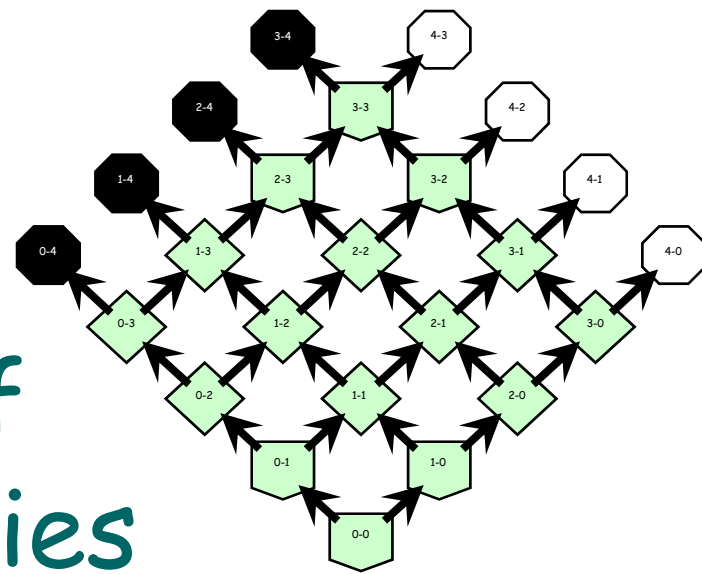


Forming a Markov Probability Model of 7-Game Playoff Series



Christopher M. Rump

Applied Statistics & Operations Research

College of Business Administration

Bowling Green State University

cmrump.with.bgsu.edu

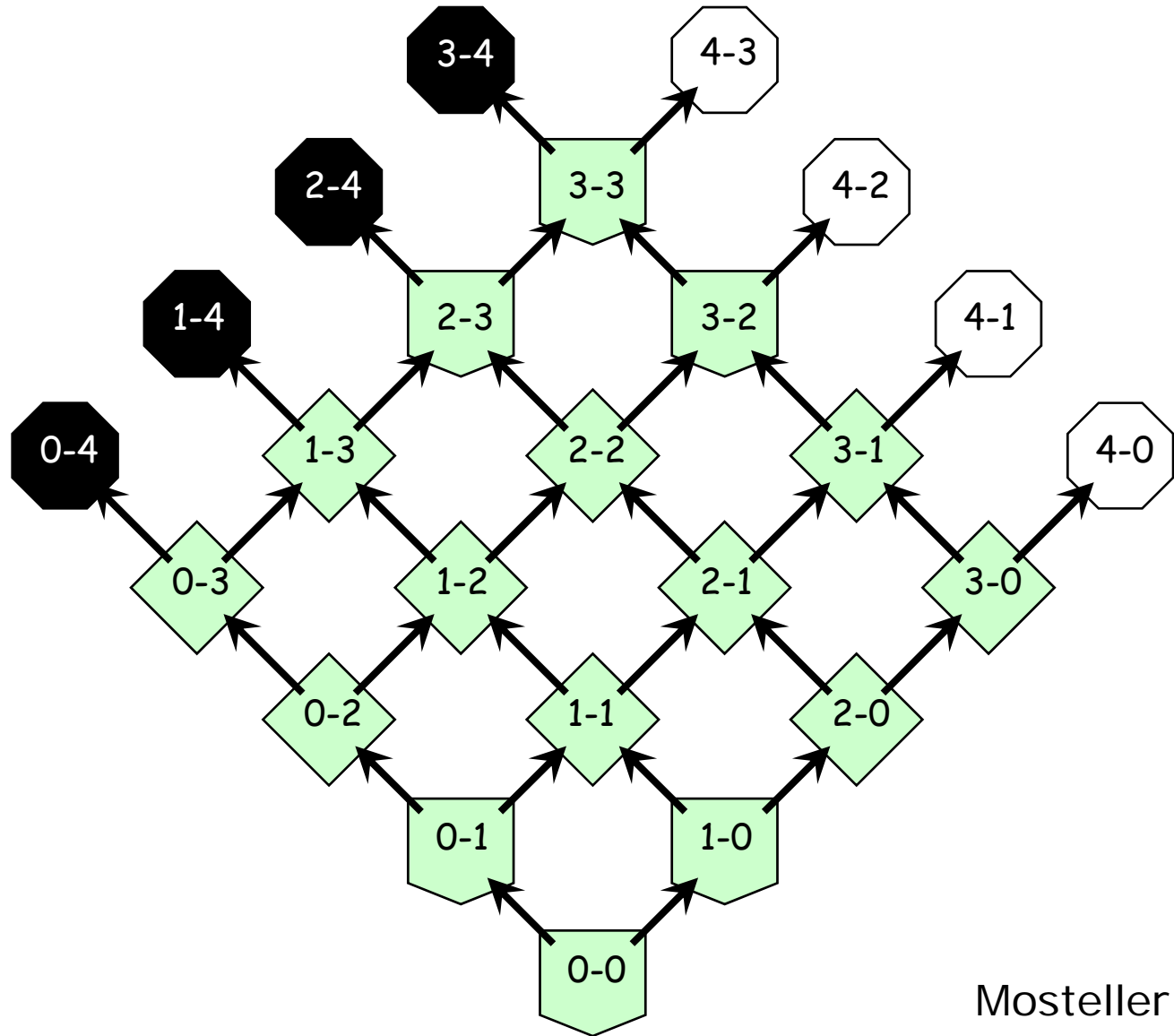


MLB Playoff Data

- Best-of-7 Game Playoffs
 - World Series since 1924 (83)
 - League Championship Series since 1985 (44)
- *Favored* team (assigned home-field advantage) won
 - 51.4% of all 736 games
 - 56.5% of 377 home games
 - 46.0% of 359 away games
- &
 - 54.3% of all 127 *series*

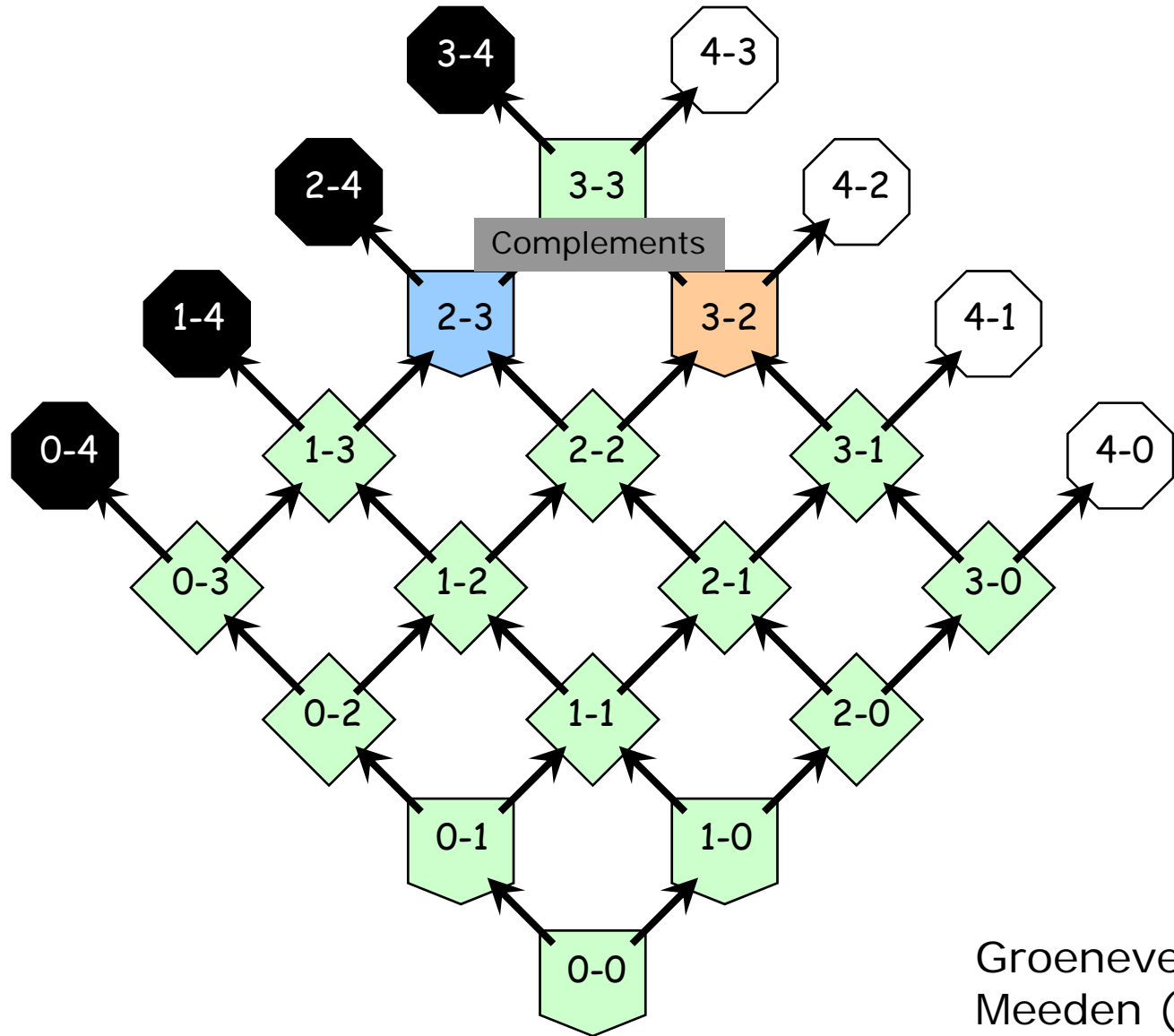
Status	Games,n	Wins,w	Pct.
3-0	14	11	.786
2-3 <small>HOME</small>	42	27	.643
0-0 <small>HOME</small>	127	75	.591
0-1 <small>HOME</small>	52	30	.577
3-3 <small>HOME</small>	46	26	.565
1-3	26	14	.538
1-0 <small>HOME</small>	75	39	.520
2-1	54	28	.519
3-1	31	16	.516
1-2	46	22	.478
3-2 <small>HOME</small>	35	16	.457
1-1	66	29	.439
2-2	48	20	.417
0-2	22	9	.409
2-0	39	14	.359
0-3	13	2	.154
Total	736	378	.514

