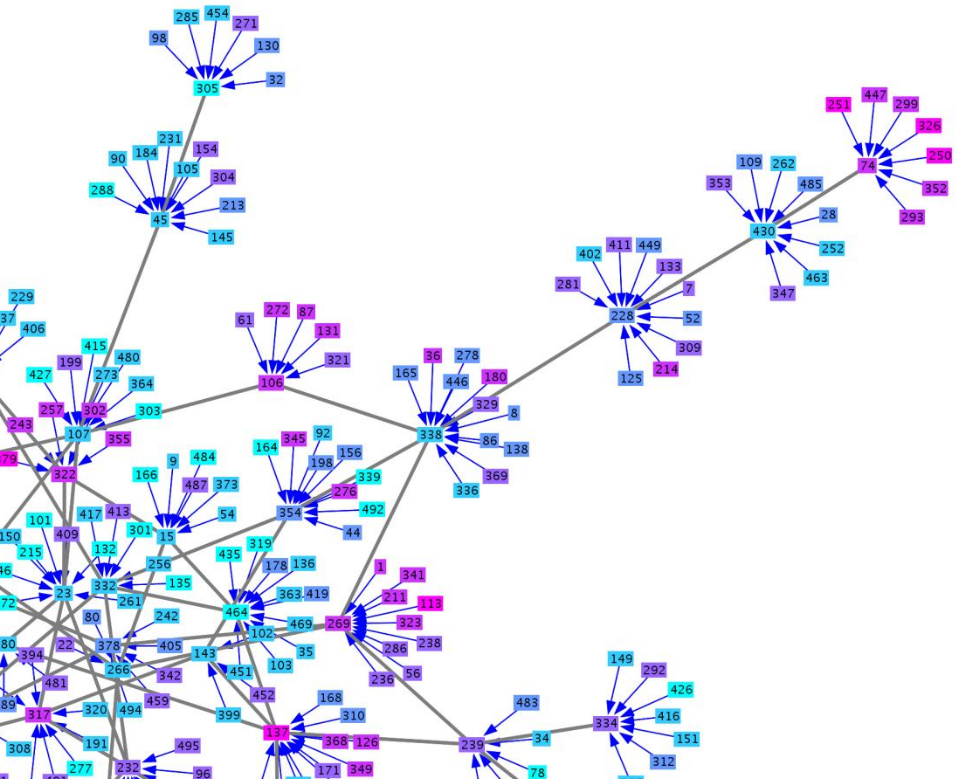
Generation of realistic consumer populations
for electrical demand simulation in the context of the smart grid



ECCS'2012

Samuel THIRIOT — EDF R&D
Enrique KREMERS — **EIFER**
Wolfgang HAUSER — Stuttgart University

Introduction

- Agent-based simulations for energy systems require as an input both:
  - the structure of the system (a network or graph) and
  - the attributes of the agents (or nodes)

- Moreover, the agents are positionned over the network according to their characteristics

- Generation of synthetic, networked populations involve two steps:
  - generate a population of agents (that represent energy consumers) with realistic attributes
  - position them over a network according to their attributes

