Spreader Selection by Community to Maximize Information Diffusion in Social Networks

Didier A. Vega-Oliveros and Lilian Berton
Department of Computer Science
ICMC, University of São Paulo
C.P. 668, CEP 13560-970, São Carlos, SP, Brazil
davo,lberton@icmc.usp.br

Abstract

Rumors or information can spread quickly and reach many users in social networks. Models for understanding, preventing or increasing the diffusion of information are of greatest interest to companies, governments, scientists, etc. In this paper, we propose an approach for maximizing the information diffusion by selecting the most important (central) users from communities. We also analyze the selection of the most central vertices of the network and considered artificial and real social networks, such as *email*, *hamsterster*, *advogato* and *astrophysics*. Experimental results confirmed the improvement of the final fraction of informed individuals by applying the proposed approach.

1 Introduction

The modeling of propagation or diffusion processes in social networks has recently received more attention, since it allows to understand how a disease can be controlled or how information spread among individuals. These diffusion process are generally analyzed employing complex network theory (Barrat et al., 2008; Castellano et al., 2009). The area of complex networks seeks to study and understand the dynamics and behavior of complex systems, from the structure of the network to the internal dynamics or interactions.

Models that describe the evolution of rumors can be adapted to analyze the spread of spam on the Internet, advertising and marketing, political ideologies or technological news between individuals (Castellano et al., 2009). In such cases, the representation in complex networks enables the analysis of traditional models and the heterogeneous structure, which has a strong influence on the information diffusion process (Moreno et al., 2004; Barrat et al., 2008; Castellano et al., 2009). Therefore, some individuals can have a higher influence than others according to the network structure. Researchers have focused on identifying the most influential vertices (Kempe et al., 2003; Kitsak et al., 2010; Lawyer, 2012; Pei and Makse, 2013; Hébert-Dufresne et al., 2013) according to topological properties. It is expected this influencers convince the largest number of individuals in the network. However, the selection of more than one of them not necessarily maximizes the expected fraction of informed individuals, compared to an uniformly random selection approach.

In this paper, we propose an approach to maximize the information diffusion considering the community structure of the network. The community symbolizes a group of individuals with a greater tendency to have more internal than external connections to other groups. The reason is that vertices belonging to the same community are likely to be more similar to each other and share similar properties and affinity. We confirmed that selecting the most influential individual from each community as initial spreaders increases more the information diffusion than selecting the most influential individuals from the whole network.

As a motivation example, let us consider a company that wants to market a new product in the blogosphere. The company could select three very influential individuals of this social network (bloggers with thousands of access) to advertise its product. However, these influencers may be popular in the same group of people. On the other hand, if the strategy is to identify the three main communities on the network and select the most influential individuals of each community, the company would achieve a variety group of users and maximize the marketing diffusion.

The main contribution of this paper is the information diffusion approach based on selecting the most influential individuals from communities.

We employed an artificial scale-free and four real networks: *email*, *hamsterster*, *advogato* and *astrophysics*. We applied the SIR model for rumor propagation selecting the initial seeds from the whole network and from the communities. The impact that the Truncate (TP), Contact (CP) or Reactive (RP) processes have in the information diffusion was analyzed. The experimental results showed that the selection of individuals from the communities as initial spreaders, the final fraction of informed individuals is improved.

The remainder of the paper is organized as follows: Section 2 introduces some definitions and measures covered in this paper, the community detection algorithm applied and the propagation process in networks. Section 3 brings some related work. Section 4 presents the proposed approach for information diffusion based on communities. Section 5 exhibits the experimental results on an artificial scale-free and four real social networks. Finally, Section 6 discusses the final remarks.

2 Theoretical background

A network is a collection of items called nodes or vertices, which are joined together by connections called links or edges. Formally we define the network $G = (V, E, W)$, where $V = \{v_1, v_2, \dots v_n\}$ is the set of N vertices, $E = \{e_1, e_2, \dots e_m\}$ is the set of M edges and $W = \{w_1, w_2, \dots w_m\}$ are the weights of the edges that determine the strength of the interaction between the respective vertices, in the case of weighted networks. In mathematical terms, an undirected and unweighted network can be represented by an adjacency matrix A, in which two connected vertices i and j are in the matrix $a_{ij} = a_{ji} = 1$, otherwise, they are equal to 0.

A path is a consecutive sequence that starts at vertex i and ends in j, so that vertices are visited more than once. The distance or length of the path is defined as the number of edges contained in the sequence. The shortest distance between two vertices is known as the shortest path or geodesic path $g_{i,j}$. A component is the largest sub-set of vertices from the network in which exist at least one path between each pair of vertices, but never connect to another component. A connected network has only one component. When the networks have more than one component, we considered the largest of them.

