Analysis of the Cuban HIV network

Chaire MMB - workshop "Connectivité en écologie"

Viet Chi TRAN - Université Lille 1 - France

with S. Clémençon, H. De Arazoza and F. Rossi

October 18, 2013

Introduction - Cuban data

Cuban social network and configuration models

Clustering the Cuban network

Cuban data

★ The AIDS epidemics is present in Cuba since 25 years and a database contains detections since 1986 with information for contact-tracing.

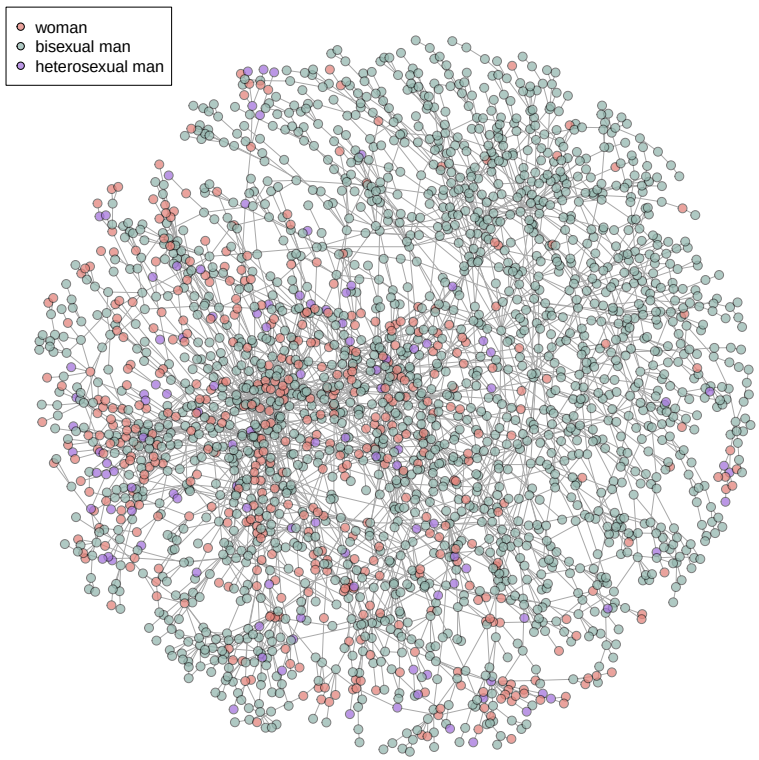
★ Mixing compartmental models with CT: De Arazoza-Lounes (2002), Clémençon-De Arazoza-T. (2008), Blum-T. (2010) with various statistical motivations.

★ We would like to take into account the network underlying AIDS epidemic in Cuba.

- understanding of the propagation mechanisms and reconstruction of the history of the disease,
- modeling the evolution of the disease and making predictions ; evaluation of public health policies and prevention strategies...

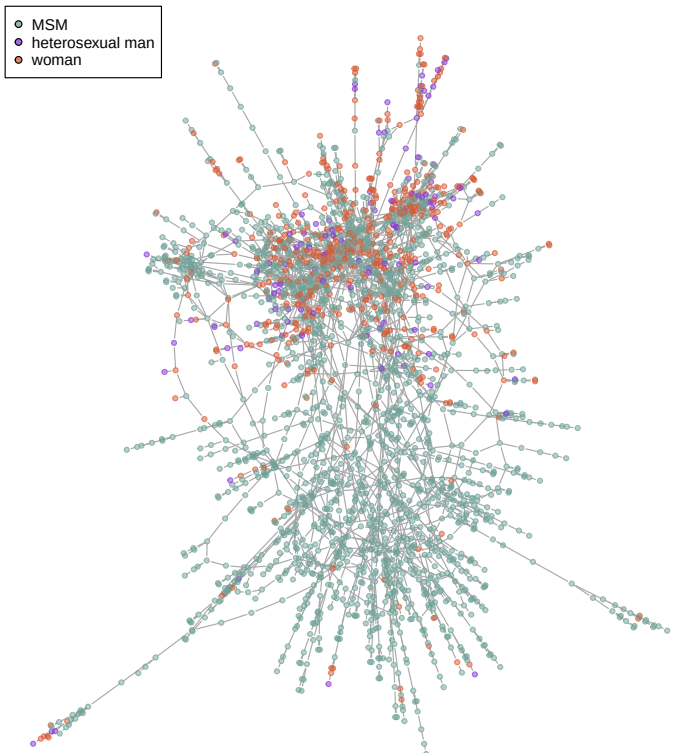
Acknowledgements: This work has been financed by ANR *Viroscopy*, Chaire MMB and ANR *MANEGE*.*Thanks to Dr. J. Perez of the National Institute of Tropical Diseases in Cuba for granting access to the HIV/AIDS database.*

Cuba CT graph



- 5389 ind., 4073 edges
- Giant component: 2386 ind. (44%), 3168 edges (78%)
- Second largest component has 17 edges.
- almost 2000 isolated ind. or couples.

Cuba CT graph



- 5389 ind., 4073 edges
- Giant component: 2386 ind. (44%), 3168 edges (78%)
- Second largest component has 17 edges.
- almost 2000 isolated ind. or couples.

Some literature on sexually transmitted diseases on networks

★ Keeling and Eames (2005), Liljeros et al. (2003)

★ Data-based studies: usually smaller populations and/or smaller giant components and/or few infected indviduals

- Bearman et al. (2004): study on the american teenage sexuality (without STD) ; 573 persons, giant=288.
- Wylie and Jolly (2001): Manitoba study ; 4544 individuals, but giant=82 persons
- Rothenberg et al. (1995): Colorado Springs study ; 2200 individuals, giant=965 persons with only a very small number of HIV positive individuals

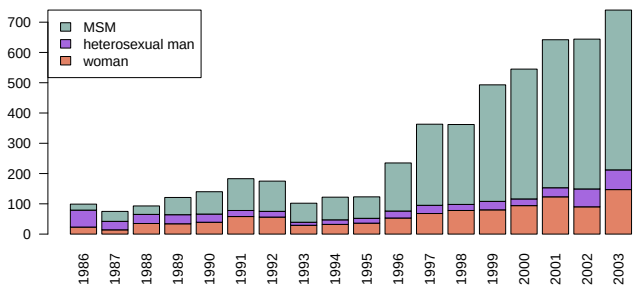
Questions

★ Does the data justify the modeling of the Cuban social network with some simple graph model (for instance CM), hence providing some simple evolution equations for the propagation of the disease ?

