Information cascades in complex networks

Author: Garcia Seuma, Ramon Marc
Facultat de física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Advisor: Diaz-Guilera, Albert

Abstract: In this paper we model a dynamic process to analyze the spreading of information in social networks. This model encompass the spontaneous tendency of any individual to take a certain action, in this case to emit information, with the influence that this action exert in other individuals. We apply this model to a scale-free network, which has the same general structure than most of real networks, including social ones, then analyzing the emergence of synchronization, a collective pulse of information, so-called cascade of information. We examine the size distributions of this phenomena in and out of the phase transition as well as the dependence of the size depending on the trigger element. Once classified all elements in a space parameter, which takes account of the specific role of each element in the community structure of the network, we find that there is an exponential dependence on the average cascade size with the parameter related with the connectivity of the elements among distinct communities.

I. INTRODUCTION

The improvement of technology devices has brought a new and broad virtual world to the society. This, togheter with the increasing capacity of storage information and the development of big data analisis, has opened to the scientific community a new framework for develope and testing models and gain insights in a broad range of different fields related to networks.

These networks represent an idealization of systems, and allows to create plausible models to understand and predict certain processes that take part in real world, like the spreading of diseases, information, fashions, ideas and computer virus. In particular, avalanches in microblogging data logs, like twitter, have been documented and offer a good chance to develope and improve dynamics related to the spread of information. The models are based on concepts like neighbours influence, i.e. social reinforcement, and propensity to transmit information.

II. CHARACTERIZING A NETWORK

Here there are presented the basic concepts about networks necessary for understanding the results presented in this paper.

A. Degree distribution

In a given network, we have N nodes and L links, which represents the interactions between nodes. An important quantity for each node is its degree, denoted by k_i, i.e. the number of neighbours of the node i. Thus, in undirected networks

$$N\cdot <k>=\sum_{i=1}^{N} k_i = 2\cdot L \tag{1}$$

It is of great importance on the study of networks to determine the specific degree distribution p_k, that in a given network provides the probability that a randomly selected node has degree k. This allows us, among other things, to analytically calculate the network's moments.

One of the simplest and most useful models is the random network model, called also Erdös-Rényi network. It consists of nodes where each node pair is connected with probability p. Thus, its degree distribution follows a binomial distribution, centered in $<k>$, or a Poisson distribution in sparse networks.

Real networks like biological, social and technological ones, display a very different kind of distribution, since most nodes have few neighbours and few others, called hubs, lay up lots of them. In random networks hubs are missed, as all degrees are near the average one. Nevertheless, they are very useful in the study of networks, due to its intuitive and simple form, in order to compare properties of networks, processes and dynamics taking part in real ones.

Real networks are better represented by scale-free networks, which means that its degree distribution follows a power law distribution. Despite the differences in its origins, networks ranging from protein interaction networks, created since about billion of years, as well as recent social networks, share common features, and we can approximate its degree distribution as

$$p_k \sim k^{-\lambda} \tag{2}$$

where λ is called the degree exponent. For a network with $\lambda < 3$, although its first moment is finite, the second one diverges in the limit of $N \to \infty$. That is the reason why we call them scale-free networks and it appears the hubs. The specific form of the distribution depends on the model used to create the network.

In this work we tested the dynamics in both Erdös-Rényi network (ER) and in a particular scale-free network called Barábasi-Albert (BA), that has $\lambda = 3$.

B. Communities

One fundamental property of networks found in real world is their tendency to show an underlying modular structure. The modules, commonly called communities, are groups of nodes that connect more densely among them. Detect and characterize those inner structure allows to reveal internal organization, thus to infer subtle relationships of great importance for the develope of the network and the processes taking part on it.

There has been proposed many algorithms to detect communities, based for example on the network clustering and the cliques. We use in this work an implemented algorithm called Louvain method, which maximizes the quantity called modularity, thus finding an optimal partition of the network.

C. Roles

In a given network, it is tipically difficult to extract specific information related to the individual function of the components of these, given the dimensions and intricacy of such networks. We use here the method presented in [3], based in the connectivity of the nodes, that yields to a cartographic representation of complex networks and classifies nodes in function of universal roles, depending on their proportion of within and between-module connections. This characterization will help us to gain insight about how these affect to the dynamics studied in this paper.

