Data Analytics, Sports and Optimization

William Pulleyblank

Mathematical Sciences

USMA, West Point

Agenda

1. Scheduling scouts for the New York Yankees
2. Data Analytics and NCAA Div.I Football
3. Restructuring the NHL

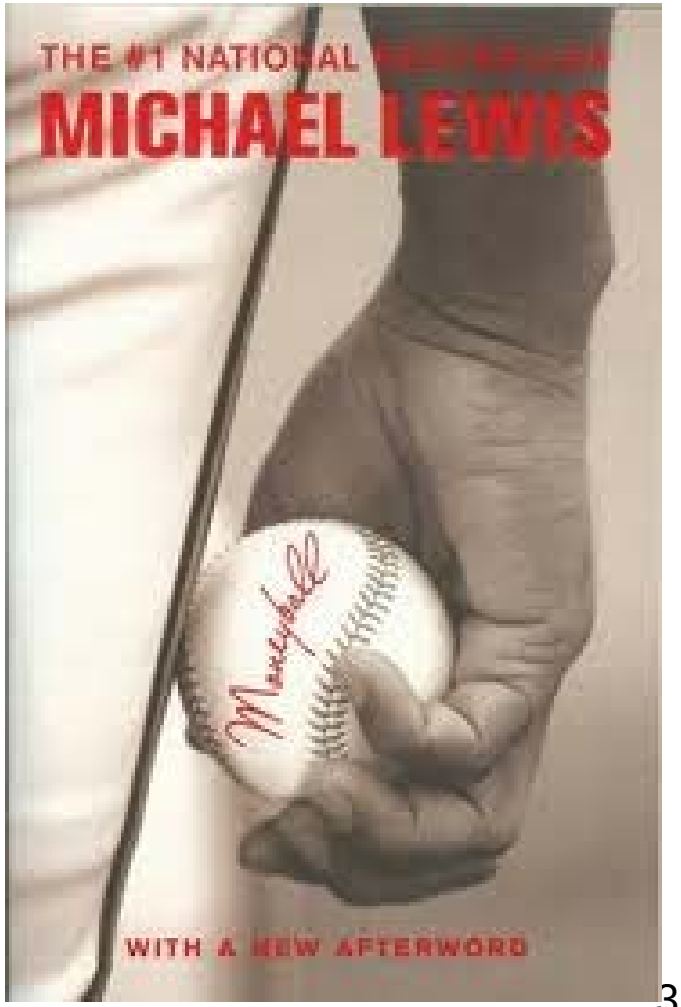
Sports analytics

Suppose you knew:

- for every team,
  - the history of every player,
- for every game,
  - conditions,
  - every event in the game,
  - the outcome of every play

And suppose you could use all this information to make all the decisions in the game -

Could it give you an advantage?



1. Scheduling Scouts for the New York Yankees

- 2LT Ryan Davis
- WRP
- MAJ Chris Marks
- LTC Mark Wood

Scouting Process

- Three ways to acquire player rights
  - Amateur Draft
  - Trades
  - Free Agency
- Draft talent evaluation is critical and difficult
  - Expenses
    - Average signing bonus for a first round draft selection is now $2 million
  - Development System of MLB
    - From 64 players selected in first two rounds of 2007, only one had played an inning in MLB at the end of 2008
    - From 917 players selected from 1981-2005, one third (32.7%) have not played an inning in MLB
  - Size
    - 50 rounds, over 1,500 players selected
      (NHL: 215 players, NBA: 60 players, NFL: 256 players)

Scouting Staff and Responsibilities

- 22 person scouting staff:
  - 1 scouting director
  - 4 national cross-checkers
  - 17 regional scouts
- Responsibilities of scouts
  - Scouting Director: overseeing player evaluation
  - National Cross-checkers: cross-checking reports of regional scouts
  - Regional Scouts: evaluate players in defined geographical regions, provide reports on possible draftees
    - Example of a defined region: Florida, Georgia, Alabama, Mississippi

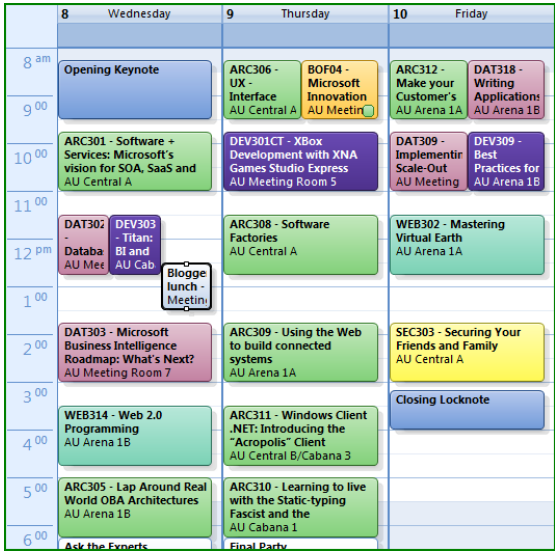
Current Process

*Dynamic, ever-changing environment requires a flexible solution

Yankee Scouting staff: 17 regional scouts, 4 national cross-checkers, 1 scouting director for 22 total

Cross-checker's job is to validate and integrate ratings across the national pool of players

Cross-checker's have to create their own schedule seeing certain caliber of players



Current system is to choose game manually and call them in

Scouting Value

What is the value of attending a game?

Two components: *Player Value* and *Confidence Rating*

Player Value

•Regional scouts grade the five tools of baseball: arm strength, fielding ability, hitting for average, hitting for power, and speed

•Numeric value ranging from 20-80

•The scale is broken down into letter grades

Confidence Rating

•Given as a percentage, describes the level of confidence the scouts have in the grade given

•Based on :
- –Number of times player seen
- –Distribution of times seen
- –Caliber of competition when player evaluated
- –Potential for player to improve or decline

Data Construction and Collection

Goal: Provide a method for data collection that allows for easy manual access to changing game values

Game

	A	B	C	D	E	F	G
1	ID	Date	Time	Team1	Team 2	Game Value	Game Area Scout
2	1	40622.00	0.46	SAN DIEGO	OREGON	48	David Keith
3	2	40622.00	0.46	NOTRE DAME	GONZAGA	40	Mike Thurman
4	3	40622.00	0.50	DAYTON	SIENA	19	Michael Gibbons
5	4	40622.00	0.50	WAYNE STATE	ARMY	6	Matt Hyde
6	5	40622.00	0.50	MARSHALL	ST BONAVENTURE	27	Michael Gibbons

