Homophily and Collaboration in Global Teams

Gábor Békés* and Gianmarco I.P. Ottaviano**

*CEU, KRTK KTI and CEPR; ** Bocconi University, Baffi-CAREFIN, IGIER, CEP and CEPR

CEU Brownbag // October 2021

Motivation: Globalization and homophily in teams

- Globalization - mixing the best of global expertise in multinational teams
  - Diversity benefits: learning, innovation
  - Hurdles: communication, trust

- Interaction between people of different cultural background key to understand function of teams
- Homophily is association of similar people (shared cultural background)
- What we know most: how teams are formed, tie-formation, friendship networks
- Our focus is collaboration (work for a common purpose)

- How do barriers related to nationality and language affect collaboration in multinational teams?

Motivation: Measuring homophily

- Homophily = Opportunity (**induced**) + Preference (**choice**)
  - Opportunity: mechanically induced - distributions across categories define the probability they choose similar others

- Challenge: partial out induced homophily to measure choice homophily in a setup with external validity to modern workplaces

  - Option A: With experiment / find a case with random team formation
  - Option B: With observational data / model baseline.

How we measure choice homophily: Setting

- Use professional football - top European leagues
- Collaboration can be measured by pass rate between a pair of players
- Team composition is exogenous to players' decisions, collaboration is individual choice.
- Observe collaboration and human characteristics, repeatedly in great detail

- Ideal setting:
  - Collaborative with well defined objectives and roles.
  - Rules are simple and known - allow calculating baseline
  - Global workplace: data from several countries, with players from 130 countries

How we measure choice homophily: Data and model

- Exhaustive dataset recording passing events from professional football
  - Five countries: Spain, Germany, France, Italy and England, 8 seasons
  - Players from 132 countries
  - all 10.7 million passes (origin and destination player ID, location)
  - all 7000 players' characteristics.
- Model baseline – a discrete choice model of players' passing behavior.
  - Pass rate for a pair of players is pinned down by their characteristics and opportunities during the matches
  - Estimate directly in model.

Related literature

- Same ethnicity workers collaborate better, learn less: Lazear (1999) Lang (1986)
- Diversity spillovers' - diverse environments improve team performance in cities, plants: Ottaviano and Peri (2006, 2005) Buchholz (2021)
- Diversity in teams: team performance and composition
  - Hockey: Kahane et al. (2013), Football: Nüesch and Haas (2013), Tovar (2020).
  - Ethnic conflict: Hjort (2014), Laurentsyeva (2019),
  - Team formation process, diversity and performance: Calder-Wang et al. (2021)
- Homophily in finding partners in scientific publishing: Freeman and Huang (2015), AlShebli and Woon (2018)
- Homophily in network formation in friendships: Currarini et al. (2009, 2010)
- Great deal more in psychology, finance, management (Lawrence and Shah, 2020; Ertug et al., 2021).

Contributions

1. Focus on everyday workplace collaboration - high skilled, lowly charged context in a developed area with no real conflicts.
2. Well defined measure of collaboration at individual level through time (not rare pair formation)
3. Careful model of baseline, both theory and empirics (go beyond randomization)
4. Rich and precise measures of individual characteristics
5. Very large, global sample - external validity

Preview of results

- Baseline difference is around 6.8%
- Once partialling out induced homophily, we find that pairs with a same nationality will pass 2.8% more (choice homophily)
  - Language plays some role
  - There is selection into same nationality pairs spending more time together, too.
  - Findings stronger for deep collaboration (intensive passing plays)

Data: sources

- Event data - each event separately recorded with a timestamp
  - originally created by OPTA (cameras+algorithms + humans)
  - Scraped from a sports website (whoscored.com)
  - Pass events separated
- Player information comes from scraping Transfermarkt, a player information database
  - Player characteristics: nationalities, valuations, position
- Massive data work
  - dealing with very large datasets
  - combing datasets: entity resolution / coreference

Data: scope

- 5 top leagues (France, Germany, Spain, Italy, England),
- 8 seasons (2011/12-2018/19) every teams play with every other twice
  - 20 (18) teams per league, 14,608 games in total
  - 800 passes/game/team
- 10.7 million passes in total
- 154 team, each with 25-30 strong squad, regular churning (twice a year)
- 7000 players in sample from 132 countries

Data: Same nationality definition

- 26% of players have two or three nationalities
  - Born in a country and moved to another as minor and got nationality with family (Argentina and Spain)
  - Parents have multiple nationalities (French and Algerian)
- Same nationality definition = two players have **a** common nationality
  - Example: French-Algerian dual citizenship player will have common nationality with *both* a French and an Algerian player.

