Understanding following patterns among Olympic medallists

JORGE P. RODRÍGUEZ

In collaboration with:
LLUÍS AROLA-FERNÁNDEZ

SocioMeeting

IFISC, Palma, 25th June 2024

The case of Simone Biles *

Google Trends Principal Explorar Tendencias actuales

simone biles
Término de búsqueda

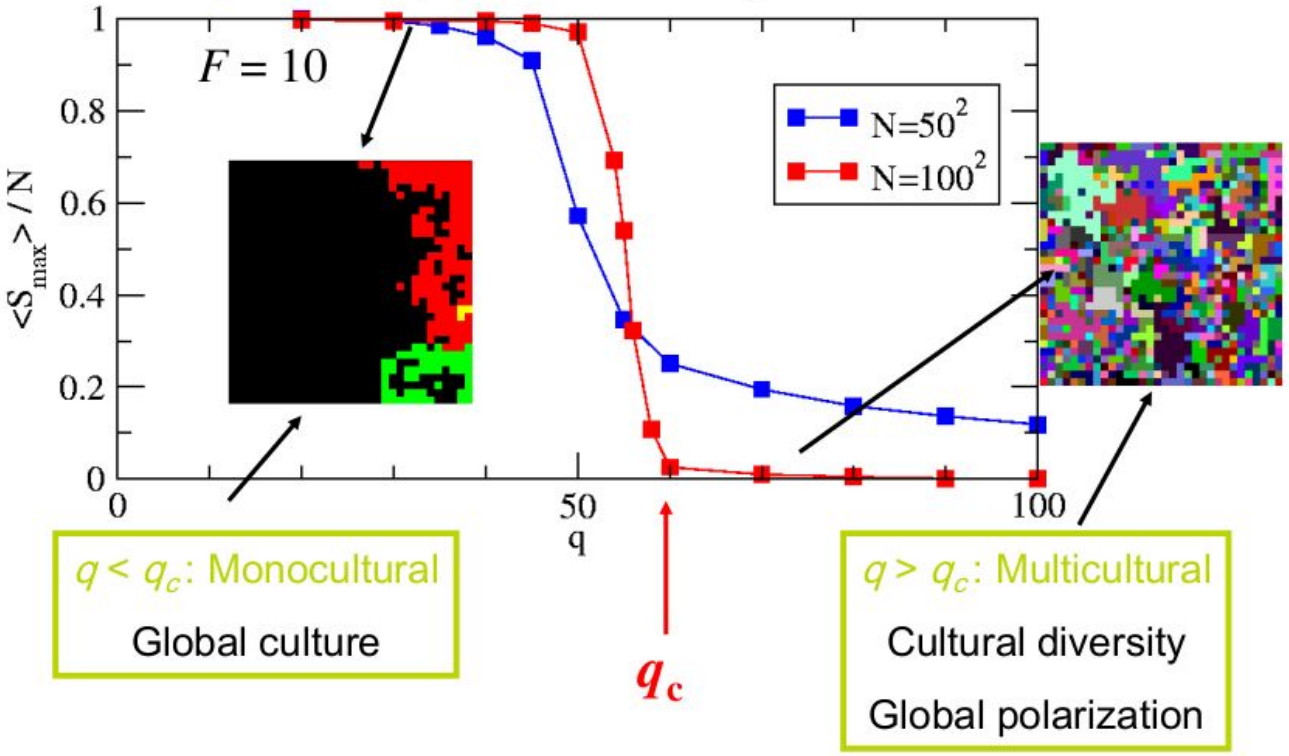
Interés a lo largo del tiempo

IFISC

Polarization-Globalization Transition

Castellano, Marsili, Vespignani, Phys. Rev. Lett. **85**, 3536 (2000)
San Miguel et al., Computing in Science and Engineering **7**, 67 (2005)

- **Order parameter:** S_{max} size of the largest homogeneous domain
- **Control parameter:** q measures initial degree of disorder.



Model of cultural dissemination

Individual states are described by vectors (F features, q possible traits per feature)

Principles:

1) Homophily (I interact more with similar peers)
2) Social influence (interactions change my state)

Source: Maxi San Miguel

Questions:

- What is the underlying structure for information flows among athletes?
- Is it reciprocal? I follow you if you follow me and vice versa?
- Are there features that drive homophilic interactions?
- Are all the features equally important? How can we model them?

1. Creating a database of **Olympic medallists** (Tokyo 2020) who have a public Twitter account

2. Querying the **Twitter** API (January 2023), searching for links among those athletes

   2.1 WARNING: querying rates were limited. Trick: typically, for celebrities, the number of followers was much higher than the number of followees. Thus, we looked for accounts in our database that followed each athlete. This decreased the scraping time to ~1 week (real life: a few failures due to account deletion delayed the process).

