Chimera Networks Generation

Venera Tomaselli - Giulio Giacomo Cantone
University of Catania

What kind of problems are we trying to tackle?

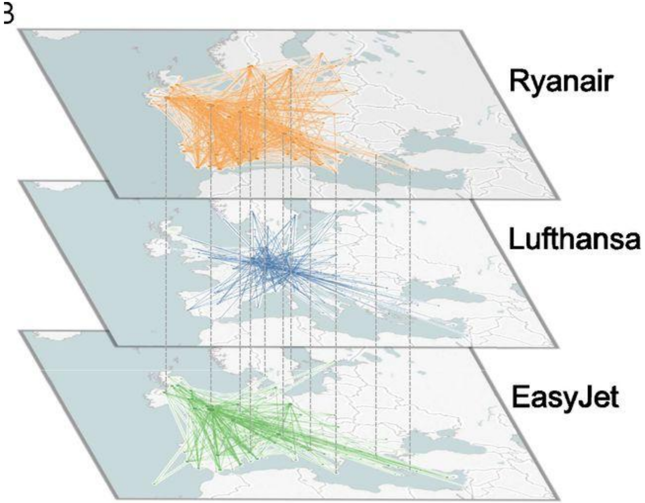
*A sufficient number of people randomly sampling 100 followers from Twitter accounts with **huge leads of followers**, can **properly classify** those followers as bots or real humans, and then true proportion of bots can be estimated.*

- Elon Musk, May 2022

But this will not work. Yet, social connections have untapped potential for social research.

We know that **it's wrong** to recruit respondents to our questionnaires among our Facebook's friends**, but** we wish it was not, because **technology can make the job easier**.

Networks with multidimensional layers



Wu et al. (2019), PNAS

A layer is the value of a nominal variable attributed to the edges.

The formation and structure of connections in multidimensional networks has **at least 1 mechanism per layer.**

Chimeric Networks try to represent complex networks (multiplex)...
...from the perspective of Population Studies

E.g., the connection with my mother is different (qualitatively and structurally) from connection with my friend.

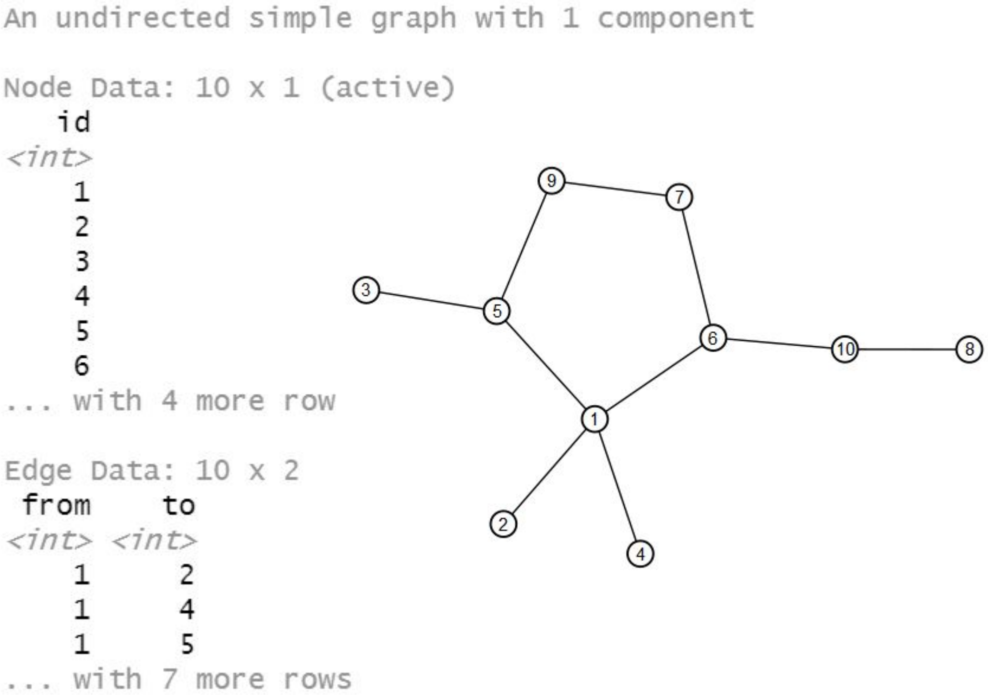
Not always these layer differences are caught through snowball sampling and respondent-driven sampling.

The research on multi-layer networks does not always answer to issues of social scientists.

