AMERICAN BASKETBALL PLAYERS

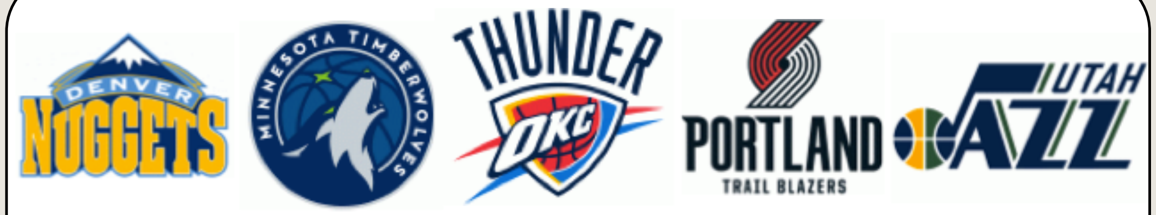
NETWORK TOUR OF DATA SCIENCE

Olivier COUQUE, Emma LEJAL GLAUDE, Matthieu SAUVÉ

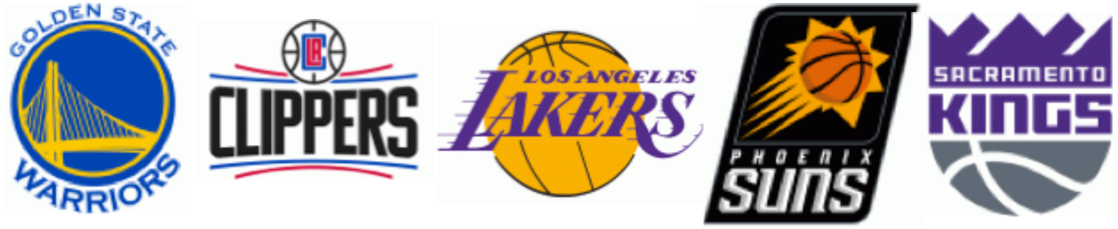
GOAL

Can we recognize the **team clusters** of the American Basketball **players** given their **Twitter** network ?

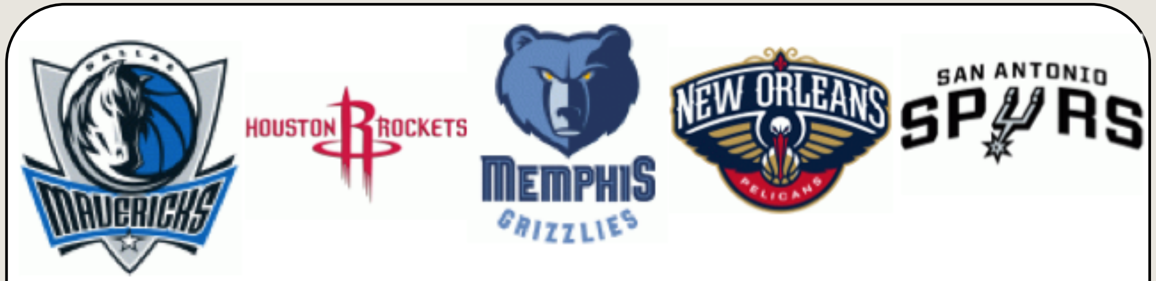
WESTERN CONFERENCE



Northwest Division

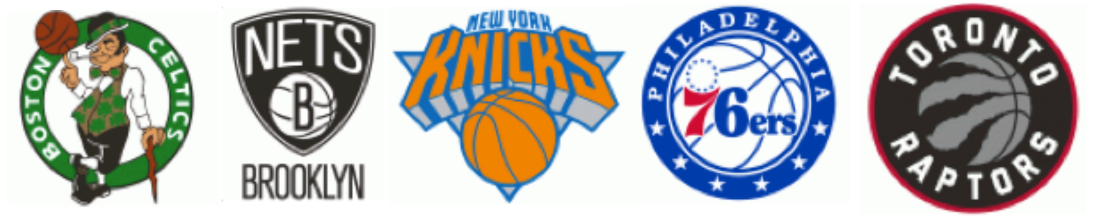


Pacific Division

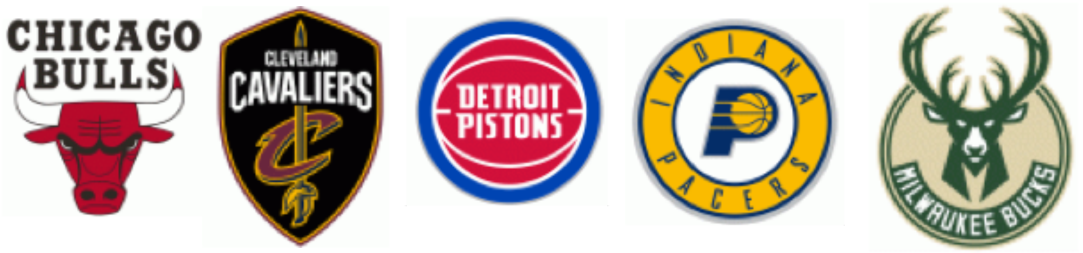


Southwest Division

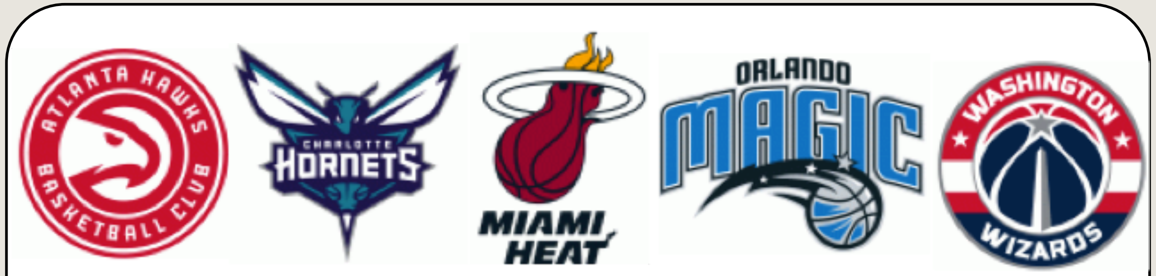
EASTERN CONFERENCE



Atlantic Division



Central Division



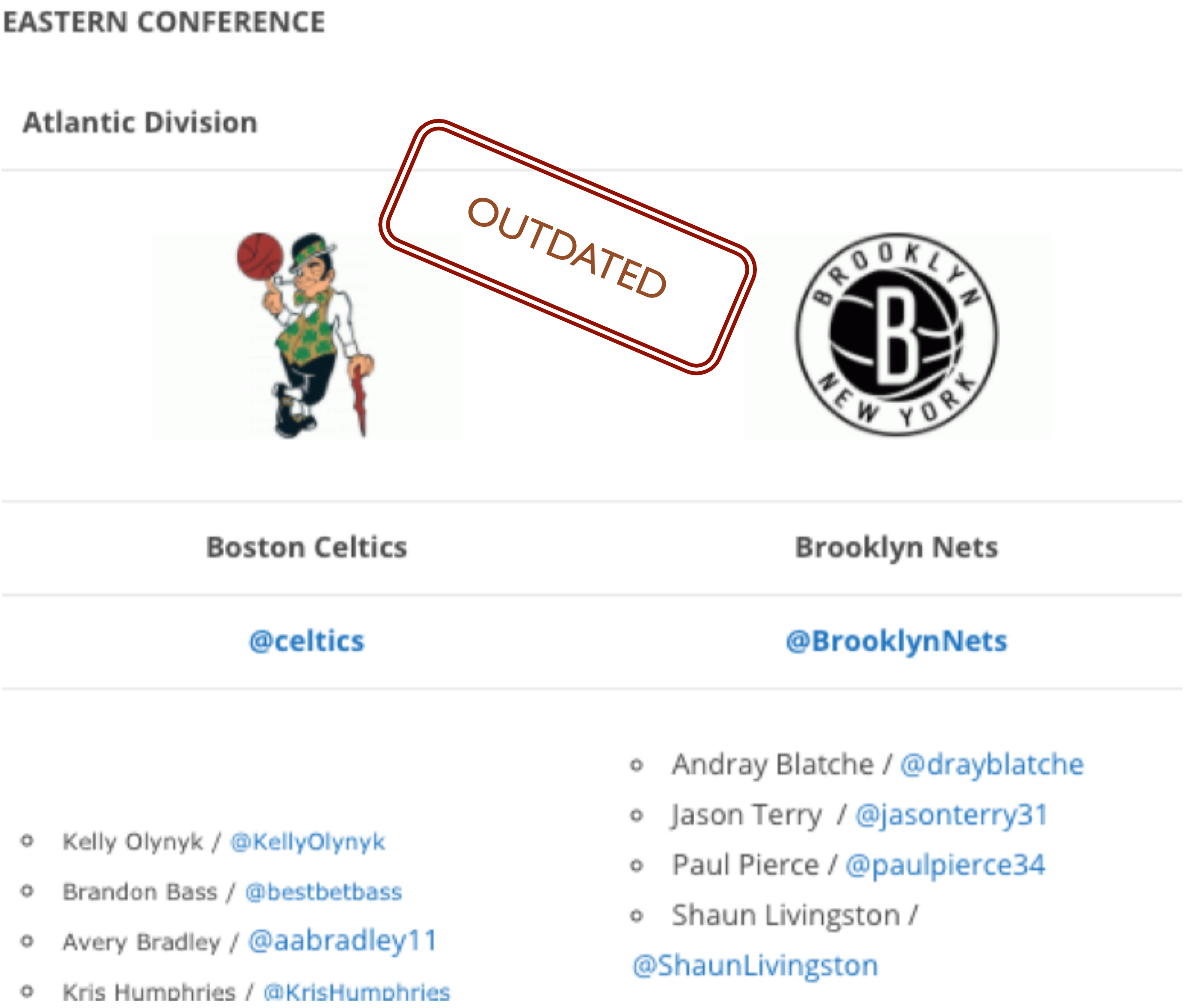
Southeast Division

DATA COLLECTION

What data do we need ?