3. For each node, we had **metadata**: sport, country, sex, award(s) and number of followees

4. Final product: network (A followed B) + metadata

The Olympic medallists following network *

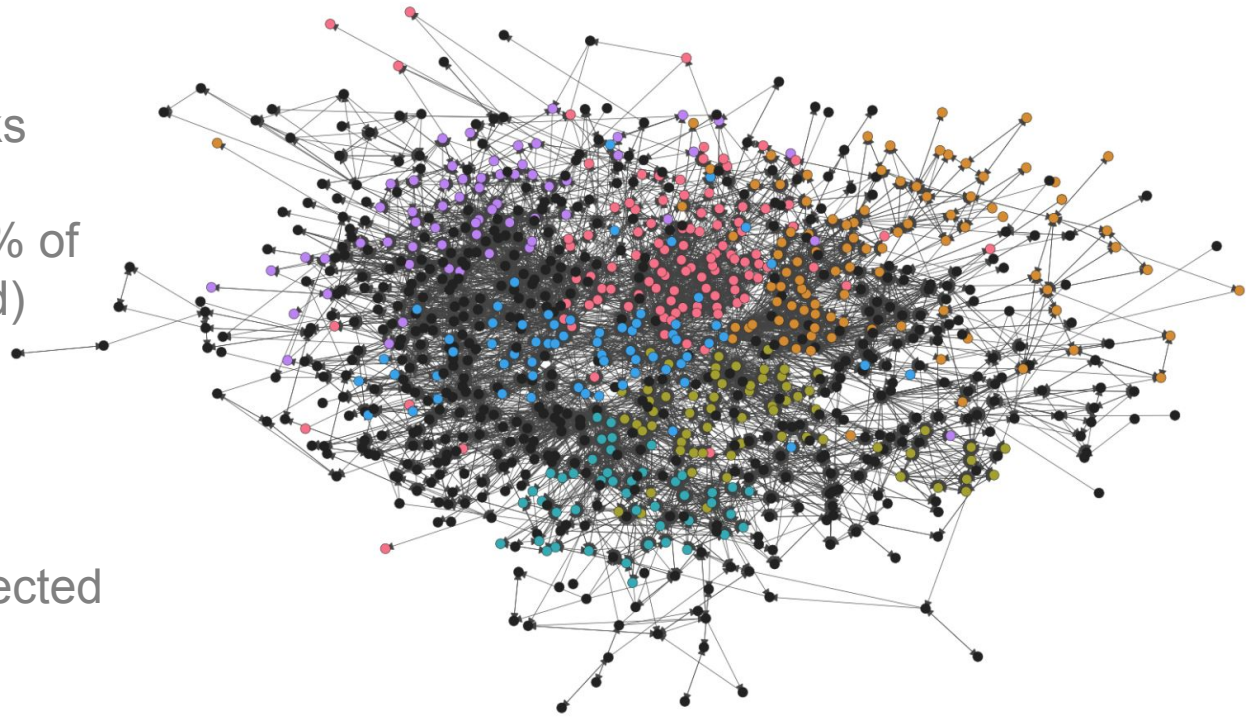
N = 964 athletes

E = 7,326 directed links

Sparse network (0.79% of links in fully-connected)

Largest connected component: L = 956

Largest strongly connected component: S = 827



Athletics
Basketball
Cycling
Handball
Football
Swimming

Rank of athletes by their number of followers (k_{in})

Rank	Name	Sport	Country	k_{in}	Medal(s)
1	Kevin Durant	Basketball	USA	60	🥇
2	Allyson Felix	Athletics	USA	56	🥇🥉
3	Teddy Riner	Judo	France	54	🥇🥉
4	Alex Morgan	Football	USA	52	🥉
5	Simone Biles	Gymnastics	USA	50	🥈🥉
6	Megan Rapinoe	Football	USA	48	🥉
7	Adam Peaty	Swimming	Great Britain	46	🥇🥇🥈
8	Nikola Karabatic	Handball	France	43	🥇
9	Noah Lyles	Athletics	USA	40	🥉
9	Tom Daley	Diving	Great Britain	40	🥇🥉

A directed following network

Directed network: A following B does not imply B following A

But 49.6% of the unique connections (undirected) were reciprocal

Reciprocity: if you follow me, will I follow you back? If I follow you, will you follow me back?