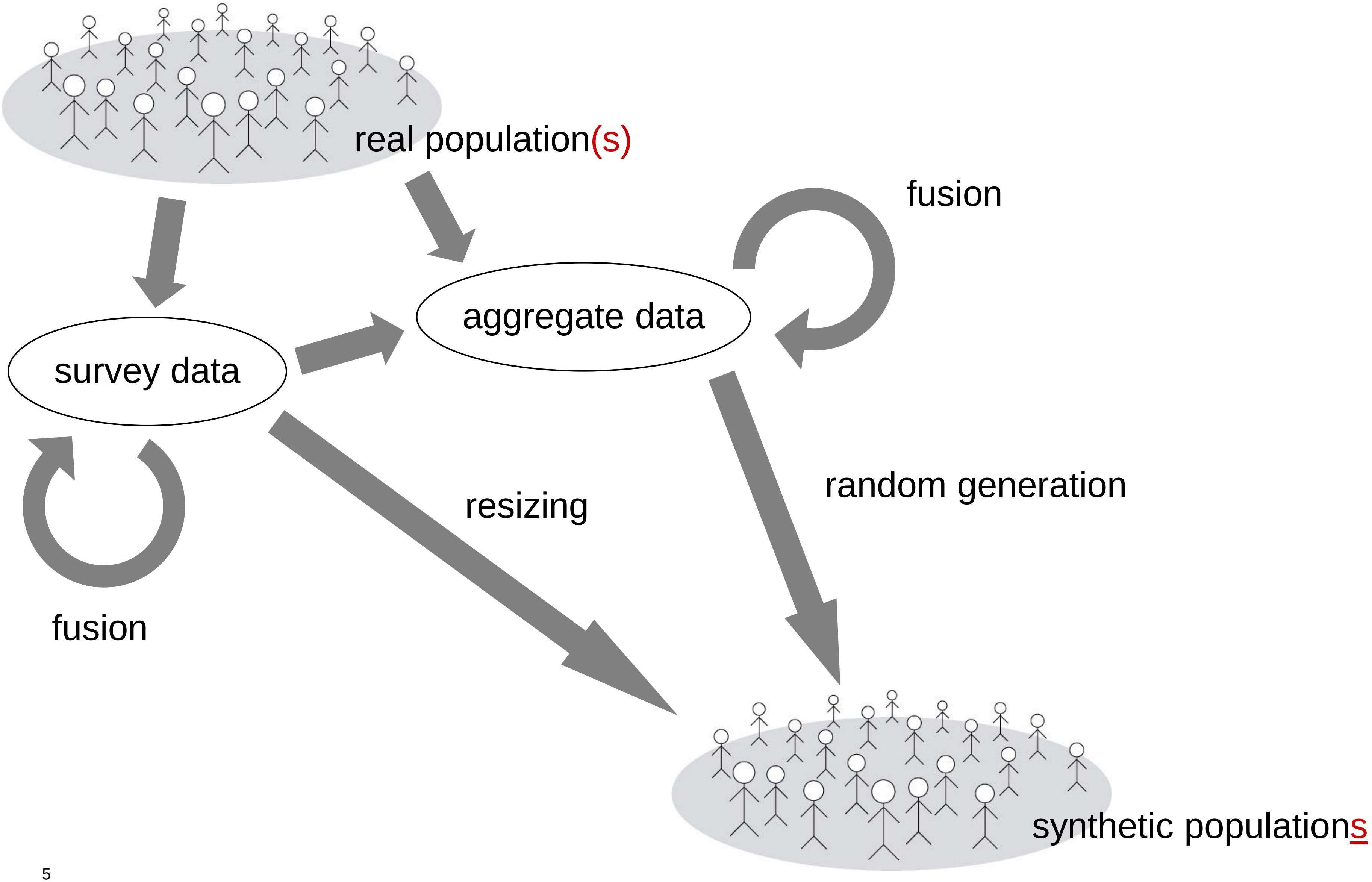
Use case

- In a previous work, authors proposed a model of energy consumption that reconstructs the load curve from the bottom up [3,6]
- Parameters of this model include a list of the devices of each household, and the corresponding usage of these appliances (the « synthetic population »)
- Data:
  - we previously conducted a survey on 769 households [3]. Data includes household usage rates, sociodemographics and lifestyle issues.
  - Objective: extrapolate « smartly » this sample to large synthetic population sizes to be used into simulations
- This case is representative of most data-driven agent-based simulations in which data is rare and costly

First approach:

Hand-made generation of attributes (without network)

Main approaches to population generation



Aggregating data

- We rejected population resizing in order to avoid errors caused by the low size of our sample.
  - for instance, this ~700 sample contains a household with 1 adult and 12 children
  - extrapolated to 10,000 households => 15 of such an household, which is actually rare in the actual population.

