Basketball lineup performance prediction

Sports Network Seminar 2022

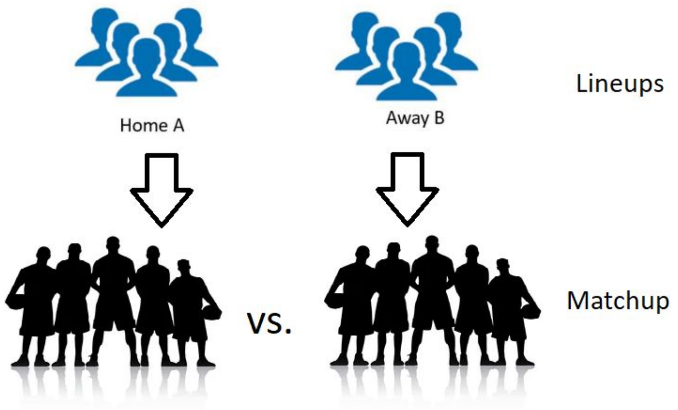
Hongruyu Chen, Oto Mraz

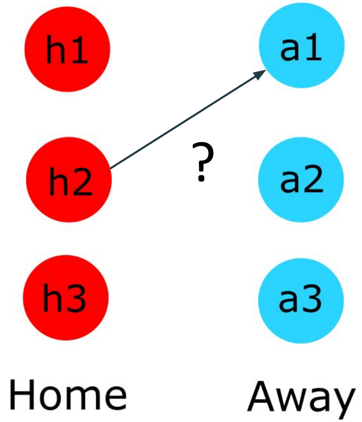
Motivation

- Advise trainer which team composition to use.
- Composition of opposing team not known in advance.
- Dynamically adjust strategy to win.
- Help predict the outcome of full matches (for betting).
- NBA season:
    - 30 teams in total (16 in playoffs, can’t afford to lose!)
    - 82 games for each team (2-4 games against all other teams)
    - Up to 15 players on a team => 3003 possible lineups to send!

Task

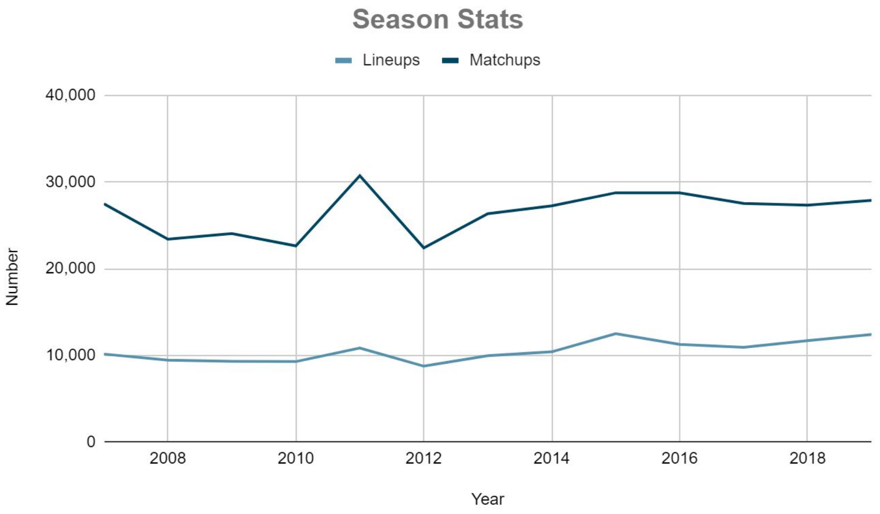
- Predicting the performance of basketball lineups in NBA games.
- **Lineup** - list of 5 players playing for a team at a given time.
- **Matchup** - 2 lineups playing against each other at a given time.
- E.g. Home and Away team have lineups h_1, h_2, h_3 and a_1, a_2, a_3.
- If h_i and a_j play against each other, who will win?