Game Value = $\sum$ Team Values = $\sum$ Scouting Values

Player Profile

A	B	D	E	F	G	H	I	J	K	L	M	N
School	PFirstLast	SYear	Group	HF	HI	Position	Role	LastRanking	#Seen	LastSeen	ConfidenceRtg	ScoutsSeen
A	Ricardo Jacquez	HS	D	5	9	RHP	Reliever	2	4	7/25/2011	0.891	2
ACADEMY OF THE ANGELS	Joseph Loftus	HS	E	6	3	3B		1	6	12/19/2011	0.846	3
ACTON-BOXBOROUGH	Scott Weismann	HS	C	6	1	RHP		5	5	12/18/2011	0.218	2
AHUNTSIC COL	Jesen Dygestile-Therrien	JC1	C+			RHP	Starter	8	7	11/21/2011	0.222	3
ALABAMA	Adam Morgan	C3	C+	6	1	LHP	Starter	8	1	12/8/2011	0.356	1
ALABAMA	Taylor Dugas	C3	C	5	7	CF		5	1	11/10/2011	0.778	1
ALABAMA-BIRMINGHAM	Jamal Austin	C3	C	5	9	CF		5	3	9/24/2011	0.949	1
ALBANY	David Kubiak	C4	D			RHP		2	5	5/7/2011	0.348	2
ALBANY	Zachary Kraham	C3	C			RHP	Starter	5	5	1/12/2011	0.185	2

Scheduling Optimization

Provides the national cross-checkers with suggested schedules from a week's slate of games

G is the set of all games played in a week

G_d is the set of games played on day d

S is the set of all scouts

$$x_{ig} = \begin{cases} 1 & \text{scout } i \text{ sees game } g \\ 0 & \text{scout } i \text{ does not see game } g \end{cases}$$

$$maximize \sum_{i \in S} \sum_{g \in G} val(g) * x_{ig}$$

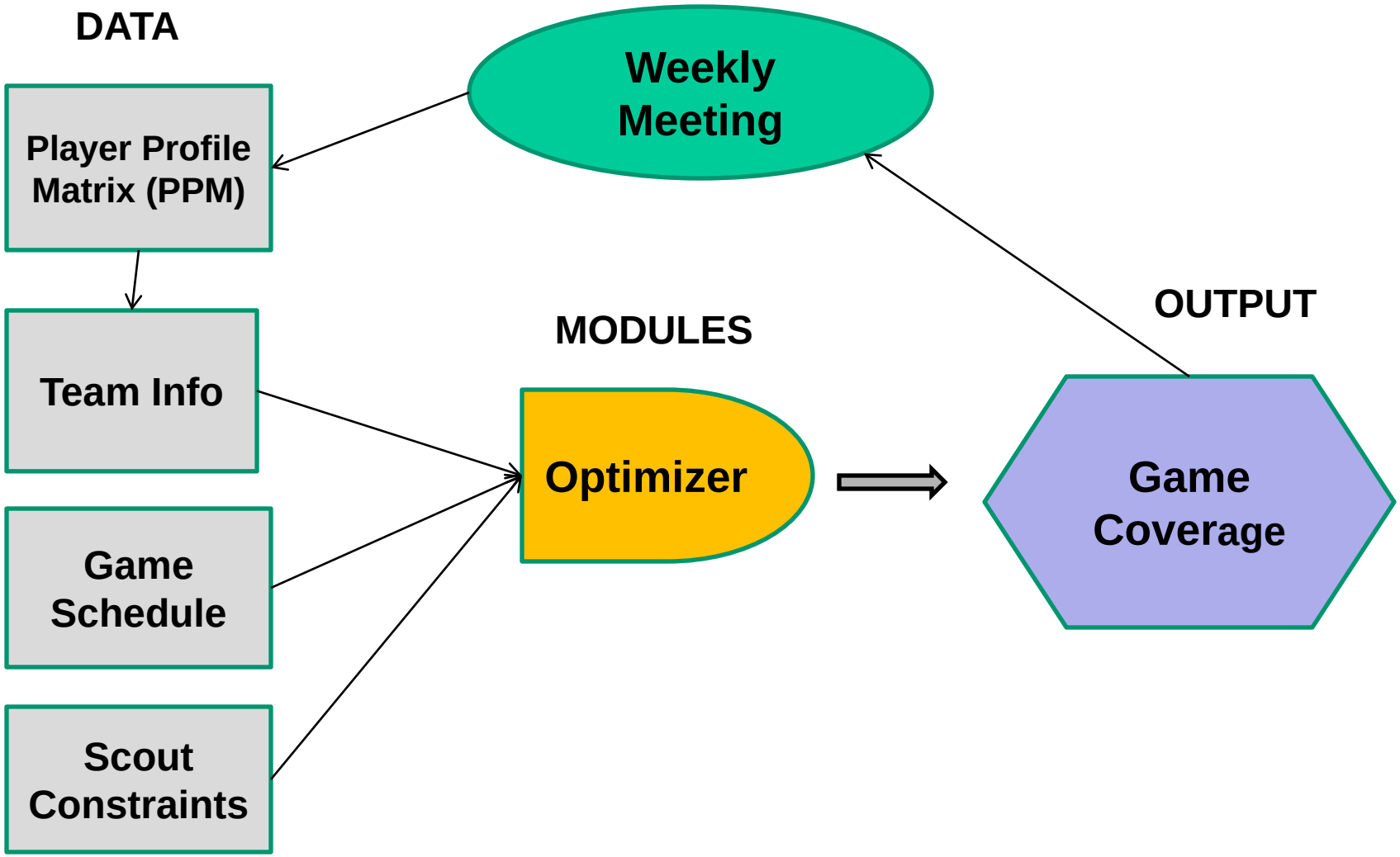
$$\sum_{i \in S} x_{ig} \leq 1 \quad for\ each\ g \in G$$

$$\sum_{g \in G_d} x_{ig} \leq 1 \quad for\ each\ i \in S\ for\ each\ d$$

Refinements

- Team value discounted by 20% if team seen in prior two weeks
- Scouts are permitted to see more than one game in a day, provided that it is feasible
- A team cannot be seen more than once by any scout in a week.