- We aggregate data by detecting correlations, based on:
  - common sense: « the number of showers may probably depends on the number of people in the household or the number of children »
  - statistical analysis: « does the income determines the number of showers ? »
  - number of samples in the cross analysis: too few correlations, reject the correlation

nrshowers_class * npers Crosstabulation									
			npers						
			1	2	3	4	5	6	Total
nrshowers_class		Count	12	22	1	6	1	0	42
		% within npers	4,2%	4,5%	,7%	4,2%	2,6%	,0%	3,8%
	0_to_4	Count	108	137	15	18	2	2	282
		% within npers	37,6%	28,0%	10,6%	12,7%	5,1%	16,7%	25,4%
	10_to_14	Count	18	188	41	25	10	3	285
		% within npers	6,3%	38,4%	28,9%	17,6%	25,6%	25,0%	25,6%
	15_to_19	Count	1	12	28	15	7	0	63
		% within npers	,3%	2,4%	19,7%	10,6%	17,9%	,0%	5,7%
	20_to_24	Count	0	3	26	25	6	1	61
		% within npers	,0%	,6%	18,3%	17,6%	15,4%	8,3%	5,5%
	25_to_29	Count	0	0	2	23	3	1	29
		% within npers	,0%	,0%	1,4%	16,2%	7,7%	8,3%	2,6%
	30_to_34	Count	0	1	0	4	1	1	7
		% within npers	,0%	,2%	,0%	2,8%	2,6%	8,3%	,6%
	35_to_39	Count	0	0	0	0	6	0	6
		% within npers	,0%	,0%	,0%	,0%	15,4%	,0%	,5%
	40_more	Count	0	0	0	1	1	1	3
		% within npers	,0%	,0%	,0%	,7%	2,6%	8,3%	,3%
	5_to_9	Count	148	127	29	25	2	3	334
		% within npers	51,6%	25,9%	20,4%	17,6%	5,1%	25,0%	30,0%
Total		Count	287	490	142	142	39	12	1112
		% within npers	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

- Example of crosstabulation using SPSS; note the low number of cases in some cells

Choice of correlations

- The survey data is processed using the SPSS software
- for instance, for computing the number of showers, we detected:
  - p(nbshowers | nbpeople): the number of showers mainly depend on the number of people
  - p(nbpeople): for generating a population, we will use the frequency distribution of the number of people
- Feedback:
  - on small samples (even for 700 households !), we have to rely only on first-order correlations
  - The use of aggregate statistics facilitates the correction of aberrant cases (by manually correcting some probabilities)

Application

- Algorithm: monte-Carlo sampling:
  - for each household to create
    - randomly select the value of the lifestyle according to p(lifestyle),
    - randomly select the value of the number of showers according to p(nbshowers | lifestyle)

- Does work, but:
  - we have to encode the generation of the population by ourself; this increase the risk of programming errors or artifacts
  - it provides no solution for the generation of the network according to agents' attributes
  - => we explore the use of a standalone generic tool for the generation of networked populations [2]

Second approach:

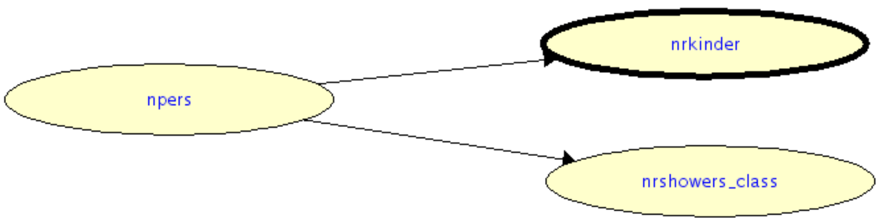
use of a generic tool to generate both attributes and a network