The degree or connectivity of vertex i, called as k_i, is the number of edges or connections that vertex i has. In the case of directed networks is the sum of the degrees of input (edges that reach the vertex) and output (edges that leave the vertex). The average degree $\langle k \rangle$ is the average of all k_i of the network. The vertices that have a very high degree in the network are called hubs.

The degree distribution of a network $P(k)$ is the probability of randomly select a vertex with degree k. Social networks present scale-free degree distribution (Barabási, 2007; Newman, 2010), with $P(k) \sim k^{-\gamma}$, in which most of the individuals have low degree, near to the average, and only a few of them have very high degree (hubs). The level of disorder or heterogeneity in vertices connections is obtained with the entropy of degree distribution. We employed the normalized version of the Shannon entropy, i.e.

$$\tilde{H} = -\frac{\sum_{k=0}^{\infty} P(k)\log(P(k))}{\log(N)}\,, \qquad (1)$$

with $0 \leq \tilde{H} \leq 1$. The maximum value for the entropy occurs when $P(k)$ shows a uniform distribution and the lowest possible value happens when all vertices have the same degree. The entropy of a network is related to the robustness and their level of resilience.

The robustness is also related to the correlation degree of the network. A network is assortative, or positive correlated, if vertices tend to connect with vertices with similar degree. A Network is dissassortative, or negative correlated, if vertices with low degree tend to connect with higher connected vertices (hubs). When networks do not present any of above patterns, they are called as non-assortative. For the calculation of the network correlation we employed the Pearson coefficient, formulated with adjacency matrix as

$$\rho = \frac{(1/M)\sum_{j>i} k_i k_j a_{ij} - \left[(1/M)\sum_{j>i}\frac{(k_i+k_j)a_{ij}}{2}\right]^2}{(1/M)\sum_{j>i}\frac{({k_i}^2+{k_j}^2)a_{ij}}{2} - \left[(1/M)\sum_{j>i}\frac{(k_i+k_j)a_{ij}}{2}\right]^2} \qquad (2)$$

where M is the total number of edges in the network. If $r > 0$, then the network is assortative. If $r < 0$ the network is dissassortative. If $r = 0$, then there is no correlation between the degree of vertices.

2.1 Centrality measures

In complex and social networks have been proposed measures to describe the importance or centrality of vertices (Costa et al., 2007) according to topological and dynamical properties. The centralities adopted in this work are briefly described as follow.

Degree centrality (*DG*) is related with the number of connections or popularity of a vertex (Costa et al., 2007) and in terms of the adjacency matrix is expressed as

$$k_i = \sum_{i \in N} a_{ij} \, . \tag{3}$$

Betweenness centrality (*BE*) quantifies the number of shortest paths that pass through a vertex j between all pair of different vertices (i, k) (Freeman, 1977). It expresses how much a vertex B_j works as bridge or is a trusted vertex in the transmission of information, i.e.

$$B_j = \sum_{i,k \in V, i \neq k} \frac{\sigma_{ik}(j)}{\sigma_{ik}} \, , \tag{4}$$

where σ_{ik} is the total number of different shortest path between i and k, and $\sigma_{ik}(j)$ is the number of times j appears in those paths.

PageRank centrality (*PR*) expresses the importance of a vertex according to the probability that other vertices have to arrive at it, after a large number of steps (Brin and Page, 1998). The idea is to simulate the surfing on the net. The user can follow the links available at the current page or jump to other by typing a new URL. In social terms, it can be approached like the more cited or renowned individuals. The formalization of the PageRank centrality is

$$\vec{\pi}^t = \vec{\pi}^{t-1} \mathbb{G} \, , \tag{5}$$

where $\vec{\pi}^t$ are the PageRank values for each vertex in the t^{th} step of navigation and $\mathbb{G}$ is known as the Google matrix. When $t = 0$ we have by default $\vec{\pi}^0 = \{1, \ldots, 1\}$. The jumps are represented by a probability α and we adopted the same value as defined in the original version (Brin and Page, 1998).

2.2 Community detection

Communities are sets of densely interconnected vertices and sparsely connected with the rest of the network (Newman, 2010). Vertices that belong to the same community, in general, share common properties and perform similar roles. Therefore, the division of a network into communities helps to understand their topological structure (structural and functional properties) and its dynamic processes, obtaining relevant information and features to the network domain.

