UNIVERSITY OF Southampton

# Predicting Winners of Competitive Events with Topological Data Analysis


Conrad D'Souza
cds1g09@soton.ac.uk

Ruben Sanchez-Garcia
R.Sanchez-Garcia@soton.ac.uk

Tiejun Ma
tiejun.ma@soton.ac.uk

Johnnie Johnson
J.E.Johnson@soton.ac.uk

Ming-Chien Sung
M.SUNG@soton.ac.uk



EPSRC
Engineering and Physical Sciences Research Council

# Research Aims

- Develop topological tools for data analysis
- Measure the underlying quality of horses based on past race performances
- Make profitable predictions from the results

Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Horse Racing Data

- Data consists of outcome of UK horse races between 2005 and 2014
- 70261 races competed in by 64691 horses
- Various performance indicators including finishing position and beaten lengths

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

# Handicapping

- Handicapping decreases the predictability of races
- Horses are given extra weight to carry to inhibit their performance
- Aim of handicappers is that races finish in a dead heat
- Account for this by estimating the results had all horses carried the same weight

cds1g09@soton.ac.uk

# Pairwise Scores

- Each race $\alpha$ , with performance indicator $P$, forms a local pairwise score matrix $Y^{\alpha}$ with

$$Y^{\alpha}{}_{ij} = P_i - P_j$$

- Information is aggregated, with respect to a reliability measure



cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Indirect Comparisons

| Horse | Finishing Position |
|---|---|
| A | 1 |
| B | 2 |

| Horse | Finishing Position |
|---|---|
| B | 1 |
| C | 2 |

Which is better? A or C?


EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Optimisation Problem

- HodgeRank finds global scores $s$ minimising the weighted square error between the induced and observed pairwise comparisons
- Want to solve the optimisation problem:

$$\min_{s \in \mathbb{R}^m} \sum_{i,j} W_{ij}(s_i - s_j - \bar{Y}_{ij})^2$$

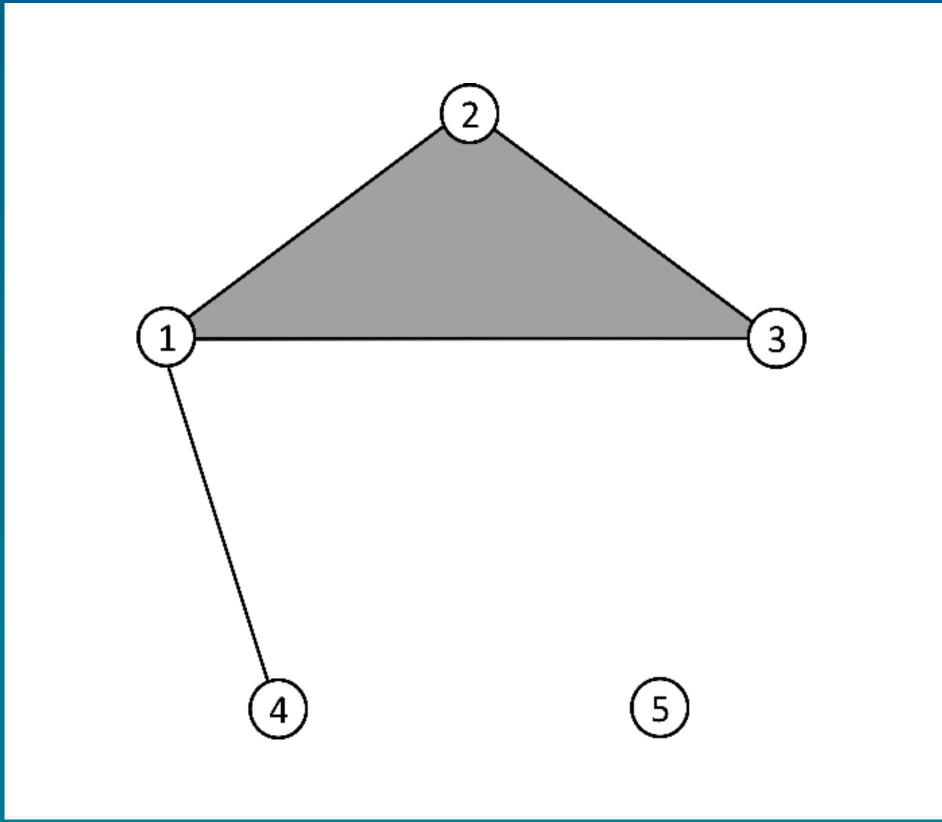
EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Flag Complex Representation