★ If not, how can we explore the data and give a description of the network ?

★ As an illustration, can we understand the relation between sexual orientation of the infected individuals and propagation of the HIV ?

	population	GCC
women	0.21	0.20
HT men	0.11	0.05
MSM	0.69	0.76



Introduction - Cuban data

Cuban social network and configuration models

Clustering the Cuban network

Joint distribution of the degrees of two neighbors

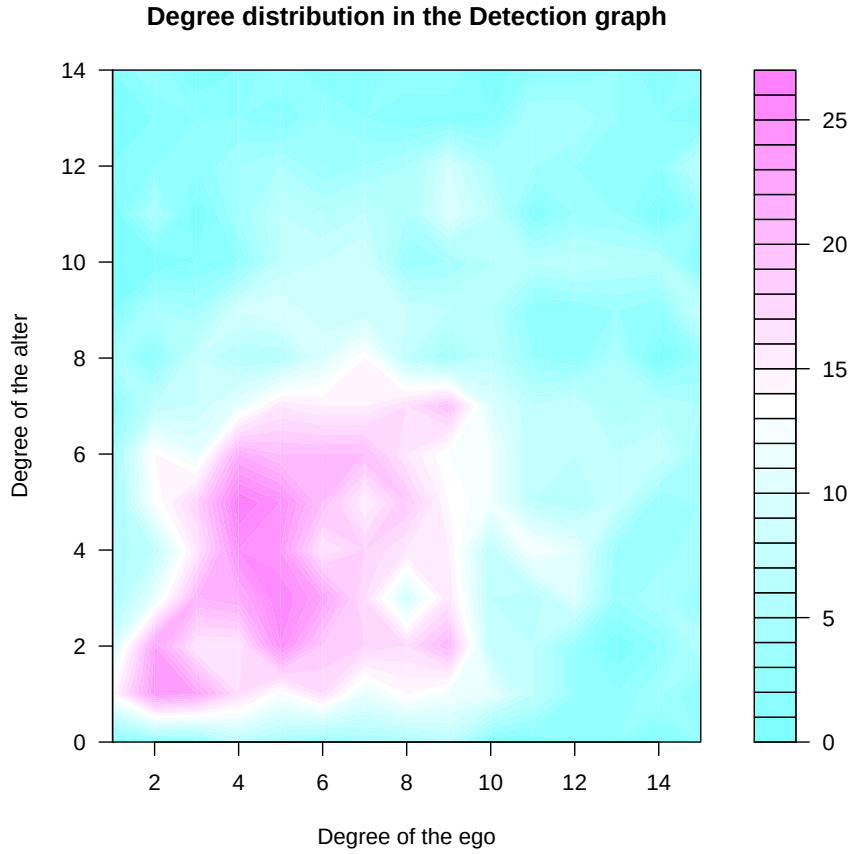


Figure: *Joint degree distribution of alter and ego for the population of MSM.*

If we restrict to the subgroup of individuals with less than 10 contacts, the independence assumption is accepted thanks to a χ^2 test.

A tree-like graph

- Indicators show an apparent weak resilience:
  - 1157 articulation points
  - 187 cliques (among them 177 triangles)
  - low assortative mixing coefficients:

Ego is a	Alter is a woman	Alter is a heterosexual man	Alter is an MSM	Total
Woman	77 (*1.9%*)	157 (*3.9%*)	408 (*10.0%*)	642 (*15.8%*)
HT man	282 (*6.9%*)	4 (*0.1%*)	20 (*0.5%*)	306 (*7.5%*)
MSM	800 (*19.6%*)	25 (*0.6%*)	2300 (*56.5%*)	3125 (*76.7%*)
Total	1159 (*28.5%*)	186 (*4.6%*)	2728 (*67.0%*)	

$$r = \frac{tr(M) - \|M^2\|}{1 - \|M^2\|}$$

where $M = (m_{i,j})_{i,j}$ and $m_{i,j}$ is the fraction of edges linking group i to group j. r=0.0512 for sexual orientation.

Degree distribution

Based on Clauset, Shalizi and Newman, we minimize the dissimilarity measure

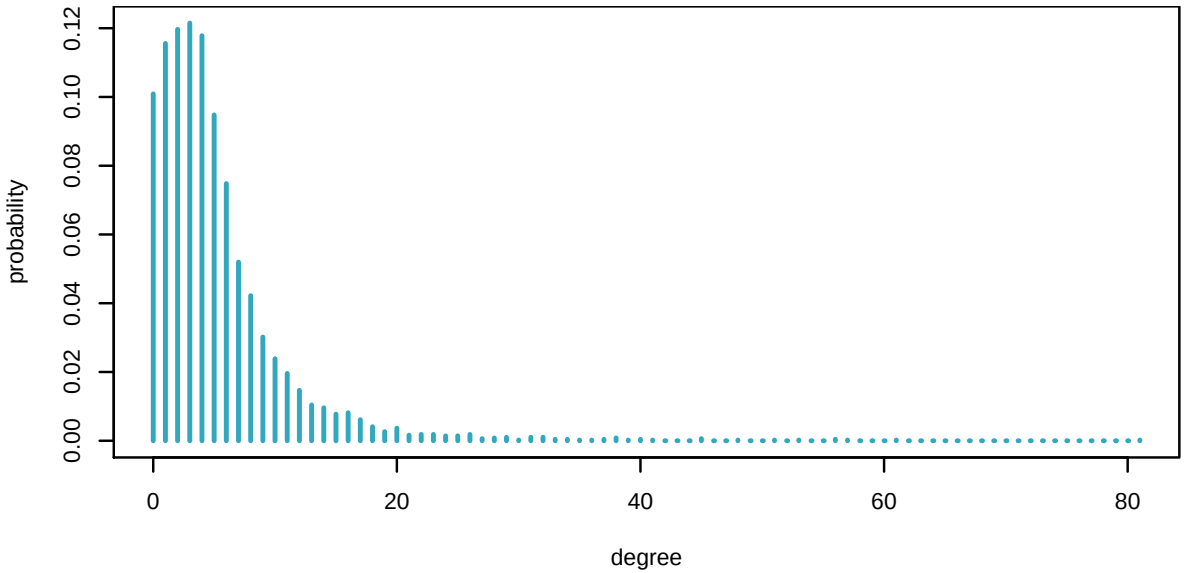
$$\mathcal{K}_{k_0}(p, \alpha) = \sum_{k \geq k_0} \frac{p_k}{c_{p,k_0}} \log \left(\frac{C_\alpha \cdot p_k}{c_{p,k_0} \cdot k^{-\alpha}} \right),$$

