A network approach to investigate the aggregation phenomena in sports

This is the author's manuscript

Original Citation:

Availability:
This version is available http://hdl.handle.net/2318/90622 since

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
A network approach to investigate the aggregation phenomena in sports

Luca Ferreri$^{1,2}$  Fabio Daolio$^4$  Marco Ivaldi$^3$
Mario Giacobini$^{1,2}$  Marco Tomassini$^4$

$^1$Complex System Unit, Molecular Biotechnology Center
$^2$Department of Animal Production, Epidemiology and Ecology, Faculty of Veterinary Medicine
$^3$Motor Science Research Center, S.U.I.S.M
University of Torino, Italy
$^4$Faculty of Business and Economics, Department of Information Systems
University of Lausanne, Switzerland

CS-SPORTS
Paris 12th August 2011
Network: definition

A network is given by a set of nodes
Network: definition

A network is given by a set of nodes and of interactions among nodes called edges.
Weighted Networks
Directed Networks
Complex Networks: Nodes Centralities

- **degree-centrality**: the importance of a node grows proportionally with its degree
Complex Networks: Nodes Centralities

- **degree-centrality**: the importance of a node grows proportionally with its degree
- **betweenness-centrality**: the importance of a node given by the number of paths of minimum length that cross the node


Complex Networks: Nodes Centralities

- **degree-centrality**: the importance of a node grows proportionally with its degree
- **betweenness-centrality**: the importance of a node given by the number of paths of minimum length that cross the node
- **eigenvector-centrality**: the importance of a node is proportional to the sum of the importance of all vertices that point to it (Newman 2003):
Complex Networks: Communities Detection

A community is defined as a subnet having few number of edges departing from it.
Complex Networks: Distance among Nodes

Distance among nodes is defined as the minimum number of edges necessary to connect two nodes.

The shortest path in network is called the **radius** of the network while the longest is the **diameter**. In real world network it has been observed the **small-world** phenomena: a small diameter compared with the number of nodes.
Complex Networks: Assortativity

We try to answer the question whether nodes prefer to connect with their similar (assortative behaviour) or not (dissasortative). In particular for node similarity we intend degree similarity.
Complex Networks: Assortativity

We try to answer the question whether nodes prefer to connect with their similar (assortative behaviour) or not (dissasortative). In particular for node similarity we intend degree similarity. Two approaches are known to detect the assortativity:
Complex Networks: Assortativity

We try to answer the question whether nodes prefer to connect with their similar (assortative behaviour) or not (dissasortative). In particular for node similarity we intend degree similarity. Two approaches are known to detect the assortativity:

• the study of the Pearson assortative coefficient, that detect the correlation among nodes;
• the study of the average degree of the nearest neighbors
Complex Networks: Assortativity

We try to answer the question whether nodes prefer to connect with their similar (assortative behaviour) or not (disassortative). In particular for node similarity we intend degree similarity. Two approaches are known to detect the assortativity:

- the study of the Pearson assortative coefficient, that detect the correlation among nodes;
- the study of the average degree of the nearest neighbors

\[ \langle k_{\text{n}} \rangle \]
We try to answer the question whether nodes prefer to connect with their similar (assortative behavior) or not (dissassortative). In particular for node similarity we intend degree similarity. Two approaches are known to detect the assortativity:

- the study of the Pearson assortative coefficient, that detect the correlation among nodes;
- the study of the average degree of the nearest neighbors
Complex Networks: Clustering Coefficient

The clustering coefficient of a node is a measure of how its neighbors are connected.
Complex Networks: Heterogenity

The degree distribution of a network is the probability for a node to have a number of edges departing from it. Complex networks could be distinct in:

- regular, having homogeneous degree distribution such as fixed, binomial, Poisson, exponential or normal;
- scale-free, having fat tailed degree distribution well described by power law distribution (at least asymptotically).
Complex Networks: Heterogeneity

The degree distribution of a network is the probability for a node to have a number of edges departing from it. Complex networks could be distinct in:

- **regular** having homogeneous degree distribution such as fixed, binomial, Poisson, exponential or normal;
- **scale-free** having fat tailed degree distribution well described by power law distribution (at least asymptotically)
Complex Networks: Heterogeneity

The degree distribution of a network is the probability for a node to have a number of edges departing from it. Complex networks could be distinct in:

- **Regular** having homogeneous degree distribution such as fixed, binomial, Poisson, exponential or normal;
- **Scale-free** having fat tailed degree distribution well described by power law distribution (at least asymptotically);

A measure of the heterogeneity is given by:

\[
\frac{\langle k \rangle}{\langle k^2 \rangle}
\]
Complex Networks: Heterogeneity

The degree distribution of a network is the probability for a node to have a number of edges departing from it. Complex networks could be distinct in:

- **regular** having homogeneous degree distribution such as fixed, binomial, Poisson, exponential or normal;
- **scale-free** having fat tailed degree distribution well described by power law distribution (at least asymptotically);

\[
P(X = h) = C h^{-\gamma}
\]

\[
\ln P(X = h) = \ln C - \gamma \ln h
\]
Bipartite Networks

Many example:

- co-authorship network;
- diseasome;
- heterosexual contact network;
- vector-borne disease network;
Bipartite Networks

![Diagram of Bipartite Network](image)
Bipartite Networks
Bipartite Networks
Bipartite Networks

The $A$-projection

the $B$-projection
Bipartite Projection is less informative
Bipartite Projection is less informative
Bipartite Projection is less informative

are both projected in
We-Sport: a sparse network

We consider a snapshot of the entire network:
- 1680 athletes
- 240 sports
- 6107 interactions

We define the density for a bipartite network as:
$$\delta = \frac{\text{edges}}{\text{sports} \cdot \text{athletes}} \approx 0.014$$

We observe only two large connected components:
- The first has 1679 athletes and 239 sports
- The second...
We-Sport: a sparse network

We consider a snapshot of the entire network:

- 1680 athletes
- 240 sports
We consider a snapshot of the entire network:
- 1680 athletes
- 240 sports
- 6107 interactions
We-Sport: a sparse network

We consider a snapshot of the entire network:

• 1680 athletes
• 240 sports
• 6107 interactions

we define the density for bipartite network as:

$$\delta = \frac{\text{edges}}{\text{sports} \cdot \text{athletes}} \approx 0.014$$
We-Sport: a sparse network

We consider a snapshot of the entire network:

- 1680 athletes
- 240 sports
- 6107 interactions

We define the density for bipartite network as:

\[
\delta = \frac{\text{edges}}{\text{sports} \cdot \text{athletes}} \approx 0.014
\]

We observe only two large connected components:

- the first have 1679 athletes and 239 sports
We-Sport: a sparse network

We consider a snapshot of the entire network:

- 1680 athletes
- 240 sports
- 6107 interactions

We define the density for a bipartite network as:

\[ \delta = \frac{\text{edges}}{\text{sports} \cdot \text{athletes}} \approx 0.014 \]

We observe only two large connected components:

- The first have 1679 athletes and 239 sports
- The second

![Graph](attachment:graph.png)
Graphical representation of bipartite We-Sport network
Graphical representation of bipartite We-Sport network
The 70 most played Sports

female
male
The 70 most played Sports: gender frequencies
A complex network

<table>
<thead>
<tr>
<th>partition</th>
<th>mode</th>
<th>median</th>
<th>$\langle k \rangle$</th>
<th>$\langle k^2 \rangle$</th>
<th>$\frac{\langle k \rangle}{\langle k^2 \rangle}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sport</td>
<td>1</td>
<td>5</td>
<td>25.54</td>
<td>$6.183 \cdot 10^3$</td>
<td>0.0041</td>
</tr>
<tr>
<td>athletes</td>
<td>1</td>
<td>3</td>
<td>3.63</td>
<td>23.78</td>
<td>0.1530</td>
</tr>
</tbody>
</table>
Degree distribution: sport nodes

\[ P(x = h) \]

\[ h \]