- For each player, we need their Twitter account → +400 players
- For each player, we need the name of their team → 30 teams
- For each team, we need the name of its division → 6 divisions
- For each division, we need the name of its conference → 2 conferences
- For each pair of players (p1, p2), we need to know if p1 follows p2 on Twitter

- First source of data : Hoopeduponline.com
  - Repartition of teams in conferences and division
  - Repartition of players in teams
  - Pair of player name and twitter account



EASTERN CONFERENCE

Atlantic Division

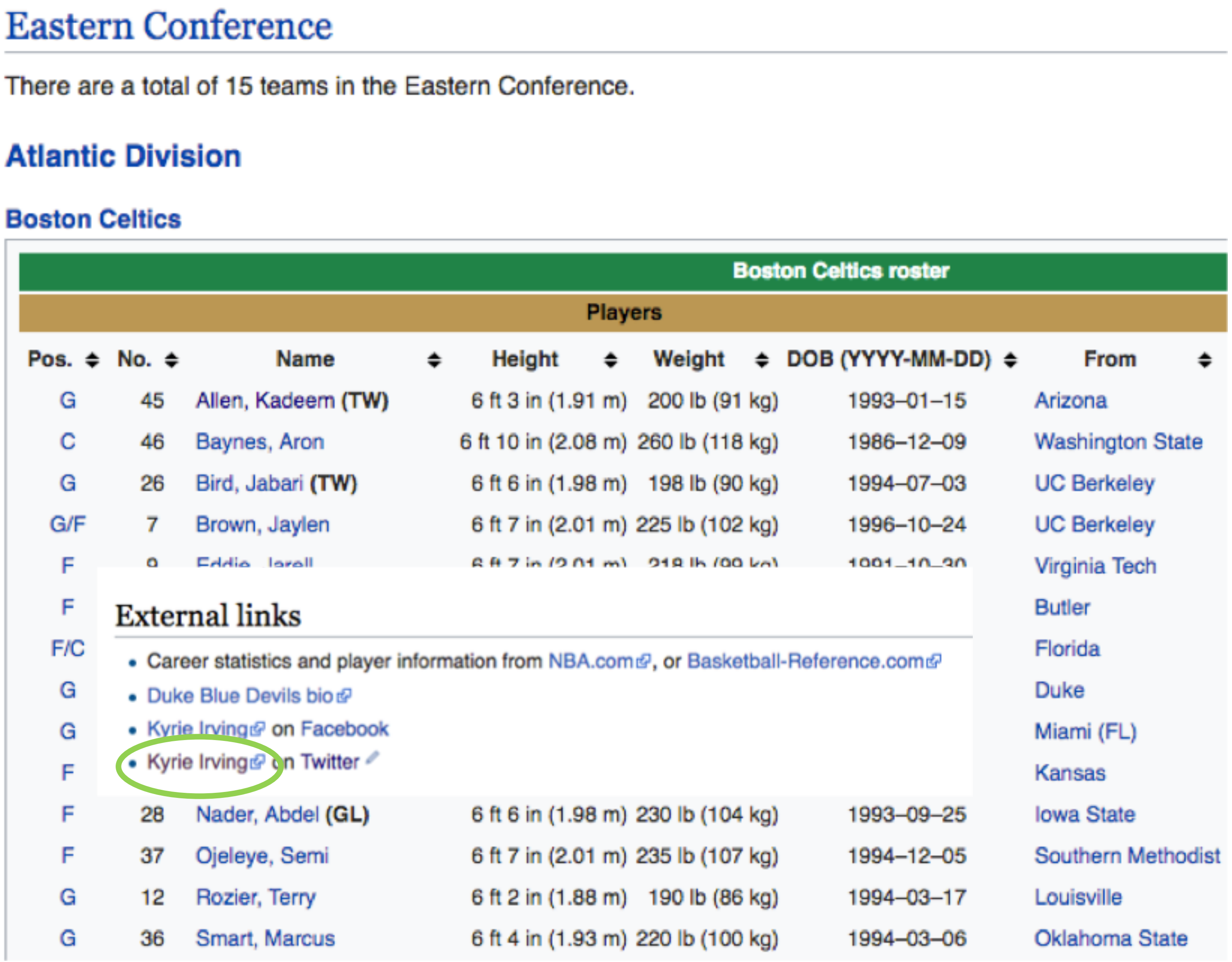
OUTDATED

Boston Celtics	Brooklyn Nets
@celtics	@BrooklynNets

- Kelly Olynyk / @KellyOlynyk
- Brandon Bass / @bestbetbass
- Avery Bradley / @aabradley11
- Kris Humphries / @KrisHumphries

- Andray Blatche / @drayblatche
- Jason Terry / @jasonterry31
- Paul Pierce / @paulpierce34
- Shaun Livingston / @ShaunLivingston

- First source of data : Hoopeduponline.com
  - Repartition of teams in conferences and division
  - Repartition of players in teams
  - Pair of player name and twitter account
- Second source of data :Wikipedia
  - Current repartition of players in teams
  - More twitter accounts in player personal pages



Eastern Conference

There are a total of 15 teams in the Eastern Conference.

Atlantic Division

Boston Celtics

Boston Celtics roster

Players

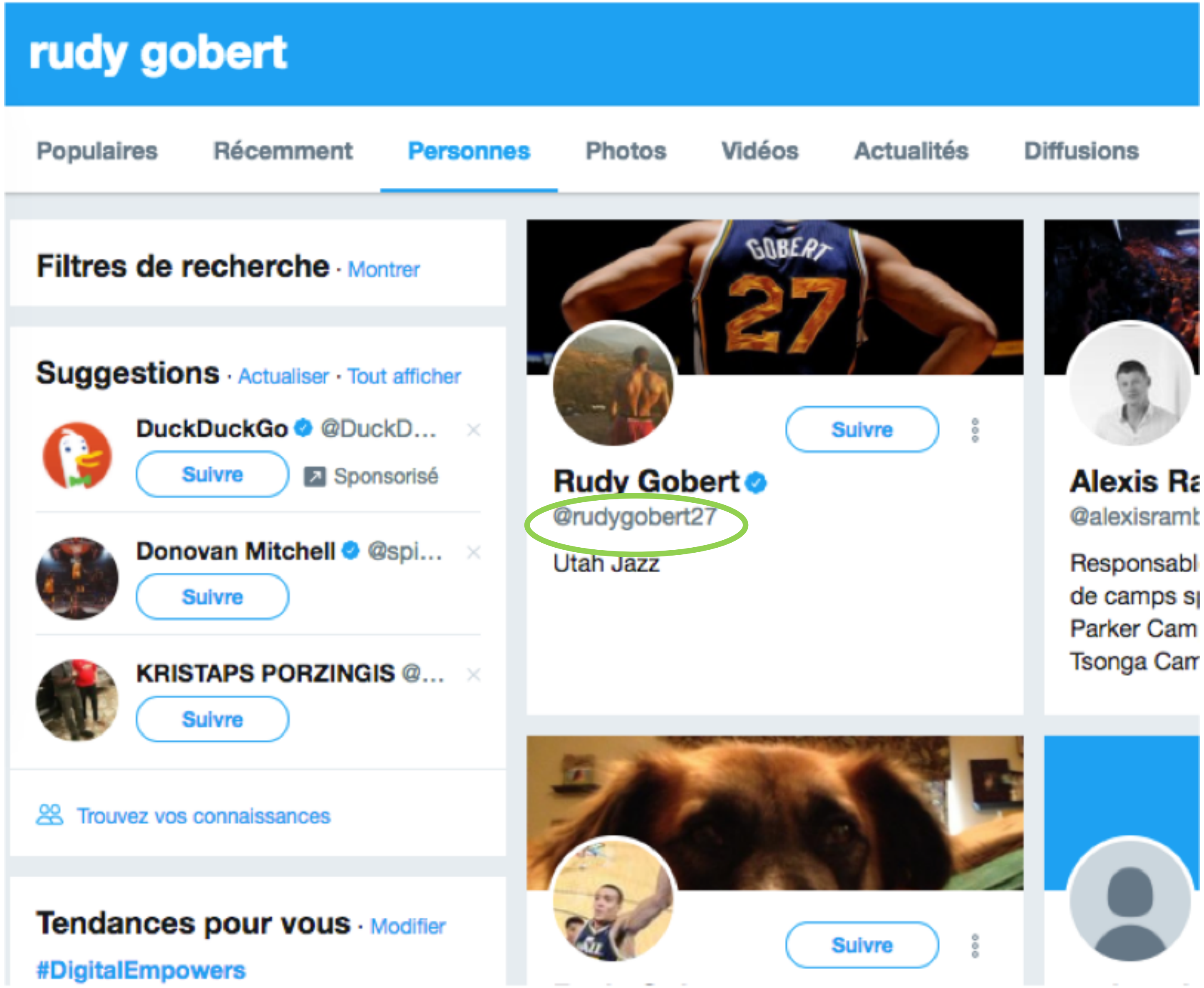
Pos. ⬍	No. ⬍	Name ⬍	Height ⬍	Weight ⬍	DOB (YYYY-MM-DD) ⬍	From ⬍
G	45	Allen, Kadeem **(TW)**	6 ft 3 in (1.91 m)	200 lb (91 kg)	1993–01–15	Arizona
C	46	Baynes, Aron	6 ft 10 in (2.08 m)	260 lb (118 kg)	1986–12–09	Washington State
G	26	Bird, Jabari **(TW)**	6 ft 6 in (1.98 m)	198 lb (90 kg)	1994–07–03	UC Berkeley
G/F	7	Brown, Jaylen	6 ft 7 in (2.01 m)	225 lb (102 kg)	1996–10–24	UC Berkeley
F	9	Eddie, Jarell	6 ft 7 in (2.01 m)	218 lb (99 kg)	1991–10–30	Virginia Tech
F						Butler
F/C						Florida
G						Duke
G						Miami (FL)
F						Kansas
F	28	Nader, Abdel **(GL)**	6 ft 6 in (1.98 m)	230 lb (104 kg)	1993–09–25	Iowa State
F	37	Ojeleye, Semi	6 ft 7 in (2.01 m)	235 lb (107 kg)	1994–12–05	Southern Methodist
G	12	Rozier, Terry	6 ft 2 in (1.88 m)	190 lb (86 kg)	1994–03–17	Louisville
G	36	Smart, Marcus	6 ft 4 in (1.93 m)	220 lb (100 kg)	1994–03–06	Oklahoma State