where $c_{p,k_0} = \sum_{k \geq k_0} p_k$ and $C_\alpha = \sum_{k \geq k_0} 1/k^\alpha$.

$$\widehat{\alpha}_{k_0} = \arg \min_{\alpha > 1} \mathcal{K}_{k_0}(p_n, \alpha).$$

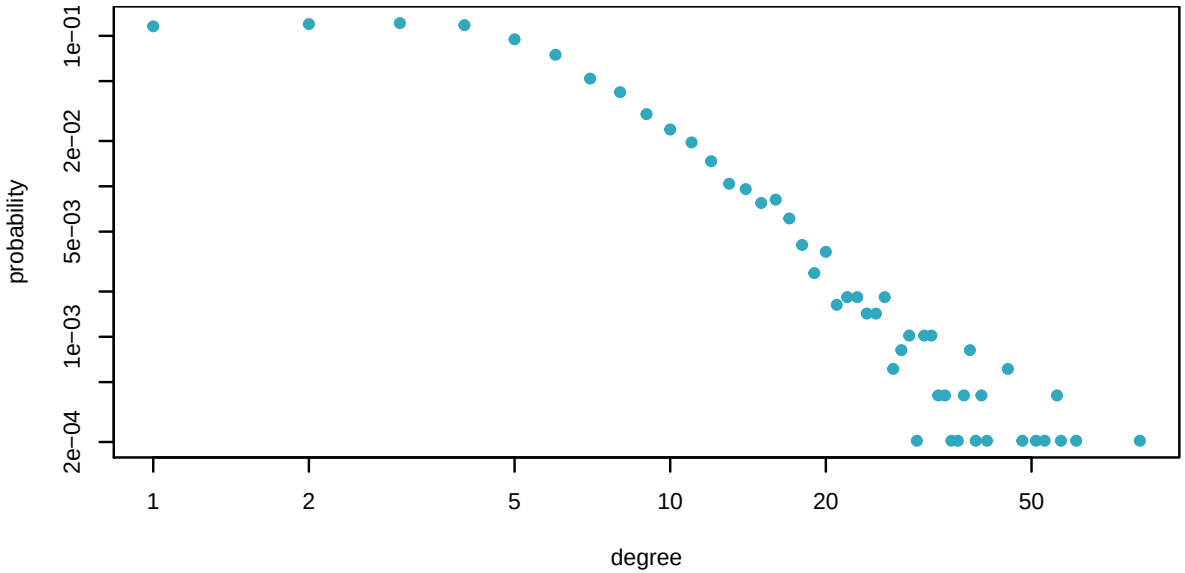
	$\widehat{k}_0$	$\widehat{\alpha}_{k_0}$	Mean	Std dev.	Min	Max
Whole population	7	3.06	6.17	5.54	1	82
Women	6	2.71	5.88	5.03	1	39
Heterosexual men	7	3.36	4.98	4.11	1	30
MSM	7	3.02	6.43	5.84	1	82

Degree distribution



	$\widehat{k}_0$	$\widehat{\alpha}_{k_0}$	Mean	Std dev.	Min	Max
Whole population	7	3.06	6.17	5.54	1	82
Women	6	2.71	5.88	5.03	1	39
Heterosexual men	7	3.36	4.98	4.11	1	30
MSM	7	3.02	6.43	5.84	1	82

Degree distribution



	$\widehat{k}_0$	$\widehat{\alpha}_{k_0}$	Mean	Std dev.	Min	Max
Whole population	7	3.06	6.17	5.54	1	82
Women	6	2.71	5.88	5.03	1	39
Heterosexual men	7	3.36	4.98	4.11	1	30
MSM	7	3.02	6.43	5.84	1	82

Degree distribution

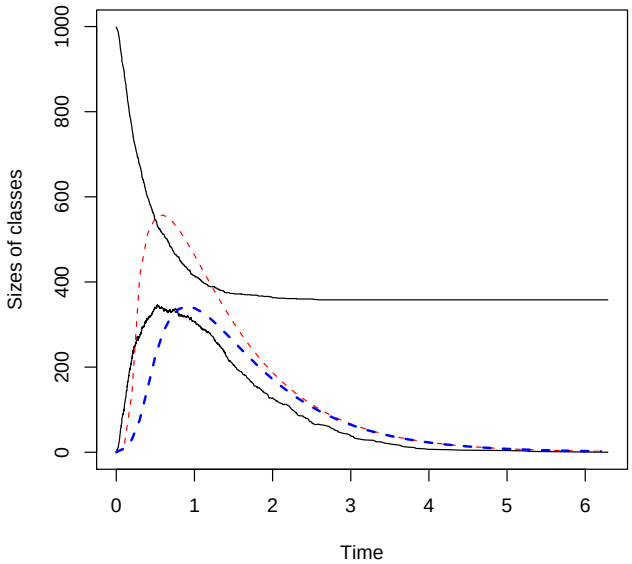
Prop: there is a giant component in a configuration model with i.i.d. degree distribution $(p_k)_{k\geq 0}$ if

$$\mathbb{E}(D(D-1))/E(D) > 1 \quad \Leftrightarrow \quad \sum_{k\geq 2} k(k-2)p_k > 0$$

where D is a r.v. with distribution (p_k).

