Foreign inventors in Europe and the US: Testing for Self-Selection and Diaspora Effects

Stefano Breschi[1], Francesco Lissoni[2,1]

[1] *CRIOS – Università Bocconi, Milano ;* [2] *GREThA – Université Montesquieu, Bordeaux IV*

MEIDE Conference 2013, Santiago de Chile – November 7-8, 2013

Motivation

- Steady increase in the global flows of scientists and engineers (S&Es) over the past 20 years…
  - in absolute terms
  - as a % of total migration flows *(Docquier & Rapoport, 2012)*
- Open questions on impact on innovation in:
  - Destination countries → migrant S&Es' skills relative to locals *(Stephan&Levin 2001, Hunt 2011, Walsh&Nagaoka, 2009)*
  - Origin countries: brain drain *or* brain gain? → role of "diasporas" *(Wadhwa et al., 2007a-c; Foley and Kerr, 2013)*
- Lack of data for micro-econometric analysis
- Prevalent focus on:
  - one destination country (the US) …
  - and its most recent origin countries: India & China

Paper's objectives

- To explore the potential of **inventor data** to address immigration & innovation key issues, such as:
- For destination countries:
  - Are migrant S&Es self-selected on (inventive) skills?
  - How many and how skilled are migrant S&Es to Europe?
- For origin countries:
  - Do migrant S&Es represent a real "diaspora" (socially connected community)?
  - If yes, does this help accessing technological information?

Outline

1) The pilot ***Ethnic-Inv*** database
   - EP-INV database on "disambiguated inventors" (from APE-INV project: http://www.academicpatenting.eu)
   - IBM-GNR database on names&surnames' "ethnicity"

2) Testing for diaspora /1: ethnicity and "closure" in network of inventors *[exploratory → do not cite]*

3) Testing for diaspora/2: ethnicity and patent citations *[exploratory → do not cite]*

The ETHNIC-INV database

INVENTORS' COUNTRY OF ORIGIN: 2 strategies for data collection

1. WIPO/PCT: information on inventors' ***nationality***
   (Fink and Miguelez, 2013 ; Miguelez, 2013)

2. ETHNIC MATCHING: assign inventors to a "country of origin" (or a "meta-country" of origin) on the basis of ***linguistic analysis of names & surnames***
   i. Kerr, 2007 (limitation: US-centric)
   ii. Agrawal et al., 2008 & 2011 (limitation: US-Indian centric)

- We go for improving 2. !!!
- Soon merge with 1. (thanks Ernest!)
- Key technical issue: DISAMBIGUATION

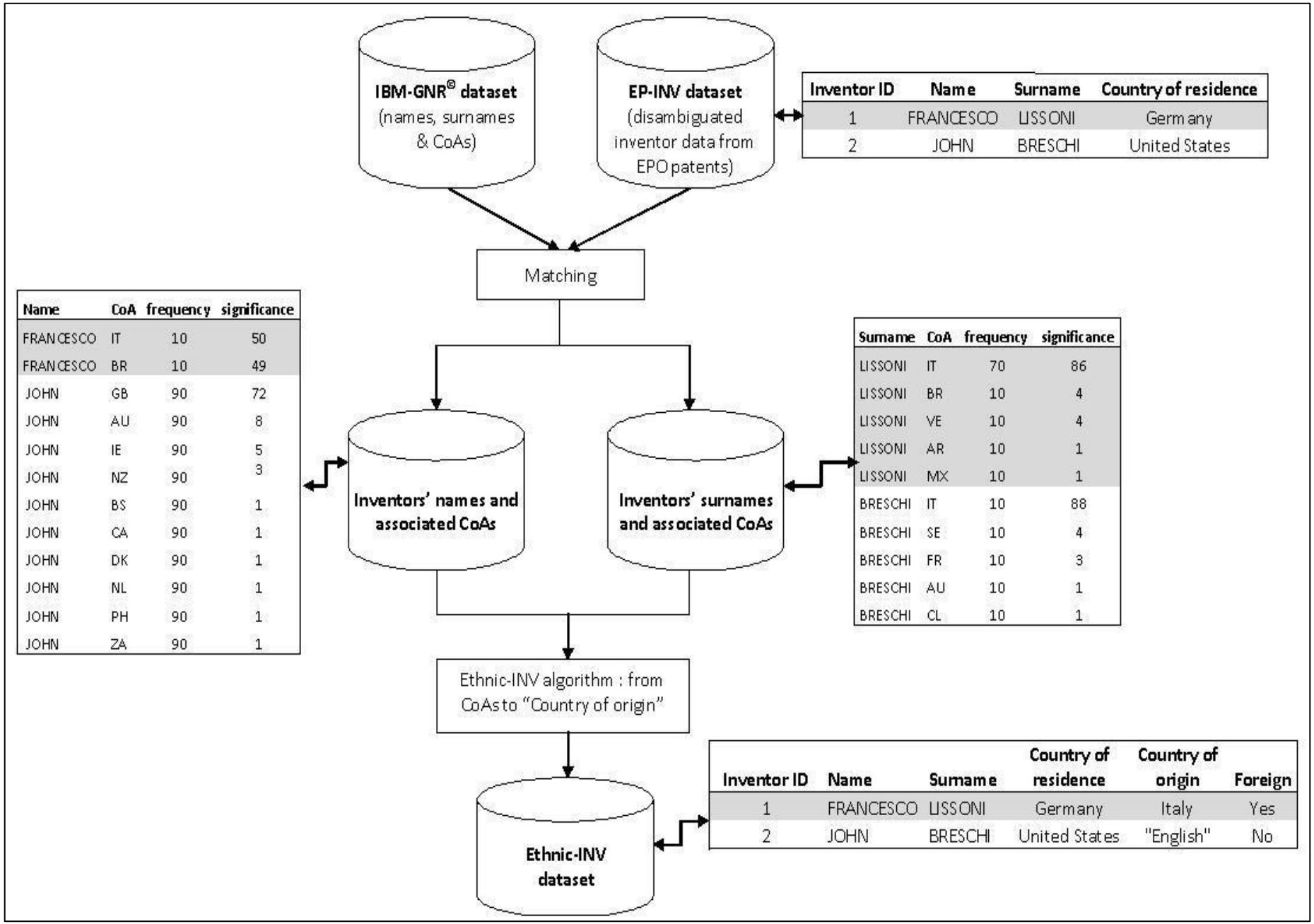
The ETHNIC-INV database:

- **EP-INV database**: ≈3 million uniquely identified inventors from EPO patents (1978-2009; Patstat 10/2011 edition)

+

- **IBM Global Name Recognition (GNR)** system: 750k full names + computer-generated variants → For each name or surname:

  1. (long) list of "countries of association" (CoAs) + statistical information on cross-country and within-country distribution
  2. elaboration on (1) with our own algorithms



Inventor ID	Name	Surname	Country of residence
1	FRANCESCO	LISSONI	Germany
2	JOHN	BRESCHI	United States

Name	CoA	frequency	significance
FRANCESCO	IT	10	50
FRANCESCO	BR	10	49
JOHN	GB	90	72
JOHN	AU	90	8
JOHN	IE	90	5
JOHN	NZ	90	3
JOHN	BS	90	1
JOHN	CA	90	1
JOHN	DK	90	1
JOHN	NL	90	1
JOHN	PH	90	1
JOHN	ZA	90	1

Surname	CoA	frequency	significance
LISSONI	IT	70	86
LISSONI	BR	10	4
LISSONI	VE	10	4
LISSONI	AR	10	1
LISSONI	MX	10	1
BRESCHI	IT	10	88
BRESCHI	SE	10	4
BRESCHI	FR	10	3
BRESCHI	AU	10	1
BRESCHI	CL	10	1

Inventor ID	Name	Surname	Country of residence	Country of origin	Foreign
1	FRANCESCO	LISSONI	Germany	Italy	Yes
2	JOHN	BRESCHI	United States	"English"	No

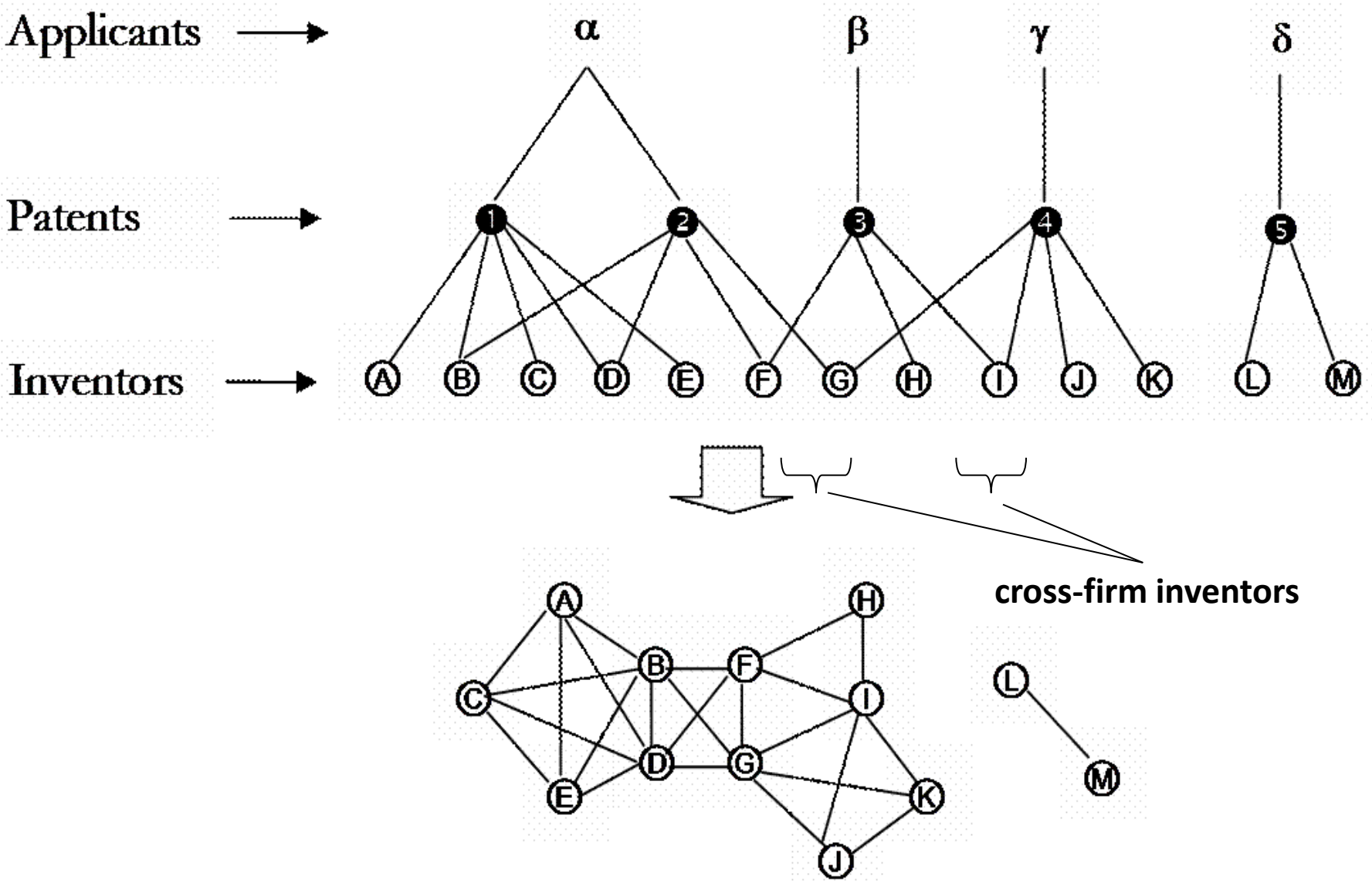
In the paper…

- Inventor-based figures are in the same order of magnitude as those for highly skilled migration BUT some problems for the European-language groups in US (→ over-estimation of migrants)