External links

- Career statistics and player information from NBA.com, or Basketball-Reference.com
- Duke Blue Devils bio
- Kyrie Irving on Facebook
- Kyrie Irving on Twitter

- First source of data : Hoopeduponline.com
  - Repartition of teams in conferences and division
  - Repartition of players in teams
  - Pair of player name and twitter account
- Second source of data : Wikipedia
  - Current repartition of players in teams
  - More twitter accounts in player personal pages
- Third source of data : Twitter
  - Last resource to complete the twitter accounts

We can now use Tweepy to get the links between our twitter accounts and complete our graph.



CLUSTERING METHODS

- K-Means
- DBSCAN
- Spectral Clustering
- Principal Component Analysis & K-means
- Gaussian Mixture Model

With the libraries sklearn and scipy

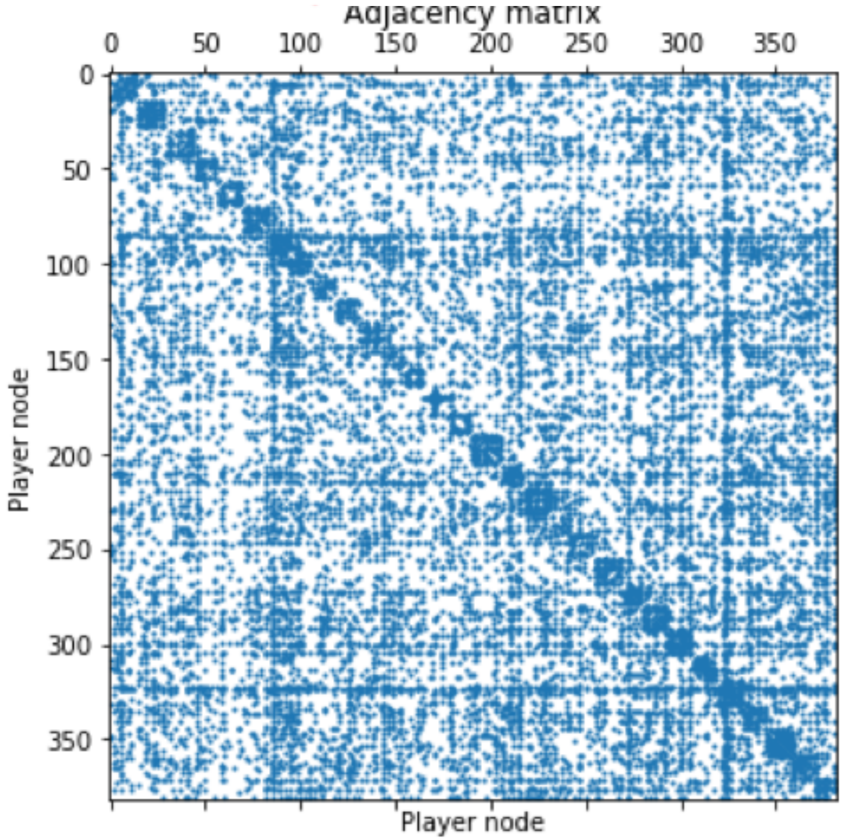
CLUSTERING METHODS

- K-Means
- DBSCAN
- Spectral Clustering
- Principal Component Analysis & K-means
- Gaussian Mixture
- Accuracy Method

1) Compute the lists of names according to ground truth
2) Extract list of names for each cluster from the computed labels
3) Compute the number of common names for each pair (correct list, computer cluster)
4) Determine which pair is the maximum for each computed cluster
5) Sum total of correct names and return as a percentage

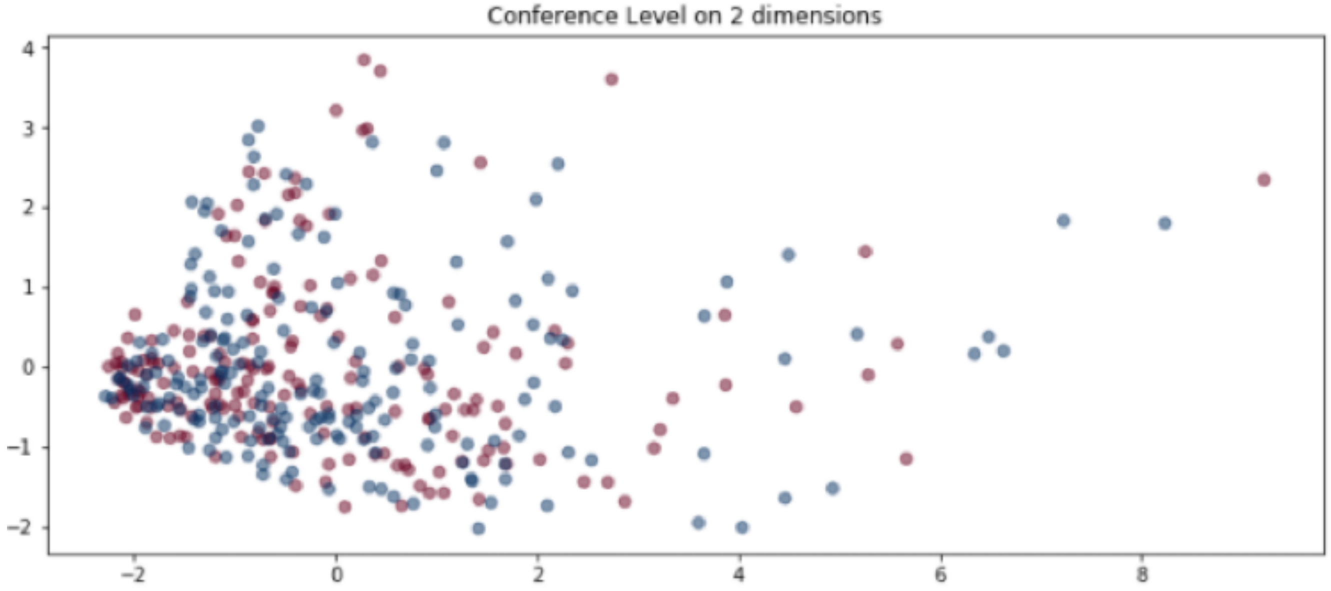
FIRST GRAPH – PLAYER-PLAYER

- Construction
  - Add all the players twitter accounts as nodes
  - Use Tweepy to add edges between (p1, p2) nodes if p1 is following p2
- Analysis
  - We remove the few isolated nodes
  - We study the degree distribution
  - The graph is highly connected



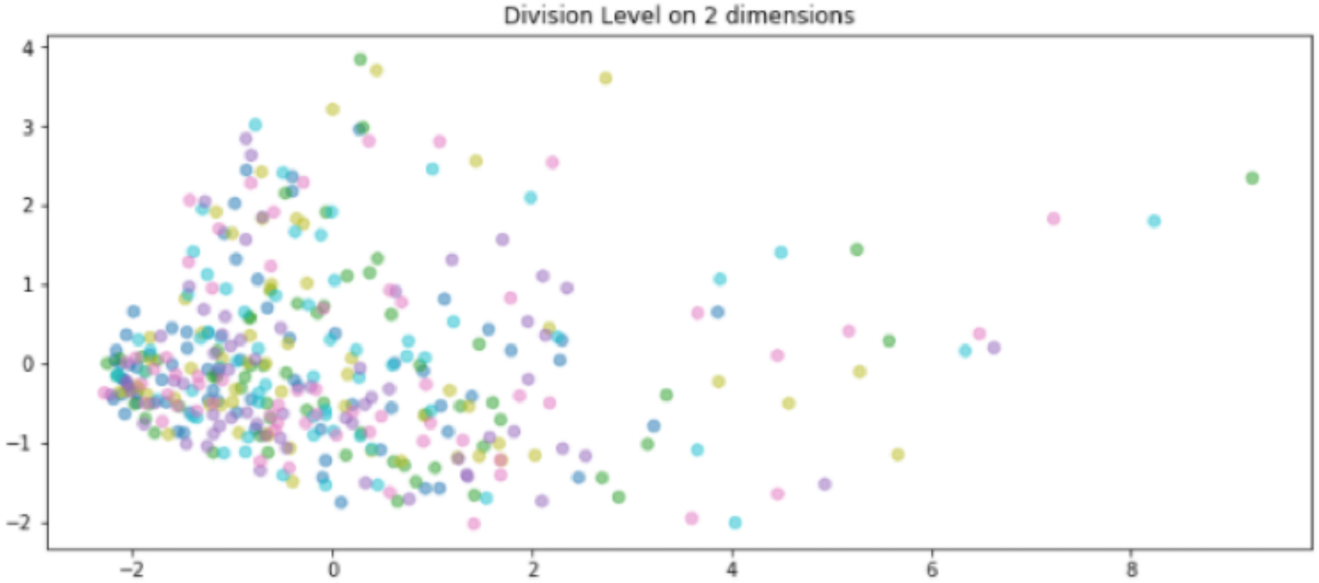
FIRST GRAPH – PLAYER-PLAYER

- Cluster per Conference
  - Best result with : K-means, 200 iterations
  - Accuracy = 53.93 %



