ON NULL MODELS FOR CONTAGION EFFECTS IN MULTIDIMENSIONAL NETWORKS

Giulio Giacomo Cantone, Venera Tomaselli

1. Introduction

A network is a structure of entities that can be connected to each other. Each network has two distinct sets of entities: nodes (or, vertexes) and edges (or, links). Nodes represent punctual entities while edges represent the *connection* between two nodes. One can say: node *i* and node *j* are *connected*. This implies that not only *i* and *j* exist as entities in a network, but also that an edge (*i*↔*j*) exists. If the connection works only on one side but not on another, then the edge is *directed* and it is represented as (*i*→*j*). The set J_i of all *j*-nodes connected to *i* is the ego-network (*of first order*) of *i*. A network allowing directed edges is a directed network. A count of edges is referred as *degree*, with the letter *k*. To say that the node *i* has *k* = 3 means that *i* has 3 connections with other nodes.

Networks are represented as mathematical graphs, that are special sets:

$$G = (V, E) \quad (1)$$

where: G is the graph, V is a set of nodes, and E is a set of edges. This representation is convenient for the abstraction of structural proprieties of classes (or “ensambles”, a word borrowed by Statistical Mechanics) of real networked topologies. In graph theory, G is often processed as an adjacency matrix: a square matrix where the indexes are the nodes and the elements are the edges. For simple graphs, 0 in the adjacency matrix represents absence of edge, while 1 would represent presence. In *weighted graphs*, the value of the element of the matrix can vary and the variation would indicate the difference in size, relevance, etc. of the relation, keeping 0 as the reference value for absence of relation. Matrices are notoriously fast structures for computation. This is a useful feature both for analysis and visual representation of networks. Indeed, in computationally intensive applications as in Statistical Mechanics, Bioinformatics, or Big Data, the ‘networks-as-graphs’ paradigm prevailed. However, as noted by Crane (2018), traditional graphical tools are not always appropriate to represent real networks and in particular social networks as networks of social actors.

The alternative representation of networks would be a relational database of 2 tables. One table has nodes as rows, the other edges as rows. In the table of the edges, one column references the first node of the pair (or for directed networks, the *sender* of the connection), and another column references the second node of the pair (or, the *receiver* node, in directed networks). Each table may have many columns, each column representing an observed attribute (a variable) of the entity.

In the literature on networks, attributes of the nodes are not particularly problematic. Indeed, as long as edges have no attributes, a network is not properly multidimensional; but if edges are nominally differentiable at least through an attribute, then the network is multidimensional. Terminology is not always established (Barrett *et al*., 2012), but a core concept is the *layer*. The layer is the subset *g* of the graph G such that all the edges of *g* (E_g) share a common nominal value in one attribute. All the nodes (V_g) connected through E_g fall within sub-graph *g*, too. Since the layer is associated with a (nominal) value of an attribute of the edgeset, often the word *layer* recalls simply that value (Kivelä *et al*., 2014, Dickison *et al*., 2016).

The present manuscript is about inference on multidimensional network data. The theoretical issues of correlations among many variables are discussed referring to the approach of neutral models for statistical testing of hypotheses. In this aim, a generative technique of multidimensional networks is proposed.

2. Multivariate models for multidimensional networks

Mathematical representation of multidimensional networks is problematic because the adjacency matrix is insufficient to represent layers. Advanced mathematical solutions to represent layers involve the employment of tensor structures, but tensor algebra is much less known than matrix algebra. Its application could alienate researchers to pursue valid research questions involving representation of social groups as networks (demography of families, organization studies, marketing, etc.). The database representation has a benefit here: it makes easy to represent both variables (attributes) of the nodes and *layer* values of edges as additional columns of the tables. This allows to run traditional multivariate analysis models, as multilevel models, across nodes and edges (Vacca *et al*., 2019).

Multilevel models are employed in population studies as a tool to avoid ecological fallacies (Gnaldi *et al*. 2018). Multilevel models account for cases where observations are *nested* within other observation, for example: in a list of high schools, these are nested within towns, and any analysis of the variance *between* high schools needs to account for the variance *between* town. Multilevel models (or, hierarchical models) are a special class of *mixed models with fixed values*.