- Relevance of immigrant inventors in Europe, not just the US
  1) Role of intra-European migration
  2) High productivity of foreign inventors
  3) These results hold also for specific countries of origin, BUT they are weaker for countries
     - ✓ where highly skilled migration is traditionally low
     - ✓ with large ethnic minorities

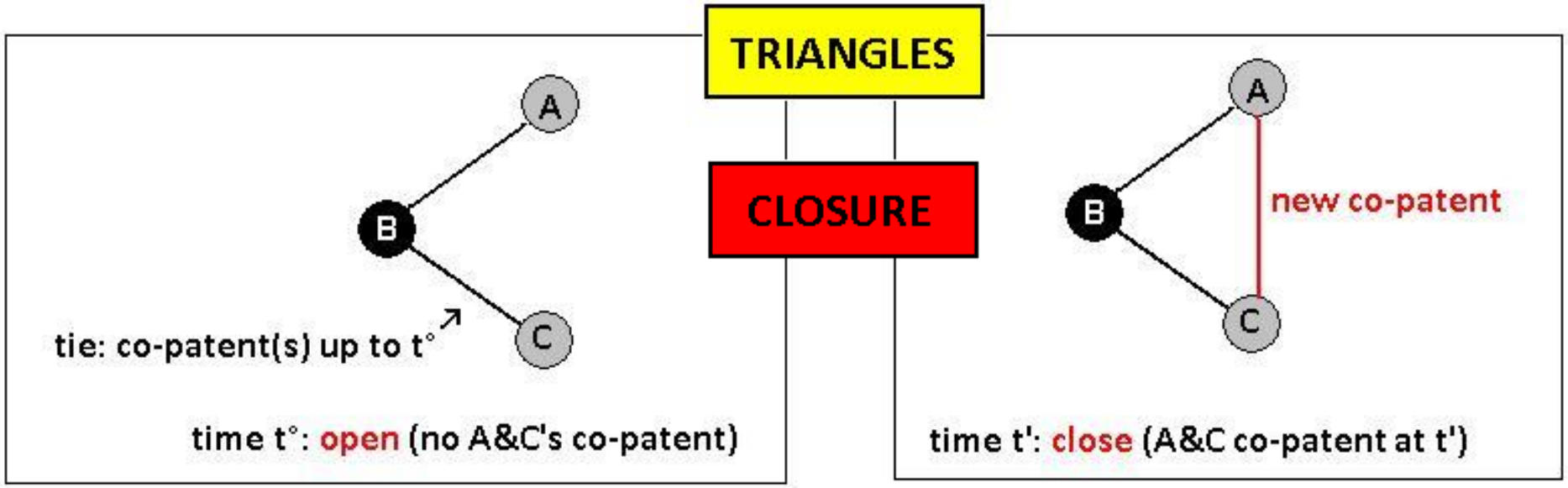
Testing for diaspora /1: ethnicity and “closure” in network of inventors

Network of inventors



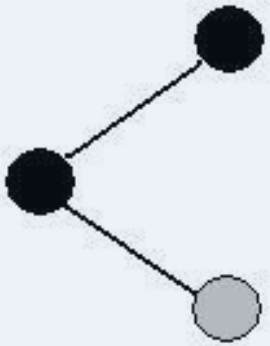
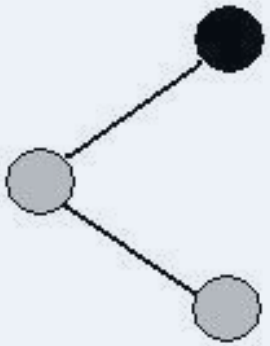
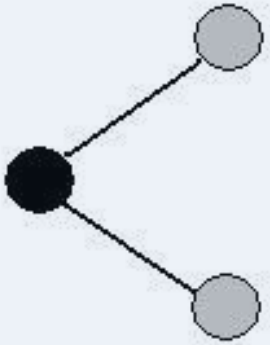
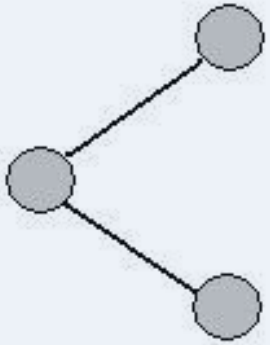
DIASPORA as “ethnic homophily” in the social network:

→ Probability(“closure”) = f (co-ethnicity, nr of paths)

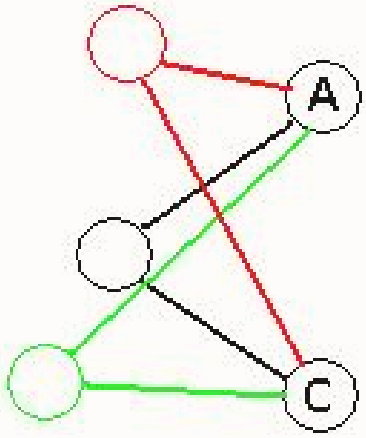


gray nodes = foreign inventors (same country of origin)

black nodes = local inventors or foreign inventors from different countries than grey's