Once classified the modules present in our network, we have N_M modules. Then we define two quantities for each node, in order to classify all them in the so-called zP parameter space.

$$z_i = \frac{\kappa_{is_i} - \bar{\kappa}_{s_i}}{\sigma_{s_i}} \tag{3}$$

Here, κ_{is_i} is the number of neighbours of the node i inside its own module s_i, $\bar{\kappa}_{s_i}$ is the average of κ_{s_i} of all nodes in s_i. The z-score takes account of the level of connection a node has inside a module. In [3] hubs are defined to be those which have $z \geq 2.5$, and non-hubs $z < 2.5$.

Then we define the participation coefficient as

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{\kappa_i} \right)^2 \tag{4}$$

where κ_{is} is the number of links of node i to nodes in module s, and κ_i the total degree of i.

There are classified seven regions (corresponding to seven roles), four for non-hubs and tree for hubs.

III. CHARACTERIZING THE DYNAMICS

Once we have established the connectivity among nodes, thus created our network, we introduce a dynamic running within it, in order to characterize a particular process taking part on it.

A. Threshold model

This is a simple but widely studied model, and captures the essence of the dynamic taking part in many spreading phenomena. Simplifying, it assigns a fixed threshold τ_i to each node i, took from a distribution between the values of zero and one. The nodes can be either in two states, active or inactive. Initially we have a selected fraction of them active. When a node i with degree k_i has a fraction of active neighbors denoted by a_i, exceeding its threshold, i.e. $a_i/k_i > \tau_i$, then it becomes also active. In each step of the process we update the state of all nodes until they reach an equilibrium.

B. Integrate-and-fire model

Despite the success of the threshold model capturing the basics of the dynamics taking part in some complex contagious phenomena, it has some limitations. For example, once we have equilibrium in the system, it remains quiet, so it is impossible to have a cascade of information were states of activeness and disactiveness are constantly happening. It also do not allow nodes to spontaneous become active, and many external stimulous affect nodes in real networks so they are not only controlled by their neighbours state. We use here a modification, introduced in [4], of the integrate-and-fire oscillator model (IFO) by Mirollo and Strogatz [5]. It is curious to remark that this model has been proven to be useful in modelling burst in neuronal systems.

The dynamics is governed by N equations, which have the form:

$$m_i = \frac{1}{w} \cdot \ln(1 + (e^w - 1) \cdot \theta_i) \tag{5}$$

where m_i is the motivation, a voltage-like state, which increases monotonically from 0 to 1 in a cycle governed by the inner time θ_i. They are coupled, hence if a particular node reaches the threshold it emits a pulse of information, and all their neighbours, which are exposed to it, increase their own motivation a certain quantity ϵ.

Thus, we see that spontaneous propensity is purely periodical if all nodes are disconnected, i.e. $\epsilon = 0$, either if the system reaches total synchronization.

IV. SIMULATION AND RESULTS

In order to define a global time in the system, cycles are introduced. One cycle is completed whether all nodes has almost activated once. Thus, we first analize how many cycles the system needs to globally synchronize depending on the parameters of the model, and then we analize the shape of the distributions related with the sizes of the cascades.