We can evaluate a partition based on the scores obtained from a quality measure. The goal is to evaluate expected features in a good community division. One of the most popular quality measures is the modularity Q (Newman, 2004). It compares the current density of intra-community and inter-community edges relative to a random network with similar characteristics. It is based on the fact that random networks have no community structure.

Given a network with c communities, the Q modularity is calculated by a symmetric matrix $N \times N$, in which elements along the main diagonal e_{ii} represent connections into the same community and elements e_{ij} represent connections between different communities i and j. Equation 6 shows the formulation of Q.

$$Q = \sum_i \left[e_{ij} - \left(\sum_j e_{ij} \right)^2 \right] \tag{6}$$

If a specific division provides less edges between communities than would be expected by random connections, the value of Q would be 0. When the network has isolated communities the value of Q would be 1. This measure is employed by several techniques to identify communities in networks systems, especially in divisive and agglomerative methods (Guimera et al., 2003; Newman, 2004; Newman, 2006).

Newman (Newman, 2004) proposes an agglomerative method that is an optimized greedy algorithm, called *fastgreedy*. The approach starts with a copy of a real network of N vertices with no connections, producing at first N communities. At each iteration, two communities c_i and c_j, which have connections in real network, are chosen in order to obtain the greatest improvement of Q (Equation 7). A pruning is performed in the search space considering only the edges that exist between communities. Therefore, execution time is reduced when considering the new Q function (Equation 7).

$$\Delta Q_{ij} = 2 \left(e_{ij} - \frac{\sum_j e_{ij} \sum_i e_{ij}}{2M} \right) \tag{7}$$

The result can also be represented as a dendrogram. Cuts at different levels of the dendrogram produce divisions with greater or lesser number of communities, and the best cut yields the largest value of Q. The algorithm at each step has $O(M + N)$. Since there are at most $N - 1$ join operations required to build a complete dendrogram, their overall complexity is $O((M + N)N)$, or $O(N^2)$, for sparse graph. Consequently, by adopting this method is more treatable the analysis of communities in larger networks.

2.3 Propagation process on networks

In classical rumor diffusion models the ignorant or inactive individuals (S) are those who remain unaware of the information, the spreaders (I) are those who disseminate the information and the stifler (R) are those who know the information but lose the interest in spreading it. All vertices have the same probability β for transmit the information to their neighbors and μ for stopping to be active.

The Maki-Thompson (MT) (Maki and Thompson, 1973) model is a spreader-centric approach employed for describing the propagation of ideas and rumors on networks. In the MT process whenever an active spreader i contacts a vertex j that is inactive, the latter will become active with a fixed probability β. Otherwise, in the case that j knows about the rumor, it means j is an active spreader or a stifler, the vertex i turns into a stifler with probability μ. The behavior when the spreader stops to propagate can be understood because the information is too much known (contacting spreaders) or without novelty (contacting stifler).

Three possible choices related with the spreader behavior during the diffusion have been reported (Borge-Holthoefer et al., 2012; Meloni et al., 2012). They are the Reactive process (RP), Truncated process (TP) and Contact Process (CP). However, a clear analysis about the impact of spreaders behavior in the propagation process has not been tackled yet. Moreover, there is not a consensus about what to employ in rumor or information diffusion and it may cause a misinterpretation of results. The three main characteristic behaviors reported to spreaders are described as follow.

- Reactive Process (RP): In each iteration, the spreaders try to pass the rumor among all their ignorant neighbors. After that, it evaluates whether it will become stifler in the next iteration or not, considering the contact with all their spreader and stifler neighbors.
- Truncated Process (TP): It consists of truncate or interrupt the contagion in the precise time. In each iteration and for each spreader, it is randomly selected one neighbor at time, and setting up the states of the contact as corresponds. The selection continues until the spreader visit all its neighbors or it becomes stifler, whichever occurs first.
- Contact Process (CP): In each time step and for each spreader, it is chosen at random a single neighbor. Then, it is resolved the transition states according to the rule that corresponds. After that, continues with the next spreader of the network of the same time step.

Different theoretical models have been proposed for modeling the rumor dynamics on networks (Moreno et al., 2004; Barrat et al., 2008; Castellano et al., 2009; Borge-Holthoefer et al., 2012). These analytical models make assumptions about the network structure, such as the degree correlation or distribution, compartments or class of vertices with same probabilities, homogeneous/heterogeneous mixing or mean field theory. Notwithstanding all of them claim that their numerical solutions agree with the MC simulations, so we adopt this approach as an exploratory research.