Networks with Tidygraph

Tidygraph is a R package that totally rewrites “igraph” to represent networks as 2 related tables (tibbles).

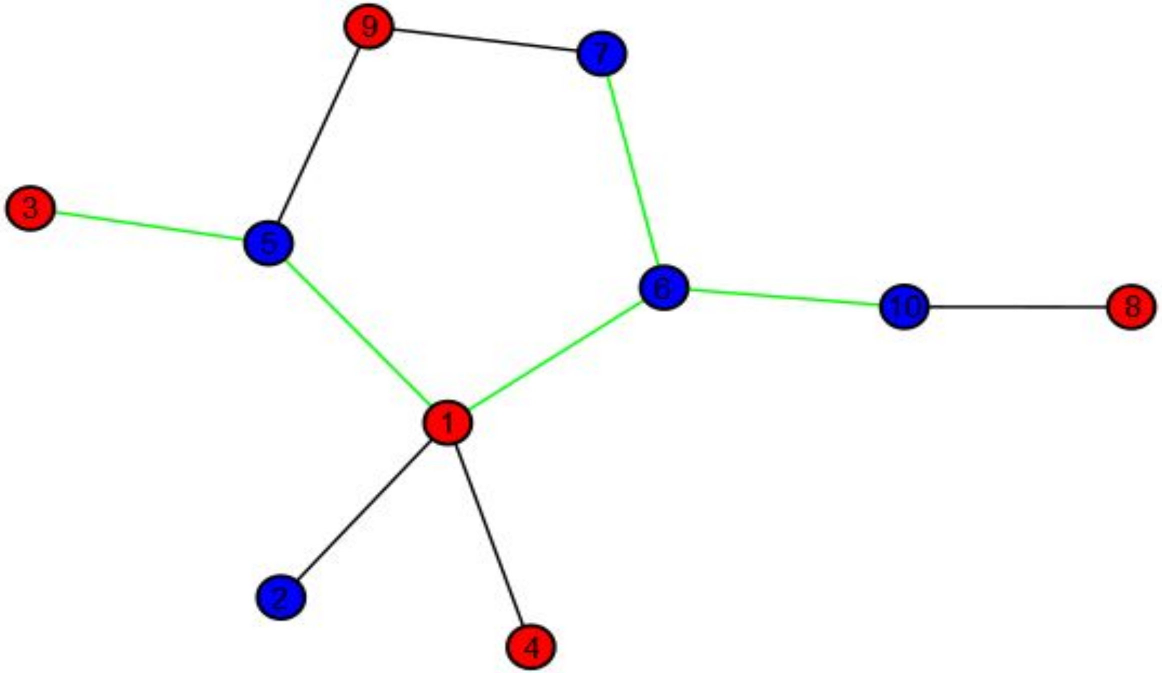
Recently social scientists and statisticians (Crane 2017) advocated for representing networks as tables.



```
An undirected simple graph with 1 component

Node Data: 10 x 1 (active)
   id
<int>
    1
    2
    3
    4
    5
    6
... with 4 more row

Edge Data: 10 x 2
 from    to
<int> <int>
    1     2
    1     4
    1     5
... with 7 more rows
```

Tidytext: very easy to add variables

```
An undirected simple graph with 1 component

Node Data: 10 x 3 (active)
   id  Wage Gender
<int> <dbl> <chr>
    1 1545. F
    2 1023. M
    3 1402. F
    4 1963. F
    5 1741. M
    6 1202. M
... with 4 more rows

Edge Data: 10 x 3
 from    to Friendship
<int> <int> <chr>
    1     2 Old
    1     4 Old
    1     5 Recent
... with 7 more rows
```



The intuition for Chimeric Networks

- If nodes of networks can be represented as one multivariate table
- …then variables (attribute) can be modeled as outcomes of a multivariate model (inferential or generative)
- Chimera Networks have **one characteristic attribute (y)** of the node that is **parameterised with** other attributes (x_*)