4 cases				
End nodes :	No-ethnic	No-ethnic	Ethnic	Ethnic
Broker:	No-ethnic	Ethnic	No-Ethnic	Ethnic

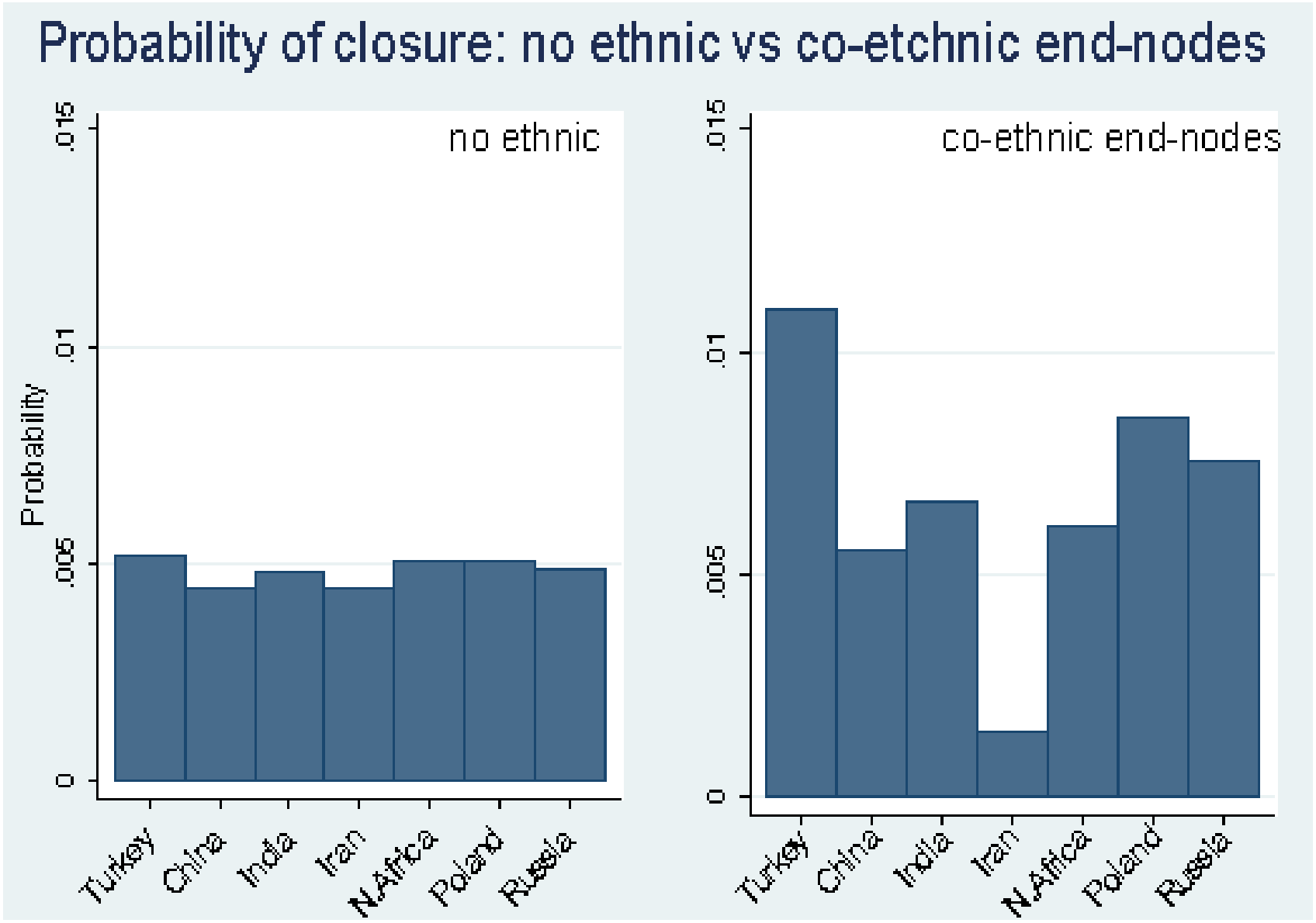


NB: A & C at t° may belong to several triangle
(be connected by multiple paths)