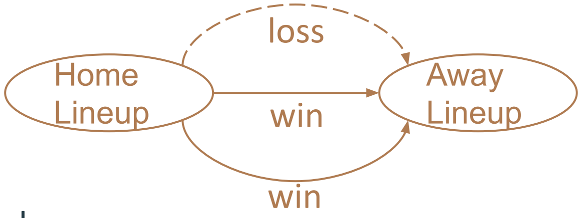


Difficulty

- The outcome depends on numerous factors (hard to quantify)
  - Strength of the individual players
  - Strength of lineup (collaboration)
  - Home or away location
  - Current score
  - Fatigue
- Magnitude of lineups and matchups



Data Preprocessing

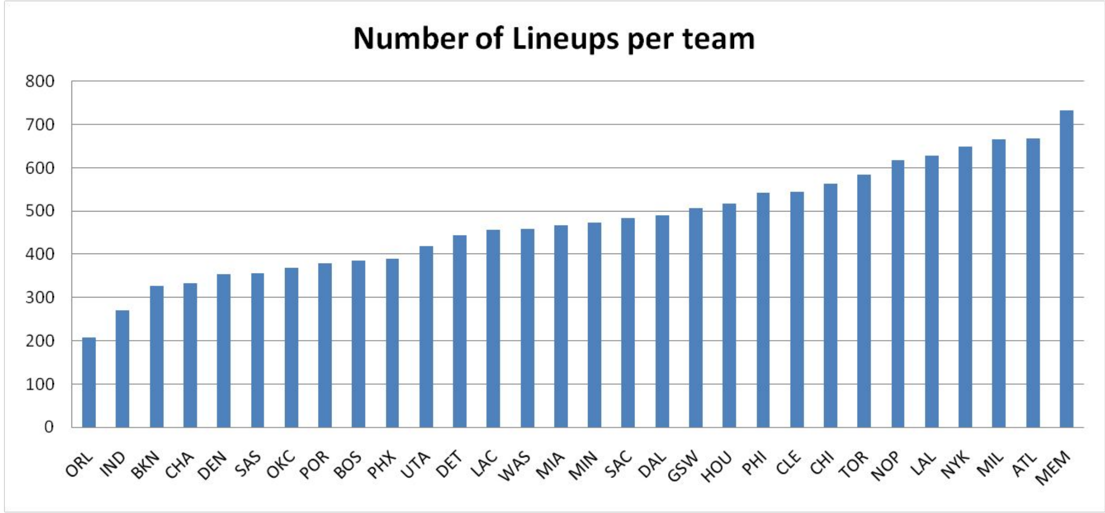


- Data collected from 2018-2019 season (regular + playoff).
- Extracted from 'play-by-play' tables.
- Collect the time and score of individual matchups.
- Aggregate the result for repeated occurrences of the same matchup (different time of the game, different games).

2nd Q			
Time	**Minnesota**	**Score**	**Golden State**
6:13.0		46-47	Defensive rebound by D. Jones
6:13.0	A. Wiggins enters the game for J. Okogie	46-47	
6:13.0		46-47	S. Curry enters the game for D. Green
6:13.0		46-47	K. Looney enters the game for K. Thompson
5:57.0		46-47	A. Iguodala misses 3-pt jump shot from 26 ft
5:55.0	Defensive rebound by K. Towns	46-47	
5:37.0	A. Tolliver misses 3-pt jump shot from 23 ft	46-47	

Data

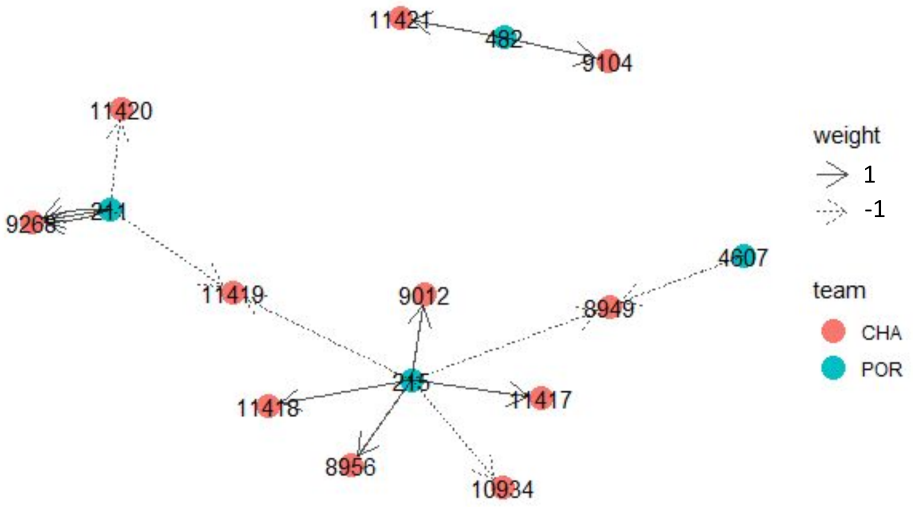
- 2018-2019 NBA Season
- 30 teams
- Lineups: 14281
- Matchups: 35922
- Matchups after removing draws: 29425