Bernoulli "Coin Flip" Model ($p=0.348$)



Probability
[0.70,1.00]
[0.60,0.70)
[0.55,0.60)
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

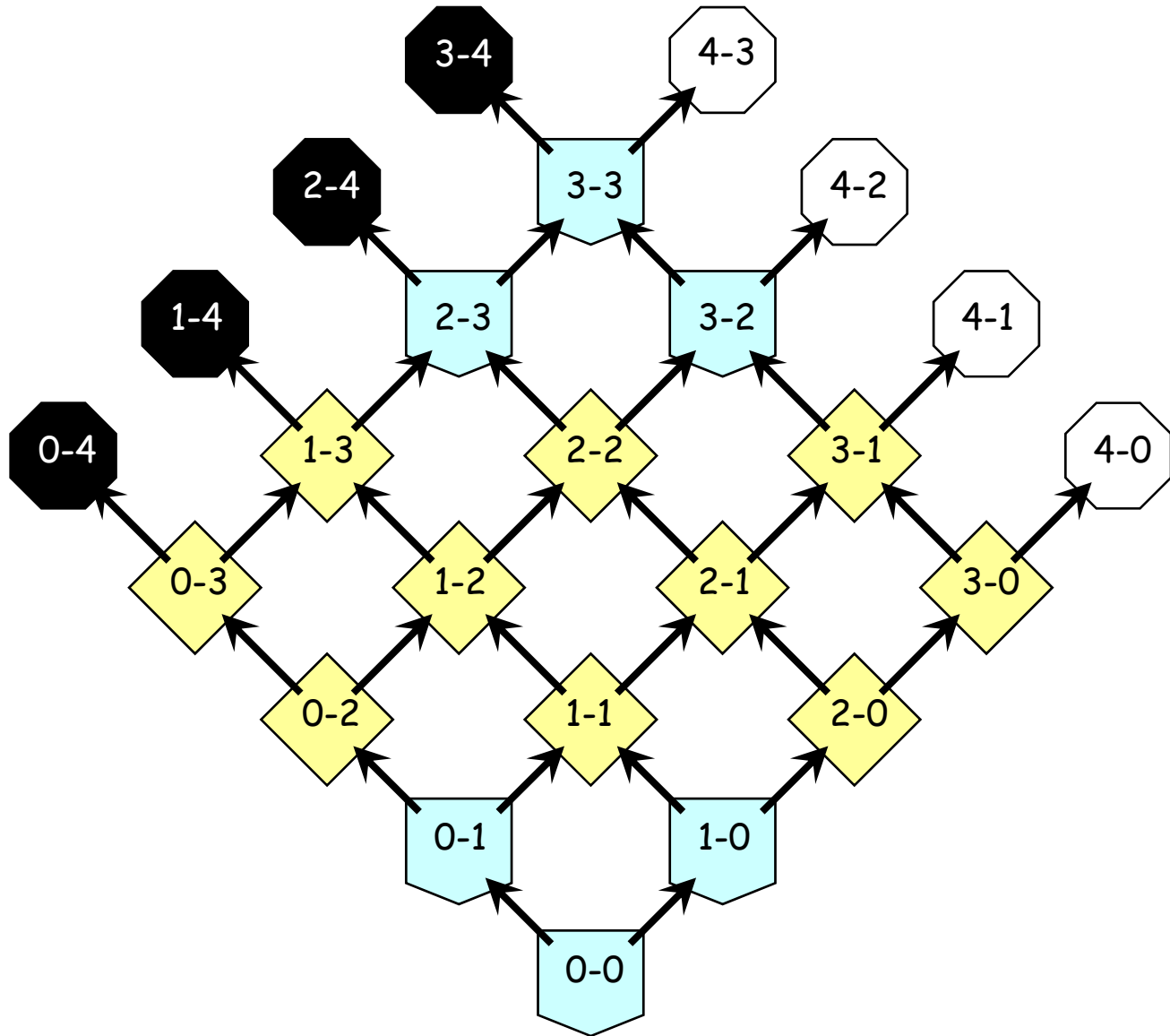
"Back-to-the-Wall" Model (p=0.580)



Probability
[0.70,1.00]
[0.60,0.70)
[0.55,0.60)
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

Groeneveld & Meeden (1975)

Home-Away Model (p=0.199)



Probability
[0.70,1.00]
[0.60,0.70)
[0.55,0.60)
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

"p-value" of fit is probability
- assuming model's correct -
of seeing actual outcomes
deviate that much or more

Modeling Goals

- Want a parsimonious probability model that best "fits" 8 series-ending frequencies
- Absorbing Markov chain with actual data
 - provides a perfect fit ($p=1$) to historical data
 - "overfit" as a predictor of future play since uses $16 > 8-1$ parameter estimates
- Partition $M = 16$ states (ordered by game-winning frequency) into $K \leq 6$ clusters
 - Possible partitions: $\binom{M-1}{K-1}$

Set Partitioning Problem

- For cluster of consecutive states i to j
 - p_{ij} as aggregated game-winning frequency
 - c_{ij} as dissimilarity within cluster
 - x_{ij} as indicator whether or not to use cluster
- Find K clusters of minimal dissimilarity:

$$\min_{\mathbf{x}} \sum_{i=1}^M \sum_{j=i}^M c_{ij} x_{ij}$$

$$\text{subject to } \sum_{i=1}^M \sum_{j=i}^M x_{ij} = K$$

$$\sum_{i=1}^s \sum_{j=s}^M x_{ij} = 1$$

$$x_{ij} \in \{0, 1\}$$

Select K of $M(M+1)/2$ clusters

Assign each state s to exactly 1 cluster

$$s = 1, \dots, M$$

$$i = 1, \dots, M, j = i, \dots, M.$$

Cluster Dissimilarity Criteria

$$p_{ij} = \frac{\sum_{s=i}^j n_s p_s}{\sum_{s=i}^j n_s} = \frac{\sum_{s=i}^j w_s}{\sum_{s=i}^j n_s} = \frac{w_{ij}}{n_{ij}}$$

○ Sum Squares (SS)

$$c_{ij} = \sum_{s=i}^j n_s (p_s - p_{ij})^2$$

○ Sum Absolutes (SA)

$$c_{ij} = \sum_{s=i}^j n_s |p_s - p_{ij}|$$

○ Clique (CL)

$$c_{ij} = \sum_{r=i}^j \sum_{s=i}^j n_s |p_s - p_r|$$

○ Star (ST)

$$c_{ij} = \min_{r=i}^j \sum_{s=i}^j n_s |p_s - p_r|$$

○ Radius (R)

$$c_{ij} = \min_{r=i}^j \max_{s=i}^j n_s |p_s - p_r|$$

○ Diameter (D)