3 Related work

Many approaches have been developed in order to understand the propagation of ideas or information through social networks (Castellano et al., 2009). Specially, characterize the individuals that are most influential in the propagation process has attracted the attention of researchers (Richardson and Domingos, 2002; Kempe et al., 2003; Kitsak et al., 2010; Pei and Makse, 2013).

The conventional approach for describing the most influential vertices is performing a microscopic analysis on the network. Vertices are classified considering their topological properties, sorted and ranked in order to generalize their ability to propagate (Kitsak et al., 2010; Hébert-Dufresne et al., 2013; Pei and Makse, 2013). However, to find the set of initial vertices that maximize the propagation capacity, the selection of the most influential spreader may produce an overlap of influence in the population (Kitsak et al., 2010; Pei and Makse, 2013).

In terms of topological properties, there not exists a consensus about what is the more accurate measure that describes the most influential vertices. Some researches claim that hubs are more representative to influence others vertices (Pastor-Satorras and Vespignani, 2001; Albert and Barabási, 2002). Vertices with higher degree are more efficient to maximize the propagation because, in general, hubs not tend to connect with each other and thus can achieve a greater number of vertices (Kitsak et al., 2010). In the case of communities, the degree proportion of a vertex i is defined as the number of edges that i has in each community. This degree proportion was found as a good descriptor of influence for communities (Lawyer, 2012).

On the other hand, the most influential vertices are described as those with the largest Betweenness centrality (Hébert-Dufresne et al., 2013), because they intermediate the communication between groups of vertices, which increase their influence. According to the authors, Betweenness centrality is a better descriptor of the most influential spreader in communities.

The PageRank is also considered a better measure to describe the most influential vertices (Cataldi et al., 2010). The reason is that it employs the random walk concept over the network to be calculated and vertices with higher values mean higher probability to be visited.

Finally, Kempe et al. (2003) propose a greedy algorithm to obtain η initial spreaders that maximizes the diffusion influence. The authors adopt a discrete optimization approach and prove that the optimization problem is NP-hard. It was implemented considering the independent and weighted cascade model that have only two states, which are different to the SIR model. The method evaluates one vertex at time to be added in the set of selected seeds. The new vertex is accepted if it is what most increment the diffusion. However, this approach has a very higher computational cost problem, although new researches try to optimize the performance (Chen et al., 2009).

4 Information diffusion by communities

Let us consider a constant population of N vertices in all time steps. Each vertex can be only in one state, that is $I_i(t) = 1$ *iff* $i \in I$, otherwise $I_i(t) = 0$, and $S_i(t)+I_i(t)+R_i(i) = 1$. The macroscopic fraction of ignorant ($\psi(t)$), spreaders ($\phi(t)$) and stifler ($\varphi(t)$) over time is calculated as $\psi(t) = \frac{1}{N}\sum_{i=1}^{N} S_i(t)$, that is similar to the other states and always fulfill $\psi(t) + \phi(t) + \varphi(t) = 1$. We assume that infection and recovering do not occur during the same discrete time window or step.

4.1 Setup

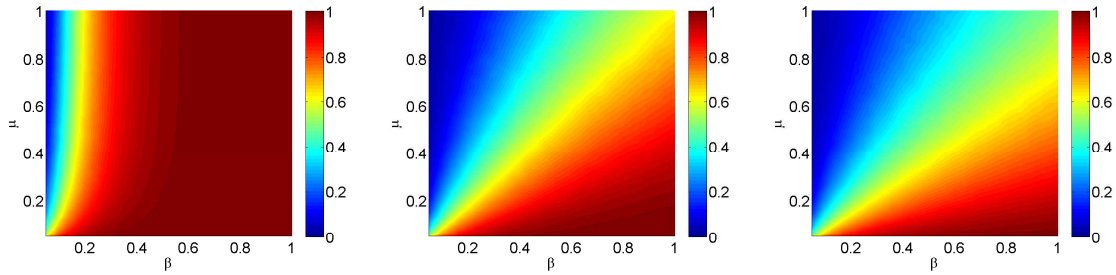
The initial setup for the propagation is $\psi(0) = 1 - \eta/N$, $\phi(0) = \eta/N$ and $\varphi(0) = 0$, where η represents the seeds or number of initial spreaders. Each simulation begins with a selection of η vertices. At each time step, all spreaders uniformly select and try to infect its neighbors with probability β, or stop the diffusion with probability μ according to the spreader behavior adopted. Successful change of state (to be spreader or to be stifler) are effective at the next iteration. The simulations run until the end of the process is reached, when $\phi(\infty) = 0$.