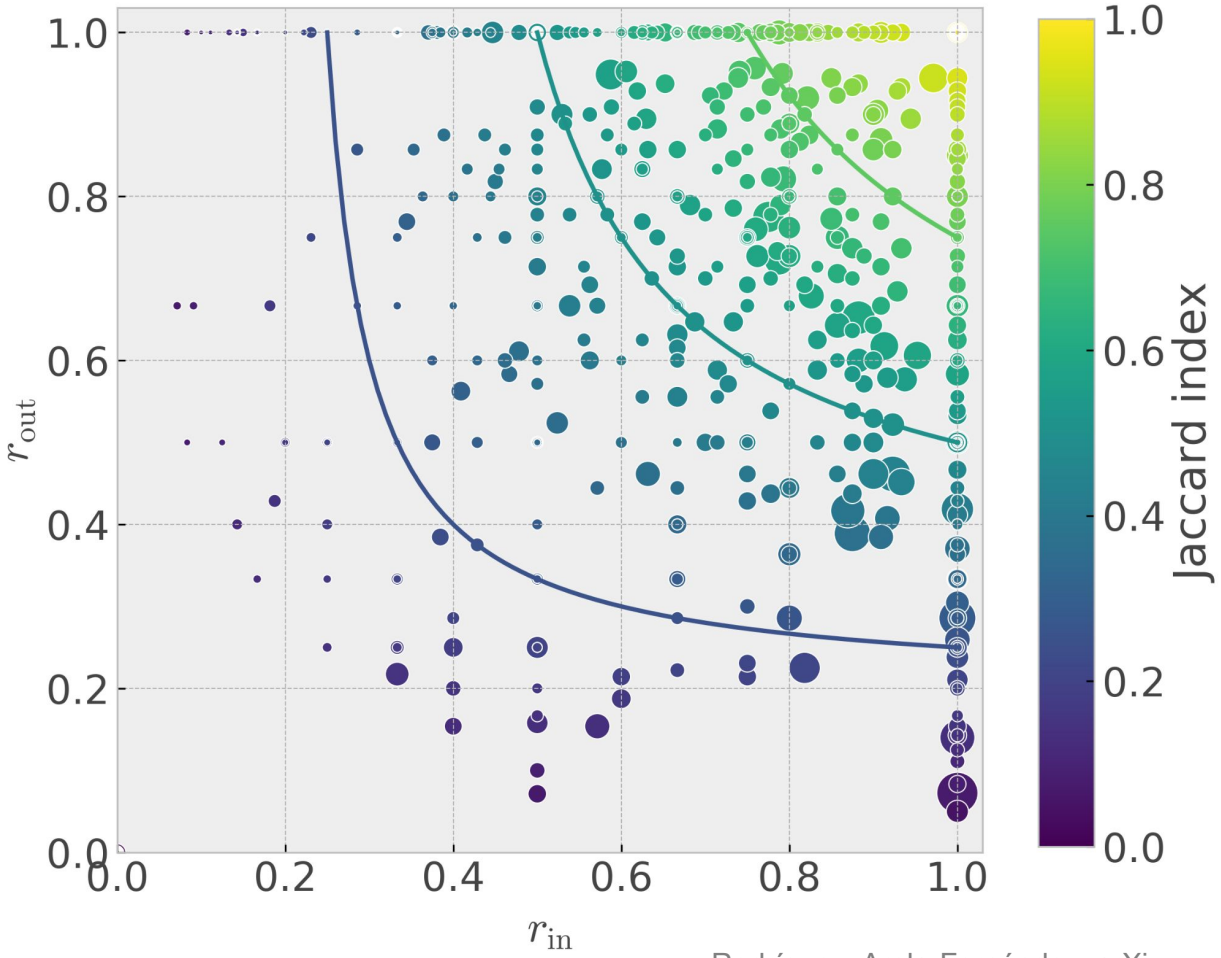
A directed following network *

Directed network: A following B does not imply B following A

But 49.6% of the unique connections (undirected) were reciprocal

Reciprocity: if you follow me, will I follow you back? If I follow you, will you follow me back?

A) Jackard index per individual between followers and followees
B) (r_{in},r_{out}) space



Symbol size: in-degree

Rodríguez, Arola-Fernández, arXiv (2024)

Sex homophily *

- 73% of the links connected athletes with the same sex

- Adjacency matrix by sex

- k_{in}/k_{out} = 1.05 (men), 0.96 (women)