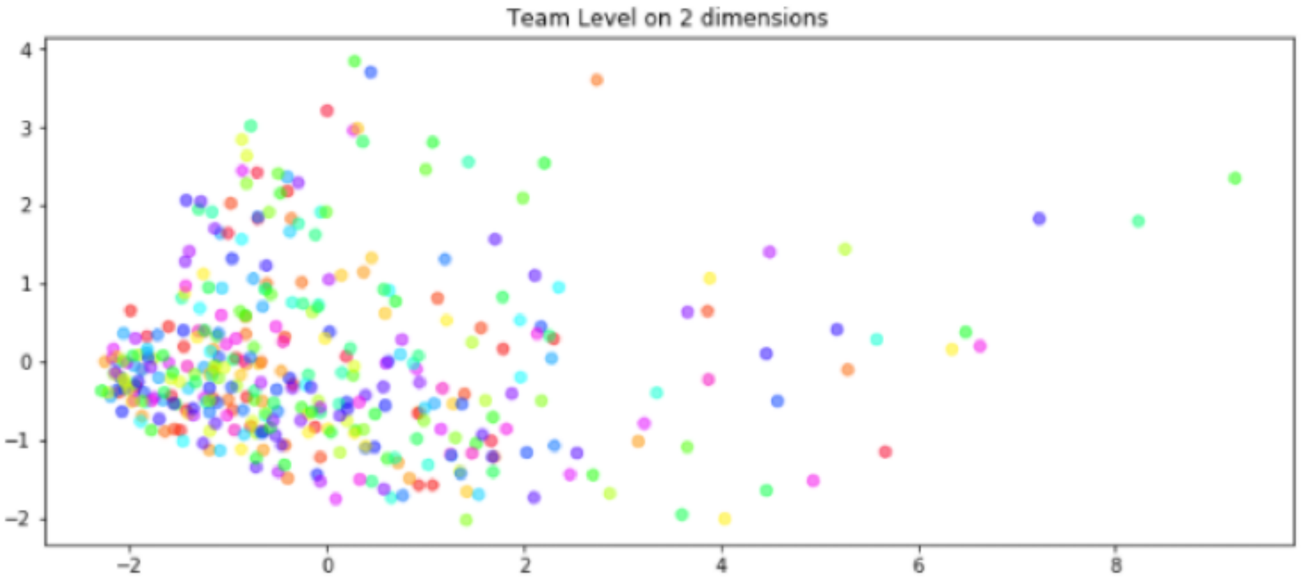
FIRST GRAPH – PLAYER-PLAYER

- Cluster per Conference
  - Best result with : K-means, 200 iterations
  - Accuracy = 53.93 %
- Cluster per Division
  - Best result with : GMM, 5862 iterations
  - Accuracy = 27.49 %



FIRST GRAPH – PLAYER-PLAYER

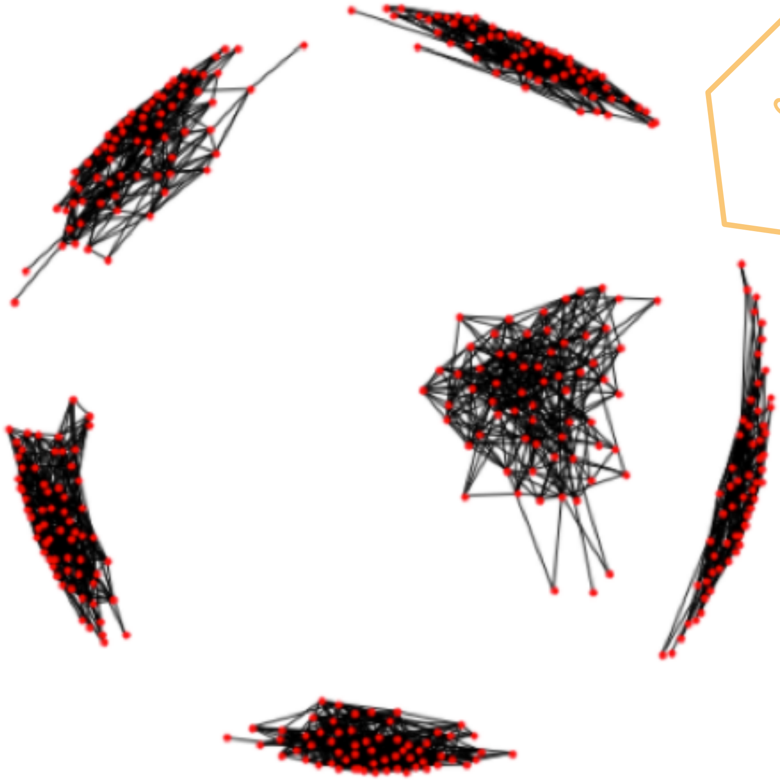
- Cluster per Conference
  - Best result with : K-means, 200 iterations
  - Accuracy = 53.93 %
- Cluster per Division
  - Best result with : GMM, 5682 iterations
  - Accuracy = 27.49 %
- Cluster per Team
  - Best result with : GMM, 4742 iterations
  - Accuracy = 27.23 %



FIRST GRAPH – PLAYER-PLAYER

Optimal Graph :

Remove all noisy (ie, cross-cluster) edges, compute the accuracy when progressively adding the noise

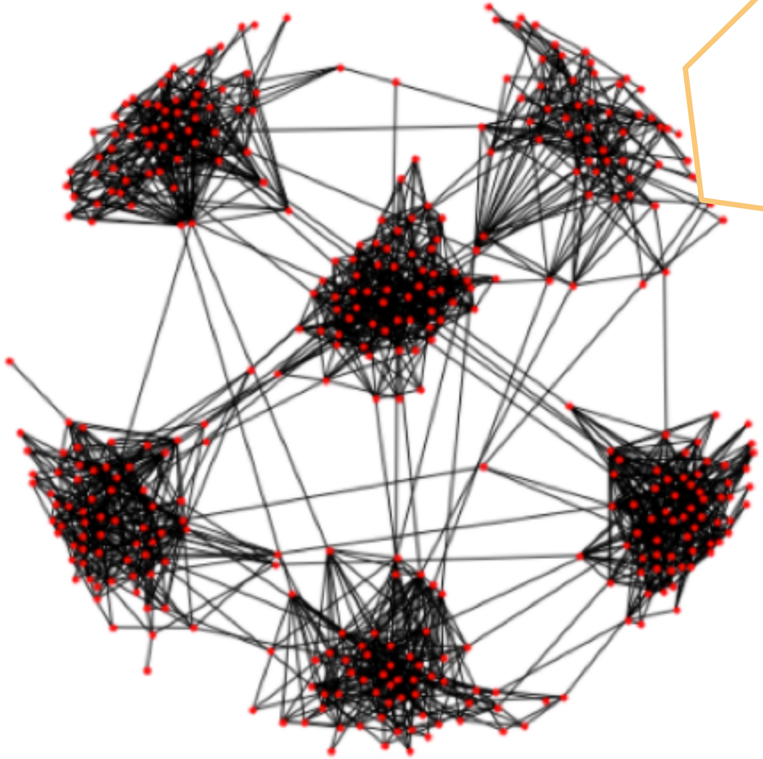


Optimal Division graph

FIRST GRAPH – PLAYER-PLAYER

Optimal Graph :

Remove all noisy (ie, cross-cluster) edges, compute the accuracy when progressively adding the noise

