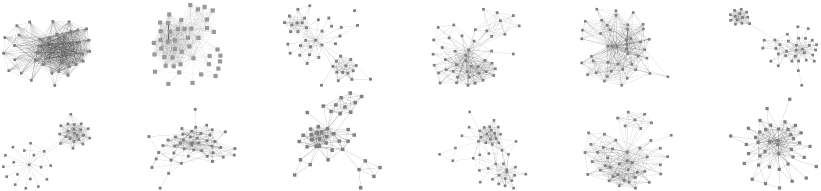
Visualizing Tendency and Dispersion in Collections of Attributed Networks

Jürgen Lerner   Ulrik Brandes   Miranda J. Lubbers
Chris McCarty   José Luis Molina

U. Konstanz   Autonomous U. Barcelona   U. Florida

XXVIII Sunbelt'08   St. Pete Beach   January 22–27, 2008

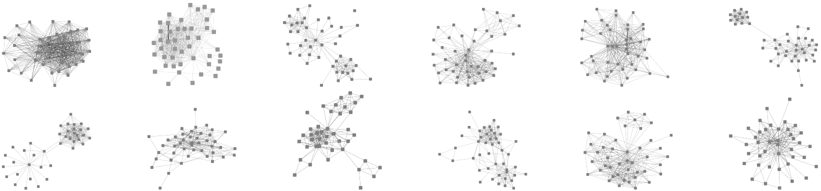
Suppose we have a sample of networks (hundreds).



- What is the **average** network of this sample?
- What is the **variability** within this sample?
- Do sub-samples have different averages?

Present a method for **visual exploration** of **collections of networks**; showing trends (statistical average) and dispersion (statistical variability).

Suppose we have a sample of networks (hundreds).



- What is the **average** network of this sample?
- What is the **variability** within this sample?
- Do sub-samples have different averages?

Present a method for **visual exploration** of **collections of networks**; showing trends (statistical average) and dispersion (statistical variability).

Illustrate this approach on a concrete application.

Understanding the **acculturation** of migrants by analyzing their personal networks.

(**Acculturation**: outcome of cultures coming into contact.)

Empirical data set by interviewing $\approx$ 500 migrants in Cataluña (Spain) and Florida (USA).

From each respondent (**ego**) we got

1. **(questions about ego)** country of origin, years of residence, skincolor, health, language skills, ...
2. **(name generator)** list of 45 alters
3. **(questions about alters)** from, lives, skincolor, ...
4. **(ties)** which alters know each other

Here: present visual exploration of this set of 500 networks.

See `www.egoredes.net` for other work using this dataset.

Reduce networks of individuals to networks of classes.

Two steps

1. Define actor **classes** based on selected attributes.
2. Define inter-class and intra-class **ties** (how strongly are two classes connected?).

Benefits

- Reduction in **size**; small but informative images.
- Enables simple and efficient **comparison** between disjoint networks.
- Allows for **averaging** over collections of networks.

Reduce networks of individuals to networks of classes.

Two steps

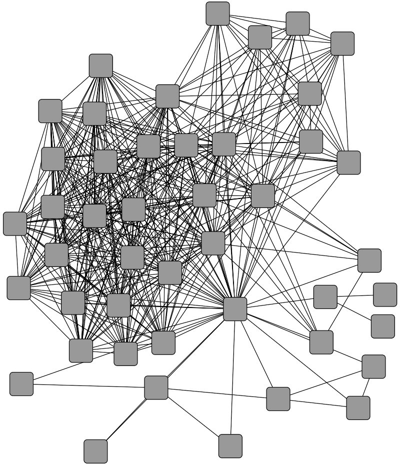
1. Define actor **classes** based on selected attributes.
2. Define inter-class and intra-class **ties** (how strongly are two classes connected?).

Benefits

- Reduction in **size**; small but informative images.
- Enables simple and efficient **comparison** between disjoint networks.
- Allows for **averaging** over collections of networks.

Defining the actor classes.

Case of an Argentinean migrant in Spain.



where are the alters originally from? where do they live?