The application of multivariate models to networks is important for the demographic data analysis. For example, in 2022 has been completed the mapping of the whole population network of the Netherlands (van der Laan *et al*., 2022): a database of more than 14 million nodes representing people inhabiting Netherlands in 2018, connected through more than 1.4 billion edges. One attribute identifies 5 macro-layers of edges: family, household, neighbours, schools, and work. But then, for each of these layers are specified more detailed classes of relationships as additional variables in the database. For example, among the edges in the layer that are labelled as "family", are nested classes of directed relationship as "parent of", "cousin of", etc (Table 1).

Table 1 – *Example of a social network represented as a relational database.*

Node ID	**Name**	**Surname**	**Job**	**High School**	**…**
1	John	Doe	ABC Inc.	Alighieri	…
2	Mary	Smith	FinanzGroup	Cervantes	…
3	Jane	Doe	NA	Shakespear	…
4	Paul	Jones	NA	Shakespear	…
5	Peter	Taylor	ABC Inc.	Shakespear	…
6	Luke	Brown	FinanzGroup	Alighieri	…
…	…	…	…	…	…

Edge ID	**From**	**To**	**Macrolayer**	**Microlayer**	**…**
1	Node 1	Node 2	Family	Married to	…
2	Node 1	Node 3	Family	Parent of	…
3	Node 1	Node 5	Work	Manager of	…
4	Node 2	Node 1	Family	Married to	…
5	Node 2	Node 3	Family	Parent of	…
6	Node 2	Node 6	Work	Manager to	…
7	Node 3	Node 1	Family	Child of	…
8	Node 3	Node 2	Family	Child of	…
9	Node 3	Node 4	School	Classmate of	…
10	Node 4	Node 3	School	Classmate of	…
11	Node 5	Node 1	Work	Managed by	…
12	Node 6	Node 2	Work	Managed by	…
…	…	…	…	…	…

In Table 1 microlayers are properly nested within macrolayers. Mixed models of networks data are found in recent developments of applied network analysis to Economics (Jochmans and Weidner, 2019). With mixed models, it is possible to combine co-occurrences of different layers and attributes of nodes into very rich multivariate models with fixed effects, too. An example: *i* nodes are associated to a *y* numeric value standing for body weight. The researcher is interested in correlation of *y* with the average $\bar{y}_J$ in the ego-networks J_i of each *i*-node. But $\bar{y}_J$ has a strong dependency on *j* nodes being men or women, so the model must correct the estimate for this fixed attribute of the j-node. Then, the researcher can observe the differences in the coefficients across layers of relationships, for example family *vs*. co-workers. These values (family, co-workers, etc.), differently than gender, are not *fixed* per *j*, since each *i* has different relationships with *j*, hence the reference to *mixed* models. Panel models are a special case of mixed models with a differentiation in time (*lagged* regression) or just with a time-point (for example, the month) as a *fixed* control variable.

3. Models of direct contagion in network data

The relevance of a regression model of the attribute y_i of *i* (ego) on the average $\bar{y}_J$ of its J_i ego-network implies that there are correlations between the value of y_i and y_j that are scientifically not trivial. The presence of positive correlations is also called assortativity, and negative correlations lead to disassortativity. Assortativity is also structurally tied to others indicators of correlation, like network clustering, etc. According to Christakis and Fowler (2013), assortativity has three explanations other than chance:

- i-egos have a preference to associate subjects with similar attributes. Sometimes this preference is called homophily, but this term is also confused with assortativity itself;
- i-egos and their J_i ego-networks might jointly experience unobserved simultaneous exposures to common omitted variables, confounding the correlation;
- and J_i ego-network induces an effect on *i*. These explanations are not mutually exclusive, but they are hard to disentangle in causal models.

Christakis and Fowler (2013) proposed an explicit model to estimate how a change over time in Y can be attributed to contagious effects:

$$g\left(E\left(Y_{i,t+1}\right)\right) = \alpha + \beta_{i,t} y_{i,t} + \beta_{J,t} \bar{y}_{J,t} + \beta_{J,t} \bar{y}_{J,t+1} + B(Z) \tag{2}$$

where:

- Y is the attribute under hypothesis of contagion
- *i* is the ego
- J is the set of its neighbours.
- *t* is a time-point, assumed as *fixed* in the model
- Z are controls variables, assumed as *fixed* in the model
- β is the coefficient of the type regression
- B are the vectors of the coefficients of the controls.