	$\widehat{k}_0$	$\widehat{\alpha}_{k_0}$	Mean	Std dev.	Min	Max
Whole population	7	3.06	6.17	5.54	1	82
Women	6	2.71	5.88	5.03	1	39
Heterosexual men	7	3.36	4.98	4.11	1	30
MSM	7	3.02	6.43	5.84	1	82

Degree distribution

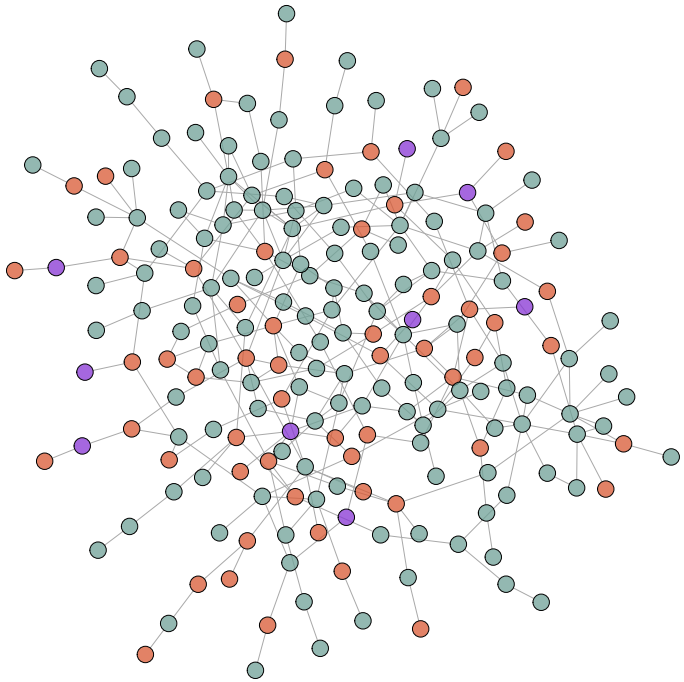


Prop: For $\varepsilon > 0$, when $n \to +\infty$, the degree distribution when after $[\varepsilon n]$ infections converges to:

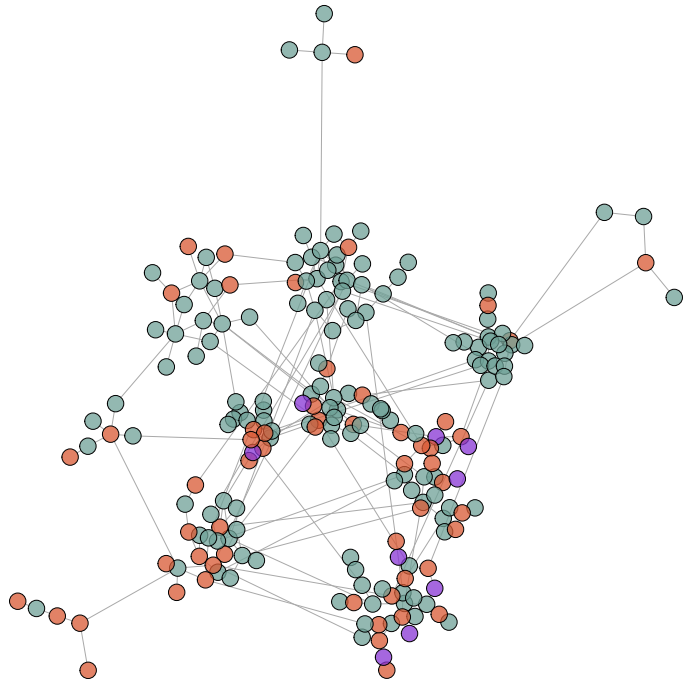
$$\frac{1}{1-\varepsilon} \sum_{k \geq 0} p_k (1 - z^{\varepsilon})^k \delta_k$$

where z^{ε} is the solution of $1 - \varepsilon = f(1 - z)$, f being the generating function of the original degree distribution.

A visual mining of large graph is necessary



Classical visualization



Hierarchical visualization

Moreover, as has been seen in a previous slide, it is impossible to see anything from the “naive” graph representation of the data.

Methodology

★ **Clustering:**

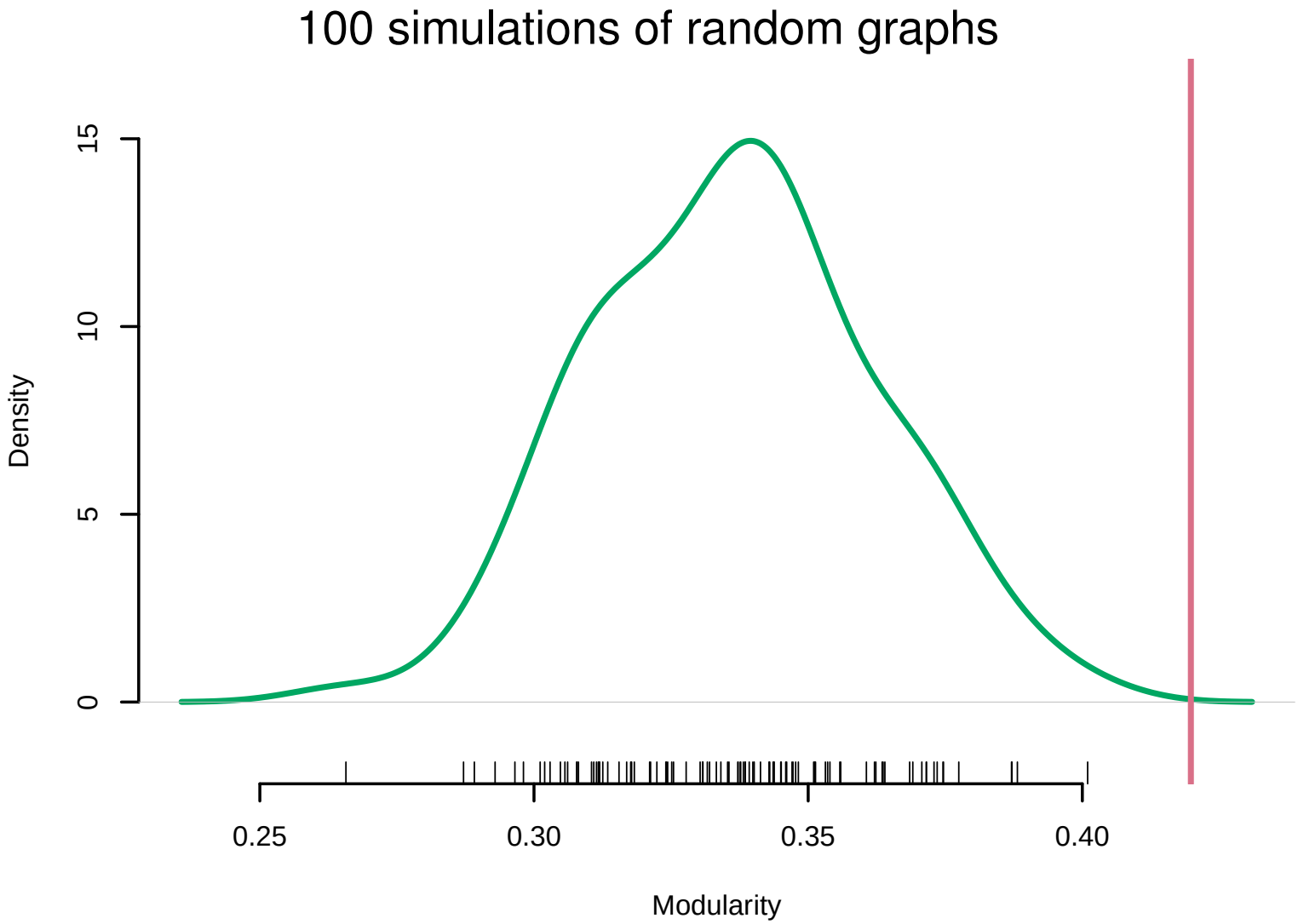
- maximization of the modularity (Girvan and Newman, 2004) :