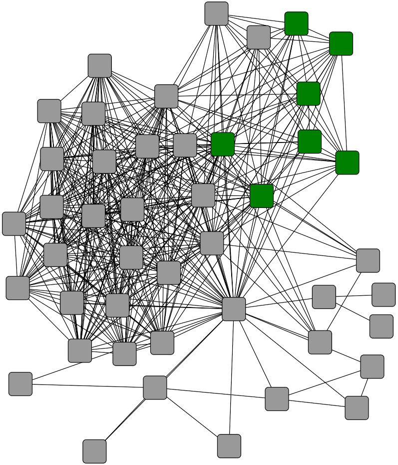
(origin) Argentineans living in Argentina

(fellows) Argentineans living in Spain

(host) Spanish alters

Defining the actor classes.

Case of an Argentinean migrant in Spain.



where are the alters originally from? where do they live?

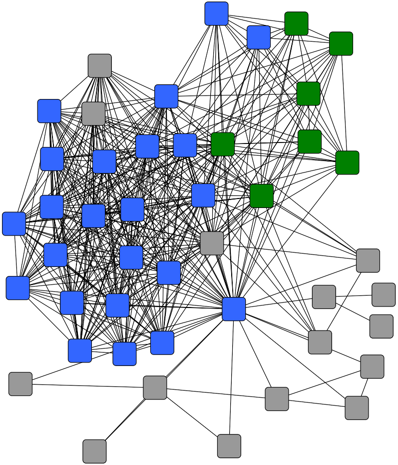
(origin) Argentineans living in Argentina

(fellows) Argentineans living in Spain

(host) Spanish alters

Defining the actor classes.

Case of an Argentinean migrant in Spain.



where are the alters originally from? where do they live?

(origin) Argentineans living in Argentina

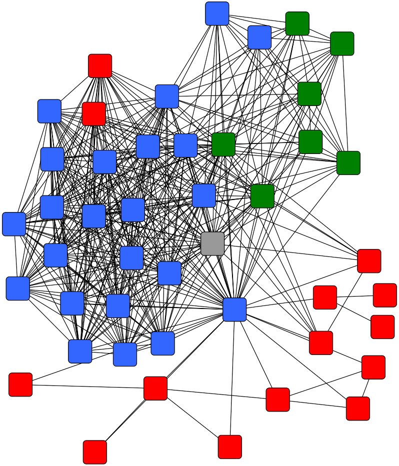
(fellows) Argentineans living in Spain

(host) Spanish alters

(transnationals) all other cases

Defining the actor classes.

Case of an Argentinean migrant in Spain.



where are the alters originally from? where do they live?

(origin) Argentineans living in Argentina

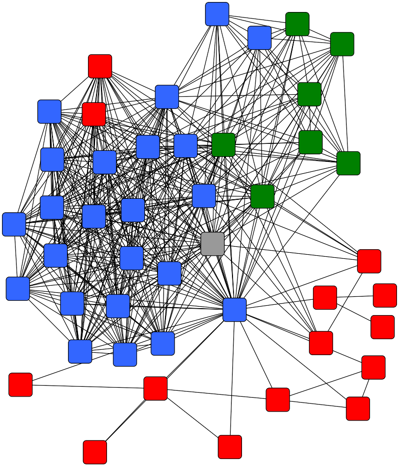
(fellows) Argentineans living in Spain

(host) Spanish alters

(transnationals) all other cases

Defining the actor classes.

Case of an Argentinean migrant in Spain.



where are the alters originally from? where do they live?

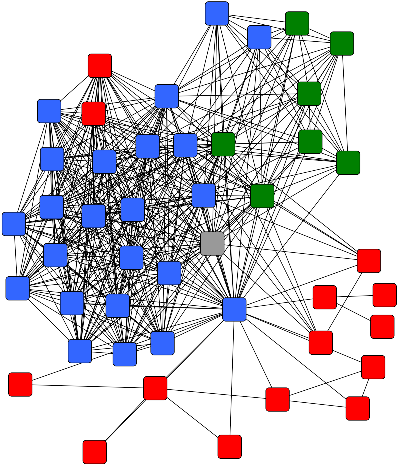
(origin) Argentineans living in Argentina