The (2) expresses the link function *g* in a generic form so it can be adapted for different data types (linear for continuous Y, logit for binary, etc). It is actually a *panel* model that lies on the same methodological foundations of *mixed* models. Christakis and Fowler (2013) acknowledged that (2) is still problematic if omitted variables are not controlled within the set Z. In other words, the problem of identification of spillover effects in networks is analogous to the notorious problem of ignorability of missing variables (Imai *et al*., 2010). A non-parametric approach in modelling contagion is in Aral *et al.* (2009). They propose to statistically match nodes from two groups:

- the null effect group of i_0 such that *i* has less than k_0 ties who shifted from y_0 to y_1 between t_0 and t_1, for example these are friends who adopted a new status y=1 in t_1 for a binary Y;
- and the alternative effect group i_0 with more than k_0 ties.

The matching algorithm minimises the global differences in all the Z between the element i_0 and the element i_0. Aral *et al.* (2009) reached the conclusion that the coefficient of the contagion effects $\beta_{J,t}\bar{y}_{J,t}$ in (2) overestimates the effect of a factor roughly ~2.

Shalizi and Thomas (2011) generalise the issue on the origin of assortativity for multidimensional networks. The idea is that more than homophily can be driven towards more than one attribute, and the co-existence of contagion dynamics and multidimensional preferential attachment would make very hard to properly estimate contagion effect. The simple example involves the difference between:

- attribute assortativity: nodes show a tendency to cluster around values of one attribute. If this attribute is *strictly nominal* these clusters will approximate sub-graphs, few ties bridge between the clusters, and layers emerge naturally. If the attribute is ordinal or metric, nodes are attached to nodes with similar values of the attribute.
- degree assortativity: a special case of a numerical attribute assortativity. It is observed when ties of the networks grow over time and new nodes have a preference to attach themselves to the more connected old nodes.

Attribute assortativity does not imply degree assortativity but for any non-Uniform distribution of an assortative attribute, it would be observed a significant positive correlation between *k* and $\bar{k}_J$ even in absence of a preference of new nodes to attach to old nodes. This is, hence, a *technical* value of degree assortativity, that is also the true null of the observed value, conditional to the observed value of attribute assortativity (Crawford *et al*., 2018). If the attribute that is the *true* source of assortativity is unobserved, then the technical effect on the coefficient could be confounded as misleading evidence for a mechanism of preferential attachment, in absence of both preferences and agency in the nodes. This is a simple case for one-dimensional networks. The general case for multidimensional networks assumes that general homophily (that is, assortativity across many attributes) confounds processes of influence at agent level.

4. Neutral models

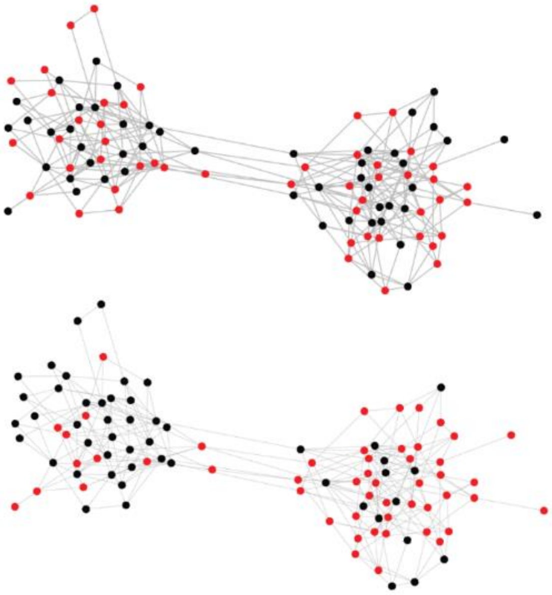
Neutral models are simulation models that include all the relevant features of complex dynamics, except one, that is suppressed (*neutered*). The absence of the neutered feature can be considered as a treatment. They were originally developed in biology to test evolutionary hypotheses. Their role in epidemiological modelling is analogous to null hypothesis in statistical testing (Gotelli and McGill, 2006).

In order to introduce the connection between proprieties of iterated simulations and null models, it is worth to mention a toy model (Figure 1) in Shalizi and Thomas (2011) because it allows to explain why pre-existing structural factors (e.g. religion) mask and confound agency (e.g. why capitalism spread?).