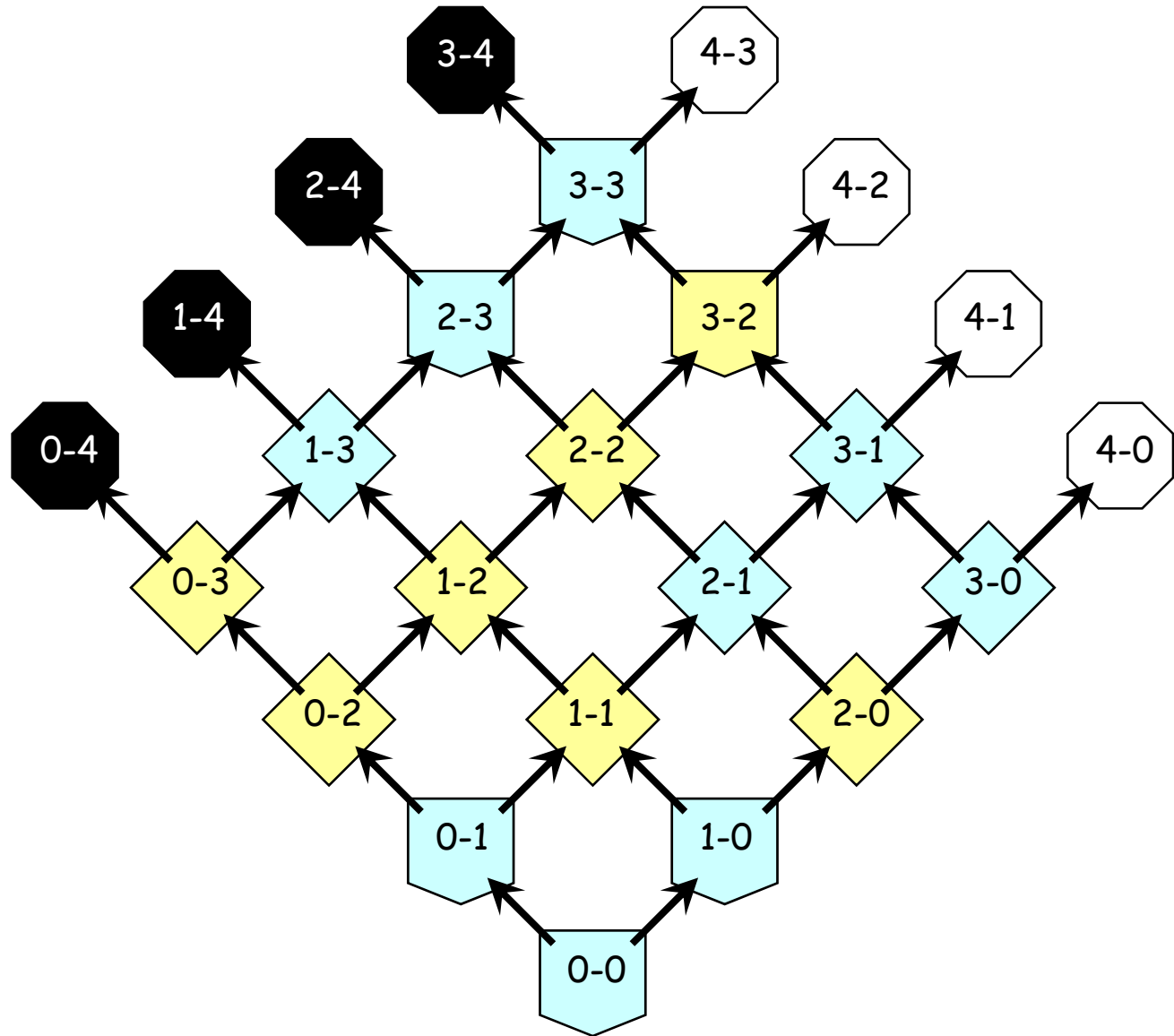
$$c_{ij} = \max_{r=i}^j \max_{s=i}^j n_s |p_s - p_r|$$

Minisum & Minimax Relative Error

K	Partition	Clustering Metric						Sum	Max
		SS	SA	Clique	Star	Radius	Diam		
2	9/7	0	0	0	0	0	0.484	0.484	0.484
3	5/4/7	0.131	0.005	0.167	0.022	0.088	0.816	1.228	0.816
3	2/6/8	0.238	0.333	0.351	0.426	0.452	0.372	2.172	0.452
4	5/4/6/1	0.066	0.024	0.256	0.033	0.188	1.119	1.685	1.119
4	1/5/7/3	0.416	0.377	0.398	0.532	0.570	0.499	2.792	0.570
5	1/4/5/5/1	0.000	0.000	0.434	0.050	0.425	0.249	1.158	0.434
6	1/4/4/5/1/1	0.022	0.000	0.728	0.023	0.135	0.235	1.142	0.728
6	1/4/4/3/3/1	0.004	0.063	0.365	0.056	0.475	0.472	1.436	0.475

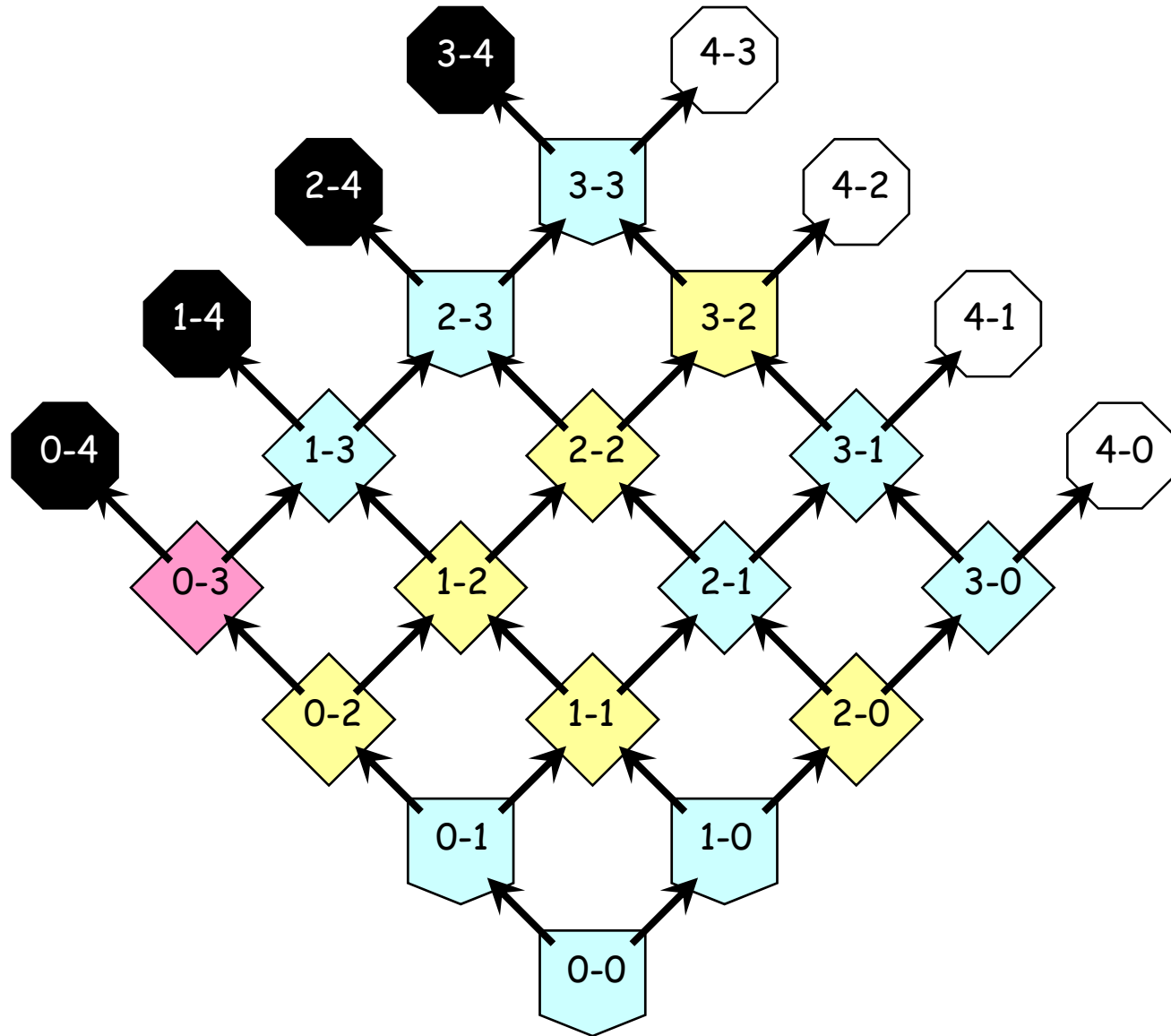
			Minisum Partitions					
State	Status	Actual	K=1	K=2	K=3	K=4	K=5	K=6
1	3-0	.786	.514	.570	.601	.601	.786	.786
2 ^{HOME}	2-3	.643	.514	.570	.601	.601	.592	.592
3 ^{HOME}	0-0	.591	.514	.570	.601	.601	.592	.592
4 ^{HOME}	0-1	.577	.514	.570	.601	.601	.592	.592
5 ^{HOME}	3-3	.565	.514	.570	.601	.601	.592	.592
6	1-3	.538	.514	.570	.522	.522	.513	.522
7 ^{HOME}	1-0	.520	.514	.570	.522	.522	.513	.522
8	2-1	.519	.514	.570	.522	.522	.513	.522
9	3-1	.516	.514	.570	.522	.522	.513	.522
10	1-2	.478	.514	.416	.416	.430	.513	.442
1 ^{HOME}	3-2	.457	.514	.416	.416	.430	.419	.442
12	1-1	.439	.514	.416	.416	.430	.419	.442
13	2-2	.417	.514	.416	.416	.430	.419	.442
14	0-2	.409	.514	.416	.416	.430	.419	.442
15	2-0	.359	.514	.416	.416	.430	.419	.359
16	0-3	.154	.514	.416	.416	.154	.154	.154

