Community Detection in Complex Networks

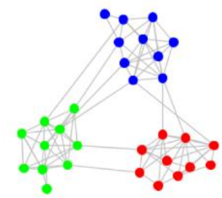
Liang Zhao

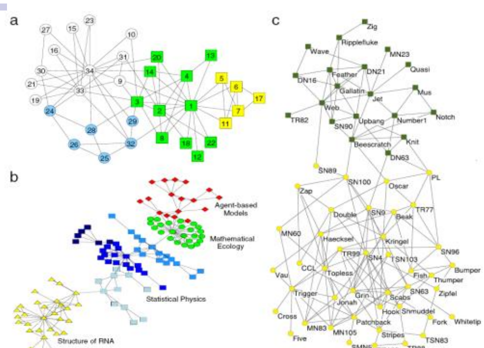
zhao@icmc.usp.br

Department of Computer Science. Institute of Mathematics and Computer Science. University of São Paulo. São Carlos-SP. Brazil

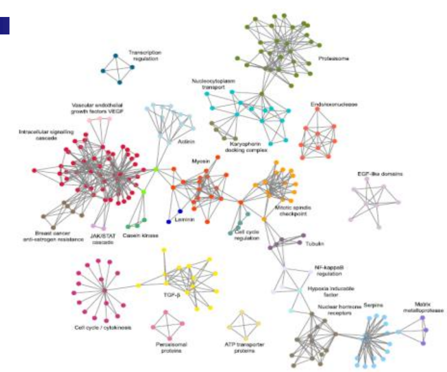
Introduction

- Communities are defined as a subgraph whose nodes are densely connected within itself but sparsely connected with the rest of the network.

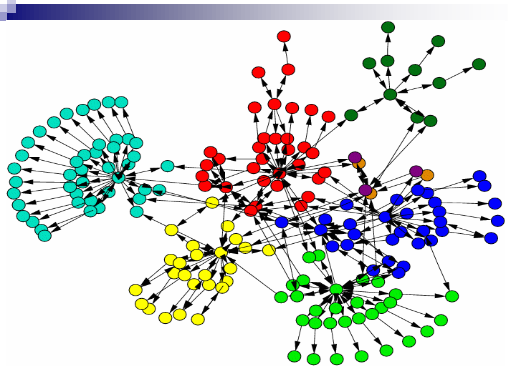




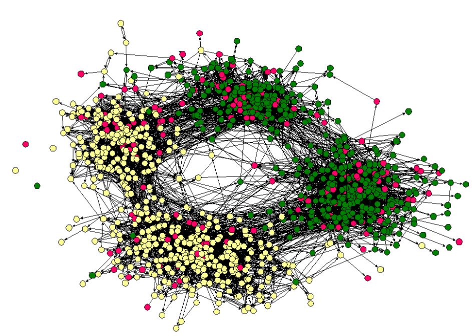
Community structure in social networks. (a) Zachary's karate club, a standard benchmark in community detection. (b) Collaboration network between scientists working at the Santa Fe Institute. (c) Lusseau's network of bottlenose dolphins.



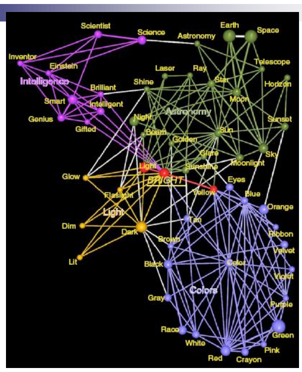
Community structure in protein-protein interaction networks. The graph pictures the interactions between proteins in cancerous cells of a rat. Communities, labeled by colors, were detected with the Clique Percolation Method by Palla et al.



Community structure in technological networks: WWW and communities (groups of homepages having topical similarities)



Network of friendship of high school students



Overlapping communities in a network of word association. The groups, labeled by the colors, were detected with the Clique Percolation Method by Palla et al.

Graph Partition Problem

Consider a graph $G(V, E)$, where V denotes the set of vertices and E the set of edges. The standard (unweighted) version of the graph partition problem is: Given G and an integer $k > 1$, partition V into k parts (subsets) $V_1, V_2, ..., V_k$ such that the parts are disjoint and have equal size, and the number of edges with endpoints in different parts is minimized. In practical applications, a small imbalance ε in the part sizes is usually allowed, and the balance criterion is

$$\max_i |V_i| \leq (1+\varepsilon)\frac{|V|}{k}.$$

Graph partitioning is known to be NP-Complete.