Directed Sign Graph

Subgraph of team **Portland Trail Blazers** (POR, home) and **Charlotte Hornets** (CHA, away)

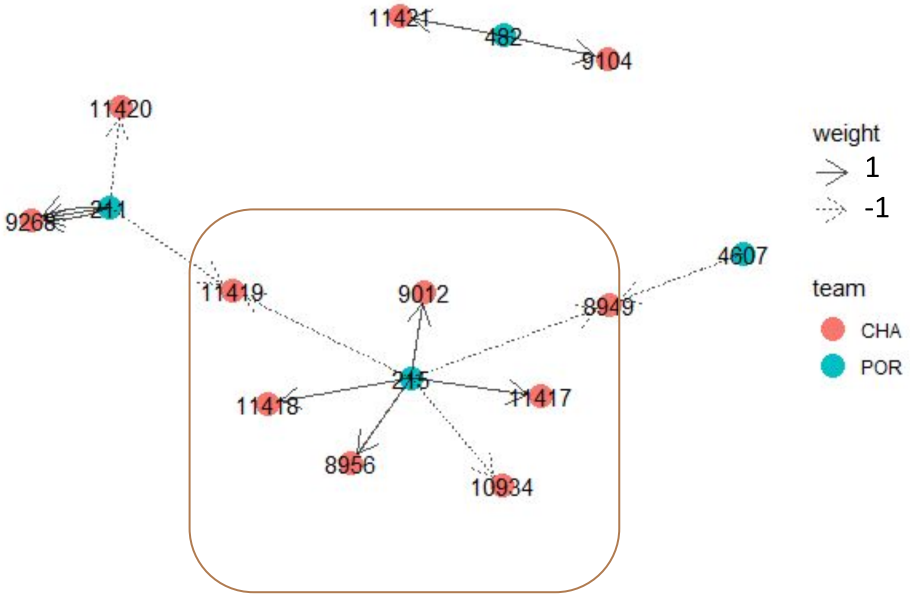
- Nodes: lineups
- Edges: matchups
- Direction: home to away
- Weight: 1(win) / -1(loss)



Network Analysis

Lineup 215

- Degree: 7
- Outdegree: 7
  - Positive outdegree: 4
  - Negative outdegree: 3
- Indegree: 0
  - Positive indegree: 0
  - Negative indegree: 0



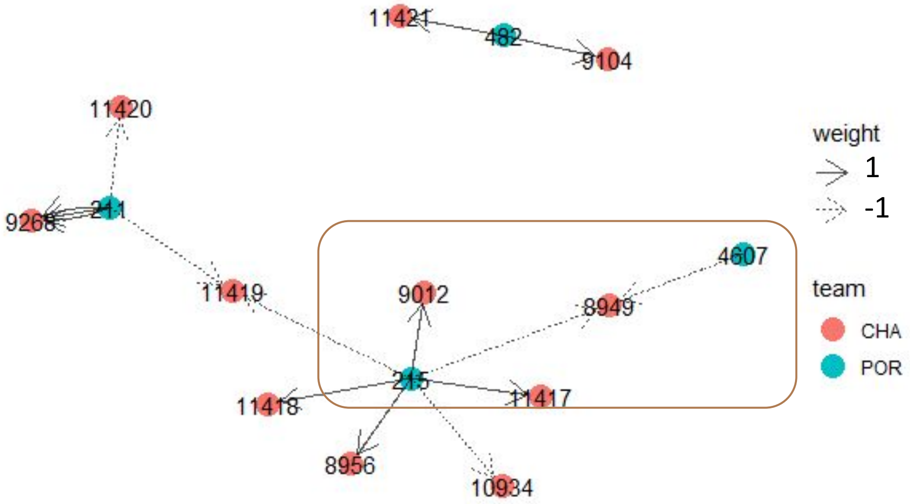
Network Analysis

Lineup 215

- Degree: 7
- Outdegree: 7
  - Positive outdegree: 4
  - Negative outdegree: 3
- Indegree: 0
  - Positive indegree: 0
  - Negative indegree: 0

