What makes a good (fantasy) football manager?
webscraping and data analysis in R

Joseph O'Brien

4-Feb-2021 | Why R? Webinar

UNIVERSITY of LIMERICK
OLLSCOIL LUIMNIGH

joseph.obrien@ul.ie
@obrienj

sfi
Fondúireacht Eolaíochta Éireann
Dá bhfuil romhainn
Science Foundation Ireland
For what's next

About Me

- PhD student at the University of Limerick
- Generally interested in understanding complex systems via mathematical modelling + data science
- Avid R user with sides of Python, Julia, and MATLAB
- **webpage:** joeyobrien.ie | **twitter:** @obrienj_ | **github:** obrienjoey

About Me

- PhD student at the University of Limerick
- Generally interested in understanding complex systems via mathematical modelling + data science
- Avid R user with sides of Python, Julia, and MATLAB
- **webpage:** joeyobrien.ie | **twitter:** @obrienj_ | **github:** obrienjoey

About Me

- PhD student at the University of Limerick
- Generally interested in understanding complex systems via mathematical modelling + data science
- Avid R user with sides of Python, Julia, and MATLAB
- **webpage:** joeyobrien.ie | **twitter:** @obrienj_ | **github:** obrienjoey

About Me

- PhD student at the University of Limerick
- Generally interested in understanding complex systems via mathematical modelling + data science
- Avid R user with sides of Python, Julia, and MATLAB
- **webpage:** joeyobrien.ie | **twitter:** @obrienj_ | **github:** obrienjoey

Outline of Today's Talk

- What is webscraping and why should we learn it?
- Key packages in R
- Demo of how it can be done
- Example project – Fantasy Premier League

Outline of Today's Talk

- What is webscraping and why should we learn it?
- Key packages in R
- Demo of how it can be done
- Example project – Fantasy Premier League

Outline of Today's Talk

- What is webscraping and why should we learn it?
- Key packages in R
- Demo of how it can be done
- Example project – Fantasy Premier League

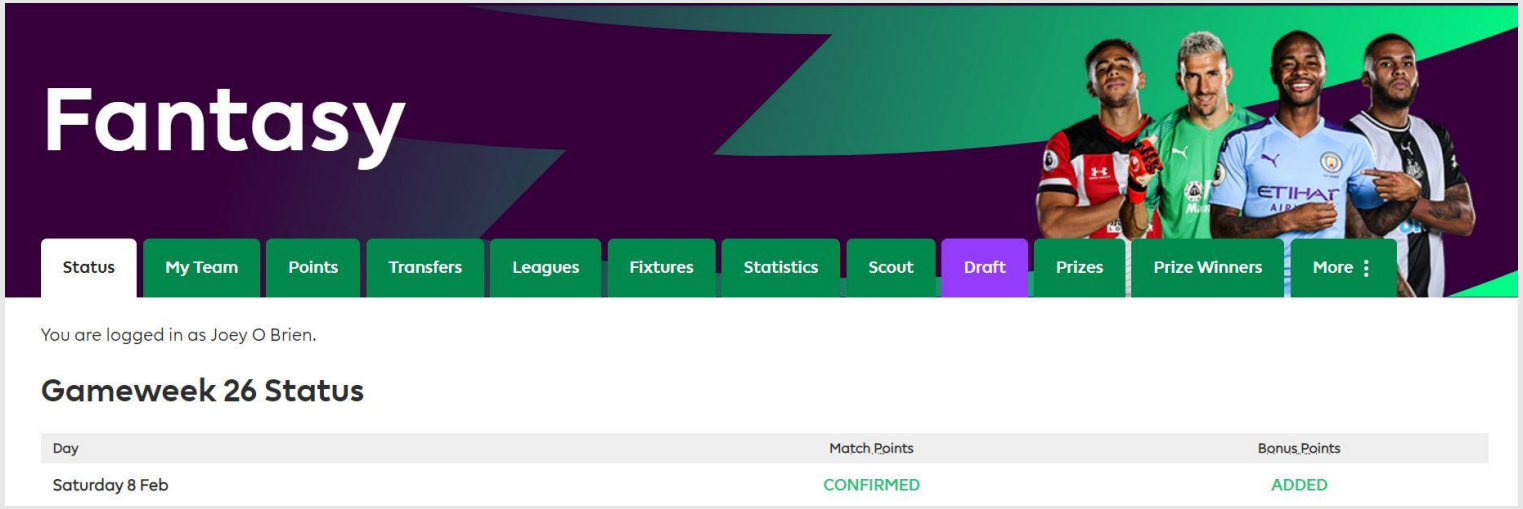
Outline of Today's Talk

- What is webscraping and why should we learn it?
- Key packages in R
- Demo of how it can be done
- Example project – Fantasy Premier League

What is Webscraping All About?

- The web is essentially a repository for data.

- In order to store and display that data the websites have to be coded - generally in **HTML** or **Javascript**.



What is Webscraping All About?

- The web is essentially a repository for data.

- In order to store and display that data the websites have to be coded - generally in **HTML** or **Javascript**.

```
<!doctype html><html lang="en"><head><meta charset="utf-8"/><meta http-equiv="x-ua-compatible" content="ie=edge"/><meta name="viewport" content="width=device-
width,initial-scale=1"/><title>Fantasy Premier League, Official Fantasy Football Game of the Premier League</title><meta name="description" content="Official Fantasy
Premier League 2019/20. Free to play fantasy football game, set up your fantasy football team at the Official Premier League site." data-react-helmet="true"/><link
rel="apple-touch-icon" sizes="180x180" href="/img/favicons/apple-touch-icon.png"/><link rel="icon" type="image/png" href="/img/favicons/favicon-32x32.png" sizes="32x32"/>
<link rel="icon" type="image/png" href="/img/favicons/favicon-16x16.png" sizes="16x16"/><link rel="manifest" href="/img/favicons/manifest.json"/><link rel="mask-icon"
href="/img/favicons/safari-pinned-tab.svg" color="#5bbad5"/><link rel="shortcut icon" href="/img/favicons/favicon.ico"/><meta name="msapplication-config"
content="/favicons/browserconfig.xml"/><meta name="theme-color" content="#ffffff"/><meta property="og:url" content="https://fantasy.premierleague.com/"/><meta
property="og:type" content="website"/><meta property="og:title" content="Fantasy Premier League, Official Fantasy Football Game of the Premier League"/><meta
property="og:description" content="Official Fantasy Premier League 2019/20. Free to play fantasy football game, set up your fantasy football team at the Official Premier
League site."/><meta property="og:image" content="/img/share/facebook-share.png"/><meta name="twitter:card" content="summary"/><meta name="twitter:site"
content="@OfficialFPL"/><meta name="apple-itunes-app" content="app-id=1138895159"/><link rel="stylesheet" href="//www.premierleague.com/resources/prod/a83e4e7-
1591/ism/css/ism.css"/><script async src="https://platform.twitter.com/widgets.js"></script><script type="text/javascript">var googletag=googletag||
{};googletag.cmd=googletag.cmd||[],function(){var t=document.createElement("script");t.async=!0,t.type="text/javascript";var e="https:"==document.location.protocol;t.src=
(e?"https:":"http:")+"//www.googletagservices.com/tag/js/gpt.js";var o=document.getElementsByTagName("script")[0];o.parentNode.insertBefore(t,o)}()</script>
<script>!function(e,t,n,c,o,a,f){e.fbq||(o=e.fbq=function(){o.callMethod?o.callMethod.apply(o,arguments):o.queue.push(arguments)},e._fbq||(e._fbq=o),
(o.push=o).loaded=!0,o.version="2.0",o.queue=[],(a=t.createElement(n)).async=!0,a.src="https://connect.facebook.net/en_US/fbevents.js",(f=t.getElementsByTagName(n)
```

What is Webscraping All About?

- The web is essentially a repository for data.

- In order to store and display that data the websites have to be coded - generally in **HTML** or **Javascript**.

- Once we become comfortable with the syntax of these languages any data on the web is obtainable!

Why Webscrape?

- Many important publications use novel datasets to determine previously unconsidered phenomena
  - Think Stanley Milgram's research into the *six degrees of separation*
  - One of the many studies into citation dynamics

- Some fields suffer from 'Zachary Karate Club Syndrome'

Why Webscrape?

- Many important publications use novel datasets to determine previously unconsidered phenomena
  - Think Stanley Milgram's research into the *six degrees of separation*
  - One of the many studies into citation dynamics

- Some fields suffer from 'Zachary Karate Club Syndrome'

Why Webscrape?

- Many important publications use novel datasets to determine previously unconsidered phenomena
  - Think Stanley Milgram's research into the *six degrees of separation*
  - One of the many studies into citation dynamics

- Some fields suffer from 'Zachary Karate Club Syndrome'

Why Webscrape?

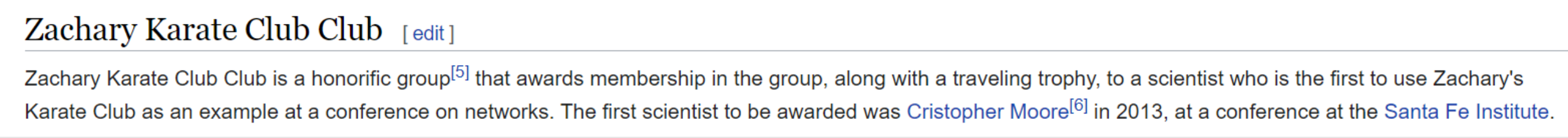
- Many important publications use novel datasets to determine previously unconsidered phenomena
  - Think Stanley Milgram's research into the *six degrees of separation*
  - One of the many studies into citation dynamics

- Some fields suffer from 'Zachary Karate Club Syndrome'

Why Webscrape?

- Many important publications use novel datasets to determine previously unconsidered phenomena
  - Think Stanley Milgram's research into the *six degrees of separation*
  - One of the many studies into citation dynamics

- Some fields suffer from 'Zachary Karate Club Syndrome'



Zachary Karate Club Club [edit]

Zachary Karate Club Club is a honorific group[5] that awards membership in the group, along with a traveling trophy, to a scientist who is the first to use Zachary's Karate Club as an example at a conference on networks. The first scientist to be awarded was Cristopher Moore[6] in 2013, at a conference at the Santa Fe Institute.

Why Webscrape?