4.2 Community selection approach

We propose to select the initial spreaders from the community division of the network. The multiple seeds are the most central vertices of each community. The community division may be calculated by some divisive or agglomerative method (Section 2.2) and here the *fastgreedy* algorithm was employed. The method is detailed as follow:

First, given a required number of η initial spreaders, we find the η main communities of the network by the *fastgreedy* algorithm. Then, each community is isolated, which produces η components. The isolation process consists in maintaining the intra-community edges and erasing the inter-community connections. For each isolated community, a specific centrality measure is calculated to all vertices. Since vertices with higher centrality are considered more suitable to influence on the network, we select the most important vertex from each community. Therefore, these vertices influence more in their own community and the overlap of influence in the population is minimized. At the end, η seeds are selected and they have the best centrality value of its community. We take the original full network, the η seeds, the parameters and execute the corresponding simulations.

For the centrality measure, the point is to find what centrality better identifies the influential spreaders, by communities and in the whole network, that maximizes the information diffusion.



(a) MT-RP (b) MT-TP (c) MT-CP

Figure 1: MT (Maki and Thompson, 1973) propagation in an artificial scale-free network with $N = 1000$, $\langle k \rangle = 8$ and $\eta = 1\%$ of initial seeds selected. The color bar shows the final fraction of informed individuals. The behavioral approaches for spreader analyzed are: (a) Reactive process (RP); (b) Truncated process (TP); and (c) Contact process (CP)

Here, the degree (*DG*), PageRank (*PR*) and Betweenness (*BE*) centralities were considered.

5 Experimental results

In this section we analyzed the information diffusion in an artificial scale-free and four real social networks. We evaluated the impact spreaders behavior have in the diffusion on the networks. Then, the results about the selection of initial spreaders by communities, best-ranked vertices of the network and random seeds were explored.

5.1 Spreader behavior analysis

We analyzed the three behavioral approaches for the spreaders and present the impact they produce on the propagation process. We considered the MT model with an artificial scale-free network of size $N = 1000$ and $\langle k \rangle = 8$. In order to understand the overall spectral effect with the parameters, the simulations were evaluating a range of β and μ in $(0, 1]$. Therefore, the differences between the approaches are evidenced. For each tuple of values (β, μ), it was selected 100 times at random $\eta = 1\%$ of initial spreader (seeds) and each time was an average over 50 executions.

The impact of the behavioral approach in the final fraction of informed individuals is shown in Figure 1. We observed that the CP approach is less redundant in the number of contacts made by spreader, producing lower fractions of informed individuals, in comparison to the other behaviors. Still, because the single contact made by iteration, the CP behavior is more similar to a propagation through the "word-of-mouth" situation.

The RP approach obtained more than 80% of informed individuals with values of $\beta \geq 0.3$, no matter values of μ. Therefore, the RP approach favors a viral diffusion on the network with lower values of β and it happens independently of which are the initial seeds. For this reason, RP is a more suitable approach to simulate broadcasting propagation.

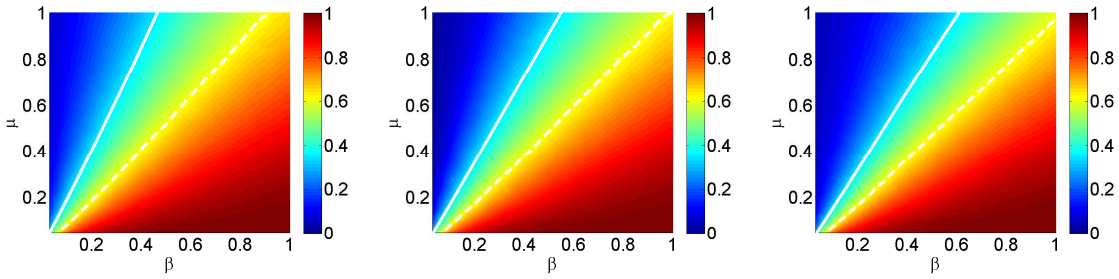
On the other hand, the TP behavior is more related to the contact network scenario, where the position and topological characteristics of seeds may have influence in the diffusion. Moreover, TP presents more balanced results, near 60% when $\beta \approx \mu$, and contacts are not as restricted as CP behavior. For this reason, we adopted hereafter the MT-TP approach as the propagation process for the analysis.

5.2 Multiple initial spreader analysis

The experiments were performed with three possibilities for choosing the initial spreader: (*i*) by randomly selecting η individuals as initial seeds in the network; (*ii*) by selecting the best- ranked η individuals with highest value of a specific centrality of the network; and finally, (*iii*) by detecting η communities on the network and for each isolated community selecting the individual with highest value of a specific centrality measure. The centrality measures selected were degree (*DG*), Betweenness (*BE*) and Pagerank (*PR*).