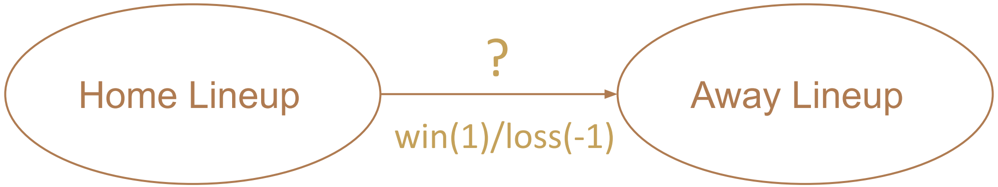
Path

215 - 8949 - 4607: undirected, length 2



Framework

Target: Given a lineup of the Home team and a lineup of the Away team, predict who wins (i.e. Sign of the edge).



Framework

Target: Given Home lineup and Away lineup, predict who wins.

↓

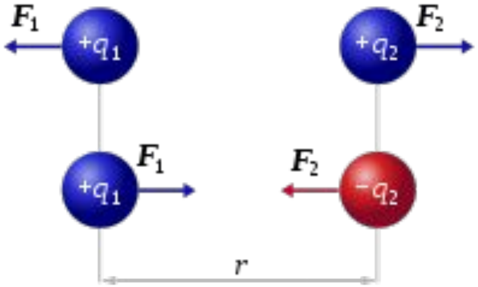
Extract features

- Train Models (train data, 80%)
  - Machine learning classification models
  - Random split / Time series
- Make Predictions (test data, 20%)
  - Binary : win/loss
  - Metric: accuracy

Extract Features — ISM Metric

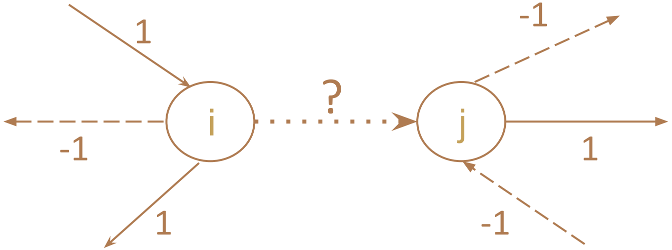
- Inverse Square Measure (ISM)
- Inspired by a physics model

$$|F_1| = |F_2| = k_e \frac{|q_1 \times q_2|}{r^2}$$



Coulomb's inverse-square law

$$ISM(i,j) = \frac{Deg(i) \cdot Deg(j)}{|SP(i,j)|^2}$$

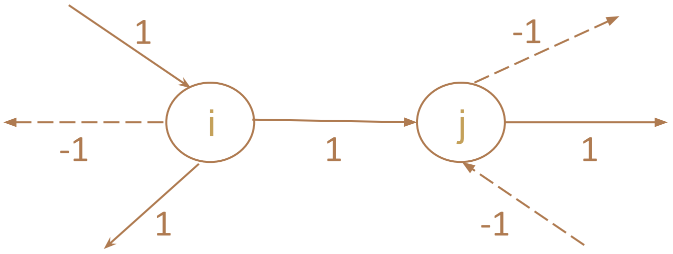


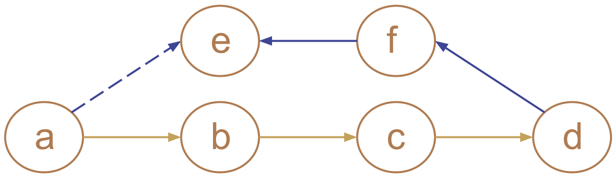
Extract Features — ISM Metric

- Inverse Square Measure (ISM)

$$ISM(i,j) = \frac{Deg(i) \cdot Deg(j)}{|SP(i,j)|^2}$$

- Variation
  - Degree (Deg) $\longrightarrow$ Positive (Negative) indegree, Positive (Negative) outdegree.
  - Shortest Path (SP) $\longrightarrow$ Directed / Undirected