- Many important publications use novel datasets to determine previously unconsidered phenomena
  - Think Stanley Milgram's research into the *six degrees of separation*
  - One of the many studies into citation dynamics
- Some fields suffer from 'Zachary Karate Club Syndrome'
- A chance to do *research* on something that really interests you

Key Packages

- In R there are two main packages used scrape data - **XML2** and **rvest**.

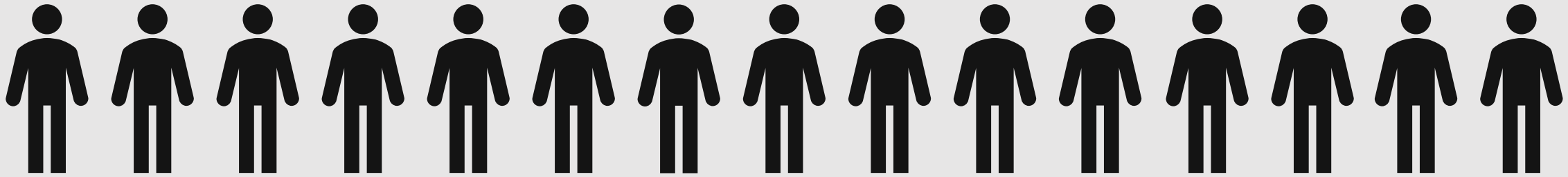
- Python equivalents – beautifulsoup4 + requests

Examples 

- Reddit data
- Wikipedia
- Fantasy Premier League

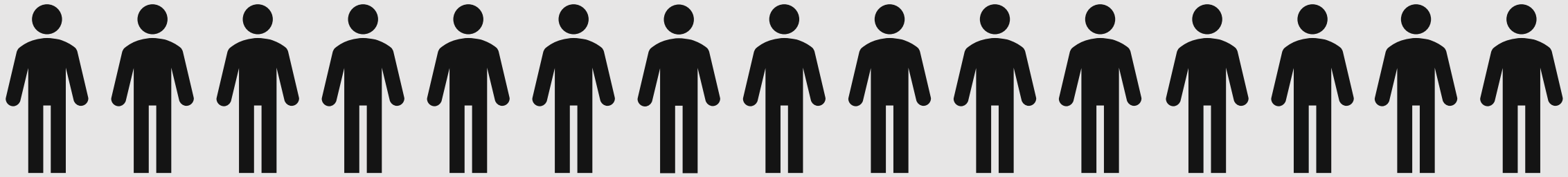
What is Fantasy Premier League?

- Online game where users are provided with a virtual £100 million budget to assemble a squad of 15 real-life Premier League footballers.
- Each week choosing 11 of them to ‘start’ for the user’s team, who then receive points based upon their statistical performance in the real game.
- Also choose one **captain** who receives double points. The challenge of the game is to **determine an optimal way of spending your budget as to maximize your points.**



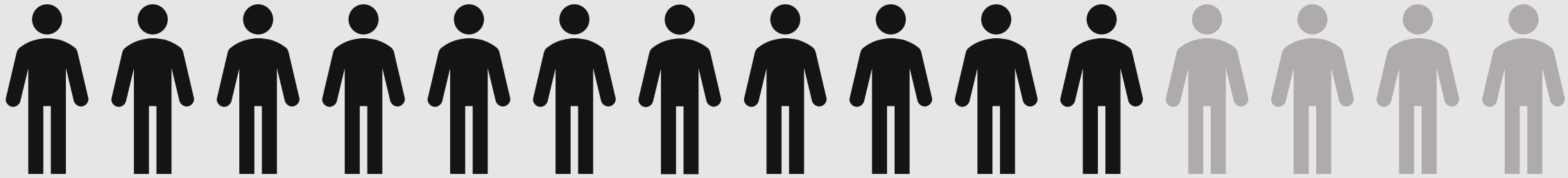
What is Fantasy Premier League?

- Online game where users are provided with a virtual £100 million budget to assemble a squad of 15 real-life Premier League footballers.
- Each week choosing 11 of them to ‘start’ for the user’s team, who then receive points based upon their statistical performance in the real game.
- Also choose one **captain** who receives double points. The challenge of the game is to **determine an optimal way of spending your budget as to maximize your points.**



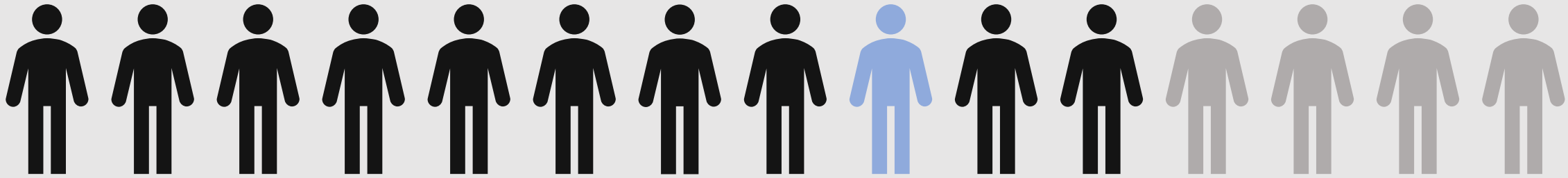
What is Fantasy Premier League?

- Online game where users are provided with a virtual £100 million budget to assemble a squad of 15 real-life Premier League footballers.
- Each week choosing 11 of them to ‘start’ for the user’s team, who then receive points based upon their statistical performance in the real game.
- Also choose one **captain** who receives double points. The challenge of the game is to **determine an optimal way of spending your budget as to maximize your points.**



What is Fantasy Premier League?

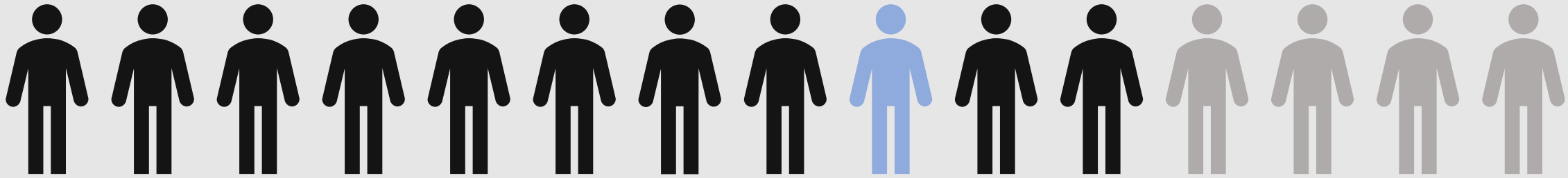
- Online game where users are provided with a virtual £100 million budget to assemble a squad of 15 real-life Premier League footballers.
- Each week choosing 11 of them to 'start' for the user's team, who then receive points based upon their statistical performance in the real game.
- Also choose one **captain** who receives double points. The challenge of the game is to **determine an optimal way of spending your budget as to maximize your points.**



@obrienj_

What is Fantasy Premier League?

- Online game where users are provided with a virtual £100 million budget to assemble a squad of 15 real-life Premier League footballers.
- Each week choosing 11 of them to ‘start’ for the user’s team, who then receive points based upon their statistical performance in the real game.
- Also choose one **captain** who receives double points. The challenge of the game is to **determine an optimal way of spending your budget as to maximize your points.**



Examples 

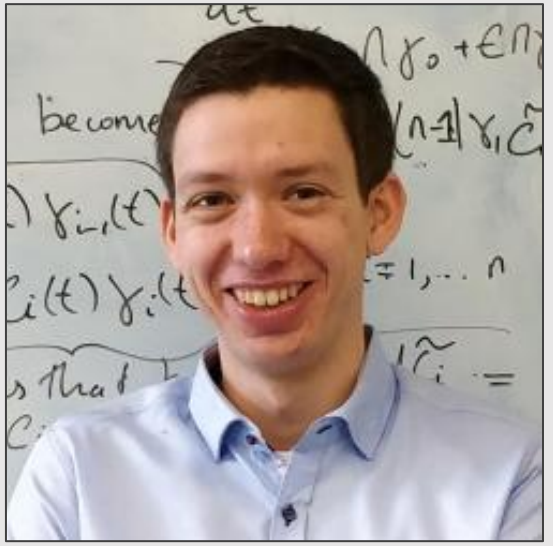
- Reddit data
- Wikipedia
- Fantasy Premier League

Example Study

James Gleeson

David O'Sullivan

University of Limerick



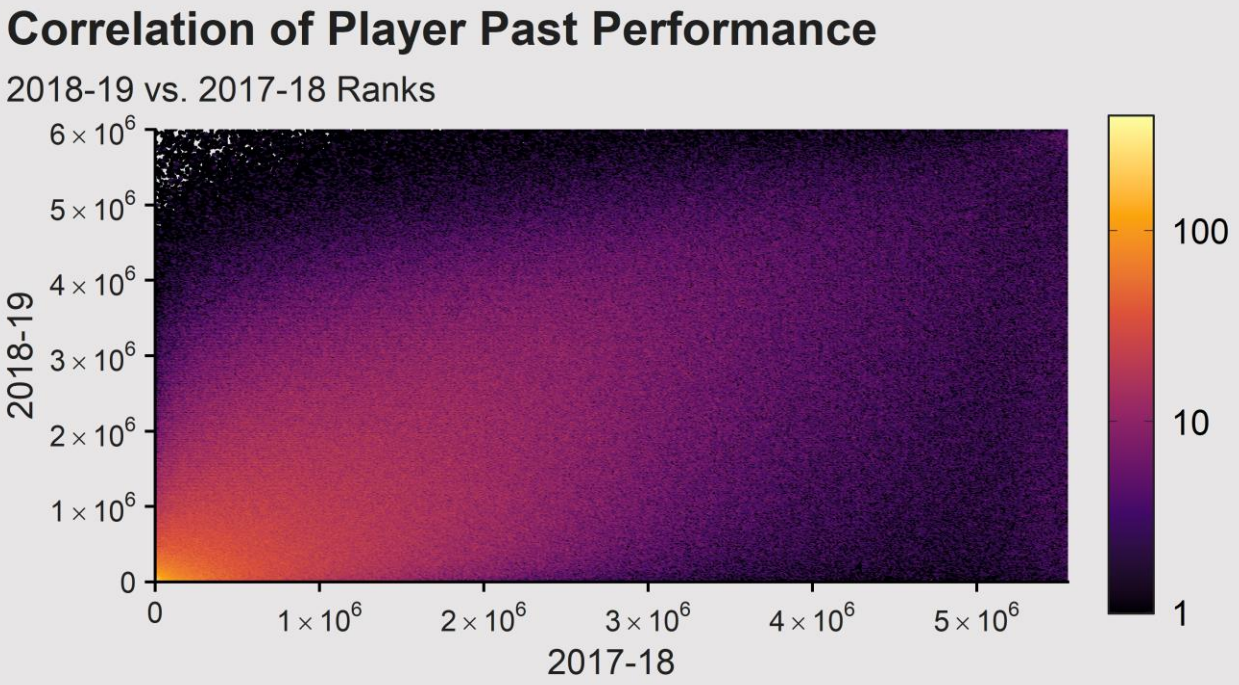
Identification of skill in an online game: The case of Fantasy Premier League