5.2.1 Real social networks

We adopted the *email* (Guimera et al., 2003), *advogato* (Kunegis, 2014a), *astrophysics* (Newman, 2001) and *hamsterster* (Kunegis, 2013; Kunegis, 2014b). All of them were assumed as undirected and unweighted networks and also it was considered the largest component for the simulations. The structural properties of the networks are summarized in Table 1, with the respective number of vertices N, the average degree $\langle k \rangle$, shortest paths



(a) MT-TP Community (b) MT-TP Random (c) MT-TP Best

Figure 2: MT-TP propagation in an artificial scale-free network with $N = 1000$ and $\langle k \rangle = 8$. The final fraction of informed individuals are shown in the color bar. The selection of $\eta = 4\%$ of seeds was made by: (a) η communities taking the individual with best PR centrality of each one; (b) uniform random selection of individuals; and (c) the η individuals with the best PR centrality of the network. Solid white lines to the left in the contour plots show the β and μ combinations that achieved 35% of informed individuals. Dashed white lines show the combinations that achieved 60% of informed individuals.

Table 1: Topological properties and results of community detection of the networks: last column, the best modularity Q and community division by fastgreedy algorithm

Network	N	$\langle k \rangle$	$\langle g \rangle$	$\tilde{H}$	ρ	FastGreedy	
						Q	Nc
email	1133	9.62	3.60	0.45	0.01	0.49	16
hamsterster	2000	16.1	3.58	0.48	0.02	0.46	57
advogato	5054	15.6	3.27	0.40	-0.09	0.34	49
astrophysics	14845	16.1	4.79	0.38	0.23	0.63	1172

average $\langle g \rangle$, normalized entropy $\tilde{H}$, pearson correlation ρ. Also, the best modularity Q value and division number of communities NC of the networks produced by the FastGreedy algorithm are reported.

email represents a social network of information exchanged by emails between members of the *Rovira i Virgili* University, Tarragona, with largest hub degree equal to 71.

hamsterster is an undirected and unweighted network based on the website data HAMSTERSTER.COM. The edges represent a relationship of family or friend among users. The largest hub has degree equal to 273.

advogato is an online community platform for developers of free software launched in 1999. Vertices are users of advogato, the directed edges represent trust relationships. The largest hub has degree equal to 807.

Finally, *astrophysics* is a collaborative network between scientists on previous studies of astrophysics reported in arXiv during January 1, 1995 until December 31, 1999. The network is weighted and directed and originally it has 16707 vertices. The largest hub of the main component has 360 connections.

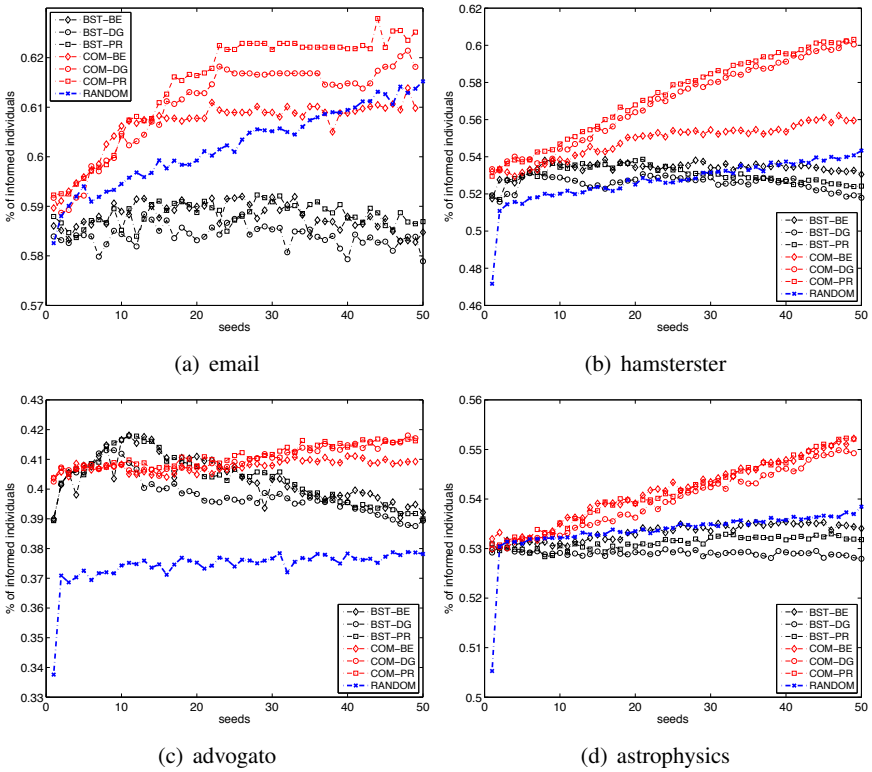
5.2.2 Information diffusion results

The final fraction of informed individuals ($\varphi(\infty)$) was averaged over 100 executions for each combination of initial seeds and parameters. This average represents the propagation capacity achieved by the selected seeds.