A. Phase transition

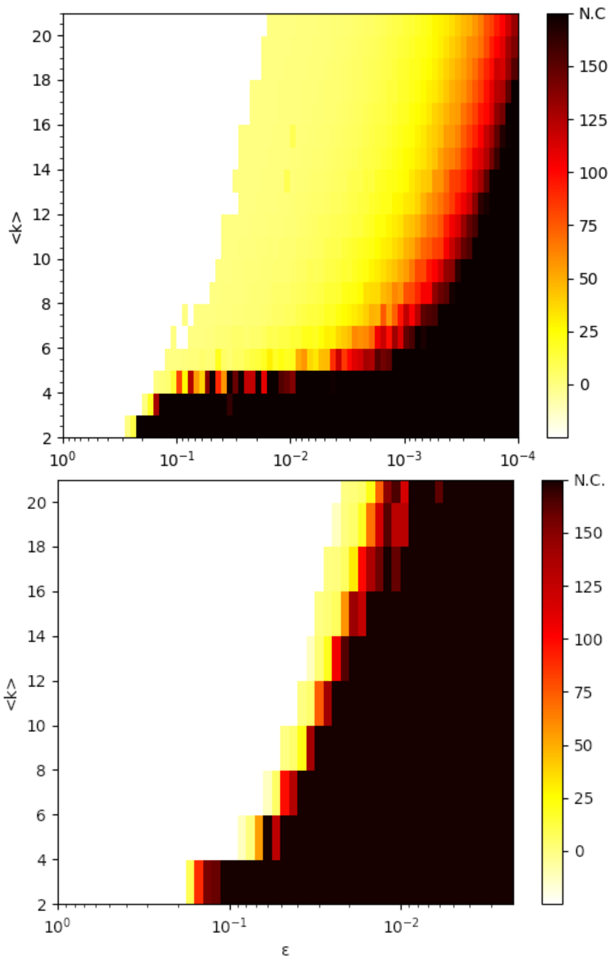

Figure 1: Phase space of the occurrence of cascades, results averaged over 25 networks of 1250 nodes with random initial energies. Up: ER network. Bottom: BA network. N.C. means that no macroscopic cascade is observed.

In the previous figure we have coded with colours the average cycles needed to observe one macroscopic cascade, i.e. of size $S \geq 0.25N$.

B. Size distributions

We show now the behaviour of the cascade size distribution in the BA networks, varying the parameter ϵ so that we cross the phase transition zone.

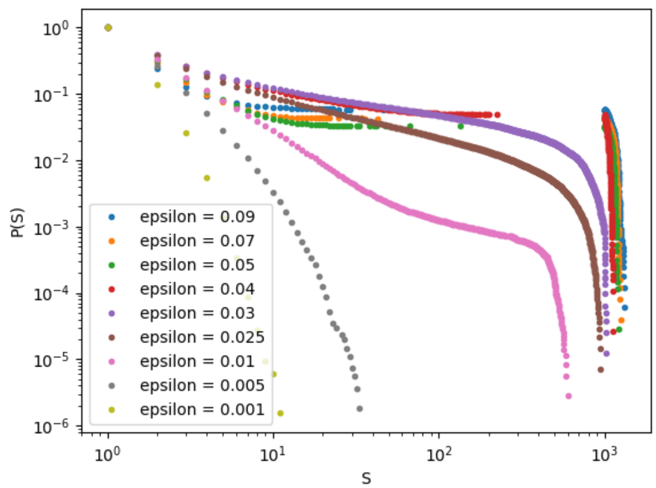

Figure 2: Cascade size distribution in a BA network with $N = 1000$ nodes and $< k >= 18$, all set to $w = 3$ but with different ϵ. The phase transition leave its trace as it appears a discontinuity. Realizations over 750 cycles.

The most remarkable fact is that this phase transition brings a discontinuity to the distribution, so above it the probability of obtaining cascades spreading among all the nodes of the system is pretty high. In weakly-coupled systems we observe a logarithmic dependence. This has been documented in real social networks [6]. So in high-activity periods in real networks as twitter [2][4], we observe an effect similar to an increase in the coupling of the nodes, as macroscopic cascades are much more probable than in stable periods. Specifically, this is captured in the model by the increase of w, which in turn is predicted in [4] to have an inverse dependence to the average inter-event times of the macroscopic cascades.

Obviously, a size of a cascade will depend on who is the trigger element of this one. First, we explore these sizes depending on the degree of the node who begins the cascade. The results show that this is not as important as it may seem, as in scale-free networks, the probability that a neighbour of any node is a hub is high, and it is more decisive for the spreading of a particular piece of information if the fire passes through a hub.

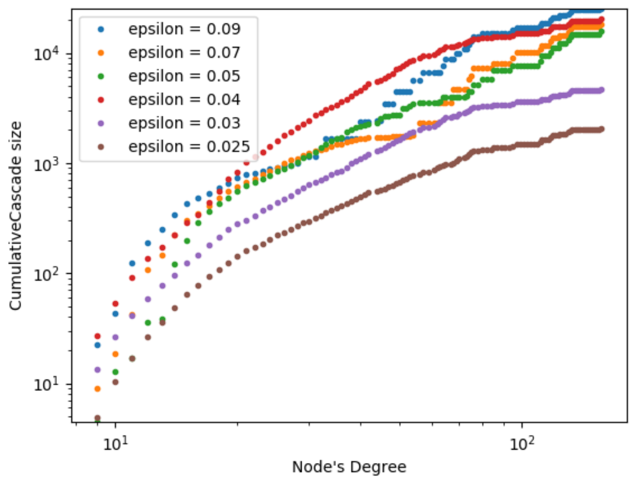

Figure 3: Average cascade size in function of the degree of the node that initiate the cascade registered, in networks with the same parameters as those described in Figure 2.