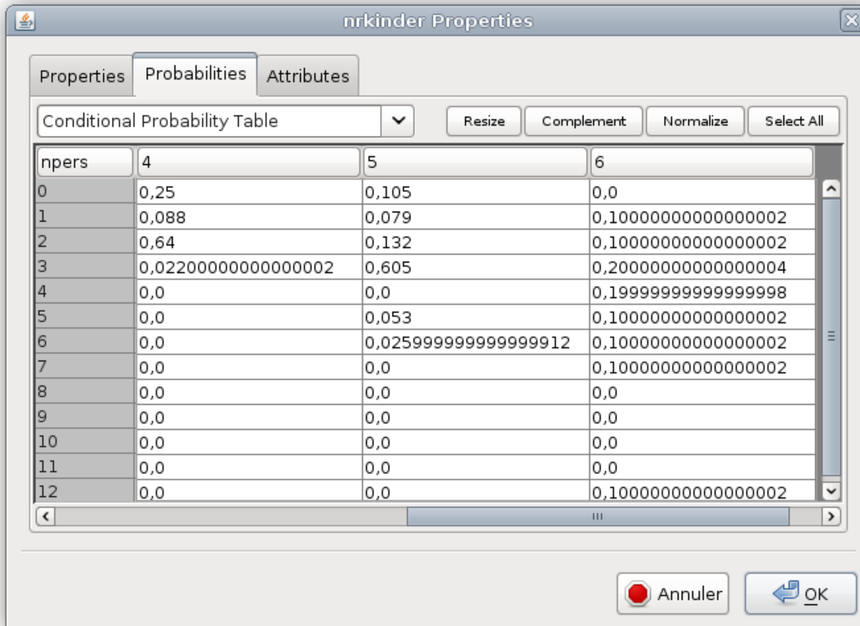
The YANG approach

- The YANG approach [2] is devoted to the generation of synthetic networked populations.
- It provides:
  - a meta-model for describing the attributes of individuals and probabilistic generation rules for creating links according to the properties of agents
  - an algorithm for the generation (we ignore the generation of networks in the frame of this study), and and open-source graphical tool for using it
  - a measure of generation errors

Encoding of attributes

- Attributes are encoded as random variables with conditional probabilities into a Bayesian network (same principle with a graphical representation)





npers	4	5	6
0	0,25	0,105	0,0
1	0,088	0,079	0,10000000000000002
2	0,64	0,132	0,10000000000000002
3	0,022000000000000002	0,605	0,20000000000000004
4	0,0	0,0	0,19999999999999998
5	0,0	0,053	0,10000000000000002
6	0,0	0,025999999999999912	0,10000000000000002
7	0,0	0,0	0,10000000000000002
8	0,0	0,0	0,0
9	0,0	0,0	0,0
10	0,0	0,0	0,0
11	0,0	0,0	0,0
12	0,0	0,0	0,10000000000000002

