IL NUOVO CIMENTO Vol. 34 C, N. 5 Settembre-Ottobre 2011
DOI 10.1393/ncc/i2011-10903-1

Communications: SIF Congress 2010

Explosive percolation in correlation-based networks

D. Remondini(*)

Dipartimento di Fisica, Università di Bologna - Bologna 40127, Italy

(ricevuto il 31 Dicembre 2010; approvato il 24 Marzo 2011; pubblicato online il 30 Giugno 2011)

Summary. — We show briefly the features of a percolation transition related to the networks obtained from a correlation matrix. The most interesting behaviour of this transition, investigated by numerical simulations with different thresholding rules, is that it shows a much faster transition from the disaggregated to the clustered phase, that resembles what has been described as an "explosive" percolation. A comparison with the "classic" random network percolation is shown, together with some applications of these concepts to the networks obtained from real data, that behave differently depending on the data intrinsic structure.

PACS 87.18.Vf – Systems biology.
PACS 87.18.Wd – Genomics.
PACS 64.60.aq – Networks.
PACS 64.60.ah – Percolation.

1. – Description

The percolation transition in Erdös-Rényi (E-R) random networks is a long-standing example of phase transitions in networks [1]. Starting from a set of disjointed nodes, links are added randomly, and in the proximity of an average connectivity of one link per node, a transition is observed for which a unique connected component develops that occupies a finite fraction of nodes in the thermodynamic limit.

Very recently it has been shown that a slight modification of the growth rule can lead to a completely different behaviour at the transition point [2]. This transition has been called explosive, since it seems to produce a finite connected component all of a sudden, changing thus the transition properties. This discontinuity has been debated [3], but at least it can be stated that anomalous critical exponents occur at such a transition.

Network approaches have spread over several fields, from sociology to economics [4] and biology [5-7] to cite only some examples. In particular, in the study of times series data, approaches based on the correlation matrices have been applied [4,8], that seem to

(*) E-mail: daniel.remondini@unibo.it

provide robust results even when data are very noisy and the problem is ill-posed (when there are many more observed variables than time points available). This is a typical situation when dealing with recently available gene expression data, that can span the whole genome (10^3–10^4 nodes) but for which very few time points are available (10^1 is a typical order of magnitude [9, 8]).

We address the problem of percolation transition in the network obtained from correlation matrices, by thresholding of the correlation coefficients. The thresholding procedure may be useful for removing noise from the data, but a problem is that a predefined threshold cannot be defined, if not based on single-node characteristics (statistical significance of the correlation). Starting from randomly generated data, we observe an anomalous transition with features similar to the explosive percolation, thus providing more information about the "null model" to be compared with real data. Moreover, applying the same method to real data, we get some relevant information about the data structure, that seem to reflect features like modularity and non-random correlations.

2. – Percolation by thresholding

The "null model" is developed from a random Gaussian matrix X, (N M-dimensional vectors) defined as

$$X = \begin{pmatrix} \vec{x}_1 \\ \dots \\ \vec{x}_N \end{pmatrix}, \tag{1}$$

with $x_i \propto N(0, \sigma)$.

Hence we define the $N \times N$ correlation matrix C as usual:

$$c_{ij} = \frac{E[(x_i - \mu_i) \cdot (x_j - \mu_j)]}{\sigma_i \cdot \sigma_j}\,. \tag{2}$$

The correlation matrix gives us the chance to build a network in three different ways, according to the thresholding ($C_T > 0$) we choose

- correlation: link if $c_{ij} \geq C_T$,
- anticorrelation: link if $c_{ij} \leq -C_T$,
- both: link if $c_{ij}^2 \geq C_T$.

Taking different threshold values we get networks with a different number of links, in this way we follow the evolution of our networks just modifying the threshold value. We start with N isolated nodes and we finally reach a fully connected network.

The network dynamics is reflected by the formation of a giant component. We compared our models with the classical percolation problem and with a peculiar explosive percolation model due to a modification of the growth rule, the so-called Product Rule (PR [2]).