2-Cluster "Game Favorite" (p=0.582)



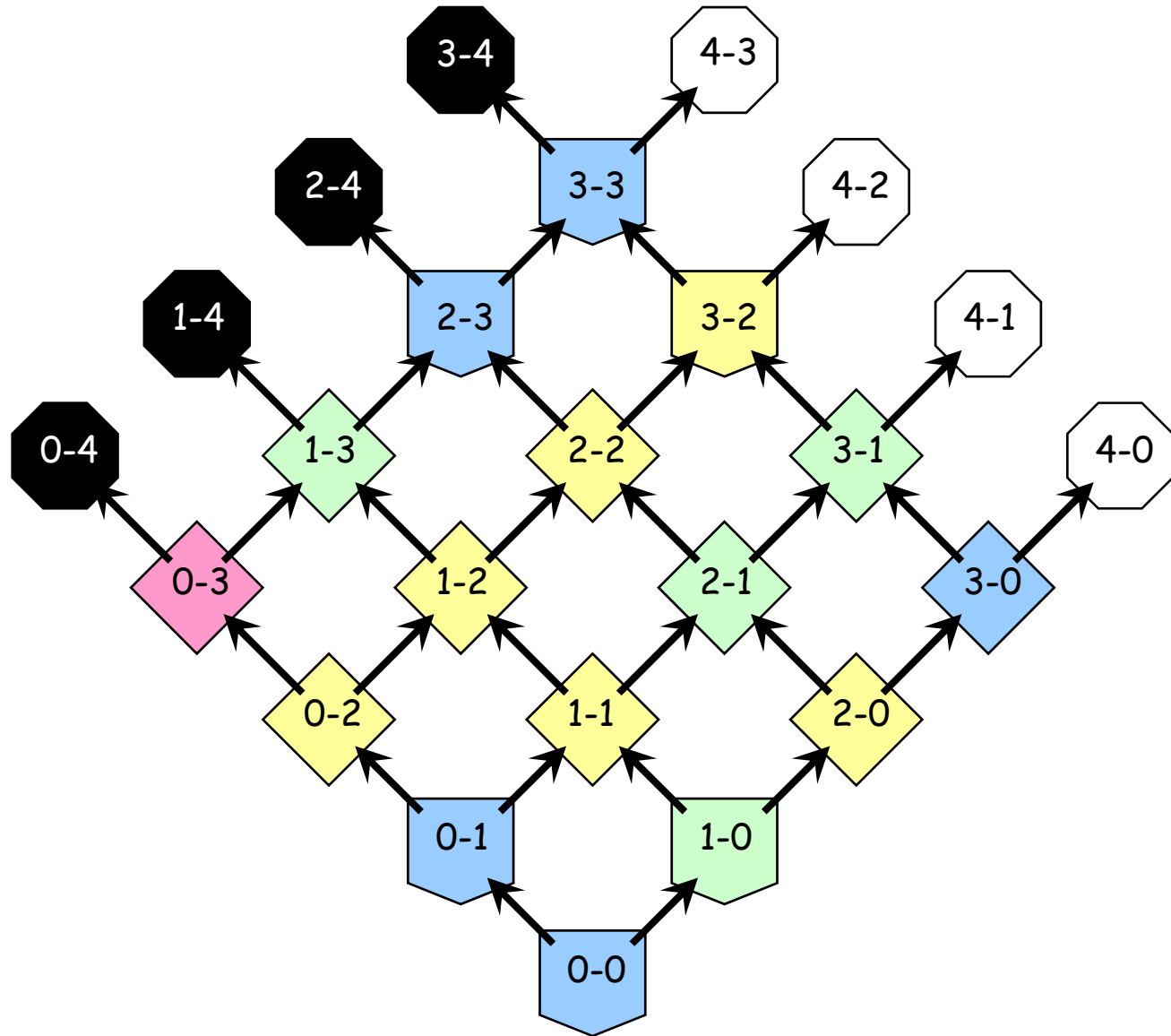
Probability
[0.70,1.00]
[0.60,0.70]
[0.55,0.60]
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

3-Cluster Minisum Solution ($p=0.724$)



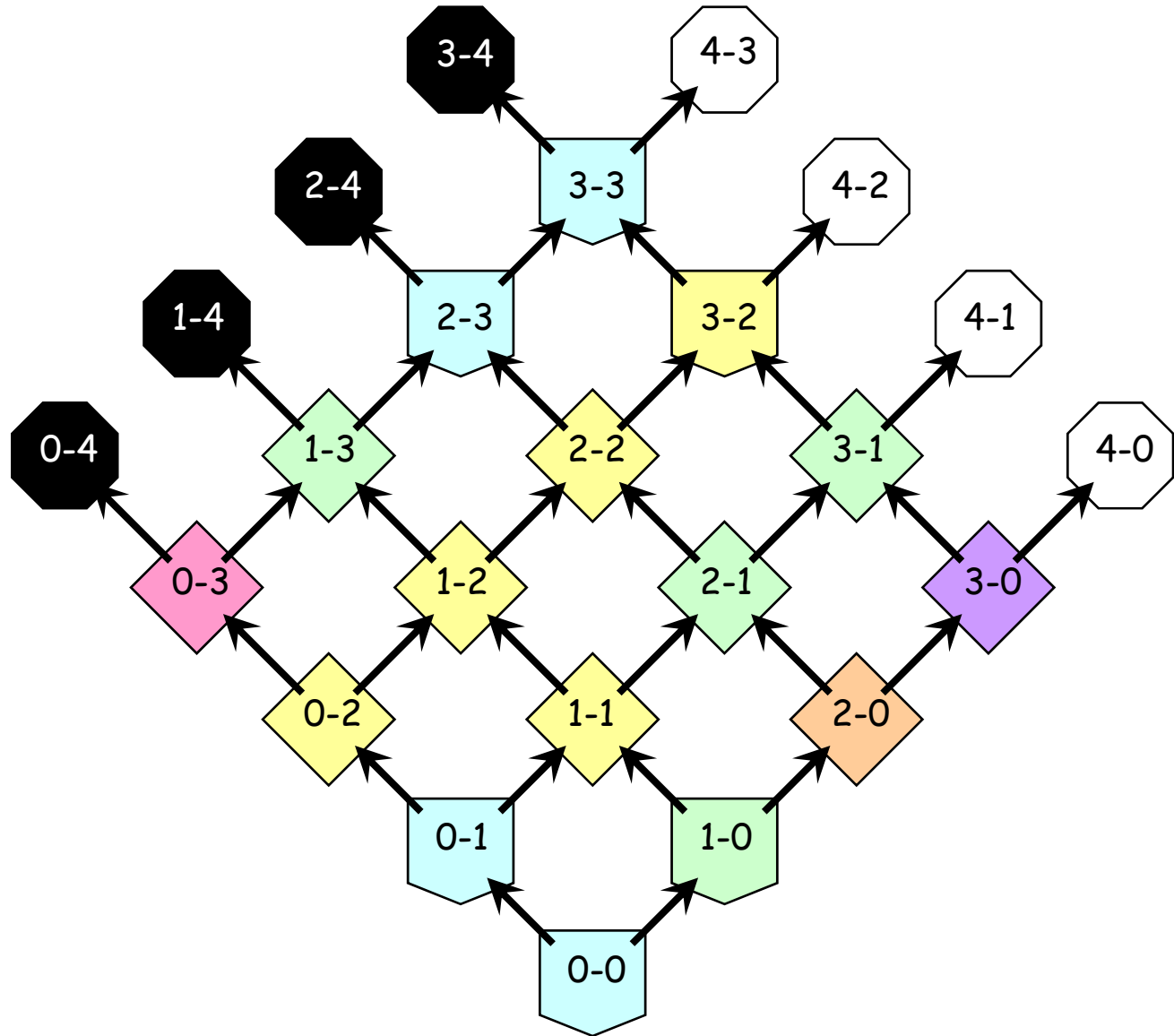
Probability
[0.70,1.00]
[0.60,0.70]
[0.55,0.60]
[0.50,0.55]
[0.40,0.50]
[0.30,0.40]
[0.00,0.30]