Data: aggregation

- From a choice model, aggregate to relative frequencies
- Aggregate to half-seasons (16-20 games), compromise between games and seasons
  - Squads are large, only 11 players at field at once, lot of variation across games, selection major issue for a single game.
- Noise is high / randomness of games
  - There is churning in mid-season
  - Assume player quality is stable within a half-season (4 months) but may vary across half -seasons.

- Key object of interest is number of passes for player pairs per minute
  - compared to total passes for the team in a game.

Model: setup

- Football team $N = 11$ players, two players indexed o, d.
- The passer's decision = problem of passing the ball to the receiver who generates the highest expected benefit for the *team*.
- Game = series of short units of time (t) up to T ('periods').
  - Players o and d are together in the football pitch for $T^{o,d}$ periods.
  - In any t, a player is identified by his ID and position.
  - Two periods: t (current period') and $t+1$ ('future period').
- A 'pass' (o, d, t) = player o ('passer') to teammate d ('receiver'). Started by o in t, received by d in $t+1$
- Passer takes into account the current and future implications for the team's payoff.

Model: The player's decision

The passer's decision = passing the ball to the receiver who generates the highest expected benefit for the team. Benefit to have + option value.

- $\ln u_t^d$ = benefit due to player d's characteristics
- z_t^d = random part (shock')
- $\widetilde{c}^{o,d}$ = challenges – 'passing cost'
- φ^d = probability of successful pass to receiver d

$$U_t^o = \ln u_t^o + \beta \max_{\{d\}_{d=0}^N} \left\{ \varphi^d E\left[U_{t+1}^d\right] - \widetilde{c}^{o,d} + z_t^d \right\}. \tag{1}$$

Model: Passing cost

Model passing cost

$$\widetilde{c}^{o,d} = \left(g^{o,d}\right)^{\gamma} \left(I^{o,d}\right)^{\lambda} \tag{2}$$

- $g^{o,d}$ captures all distance-related frictions
- $I^{o,d}$ captures all non-distance-related frictions.
  - This is where we may see homophily - same nationality indicator
  - May be high if o and d find it hard to collaborate
  - In model, assume separability (true empirically)

Model: The player's decision 2

- The probability that player o with ball in t successfully passes to teammate d depends on
  - the (relative) value of both players
  - ability to pass/receive a pass successfully
  - cost of the pass

- Aggregation to T - probabilities to relative frequency
- Probability = the average share of successful passes that player o makes to player d per episode over a half-season
  - in the subset of time (passing episodes) $T^{o,d}$ when both o and d are fielded and player o has ball possession

Model Estimation

- Pass rate = f(player characteristics, position, friction)
- Homophily: same nationality interaction

- Poisson model for count of passes conditional on time (passes) they spent together
  - Poisson (PPML-FE) with many fixed effects has several advantages over $\ln count$ (Fally, 2015; Santos-Silva and Tenreyro, 2021)
  - Result is robust to OLS with $\ln count$

- Results roadmap
  - Core results of model estimation
  - Deeper collaboration
  - The role of managers, selection into the team
  - The role of language

Estimated models 1: Poisson with player characteristics

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|\ldots) = exp(\gamma SameNat_{p1,p2,t} + \lambda PassDist_{p1,p2,t} + \\ +1 \ln minutes\_shared + \theta_1 playerchar_{p1,t} + \theta_2 playerchar_{p2,t}) \quad (3)$$

- SameNat=Same Nationality Indicator (in $I_{p1,p2,t}$)
- PassDist=ln(Average pass distance) (in $g_{p1,p2,t}$)
- For both players
  - $playerchar$: $valuation \times half\_season$, $position \times half\_season$, $nationality \times half\_season$
  - $team \times half\_season$ dummies.

Estimated models 1: Poisson with player fixed effects

Second, we estimate a version of the Poisson model with $player_1 \times half\_season$ and $player_2 \times half\_season$ fixed effects:

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|\ldots) = exp(\gamma SameNat_{p1,p2} + \lambda PassDist_{p1,p2,t} + \\ +1 \ln minutes\_shared + \gamma^1_{p1,t} + \gamma^2_{p1,t}) \tag{4}$$

- SameNat=Same Nationality Indicator ($I_{p1,p2,t}$)
- PassDist=ln(Average pass distance) ($g_{p1,p2,t}$)
- $\gamma^1_{p1,t}$ and $\gamma^1_{p1,t}$ are $player_1 \times half\_season$ and $player_2 \times half\_season$ fixed effects

Model estimation: Results (Poisson)