$$Q = \frac{1}{2m} \sum_{l=1}^{L} \sum_{i,j \in C_l} \left(G_{ij} - \frac{d_i d_j}{2m} \right)$$

+ favor dense clusters and produces interesting partitions for visualization (Fortunato 2010)
+ the optimisation is an NP-hard problem but high quality sub-optimal solutions can be obtained by annealing (Rossi Villa-Vialaneix 2010) or other methods (Noak Rotta, 2009)

- Clustering significance:

+ compute the modularity of the partition that is obtained
+ simulate configuration models with same degrees and compute modularity.



The partition that is obtained is statistically significant.

★ **Hierarchical clustering:**

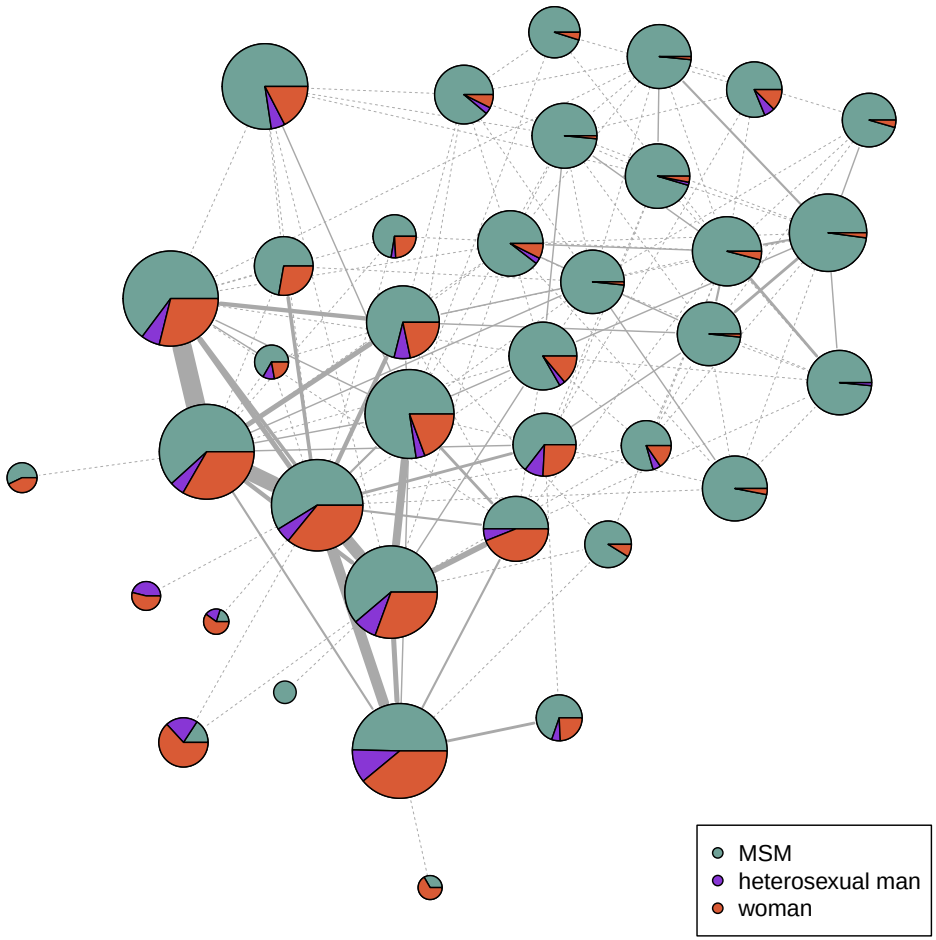
- If the first clustering is relevant, and if the classes have large sizes, we can refine the partition.
  - + Reiterate the clustering for each element of the partition, without taking inter-cluster connections.
  - + Test the significativeness of the cluster's partition
  - + Test the significativeness of the global clustering of the graph.

★ **Coarsening:**

- merge clusters that induce the least reduction in modularity as long as we remain above the original graph.