Figure 1 represents this argument: the nodes are cities and the red colour indicates a high concentration of factories; the edges are natural trade routes; the two clusters are a Catholic nation and a Protestant one. The network at top is observed at Year 1600, the bottom one at Year 1900. Ideally, the argument of the authors is that there is no need of a causal impact of religion to cause the polarised spread of capitalism towards Protestant (or, Catholic) cities: natural trade routes (i.e. topology) can explain it already in its neutered state. In the toy model, either of the two clusters (Protestant and Catholic nations) could experience with the same probability a global rise in factories (red nodes). Reiterating the model many times, half of the time Catholics would experience the spread of Capitalism, and the other half the Protestants. Religion would look a relevant factor for how the Capitalism spreads only because "history cannot repeat itself", or, alternatively, because trade routes are not accounted (omitted variables) in the original models of spread of industrialization. One can notice that this is the inverse case of confounding between structure and agency than the one presented in Aral *et al.* (2009).

Simulation models can use estimates of coefficients from a regression model as parameters for agency (or, contagion) in multidimensional networks, and then can differentiate parameterisation between the neutral model (null hypothesis) and the alternative parameterisation. Simulation models iterate until is possible to infer steady stochastic averages of the iterating time series. The series themselves are the result of the simulation. If results of the neutral model *vs*. the alternative show no significant differences, then it possible to conclude that the suppressed feature had a no causal role in the final output. However, if differences emerge in the time series, it is possible to characterise the causal role of the feature in the model.

Figure 1 – A 2-clusters network: before (top) and after (bottom) a contagion process.



A network of 100 nodes structured in two clusters (communities). It is represented before the start (top) and after the end (bottom) of process of contagion. For 1000 iterations, a node is picked. With a probability it assumes the colour of another adjacent node. This probability is independent to the cluster of the node, hence the propensity for a node to adopt the colour of the majority of its cluster is null (neutered). Before the first iteration, there is no association between clusters and colours, but after the 1000th iteration, clusters and colours are correlated even in a neutral model. This as a by-effect of the fact the clusters exists, even in absence of a link to the probability to change colour (Shalizi and Thomas 2011, p. 24).

Can this methodology be extended for multidimensional network? Yes, with a *caveat*. There are two general approaches for generation of a multidimensional network: formation through union of simpler networks and procedural formation. The latter implies that the network growths iteratively and new nodes, edges, and layers happen as statistical events over time, with specified probabilities. Procedural multidimensional models have issues regarding correct posterior parameterisation of homophily in neutral models (Dickison *et al*., 2016). The procedure of formation implies a micro-model of *agency* of the new nodes which have preferential attachment (whatever it is) towards old nodes. Any parameterisation of *agency* would bring *technical* alterations in the joint parameterisation of the null hypothesis, for the reasons explained in section 3. In other words, the implicit micro-model of multidimensional agency of nodes in attachment could mask and confound the model of contagion that is tested, instead. For any non-specific hypothesis on the agency in attachment, union of simpler networks is a safer choice.

5. Chimera networks: generating multidimensional networks as union of simpler models

We refer to generative methods for multidimensional networks as ‘chimera’. The method involves a union of layers, disjointly generated before, hence truly statistically *independent*. The idea is to generate only one set of nodes, and many sets of edges. Each set of edges is a layer of the chimera edgeset. If the networks are recorded as tables, the union of the edges can be coded with commands common to all programming languages for data analysis, as *append* or *join*[1]. Layers are generated through a one-dimensional technique, that can be procedural or not. In both cases, parameters for formation of the layer as one-dimensional edgeset should be kept as an attribute in table of the nodes, as designed in section 1.

For example, a layer can be generated through a stochastic blockmodel (SBM). This is not a procedural generative model. In a SBM, each node is assigned to a block. The probabilities of two nodes to be randomly connected are parameterised through a *mixing* matrix: a square matrix mapping all the possible combinations of pairs of blocks (Faust and Wasserman, 1992, Latouche *et al.*, 2011). The information about the *block* of nodes *i* is stored as an attribute of the nodeset in the databases.

One of the benefits of the chimera method is that it allows unbiased generation (i.e. draw) of random **y** attributes (characteristic attribute of the chimera), parameterised after other **x** attributes that are inherited from the disjoint layers: one generates many layers, each is a different SBM; the nodes will be associated to a set of variables D_X: X_1, X_2, … X_d, each variable being a vector of parameters regarding a layer. It follows that:

$$f_Y(x_{1,i}, x_{2,i}, \dots x_{d,i}) = y_i \tag{3}$$

allows to model the variable Y of a *characteristic attribute* of the chimera network.