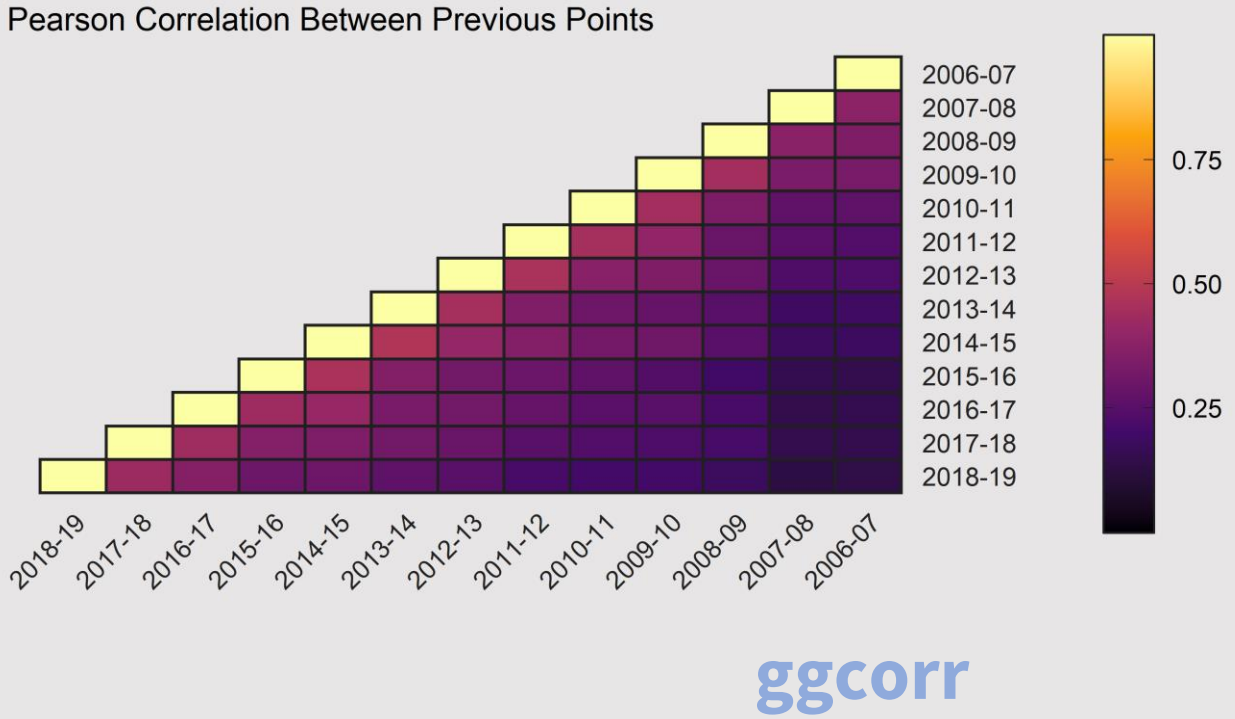
Joseph D. O'Brien, James P. Gleeson, and David J. P. O'Sullivan
MACSI, Department of Mathematics and Statistics,
University of Limerick, Limerick V94 T9PX, Ireland
(Dated: September 3, 2020)

In all competitions where results are based upon an individual's performance the question of whether the outcome is a consequence of skill or luck arises. We explore this question through an analysis of a large dataset of approximately one million contestants playing *Fantasy Premier League*, an online fantasy sport where managers choose players from the English football (soccer) league. We show that managers' ranks over multiple seasons are correlated and we analyse the actions taken

PLOS ONE (in press) 2021
arXiv: 2009.01206

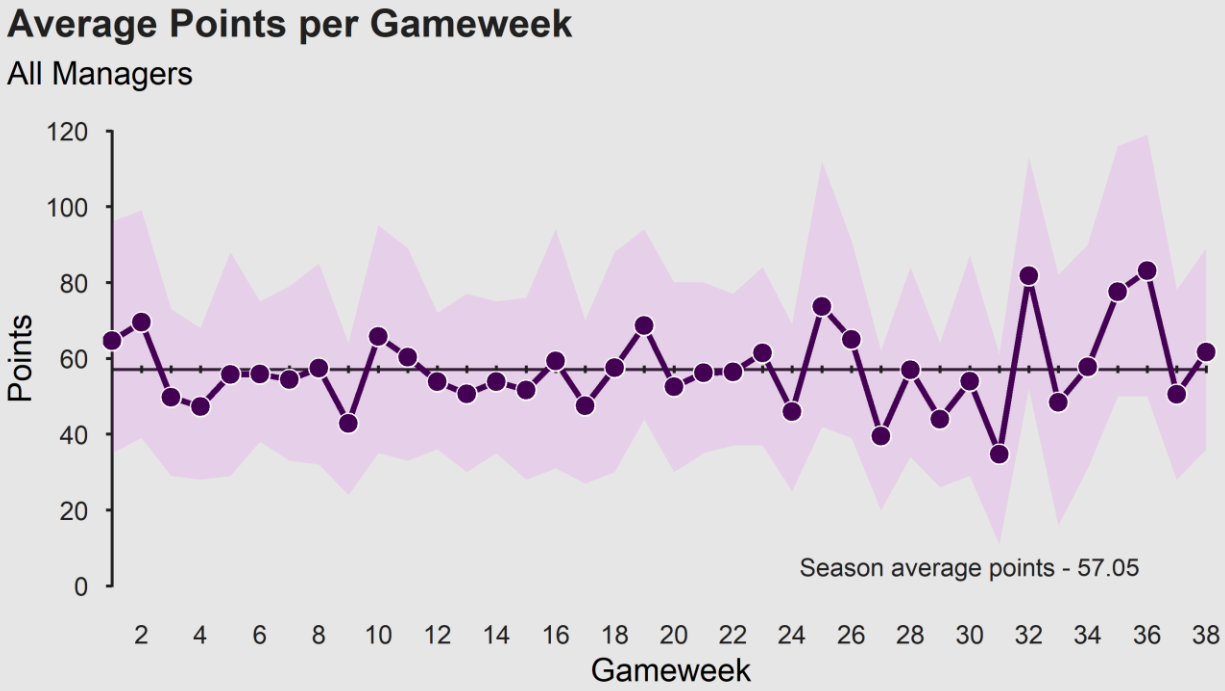
Consistency Between Performances?

