ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1115

Community Detection based on a Naming Game

Thaís Gobet Uzun and **Carlos Henrique Costa Ribeiro** [1]

Abstract. Opinion exchange in a social network can profoundly affect its structure, given that agreeing with another person bounds he/she more closely and vice versa. In this work, we show that communication interactions can form and reveal the community structure of a network, as we present a new model where agents change their links and behaviors according to the local history of communication successes. Our simulations show that a local node parameter based on such history changes relatively to the node proportion of extra-community connections, and that (adaptive) edge weights tend to get high for intra-community and low for extra-community connections, also as a consequence of the history of communication successes. In non-convergent executions, the model gets trapped on a regime where clusters of agents agreeing with different words emerge, corresponding to the existing communities in the network.

1 INTRODUCTION

The organization of nodes in communities is a relevant feature in real networks. These communities, depending on the nature of the problem, can be considered as almost independent partitions of a graph, playing different roles and influencing the functionality of the system [4][7]. The definition of community, however, has been extensively discussed and no definition is universally accepted, as it often depends on the application[4]. In this work, the intuitive definition of community - group of nodes that have more edges between them than edges linking those nodes to the rest of the graph - will be considered. In fact, this is the reference for most existing community definitions, and in most cases, communities are algorithmically defined without a precise *a priori* concept [4].

One of the developed community detection algorithms is the *Speaker-listener Label Propagation Algorithm* - SLPA [7], recently introduced in the literature. In SLPA, each agent maintains a memory with a counter of the occurrences of labels. At each time step, one node is chosen as a listener and every neighbor chooses a label from its memory with probability proportional to the label's occurrence, and sends it to the listener. The listener then adds the most popular received label to its memory. The resulting list of received labels and occurrences is interpreted as pertinence levels of the node to the community associated with each label. This algorithm can be used also to detect overlapping communities, and it has been shown to produce one of the best performances reported in the literature [7].

The introduction of a memory and the presence of speakers and listeners resembles the *Naming Game* (NG), a well-known opinion dynamics model . Such resemblance raises the question if the linguistic dynamics of a population can form or reinforce an existing community structure. In this paper we tackle this question.

2 THE NAMING GAME

The minimal Naming Game [1] (NG) was developed to investigate how a convention on the use of a vocabulary emerges through local interactions under relatively simple rules. The game is played by N agents that have a local memory initially empty. At each time step, an agent is chosen as speaker and, among its neighbors, an agent is chosen as listener. The speaker randomly chooses a word from its memory, or — if it is empty — it invents a word and transmits it to the hearer. If the hearer has the word in its memory the communication is a success and both agents delete all but the transmitted word from their memories. If the hearer does not have the word in its memory, the communication is a failure and the hearer adds the transmitted word to its memory. The game ends when a convergence state, where all agents have only one and the same word in their memories, is reached. This minimal Naming Game can be shown to converge for various network topologies, with the exception of networks with high community structure [6] when the game gets trapped in a state where several clusters of nodes agreeing with different words coexist.

One NG-based model is the NG with Adaptive Weights [5] (NG-AW), that offers the possibility of non convergence even in fully connected networks. In NG-AW, the choice of a hearer is proportional to the edge's success rate plus a parameter ϵ, that can be set to result in either convergence or non convergence of the game. Crucially, the weight of the edge, *i.e.* its success-attempts ratio, is responsible for the history-based determinism in the communication, whereas ϵ is responsible for the randomness in choosing the hearer. When ϵ is large, the choice of the hearer is less biased by a history of past successes, and the game resembles NG in behavior. When ϵ is small, the game forces the agent to communicate preferably through edges that had many successes in the past, thus entering a multilanguage regime.

In the community detection scope, as communication flows through large weight edges, interactions will take place mostly inside communities, resulting in a regime where agents belonging to the same community share the same word. However, simulations show that the percentage of explored edges (*i.e.*, edges that have any communication attempts), is very small under small values of ϵ. This means that the algorithm constrains the communication to a few links randomly chosen in the first steps of the game. As ϵ increases, more edges are explored in the game but the convergence probability also increases, in a rate that depends on the network topology. For networks with high community structure, it is possible to find a small interval of epsilon values that results in high edge exploration and relatively low convergence probability.

NG-AW is based on the assumption that communicating agents interact preferably with agents that share the same "opinions". However, we can speculate that communicating agents behave differently in this aspect. For instance, in human societies some people have more patience than others to discuss opposite opinions, and a per-

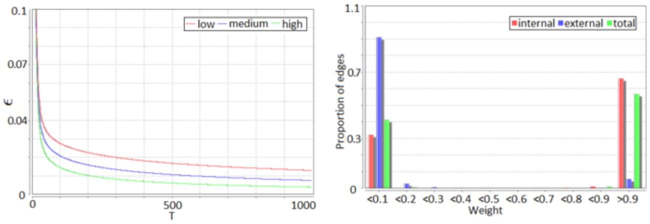
[1] Instituto Tecnológico de Aeronáutica - São José dos Campos - Brazil, thaisgu@ita.br, carlos@ita.br. Thais G. Uzun thanks CNPq (proc. no. 143356/2011-9), and Carlos H. C. Ribeiro thanks CNPq (proc. no. 303738/2013-8).