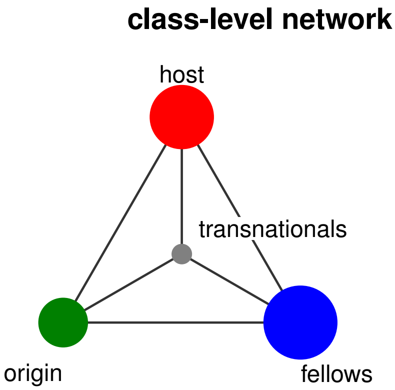
(fellows) Argentineans living in Spain

(host) Spanish alters

(transnationals) all other cases

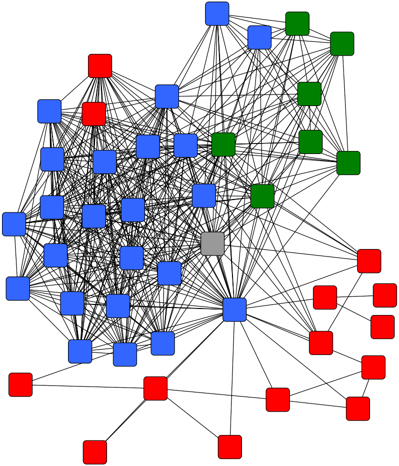
Layouting the actor classes.



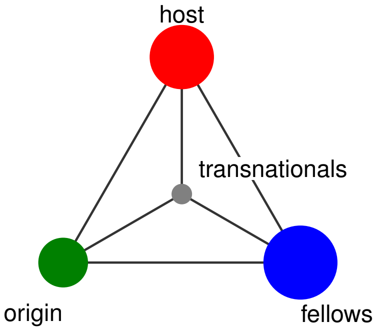


tie weights?

Layouting the actor classes.



class-level network



tie weights?

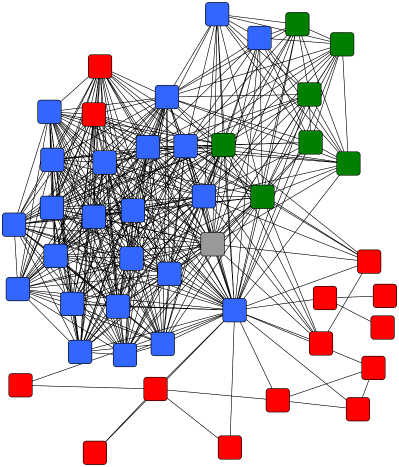
Normalization of the tie weight.

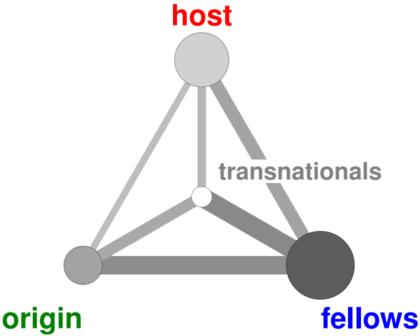
How strongly are two classes connected?

Given: network $G = (V, E)$ and two actor classes A and B

1. **(un-normalized count)** $\#\{(a, b) \in E\,;\ a \in A,\ b \in B\}$
   $\Rightarrow$ *larger classes will be stronger connected*
2. **(density)** $\frac{\#\{(a,b) \in E\,;\ a \in A\, b \in B\}}{|A| \cdot |B|}$
   $\Rightarrow$ *tends to zero when class sizes increase (sparsity)*
3. **(avg. number of B-neighbors)** $\frac{\#\{(a,b) \in E\,;\ a \in A\, b \in B\}}{|A|}$
   $\Rightarrow$ *asymmetric class-level network*
4. **(symmetric normalization)** $\frac{\#\{(a,b) \in E\,;\ a \in A\, b \in B\}}{\sqrt{|A| \cdot |B|}}$
   $\Rightarrow$ this is what we take (appropriate for sparse networks)