$$y = f(x_1, x_2, \ldots)$$

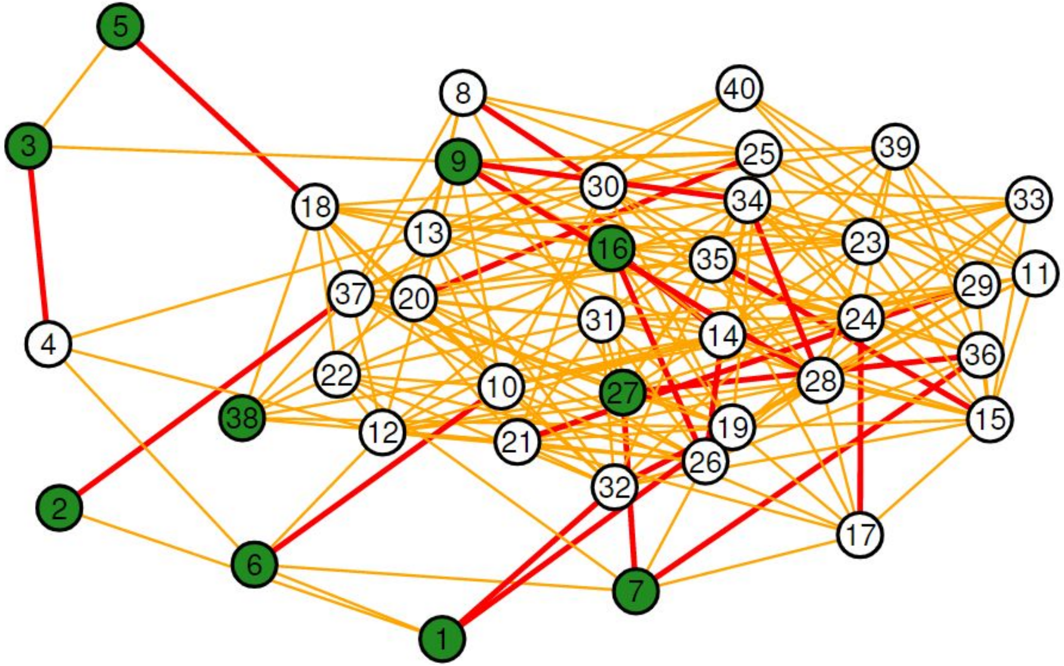
Questions for research

Question 1: how are distributed other *x*-variables and how *x* depends by statistics across their links?

E.g., does my wage depend by the number of my friends?
also, by how old are the friendships?

Question 2 : if *x-variables* are structurally dependent by the connections (e.g., preferential attachment), can these ties also *y = f(X)* be dependent by the connections of the node? How?

Chimeric network modeling helped us to generate toy models of cross-sectional epidemic networks of smokers and non-smokers



Toy model with few nodes.

Green nodes are smokers.
Red edges are household connections.

The code scaled up with almost instantaneous generation up to 100k nodes.

Parameters of the toy model of smokers

The binary smoker/non-smoker status in the nodes depends by few parameters.

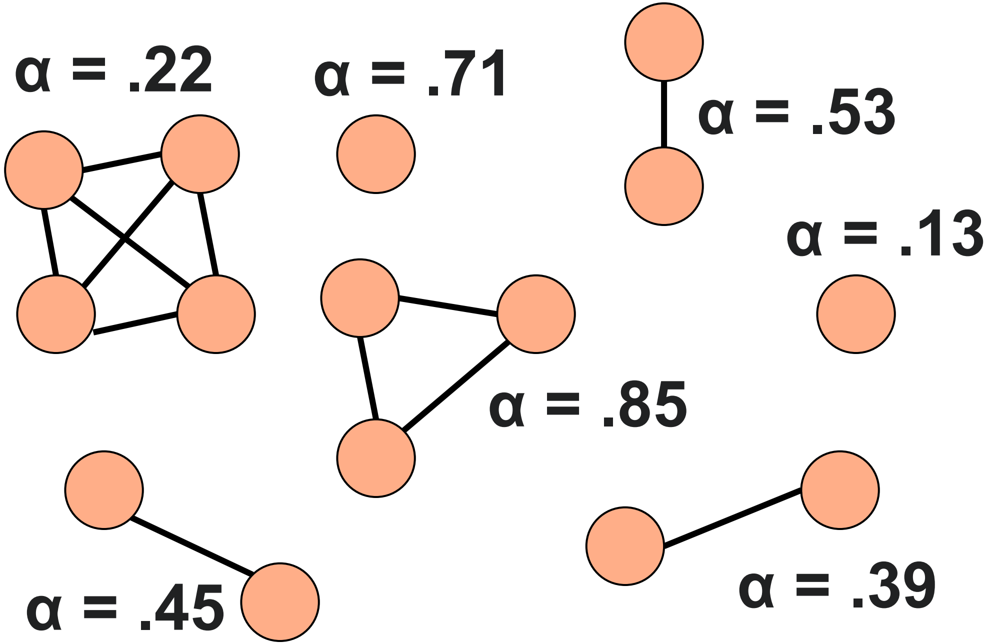
We called this particular Chimera Model the “Cliques-and-blocks”

