Activity Spreading in Modular Networks (Extended Abstract)

Aram Galstyan and **Paul R. Cohen**
USC Information Sciences Institute
Center for Research on Unexpected Events (CRUE)
4676 Admiralty Way, Marina del Rey, CA 90292
{galstyan,cohen}@isi.edu.

Many relational classification problems involve estimating joint probability distributions over sets of nodes (and possibly edges) in relational graphs. Since exact estimation is infeasible even for moderately sized graphs, a variety of approximate methods have been proposed. One such approach is relaxation labeling, or label propagation, which works iteratively by propagating the labels (or associated probabilities) of initially labeled nodes throughout the networked data. Different label propagation techniques have been used for approximate inference making in graph–based semi–supervised learning problems.

From the perspective of network dynamics, label propagation corresponds to some kind of an activation spreading in the network. Clearly, such an activation process is affected by the structural and statistical properties of the network. There has been an extensive amount of work on modeling various dynamical processes on complex networks. For instance, recent research has examined the role of scale–free degree distribution on the critical behavior of various epidemic models. Here we focus on another important property of networks, *modularity*, which is the tendency of nodes to partition themselves into clusters. Specifically, we are interested in how the activation process is affected by the network modularity.

We consider a random graph consisting of $N = N_a + N_b$ nodes of two different type, a and b. The probabilities of edges between nodes of different types are γ_{aa}, γ_{bb} and $\gamma_{ab} = \gamma_{ba}$, and the average connectivity between nodes of the respective types are then $z_{aa} = \gamma_{aa}N_a$, $z_{bb} = \gamma_{bb}N_b$, $z_{ab} = \gamma_{ab}N_b$ and $z_{ba} = \gamma_{ab}N_a$. We want to find out how the modularity of the network, as described by the coupling between the groups, affects the cascading process.

Each node is either active or passive (e.g., *labeled* and *unlabeled*). Initially, only a small fraction of *seed* nodes are active. During the activation process, a passive node will be activated with probability that depends on the state of its neighborhs. In Watt's original model (Watts 2004) this probability is $p = \Theta(h_i/k_i - \phi)$, where Θ is the step function, h_i and k_i are the number of active neighbors and the total number of the neighboring nodes, respectively, and ϕ_i is the activation threshold for the i–th node. Here we use a threshold condition on the $number$ of active neighbors rather than their fraction: $p = \tau^{-1}\Theta(h_i - H_i)$, where τ determines the time–scale of the activation process. We will assume that all nodes have the same activation threshold, $H_i = H$ for all i. Note that this activation mechanism is reminiscent of simple relational neighbor classifier(Macskassy & Provost 2007).

Let ρ_a^0 and ρ_b^0 be the fraction of seed nodes in each population. Further, let $P_a(h;t)$ and $P_b(h;t)$ be the probability distribution that a randomly chosen node of corresponding type is connected with exactly h active nodes at time t. Clearly, $P_a(h;t=0)$ and $P_b(h;t=0)$ are Poisson distributions with means $z_{aa}\rho_a^0 + z_{ab}\rho_b^0$ and $z_{bb}\rho_b^0 + z_{ba}\rho_a^0$, respectively. To study the activation dynamics, we need to estimate these distributions for later times. This is particularly straightforward to do within the *annealed approximation*, e.g., by "rewiring" the network after each iteration (Derrida & Pomeau 1986; Derrida & Stauffer 1986). Indeed, since all edges of corresponding type are equally likely, it is easy to see that $P_a(h;t)$ and $P_b(h;t)$ are still given by Poisson distribution, with the means that now depend on the fraction of active nodes $\rho_a(t)$ and $\rho_b(t)$: $P_{a,b}(h;t) = Poisson(\lambda_{a,b}(t))$, where $\lambda_a = z_{aa}\rho_a(t) + z_{ab}\rho_b(t)$ and $\lambda_b = z_{bb}\rho_b(t) + z_{ba}\rho_a(t)$. One can then show that in the continuos time limit the activation dynamics is governed by the following equation (Galstyan & Cohen 2007):

$$\tau\frac{d\rho_{a,b}}{dt} = 1 - \rho_{a,b} - (1-\rho_{a,b}^0)Q(H;\lambda_{a,b}) \qquad (1)$$

where $Q(n,x) = \sum_{k<n} e^{-x}x^k/k!$ is the regularized gamma function.

Let $\rho(t) = \alpha\rho_a(t) + (1-\alpha)\rho_b(t)$, $\alpha = N_a/(N_a+N_b)$, be the fraction of active nodes in the whole network. In Figure 1 we compare the solutions obtained from Equations 1 with the results of simulations on randomly generated graphs for the same network parameters but two different values of the threshold parameter. The parameters of the network are $N_a = 5000$, $N_b = 15000$, $z_{aa} = z_{bb} = 15$, $z_{ab} = 4$. The fraction of seed nodes is $\rho_a^0 = 0.1$, and $\tau^{-1} = 0.1$. The simulations are averaged over 100 random realizations.