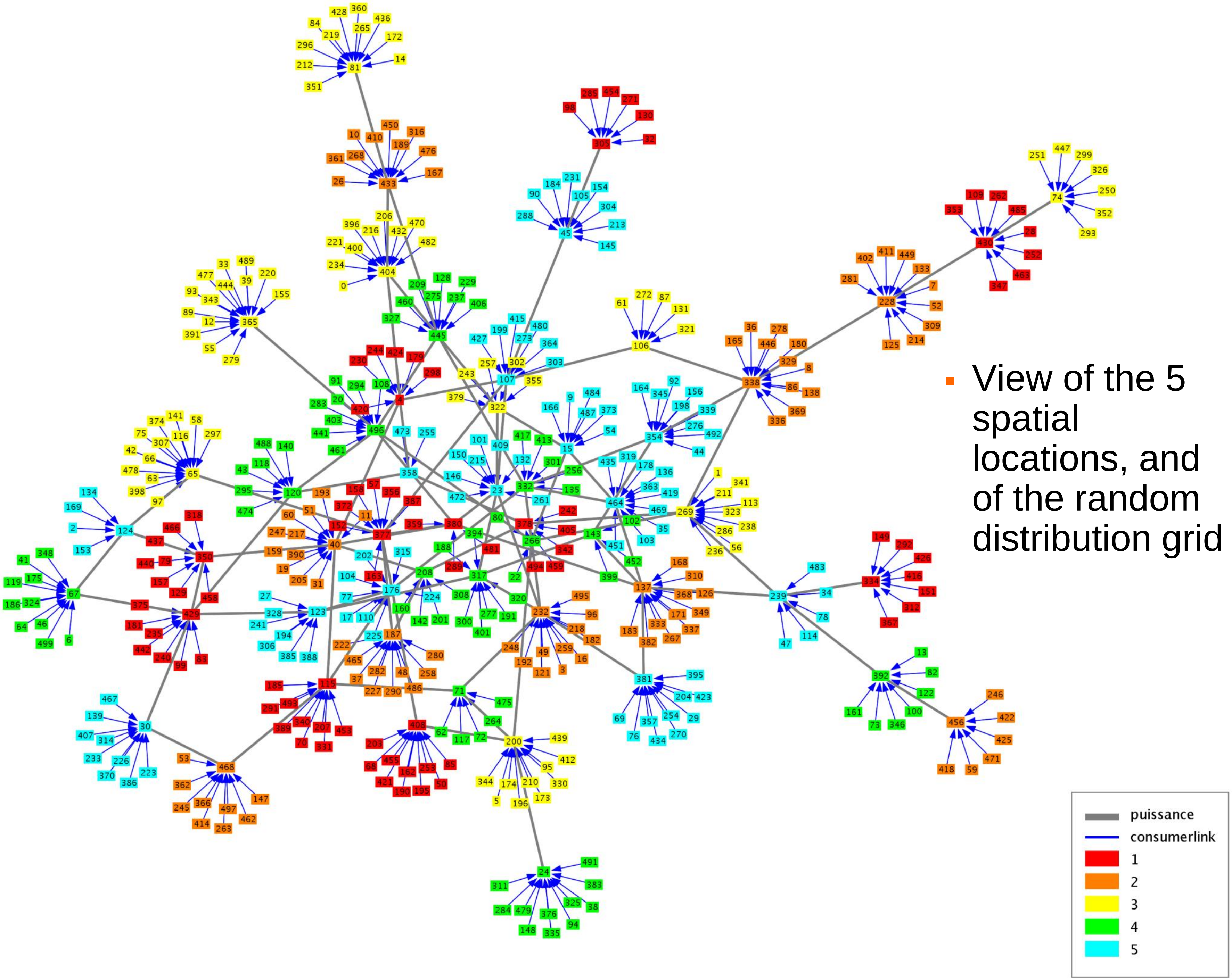
- Then YANG uses the same Monte-Carlo principe as ours for generating the population

Generation of networks

- principle: households are not positioned randomly over the distribution network; for instance, rural areas often contain bigger households
- YANG enables the description of rules for creating links; these rules describe the probability to create a link given the attributes of individuals