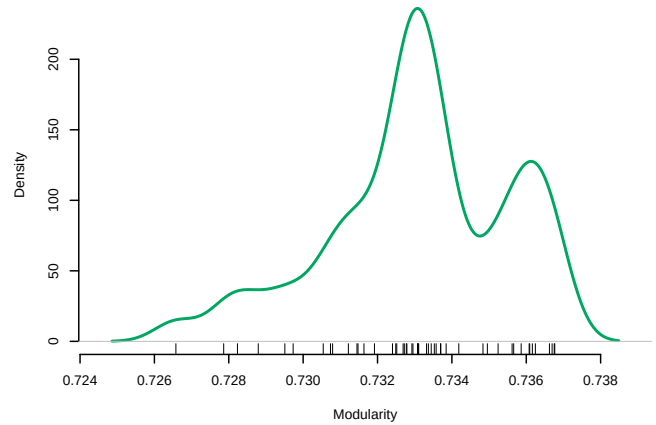
★ **Visualization**

- Fruchterman Reingold algorithm to display the network of clusters

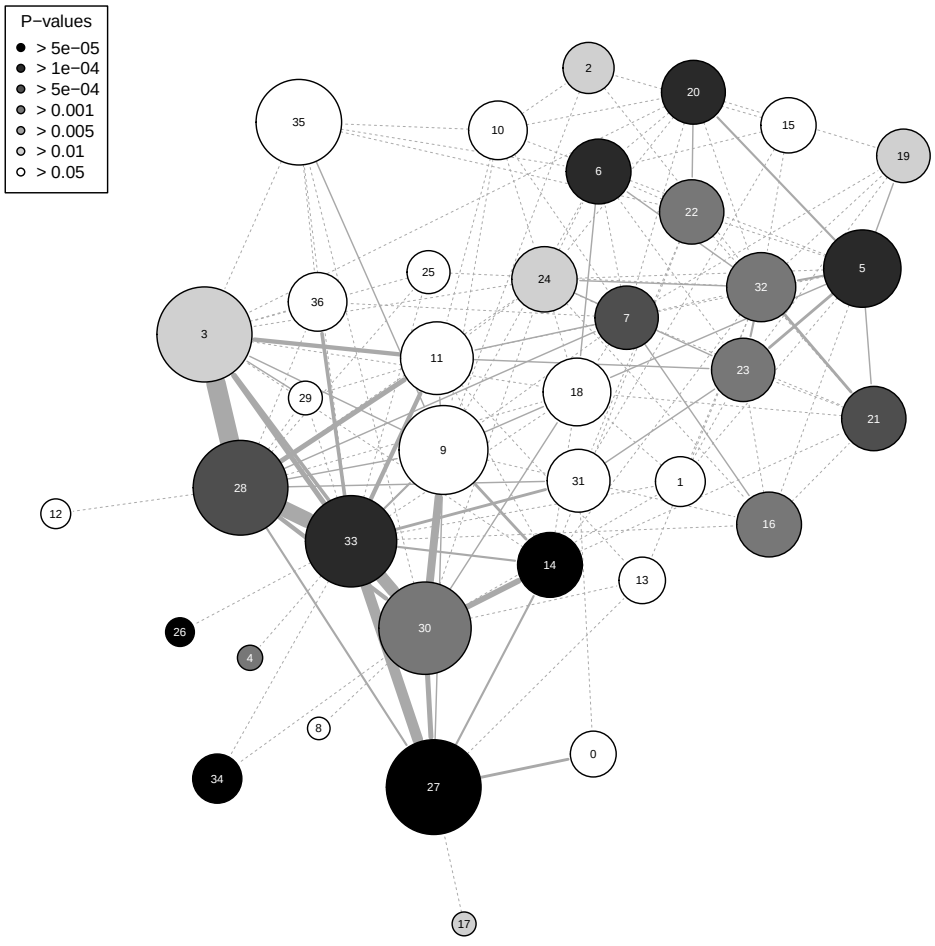


$\Rightarrow$ 37 classes (89.5% of internal links)

$\Rightarrow$ modularity $\simeq 0.85$

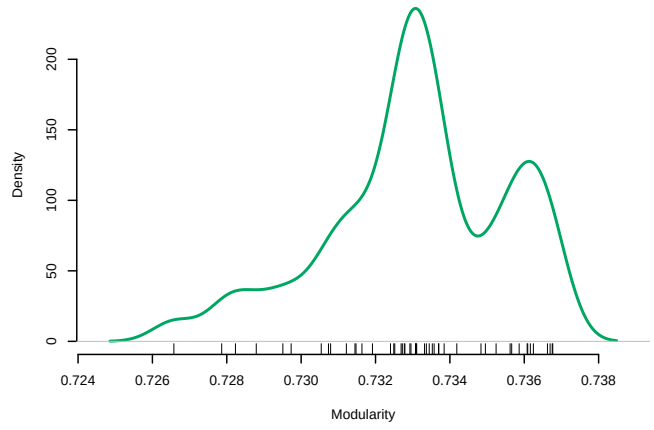


$\Rightarrow$ random modularity ≤ 0.74



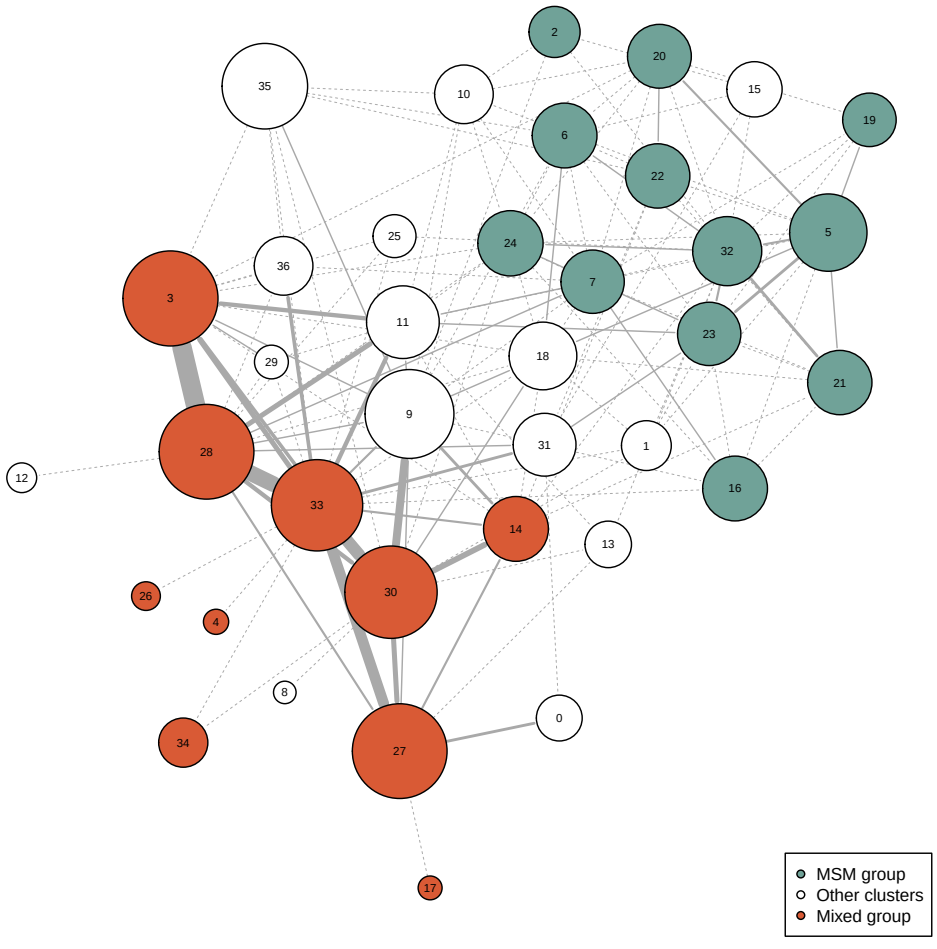
⇒ 37 classes (89.5% of internal links)