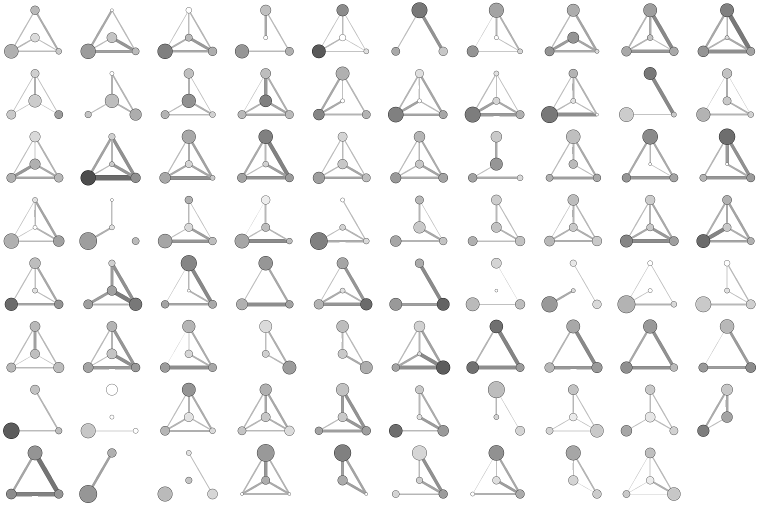
Graphical representation of class size and tie weight.





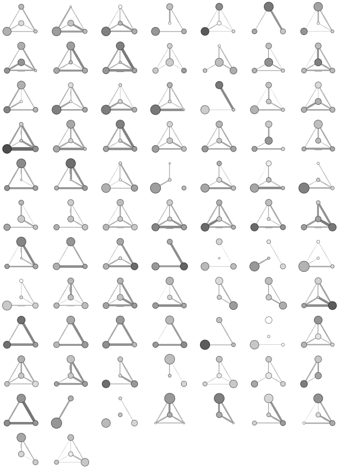
Summarizes essential network properties, while reducing complexity.

Facilitates comparing class-level networks of many individuals on small space (79 Argentineans).

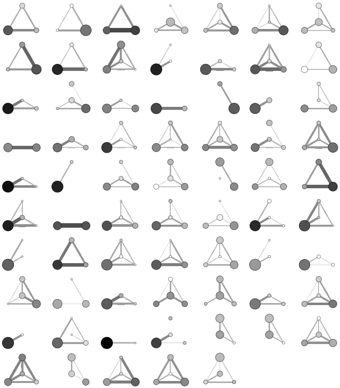


We could (in principle) also compare two populations.

from Argentina (N=79)



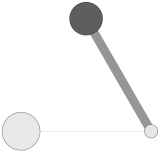
from Senegal/Gambia (N=68)



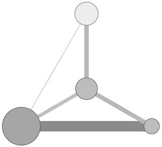
Would rather like to summarize them first. ($\Rightarrow$ average)

To compare populations we need to define **average** and **variability** of networks.

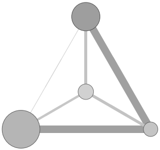
Arithmetic mean (or **median**) of class-level networks is defined componentwise (on class-sizes and tie-weights).



$+$



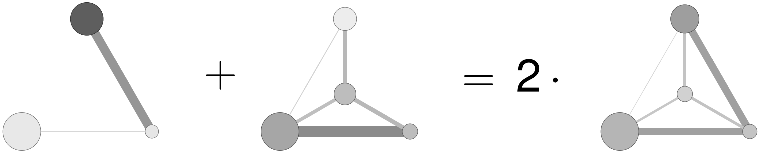
$= 2 \cdot$



Standard deviation, **quartiles**, **percentiles**, etc, of a set of networks is defined similarily (componentwise).

To compare populations we need to define **average** and **variability** of networks.

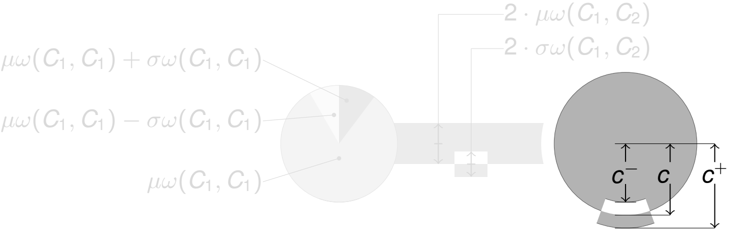
Arithmetic mean (or **median**) of class-level networks is defined componentwise (on class-sizes and tie-weights).



Standard deviation, **quartiles**, **percentiles**, etc, of a set of networks is defined similarily (componentwise).