Weekly Process



Results

Cross-Checker 1		Cross-Checker 2		Cross-Checker 3	
Game	Value	Game	Value	Game	Value
46	75	27	74	24	75
54	41	51	57	52	56
87	73	62	74	71	73
117	80	116	73	127	74
143	74	148	76	152	77
160	54	180	79	165	76
219	75	233	77	237	74
Total	472	Total	510	Total	485
Average	67.42857	Average	72.85714	Average	69.28571

24	40622.00	0.54	KENTUCKY	Tennessee	75	Michael Gibbons

KENTUCKY	Alex Meyer	MGIS-74AV52	C3	A-	6	7	RHP	Reliever	16	2	2/24/2011	0.466
KENTUCKY	Braden Kapteyn	MGIS-75HPVH	C3	D	6	4	RHP	Reliever	2	8	4/20/2011	0.444

TENNESSEE	Matthew Ramsey	DJSK-76GPAW	C2	A	5	11	RHP	Reliever	18	6	1/16/2011	0.160
TENNESSEE	Steven Gruver	MGIS-74B22C	C3	B	6	2	LHP	Starter	12	10	11/19/2011	0.039
TENNESSEE	William Locante	SLOP-87DNBY	C3	A-	6	0	LHP	Reliever	16	0	3/10/2011	0.604

Validation : Working with Scouting Staff for the NY Yankees

Conclusion

- Further Actions
  - Expand this process to produce schedules for the regional scouts
  - Impose soft constraints on the players different scouts evaluate
  - Address the distance traveled by each cross-checker

2. Data Analytics and NCAA Div. I Football

- 2LT Kirby Kastner
- 2LTJeremy Maness
- WRP
- Dr. Inderpal Bhandari
- Dr. Brian MacDonald
- MAJ Michael Landin

Army Football

Recent History

- 2010 Season and Bowl Game
- 2011 Season
- Triple Option Offense
- Top Ten Rushing Team
  - 3,000+ yards in 2010
  - 4,000+ yards in 2011
- Inconsistent Defense
- Army-Navy – the last regular season college football game each season

Goal

- Analyze Army Football game data and provide useful information to the Army coaching staff
- Coaching staff will use results to improve strategy and game preparation
- Win Commander-in-Chief Trophy
- Beat Navy

Advanced Scout

- I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pratap and K. Ramanujam, Advanced Scout: Data Mining and Knowledge Discovery in NBA Data, *Data Mining and Knowledge Discovery* **1** (1997) 121-125.

- United States Patent 7,110,998, Bhandari , et al. September 19, 2006 **-** *Method and apparatus for finding hidden patterns in the context of querying applications*

Methodology

- Data Collection and Organization
- Feature Selection
- Development of Intuitive User Interfaces
- Evaluation and Refinement of Predictive Models
- Evaluation of Solutions

Data- Offense

DEF	FORM	PLAY	BCPOS	BCREC	DIST	DN	FP	Gain	Hash	Play#
odd	bone rt	sherman 12		8	1	4	9	2		35
odd	bone rt	jim 12		36	4	4	40	2		42
odd	zoom rt				9	4	28			51

PLAYSUMM	QB	QTR	RES	RP	Ser#	Series	TACK1	TACK2	TACKPOS
Steelman, Trent rush up middle for 2 yards to the NIU7, 1ST DOWN ARMY (PROGAR, Sean;SCHILLER, Pat).		2		R	11	7	95	53	
Crucitti, Jon rush R for 2 yards to the NIU38 (DELEGAL, Jordan;MELVIN, Rashaan).		3		R	4	8	29	11	
PENALTY ARMY false start 5 yards to the NIU33.		3		Pen	9	9			

- Used all games from the 2010 and 2011 seasons (25 games)
- Over 1,700 plays and 21 data fields
- Bucketed the distance field to better facilitate *VirtualMiner* analysis.

Data - Defense

BKSET	COMMENT1	COVER	FORM	FRONT	Pers	PLAY	STR	DIST	DN	FP	Gain	Hash	Play#
PSTL		0	BUNCH	QWACK	10		L	10	1	-33	0	L	1
PSTL		0	BULL	QWACK	10	ZN SP	R	10	2	-33	2	L	2
PSTL	PUNT	0	BUNCH	NOSE	10		R	8	3	-35	0	L	3

PLAYSUMM	QB	QTR	RES	RP	Ser#	Series	Series End	TACK1	TACK2	Yards To GL
Higgins, Ryan pass incomplete to Mays, Aaron.	3	1	i	P	1	1				67
Whiting, Darryl rush for 2 yards to the FOR35 (Watts, Zach).		1		R	2	1		40		67
Higgins, Ryan pass incomplete to Pierre, Brad.	3	1	i	P	3	1	PUNT			65

Pivot Tables and Hypothesis Testing

- Very basic data analysis, no data mining
- Should Army pass more on first down?

	Runs		**Passes**				**All Plays**	
	Mean	St. Dev.	Average (Comp)	St. Dev. (Comp)	Average (Total)	St. Dev. (Tot)	Average	St. Dev.
First Down	5.52	6.819	**16.256**	10.161	8.806	11.036	5.817	7.535
Second Down	5.46	7.068	13.839	11.243	6.042	10.097	5.48	7.629
Third Down	4.34	6.032	13.875	8.253	8.325	9.357	4.719	7.255
Fourth Down	3.91	5.691	16	0	4	8	3.268	6.116

Hypothesis Testing

- Developed in consultation with Army defensive coaching staff
- Predicting plays after a sack or tackle for a loss
- Compare to baseline analysis

Baseline Analysis

	1st Down		*2nd Down*		*3rd Down*		*4th Down*	
Personnel Grouping	**% Run 474 total**	**% Pass 286 total**	**% Run 353 total**	**% Pass 218 total**	**% Run 131 total**	**% Pass 198 total**	**% Run 20 total**	**% Pass 21 total**
10	35	65	30	70	16	84	33	67
11	52	48	**54**	**46**	**25**	**75**	45	55
12	71	29	73	27	77	23	67	33
21	64	36	71	29	58	42	25	75
All	62	38	**62**	**38**	**40**	**60**	49	51

- 2nd and 3rd down play calls differ from normal game trends
- Plays run by 11 personnel group differ from normal game trends
- Number of 4th down plays called after a sack/tackle for a loss is insignificant
- Distance to a first down would be another good factor to test

Plays After Sack/Loss

- What will the offense do after Army sacks the quarterback or makes a tackle for a loss?
- Table depicts plays called after a sack or tackle for a loss