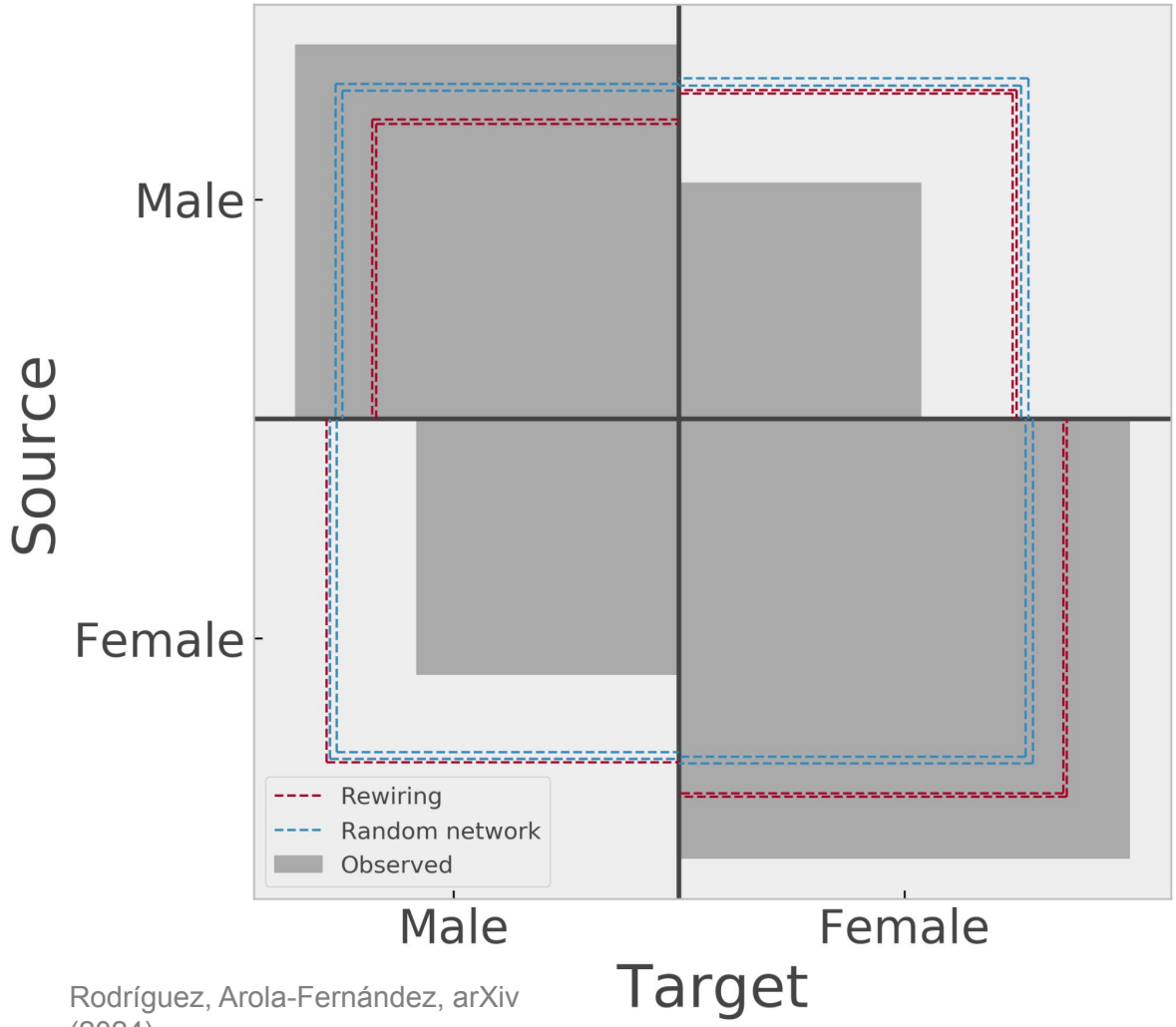
- Reference null models:

A) Rewiring: keeping individual properties, individual in- and out-degrees

B) Random directed network: keeping individual properties, average degree (in- and out-)

- Different diagonal behaviour: similar compartment size (M/F), but k_{in}(F)k_{out} (F) = 1.6 k_{in}(M)k_{out} (M)

-Homophily: reference models underestimate diagonal terms



Rodríguez, Arola-Fernández, arXiv (2024)