Michael R. Garey, David S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, W.H.Freeman & Co Ltd, 1979.

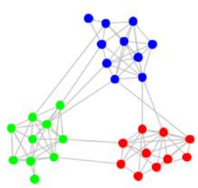
Community Detection Techniques

- Betweenness Measure

$$c_B(v) = \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v)$$

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$\delta_{st}(v)$ denote the fraction of shortest paths between s and t that contain vertex v and σ_{st} denotes the number of all shortest-path between s and t.



M. E. J. Newman and M. Girvan, Physical Review E, vol. 69, pp. 026113, 2004.

Community Detection Techniques

- Modularity Measure

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

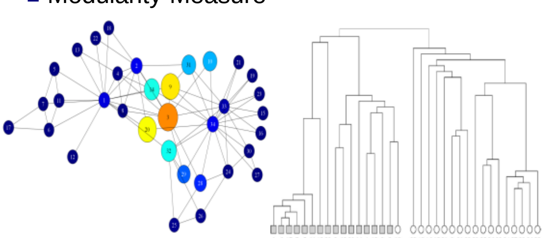
$$A_{ij} = \begin{cases} 1 & \text{if there is an edge joining vertices } i, j, \\ 0 & \text{otherwise.} \end{cases}$$

where the sum runs over all pairs of vertices, A is the adjacency matrix, m the total number of edges of the graph. The δ-function yields one if vertices i and j are in the same community ($C_i = C_j$), zero otherwise. k_i and k_j are degree of node i and j, respectively.

M. E. J. Newman, Physical Review E, vol. 69, pp. 066133, 2004.
A. Clauset, M. E. J. Newman, and C. Moore, Physical Review E, vol. 70, pp. 066111, 2004.

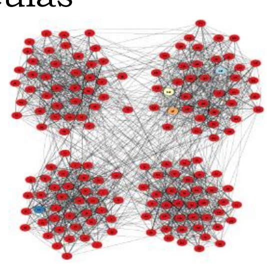
Community Detection Techniques

- Modularity Measure



M. E. J. Newman, Physical Review E, vol. 69, pp. 066133, 2004.

Técnica de Competição de Partículas



M. G. Quiles, L. Zhao, R. L. Alonso, and R. A. F. Romero, "Particle competition for complex network community detection," *Chaos*, vol. 18, no. 3, p. 033107, 2008.

12

Dinâmica de Partículas

$$\rho_j^{v}(t+1) = v_i, \tag{1}$$

$$\rho_j^{\omega}(t+1) = \begin{cases} \rho_j^{\omega}(t) & \text{if } v_i^{\rho}(t) = 0, \\ \rho_j^{\omega}(t) + (\omega_{\max} - \rho_j^{\omega}(t))\Delta_{\rho} & \text{if } v_i^{\rho}(t) = \rho_j \neq 0, \\ \rho_j^{\omega}(t) - (\rho_j^{\omega}(t) - \omega_{\min})\Delta_{\rho} & \text{if } v_i^{\rho}(t) \neq \rho_j \neq 0, \end{cases} \tag{2}$$

Dinâmica de Vértices

$$v_i^{\rho}(t+1) = \begin{cases} v_i^{\rho}(t) & \text{if } v_i^{\gamma} = 0, \\ \rho_j & \text{if } v_i^{\gamma} = 1 \text{ and } v_i^{\omega}(t) = \omega_{\min}, \end{cases} \tag{3}$$

$$v_i^{\omega}(t+1) = \begin{cases} v_i^{\omega}(t) & \text{if } v_i^{\gamma} = 0, \\ \max\{\omega_{\min}, v_i^{\omega}(t) - \Delta_v\} & \text{if } v_i^{\gamma} = 1 \text{ and } v_i^{\rho}(t) \neq \rho_j, \\ \rho_j^{\omega}(t+1) & \text{if } v_i^{\gamma} = 1 \text{ and } v_i^{\rho}(t) = \rho_j, \end{cases} \tag{4}$$

Política de Movimentação de Partículas

- Caminhada Aleatória
  - A partícula aleatoriamente seleciona um vizinho para visitar:

$$p(v_j \mid \rho_k = v_i) = \frac{A_{ij}}{\sum_{q=1}^{n} A_{iq}}$$

- Caminhada Determinística
  - A partícula prefere visitar o vizinho que ela tem a maior dominância:

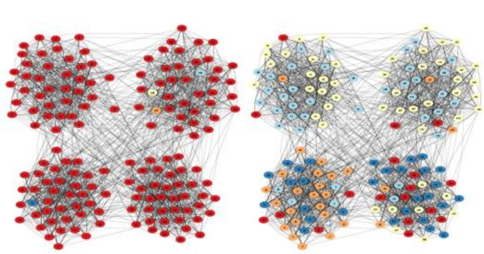
$$p(v_j \mid \rho_k = v_i) = \frac{A_{ij} v_i^{\omega_j}}{\sum_{q=1}^{n} A_{iq} v_i^{\omega_j}}$$

As partículas devem apresentar os dois movimentos, a fim de alcançar um equilíbrio entre o comportamento exploratório e defensivo

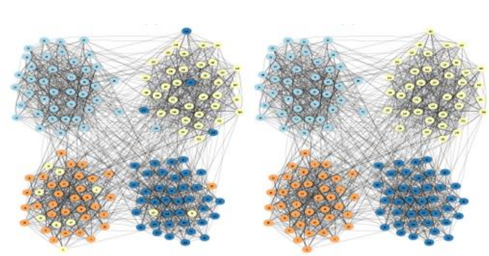
Caminhada Aleatório-Determinística

- Definimos uma probabilidade $0 \leq p_{det} \leq 1$. Em cada iteração, cada partícula tem probabilidade p_{det} para realizar uma caminhada determinística e $1\text{-}p_{det}$ para fazer uma caminhada aleatória
- Em particular, o movimento de uma partícula é completamente aleatória se $p_{det} = 0$ e é completamente determinística se $p_{det} = 1$

Estudo Numérico



Estudo Numérico



Estudo Numérico

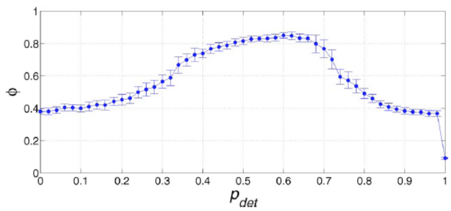


FIG. 2. (Color online) Correct community detection rate ϕ vs probability of determinism p_{det}. In these simulations, $N=128$, $M=4$, $\langle k\rangle=16$, and $z_{\text{out}}/\langle k\rangle=0.5$. Each point in the trace is averaged by 200 realizations. The error bars represent standard deviations.

19

Estudo Numérico

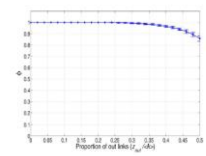


FIG. 8. (Color online) Community detection rate using the proposed technique applied to clustered random networks with $N=128$, $M=4$, and $\langle k\rangle=16$. The x axis represents the proportion of out links $z_{\text{out}}/\langle k\rangle$ and the y axis represents the portion of nodes correctly classified. Each point in the trace is average by 200 realizations. The error bars represent the standard deviations. $p_{\text{det}}=0.6$. For $z_{\text{out}}/\langle k\rangle\leq 0.1$, $\Delta_p=0.6$; for $z_{\text{out}}/\langle k\rangle>0.1$, $\Delta_p=0.4$.

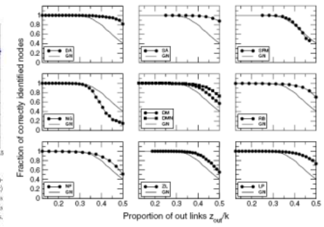


Figure 2. Comparing algorithm sensitivity using ad hoc networks with predetermined community structure. The x-axis is the proportion of connections to outside communities z_{out}/k and the y-axis is the fraction of nodes correctly identified by the method measure as described in [24]. The labels here correspond to the different methods and are listed in table 1.

Danon, L., A. Díaz-Guilera, J. Duch, & A. Arenas, "Comparing community structure identification". *Journal of Statistical Mechanics: Theory and Experiments*, P09008, 1–10, 2005.

20

Estudo Numérico

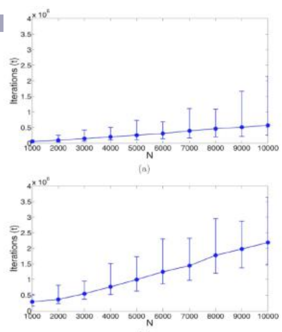


FIG. 9. (Color online) Number of iterations required to community detection vs network size N. In these simulations, the following parameters are assumed: $M=4$, $\langle k\rangle=0.5N/M$, and $p_{\text{det}}=0.6$. Each point in the trace is averaged by 200 realizations. The error bars represent the maximum and minimum number of iterations need to achieve $\langle v^{\infty}\rangle\geq 0.9$. (a) $z_{\text{out}}/\langle k\rangle=0.2$. (b) $z_{\text{out}}/\langle k\rangle=0.4$.