1 2 3 4 5 6

1 2 3 4 5 6 7 8

17 of 1
Degree distribution: sport nodes

$P(x \geq h)$

$h$

$P(x \geq h)$ vs $h$
Degree distribution: sport nodes logarithmic scale
Degree distribution: sport nodes logarithmic scale

$P(x \geq h)$

$\alpha = 1.96$

$x_{\min} = 15$

$p$-value $= 0.5670$
Degree distribution: sport nodes logarithmic scale

\[ P(x \geq h) \]

\[ h \]

\[ \alpha = 1.96 \]

\[ x_{\text{min}} = 15 \]

\[ p\text{-value} = 0.5670 \]

<table>
<thead>
<tr>
<th></th>
<th>bin.neg. LR</th>
<th>Poisson LR</th>
<th>exponential LR</th>
<th>Weibull LR</th>
<th>log-normal LR</th>
<th>Yule LR</th>
<th>power law + cutoff LR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.14 0.00</td>
<td>4.09 0.00</td>
<td>4.31 0.00</td>
<td>0.13 0.89</td>
<td>-0.35 0.72</td>
<td>-0.004 0.94</td>
<td>-0.53 0.30</td>
</tr>
</tbody>
</table>
Degree distribution: athletes nodes

\[ P(x = h) \]

\( h \)
Degree distribution: athletes nodes
Degree distribution: athletes nodes logarithmic scale

$\alpha = 3.49$

$p$-value $= 0.0730$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>LR $p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>0.03</td>
</tr>
<tr>
<td>Poisson</td>
<td>0.00</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.93</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.00</td>
</tr>
<tr>
<td>Log-normal</td>
<td>0.63</td>
</tr>
<tr>
<td>Yule</td>
<td>0.00</td>
</tr>
<tr>
<td>Power law + cutoff</td>
<td>0.00</td>
</tr>
</tbody>
</table>

$\lambda_{min} = 6$
Degree distribution: athletes nodes logarithmic scale

\[ p(x \geq h) \]

\[ h \]

\[ 10^0 \quad 10^1 \quad 10^2 \]

\[ 10^{-4} \quad 10^{-3} \quad 10^{-2} \quad 10^{-1} \quad 10^0 \]

\[ \alpha = 3.49 \]

\[ x_{\text{min}} = 6 \]

\[ p\text{-value} = 0.0730 \]
Degree distribution: athletes nodes logarithmic scale

\[ p(x \geq h) \]

\[ h \]

\[ 10^0 \quad 10^1 \quad 10^2 \]

\[ 10^0 \quad 10^{-1} \quad 10^{-2} \quad 10^{-3} \quad 10^{-4} \]

\[ \alpha = 3.49 \]

\[ x_{\text{min}} = 6 \]

\[ p\text{-value} = 0.0730 \]

<table>
<thead>
<tr>
<th>Model</th>
<th>athletes - nodes</th>
<th>bin.neg.</th>
<th>Poisson</th>
<th>exponential</th>
<th>Weibull</th>
<th>log-normal</th>
<th>Yule</th>
<th>power law + cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>3.73</td>
<td>0.00</td>
<td>3.09</td>
<td>0.00</td>
<td>-0.83</td>
<td>0.40</td>
<td>-2.16</td>
<td>0.03</td>
</tr>
<tr>
<td>p</td>
<td>-2.12</td>
<td>0.03</td>
<td>-3.94</td>
<td>0.00</td>
<td>-6.62</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The maximum distance between every pairs of nodes in a graph is defined as the **diameter** of the graph. We observe a diameter of 8 but on average the shortest path between nodes is 3.33. We are in presence of so called **small-world** phenomena.
The distance distribution

the athletes-athletes distance

![Graph showing the distribution of athletes-athletes distance. The graph has bars for distances of 2 and 4, with corresponding heights of 1/3 and 2/3.]
The distance distribution

the sport-sport distance
The distance distribution

the sport-athletes distance
Assortativity in bipartite

We-sport network shows a disassortative behaviour: the Pearson coefficient is $-0.2425$. Moreover if we calculate the nearest neighbor degree:
Assortativity in bipartite

