do the trip reductions come at the expense of a few unlucky teams?
- For 2006: Regional trips over 2000km?
  - Actual bracket: 9
  - Optimized bracket: 5
  - Worst travel times are about the same
- For 2007: similar results (Actual = 6 trips, Optimized = 3 trips).

Long Trips: 2008

- Actual bracket:
  - Rider – Fullerton 3884
  - Virginia – Fullerton 3553
  - Stony Brook – Arizona St. 3509
  - Arizona – Michigan 2616
  - Arkansas – Stanford 2489
  - Vanderbilt – Arizona St. 2311
  - St. John's – Rice 2305
- Optimized bracket:
  - Bucknell – Long Beach St. 3736
  - Arizona – Michigan 2616
  - Arkansas – Stanford 2489
  - Eastern Illinois – Arizona St. 2226

Our Tournament

1	Miami, FL	Arizona
	Florida	Michigan
	Charlotte	Kentucky
	Bethune-Cookman	Eastern Michigan
2	North Carolina	Coastal Carolina
	South Carolina	UNC-Wilmington
	Elon	Virginia
	Rider	Columbia
3	Arizona St.	Long Beach St.
	U.C. Irvine	San Diego
	Texas Christian	California
	Eastern Illinois	Bucknell
4	Florida St.	Oklahoma St.
	Southern Mississippi	Vanderbilt
	Alabama	Oral Roberts
	James Madison	Western Kentucky

Our Tournament

5	Cal St. Fullerton	Stanford
	UCLA	Pepperdine
	Oklahoma	Arkansas
	Fresno St.	U.C. Davis
6	Rice	Texas A&M
	Texas	Dallas Baptist
	New Orleans	Houston
	Texas Southern	Sam Houston St.
7	LSU	Nebraska
	Missouri	Wichita St.
	Tulane	Mississippi
	Lipscomb	Illinois-Chicago
8	Georgia	North Carolina St.
	Georgia Tech	East Carolina
	Louisville	St. John's
	Mt. Saint Mary's	Stony Brook

Caveats & Disclaimers

- Solutions are provably within 1% of optimality, and are found very quickly
- Ideally, they would be used in an iterative "man-in-the-loop" fashion:
  - Obtain a bracket
  - Disallow "bad" pairings, re-solve, and repeat
    - Perhaps try to avoid rematches?
    - But even now, Rice and Texas A&M always seem to be paired up, for instance…
  - Can also be used as a what-if tool
    - "How much worse is the solution if we split these teams?"

Impact and Summary

- Quickly finding a feasible solution can be hard.
- Time pressures can force bad decisions in manual solutions, and lead to premature design of tournaments.
- Manual tournament design lends itself to biases and suboptimal decisions.
- With college baseball becoming more popular, and travel costs increasing, an optimization tool should be used to **help** the selection committee.