	2nd Down		*3rd Down*	
Personnel Grouping	**% Run 22 total**	**% Pass 28 total**	**% Run 3 total**	**% Pass 28 total**
10	25	75	0	100
11	35	65	11	89
12	71	29	50	50
21	60	40	0	100
All	**44**	**56**	**10**	**90**

Success Measures

- What criteria should be used to categorize a play as successful?

$$gain < \frac{distance\ remaining}{(downs\ remaining - 1)}$$

- Defense attribute: ‘Maximum allowable gain’
- Offense attribute: ‘Minimum required gain’
- Success for Defense, Failure for offense
  - Turnover
  - Punt
  - Turnover on downs
  - Safety

VirtualMiner Software I

- Defines a measure called “interestingness” to find nonobvious patterns in data.
- Data consists of a number of records, each composed of a set of attributes.
- An event is defined by giving a ‘result’ attribute and a set of ‘input’ attributes plus sets of values for the attributes.
- Example event:
  - result : successful play.
  - input attributes and values: 3rd down; ball on left hash mark, less than 5 yards for first down, running play

- For any event *E*, *f(E)* is the fraction of records whose attributes match *E*’s specification.

VirtualMiner Software II

- Given an event $E = (x_1, x_2, x_3, \ldots, x_n)$, with result x_1, the interestingness $I_1(E)$ of E is the difference between $f(E)$ and $f((x_1))*f((x_2, x_3, \ldots, x_n))$.
- Motivation: $I_1(E) = 0$ if (x_1) and $(x_2, x_3, \ldots, x_n)$ are statistically independent. $I_1(E) > 0$ means event occurs more often, $I_1(E) < 0$, less often.
- Goal is to find events which makes $|I_1(E)|$ large.
- An event E is maximal if adding or removing attributes lowers the interestingness.
- VirtualMiner software finds interesting events.

VirtualMiner Results Defense

- Unsuccessful plays
  - 1st down, 11 personnel group, maximum allowable gain of 8 yards
    - Interestingness= 0.0023
  - Right hash mark, 2nd down
    - Interestingness= 0.0028
  - 1st down, 21 personnel group, maximum allowable gain of 8 yards
    - Interestingness= -0.0022

VirtualMiner Results Offense

- Rushing plays have higher success rating on short yardage.
- Instances of successful plays, based on formation and defensive alignment discovered.
- Passing plays have lower success rating at the end of the half or the game.
- Plays on which Steelman, the QB, carries the ball against an “odd” defensive front have a lower success rating.
- Army had a high success rating against Fordham (2011) but a lower success rating against Notre Dame (2010).

Future Work

- More comprehensive data – more games, more opponent data
- Incorporate additional data from coaching staff in analysis.
- Incorporate results of analysis in coaches preparation.
- Use *VirtualMiner* as an interactive exploratory tool.

3. Restructuring the NHL

- Dr. Brian MacDonald
- WRP

The National Hockey League

- Thirty teams in Canada and United States
- League is divided into two conferences; each conference is divided into three divisions, each consisting of five teams.

- Each season, each team plays each team in same division 8 times, each team in same conference but different division four times; each other team 2/3 times (two divisions once, one zero.)

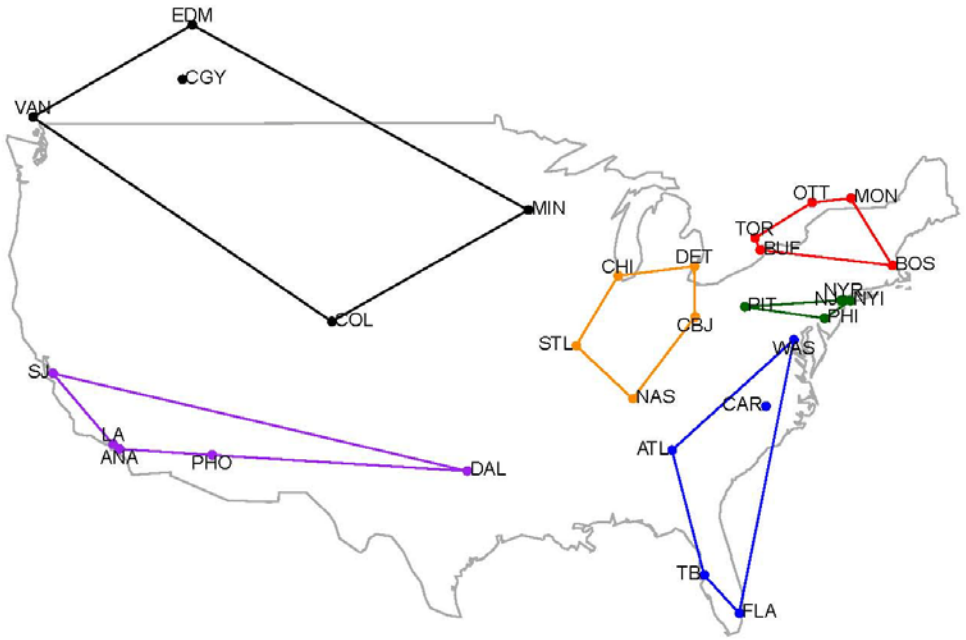
- The goal is to minimize travel and respect “traditional rivalries”

NHL prior to 2011-12

League Prior Divisions

Previous Configuration

Cost relative to 'best' alignment: 16300 miles, $179000 to 293000.
Savings over the current 6 division aligment: 37200 miles, $0.41−0.67mil
Savings over proposed 4 conference alignment: 86100 miles, $0.95−1.55mil



Teams with worst travel: VAN: 62900. SJ: 60300
Cost for West Coast and Florida teams relative to 'best' alignment: 14200 miles, $156200−255600