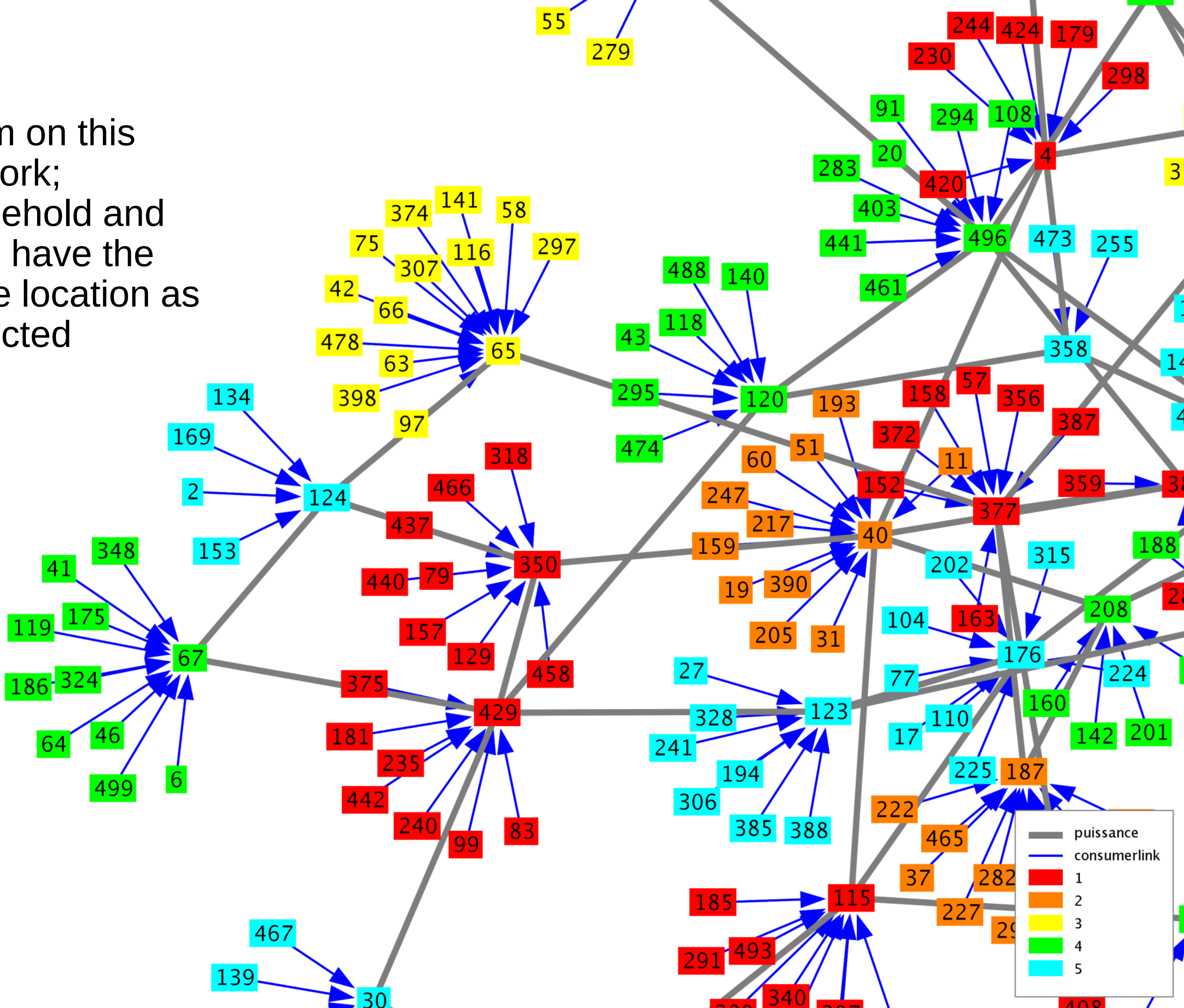
- First experiment:
  - we add agents that represent distribution hubs of the network,
  - all the agents are spatialized over 5 fictive spatial areas,
  - households are connected to distribution hubs of their area,
  - distribution hubs are interconnected by a random electric grid
  - household size depends on the spatial location