We evaluated the relation between the parameters and the selection of the initial spreaders in an artificial network. In this experiment the *PR* was defined as the centrality measure employed to find the seeds in the communities and the whole network. A value of $\eta = 4\%$ of initial spreaders was adopted for a scale-free network of size $N = 1000$, $\langle k \rangle = 8$, $\langle g \rangle = 3.19$, $\tilde{H} = 0.33$ and $\rho = -0.04$.

The propagation capacity $\varphi(\infty)$ was affected according to the initial seeds (Figure 2). The solid and dashed white curves represent the combination of β and μ parameters that obtained 35% and 60% of informed individuals respectively. We observed that these curves show a well defined linear pattern, which means any proportion of $\lambda = \beta/\mu$ will obtain equivalent $\varphi(\infty)$ results.

The selection of seeds by communities (Figure 2(a)) improved the diffusion on the network in comparison with the Random seeds (Figure 2(b)) and Best-ranked vertices (Figure 2(c)). This result is corroborated by the increase of the white lines



(a) email (b) hamsterster

(c) advogato (d) astrophysics

Figure 3: Propagation capacity of MT-TP model in the four real networks given the selection of seeds by communities (red points), best-ranked (black points), and randomly (blue points).

Table 2: Average of propagation capacities for the full range of $\eta \in [1, 50]$, for each network achieved by seeds: (second big column) selecting the most important individuals from communities; (third big column) selecting the best-ranked individuals of the network; and (last column) randomly selecting the initial seeds. The adopted measures were Betweenness (*BE*), degree (*DG*) and PageRank (*PR*) centralities

Network	Community			Best ranked			Random selection
	BE	*DG*	*PR*	*BE*	*DG*	*PR*	
email	0.6065	0.6105	**0.6150**	0.5880	0.5840	**0.5884**	0.6023
hamsterster	0.5485	0.5693	**0.5728**	**0.5306**	0.5226	0.5271	0.5273
advogato	0.4077	0.4102	**0.4112**	0.3993	0.3958	**0.4007**	0.3805
astrophysics	**0.5417**	0.5398	0.5415	**0.5321**	0.5278	0.5301	0.5337

slope. However, a little decrease in the lines slope is evidenced in the *MT-TP Best* case with respect to the *MT-TP Random* case.

Consequently, we sought to analyze the impact of η and centrality measures in the selection of seeds in the diffusion process. We varied the number of communities and seeds from 2 to 50 and fixed $\beta = 0.3$ and $\mu = 0.2$ for all simulations. The real social networks described and the MT-TP propagation model were considered in the analysis (Figure 3). The random selection of initial spreader (blue points, *RANDOM*) or best-ranked vertices (black points, *BST-***) of *DG*, *BE* or *PR* centrality, produced a constant propagation capacity ($\varphi(\infty)$). In some case, random selection of seeds reached a higher propagation capacity than the selection of best-ranked vertices. For a larger number of initial spreaders, $\varphi(\infty)$ tend to fall when the best-ranked vertices are selected.

On the other hand, when the community detection was performed and individuals with highest values of *DG*, *BE* or *PR* in each community (red points, *COM-***) were selected, the propagation capacity was improved and achieved the best results. Therefore, more individuals were informed in the network by the community selection, with the same propagation constraint (number of seeds).

In terms of the topological measures, we observed that vertices with highest PageRank cen-

trality in the communities (*COM-PR*) obtained in average the best propagation results (Table 2). Even in the selection of the best-ranked vertices, the PageRank was notable. Another important point is that often, the uniformly random selection of initial spreader could be a better option than select the most central vertices (best-ranked) of the network. This is contrary what is currently expected and adopted in marketing campaigns, for instances. For all networks and for all size of seeds, we evidenced that starts the diffusion from the best-ranked vertices produces lower influence, or final fraction of informed individuals, than purely select vertices at random; in some cases, the best-ranked selection achieved the worst results. However, the selection of initial spreaders by communities showed, independently of the centrality measure, higher results.