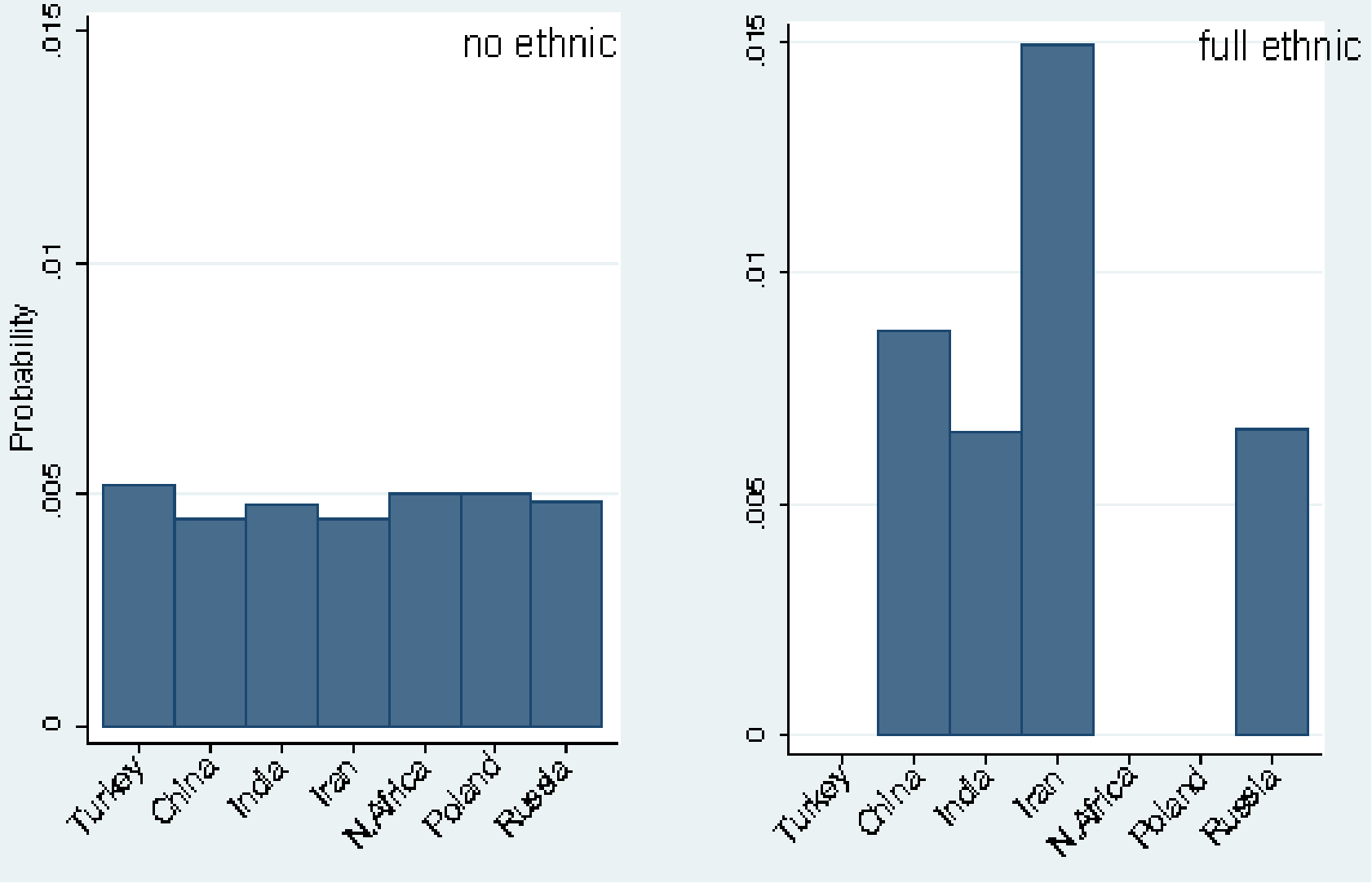
Table 9 - Probability of triadic closure, by country of origin (country of destination: US; years: 1985-2005) - LOGIT regression

	(1)	(2)
Nr of triangles	0.0516***	0.0521***
(1) Dummies for co-ethnicity in triangles (no co-ethnicity as ref.)		
End-Tie	0.263***	
Broker	0.391***	
Full	0.551***	
(2) Dummies for countries of origin (Turkey as reference)		
China	-0.164**	
India	-0.160**	
Iran	-0.233***	
N.Africa	-0.0940	
Poland	-0.0841	
Russia	-0.105	
*Interactions (1) * (2)*	N	Y
Constant	-5.306***	-5.347***
Observations	3,415,713	3,413,660
Pseudo - R2	0.01	0.01
Log likelihood	-111273	-111208
DF	10	24
Chi2	2501	2609

*(*** p<0.01, ** p<0.05, * p<0.1)*


Probability of closure: no ethnic vs co-etchnic end-nodes
no ethnic
co-ethnic end-nodes
Probability
0
.005
.01
.015
Turkey
China
India
Iran
N.Africa
Poland
Russia

Probability of closure: no ethnic ties vs full ethnic triangle



Testing for diaspora /2: ethnicity and patent citations

Do “ethnic tie” favour access to knowledge?

Probability (co-ethnic citation) =
= f (co-ethnic controls, spatial & social distance)

CASES: citing-cited patent pairs, where citing patents include at least one “ethnic” inventor (excluding self-citations)

CONTROLS: one for each cited patent →random draw:
- same priority year
- same IPC classification (12 digit)
- excluding patents by the same inventor

SOCIAL DISTANCE btw patents: MIN(social distance between their inventors)