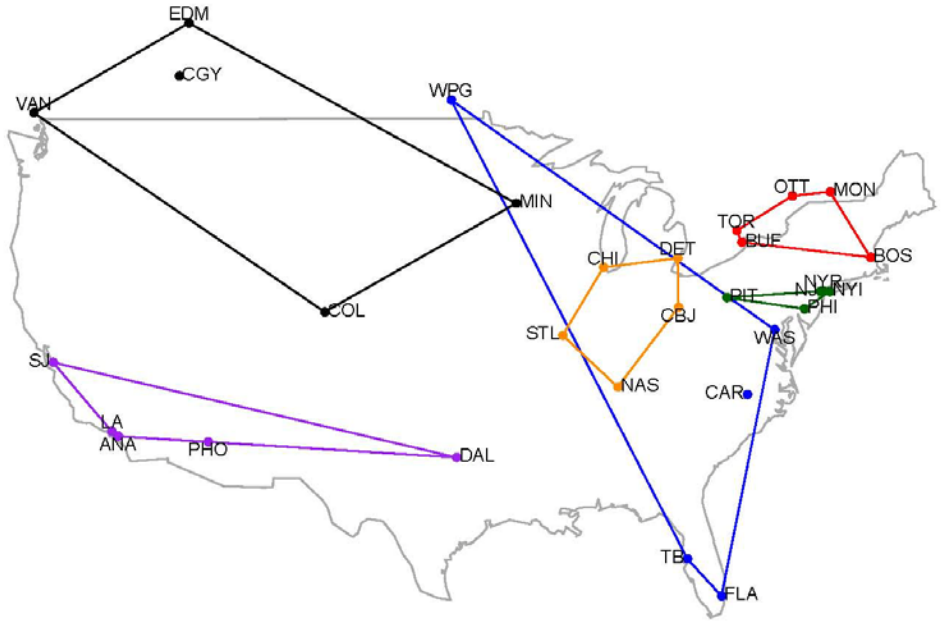
© 2012 William Pulleyblank and Brian Macdonald

NHL at present

League Divisions

Current Configuration

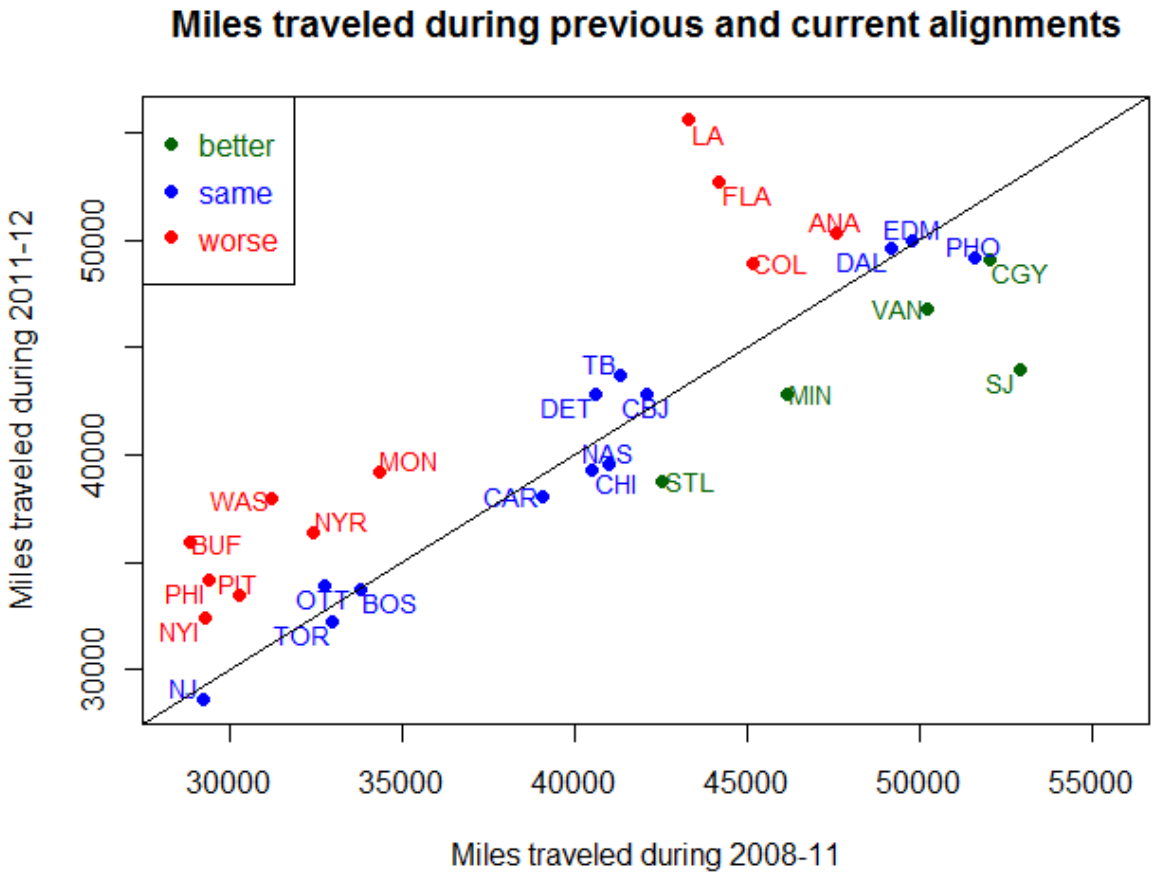
Cost relative to 'best' alignment: 51800 miles, $570000 to 932000.
Savings over the current 6 division aligment: 1700 miles, $0.02−0.03mil
Savings over proposed 4 conference alignment: 50600 miles, $0.56−0.91mil



Teams with worst travel: VAN: 62200. SJ: 60000
Cost for West Coast and Florida teams relative to 'best' alignment: 10300 miles, $113300−185400

© 2012 William Pulleyblank and Brian Macdonald

Effect of change on distance traveled



How should the NHL restructure?