6 Final remarks

In this work, we proposed a method for maximizing the information diffusion on networks. First, we analyzed the impact of the spreader behavior in the propagation and confirmed that the Truncate Process (TP) is more suitable to simulate information diffusion on networks. We applied community detection and targeted the most influential vertices from these communities as initial seeds. Experimental results on an artificial scale-free and four real social networks confirmed the increase in the final fraction of informed individuals. Moreover, it was found that the PageRank centrality in communities was a better choice in terms of efficiency and influence maximization.

A brief overview about complex network measures, community detection and information propagation was introduced. We present our proposal to select initial spreaders by communities. There is still an open problem related to an exact definition of what is considered a community and what would be the ideal division. Nevertheless, we varied the number of communities from 2 to 50 and in general (for every community division) our proposal achieved better results versus propagation without considering the community structure.

In future work, other measures for selecting influential individuals on networks could be explored, in addition to *DG*, *BE* and *PR* applied here. Also, other models of propagation and network topologies could be tested, as well as novel strategies taking into account community information.

7 Acknowledgments

This research was partially supported by National Council for Scientific and Technological Development (CNPq) grant: 140688/2013-7 and São Paulo Research Foundation (FAPESP) grant: 2011/21880-3.

References

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, jan.

A.-L. Barabási. 2007. The architecture of complexity: From network structure to human dynamics. *IEEE Control Systems Magazine*, 27(4):33–42.

Alain Barrat, MarseilleMarc Barthélemy, and Alessandro Vespignani. 2008. *Dynamical Processes on Complex Networks*. Cambridge University Press.

Javier Borge-Holthoefer, Sandro Meloni, Bruno Gonçalves, and Yamir Moreno. 2012. Emergence of Influential Spreaders in Modified Rumor Models. *Journal of Statistical Physics*, 151(1-2):383–393, September.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, V:107–117.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical Physics of Social Dynamics. *Reviews of Modern Physics*, 81(2):591–646, may.

Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining - MDMKDD '10*, pages 1–10, New York, New York, USA, jul. ACM Press.

Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 199, New York, New York, USA, jun. ACM Press.

L. D. F. Costa, F. A. Rodrigues, G Travieso, and P. R. Villas Boas. 2007. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56:167–242.

L C Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41.

R Guimera, L Danon, A Diaz-Guilera, F Giralt, and A Arenas. 2003. Self-similar community structure in a network of human interactions. *Physical Review E*, 68:2003.

Laurent Hébert-Dufresne, Antoine Allard, Jean-Gabriel Young, and Louis J Dubé. 2013. Global efficiency of local immunization on complex networks. *Scientific reports*, 3:2171, January.

David Kempe, Jon M. Kleinberg, and Éva Tardos. 2003. Maximizing the Spread of Influence Through a Social Network. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *KDD*, pages 137–146. ACM.

Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. 2010. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, August.

Jrme Kunegis. 2013. KONECT – The Koblenz Network Collection. In *Proc. Int. Web Observatory Workshop*, pages 1343–1350.

Jrme Kunegis. 2014a. Advogato network dataset – KONECT, October.

Jrme Kunegis. 2014b. Hamsterster full network dataset – KONECT, jan.

Glenn Lawyer. 2012. Measuring node spreading power by expected cluster degree. page 4, September.

D. P. Maki and M Thompson. 1973. *Mathematical Models and Applications, with Emphasis on the Social, Life, and Management Sciences*. Prentice-Hall.

Sandro Meloni, Alex Arenas, Sergio Gmez, Javier Borge-Holthoefer, and Yamir Moreno. 2012. Modeling epidemic spreading in complex networks: Concurrency and traffic. In My T. Thai and Panos M. Pardalos, editors, *Handbook of Optimization in Complex Networks*, Springer Optimization and Its Applications, pages 435–462. Springer US.

Yamir Moreno, Maziar Nekovee, and Amalio F. Pacheco. 2004. Dynamics of rumor spreading in complex networks. *Physical Review E*, 69(6):066130, jun.

M. E. J. Newman. 2001. The structure of scientific collaboration networks. In *Natl. Acad. Sci. USA*, number 98, pages 404 – 409.

M E J Newman. 2004. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(3):66133.

M E J Newman. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):36104.

Mark Newman. 2010. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, 86(14):3200–3203, April.

Sen Pei and Hernán A Makse. 2013. Spreading dynamics in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(12):P12002, December.

Matthew Richardson and Pedro Domingos. 2002. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, page 61, New York, New York, USA, jul. ACM Press.