- Horses form the 0-skeleton
- Pairwise scores form the 1-simplices

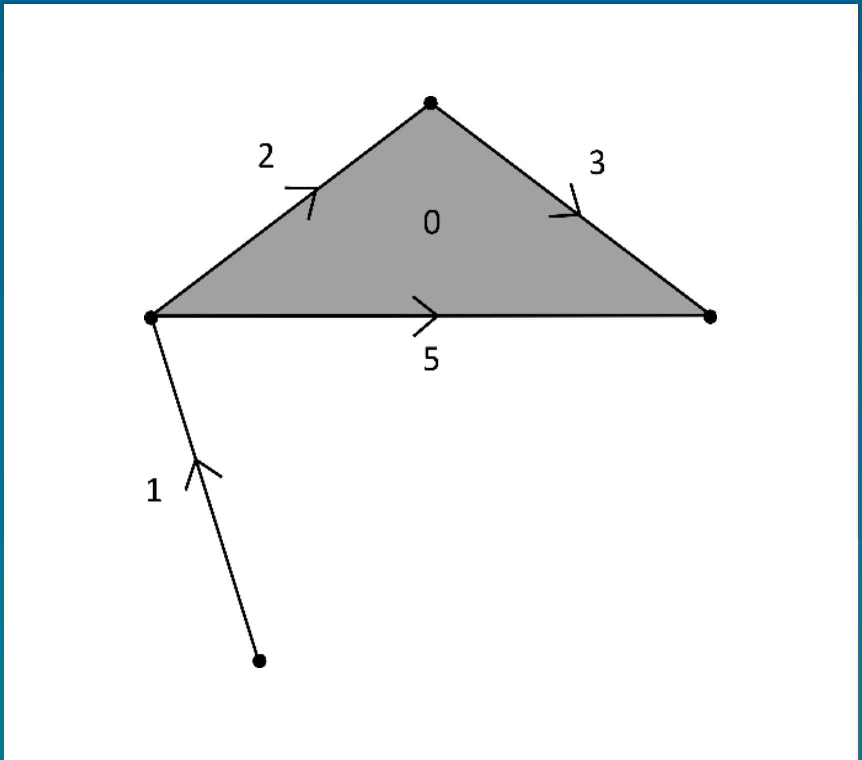



EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

# Cochains

- $k$-cochains are real-valued functions on the $k$-simplices
- Reversing direction negates the value of the cochain
- Set of all $k$-cochains is denoted $C^k$




cds1g09@soton.ac.uk

EPSRC
Engineering and Physical Sciences Research Council

UNIVERSITY OF Southampton

# Inner Products

- Choose unweighted Euclidean inner products on $C^0$ and $C^2$
- Equip $C^1$ with a weighted Euclidean inner product

$$\langle f, g \rangle_{C^1} = \sum_{i \in E} W(i) f(i) g(i)$$

cds1g09@soton.ac.uk

EPSRC
Engineering and Physical Sciences Research Council

# Coboundary Operators

- $k$-th coboundary operator is a linear map

$$\delta_k : C^k \rightarrow C^{k+1}$$

- Adjoint of the $k$-th coboundary operator is a linear map

$$\delta_k{}^* : C^{k+1} \rightarrow C^k$$

- $k$-th combinatorial Laplacian is a map

$$\Delta_k = \delta_k{}^* \circ \delta_k + \delta_{k-1} \circ \delta_{k-1}{}^* : C^k \rightarrow C^k$$