Simultaneous **visualization** of average and variability.

class size represented by **radius**
$\Rightarrow$ variance represented by varying radius

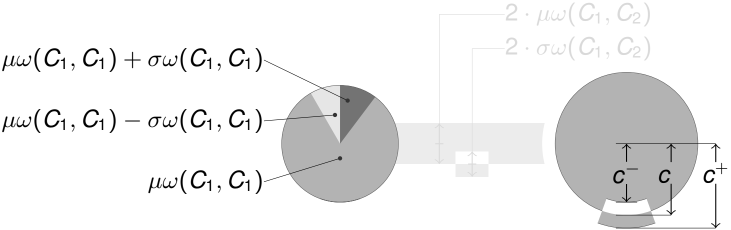


c = average class size; $c^\pm$ = average $\pm$ standard deviation

$\mu\omega(C_1, C_1)$ = average tie weight; $\sigma\omega(C_1, C_1)$ = standard deviation

Simultaneous **visualization** of average and variability.

intra class tie weights represented by **color (darkness)**
$\Rightarrow$ variance represented by varying darkness

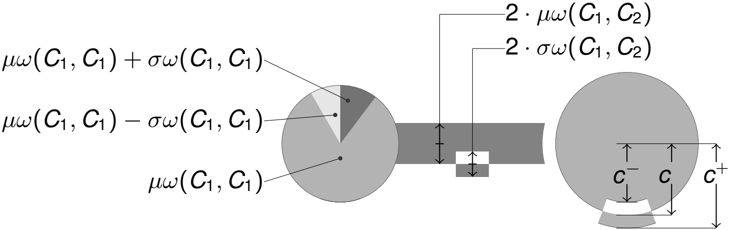


$c =$ average class size; $\quad c^{\pm} =$ average $\pm$ standard deviation

$\mu\omega(C_1, C_1) =$ average tie weight; $\quad \sigma\omega(C_1, C_1) =$ standard deviation

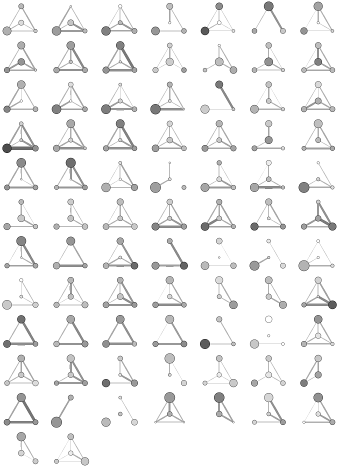
Simultaneous **visualization** of average and variability.

inter class tie weights represented by **edge thickness**
$\Rightarrow$ variance represented by varying thickness

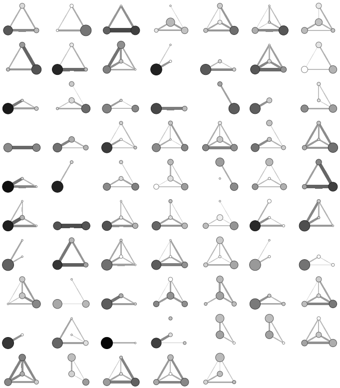


c = average class size; $\quad c^{\pm}$ = average $\pm$ standard deviation
$\mu\omega(C_1, C_1)$ = average tie weight; $\quad \sigma\omega(C_1, C_1)$ = standard deviation

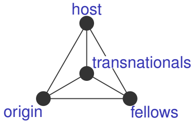
from Argentina (N=79)



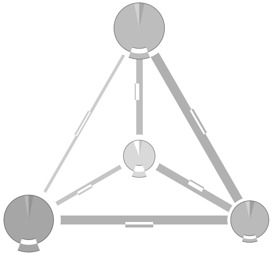
from Senegal/Gambia (N=68)



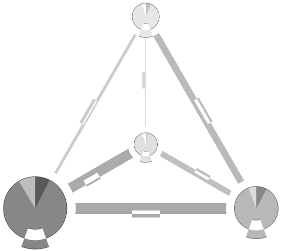
Now, the comparison of different populations can be done more conveniently.



from Argentina (N=79)