National homophily *

- 74% of links connected athletes from the same country

- Considering rewiring null model

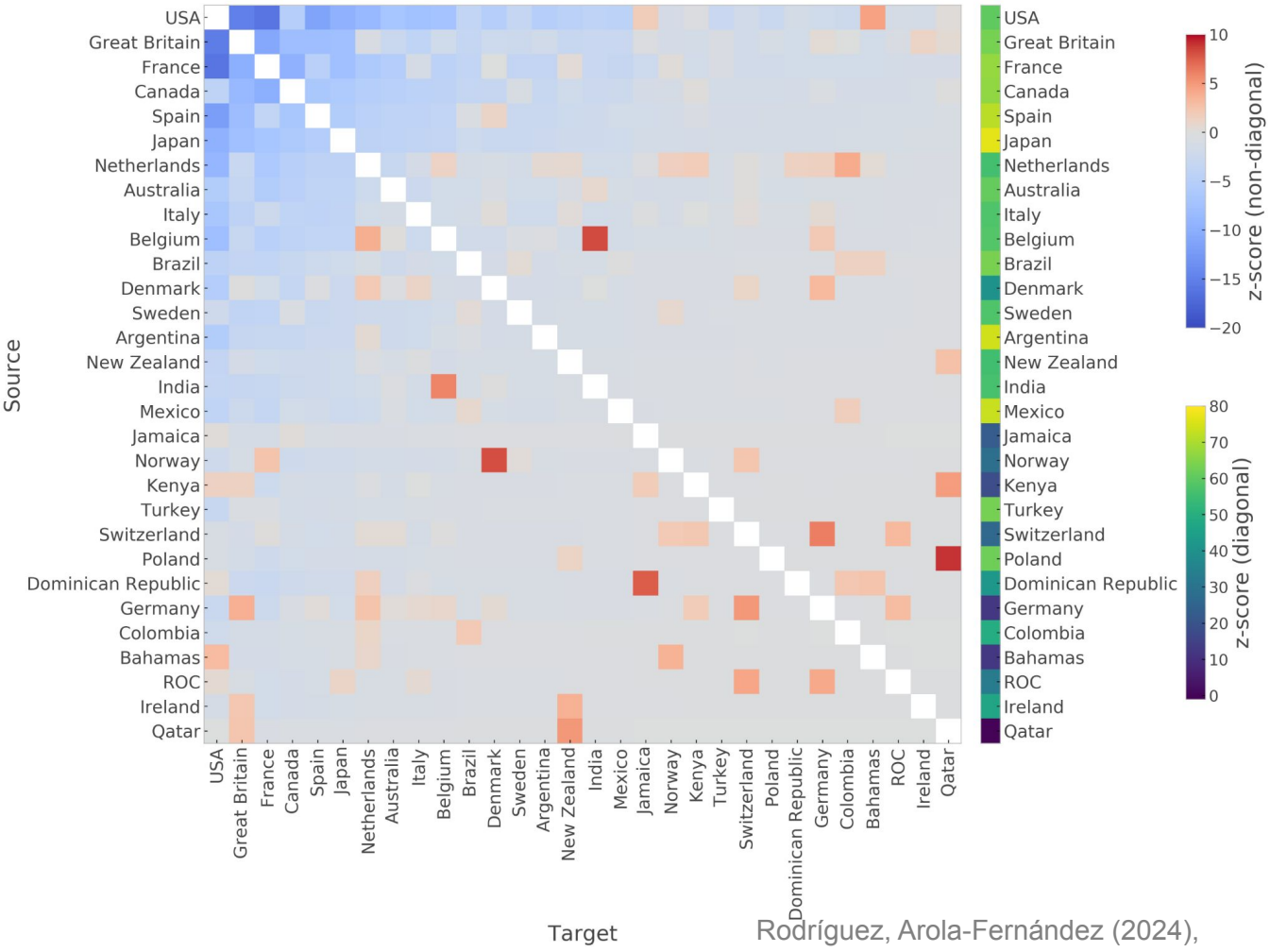
- Adjacency matrix by countries (top 30 with highest k_{in})

- Compare the observed links versus the rewiring with a *z*-score

$$z_{ij} = \frac{w_{ij} - \langle w_{ij}^r \rangle}{\sigma(\langle w_{ij}^r \rangle)}$$

- Homophily: diagonal and off-diagonal entries had different scales

- Diagonal terms: all had a z_{ii}>10 except for Qatar



Rodríguez, Arola-Fernández (2024), arXiv

AIM: to model connections, especially the impact of features (sport, country, sex) on them

1) Create homogeneous groups (same features “genotype”)
2) Extract connections among homogeneous groups *i* and *j*: E_{ij}, with mass m_i and m_j
3) Regression of a gravity-like model. Directed network may not be symmetric, so we include exponents for source and target mass:

$$E_{ij} = \frac{m_i^{\alpha} m_j^{\beta}}{f(d_{ij})}$$

4) The distance depended on linear combinations of similar/different features across groups:

$$d_{ij} = w_{\mathrm{sx}} d_{ij}^{\mathrm{sx}} + w_{\mathrm{c}} d_{ij}^{\mathrm{c}} + w_{\mathrm{sp}} d_{ij}^{\mathrm{sp}}$$

5) Regression of 5 parameters (can reduce to 4 considering *E*), unknown functional form of *f*

$$E_{ij} = \frac{m_i^{\alpha} m_j^{\beta}}{f(d_{ij})}$$

$$E_{ij} = \frac{m_i^{\alpha} m_j^{\beta}}{f(d_{ij})}$$

Consider sets of interactions with the same distance

$$\log E_{ij} = \alpha \log m_i + \beta \log m_j + K$$

For each set, obtain the estimations of α and β

Select the α and β values with the highest r^2 in this linear regression

We obtained **α = 0.50** and **β = 1.07**

Larger groups tend to have less external interactions

$$E_{ij} = \frac{m_i^{\alpha} m_j^{\beta}}{f(d_{ij})}$$

- We use the obtained values of α = 0.50 and β = 1.07
- First, we consider *f* as a power of *d*. The correlation of the regression increased with the exponent of *d*
- Thus, we used an exponential function $f(d_{ij}) = \exp(K' d_{ij})$