4-Cluster Minisum Solution ($p=0.806$)



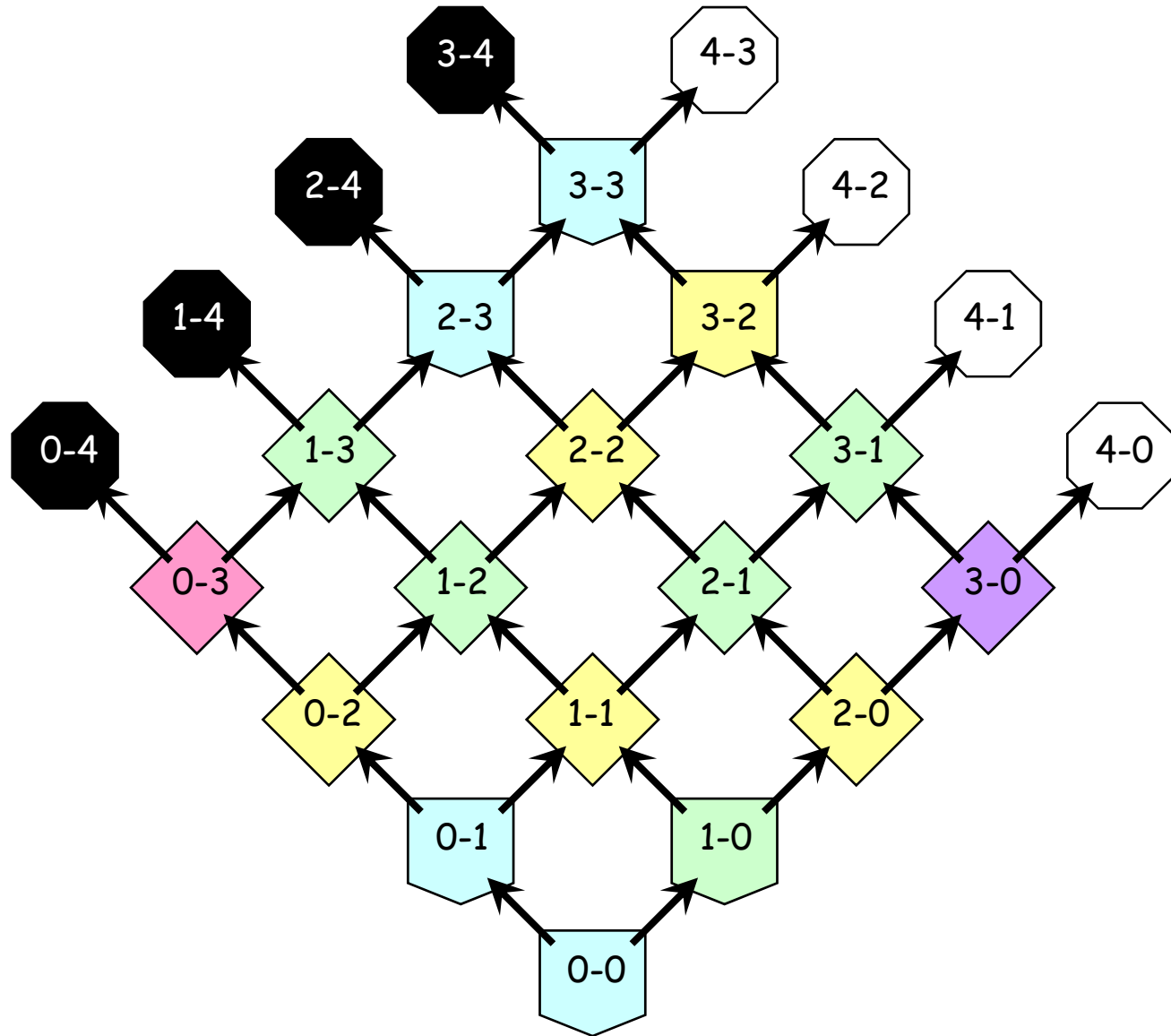
Probability
[0.70,1.00]
[0.60,0.70]
[0.55,0.60]
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

6-Cluster Minisum Solution ($p=0.429$)



Probability
[0.70,1.00]
[0.60,0.70]
[0.55,0.60]
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

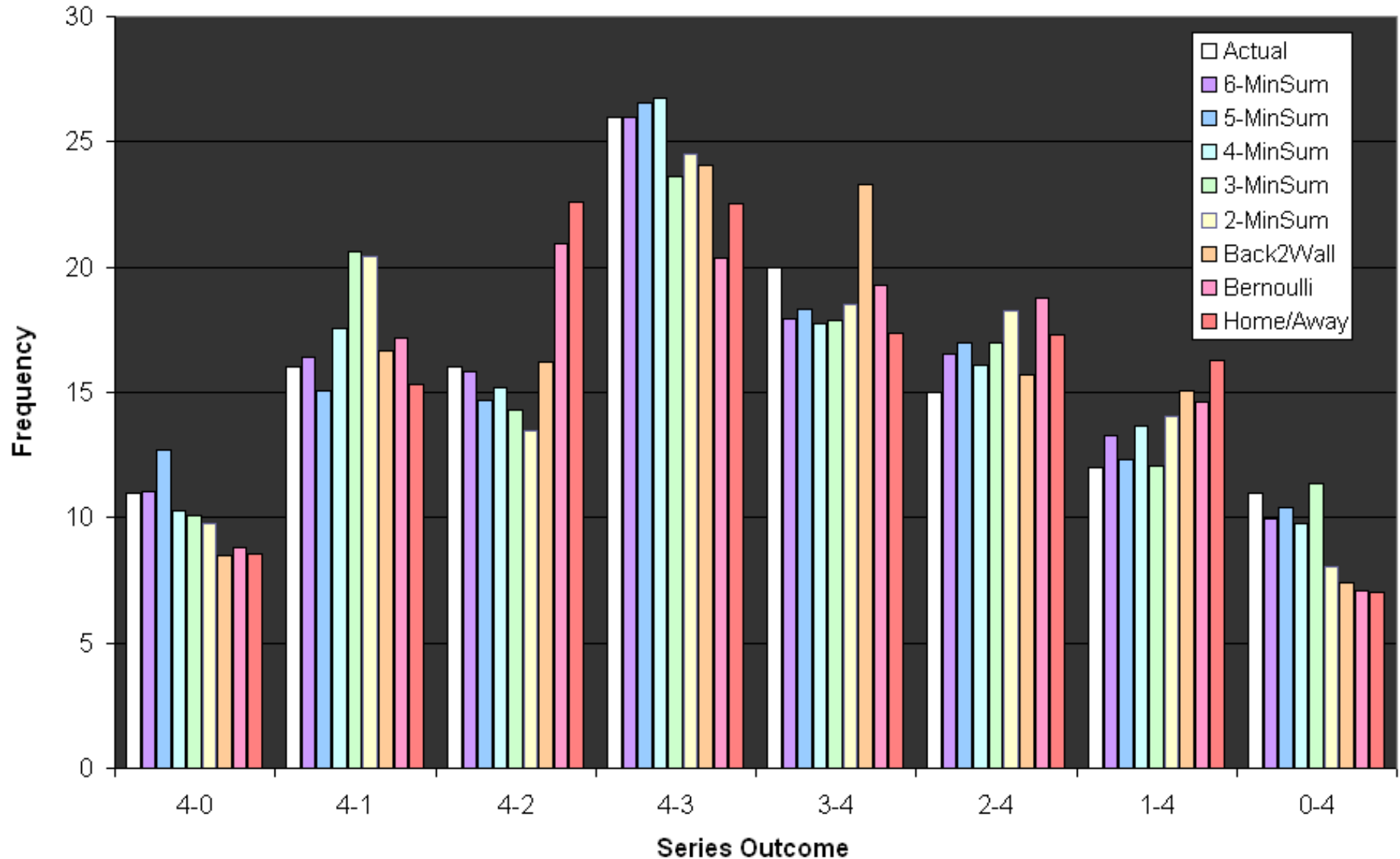
5-Cluster Minisum Solution ($p=0.657$)



Probability
[0.70,1.00]
[0.60,0.70]
[0.55,0.60]
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