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Hodge Decomposition Theorem

$C^k$ admits an orthogonal decomposition

$$C^k = im(\delta_{k-1}) \oplus ker(\Delta_k) \oplus im({\delta_k}^*)$$

and

$$ker(\Delta_k) = ker(\delta_k) \cap ker({\delta_{k-1}}^*)$$

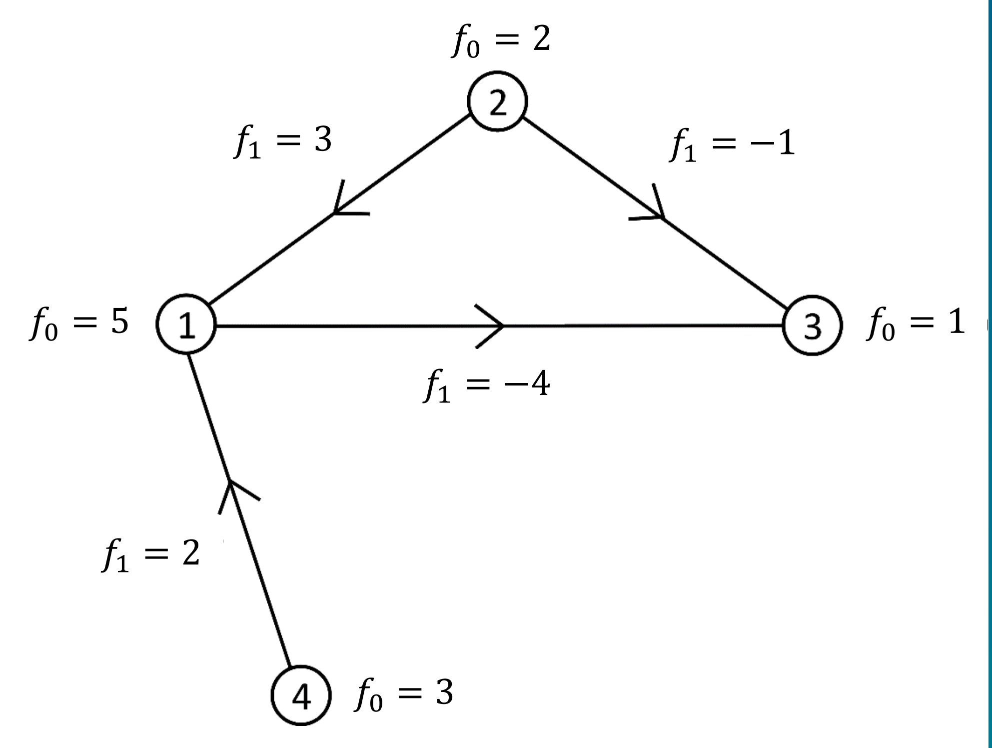
Split $\bar{Y}$ into orthogonal cochains according to the Hodge Decomposition Theorem

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

# Globally Consistent

- $im(\delta_0)$ are globally consistent cochains
- Any $f_1 \in im(\delta_0)$ has the form $f_1(i,j) = f_0(j) - f_0(i)$ for some $f_0 \in C^0$




cds1g09@soton.ac.uk

UNIVERSITY OF Southampton
EPSRC
Engineering and Physical Sciences Research Council

# Optimisation Problem

- Since $im(\delta_0)$ is the set of all globally consistent pairwise scores, the optimisation problem can be written as:

$$\min_{s \in \mathbb{R}^m} \sum_{i,j} W_{ij}(s_i - s_j - \bar{Y}_{ij})^2 = \min_{s \in \mathbb{R}^m} |\delta_0 s - \bar{Y}|_{2,W}^2$$

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Optimisation Problem Solved

Solutions to the optimisation problem satisfy

$$\Delta_0 s = {\delta_0}^* \bar{Y}$$

and the minimum norm solution is given by

$$s' = {\Delta_0}^{\dagger} {\delta_0}^* \bar{Y}$$

$s'$ is unique up to an additive constant

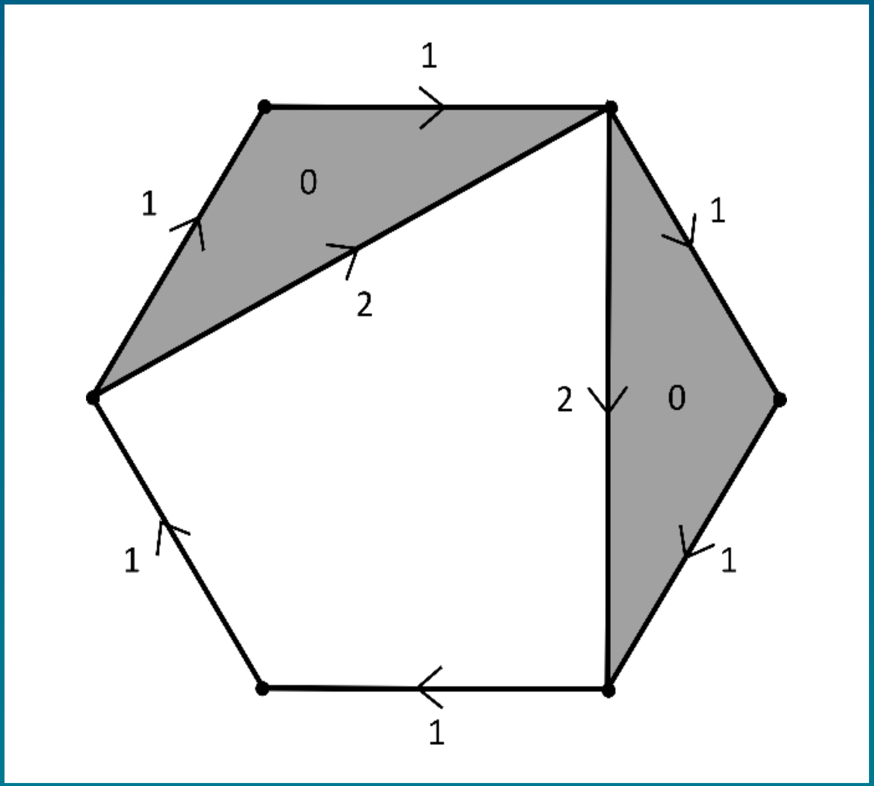
EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Inconsistencies