C. Roles

Rather than with the degree of the trigger node, we observe a strong influence on the role of the nodes. In order to explore the importance of the role in the spreading of information, we first compute the zP point related to each node in the given network, with equations (4) and (5).

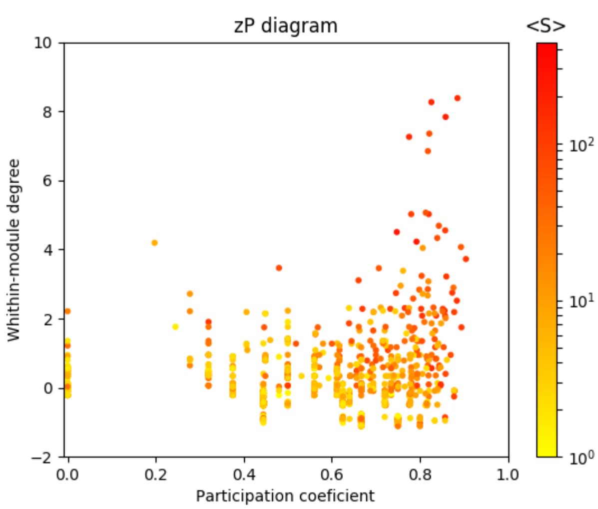

Figure 4: zP parameter space for a network of 1500 nodes coded with colours depending on the average cascade size documented in 750 cycles.

The most immediat information we obtain is the zP parameter space, in which there is located the values of all nodes, with a colour variable that depends on the average cascade size of the node in question. This is representative for what we are looking for, but there is no clue information revealed beyond an obvious tendency to document greater cascades in the nodes situated at the top right and up of the parameter space.

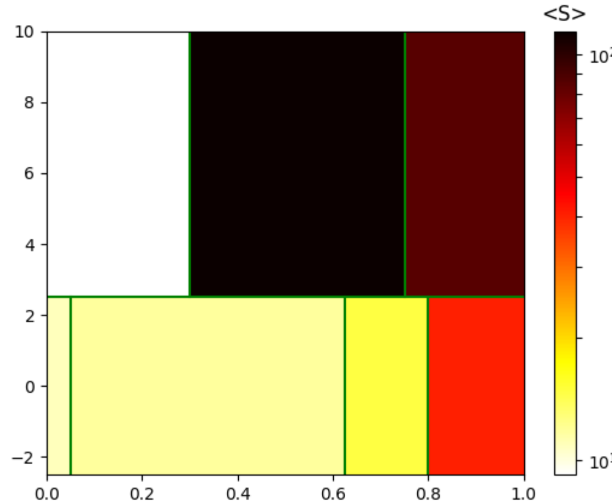

Figure 5: Averaged cascades in surfaces of the different classified roles for the networks of figures 2 and 3.

We can also differentiate all zones in which there are classified different roles in order to see whether we have a general behaviour of the cascade size averaged over all nodes belonging to the same role, as well as the deviation. We observe a clear tendency to have greater cascades with the increasing value of P, rather than with z, highlighting the importance of the connectivity of the nodes in the dynamic.

We must remark that here we have a lack of the so-called provincial hubs, the ones located at the top-left region, because the structure of the BA network has no real modular structure, although it appears a similar one due to its small size.

So far, we have seen a qualitative behaviour. In order to extract some useful insight, we found that the averaged sizes of the cascades follow an exponential growth with the P-score. Taking the average over all z-score of the hub and non-hub region and then computing its average cascade size, we observe that both regions follow the same pattern of growth.

This is interesting, in the sense that in a modular-idealized network, i.e. one with all-neighbours belonging to the same community, if we add progressively one node to each community, the probability of a certain fixed cascade size increments linearly with the number of communities that connect this node. Given the construction of the P function, in this idealized network the averaged cascade size would grow faster than exponential, thus BA networks relax this growth.