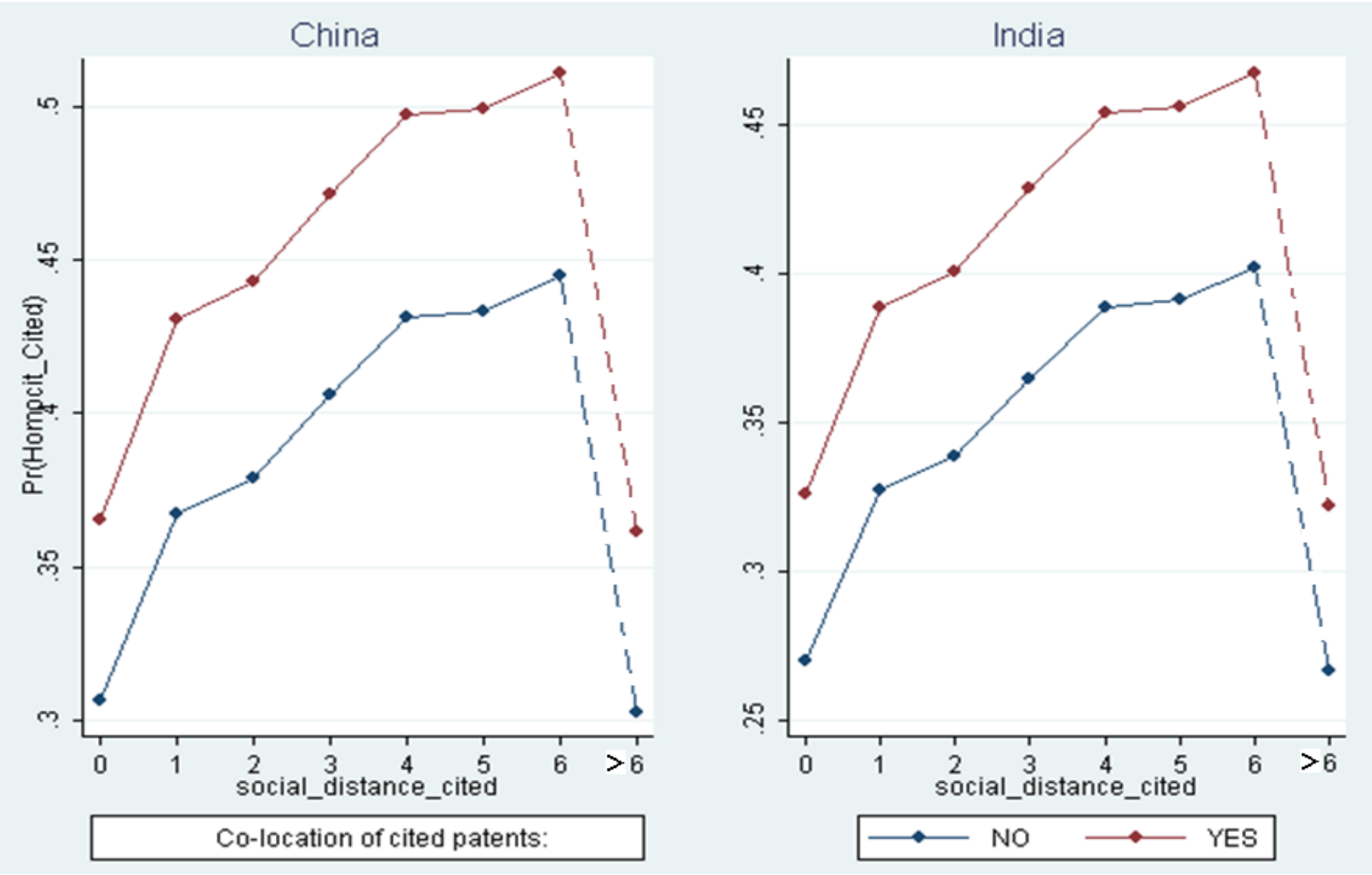
PHYSICAL DISTANCE btw patents: MSA CO-LOCATION of at least one inventor per patent

Probability of homophily in citations → LOGIT regressions (est. coefficients ; US as destination country)

Control is co-ethnic	1.246***
Countries of Origin (Turkey as reference)	
China	2.869***
India	2.881***
Iran	0.968***
North-African	1.097***
Poland	1.446***
Russia	1.450***
Cited patent is co-located	0.266***
Control is co-located	-0.0557***
Social distance of cited patent (>6 as reference)	
0	0.0199
1	0.293***
2	0.340***
3	0.457***
4	0.568***
5	0.572***
6	0.621***
Social Distance of Controls (>6 as reference)	Y
Nr controls	0.000236***
Constant	-5.057***
Observations	410,907
Log likelihood	-129926
DF	24
Chi2	27462

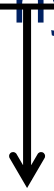
*** p<0.01, ** p<0.05, * p<0.1

Estimated probability of homophily in citations, as function of: social distance & co-location (CHINESE & INDIAN inventors)



Conclusions

- We see a potential of “ethnic” inventor data
  - Measurement of brain circulation phenomena
  - Test of “di~~aspora” hypotheses~~



- “Closure” & Citation patterns suggest:
  1) Different impact of co-ethnicity by country of origin
  2) Ethnic and social (co-inventorship) ties are substitutes, with social ties bearing a stronger effect
  3) Ethnic tie and spatial distance (co-location) are substitutes

Further research

- Refine/Extend the “country of origin” algorithm
- Extend analysis to USPTO and PCT data
- Experiment with alternative “country of association” data besides IBM/GNR
- Collect information for samples of inventors
- Split the paper in N=nr research questions

	Countries of association						
Thank you	UK	US	AU	CA	BA	IE	…
Obrigado	BR	PT	…				
Grazie	IT	..					
Merci	FR	CA(Q)	BE(W)	CI	…		

BACK-UP SLIDES

DISAMBIGUATION

In a nutshell:

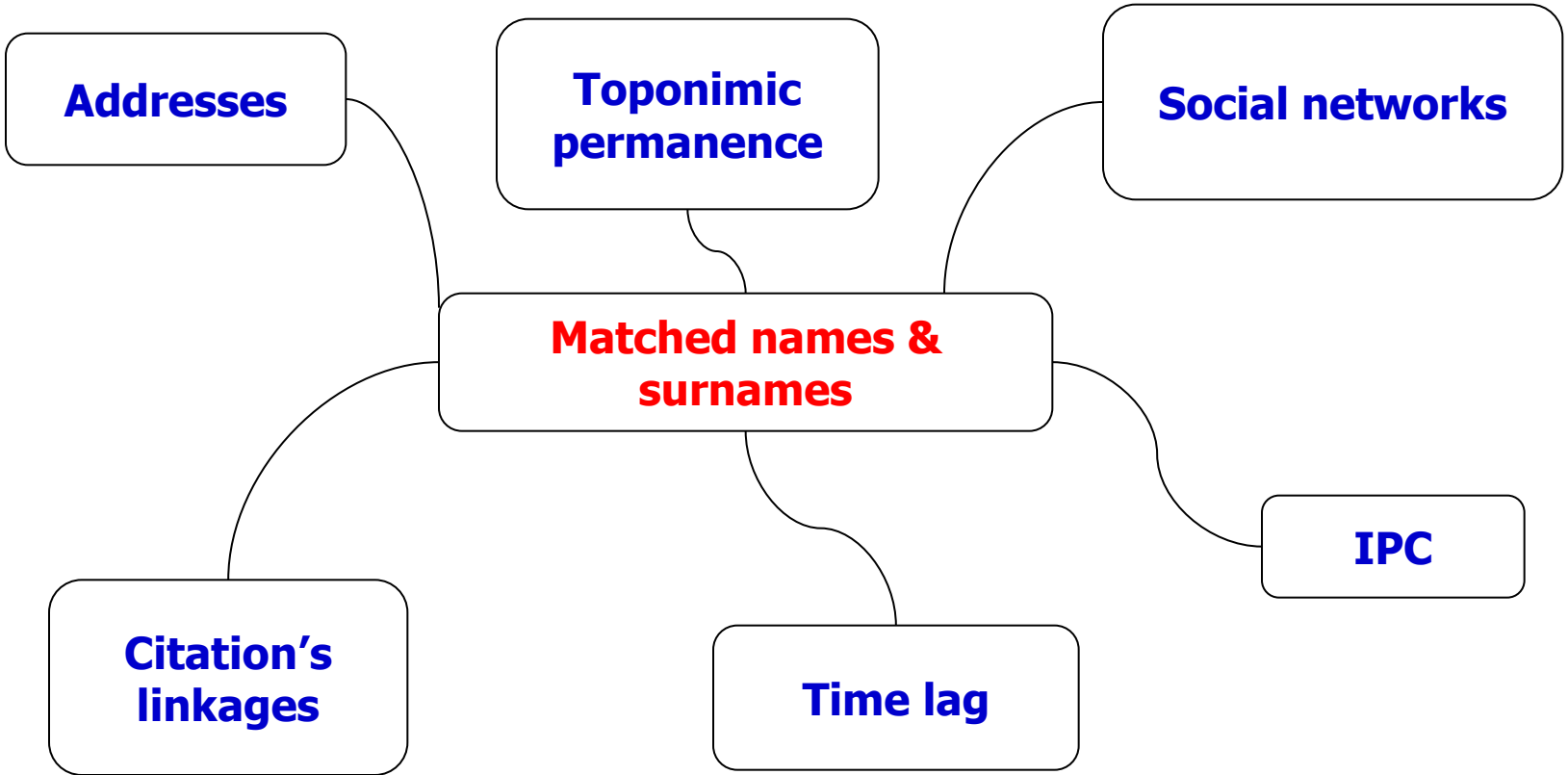
FULL NAME	Address	CY	Unique IDs…?	
David John Knight	3 PeachTree Rd, Atlanta GA	US	1	1
David John Knight	12 Oxford Rd, Manchester	UK	2	1
David J. Knight	Georgia Tech Campus	US	1	1
Knight David John	3 PeachTree Rd, Atlanta GA	US	1	1

DISAMBIGUATION STEPS:

1) Matching by name & surname:

INNAME	INNAME
KNIGHT DAVID JOHN	KNIGHT JOHN D.
KNIGHT DAVID JOHN	KNIGHT D. J.
KNIGHT DAVID JOHN	KNIGHT DAVID M.
KNIGHT DAVID JOHN	KNIGHT JOHN
KNIGHT DAVID JOHN	KNIGHT DAVID L.
KNIGHT DAVID JOHN	KNIGHT DAVID

2) Filtering by information-from-patent:



→Trade-offs between "precision" and "recall"

$$Precision = \frac{tp}{tp+fp} \qquad Recall = \frac{tp}{tp+fn}$$

where:

$$\begin{cases} tp = number\ of\ true\ positives \\ tn = number\ of\ true\ negatives \\ fp = number\ of\ false\ positives \\ fn = number\ of\ false\ negatives \end{cases}$$

→Precision and Recall vary by ethnic group (linguistic rules, naming conventions, frequency of names and surnames)

E.g.: East-Asians → low precision/high recall

Russians → high precision/low recall

Table 1: Inventors* in the Ethnic-Inv database, by country of residence (selected countries only)

	Nr	%
Austria	16,608	0.9
Belgium	20,499	1.2
Denmark	14,103	0.8
Finland	17,433	1.0
France	114,254	6.4
Germany	252,823	14.3
Great Britain	86,219	4.9
Italy	47,318	2.7
Netherlands	46,943	2.7
Spain	17,100	1.0
Sweden	31,617	1.8
Switzerland	35,510	2.0
Japan	504,431	28.4
South Korea	42,690	2.4
US	526,850	29.7
Total	*1,774,398*	*100*

** Inventors active in n>1 countries are counted n times*

Table 3: Inventors of foreign origin as % of resident inventors: estimates from WIPO-PCT and Ethnic-Inv (selected countries of destination and origin*)

	Foreign *nationals* as % of resident inventors (WIPO-PCT, 1991-2010)	Foreign *origin* inventors as % of residents (Ethnic-Inv, 1985-2005); by calibration of the Ethnic-INV algorithm §		
	(1)	(2)	(3)	(4)
France	1.71	2.26	2.92	2.58
Germany	1.54	2.10	2.52	2.31
Great Britain	2.21	3.34	3.97	3.67
Netherlands	2.89	3.47	3.94	3.68
United States	7.02	12.43	15.24	13.79