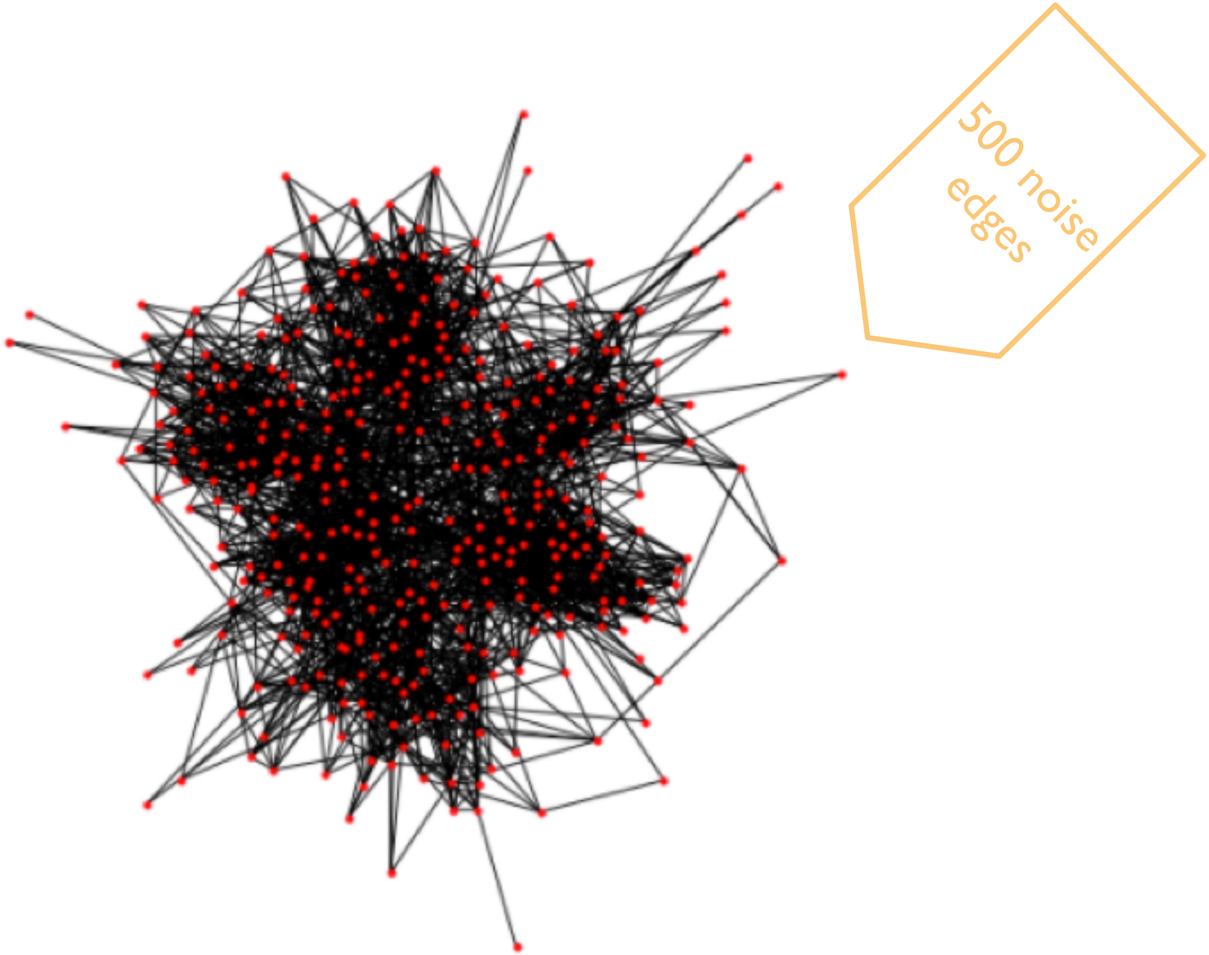


50 noise edges

FIRST GRAPH – PLAYER-PLAYER

Optimal Graph :

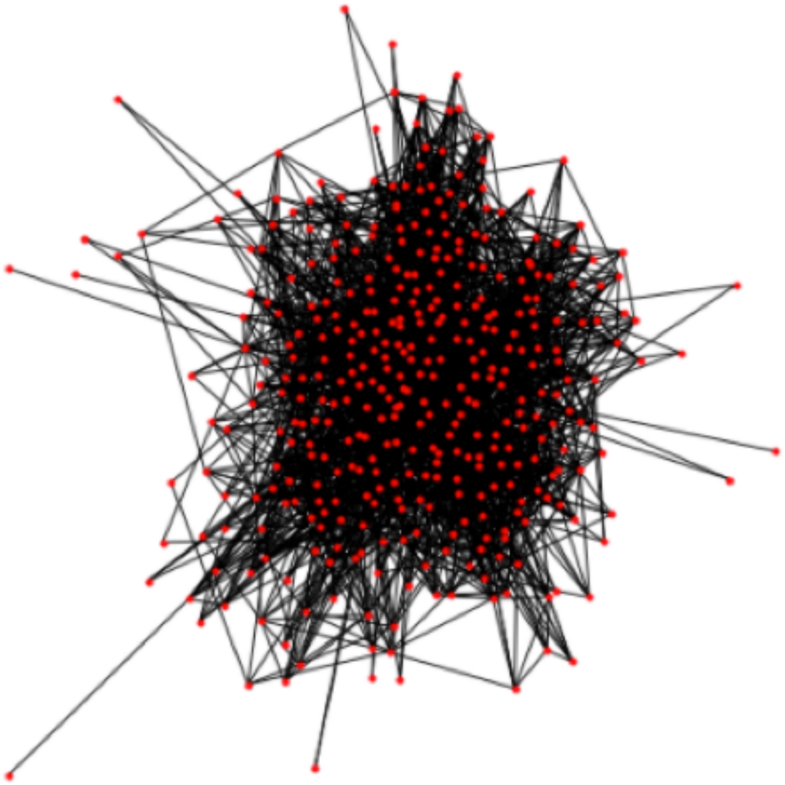
Remove all noisy (ie, cross-cluster) edges, compute the accuracy when progressively adding the noise



FIRST GRAPH – PLAYER-PLAYER

Optimal Graph :

Remove all noisy (ie, cross-cluster) edges, compute the accuracy when progressively adding the noise



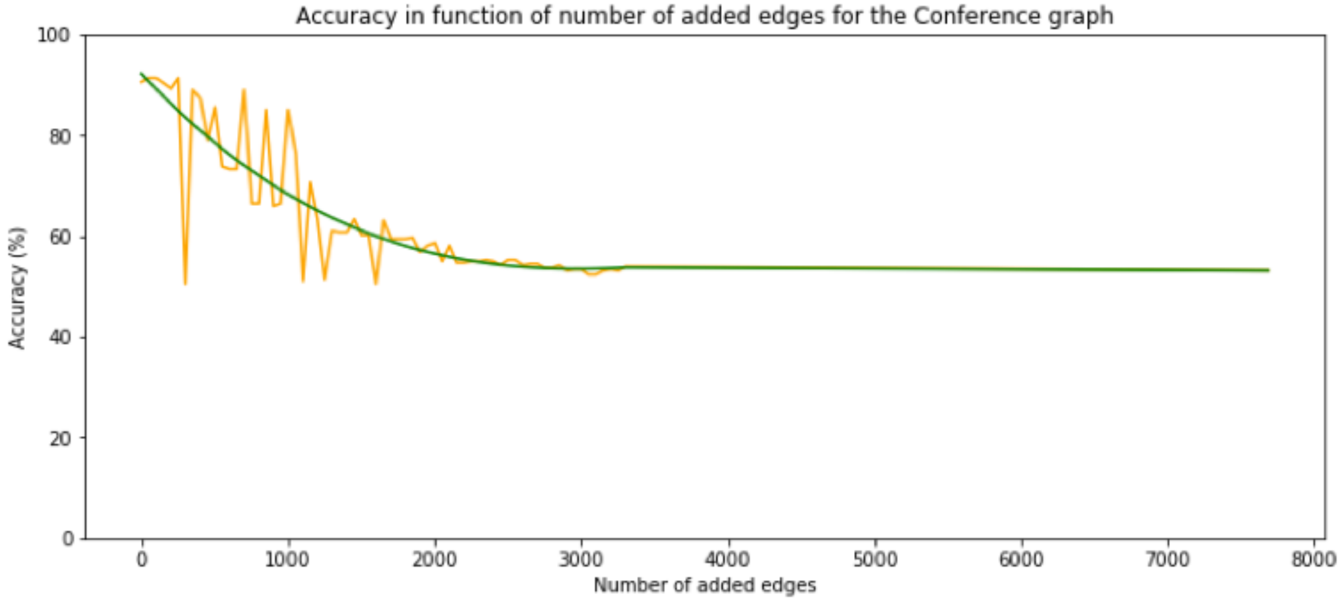
1000 noise edges

FIRST GRAPH – PLAYER-PLAYER

Optimal Graph :

Remove all noisy (ie, cross-cluster) edges, compute the accuracy when progressively adding the noise

- Conference Level
  - Noise edges : 43. 04 %

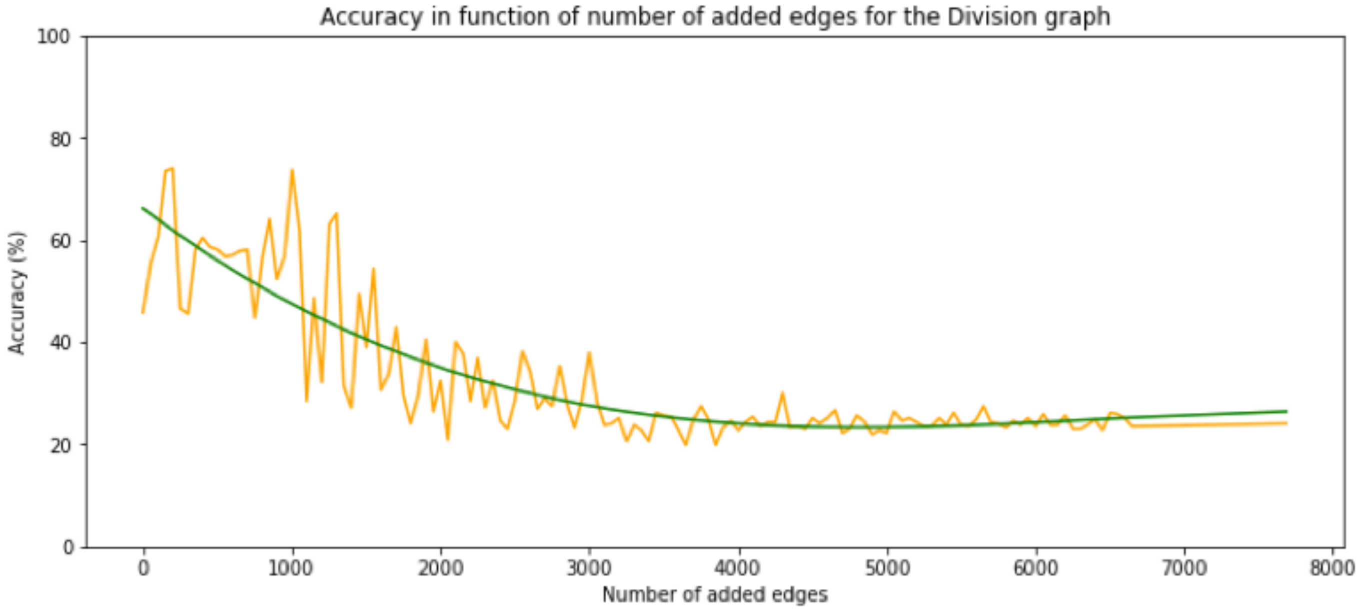


FIRST GRAPH – PLAYER-PLAYER

Optimal Graph :