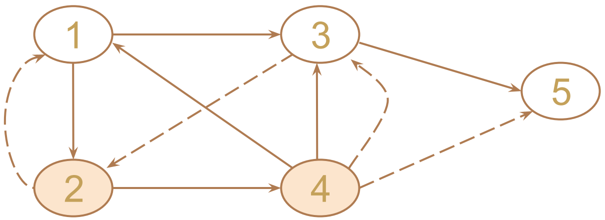


Extract Features — ISM Metric

- Inverse Square Measure (ISM)

$$ISM(i,j) = \frac{Deg(i) \cdot Deg(j)}{|SP(i,j)|^2}$$

- Variation
  - Degree (Deg) $\longrightarrow$ Positive (Negative) indegree, Positive (Negative) outdegree.
  - Shortest Path (SP) $\longrightarrow$ Directed / Undirected
- 16 perspectives $\longrightarrow$ 16-dim features for each ordered pair. E.g, for (2,4).



Full graph

$$ISM_{PinNout}(2,4) = \frac{Deg_{Pin}(2) \cdot Deg_{Nout}(4)}{|SP(2,4)|^2}$$

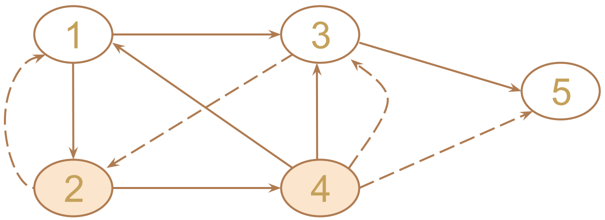
Extract Features — ISM Metric

- Inverse Square Measure (ISM)

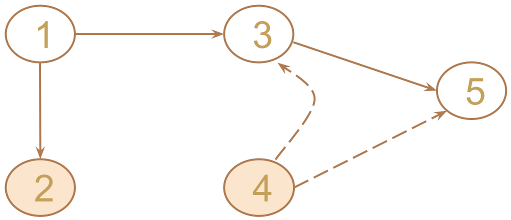
$$ISM(i,j) = \frac{Deg(i) \cdot Deg(j)}{|SP(i,j)|^2}$$

- Variation
  - Degree(Deg) $\longrightarrow$ Positive(Negative) indegree, Positive(Negative) outdegree.
  - Shortest Path(SP) $\longrightarrow$ directed / undirected
- 16 perspectives $\longrightarrow$ 16-dim features for each ordered pair. E.g, for (2,4).

$$ISM_{PinNout}(2,4) = \frac{Deg_{Pin}(2) \cdot Deg_{Nout}(4)}{|SP(2,4)|^2} = \frac{1 \cdot 2}{3^2} = \frac{2}{9}$$



Full graph



Positive In (2) Negative Out (4) graph

Models

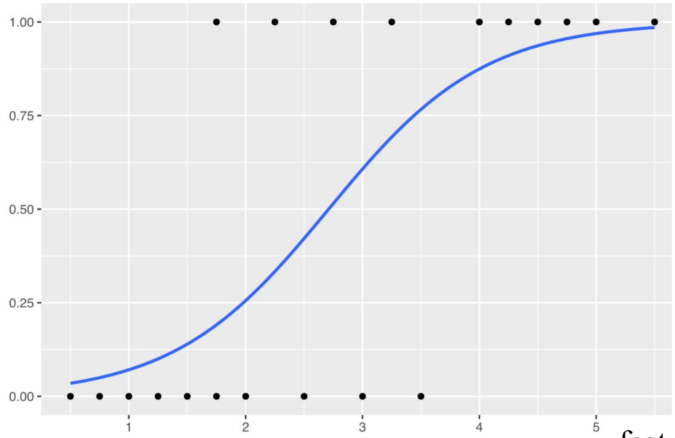
Step 1: Data split: train data + test data