* Arabic (meta), Chinese (meta), Indian (meta), Iran, Poland, Romania, Russian (meta), Turkey, Vietnam

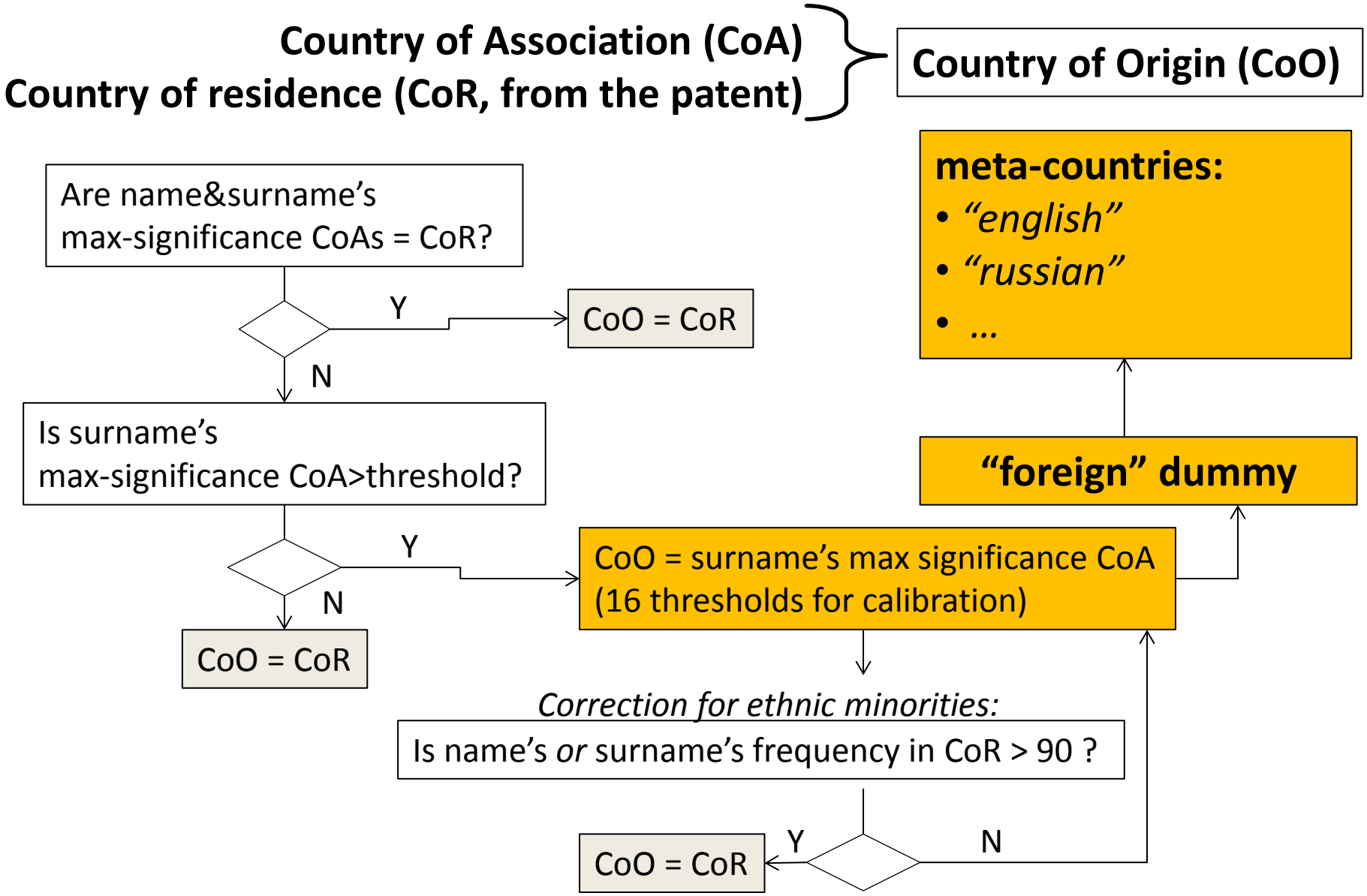
§ Column (2): max precision ; Column (3) : balanced; Column (3) : max recall [see section 4.1.2 for details]

Foreign origin and outstanding productivity: Logit regression on cohorts of immigrants (dep. variable: inventor's probability to fall in top 5% of distribution by nr of patents; ORs)

	(1) US	(2) Europe§	(3) Germany	(4) France	(5) UK	(6) Netherl.
Chinese	1.520***	1.418***	1.271	1.089	1.525**	1.673**
Iran	1.514***	1.208	1.093	0.827	0.934	2.514
Poland	1.187**	1.290**	1.090	0.778	1.456	2.614**
Romania	1.367*	1.582**	1.548	0.554	3.209**	5.994***
Russian	1.292***	1.450***	1.829***	1.165	2.020***	2.079***
Turkey	1.856***	1.072	1.005	1.603	1.409	1.256
Indian	1.561***	1.436***	1.335	1.380	1.213	2.586***
Arabic	1.617***	1.243*	1.604	0.914	1.178	2.431*
Other_foreign	1.150***	1.246***	1.115***	1.225***	1.498***	1.256***
Year controls	y	y	y	y	y	y
Technology controls	y	y	y	y	y	y
Constant	0.00494***	0.00443***	0.00456***	0.00390***	0.00380***	0.00335***
Observations	526,411	699,944	252,644	114,193	86,178	46,908

§Europe= Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Spain, Sweden, Switzerland, UK

Use IBM-GNR information:



Calibration of the Ethnic-INV algorithm (country level)