Remove all noisy (ie, cross-cluster) edges, compute the accuracy when progressively adding the noise

- Conference Level
  - Noise edges : 43. 04 %
- Division Level
  - Noise edges : 73.38 %

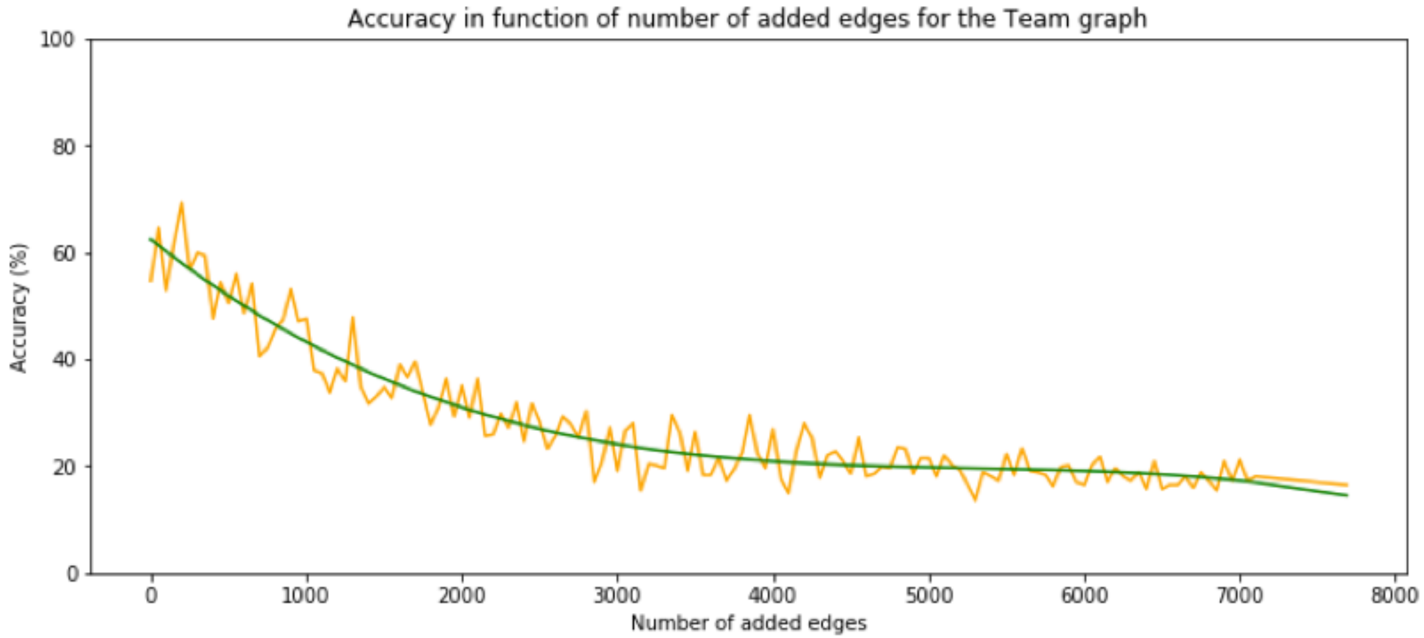


FIRST GRAPH – PLAYER-PLAYER

Optimal Graph :

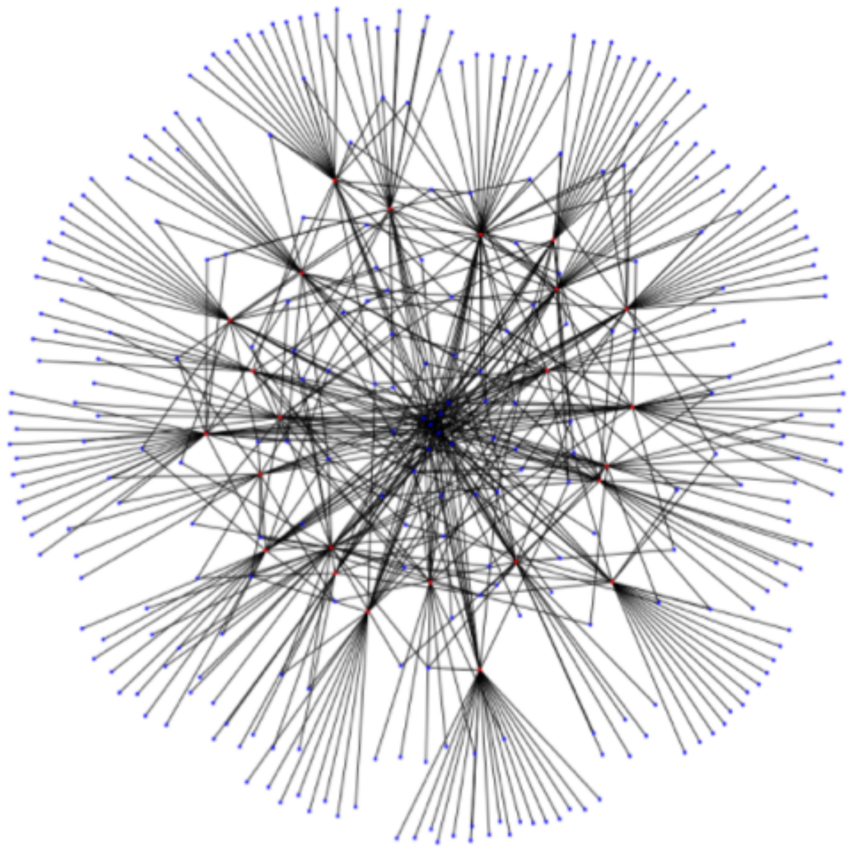
Remove all noisy (ie, cross-cluster) edges, compute the accuracy when progressively adding the noise

- Conference Level
  - Noise edges : 43. 04 %
- Division Level
  - Noise edges : 73.38 %
- Team Level
  - Noise edges : 85.44 %



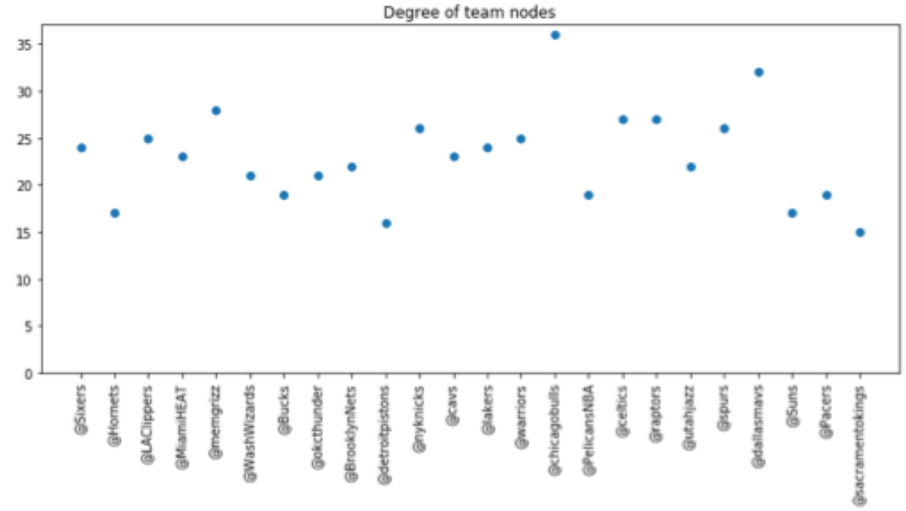
SECOND GRAPH – PLAYER-TEAM

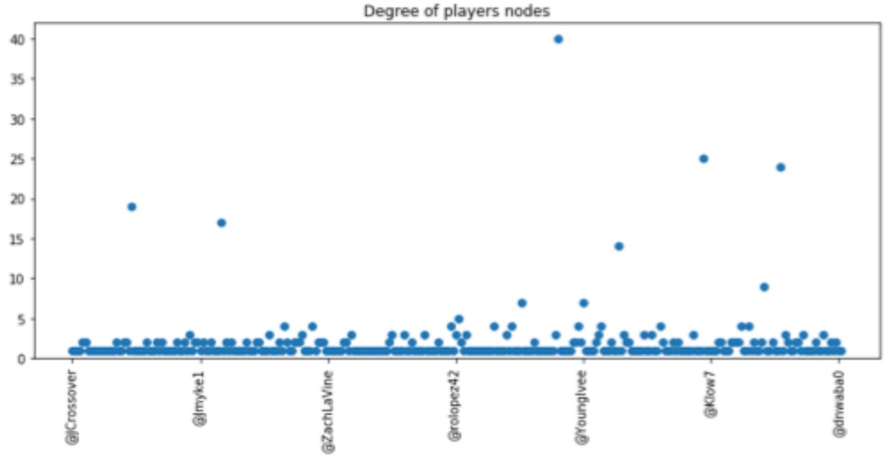
- Construction
  - Make each player twitter a node and each team twitter a node
  - For all players, check which ones of the 30 teams they follow
  - Add the edges accordingly