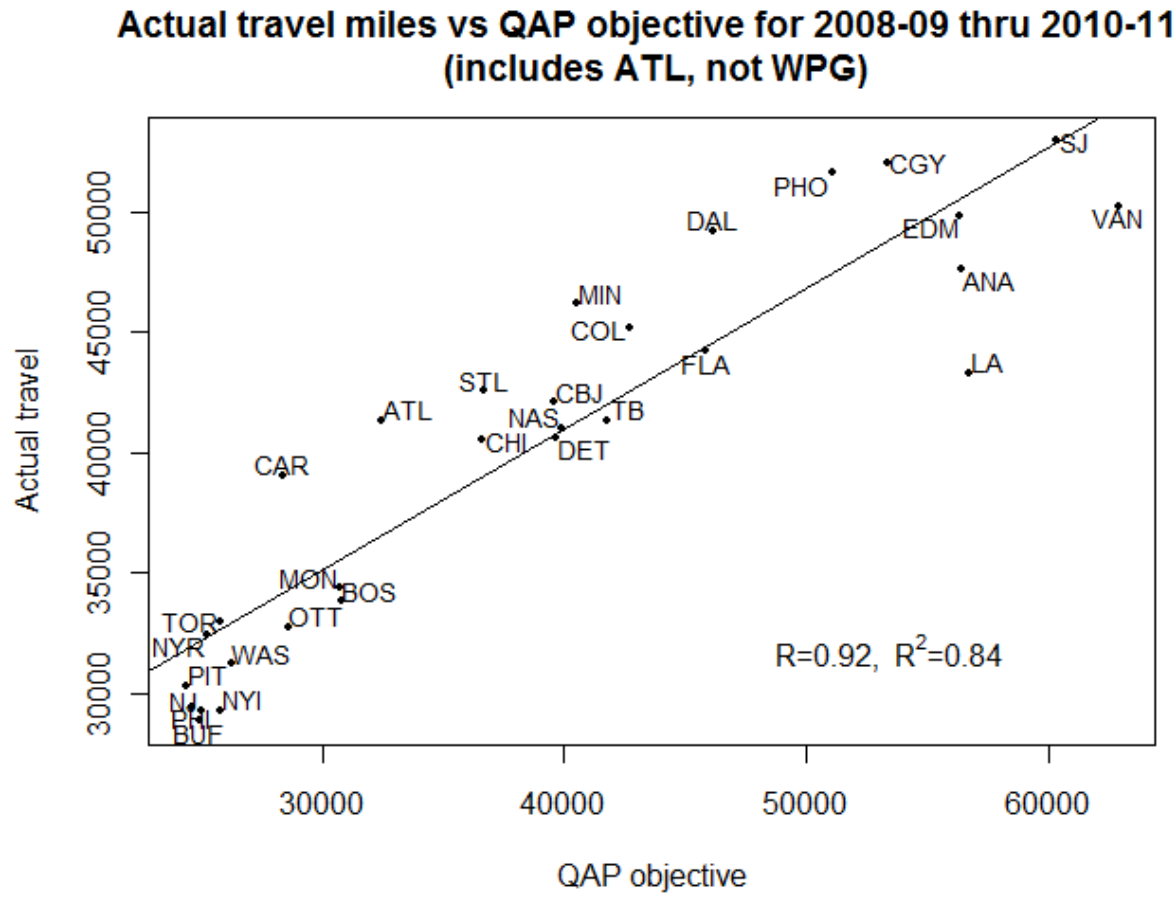
- Teams play the same number of home and away games against other teams – except for the teams in the other conference. For these, they play one division at home and one on the road.
- Problem: Annual schedule is created each season, based on league structure. A road trip will normally visit three to five cities, usually arranged chosen for efficient travel.

A surrogate objective function

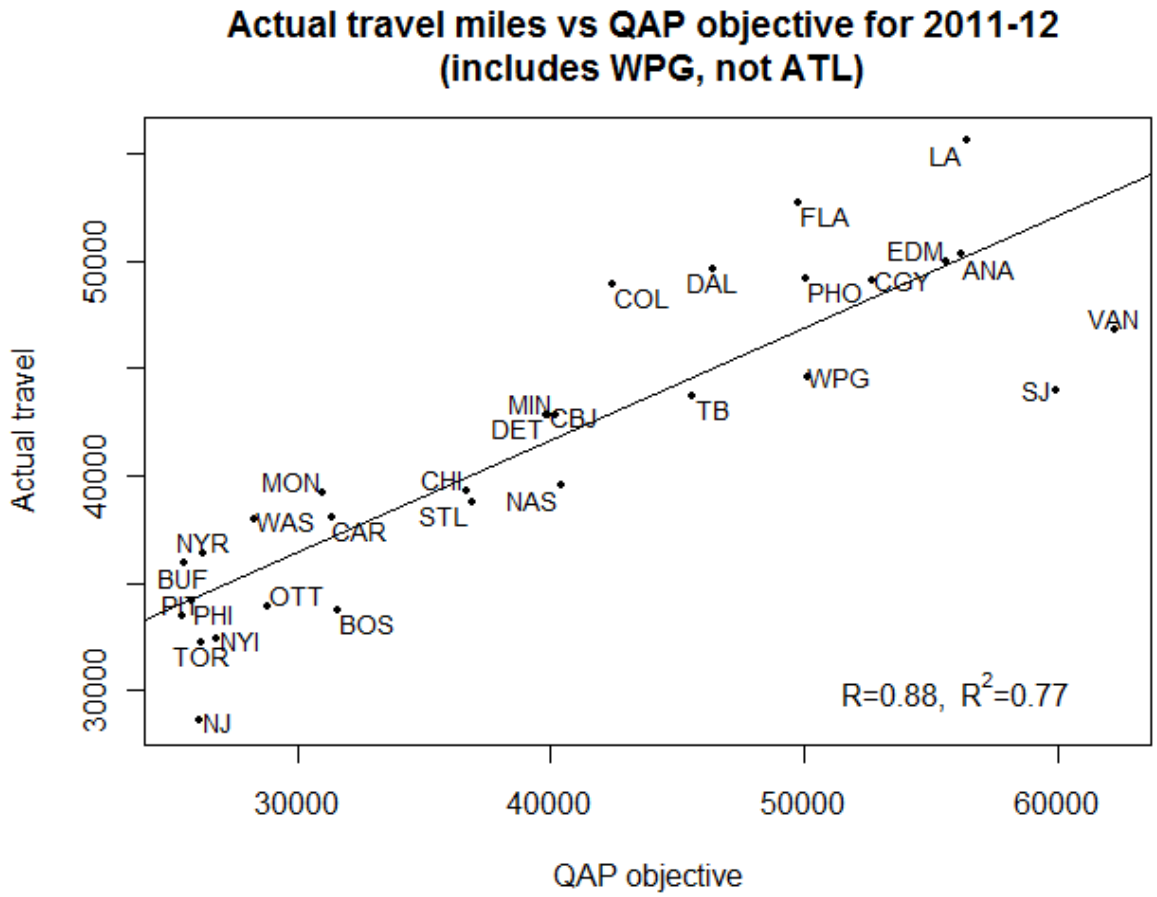
- For each pair (*i*,*j*) of cities, let *d*(*i*, *j*) be the distance from *i* to *j*. Let *s*(*i*, *j*) be the number of games that *i* plays in city *j.*
- The cost of league *L* is defined to be

$$c(L) = \text{SUM}\,(d(i,j)*s(i,j)\text{: all } (i,j)).$$

Historical quality of *c(L)*



Quality of c(L) post 2011-12



Minimizing *c(L)* is a Quadratic Assignment Problem

- Let D be the 30 by 30 intercity distance matrix.
- Let S be the “number of away games” matrix