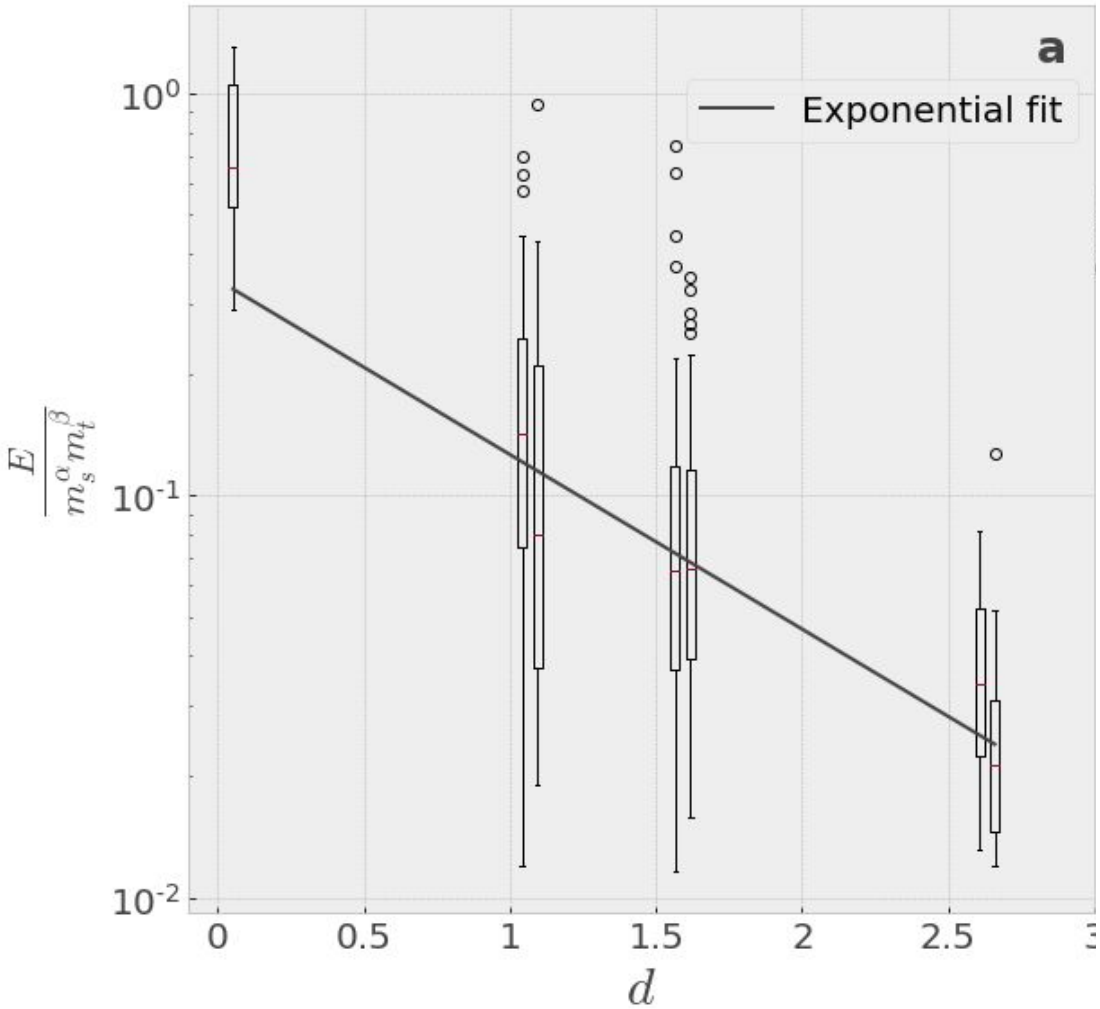
$$\log \frac{m_i^{\alpha} m_j^{\beta}}{E_{ij}} = w_{\mathrm{sx}} d_{ij}^{\mathrm{sx}} + w_{\mathrm{c}} d_{ij}^{\mathrm{c}} + w_{\mathrm{sp}} d_{ij}^{\mathrm{sp}} + K'$$

- We obtained $\boldsymbol{w_{sx} = 0.053}$, $\boldsymbol{w_c = 1.04}$, $\boldsymbol{w_{sp} = 1.57}$
- **Robustness test**: repeat the regression removing iteratively each “genotype”, obtaining:

$$w_{sx} = 0.07 \pm 0.13,\ w_c = 1.08 \pm 0.19,\ w_{sp} = 1.58 \pm 0.16$$

7 possible distances (3 features, 2^3-1)