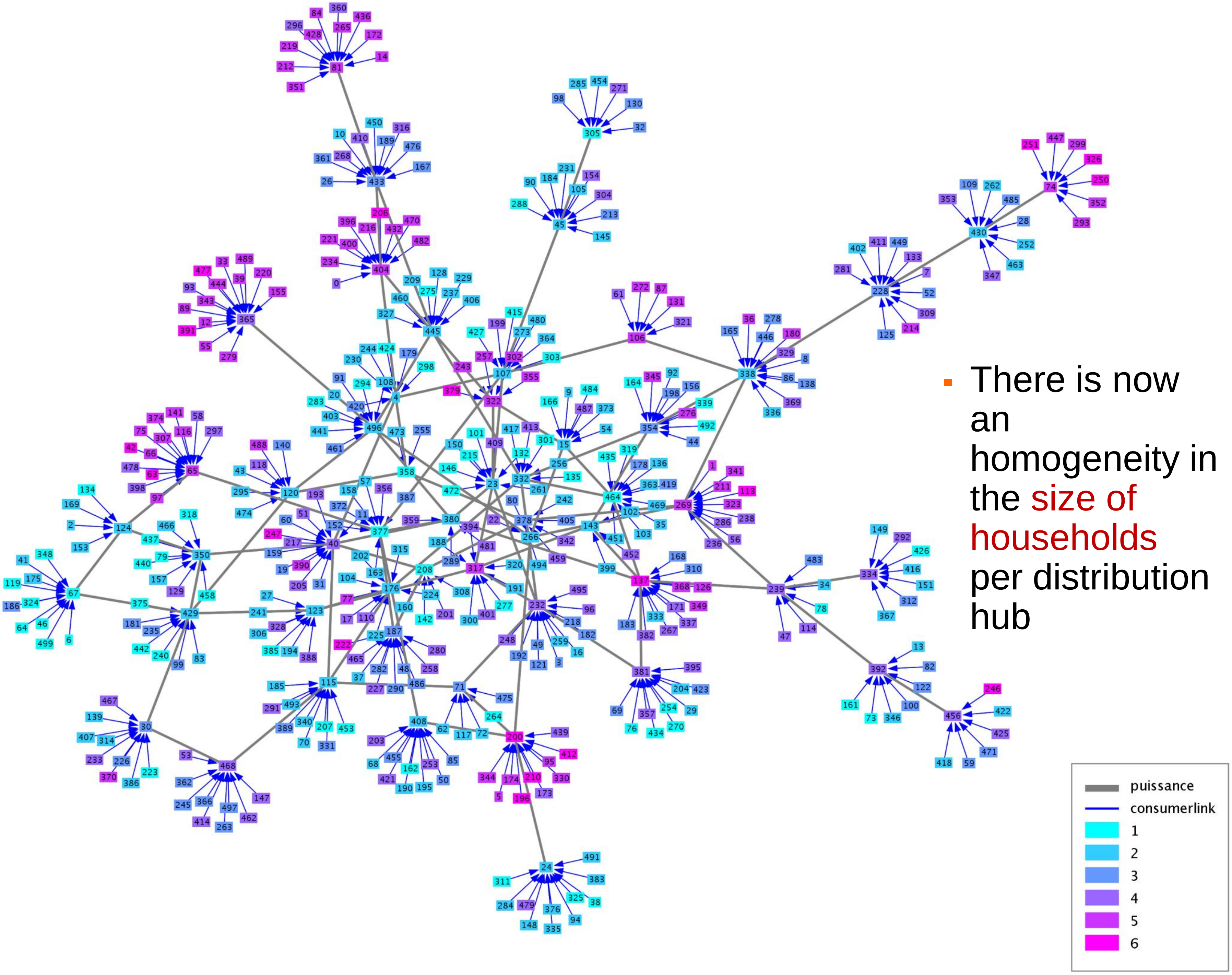
- Let’s browse an example of generated network



