From Migration Corridors to Clusters:

Using Google+ Data for Migration Studies

Johnnatan Messias[1] Fabricio Benevenuto[1]
Ingmar Weber[2] Emilio Zagheni[3]

[1]Universidad Federal de Minas Gerais, Belo Horizonte

[2]Qatar Computing Research Institute, Doha

[3]University of Washington, Seattle

eScience lightning talks
UW, Seattle, Tue, May 31, 2016

Data about international migration are bad

Different degrees of 'badness of data'

- Stocks of migrants $\rightarrow$ based on census data, **not-too-bad**

Different degrees of 'badness of data'

- Stocks of migrants → based on census data, **not-too-bad**
- Flows of migrants → come from surveys, registration systems or indirect methods, but often inconsistent → **pretty bad**

Different degrees of 'badness of data'

- Stocks of migrants → based on census data, **not-too-bad**
- Flows of migrants → come from surveys, registration systems or indirect methods, but often inconsistent → **pretty bad**
- Migration histories (residential history for the same group of individuals over time) → **practically inexistent**

Web data can complement existing data sources

- Most of the work in this area has dealt with improving estimates of flows:
  - Geolocated Yahoo! logins (Zagheni, Weber and State 2012,2013)
  - Geolocated Twitter data (Hawelka et al. 2014; Zagheni et al. 2014)
  - Professional histories of LinkedIn users (State et al. 2014)
  - Skype calls and networks (Kikas et al. 2015)
  - Facebook logins (Hofleitner, Ruths et al.)
  - ...

Web data can complement existing data sources

- Most of the work in this area has dealt with improving estimates of flows:
  - Geolocated Yahoo! logins (Zagheni, Weber and State 2012,2013)
  - Geolocated Twitter data (Hawelka et al. 2014; Zagheni et al. 2014)
  - Professional histories of LinkedIn users (State et al. 2014)
  - Skype calls and networks (Kikas et al. 2015)
  - Facebook logins (Hofleitner, Ruths et al.)
  - ...
- One of the goals of my research is to combine traditional and new data sources within a solid statistical framework (see poster)

Web data can complement existing data sources

- Most of the work in this area has dealt with improving estimates of flows:
  - Geolocated Yahoo! logins (Zagheni, Weber and State 2012,2013)
  - Geolocated Twitter data (Hawelka et al. 2014; Zagheni et al. 2014)
  - Professional histories of LinkedIn users (State et al. 2014)
  - Skype calls and networks (Kikas et al. 2015)
  - Facebook logins (Hofleitner, Ruths et al.)
  - ...
- One of the goals of my research is to combine traditional and new data sources within a solid statistical framework (see poster)

$\Rightarrow$ For this paper, the focus is on pseudo-migration histories of Google+ users and on how countries are clustered together by migration flows in different ways

Google+ Data Set

- Data originally collected by Gabriel Magno in 2012 to study gender differences in online social networks
- We considered the Google+ field ("Places where I lived") mapped to countries
- We used the subset of users who have lived in at least 2 countries (n $\approx$ 1.6 million users). 270,000 users have lived in 3 countries.

No obvious relationship between pairs of countries and triples of countries people have lived in

		Countries Lived In				Bilateral Flows
		A	B	C	D	
Scenario 1	M1	x	x	x		(A,B), (A,C), (B,C)
	M2	x			x	(A,D)
	M3		x		x	(B,D)
	M4			x	x	(C,D)
Scenario 2	M1		x	x	x	(B,C), (B,D), (C,D)
	M2	x	x			(A,B)
	M3	x		x		(A,C)
	M4	x			x	(A,D)

Illustrative example: a) More people have lived in three countries than expected from bilateral flows

$$\overbrace{\underbrace{\text{India - Singapore;}}_{7{,}665}\ \underbrace{\text{India - UAE;}}_{17{,}584}\ \underbrace{\text{Singapore - UAE;}}_{1{,}970}}^{805}$$

- Baseline model:

$$\text{Ranking for} freqABC \approx$$
$$min(freqAB, freqAC, freqBC)$$
$$\times mean(freqAB, freqAC, freqBC)$$

- Expected ranking for people who have lived in the 3 countries based on bilateral flows of Google+ users = # 682
- Actual ranking in Google+ data set = # 200

Illustrative example: b) Less people have lived in three countries than expected from bilateral flows

$$\overbrace{\underbrace{\text{Brazil - USA;}}_{46{,}784}\ \underbrace{\text{Mexico - USA;}}_{67{,}065}\ \underbrace{\text{Brazil - Mexico;}}_{14{,}593}}^{1{,}386}$$

- Expected ranking for people who have lived in the 3 countries based on bilateral flows of Google+ users = # 12
- Actual ranking in Google+ data set = # 80
- Conditional on the bilateral flows, we would have expected more users to have lived in all the three countries

Discussion

- Countries with similar bilateral flows may experience quite different dynamics as a migration system... Why?
- New types of data:
  - Large samples (important for rare phenomena like migrations)
  - Qualitatively different information (migration histories vs flows)
  - Push for new theories
- Limitations:
  - Google+ users are a non-representative sample
  - Data quality issues

$\Rightarrow$ Combining traditional and new data sources is key