- Random 80% train + 20% test
- Time series First 80% train + Rest 20% test (we predict results based on the past)

Step 2: Apply machine learning models on train data

- Logistic Regression (paper's approach)

Prob. of win



feature

Logistic regression

Models

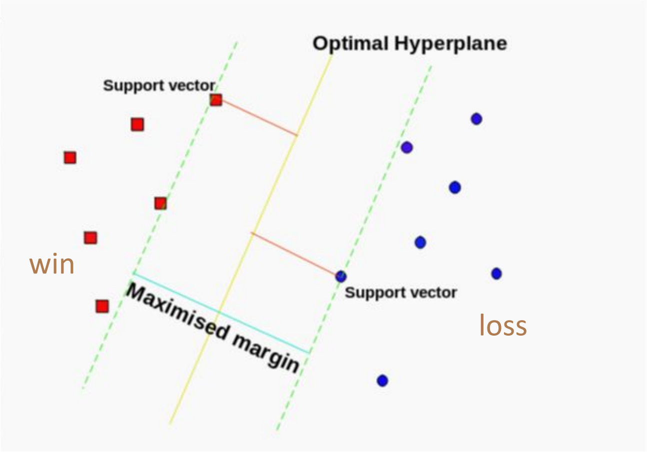
Step 1: Data split : train data + test data

- Random 80% train + 20% test
- Time series First 80% train + Rest 20% test

Step 2: Apply machine learning models on train data

- Logistic Regression (paper's approach)
- Support Vector Machine
- Gaussian Process
- Random Forests
- …

Step 3: Prediction on test data



SVM

Our Methods - ISM Variations

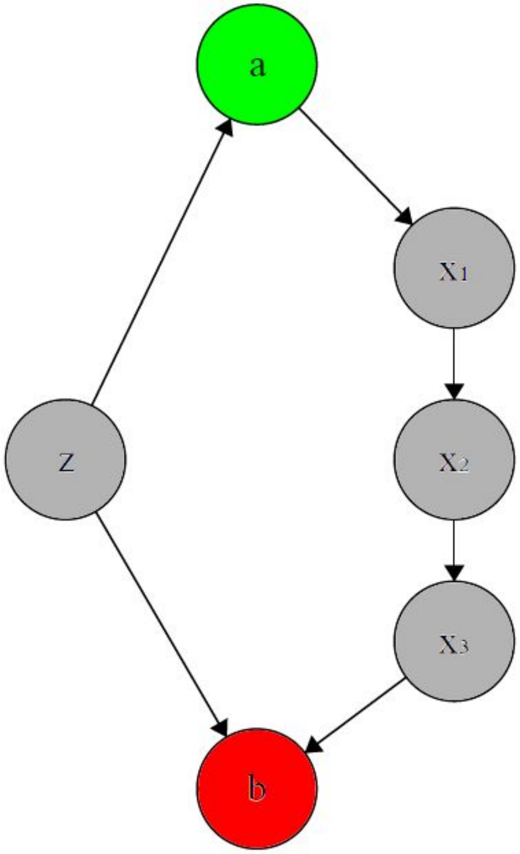
- Paper's approach: quadratic penalization of distance

$$ISM(i,j) = \frac{Deg(i) \cdot Deg(j)}{|SP(i,j)|^2}$$

- Our findings: linear penalization improves accuracy by ~ 1%

$$ISM(i,j) = \frac{Deg(i) \cdot Deg(j)}{|SP(i,j)|}$$

- 2 options for computing shortest path:
  - For indegree(a), indegree(b) must use undirected
  - For indegree(a), outdegree(b) can use directed
  - May carry more information
  - Often results in a longer path