- $\ker(\Delta_1)$ are partial inconsistencies
- Every triple of horses are consistent but larger cycles are inconsistent

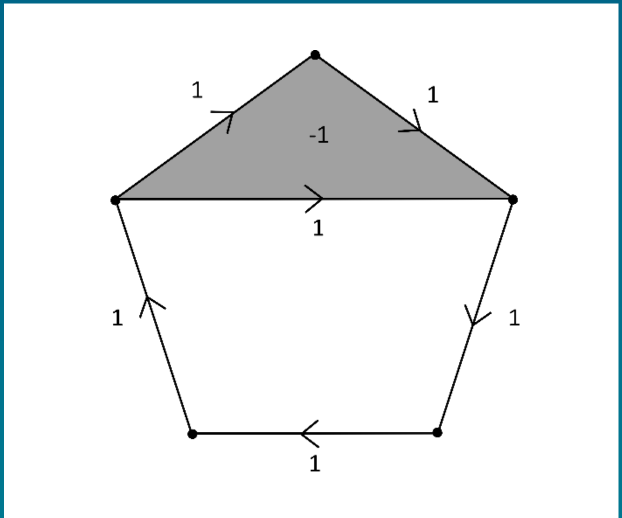



EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

# Inconsistencies

- $im(\delta_1{}^*)$ are complete inconsistencies
- These are functions which are inconsistent on every level




cds1g09@soton.ac.uk

# Partial Inconsistency Weights

- Measure how far a 1-simplex is from local consistency by

$$\phi_1(i) = |(\delta_1{}^*\delta_1\bar{Y})(i)|$$

- Reweight complex by

$$W(i) = \frac{1}{\phi_1(i) + 1}$$

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

# Predicting Winners

- Train and test a conditional logit model using public odds as a predictor
- Assess impact of adding global score variable to the model
- Measure goodness-of-fit by McFadden's $\tilde{R}^2$
- Log Likelihood Ratio Test determines benefit of adding signal variable

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

# Conditional Logit Model

- Generate vector of winning probabilities for each horse in each race

$$\boldsymbol{p}_i^\alpha = (p_1^\alpha, \ldots, p_n^\alpha)$$

- Probabilities based on predictive variables

$$\boldsymbol{x}_i^\alpha = (x_1^\alpha(1), \ldots, x_1^\alpha(m))$$

- Representative utility for horse $i$ in race $\alpha$

$$u_i^\alpha = \sum_{k=1}^{m} \beta(k) x_i^\alpha(k) + \varepsilon_i^\alpha$$

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton
EPSRC
Engineering and Physical Sciences Research Council

UNIVERSITY OF Southampton

# Conditional Logit Model

- Assuming error terms are identically and independently distributed via a double exponential distribution, probabilities are given by

$$p_i^\alpha = \frac{\exp[\sum_{k=1}^{m} \beta(k) x_i^\alpha(k)]}{\sum_{i=1}^{n_\alpha} exp\left[\sum_{k=1}^{m} \beta(k) x_i^\alpha(k)\right]}$$

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Kelly Wagering Strategy

- Fractional Kelly wagering strategy employed to assigns bets
- A fraction of the initial capital is bet, given by

$$f = \frac{p(b+1)-1}{b}$$

where $b$ is the odds ratio $b\!:\!1$ and $p$ is the estimated probability of winning