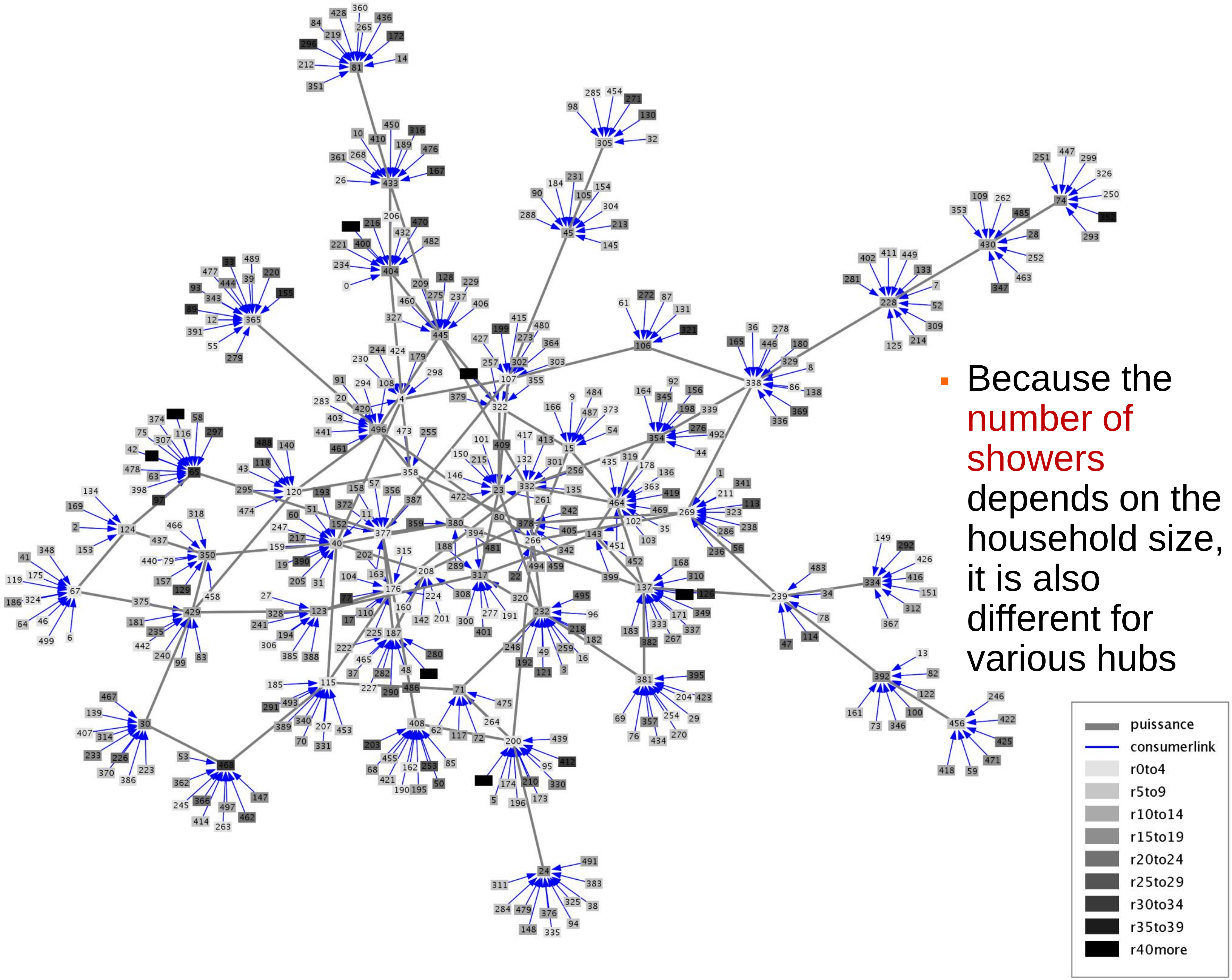
- View of the 5 spatial locations, and of the random distribution grid

- Zoom on this network; household and hubs have the same location as expected





- There is now an homogeneity in the size of households per distribution hub



- Because the number of showers depends on the household size, it is also different for various hubs

Discussion

and future work

Discussion

- In short:
  - the use of aggregate statistics facilitate the correction of data (reduce the statistical impact of aberrant cases)
  - however, on such a small sample, we can rarely correlate more than two variables
  - the use of a generic tool facilitates prototyping of networks by positionning the agents according to their characteristics

- This opens the way to the reproduction of load curves at the meso level (hubs, spatial areas);
- Next steps:
  - retrieve data on the correlation between household types and their location over a distribution network,
  - reproduce the differences in lifestyles at this local scale (heterogeneity in building isolation, work periods, etc.) and its impact on the load curve

Thanks.

questions/comments/critics
warmly welcomed !