SECOND GRAPH – PLAYER-TEAM

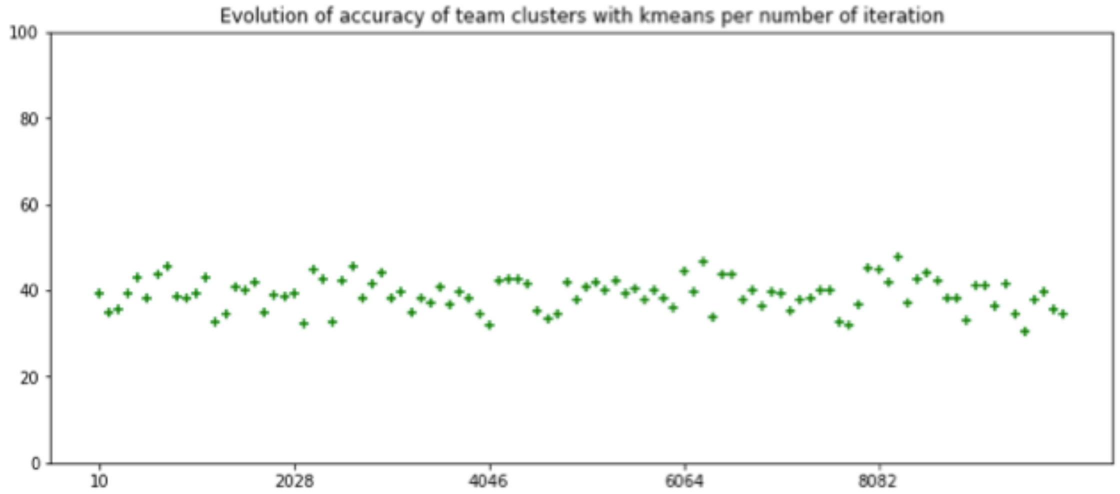
- Construction
  - Make each player twitter a node and each team twitter a node
  - For all players, check which ones of the 30 teams they follow
  - Add the edges accordingly
- Analysis
  - Each team has at least 15 followers among the players
  - Most players follow up to 3 teams





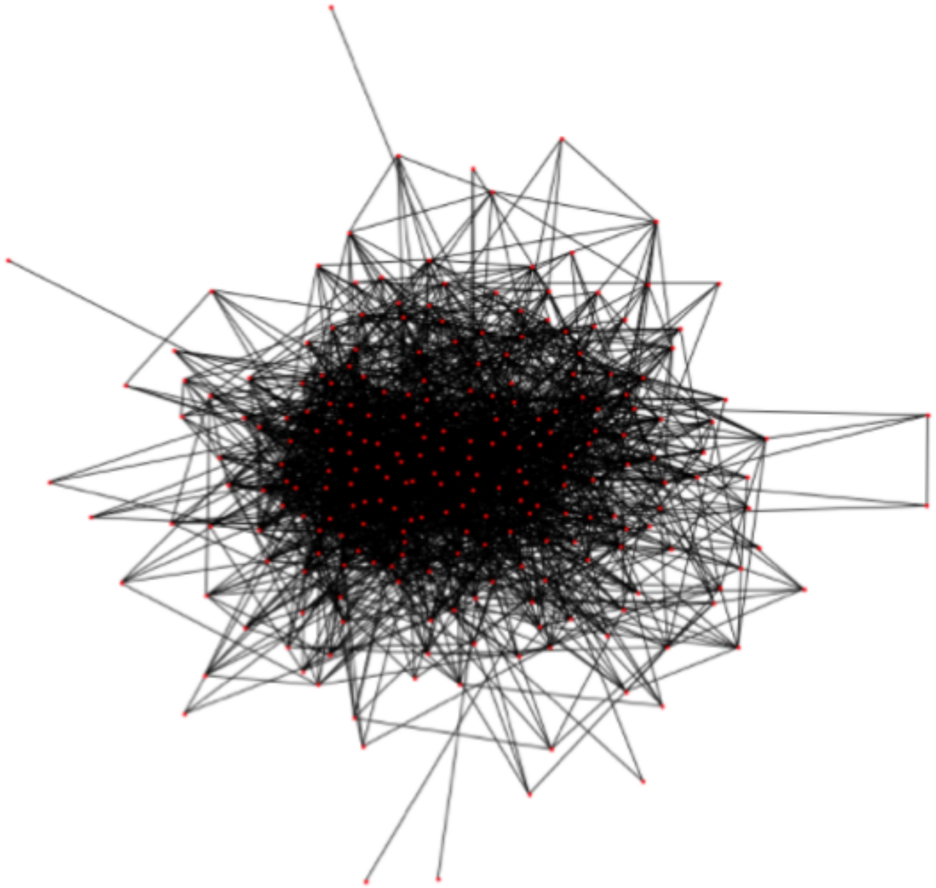
SECOND GRAPH – PLAYER-TEAM

- Construction
  - Make each player twitter a node and each team twitter a node
  - For all players, check which ones of the 30 teams they follow
  - Add the edges accordingly
- Analysis
  - Each team has at least 15 followers among the players
  - Most players follow up to 3 teams
- Clustering per Team
  - Best results with : Kmeans, 8284 iterations
  - Accuracy : 47.85 %



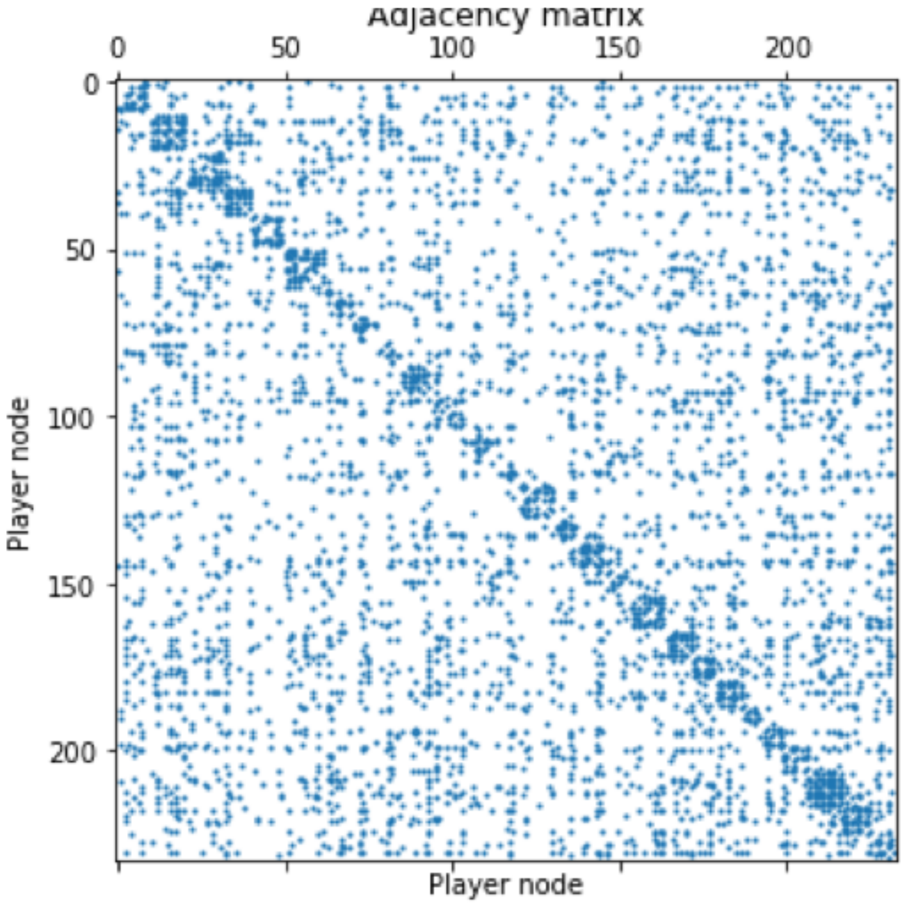
THIRD GRAPH – ROOKIE

- Construction
  - Make a sublist of players that joined a team since 2013 → 242 «Rookie» players
  - Remove the nodes of «old» players form the graph



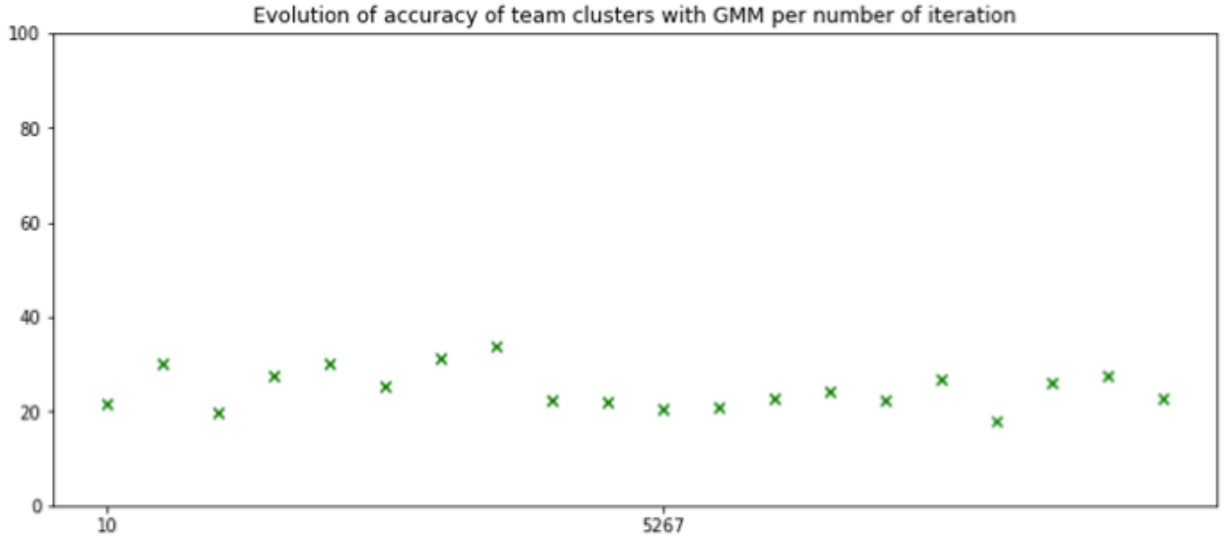
THIRD GRAPH – ROOKIE

- Construction
  - Make a sublist of players that joined a team since 2013 → 242 «Rookie» players
  - Remove the nodes of «old» players form the graph
- Analysis
  - Most degrees are under 50
  - The adjacency matrix is more sparse