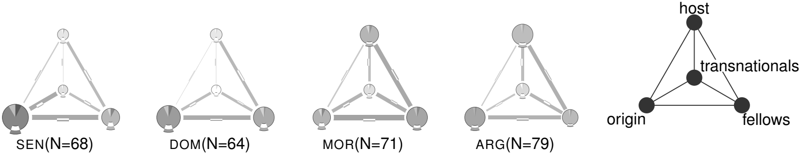


from Senegal/Gambia (N=68)

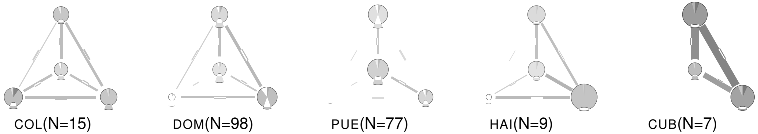


Average over migrants with the same country of origin/ host country.

migration to Spain:



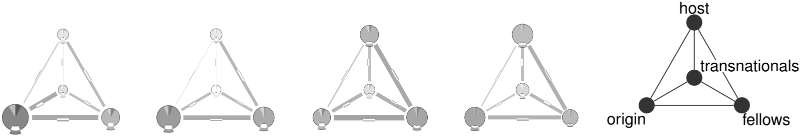
migration to the USA:



Which difference is statistically significant?

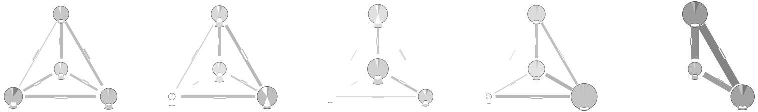
Average over migrants with the same country of origin/ host country.

migration to Spain:



SEN(N=68) DOM(N=64) MOR(N=71) ARG(N=79)

migration to the USA:



COL(N=15) DOM(N=98) PUE(N=77) HAI(N=9) CUB(N=7)

Which difference is statistically significant?

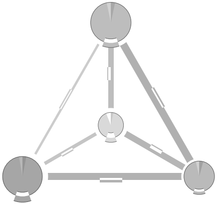
Significance of difference in averages.

H_0: $\mu = \mu'$;

H_1: $\mu \neq \mu'$;

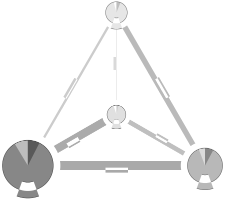
tested componentwise

Argentina



significance graph

Senegal/Gambia

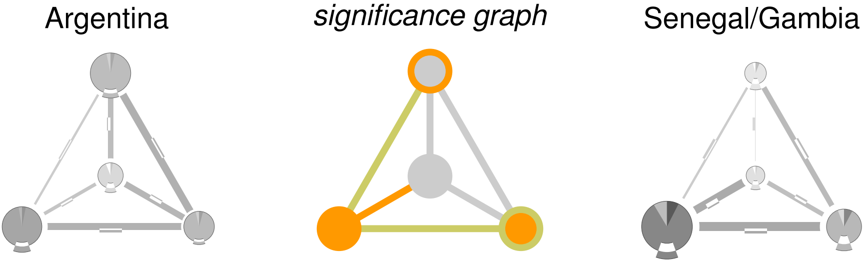


($p < 0.01$) ($p < 0.05$) (not significant)

Significance of difference in averages.

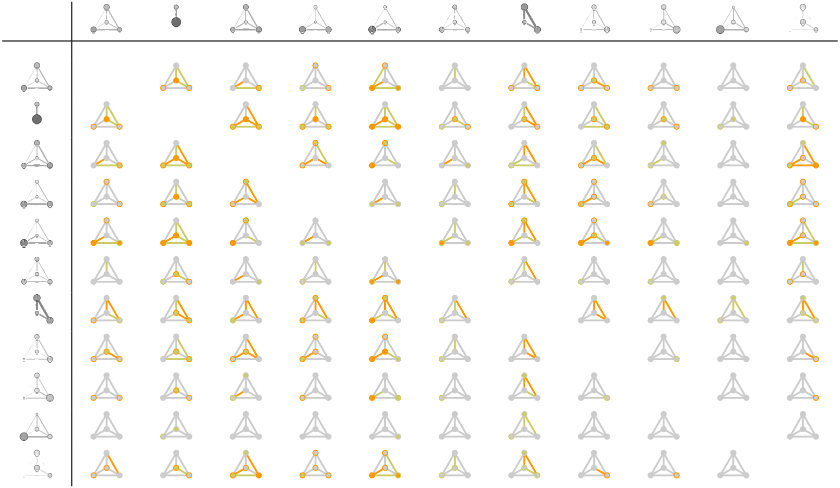
H_0: $\mu = \mu'$;
H_1: $\mu \neq \mu'$;