Not big differences when the sex impacts the distance



Model behaviour: returning to the network

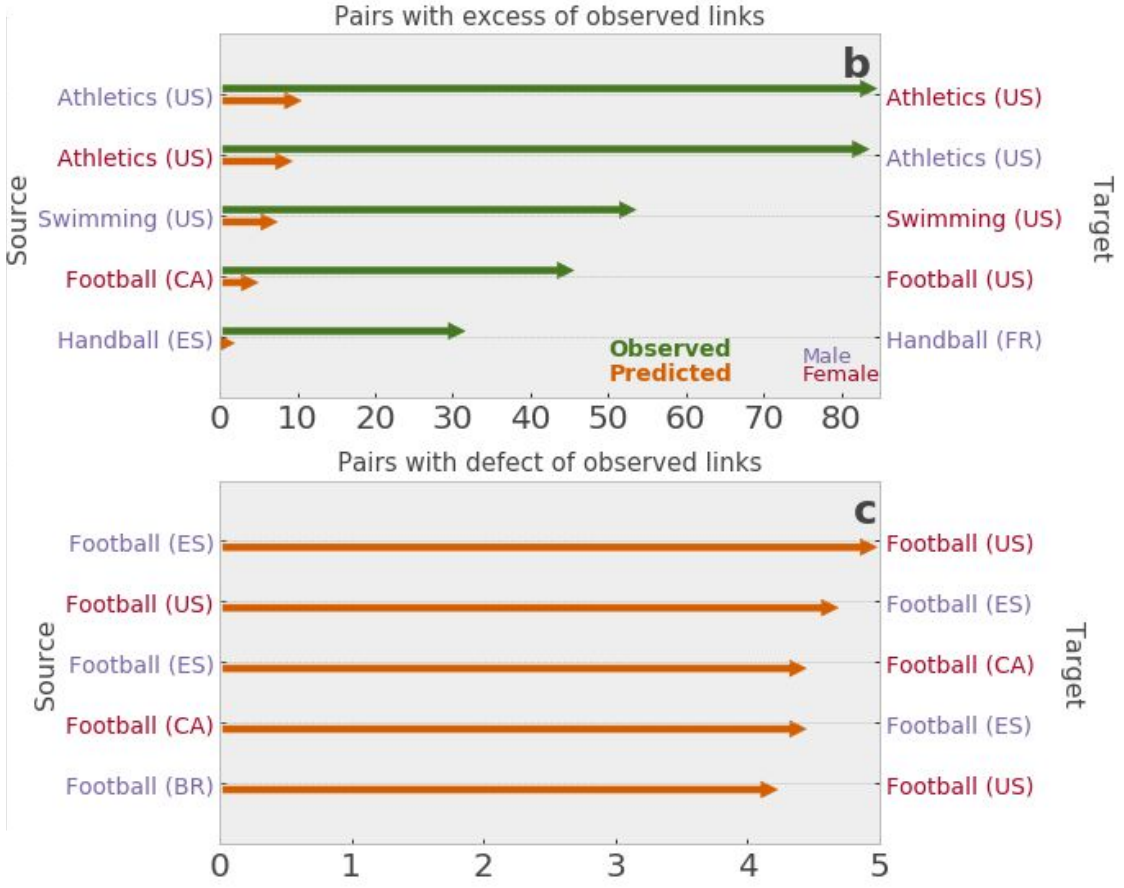
Excess and defect of links

Run multiple realizations of the model

Obtain *p*-values for each link: if we predict the same or more links than observed, increase the *p*-value

Excess of observed links: low *p*-value, genotypes with high masses (US athletics and swimming, team sports)

Defect of observed links: high *p*-value (football between men and women, different countries due to non-overlap in awarded teams)



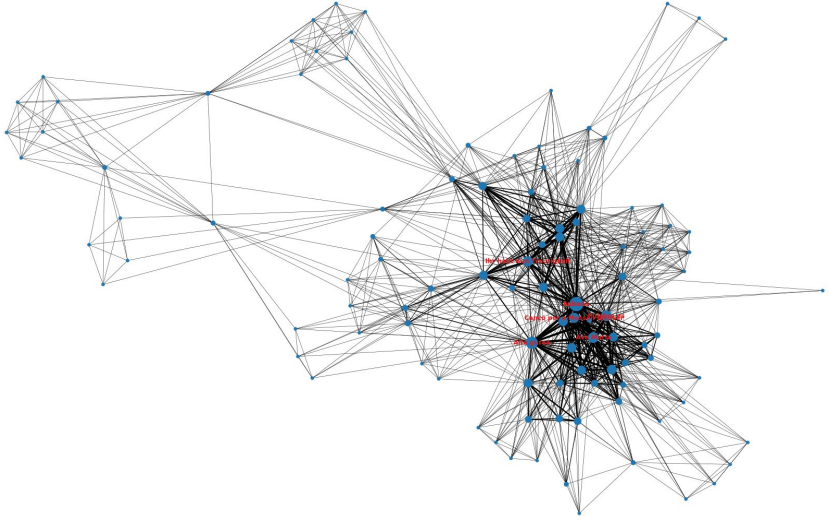
- **Sport** data is unique opportunity for **large-scale data analysis**. High diversity of available data: movement tracking, performance, records, social behaviour…
IFISC has the expertise for such analyses: random walks (with first passage times for records), complex networks, coupling dynamics, modelling approaches, AI and a large etcetera.