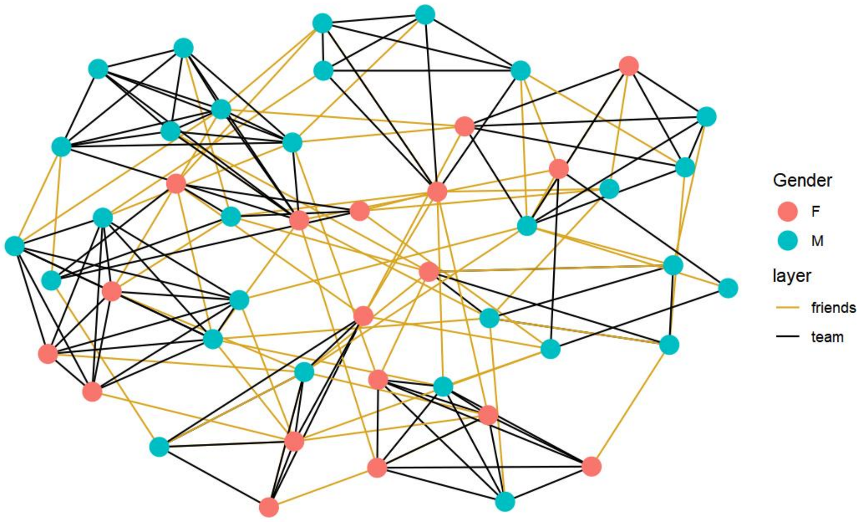
In the example (Figure 2), a multilayer toy network of 47 nodes is generated through union of layers. It represents a workplace where people are co-workers grouped in teams. Each member of a team is connected to any other member of the same team (the black edges in Figure 2). This formation is called ‘archipelago’. The expected team size is distributed as

[1] Our personal suggestion is to adopt the command *tidygraph::graph_join()* in language *R*. The package *tidygraph* has been developed as a wrapper of software *iGraph*. *Tidygraph* re-arranges the structure of a mathematical graph object as a relational database.

$$Poisson(\lambda = 3) + 1 \tag{4}$$

so teams with no members are not allowed in the model.

Figure 2 – Chimera network: co-workers.



Nodes are then randomly split between women and men. Genders work as *blocks* for a SBM layer. The SBM layer represents friendships outside the team, and it is parameterised with an expected value of connections *per* node equal to 3. While in the toy model the average is indeed ~3, it can be noticed in Figure 2 that most of light edges (*friends* layer) connects two woman (F) nodes. This is the result of the parameterisation of the mixing matrix of the SBM, that is:

Table 2 – *Mixing matrix of Stochastic Blockmodel in 2-blocks*

	F	**M**
F	.6	.1
M	.1	.2

With an expectation of $\mathbf{3 \cdot 47 = 141}$ friendships, with .6 probability a uniformly random drawn woman is attached to any other uniformly random drawn woman, with .2 a uniformly random drawn man to a uniformly random drawn woman (or, *viceversa*), and with .2 a uniformly random drawn man to any other uniformly random drawn man.

A chimera allows to perform easily many neutral models. Given D_X variables, one X *explanans* is selected to be tested.

Then, it is possible to randomly shuffle all and only the edges of the layer associated to X, while keeping the values $x \in X$. This operation is equivalent to generate a *null* model where the *agency* of agents does not depend on the correlation between their social topology and the social structure of the layer (e.g. airports being built near polluted areas). It is possible also to randomly shuffle the values $x \in X$. This operation keeps the topology, but assumes that there are no differences in an *explanans* (e.g. pollution, as if pollution was the same in all the areas). Furthermore the models enable neutralizing the effect $f_X(X)$ on Y altering the function in (3). This operation keeps both the topology and the feature, but re-models an alternative scenario of the impact of the feature, i.e. to test not the effect of the feature X on the outcome Y, but the sensibility of Y to the analytical choices regarding how to model f_Y.

References

ARAL S., MUCHNIK L., SUNDARARAJAN A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks, *Proceedings of the National Academy of Sciences*, Vol. 106, No. 51, pp. 21544–21549. https://doi.org/10.1073/pnas.0908800106

BARRETT L., HENZI S.P., LUSSEAU D. 2012. Taking sociality seriously: The structure of multi-dimensional social networks as a source of information for individuals, *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 367, No.1599, pp. 2108–2118. https://doi.org/10.1098/rstb.2012.0113

CHRISTAKIS N., FOWLER J.H. 2013. Social contagion theory: examining dynamic social networks and human behavior, *Statistics in Medicine*, Vol. 32, No.4, pp. 556-577.