⇒ modularity $\simeq 0.85$



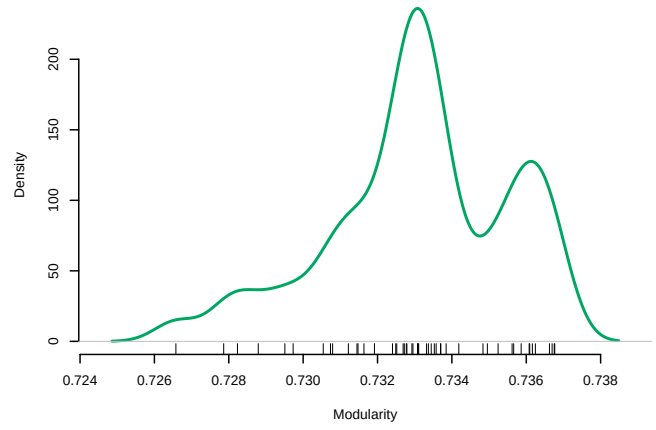
⇒ random modularity ≤ 0.74

⇒ hierarchical visualization of the sexual orientation



⇒ 37 classes (89.5% of internal links)

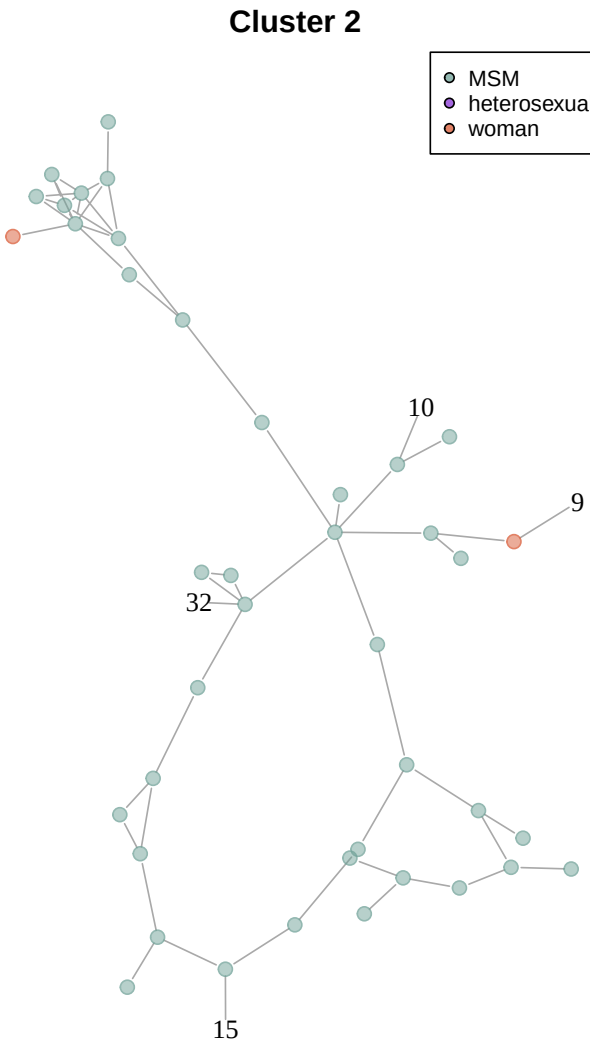
⇒ modularity $\simeq 0.85$

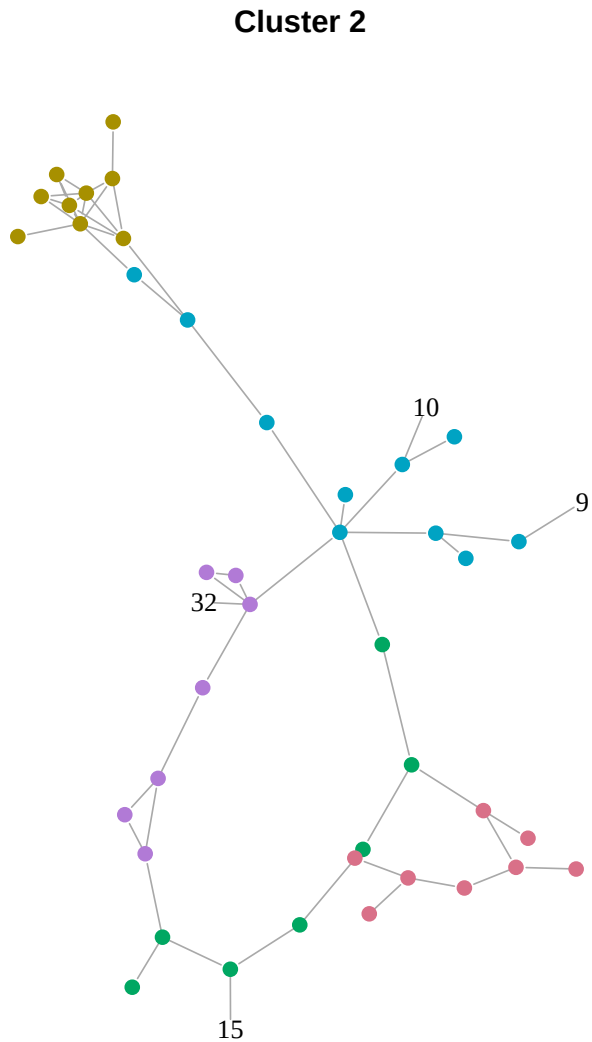


⇒ random modularity ≤ 0.74

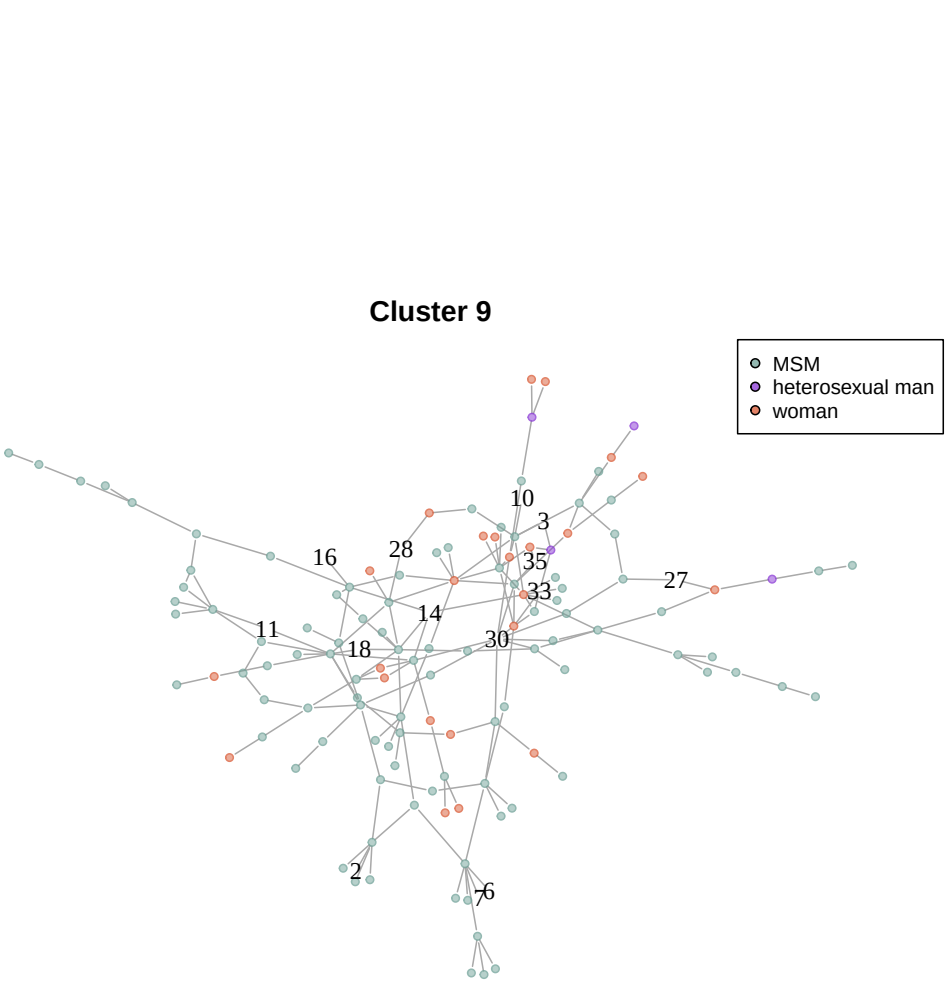
⇒ hierarchical visualization of the sexual orientation