4	2	2	1		
2	4	2		1	
2	2	4			1
	1		4	2	2
		1	2	4	2
1			2	2	4

Away Game Matrix

All entries 0 on Main diagonal

$S =$

4	2	2	1		
2	4	2		1	
2	2	4			1
	1		4	2	2
		1	2	4	2
1			2	2	4

QAP

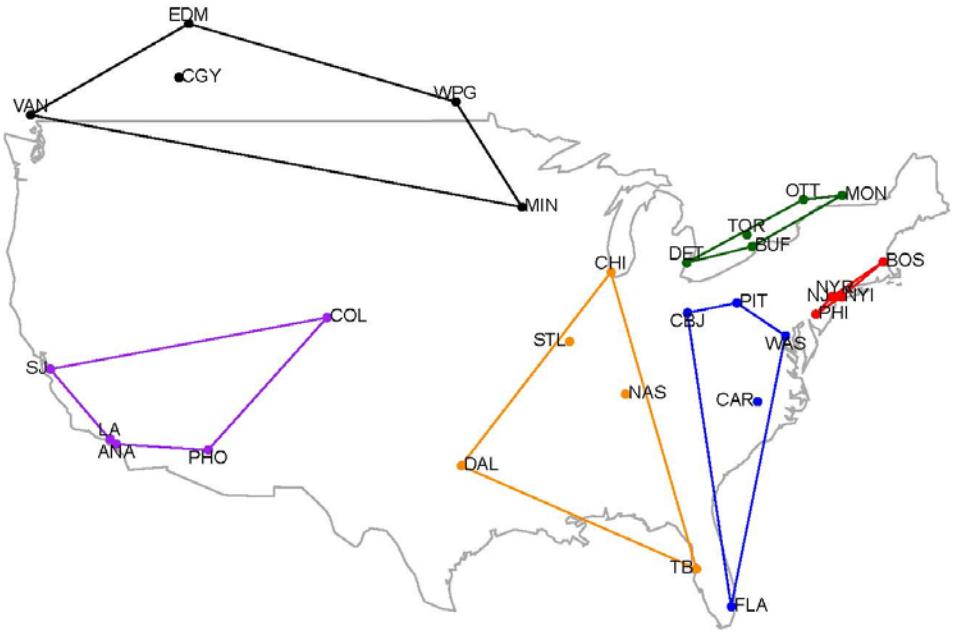
- Find a permutation π of 1,…, 30 which minimizes

  SUM ($d(i,j)$*$s(\pi(i), \pi(j))$: all (i,j)).

Best Alignment

Config 1

Cost relative to 'best' alignment: 0 miles, $ 0− 0.
Savings over the current 6 division aligment: 53600 miles, $0.59−0.96mil.
Savings over proposed 4 conference alignment: 102500 miles, $1.13−1.84mil



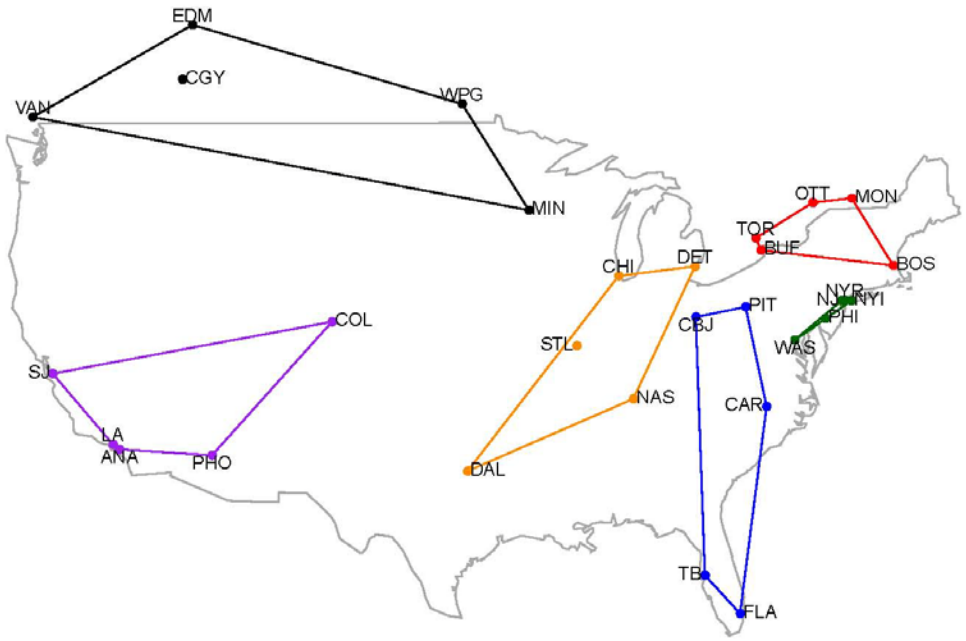
Teams with worst travel: VAN: 61926. TB: 59247
Teams with worst travel relative to their own personal min: TB: 17707. NAS: 12036
Cost for West Coast and Florida teams relative to 'best' alignment: 4000

© 2012 William Pulleyblank and Brian Macdonald

Best keeping TB and FLA together

Config 4

Cost relative to 'best' alignment: 1211 miles, $13000−22000.
Savings over the current 6 division aligment: 52300 miles, $0.58−0.94mil.
Savings over proposed 4 conference alignment: 101200 miles, $1.11−1.82mil



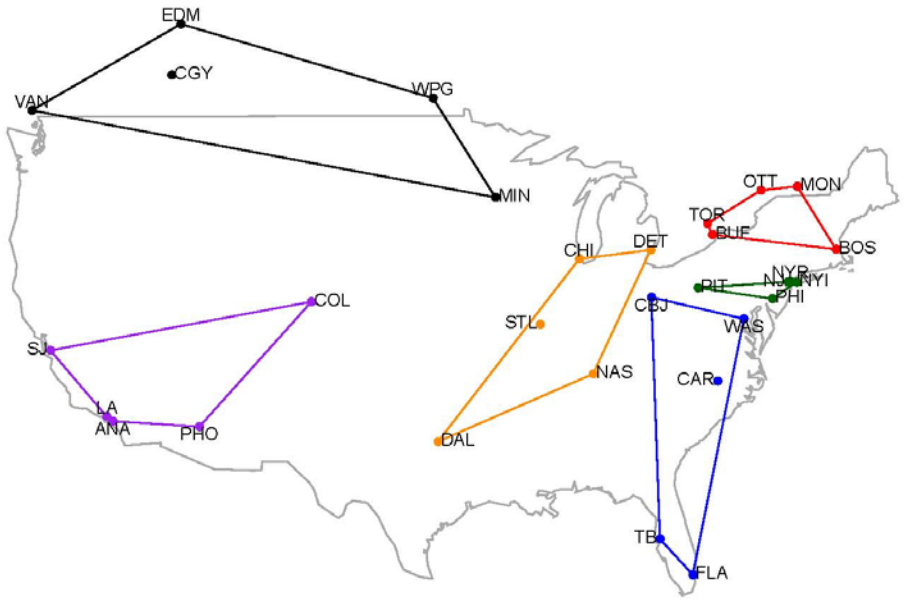
Teams with worst travel: VAN: 61038. SJ: 58633
Teams with worst travel relative to their own personal min: DET: 18415. NAS: 11674
Cost for West Coast and Florida teams relative to 'best' alignment: 0