tested componentwise



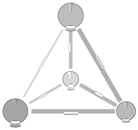
($p < 0.01$) ($p < 0.05$) (not significant)

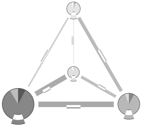
Pairwise comparison between migrants with the same country of origin / host country.

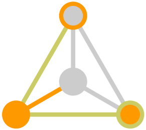


Just a thought: What do we gain by **visualizing** these variables in the way we do?

	Argentina	Senegal/Gambia
$\lvert C_1 \rvert$	13 (22|8)	22 (36|9)
$\lvert C_2 \rvert$	8 (11|5)	10 (20|3)
$\lvert C_3 \rvert$	14 (21|9)	4 (10|1)
$\lvert C_4 \rvert$	5 (10|2)	3 (7|1)
$\omega(1,1)$	5.5 (7.8|3.6)	10.0 (19.0|3.1)
$\omega(1,2)$	1.6 (4.0|0.3)	2.3 (5.9|0.4)
$\omega(1,3)$	0.3 (0.8|0.0)	0.2 (1.6|0.0)
$\omega(1,4)$	0.3 (1.1|0.0)	1.8 (3.8|0.0)
$\omega(2,2)$	3.4 (6.3|1.6)	3.7 (9.8|0.9)
$\omega(2,3)$	1.7 (3.3|0.6)	1.0 (2.4|0.0)
$\omega(2,4)$	1.0 (2.1|0.0)	0.7 (2.1|0.0)
$\omega(3,3)$	3.1 (5.3|1.6)	0.5 (2.8|0.0)
$\omega(3,4)$	0.8 (1.3|0.0)	0.0 (1.0|0.0)
$\omega(4,4)$	1.1 (2.2|0.0)	0.7 (1.9|0.0)





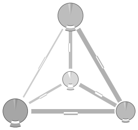


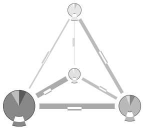
Shows the relations between variables.

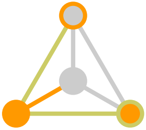
$\Rightarrow$ **(network of variables)**

Just a thought: What do we gain by **visualizing** these variables in the way we do?

	Argentina	Senegal/Gambia
$\|C_1\|$	13 (22|8)	22 (36|9)
$\|C_2\|$	8 (11|5)	10 (20|3)
$\|C_3\|$	14 (21|9)	4 (10|1)
$\|C_4\|$	5 (10|2)	3 (7|1)
$\omega(1,1)$	5.5 (7.8|3.6)	10.0 (19.0|3.1)
$\omega(1,2)$	1.6 (4.0|0.3)	2.3 (5.9|0.4)
$\omega(1,3)$	0.3 (0.8|0.0)	0.2 (1.6|0.0)
$\omega(1,4)$	0.3 (1.1|0.0)	1.8 (3.8|0.0)
$\omega(2,2)$	3.4 (6.3|1.6)	3.7 (9.8|0.9)
$\omega(2,3)$	1.7 (3.3|0.6)	1.0 (2.4|0.0)
$\omega(2,4)$	1.0 (2.1|0.0)	0.7 (2.1|0.0)
$\omega(3,3)$	3.1 (5.3|1.6)	0.5 (2.8|0.0)
$\omega(3,4)$	0.8 (1.3|0.0)	0.0 (1.0|0.0)
$\omega(4,4)$	1.1 (2.2|0.0)	0.7 (1.9|0.0)

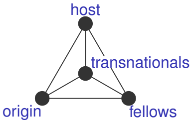






Shows the relations between variables.
$\Rightarrow$ **(network of variables)**

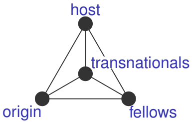
Subdividing / refining actor classes.