CRANE H. 2018. *Probabilistic Foundations of Statistical Network Analysis*. Oxford (UK): Chapman and Hall/CRC.

CRAWFORD F.W., ARONOW P.M., ZENG L., LI J. 2018. Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling, *American Journal of Epidemiology*, Vol. *187 No.*1, pp. 153–160. https://doi.org/10.1093/aje/kwx208

DICKISON M.E., MAGNANI M., ROSSI, L. 2016. *Multilayer Social Networks*. Cambridge (UK): Cambridge University Press.

FAUST K., WASSERMAN S. 1992. Blockmodels: Interpretation and evaluation, *Social Networks*, Vol. 14 No.1, pp. 5–61. https://doi.org/10.1016/0378-8733(92)90013-W

GNALDI M., TOMASELLI V., FORCINA A. 2018. Ecological Fallacy and Covariates: New Insights based on Multilevel Modelling of Individual Data, *International Statistical Review*, Vol. 86 No.1, pp. 119–135. https://doi.org/10.1111/insr.12244

GOTELLI N.J., MCGILL B.J. 2006. Null Versus Neutral Models: What’s The Difference?, *Ecography*, Vol. 29 No.5, pp. 793–800. https://doi.org/10.1111/j.2006.0906-7590.04714.x

IMAI K., KEELE L., YAMAMOTO T. 2010. Identification, inference and sensitivity analysis for causal mediation effects, *Statistical Science*, Vol. 25, No.1, pp. 51–71. https://doi.org/10.1214/10-STS321

JOCHMANS K., WEIDNER M. 2019. Fixed-Effect Regressions on Network Data, *Econometrica*, Vol. 87 No.5, pp. 1543–1560. https://doi.org/10.3982/ECTA14605

KIVELÄ M., ARENAS A., BARTHELEMY M., GLEESON J.P., MORENO Y., PORTER, M.A. 2014. Multilayer networks, *Journal of Complex Networks*, Vol. 2 No.3, pp. 203–271. https://doi.org/10.1093/comnet/cnu016

LATOUCHE P., BIRMELÉ E., AMBROISE C. 2011. Overlapping stochastic block models with application to the French political blogosphere, *The Annals of Applied Statistics*, Vol. *5*, No.1, pp. 309–336.
https://doi.org/10.1214/10-AOAS382

SHALIZI C.R., THOMAS A.C. 2011. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies, *Sociological Methods & Research*, Vol. 40 No.2, pp. 211–239.
https://doi.org/10.1177/0049124111404820

VACCA R., STACCIARINI J.-M.R., TRANMER M. 2019. Cross-classified Multilevel Models for Personal Networks: Detecting and Accounting for Overlapping Actors, *Sociological Methods & Research*.
https://doi.org/10.1177/0049124119882450

VAN DER LAAN J., DE JONGE E., DAS M., TE RIELE S., EMERY T. 2022. A Whole Population Network and Its Application for the Social Sciences, *European Sociological Review*. https://doi.org/10.1093/esr/jcac026

SUMMARY

Multidimensional networks are networks where edges are differentiated with different nominal classes, called layers. Inference of contagion effects has issues both in simple networks with only one layer and in multidimensional networks. However the inherent complexity of multidimensional networks makes almost impossible, at least with traditional approached based on regression models, a reliable inference of the "contagiousness" of a feature within a network. In the first part of the manuscript are provided introductory notions to run regression models and simulation models of multidimensional networks. The approach only requires knowledge of tabular data and mixed models of regression and not of tensor algebra, so the approach should be more congenial to social scientists. In the second part, it is introduced the concept of neutral model as a peculiar case of null model for statistical inference. Finally, given the aforementioned concerns, it is discussed why methods based on union of independent layers (chimera networks) are generally better than procedural model for parameterisation of neutral models of multidimensional networks. An example of chimera as a join of two blockmodels is provided.

Giulio CANTONE, University of Catania, giulio.cantone@phd.unict.it
Venera TOMASELLI, University of Catania, venera.tomaselli@unict.it