We have observed that, for the correlation-based percolation, the properties depend on the number of dimensions of the vectors (M), since for high-dimensional vectors ($M \gg 1$) the Gaussian limit for the distribution of c_{ij} is recovered, and the percolation becomes indistinguishable from the classical one. We fixed $M = 5$ so to appreciate the differences

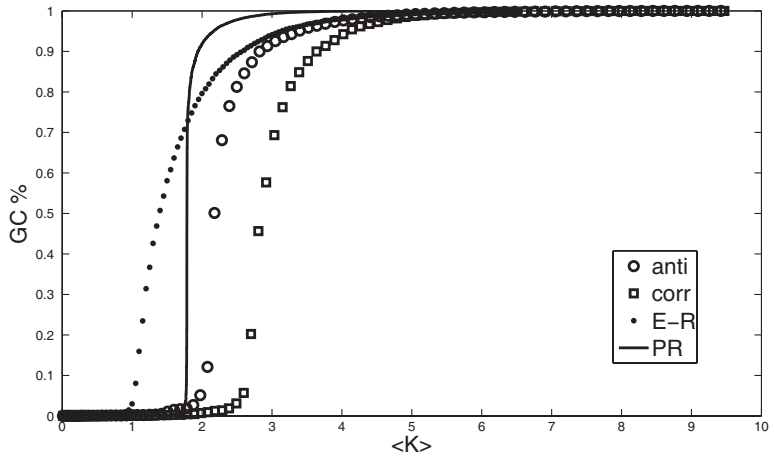

Fig. 1. – Percolation for different models with $N = 20000$ and $M = 5$: Giant Component size for Erdös-Rényi, Correlation, Anticorrelation, Product Rule models.

between the correlation coefficients distribution and a normal Gaussian distribution, and also in accordance with the size of the real-data set vectors as shown below.

As shown if fig. 1, E-R percolation is the less steep, and PR is the steepest. Giant cluster percolations for the correlation and the anticorrelation cases are less steep than PR, but start later as compared to E-R, thus they have a shorter "latent phase" between giant cluster onset and its occupation of the whole network.

3. – Example with real data

In previous papers we studied the dynamics of a gene expression time series network [8]; we compared two data sets of gene expression obtained from a set of microarray experiments using genetically engineered rat fibroblast cell lines, in which a master gene *c-myc* was, respectively, silenced (producing the N data set) and overexpressed (the T data set). We had 5 time points for both data sets (thus $M = 5$ according to the notation introduced before) and 8799 probes for each array.

In the correlation-based model, the similarity measure for the expression dynamics of two genes is given by the correlation between the two expression-level time series. For a given data set, if we define x_{lj} as the expression level of a gene with label l at time j, then the similarity between two genes with labels l and r, respectively, is given by

$$c_{lr} = \frac{\sum_j (x_{lj} - \mu_l)(x_{rj} - \mu_r)}{\sigma_l \sigma_r}, \tag{3}$$

where μ_l and μ_r are the averages in time of the expression levels for the two genes, and σ_l and are σ_r their standard deviations.

The correlation approach can be motivated by the hypothesis that genes belonging to the same activation (or inhibition) pathway should present similar (or opposite) expression profiles in time.

Hence, once we have the correlation matrix the adjacency matrix is obtained by considering thresholding on c_{ij}^2, as we explained previously.

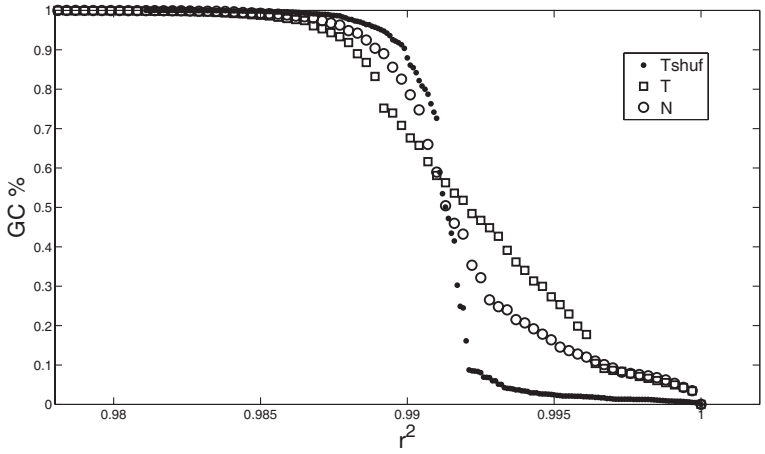

Fig. 2. – Percolation for different data sets: N, T and T reshuffled, whole array with 8799 probes. Plot of Giant Cluster (GC) relative size (with respect to the whole network) as a function of r^2 correlation value.