We-sport network shows a disassortative behaviour: the Pearson coefficient is $-0.2425$.
Moreover if we calculate the nearest neighbor degree:
The joint probability
2-length assortativity in bipartite

We want to try to answer the question:

*do people choose sports that connect them with similar people or not?*

Therefore we analyze the 2-length assortativity: we observe a weak assortative behavior for athletes-nodes (0.0326) and stronger for sport-nodes (0.2620)
2-length assortativity in bipartite

![](image)
2-length assortativity in bipartite

$k$-athletes

$k_{n+2}-l$-sport
The Clustering Coefficient for Biparite Networks

Again in order to understand the aggregation behavior of athletes we try to understand if people prefer to connect with other sharing the same sport’s preference. Hence we define a similarity matrix $cc$ which counts for each couple of athletes the number of sports they share:

$$|N(v) \cap N(u)|$$

then we can normalize that matrix. Le Blond et al., Latapy et al., and Borgatti suggest the three following denominator:

- $\min(|N(v)|, |N(u)|)$ for $cc\bullet(u, v)$
The Clustering Coefficient for Biparite Networks

Again in order to understand the aggregation behavior of athletes we try to understand if people prefer to connect with other sharing the same sport’s preference. Hence we define a similarity matrix \( cc \) which counts for each couple of athletes the number of sports they share:

\[
|N(v) \cap N(u)|
\]

then we can normalize that matrix. Le Blond et al., Latapy et al., and Borgatti suggest the three following denominator:

- \( \min(|N(v)|, |N(u)|) \) for \( cc_\bullet(u, v) \)
- \( \max(|N(v)|, |N(u)|) \) for \( cc_\circ(u, v) \)
The Clustering Coefficient for Biparite Networks

Again in order to understand the aggregation behavior of athletes we try to understand if people prefer to connect with other sharing the same sport’s preference. Hence we define a similarity matrix $cc$ which counts for each couple of athletes the number of sports they share:

$$|N(v) \cap N(u)|$$

then we can normalize that matrix. Le Blond et al., Latapy et al., and Borgatti suggest the three following denominator:

- $\min (|N(v)|, |N(u)|)$ for $cc_\bullet(u, v)$
- $\max (|N(v)|, |N(u)|)$ for $cc_\bullet(u, v)$
- $|N(v) \cap N(u)|$ for $cc_\bullet(u, v)$
The clustering coefficient II

From the similarity matrix we can calculate the clustering coefficient of each node:

\[ cc(v) = \frac{\sum_{u \in N(N(v))} cc(v, u)}{|N(N(v))|} \]

and from that the clustering coefficient of A-partition:

\[ cc = \frac{1}{|A|} \sum_{v \in A} cc(v) \]

<table>
<thead>
<tr>
<th>graph</th>
<th>\text{cc}_c</th>
<th>\text{cc}_c^-</th>
<th>\text{cc}_c^-</th>
</tr>
</thead>
<tbody>
<tr>
<td>athletes</td>
<td>0.6628</td>
<td>0.2672</td>
<td>0.2315</td>
</tr>
<tr>
<td>sport</td>
<td>0.4126</td>
<td>0.0615</td>
<td>0.0536</td>
</tr>
</tbody>
</table>
The clustering coefficient II

The athletes case

\[ \langle cc \rangle \]
The clustering coefficient II

The sport case

\[ \langle cc \rangle \]

\[ k \]

\( CC \)

\( CC_{\cdot} \)

\( CC_{\cdot} \)
The centrality

2-mode Key Sports Analysis
Betweenness Centrality
Eigenvector Centrality
An application: which is the best sport to meet girls?
Contacts

for further informations:

www.we-sport.com

or contact us:

luca.ferreri@unito.it
fabio.daollio@unil.ch