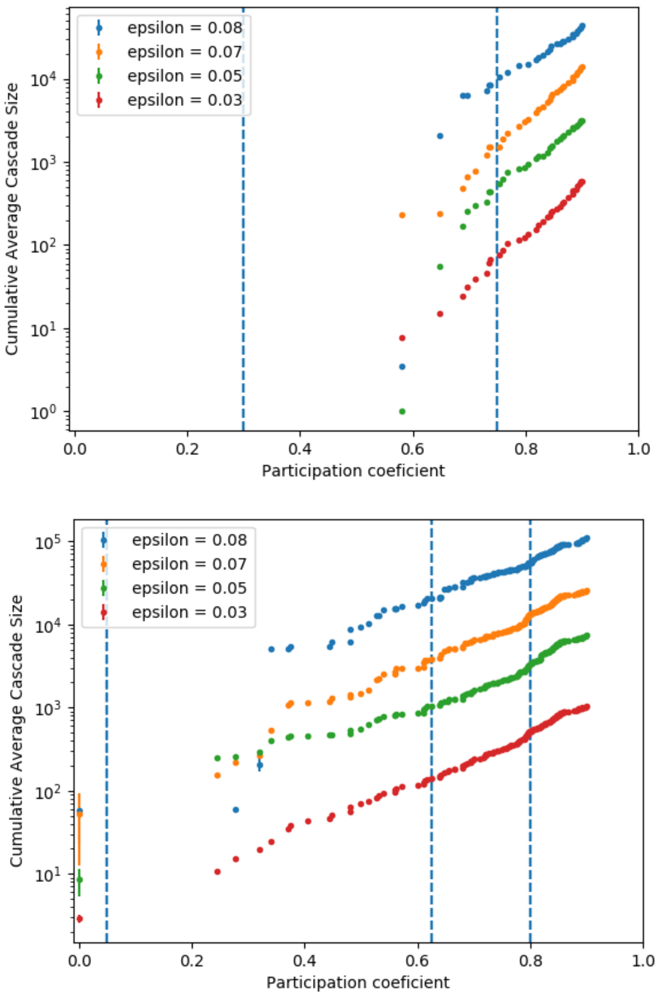

Figure 6: Cumulative average cascade size in function of P for the region of hubs (up), and non-hubs (bottom).

If we bring *epsilon* near the phase transition, the fluctuation of the exponent coefficient grows rapidly, indicating a possible discontinuity of this parameter in the phase transition.

V. CONCLUSIONS

So far we have seen some important traits of a spreading phenomenon, we had been able to model it with a dynamic based on the pulsing of information as well as a coupling which is entirely determined by the connectivity of the network to study. Gathering all the results, we can state that either in random or scale-free networks the synchronization of the nodes is an emergent phenomenon that passes through a phase transition, well established either in the *epsilon*-$< k >$ diagram as well as in the cascade size distribution. In BA networks, nevertheless, the phase transition appears more abruptly, and the variation with the size of the network is much more important, because the network structure is depending on it. We have also found an interesting result related to the importance of the properties of the node who seeds the cascade. The role of the nodes it is much more important than the degree itself, thus nodes in contact with more communities tend to achieve larger cascades, which is captured in the exponential dependence of the sizes with the P-score. Finally, the aim of this model is to apply it to real data, in real networks with true community structure, to test and ajust the parameters and dependences of the model to real ones. Therefore, the extension of the work would be to test it on real networks, especially to the ones involved in the last important burst of activity in Twitter, related to the social movement in Catalonia, which represents a recent and near high-activity social period.

Acknowledgments

I would like to thank my advisor Albert Diaz-Guilera as well as Emanuele Cozzo for their guidance and support.

[1] Barabási, A.-L., Pósfai, M. (2016). *Network science*. Cambridge: Cambridge University Press. ISBN: 9781107076266 1107076269

[2] Borge-Holthoefer J, Rivero A, García I, Cauhé E, Ferrer A, et al. (2011) *Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study*. PLOS ONE 6(8): e23883. https://doi.org/10.1371/journal.pone.0023883

[3] Roger Guimerà and Luís A. Nunes Amaral, *Cartography of complex networks: modules and universal roles*, J Stat Mech. 2005 February 1

[4] P. Piedrahita, J. Borge-Holthoefer, Y. Moreno and A. Arenas, *Modeling self-sustained activity cascades in socio-technical networks*, P. Piedrahita et al 2013 EPL 104 48004

[5] Mirollo R. and Strogatz S., SIAM J. Appl. Math., 50 (1990) 1645.

[6] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. *Everyone's an influencer: quantifying influence on twitter*. Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 65-74, 2011.