The agreement between the analytical prediction and results of the simulations is quite good. The network settles to the same steady state for both values of the threshold parameter H: that is, all of the nodes are activated at the end of the cascading process. However, the transient dynamics depend on the threshold parameter H. For $H = 2$, activation

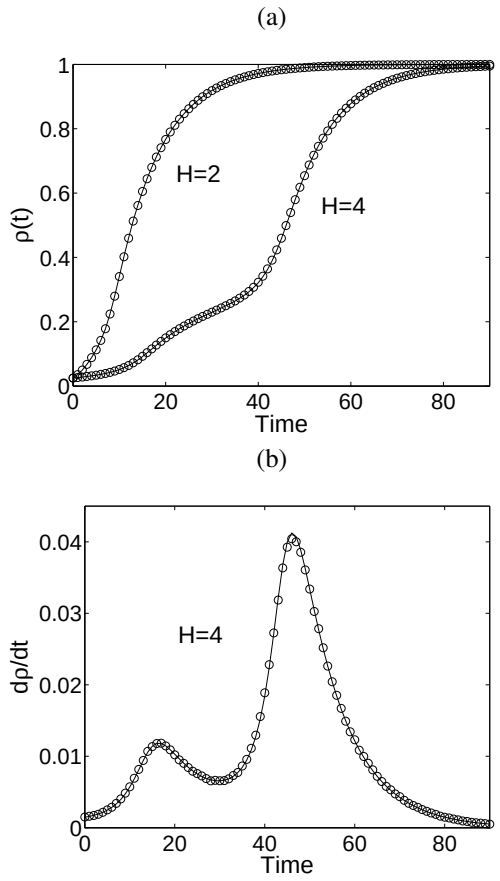

Figure 1: Analytical (solid lines) and simulation (circles) results for the activation dynamics. The upper panel shows the fraction of active nodes vs time for threshold parameter $H = 2$ and $H = 4$. The lower panel shows the activation rate $d\rho/dt$ vs time for $H = 4$.

spreads very quickly through both communities and after a short interval all of the nodes are activate. For $H = 4$, on the other hand, the fraction of active nodes seems to saturate, then, in later iterations, $\rho(t)$ increases rapidly and eventually all the nodes become active. In Figure 1(b) we plot the rate of activation process $d\rho/dt$ vs time for $H = 4$. Apparently, the peak rates of activation in the two communities are separated in time. We call this phenomenon *two-tiered dynamics*.

Our analysis of the activation dynamics presented in (Galstyan & Cohen 2007) revealed that the activation spreading in the two–cluster Erdos–Renyi graph defined above is characterized by a doubly–critical behavior. Consider, for instance, activity spreading in a single population: Our calculations show that for a fixed fraction of the seed nodes there is a critical connectivity z^c_{aa} so that for $z_{aa} < z^c_{aa}$ the activation dynamics dies out, while for $z_{aa} > z^c_{aa}$ it spreads throughout the network. In the limit of small ρ^0_a the critical connectivity scales as $z^c_{aa} \propto (\rho^0_a)^{-\frac{H-1}{H}}$. Also, at the critical point the convergence time of the activation process diverges as $T_{conv} \propto (z - z^c_{aa})^{-1/2}$. Similar results hold for the second population: Namely, for a fixed within–group connectivity z_{bb} there is a critical across–group connectivity $z^c_{ab}(\rho_a)$ so that for $z_{ab} > z^c_{ab}(\rho_a)$ the activity will surely propagate through class b provided that at least fraction ρ_a of a nodes have already been activated. Furthermore, it is easy to see that the critical value $z^c_{ab}(\rho_a = 1)$ corresponds to the marginal case where the activation will not spread to b nodes at all. At this marginal point, the two–tiered dynamics is most pronounced: This is because the convergence time for the b nodes, and thus, the separation between two peaks, is infinite (in other words, the second peak never develops).

We have shown that the two–tiered activation dynamics can be used for building an efficiently parameter–free classifier in semi–supervised settings (Galstyan & Cohen 2005). Our results show that such an algorithm achieves a good classification accuracy provided that the overlap between two classes is not very strong. Furthermore, an important advantages of our approach is that it requires initial knowledge only about the class of interest, while most of the other homophily based classification algorithms require labeled instances from both classes. This might be very important when the class of interest is just a tiny fraction of a much larger number of benign entities, so that providing an adequate number of negative examples is very costly. We also note that the algorithm can be used both for explicit classification and for ranking entities according to their similarity to the class of the interest. A good criterion for ranking entities is the activation time (i.e., nodes that are more similar to the class of the interest are activated earlier).

References

Derrida, B., and Pomeau, Y. 1986. Random networks of automata: A simple annealed approximation. *Europhysics Letters (EPL)* 1(2):45–49.

Derrida, B., and Stauffer, D. 1986. Phase transitions in two-dimensional kauffman cellular automata. *Europhysics Letters (EPL)* 2(10):739–745.

Galstyan, A., and Cohen, P. R. 2005. Inferring useful heuristics from the dynamics of iterative relational classifiers. In *Proceedings of IJCAI-05, 19th International Joint Conference on Artificial Intelligence*.

Galstyan, A., and Cohen, P. 2007. Cascading dynamics in modular networks. *Phys. Rev. E* 75:036109.

Macskassy, S., and Provost, F. 2007. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research (JMLR)* 8:935–983.

Watts, D. 2004. A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci. U.S.A.* 99:5766.