© 2012 William Pulleyblank and Brian Macdonald

Best satisfying “traditional” rivalries

Configuration 11

Cost relative to 'best' alignment: 2005 miles, $22000−36000.
Savings over the current 6 division aligment: 51600 miles, $0.57−0.93mil.
Savings over proposed 4 conference alignment: 100500 miles, $1.1−1.81mil



Teams with worst travel: VAN: 61000. SJ: 58600
Cost for West Coast and Florida teams relative to 'best' alignment: 0 miles, $ 0− 0

© 2012 William Pulleyblank and Brian Macdonald

New Structure Proposed by NHL

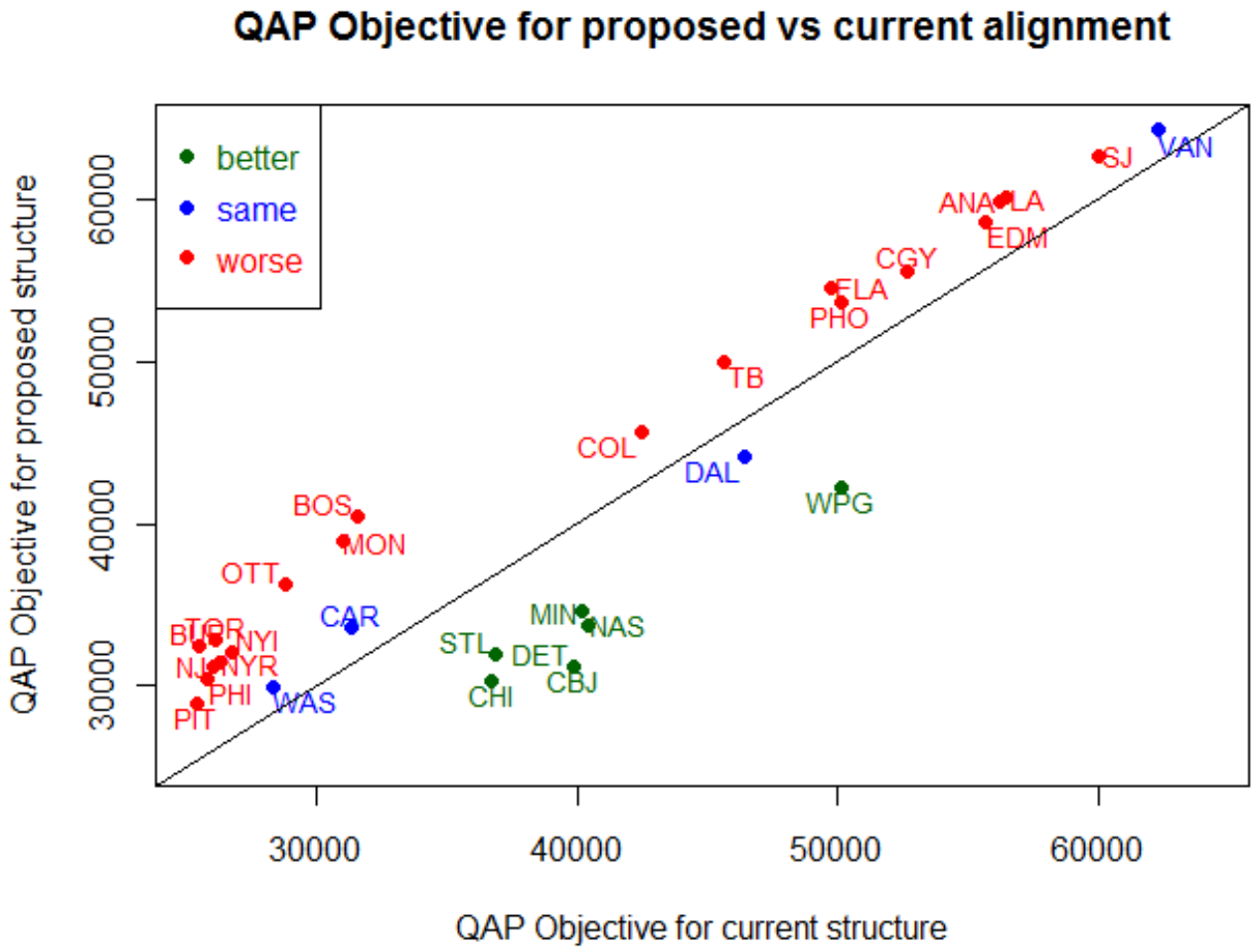
- Four Divisions, two with 7 teams, two with 8 teams.
- Each team plays each team in the other conferences twice, once at home and once away. This accounts for 46 games for a team in a seven team conference and 44 games for each team in an eight team conference.
- Each team in a seven team conference would play each of the six other teams in the conference six times, three at home and three away.
- Each team in an eight team conference would play four of the other seven teams five times and the other three teams six times.

New Away Game Matrix

	7 teams	7 teams	8 teams	8 teams
7 teams	3	1	1	1
7 teams	1	3	1	1
8 teams	1	1	3(3) 2.5(4)	1
8 teams	1	1	1	3(3) 2.5(4)

All entries 0 on Main diagonal

Proposed vs. Current



QAP and TSP?

- QAP is in $\mathcal{NP}$.
- TSP is $\mathcal{NP}$-complete.
- Bill Cook routinely solves TSPs with hundreds of thousands of cities.
- Is there a “reasonable” transformation of an instance of our QAP to an instance of TSP which Bill can solve?

Thanks

And congratulations, Bill!