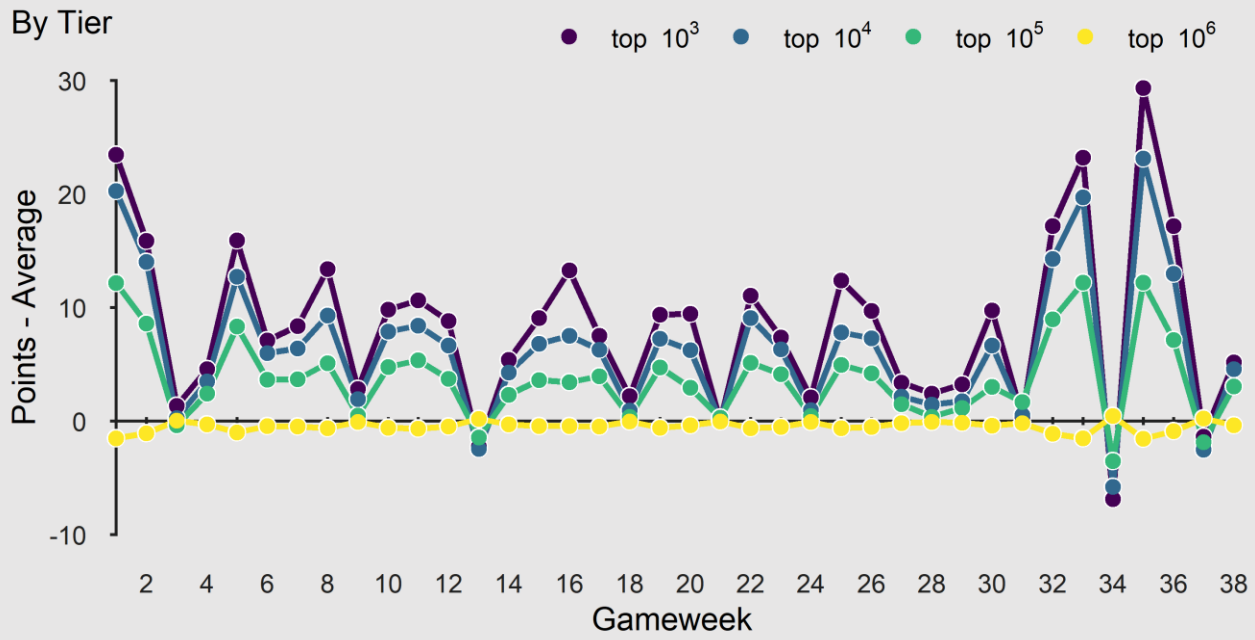




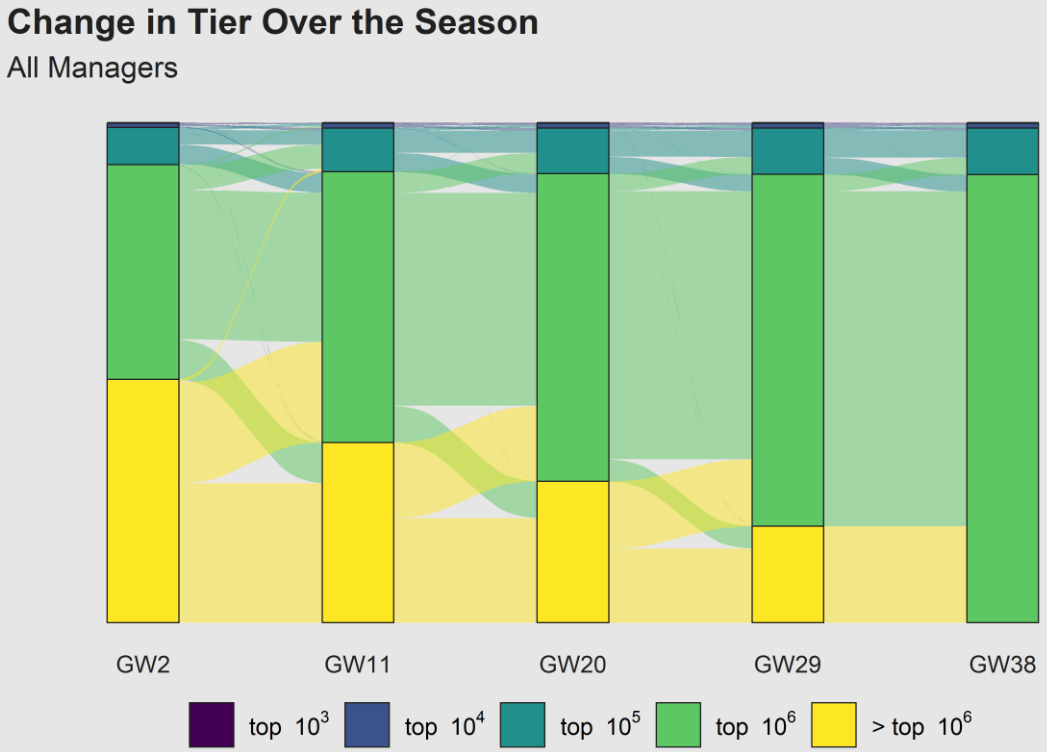
ggcorr

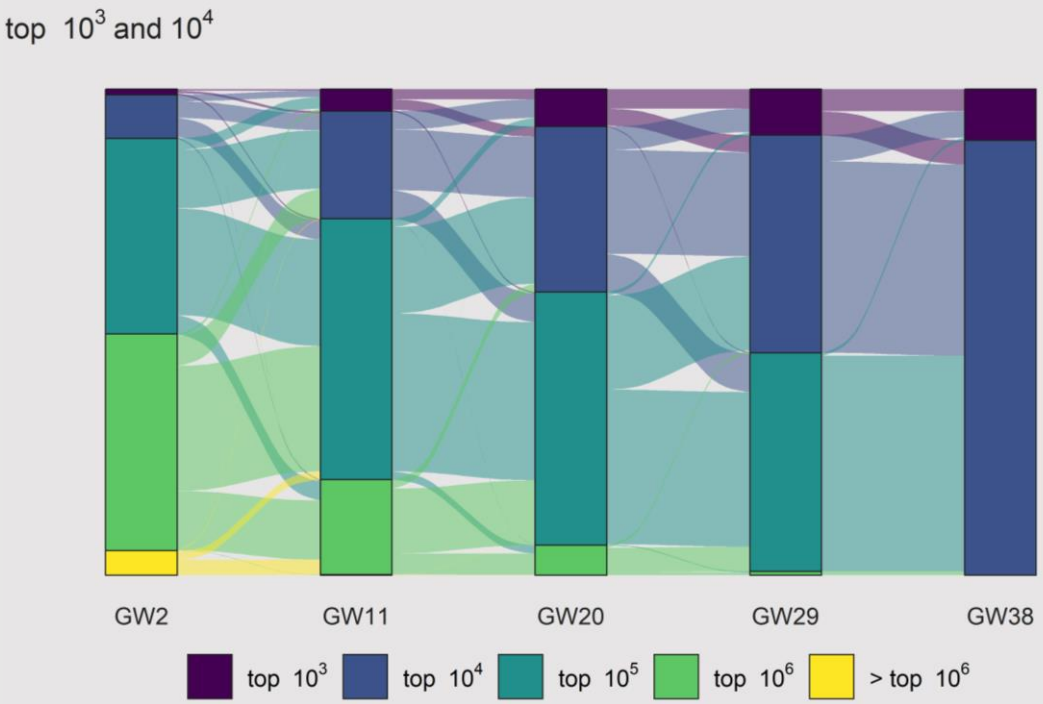
Tiers among managers





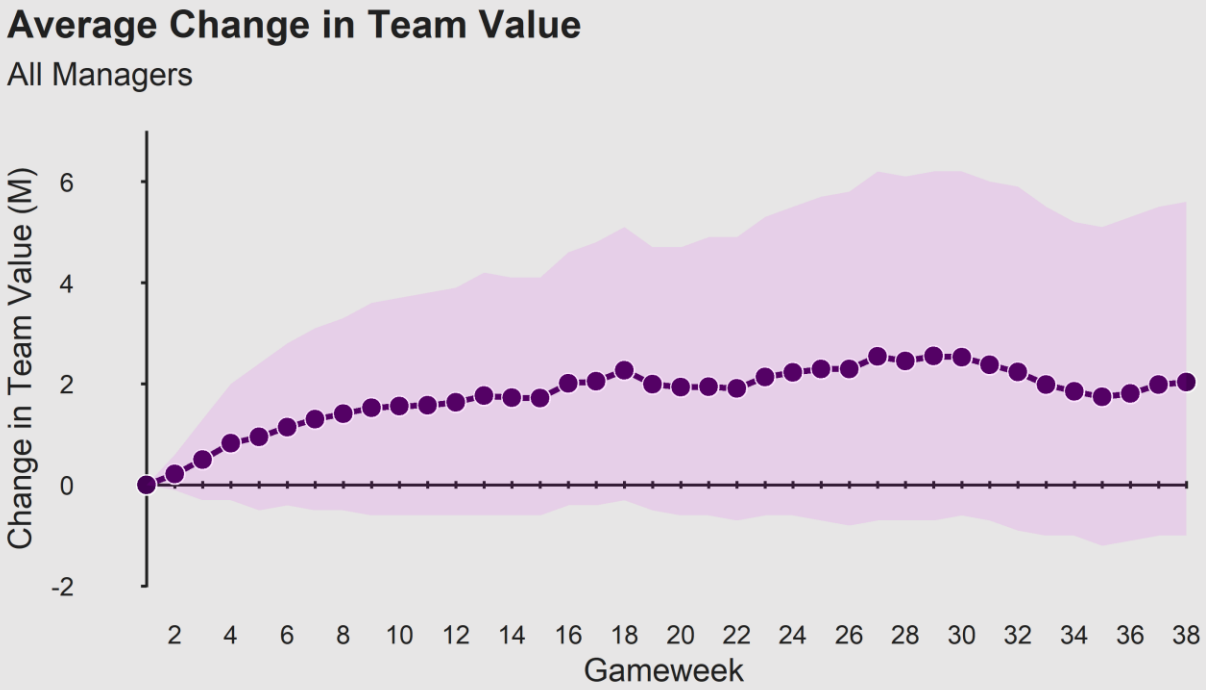
Tiers among managers

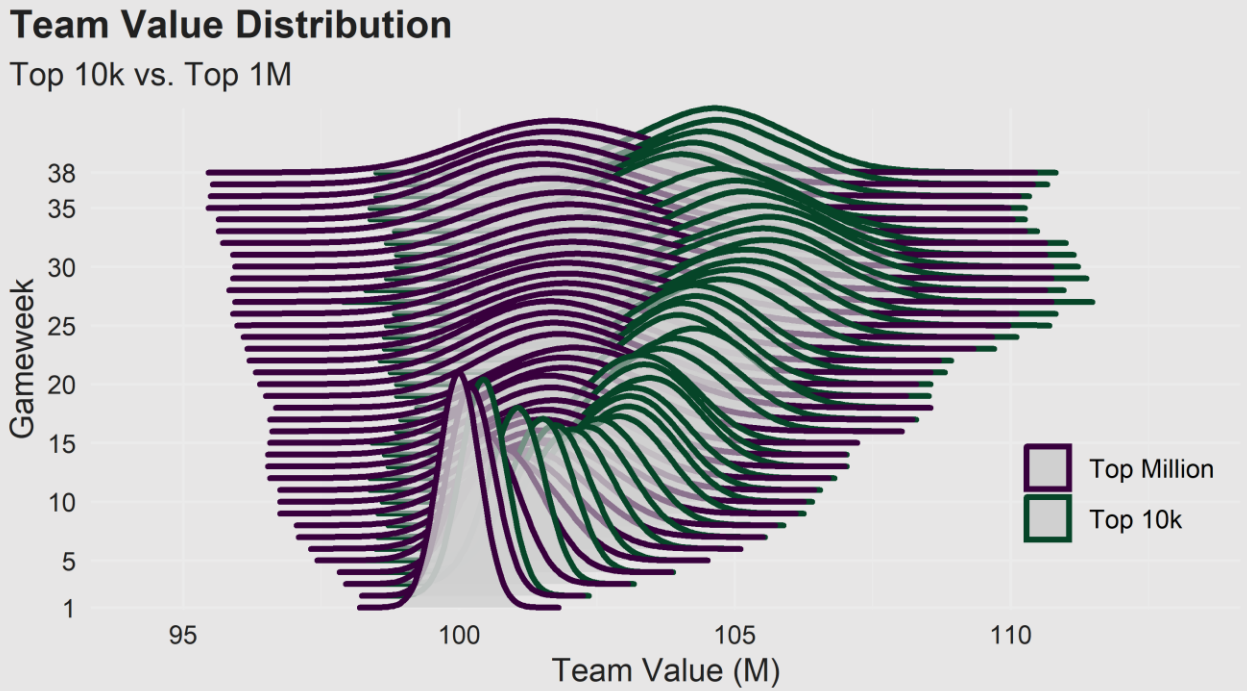




+geom_stratum

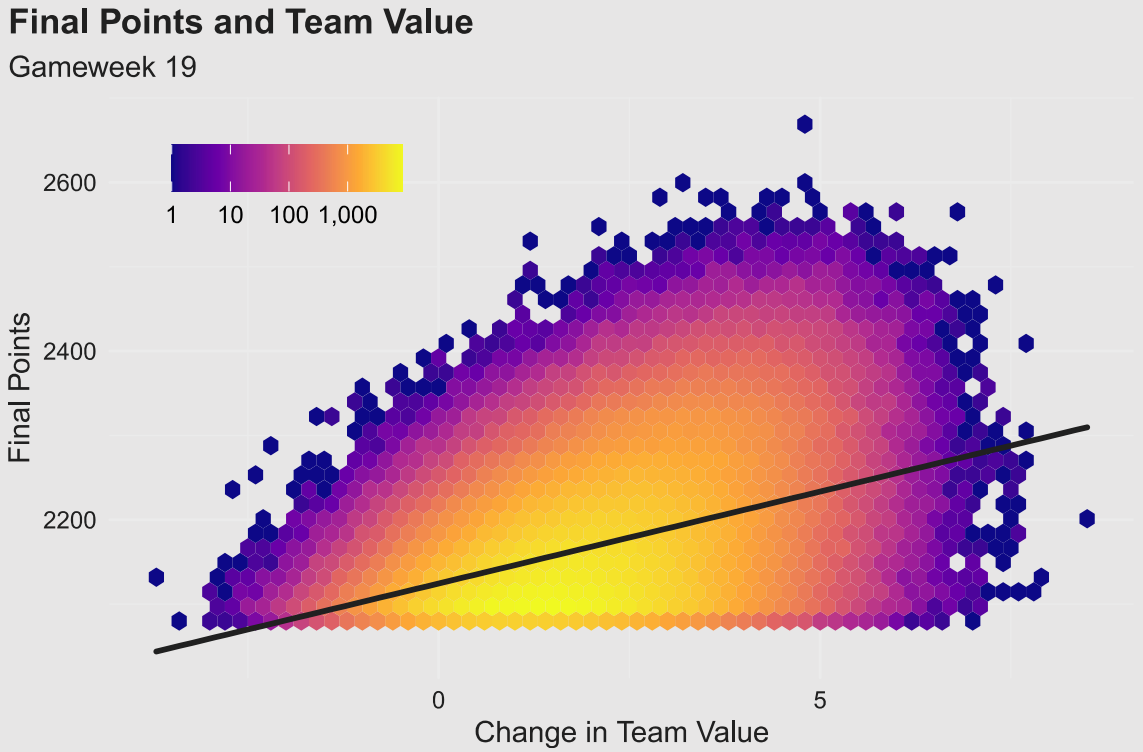
Difference in skill - money