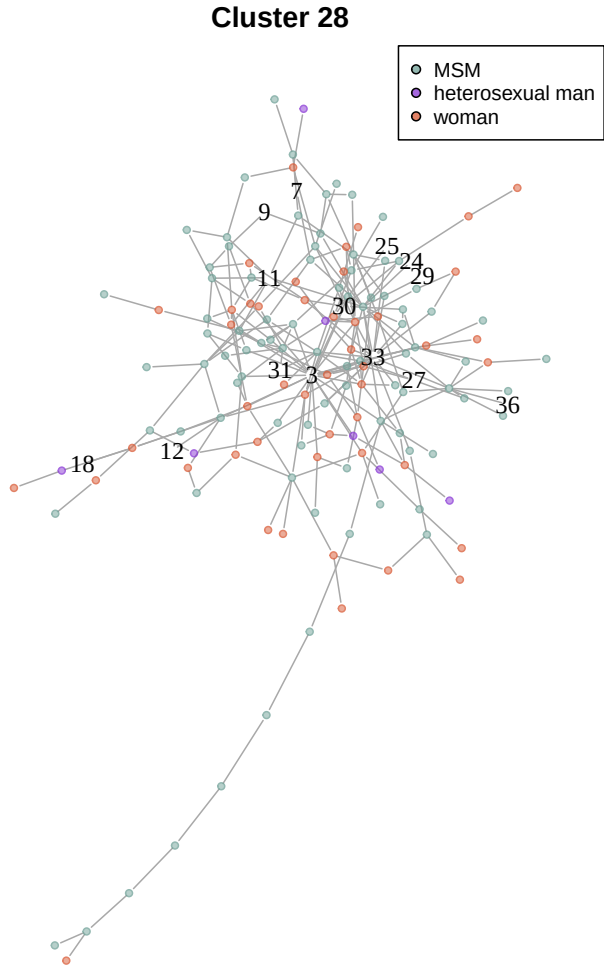
Results for the Cuban data





Results for the Cuban data





Cluster	Vertices	Edges	Sub-clusters	p-value	type
3	142	195	11	0.0122	Mixed
5	94	101	None	0.0002	MSM
9	124	143	None	0.7597	Typical
11	83	93	10	0.4439	Typical
27	141	236	11	0.0001	Mixed
28	141	184	10	0.0007	Mixed
30	134	177	17	0.0014	Mixed
33	131	214	None	0.0003	Mixed
35	115	155	10	0.7890	Typical
36	54	66	None	0.1130	Typical