Which category best describes your skin color?

16 classes ⇒ 152 variables; (beware of **overfitting**)

Subdividing / refining actor classes.

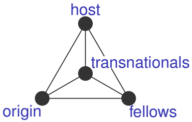


Which category best describes your skin color?

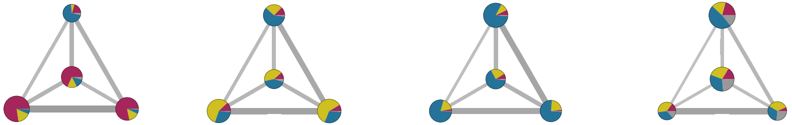
BLACK(N=77) **BROWN**(N=199) **WHITE**(N=128) **OTHER**(N=44)

16 classes ⇒ 152 variables; (beware of **overfitting**)

Subdividing / refining actor classes.



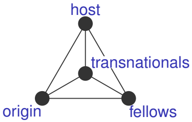
Which category best describes your skin color?



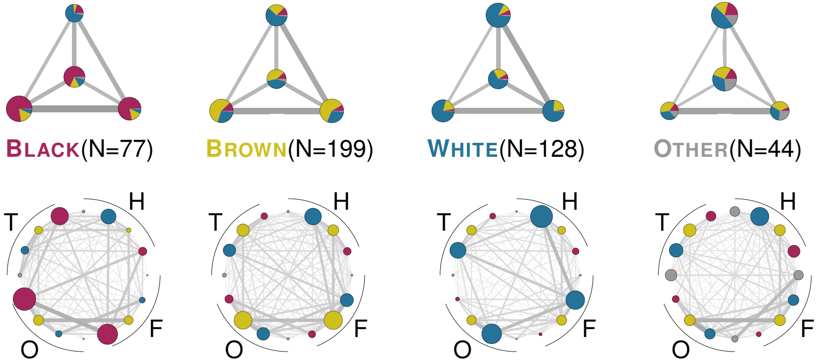
BLACK(N=77) **BROWN**(N=199) **WHITE**(N=128) **OTHER**(N=44)

16 classes ⇒ 152 variables; (beware of **overfitting**)

Subdividing / refining actor classes.

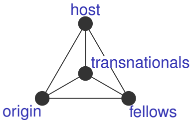


Which category best describes your skin color?

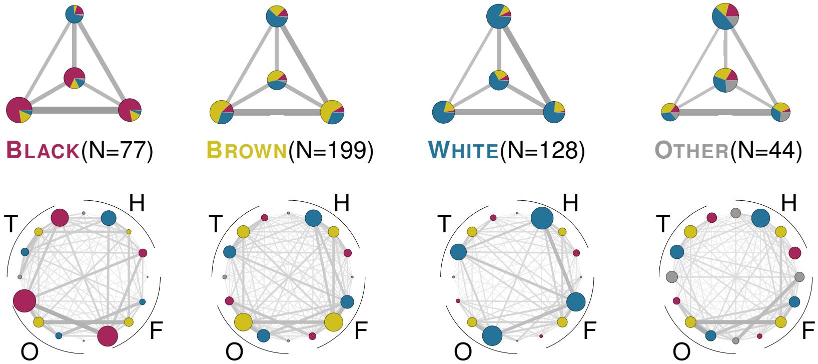


16 classes ⇒ 152 variables; (beware of **overfitting**)

Subdividing / refining actor classes.



Which category best describes your skin color?



16 classes $\Rightarrow$ 152 variables; (beware of **overfitting**)

Conclusions.

Presented a framework for averaging over **collections** of networks and for visualizing them.

- **General applicability**
  - Suitable for all sets of networks with actor attributes.
  - Efficient: suitable for large networks and many networks (small number of classes).

- **Visualization is useful during analysis**

  Large number of variables and relations between variables
  $\Rightarrow$ keep the overview using appropriate network graphics.

Further details in: Ulrik Brandes, Jürgen Lerner, Miranda J. Lubbers, Chris McCarty, and José Luis Molina. Visual Statistics for Collections of Clustered Graphs. In: *Proc. IEEE Pacific Visualization Symp.* (2008), to appear.