Dep var: pass_count	(1)	(2)	(3)
Shared nationality (0/1)	0.0681***	0.0240***	0.0288***
	(0.0111)	(0.0062)	(0.0067)
Average length of passes (ln)		-1.033***	-1.159***
		(0.0125)	(0.0138)
Pseudo R^2	0.05837	0.75632	0.80188
FE: team * season_half	✓	✓	✓
FE: p1_features * season_half		✓	
FE: p2_features * season_half		✓	
FE: player_id1 * season_half			✓
FE: player_id2 * season_half			✓

Poisson regression model. N= 335,610. Standard errors, clustered at player 1 level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Player features: position, value, total passes, citizenship. Exposure=ln shared minutes

Result discussion

- (Unconditional) Same nationality players tend to pass 6.8% more compared to different nationality players
- Partialing out baseline homophily: it is around 2.8%
  - Robust to a variety of specifications I.- confounders/selection
    - More variables on passes (direction)
    - Physical differences
    - Assortative matching
    - Experience with club
  - Robust to a variety of specifications II.- functional form
    - Allowing minutes coefficient to vary (1.07)
    - log count as dependent variable.

Deeper collaboration

- Passing is collaboration
- Deeper collaboration = pass sequences (like ABAB)
- For deeper collaboration, trust/understanding/taste should be more important
- Instead of pass count: count of pass sequences

Model estimation: Deep collaboration

Dep. var.	Count of pass sequences			
	all_count	complex_count	all_count	complex_count
	(1)	(2)	(3)	(4)
Shared nationality (0/1)	0.0205***	0.0450***	0.0249***	0.0531***
	(0.0058)	(0.0097)	(0.0062)	(0.0103)
Average length of passes (ln)	-0.8927***	-2.079***	-1.008***	-2.412***
	(0.0116)	(0.0177)	(0.0129)	(0.0197)
Pseudo R^2	0.75035	0.56048	0.79217	0.62074
teamid-time fixed effects	✓	✓	✓	✓
FE: p1_features * season_half	✓	✓		
FE: p2_features * season_half	✓	✓		
FE: player_id1 * season_half			✓	✓
FE: player_id2 * season_half			✓	✓

Poisson regression model. N= 335,610. Standard errors, clustered at player 1 level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Exposure=ln shared minutes

Endogeneity of time spent together

- We estimated homophily effect in the model
  - Taking minutes spent together as given
- But minutes playing together on the pitch may be endogenous: coach selects teams based on expected play together.
- In this sense: time spent together is a mechanism for choice homophily to effect collaboration
- Minutes played together is also a function of shared nationality
- Homophily premium is 3.8% once minutes spent together is not partialled out.

Model estimation: endogeneity of time together

	pass_count (1)	minutes_shared (2)	pass_count (3)
Shared nationality (0/1)	0.0288***	0.0089***	0.0370***
	(0.0067)	(0.0030)	(0.0076)
Average length of passes (ln)	-1.159***		-1.261***
	(0.0138)		(0.0146)
Pseudo R^2	0.80188	0.88506	0.74159
FE: player_id1 * season_half	✓	✓	✓
FE: player_id2 * season_half	✓	✓	✓

Poisson regression model. N= 335,610. Standard errors, clustered at player 1 level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Exposure=ln shared minutes

Nationality or language?

- Maybe it's all about communication: language may be crucial
- Common language if there exists a common official (or widely spoken) language
- Identification: non national, but same language speaker

Model estimation: Language

	pass_count (1)	minutes_shared (2)	pass_count (3)
Shared nationality (0/1)	0.0326***	0.0107***	0.0418***
	(0.0071)	(0.0031)	(0.0080)
Shared only language (0/1)	0.0167*	0.0072**	0.0208**
	(0.0087)	(0.0035)	(0.0098)
Average length of passes (ln)	-1.159***		-1.261***
	(0.0138)		(0.0146)
Pseudo R^2	0.80189	0.88506	0.74160
FE: player_id1 * season_half	✓	✓	✓
FE: player_id2 * season_half	✓	✓	✓

Poisson regression model. N= 335,610. Standard errors, clustered at player 1 level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Player features: position, value, total passes, citizenship. Exposure=ln shared minutes

Other issues (+pipeline)

- Heterogeneity
  - Largest impact when *both* players are young
  - Effect does not go away with experience in team
  - Effect seems higher for better players
  - League rules (EU, non-EU) matter, especially in Spain.
  - No interaction between pass type (distance) and same nationality
- Pipeline
  - Allow language to be learnt over time
  - More on experience (in league, etc)

Results summary

- Evidence of homophily: player pairs of same nationality pass more
  - More likely engaged in deeper collaboration
  - Shared language has about half the impact of same nationality.
- Additional channel exists via how frequently they play together
- Homophily is pervasive even in teams of
  - very high skill individuals
  - with clear common objectives and aligned incentives
  - and involved in well defined tasks
  - activities not particularly language-intensive.