1. A parameter fixed within “households” (highly validated hypothesis) in the *unit interval* (0,1)
2. A random ordinal distribution associated to blocks. Higher values of blocks represent lifestyles with high level of risk of smoking status
3. 1. and 2. are mixed into an individual level of risk *p* (still unit interval)
4. Mixture model for y, with Bernoulli variance $p(1-p)$

Households are modeled as a "archipelago of cliques"

Each clique, being a household, has a fixed value (0 *within* variance) for all the members of the clique.

Across different households, the value is random (non-0 *between* variance).



Now a second layer of blocks is “grafted”

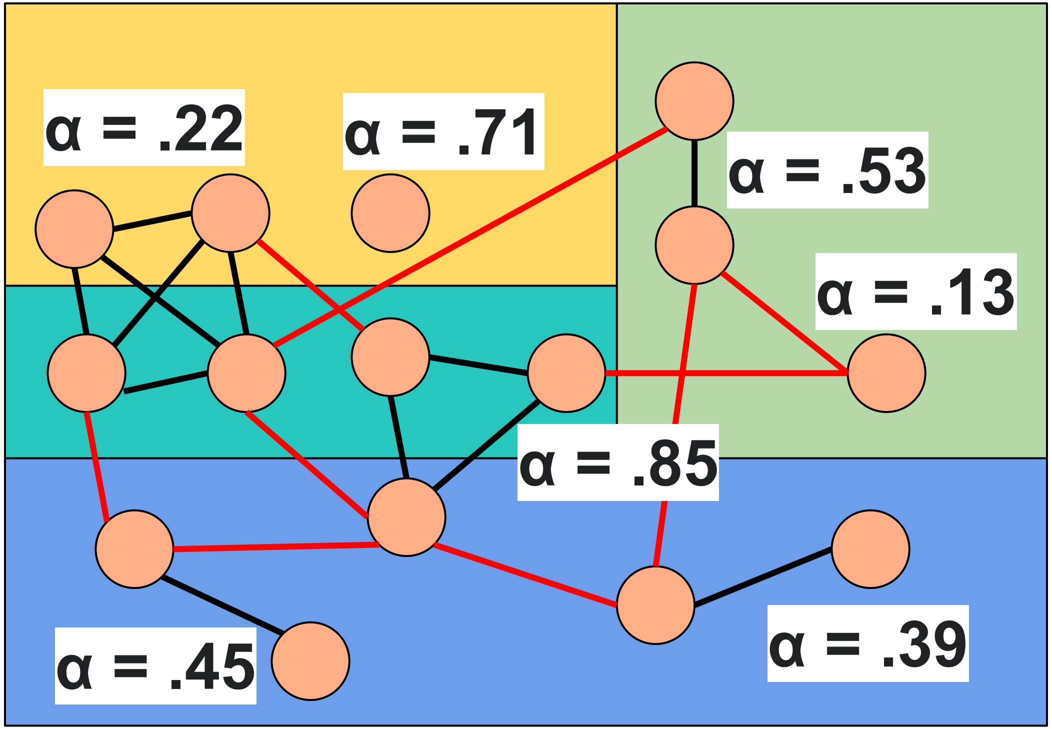
β_1 = .8

β_2 = .6

β_3 = .4

β_4 = .2

$$p = f(\alpha, \beta)$$



Structural dependency of Y

We achieved:

- Usually networks are generated procedurally. Chimeric models can procedurally generate **layers** but then…
- …once all the layers are aggregated in the edgeset, values in Y (e.g. smoker/non-smoker) are not tied anymore to a procedural structure. This makes generation more structurally flexible.
- E.g., in the smoker networks we generated homophily between smokers without modularity, i.e. excessive clustering of smokers

Future developments

- Chimeric Networks are an attempt to represent complex networks with more traditional statistical models (mixture models, possibly multilevel models). We hope this to be helpful for population studies, since…
- … Networks are populations! But most of the “hard science” is focused on high-level dynamics and not on problems of unbiased sampling, etc.

References

Harry Crane - Probabilistic Foundations of Statistical Network Analysis

Also:
Raffaele Vacca (UniMi) applied Mixed models to network data.