Minisum Model Performance

Predicted Outcomes of MLB 7-Game Playoff Series





Conclusions

- Home-Away model is good for NBA not MLB where favored team falters leading 3-2
- Markov clustering approach
 - All but 5-cluster solution are *nested* partitions
 - 2-cluster "game favorite" solution offers good tradeoff between simplicity & efficacy:
partitions games by whether or not favored team (with home-field advantage in series) is favorite for that game (more likely winner)
 - 4-cluster minisum solution is a refined version with best statistical fit ($p=0.806$)

References

- Mosteller (1952) The World Series Competition, *Journal of the American Statistical Association*, 47(259), 355-380.
- Groeneveld & Meeden (1975) Seven Game Series in Sports, *Mathematics Magazine*, 48(4), 187-192.
- Bassett & Hurley (1998) The Effects of Alternative HOME-AWAY Sequences in a Best-of-Seven Playoff Series, *The American Statistician*, 52(1), 51-53.
- Rump (2006) The Effects of Home-Away Sequencing on the Length of Best-of-Seven Game Playoff Series, *Journal of Quantitative Analysis in Sports*, 2(1), Article 5.
(www.bepress.com/jqas/vol2/iss1/5)
- Rump (2008) Data Clustering for Fitting Parameters of a Markov Chain Model of Multi-Game Playoff Series, *Journal of Quantitative Analysis in Sports*, 4(1), Article 2.
(www.bepress.com/jqas/vol4/iss1/2)

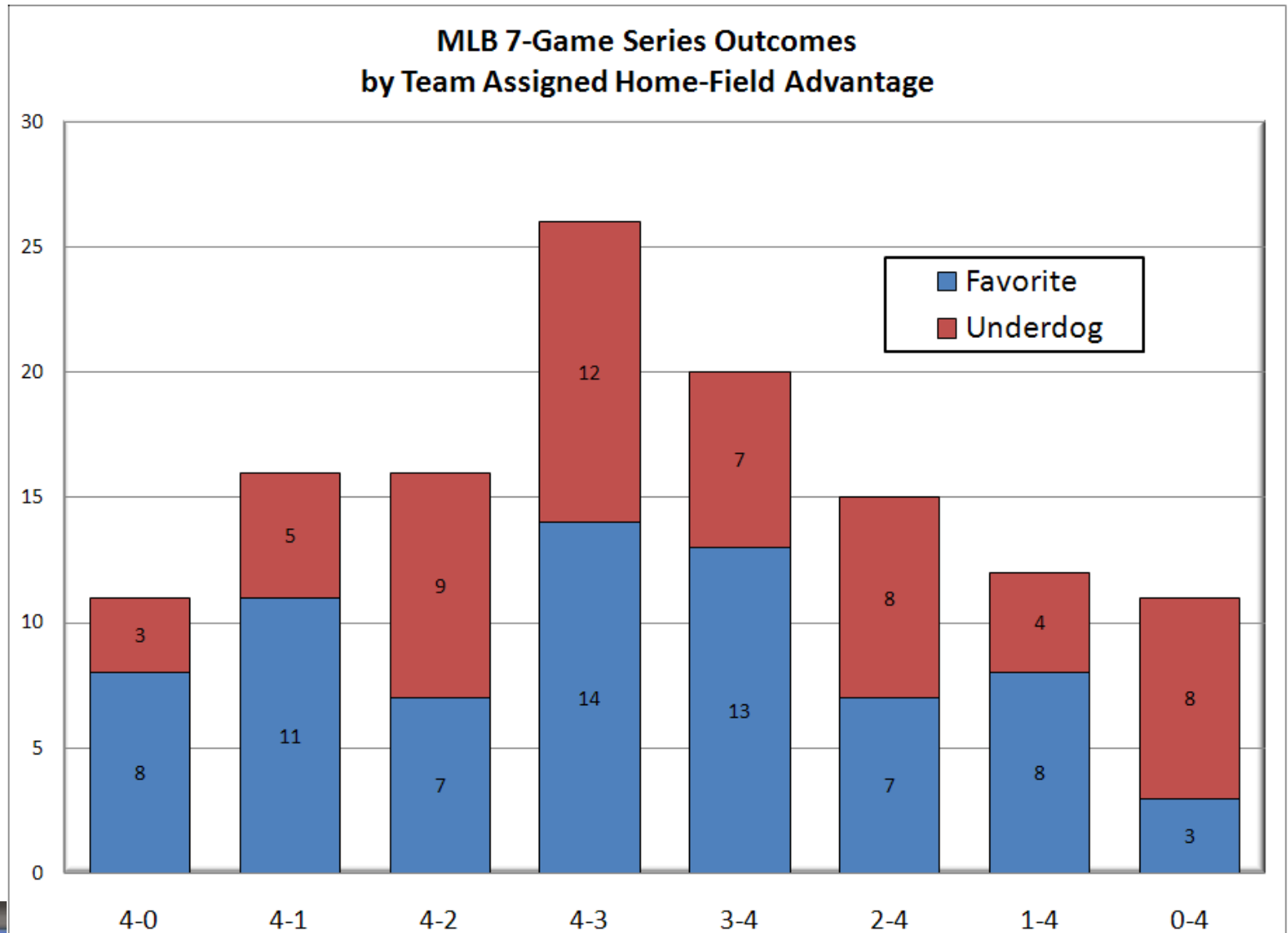
Questions or Comments?



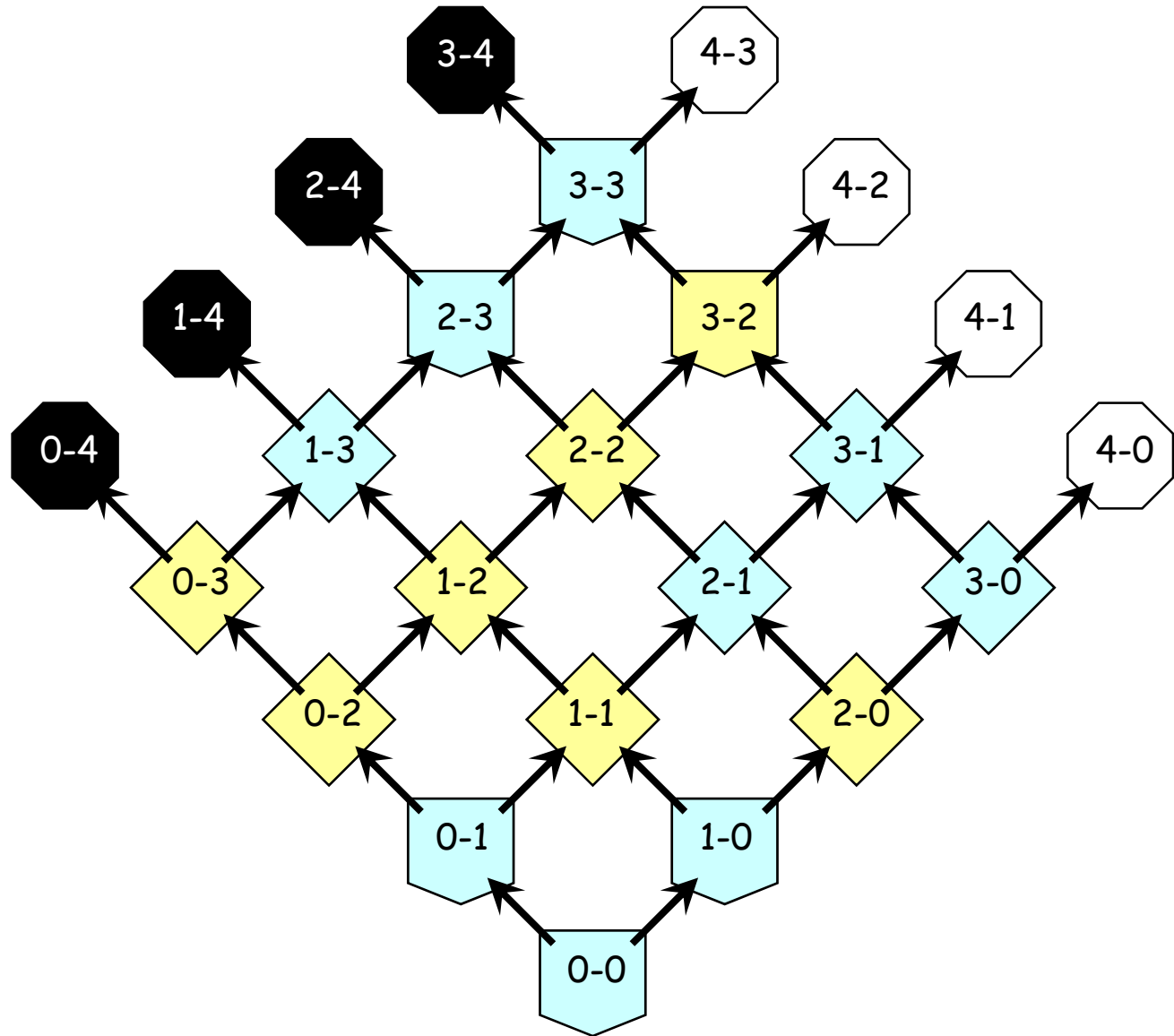
Extensions

- Segregate series data
 - by relative strength of favored team
 - Does favored team have
 - ≥ 0 more wins (a "favored favorite") - 71 series
 - or not (a "favored underdog") - 56 series
- Complementary probability clustering
 - as in Groeneveld & Meeden (1975)
 - For "non-favorite" states ($p_s < 0.5$), cluster according to complementary probability $1-p_s$