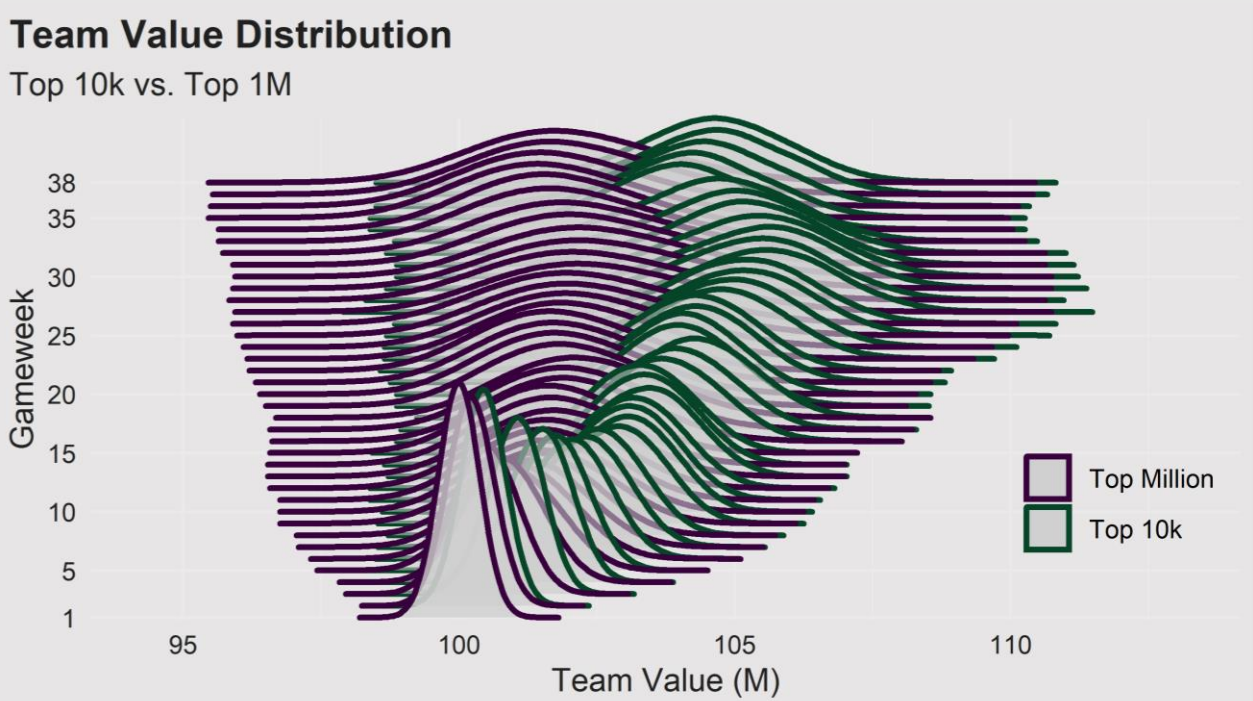




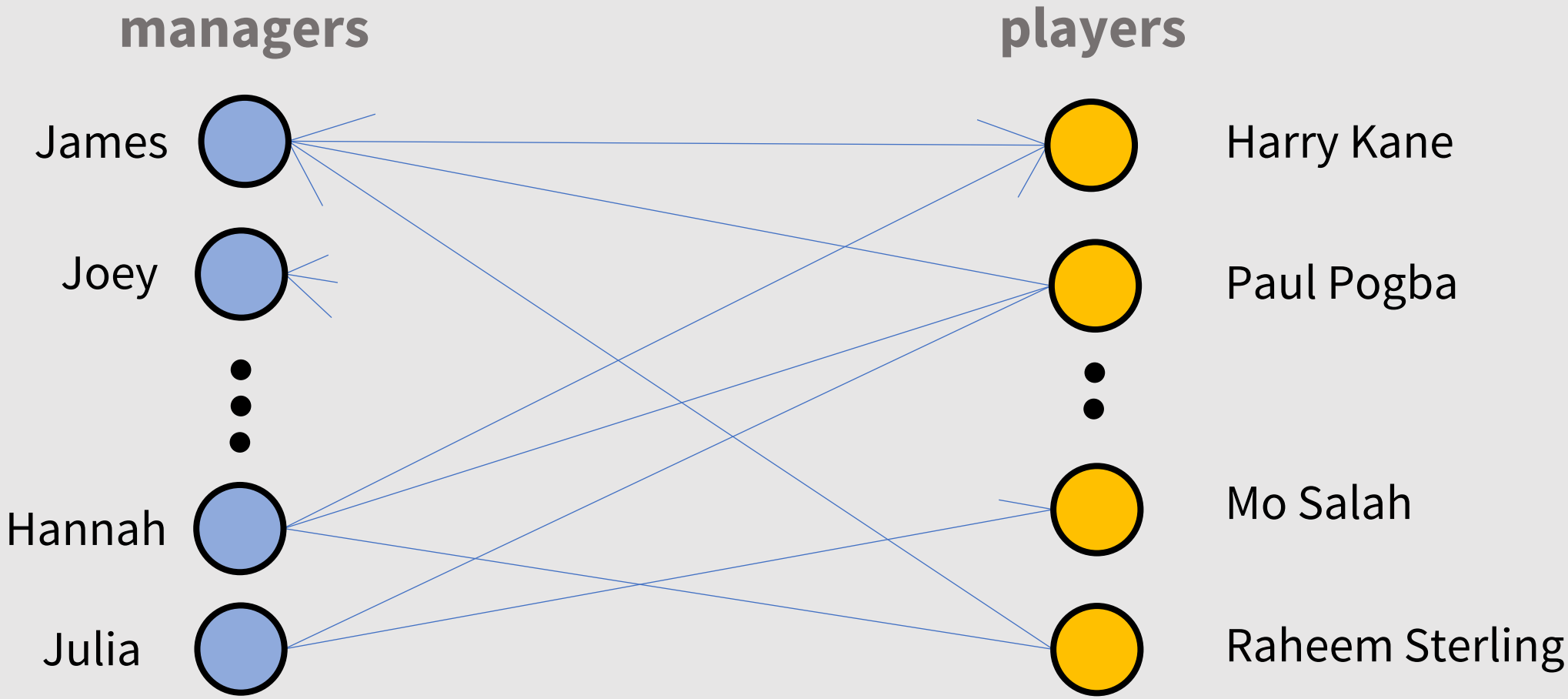
ggridges

Difference in skill - money

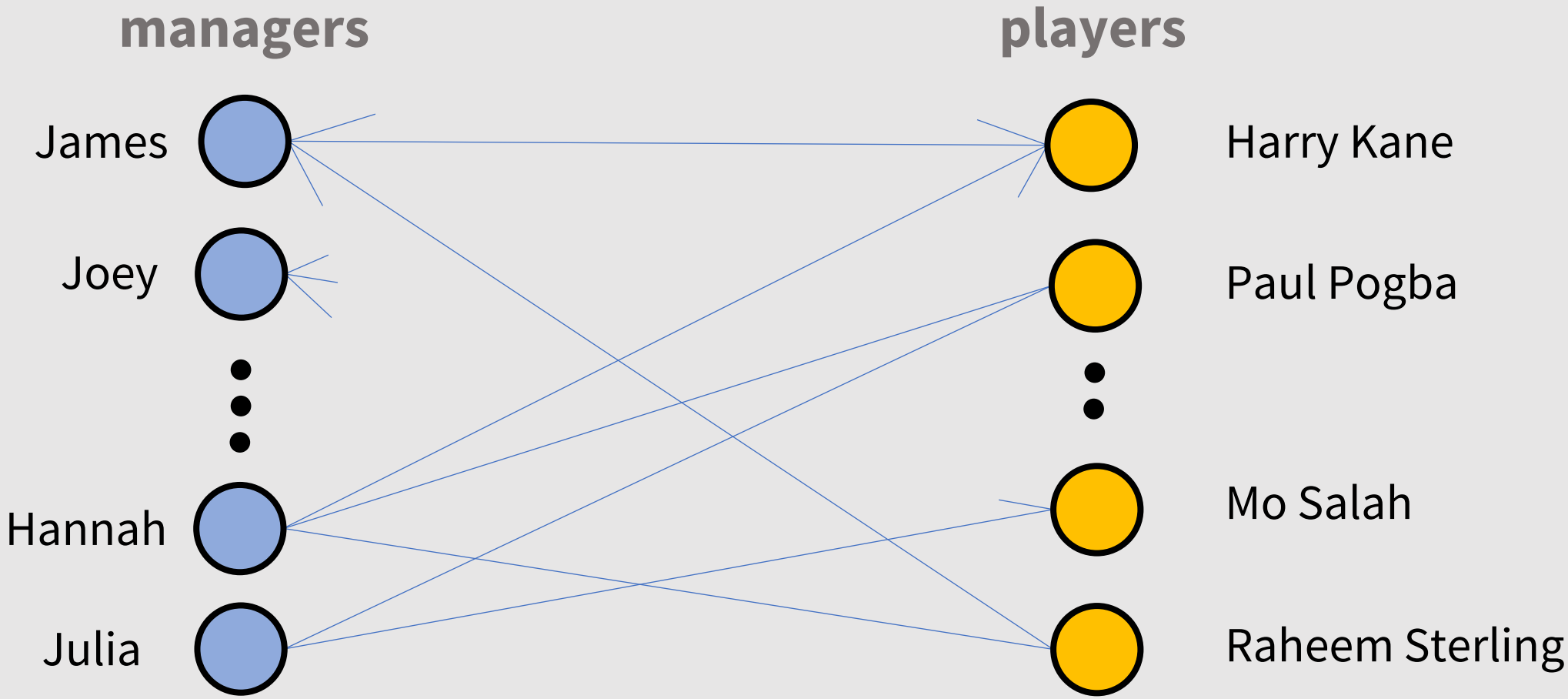




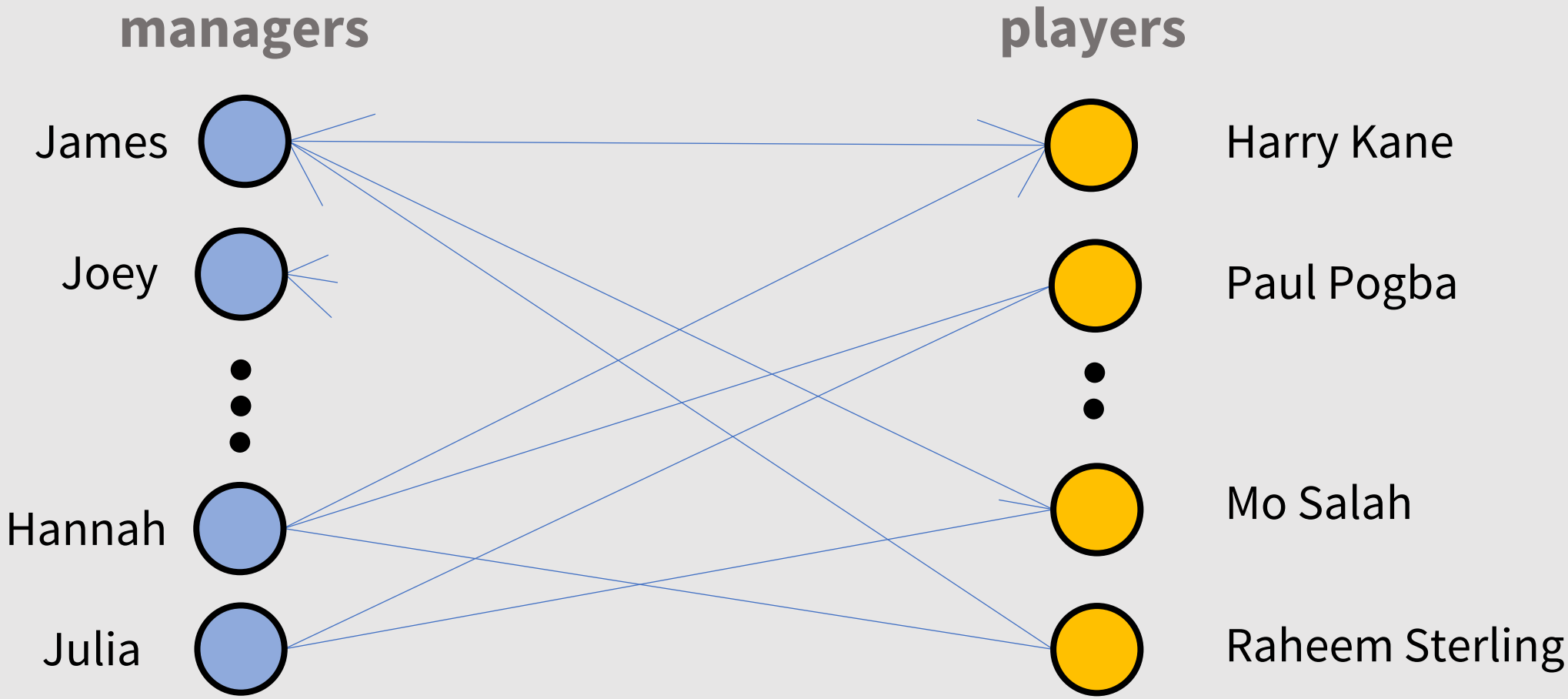
Network analysis



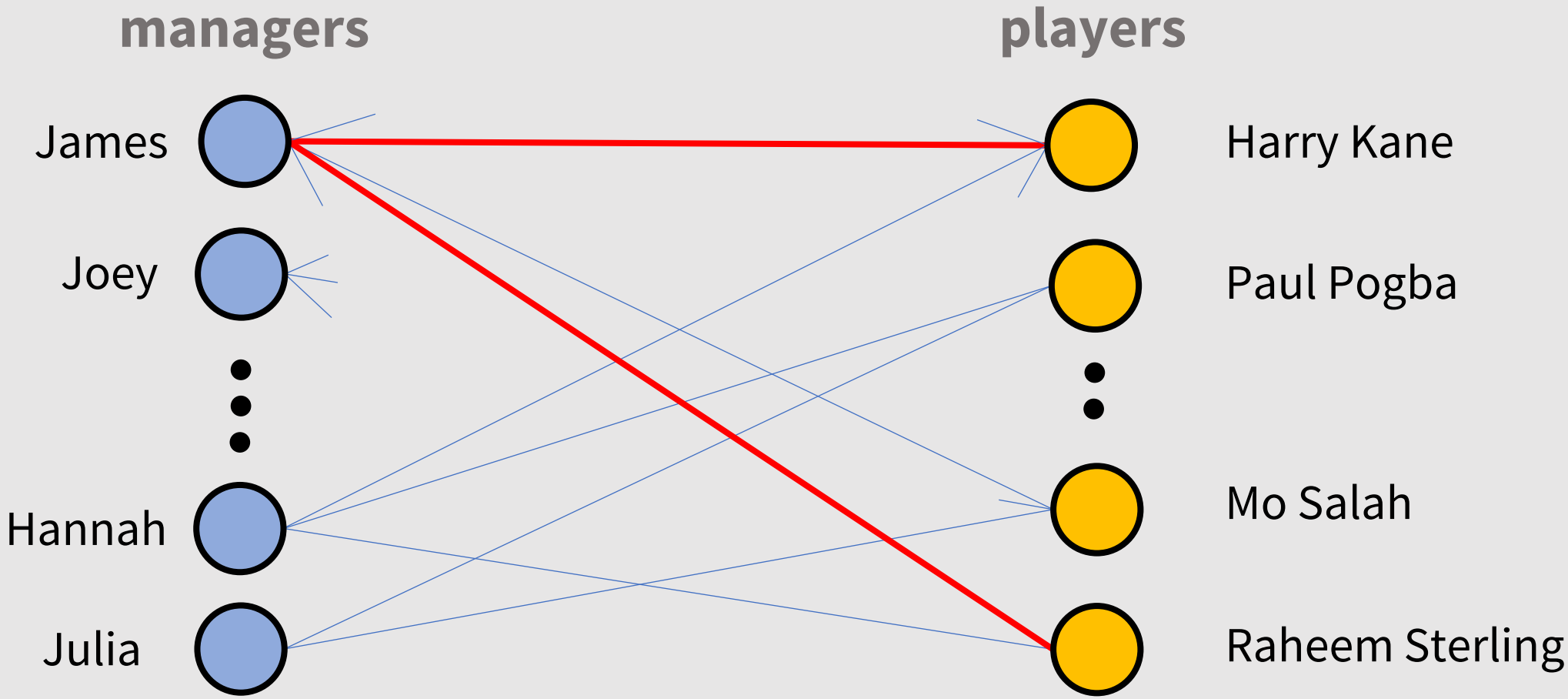
Network analysis



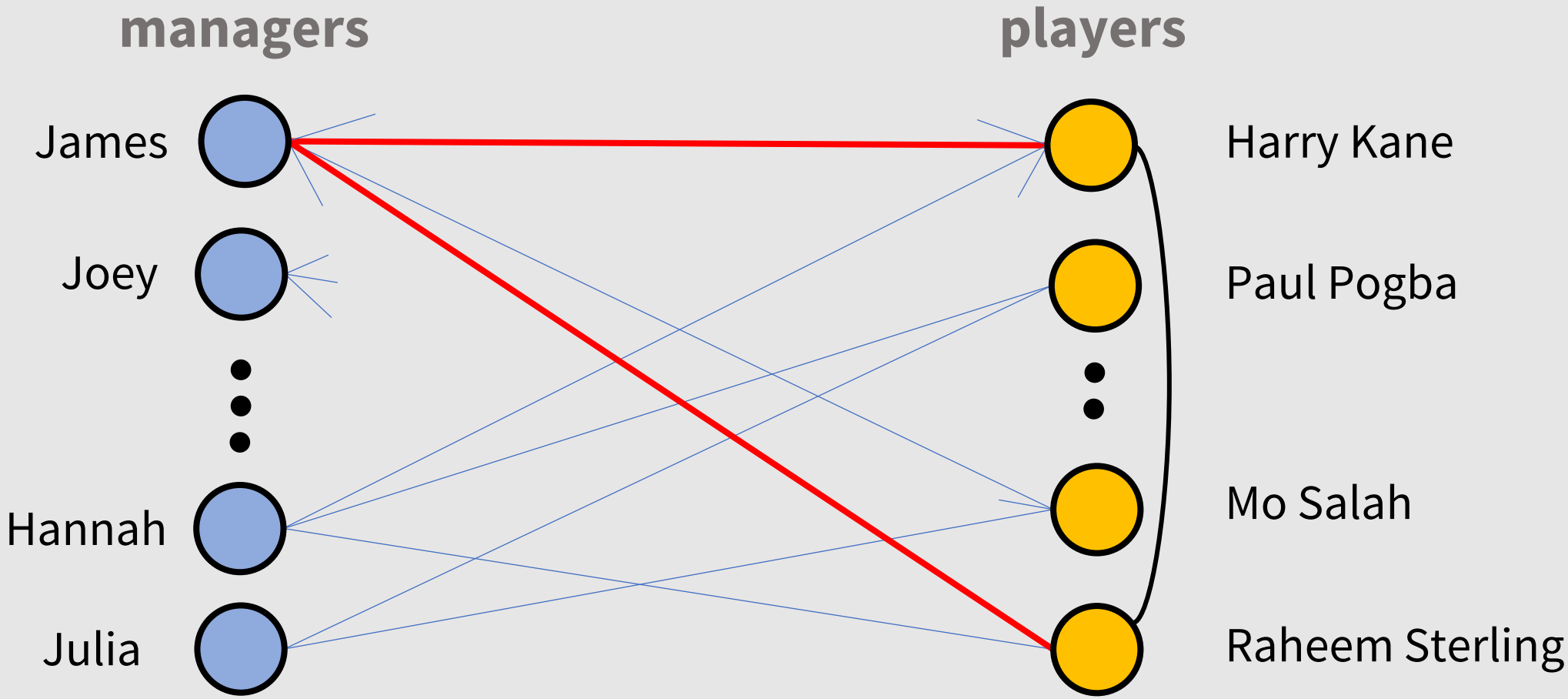
Network analysis



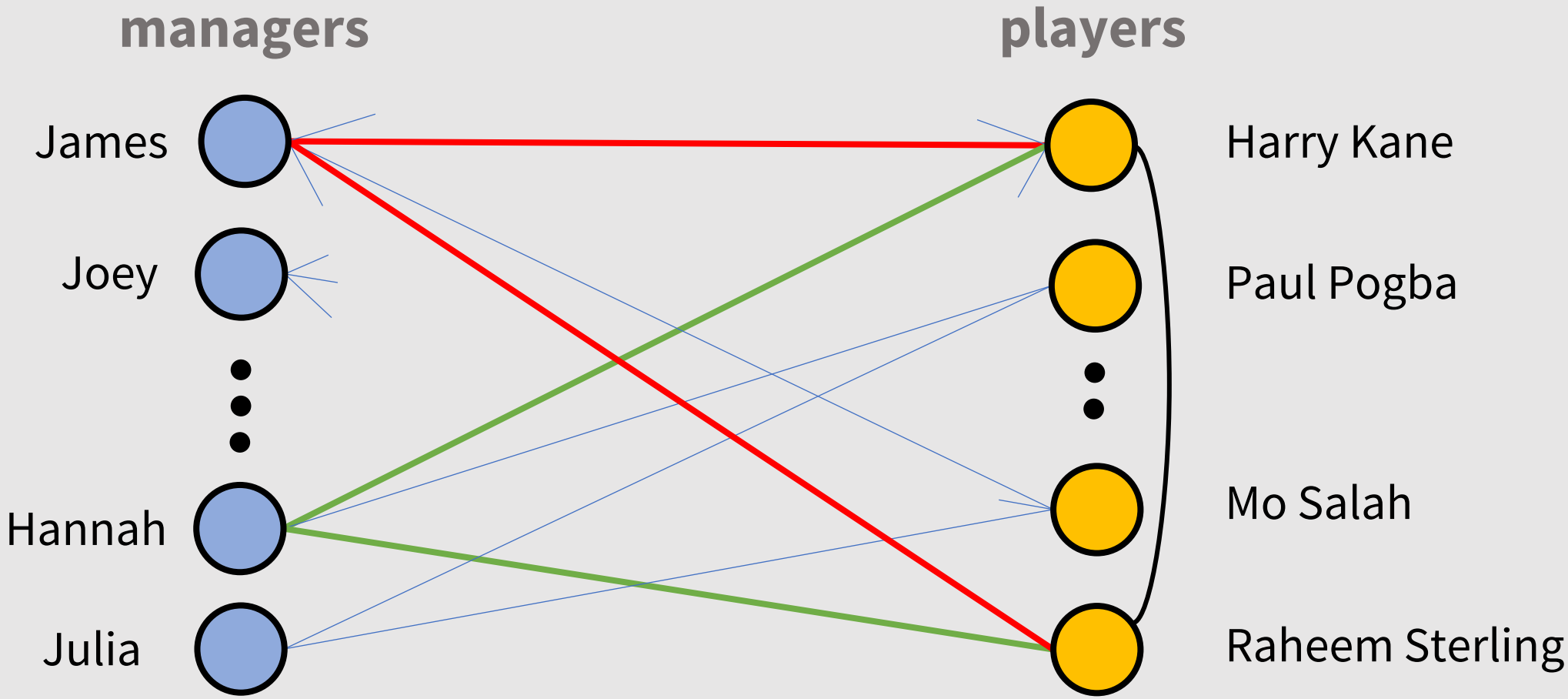