- The connections of high performance Olympic athletes in Twitter displayed **homophilic patterns**

- Our gravity model approach discarded sex as a **driver of interactions** across homogeneous groups, highlighting the influence of **sport and country**

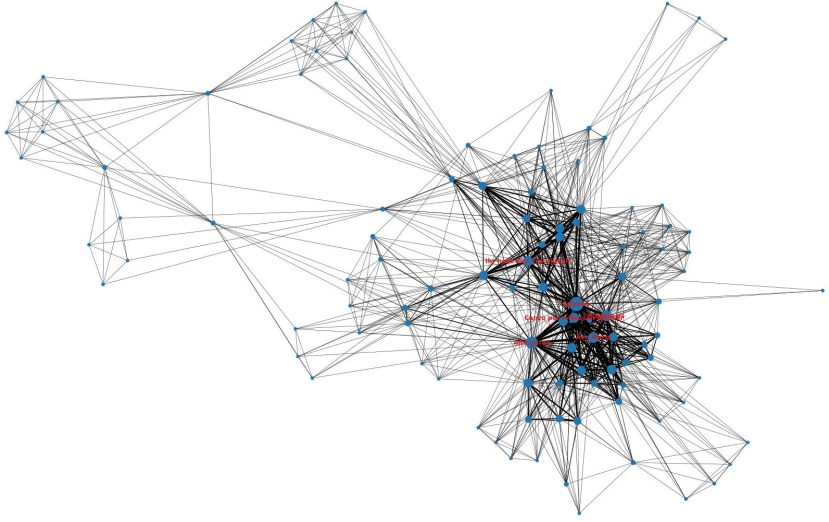
Co-programmed pieces in **choral music**
Coral Universitat de les Illes Balears
(2014-2023).

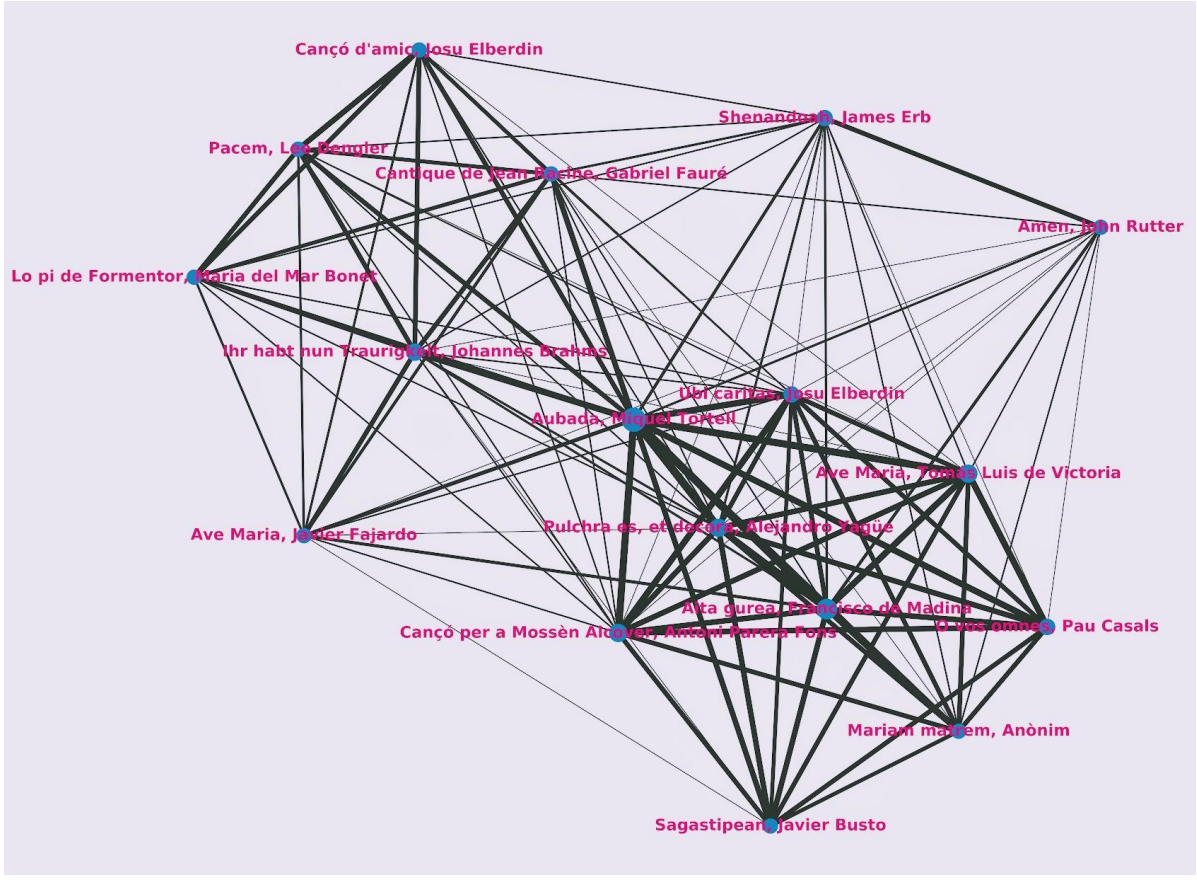
34 concerts
115 pieces
78 composers



Co-programmed pieces in **choral music**
Coral Universitat de les Illes Balears
(2014-2023).

34 concerts
115 pieces
78 composers





Lluís Arola-Fernández

arXiv: 2405.10798 (2024)

jorgep.rodriguez

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

IFISC

Govern de les Illes Balears
Conselleria d'Economia, Hisenda i Innovació

THANK YOU

Gracias

Gràcies

謝謝

for your attention

Grazie

Danke