Favored Team Disparity

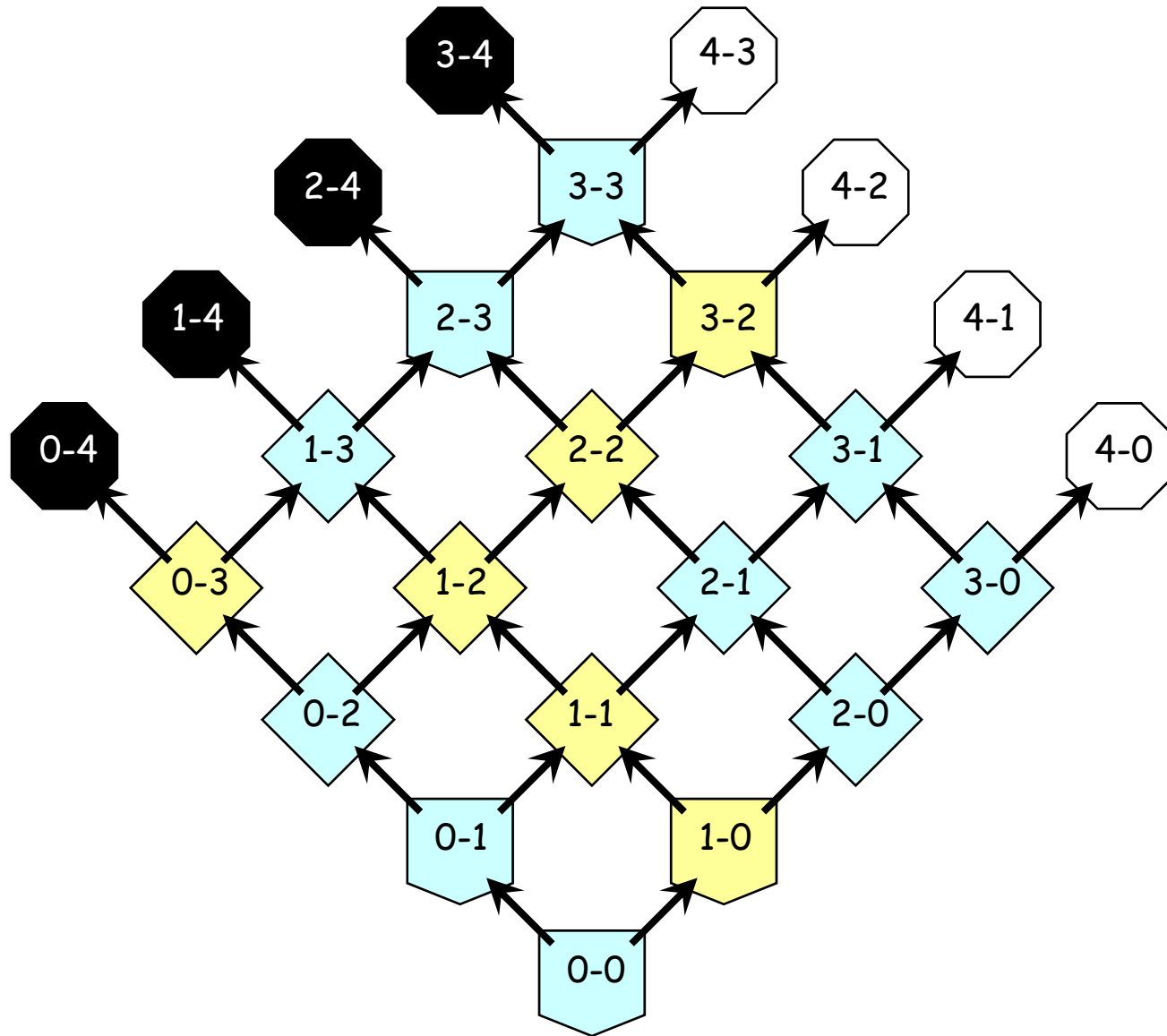


2-Cluster "Game Favorite" (p=0.582)



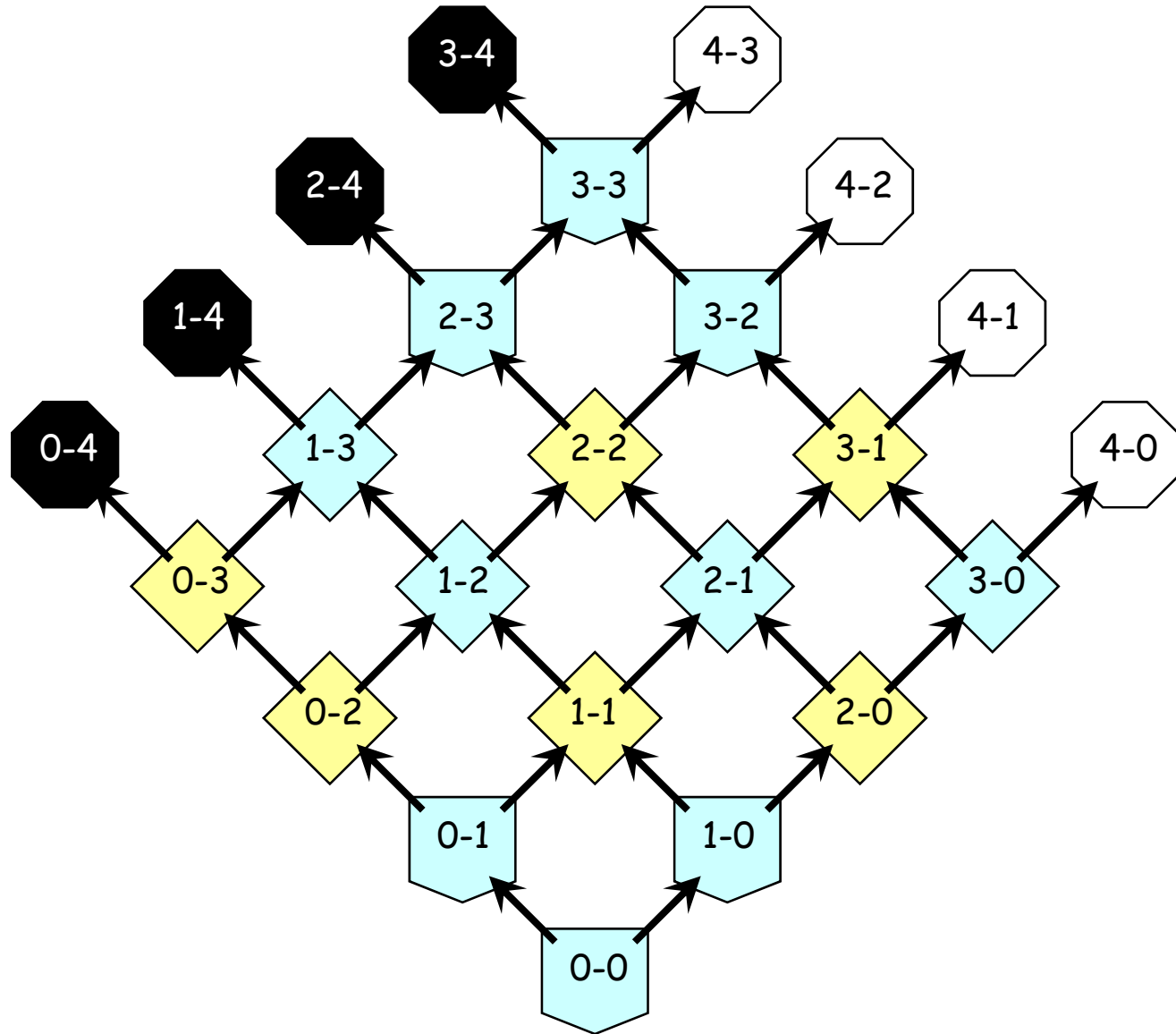
Probability
[0.70,1.00]
[0.60,0.70)
[0.55,0.60)
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