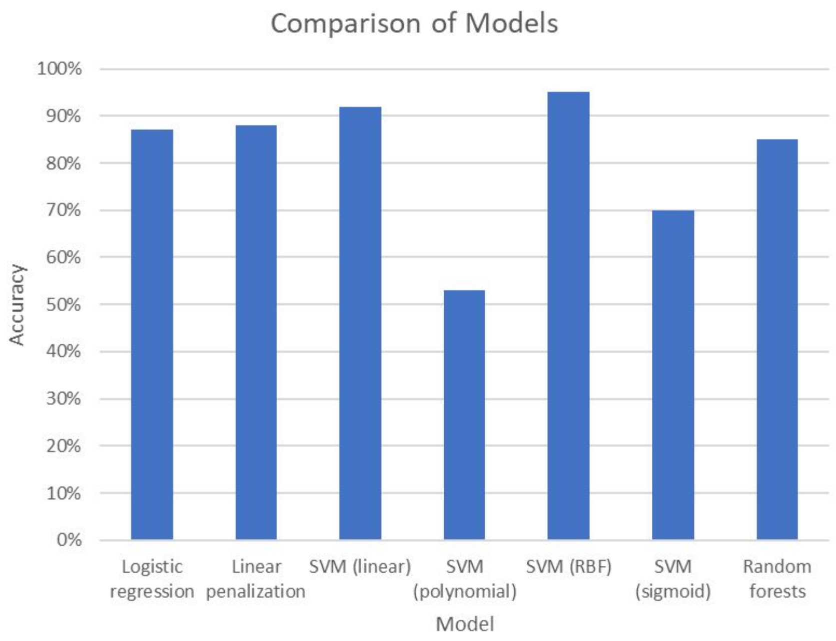


Our Methods - ML models

- Reproducing logistic regression (with ISM variation).
- Dataset has a large number of training samples.
- Small amount of features (16).
- SVMs (with appropriate kernel) more suitable.
  - Lower risk of overfitting.
  - Better at capturing data shape.
  - More stable solutions.
- Random forests suitable (lots of training data).

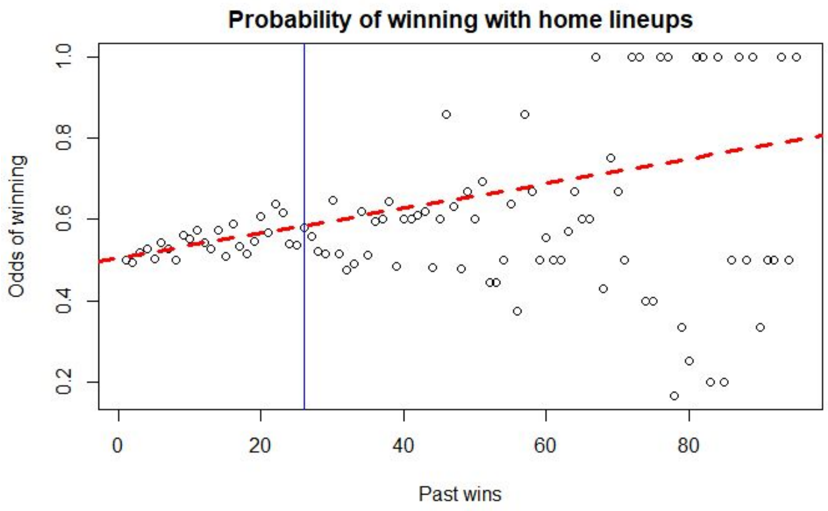
Comparison

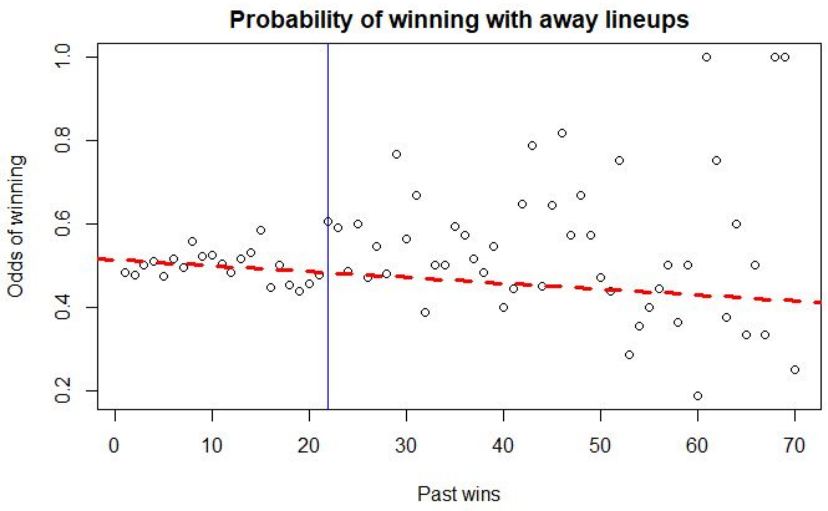
- Paper’s approach (logistic regression): 87% accuracy
- Linear ISM penalization: 88%
- SVMs:
  - Linear kernel: 92%
  - Polynomial kernel: 53%
  - **RBF kernel: 95% (most successful)**
  - Sigmoid: 70%
- Gaussian processes computationally infeasible
- Random forest: 85%