Network analysis



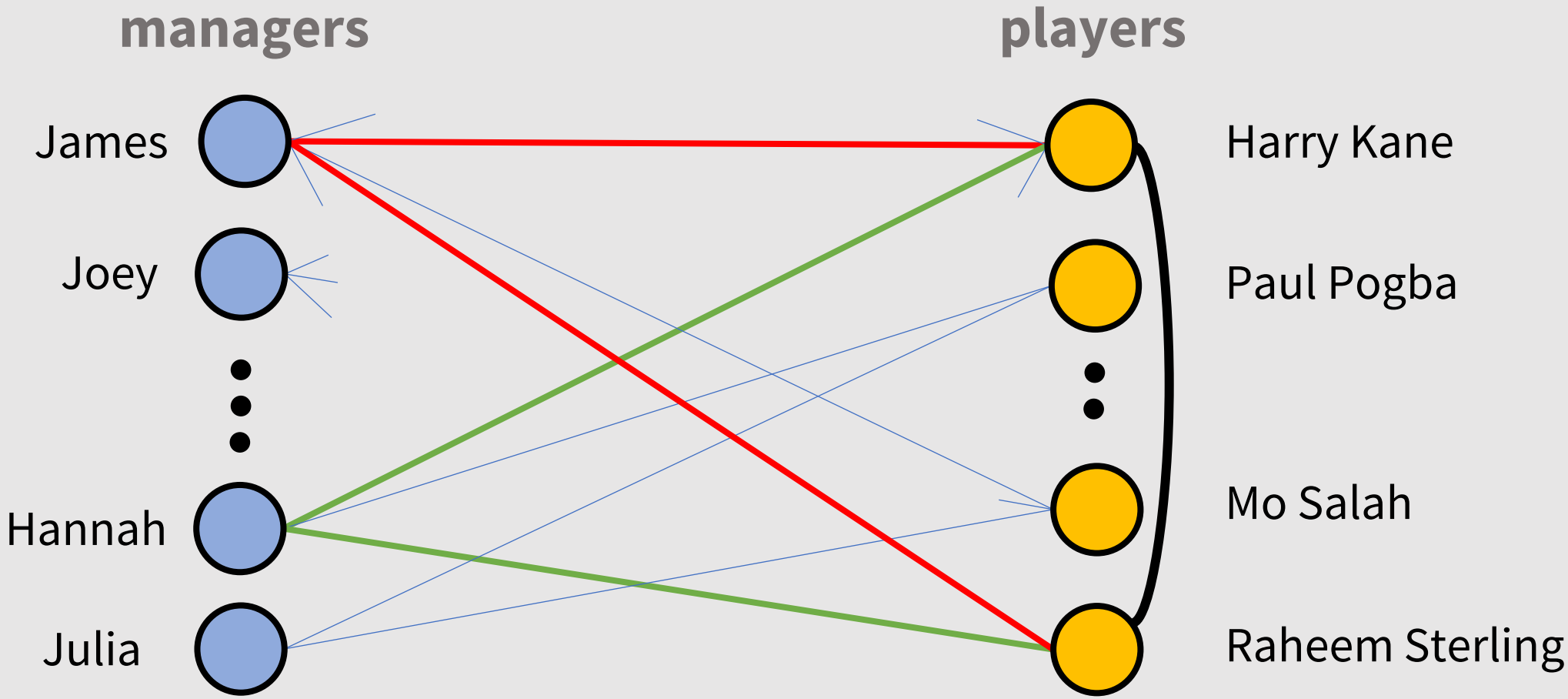
Network analysis



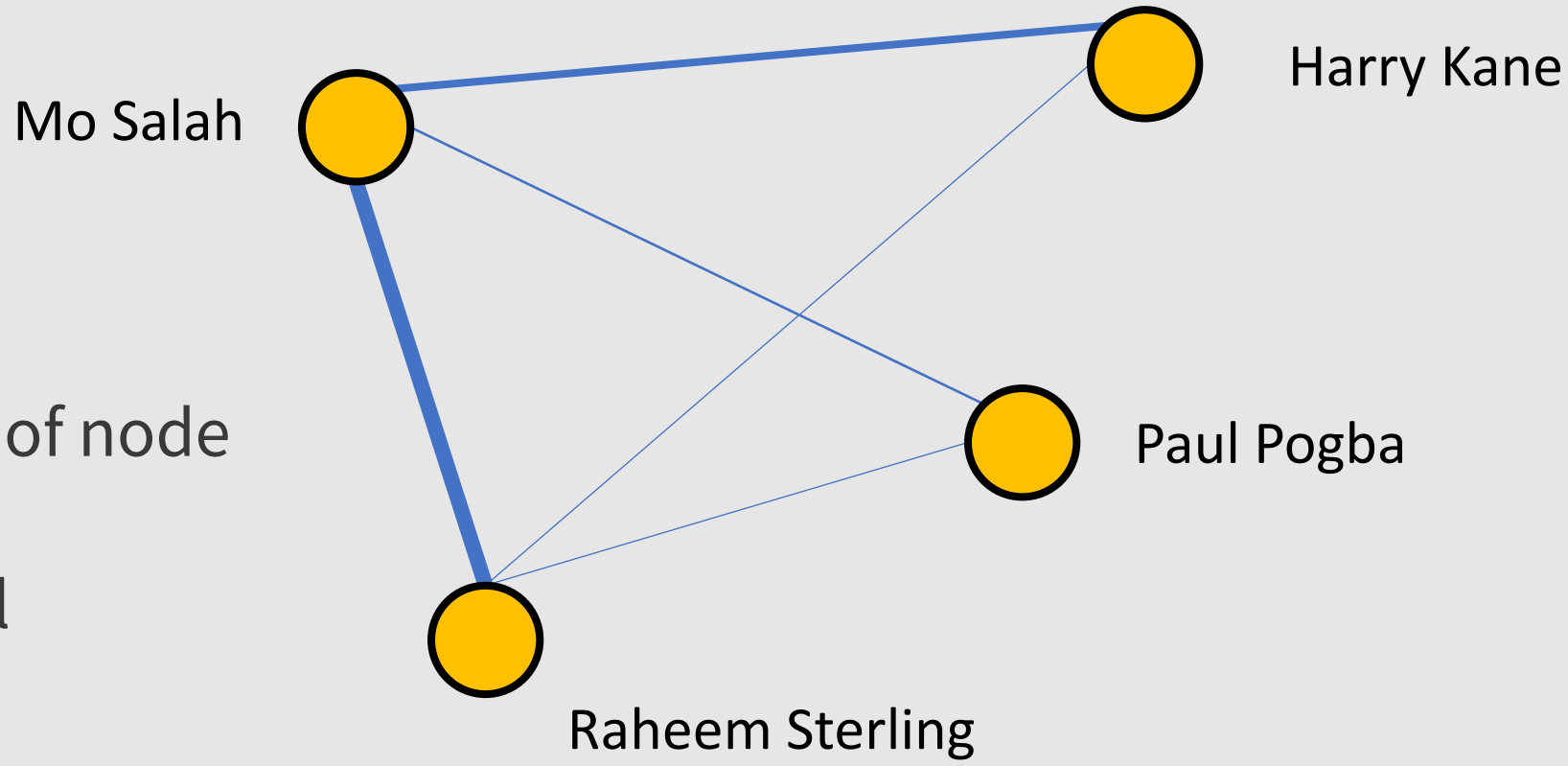
Network analysis



Network analysis



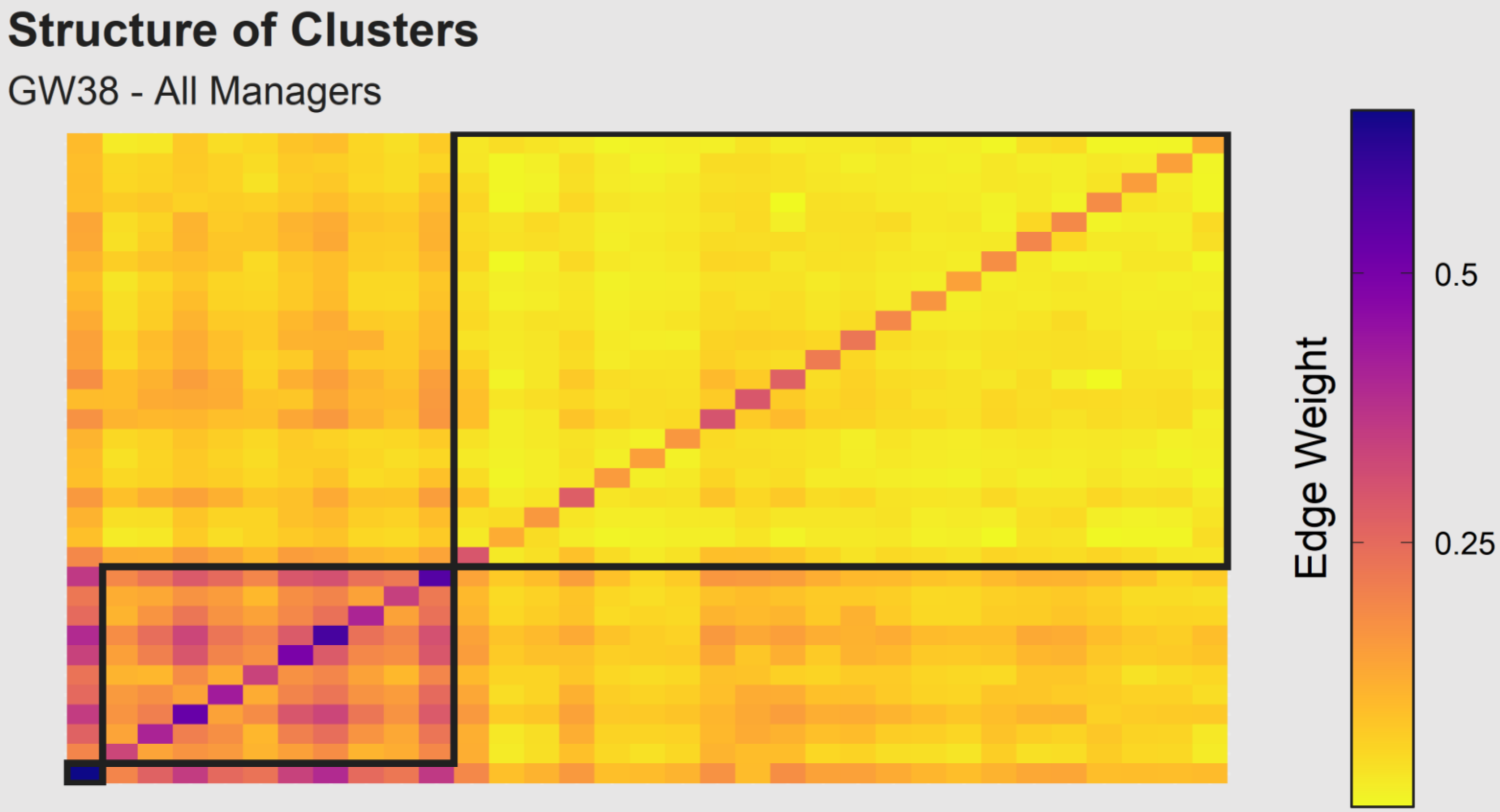
Network analysis



- One type of node
- Weighted
- Temporal
- **igraph**

Template team

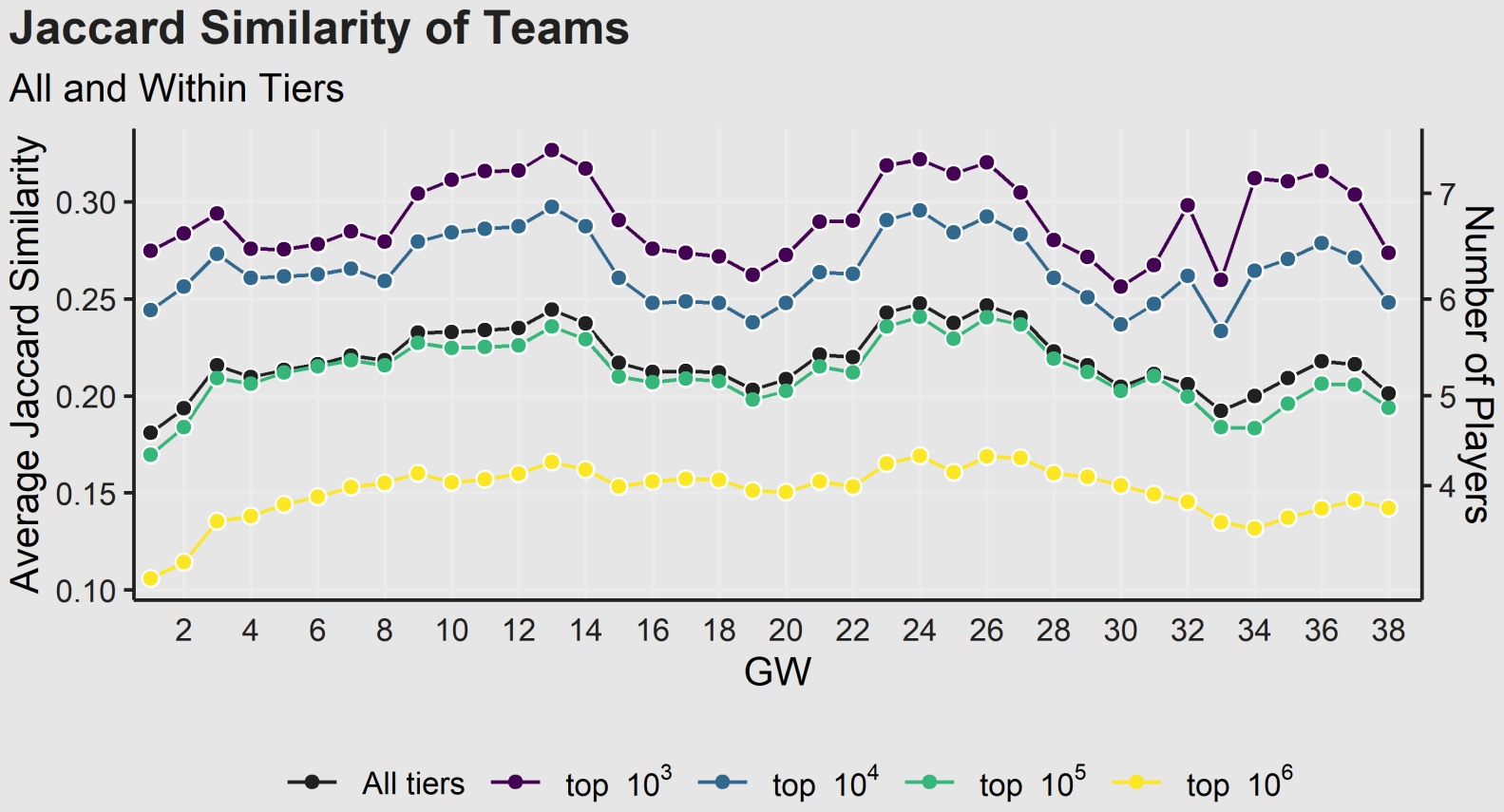
- From the network we can identify **clusters of players** based upon their section frequency.
- Find that four clusters can describe the different groups with **three of them** containing only **~30 players** (out of >600).



Template team

- We determine the similarity between two teams A and B through the **Jaccard Similarity** measure.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



In Conclusion

- It isn't one-size-fits-all

In Conclusion

- It isn't one-size-fits-all

In Conclusion

- It isn't one-size-fits-all

‘Take these three items right here. You can have this. WD-40, vise grips, and some duct tape. Any man worth his salt can do half the household chores with just those three things.’

Walt Kowalski

Thank you for listening!

@obrienj_