son might tend to loose patience as fewer people agree with him/her through time. Based on this observation, we propose a new game model, the Naming Game with Local Exploration Factor (NG-LEF), where every node has its own ϵ, the parameter responsible for the degree of randomness in the choice of the hearer - here called exploration factor - that decays as the agent has failed communications in the game. In this game, node memories are initially empty, all agents hold an initial $\epsilon(0)$, and at each time step the speaker *i* is chosen randomly and chooses a hearer *j* with probability proportional to $p_{ij} = \frac{w_{ij}+\epsilon_i}{\sum_{k=neighbors}(w_{ik}+\epsilon_i)}$, where $w_{ij} = \frac{Successes_{ij}}{Attempts_{ij}}$. The speaker then randomly chooses a word from its memory to communicate, or — if the memory is empty — invents a word, and transmits it to the hearer. If the hearer has the word the communication is a success, and both agents erase all but the transmitted word from their memories. If the hearer does not have the word the communication is a failure, both agents decrease their ϵ in 10% and the hearer adds the word to its memory. A resulting hypothesis is that peripheral nodes — with many connections with nodes from other communities — must receive many different words and thus have a low success rate. As a node has more failures, its ϵ value decays faster than for nodes with more internal connections, decreasing the randomness of this node's hearer choice and interacting preferably with agents in its own community. More internal nodes, with less failures, will make choices more randomly, so the nodes will behave according to its relative position in the network.

3 RESULTS AND CONCLUSIONS

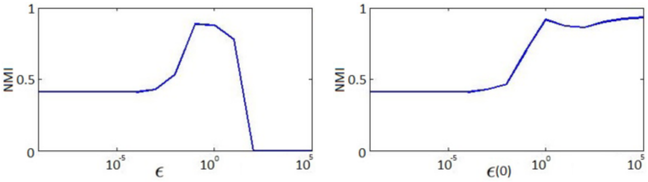
We consider the benchmark of Condon and Karp [2] with $N = 200$ agents, $n_c = 4$ communities, $p_{in} = 0.85$ and $p_{out} = 0.05$, a network with extremely high community structure for preliminary tests. In non convergent executions, as the total number of words existing in the network converges to N, the number of different words converges to n_c, indicating that it has reached a regime where each node has one word and there are n_c clusters coexisting (not shown) with $\epsilon(0) = 1$. To analyse the local behavior, we focus on two parameters: ϵ_i for each node i and the weight w_{ij} for each pair of nodes i, j. To verify the variation of ϵ_i relative to the node position in the network, we divided the nodes into 3 classes: *high* - more than 15 % of external edges; *medium* - 10 to 15 % external edges; and *low* - less than 10 % of external edges. Figure 1(a) shows the behavior of *high*, *medium* and *low* nodes regarding the variation of ϵ along time, considering $\epsilon(0) = 1$. The plot is zoomed at 10 % of $\epsilon(0)$, as it is typically a small value with fast decay. We note that different classes of nodes get different average values for ϵ_i along the game, as expected. Moreover, focusing on the weights of the resulting network, it is possible to observe in Figure 1(b) that these tend to be either close to 0, for the external weakened edges, or 1, for the internal reinforced edges, confirming the existing community structure. Also, the game gets trapped more frequently in a non-convergent state with full edge exploration than NG-AW (not shown). We adopted the unit $T = N$ communication attempts as defined in [5].

It is possible to interpret the words in an agent's memory as community labels propagating through the network. As on average each node holds a single word after the game, in a post-processing step, in case of non-convergence, each node is assigned to the word in its memory and if has more than one, the word is chosen randomly from the agent's memory. To verify the accuracy of this community identification method, we use the *Normalized Mutual Information* (NMI) [3], that takes its maximum value of 1 if the partitions are identical and takes its minimum 0 if the partitions are disjoint.



(a) Evolution of ϵ (b) Weights Distribution Histogram

Figures 1(c) and 1(d) show the NMI at the end of the games as a function of the parameters ϵ and $\epsilon(0)$ for NG-AW and for NG-LEF, respectively. For NG-AW, small values of ϵ force most communications to flow through the same paths, resulting in many small groups with the same word and a large number of assigned communities, producing small NMI. As ϵ increases, the choice of hearer gets less biased, producing a maximum NMI value when the network is sharing n_c words. This coincides with the interval with few convergence and high exploration rate. As ϵ becomes larger, the history of successes has lower influence on the choice of hearer, therefore the communication becomes more random and the game always converges, thus NMI=0.



(c) NMI as function of ϵ, NG-AW (d) NMI as function of $\epsilon(0)$, NG-LEF

For NG-LEF, lower values for NMI under low values of $\epsilon(0)$ are explained by the same reason: $\epsilon(0)$ is too small to explore the network. The increase of NMI happens, however, when $\epsilon(0)$ is large enough not to explore the network randomly, as in NG-AW, but to organize the nodes by position in the network. As every node i has its ϵ_i value, the individual decrease of ϵ_i according to the node failure rate makes different nodes have different behaviors, the more peripheral nodes behaving deterministically, and the more internal ones more randomly. This self-organization aspect is maintained for larger values of $\epsilon(0)$ as the decrease in ϵ_i for peripheral nodes tend to be faster than for internal ones. This behavior makes the model suitable for community-based applications, as it categorizes the nodes into communities under less restrictions on the input parameter.

As future work we will test other network topologies, including networks with overlapping communities, motivated by the existence of multiple words in memories during the game. It might also be interesting, due to the local nature of this algorithm, to apply it for community detection in mobile and/or evolving networks.

REFERENCES

[1] Baronchelli, A., Felici, M., Caglioti, E., Loreto, V., Steels, L. J. Stat. Mech., 2006.
[2] Condon, A., Karp, R. M., Random Struct. Algor. 18, 116, 2001
[3] Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A., J. Stat. Mech., 2005.
[4] Fortunato, S., Phys. Rep. 486, 75-174, 2010.
[5] Lipowska, D., Lipowski, A., MIT Press, Artificial Life, 2012.
[6] Lu, Q., Korniss, G., Szymanski, B., J. Econ. Int. Coord, v. 4, p. 221, 2009.
[7] Xie, J., Kelley, S., Szymanski, B., ACM Comp. Surveys 45(4), 43, 2013.