Model assumption - Preferential Attachment

Is a team that has won many times **more likely** to win?





$$\text{Odds of winning} = 0.506 + 0.003 * \text{Past wins}$$

$$\text{Odds of winning} = 0.513 - 0.001 * \text{Past wins}$$

Critique - Revisit Framework

Target: Given Home lineup and Away lineup, predict who wins.

↓

Extract features

- Train Models (train data, 80%)
  - Machine learning classification models
  - Random split / Time series
- Make Predictions (test data, 20%)
  - Binary : win/loss
  - Metric: accuracy

Critique - Revisit Framework

Target: Given Home lineup and Away lineup, predict who wins.

Extract features

Problem: Using all data!

Train Models (train data, 80%)

- Machine learning classification models
- Random split / Time series

Make Predictions (test data, 20%)

- Binary : win/loss
- Metric: accuracy

Critique - Revisit Feature Extraction

- Inverse Square Measure (ISM)

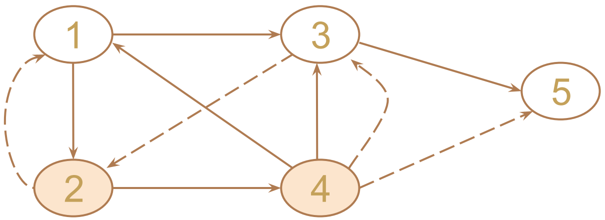
  16 perspectives $\longrightarrow$ 16-dim features

  Example:

Target (test data) : Assume 2 is home, 4 is away, predict the sign (1/-1). $2 \xrightarrow{?} 4$

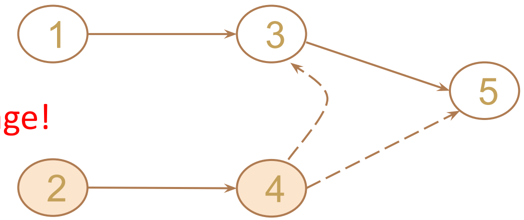
Shouldn't include all the edges from 2 to 4 when extracting features!

$$ISM_{PoutNout}(2,4) = \frac{Deg_{Pout}(2) \cdot Deg_{Nout}(4)}{|SP(2,4)|^2}$$



Full graph

Information leakage!



Positive Out (2) Negative Out (4) graph

Conclusion & Future Work

- Demonstrated the strength of ISM metric.
- Tested out several metric variants (*Best: undirected linearly penalized*).
- Preferential attachment wasn’t found.
- SVM with RBF kernel proves to be more robust.
- Feature set is relatively small (only 16 predictors).
- Information leakage might be a problem.
- Further extension by additional features (e.g. plus/minus points, rebounds, etc.)

References

1. Ahmadalinezhad, Mahboubeh, and Masoud Makrehchi. "Basketball lineup performance prediction using edge-centric multi-view network analysis." *Social Network Analysis and Mining* 10.1 (2020): 1-11.
2. Ahmadalinezhad, Mahboubeh. *Link mining in signed social networks*. Diss. University of Ontario Institute of Technology (Canada), 2020.
3. Ahmadalinezhad, Mahboubeh, and Masoud Makrehchi. "Sign prediction in signed social networks using inverse squared metric." *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, Cham, 2018.
4. Junzhou Zhao, John C. S. Lui, D. Towsley, Xiaohong Guan and Yadong Zhou, "Empirical analysis of the evolution of follower network: A case study on Douban," *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2011, pp. 924-929, doi: 10.1109/INFCOMW.2011.5928945.