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

# Results

- Model trained over 2011 to 2013 and evaluated in 2014
- LLR test significant at the 0.1%
- $\tilde{R}^2$ increase of 0.227% over public odds model with $\tilde{R}^2$ of 0.16547

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton

# Results

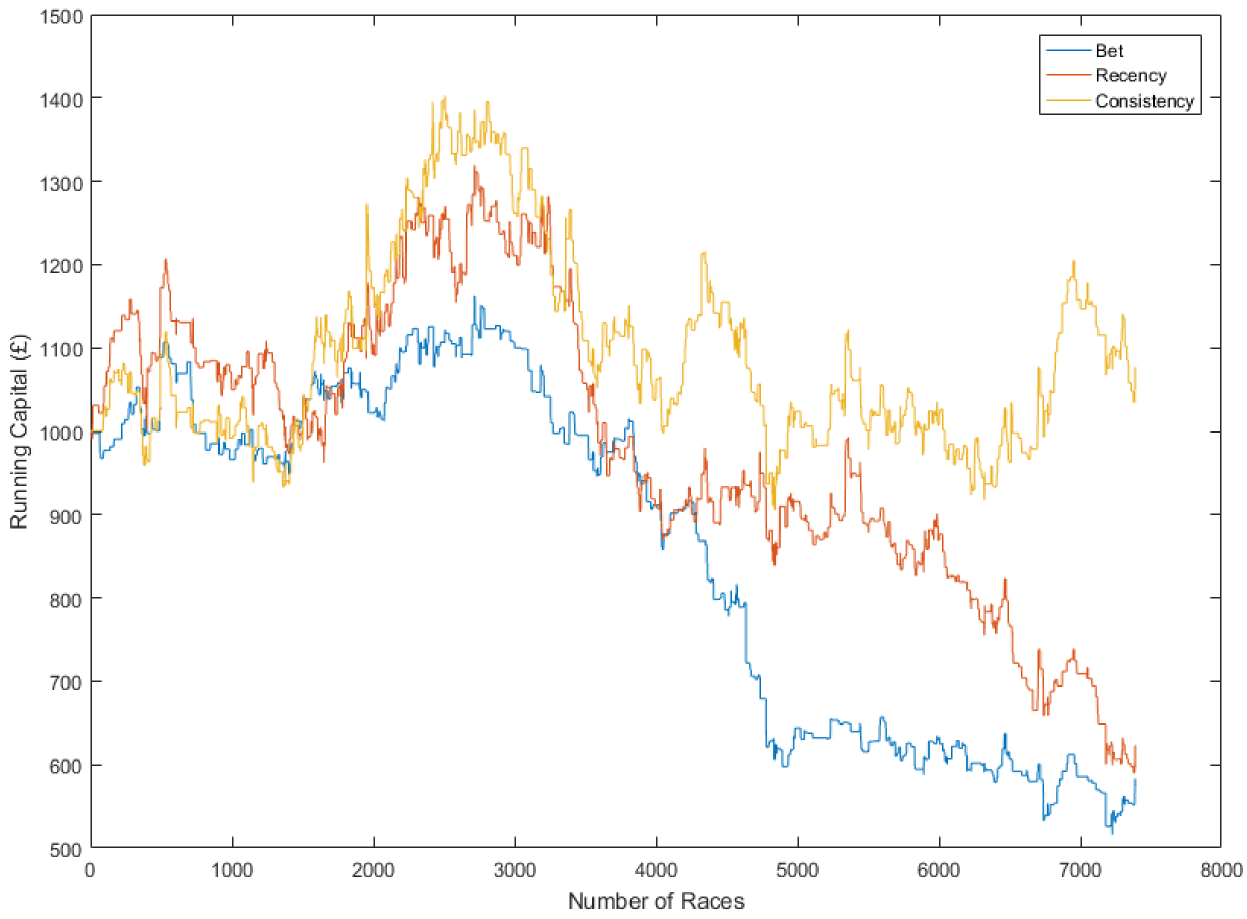
Betting simulation results (initial capital of £1000)

| Model | Profit (£) | Rate of Return (%) |
|---|---|---|
| excl. scores | -424.67 | -7.40 |
| incl. scores | 52.05 | 0.48 |

EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk

UNIVERSITY OF Southampton




EPSRC
Engineering and Physical Sciences Research Council

cds1g09@soton.ac.uk