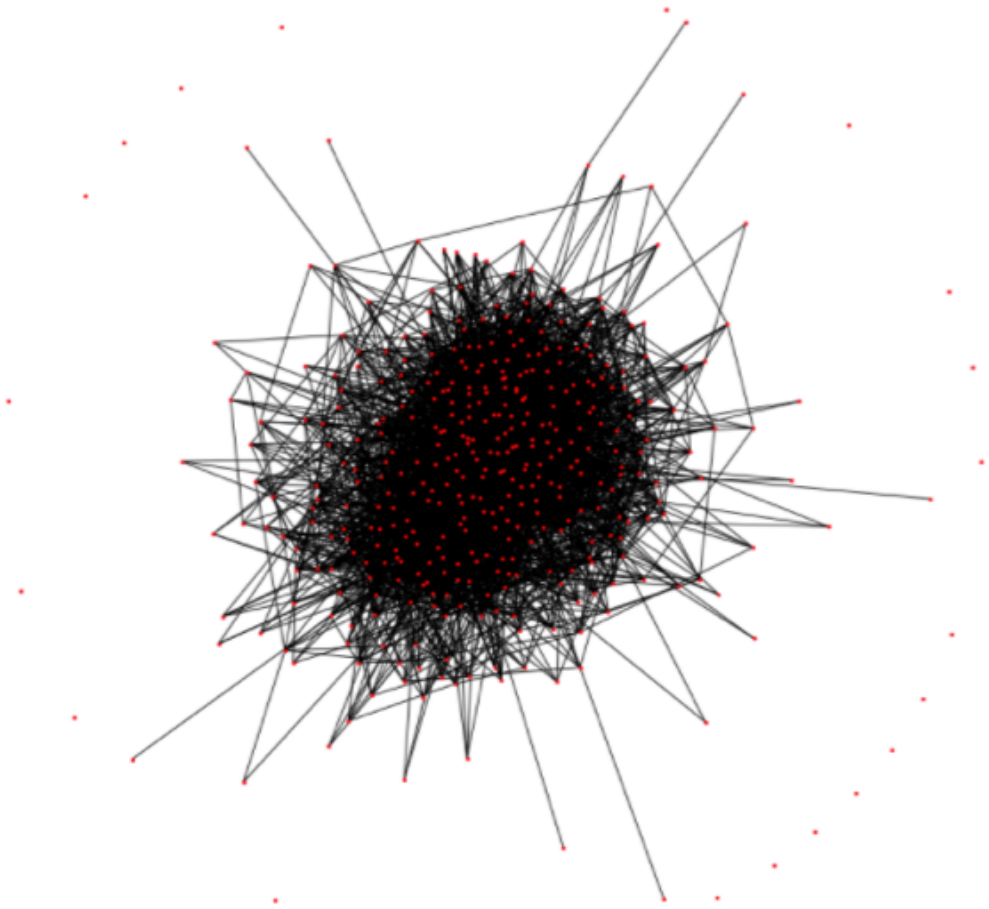
THIRD GRAPH – ROOKIE

- Construction
  - Make a sublist of players that joined a team since 2013 → 242 «Rookie» players
  - Remove the nodes of «old» players form the graph
- Analysis
  - Most degrees are under 50
  - The adjacency matrix is more sparse
- Clustering per Team
  - Best result : GMM, 3690 iterations
  - Accuracy : 33.91 %



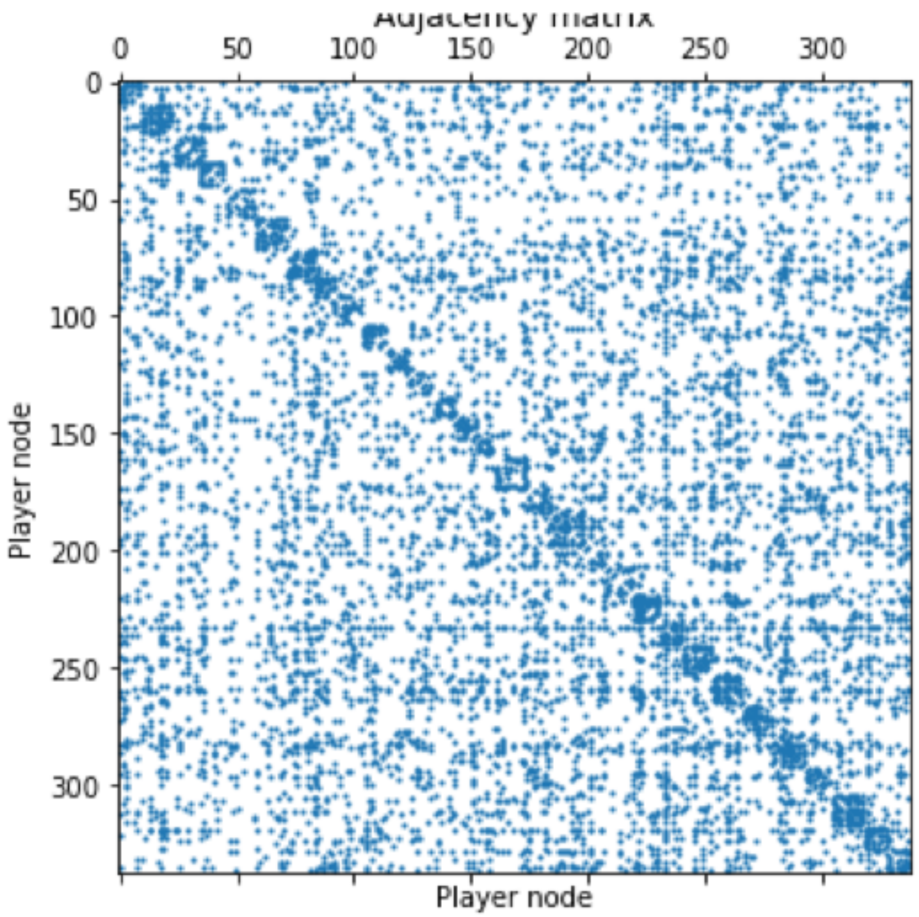
FOURTH GRAPH – DOUBLE LINK

- Construction
  - Check in the list of edges if players p1 and p2 if there is an edge (p1, p2) and (p2, p1) → draw only if they both exist



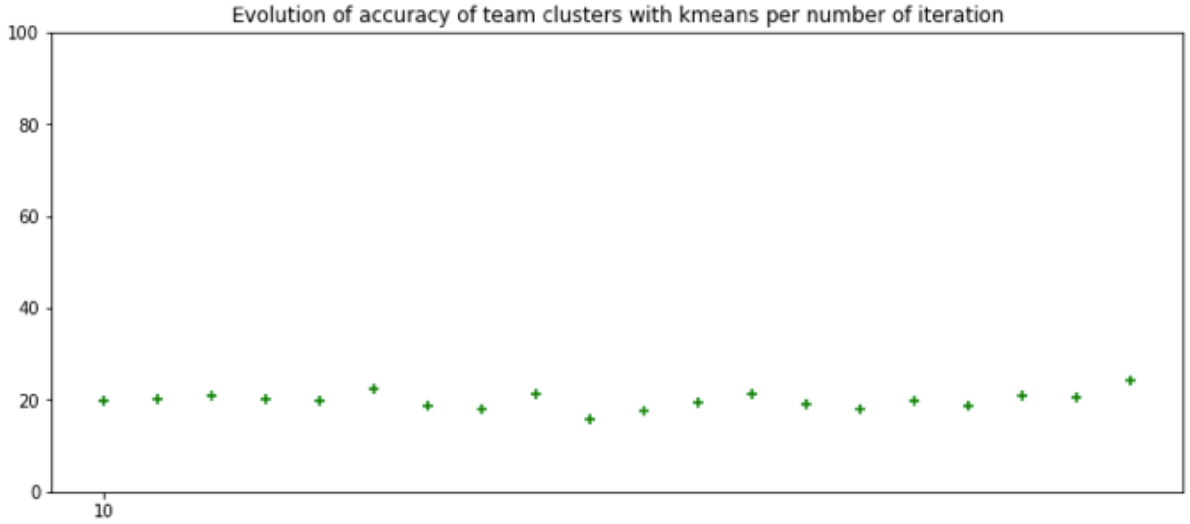
FOURTH GRAPH – DOUBLE LINK

- Construction
  - Check in the list of edges if players p1 and p2 if there is an edge (p1, p2) and (p2, p1) → draw only if they both exist
- Analysis
  - All players of biggest clique are All Star players
  - Adjacency matrix is more sparse



FOURTH GRAPH – DOUBLE LINK

- Construction
  - Check in the list of edges if players p1 and p2 if there is an edge (p1, p2) and (p2, p1) → draw only if they both exist
- Analysis
  - All players of biggest clique are All Star players
  - Adjacency matrix is more sparse
- Clustering per Team
  - Best results : Kmeans, 10000 iterations
  - Accuracy : 24.26 %



CONCLUSION

- Twitter network is insufficient for our 3 levels of clustering.
- This is visible from the adjacency matrix of the graph, the PCA visualisation and the ratio of team connections (around 14 %)
- Our main problem is that the connectivity was too high
- This is due to the fact that NBA players move often from one team to another and know each other from University League or because of their «Star status»
- The best strategy to sparsify was to take the new players
- We might want to use more information about the players as features