21

Técnica de Detecção de Comunidades via Sincronização

Definição dos Osciladores (I&D)

$$\frac{dv_i}{dt} = -v_i + I_i(t) + E_i(t) - Y_i(t)$$

- v_i - potencial do neurônio
- $I_i(t)$ - estimulação externa
- $E_i(t)$ - termo de acoplamento excitatório - cooperação
- $Y_i(t)$ - termo de acoplamento inibitório - competição

M. Quiles, L. Zhao, F. A. Breve, "Label Propagation Through Neuronal Synchrony". In: *2010 International Joint Conference on Neural Networks* (IJCNN'2010), Barcelona. v. 1. p. 2517-2524, 2010.
M. Quiles, L. Zhao, F. A. Breve, and R. A. F. Romero, "A network of integrate and fire neurons for visual selection". *Neurocomputing*, v. 72, p. 2198-2208, 2009.

22

Definição do Modelo

Termo de Acoplamento Excitatório

$$E_i(t) = \sum_{j\in\Delta_i} \omega_{ij}\delta(t-t_j)$$

$\omega_{ij} = \frac{c}{|\Delta_i|}$

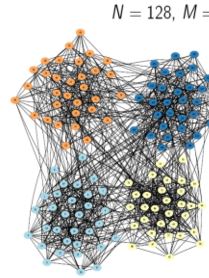
Termo de Acoplamento Inibitório

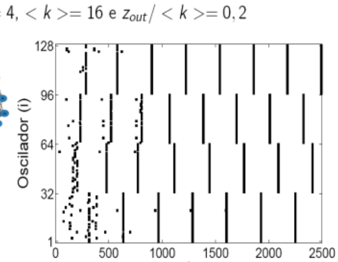
$$Y_i(t) = \frac{c}{N}\sum_{j=1;j\neq i}^{N} \delta(t-t_j)$$

23

Estudo Numérico

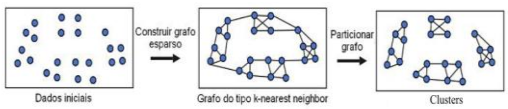
$N=128$, $M=4$, $<k>=16$ e $z_{out}/<k>=0,2$





24

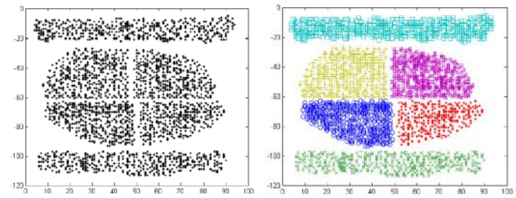
Agrupamento de Dados via Detecção de Comunidades



T. B. S. Oliveira, LIANG ZHAO, K. Faceli e A. C. P. L. F. de Carvalho, "Data Clustering Based on Complex Network Community Detection". In: *Proceedings of 2008 IEEE World Congress on Computational Intelligence* (WCCI 2008), Hong Kong, IEEE Computer Society, vol. 1. pp. 2121-2126, 2008.

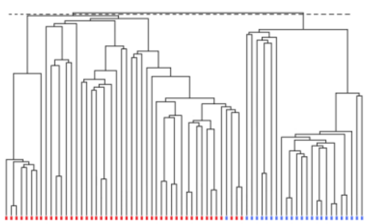
25

Agrupamento de Dados via Detecção de Comunidades



26

Agrupamento de Dados via Detecção de Comunidades



Dendrograma do resultado da simulação para a estrutura E1 do conjunto de dados Golub. A linha pontilhada corta o dendrograma em dois grupos que representam os dois tipos de leucemia: ALL (itens de dados em vermelho) e AML (itens de dados em azul).

27

Características da Técnica de Competição de Partículas

- O processo de competição de partículas é similar a muitos processos naturais
- Possui alta precisão de detecção de comunidades e ao mesmo tempo tem baixa ordem de complexidade computacional
- Revelou um novo tipo de "ressonância estocástica"
- Oferece uma alternativa para aprendizado competitivo diminuindo efeito de "caixa preta"
- Uma desvantagem – o modelo possui vários parâmetros

28