We checked both the whole 8799 probe set, and a subset of 1191 probes obtained by statistical analysis over N and T cases (2-way ANOVA, as explained in [8]). As a comparison, we randomly reshuffled data in both data sets, and performed the same analysis as for the unshuffled data (only the T reshuffled data set is shown, but both N and T produce similar results, thus demonstrating that reshuffling destroys any information contained in the time series data).

As shown in figs. 2 and 3, we observe very different percolation dynamics between the three data sets: high values of correlation last longer for the real data, while reshuffled data resemble percolation with randomly generated data (not shown). Moreover, whereas

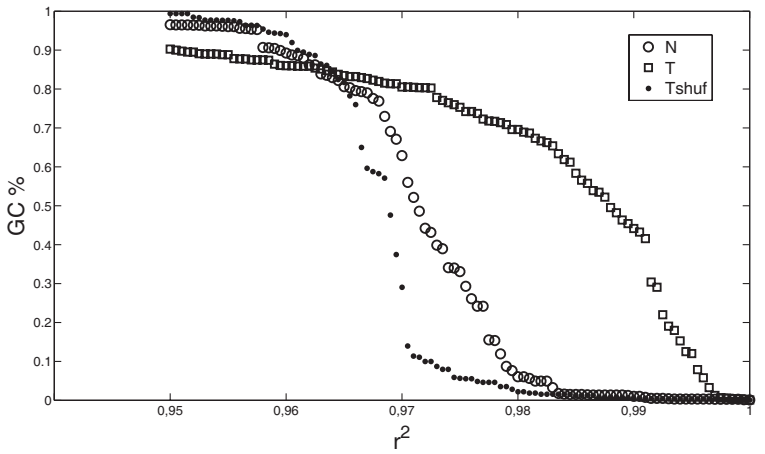

Fig. 3. – Percolation for different data sets: N, T and T reshuffled, selection of 1191 significant probes. Plot of Giant Cluster (GC) relative size (with respect to the whole network) as a function of r^2 correlation value.

the shuffled data giant component grows more continuously, in real data we observe some discontinuous "jumps" in its size (see, in particular, fig. 3, T data set for r^2 values close to 0.99). This discontinuities may resemble the formation of superparamagnetic domains during cooling [10], thus it should be interesting to further investigate the cluster size and distribution closer to these values in order to search for modules inside the whole network.

4. – Conclusions

We have observed percolation dynamics in the growth of a network based on the correlation of randomly generated data, that seems to differ from classical percolation, and results to be more similar to the so-called explosive (or better anomalous) percolation models. We investigated the behaviour of real biological data (high-throughput gene expression time series), obtained in an experiment in which a very important biological mechanism is switched on and off. We observe striking differences in percolation dynamics also in this case, both at the level of global gene expression level, and in a subset of genes selected by their significant response to the switching. Some hints point out at a possible role of percolation dynamics in finding modules and structures inside such networks, that deserves further investigation.

* * *

DR acknowledges Progetto Strategico d'Ateneo 2006 "p53 e patologie non neoplastiche nell'anziano: uno studio multidisciplinare sul ruolo del polimorfismo al codone 72 del gene TP53", Bologna University. DR also would like to thank G. Menichetti for her precious contributions to the development of this research project.

REFERENCES

[1] Erdös P. and Rényi A., *Publ. Math. Inst. Hung. Acad. Sci.*, **5** (1960) 17.
[2] Achlioptas D. *et al.*, *Science*, **323** (2009) 1453.
[3] da Costa R. A. *et al.*, *Phys. Rev. Lett.*, **105** (2010) 255701.
[4] Tumminello M. *et al.*, *J. Econ. Behav. Organ.*, **75** (2010) 40; doi:10.1016/j.jebo.2010.01.004.
[5] Wuchty S. *et al.*, *The Architecture of Complex Networks*, in *Complex Systems Science in Biomedicine*, edited by Deisboeck T. S. and Yasha Kresh J. (Springer) 2006, pp. 165-181.
[6] Milo R. *et al.*, *Science*, **298** (2002) 824.
[7] Tieri P. *et al.*, *Bioinformatics*, **21** (2005) 1639.
[8] Remondini D. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 6902.
[9] Iyer V. R. *et al.*, *Science*, **283** (1999) 83.
[10] Domany E., *Physica A*, **263** (1999) 158.