"Game Favorite" Solution for Favored Favorite Series ($p=0.806$)



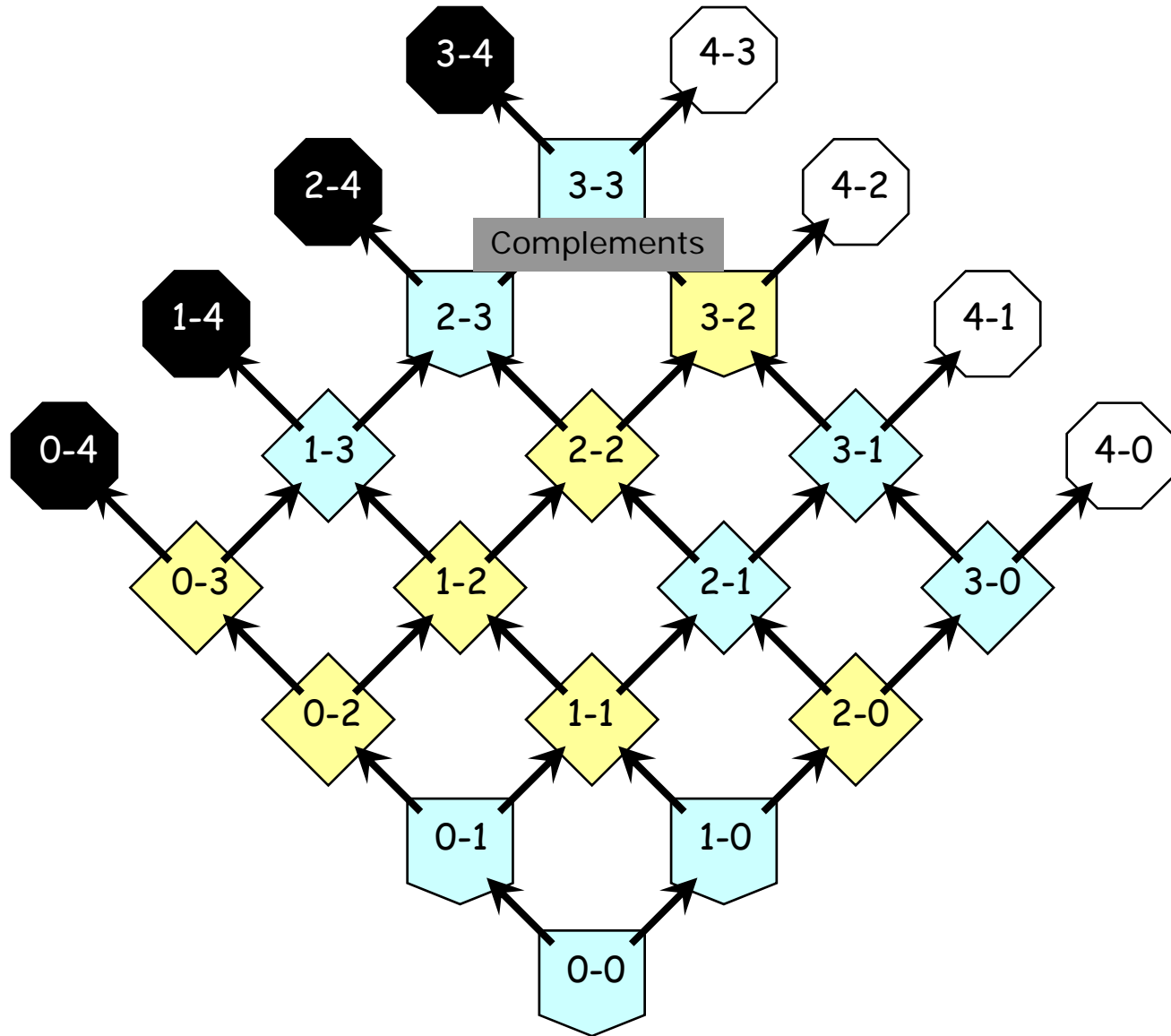
Probability
[0.70,1.00]
[0.60,0.70)
[0.55,0.60)
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

"Game Favorite" Solution for Favored Underdog Series ($p=0.738$)



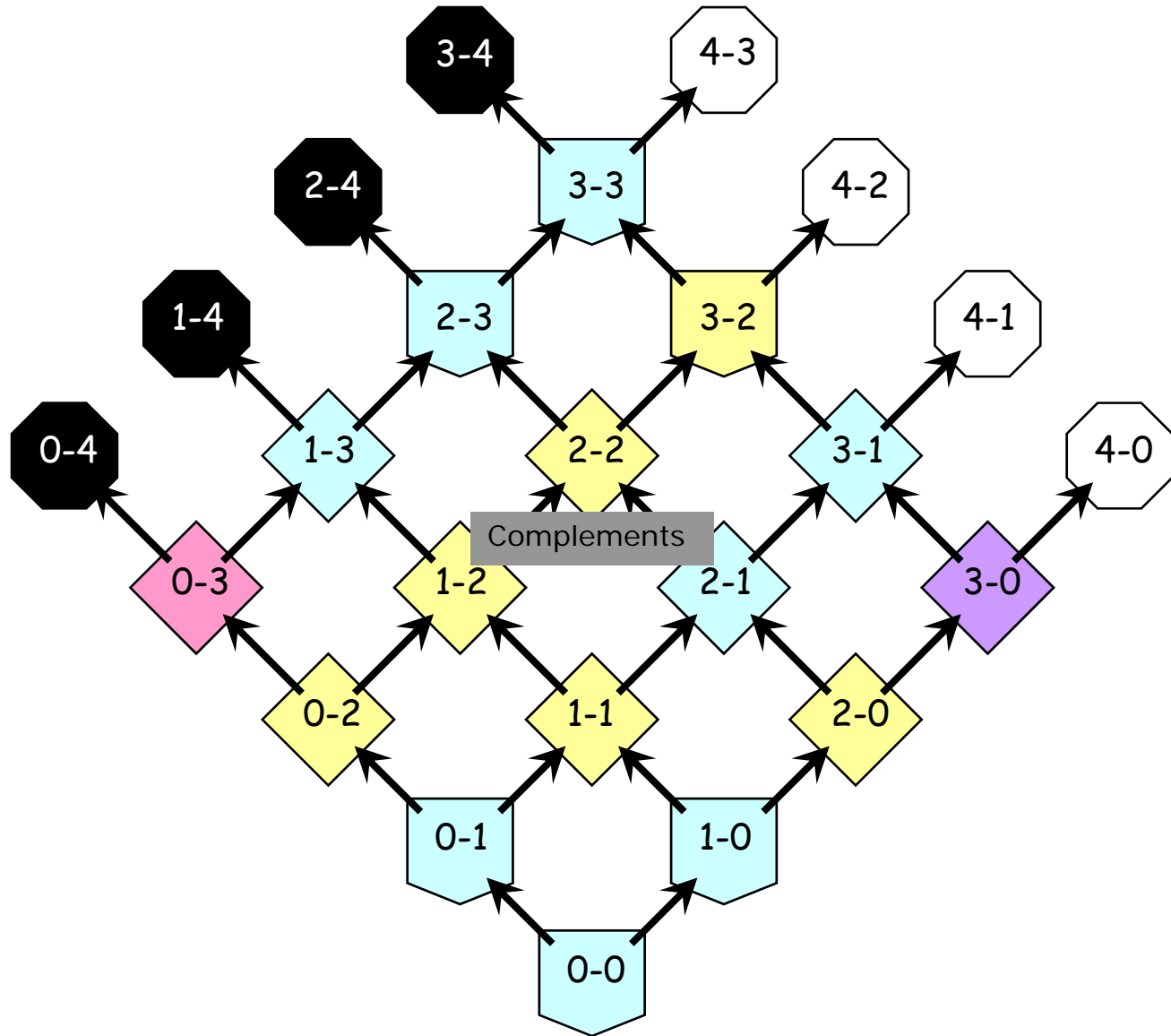
Probability
[0.70,1.00]
[0.60,0.70)
[0.55,0.60)
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

1 Complementary Cluster (p=0.688)



Probability
[0.70,1.00]
[0.60,0.70)
[0.55,0.60)
[0.50,0.55)
[0.40,0.50)
[0.30,0.40)
[0.00,0.30)

2 Complementary Clusters (p=0.790)



Probability
[0.70,1.00]
[0.60,0.70]
[0.55,0.60]
[0.50,0.55]
[0.40,0.50]
[0.30,0.40]
[0.00,0.30]