Detecção de comunidades em redes complexas: um modelo de correlação oscilatória

Marcos G. Quiles, Liang Zhao, Fabricio A. Breve & Roseli A. F. Romero

[1]Departamento de Ciências de Computação
Instituto de Ciências Matemática e de Computação
Universidade de São Paulo - São Carlos, SP, Brasil

{quiles, zhao, fabricio, rafrance}@icmc.usp.br.

Abstract. *One salient feature of complex networks is the presence of communities, or groups of densely connected nodes. Community detection can not only help to understand the topological structure of complex networks, but also provide new techniques for real applications, such as data mining. In this paper, we propose a new model for community detection based on the oscillatory correlation theory. This model has been applied to artificial and real networks and the results show its good performance and precision.*

Resumo. *Uma característica saliente em redes complexas é a presença de comunidades, ou grupos de vértices densamente conectados. A detecção de comunidades pode além de auxiliar na compreensão da estrutura topológica da rede também fornecer novas técnicas para aplicações reais, como por exemplo, em mineração de dados. Neste artigo, um novo modelo para detecção de comunidades baseado na teoria da correlação oscilatória é proposto. Este modelo foi aplicado em diversas redes artificiais e reais e os resultados obtidos mostram seu bom desempenho e precisão.*

1. Introdução

Uma característica notável observada em diversas redes complexas é a presença de estruturas modulares locais conhecidas como *comunidades* [Newman 2004a, Danon et al. 2005]. Tais comunidades podem ser definidas como grupos de vértices da rede densamente conectados, enquanto que conexões entre vértices pertencentes a grupos (comunidades) diferentes são esparsas [Newman & Girvan 2004]. Essas comunidades representam padrões de interação entre os vértices da rede e sua identificação é importante no entendimento dos mecanismos de crescimento e formação desta [Clauset 2005]. Além disso, um fator importante referente a estrutura das comunidades está na similaridade das características dos vértices que as compõem. Assim, por meio da identificação e estudo das comunidades é possível obter informações pertinentes ao domínio da rede. Por exemplo, observando-se a estrutura de ligações entre páginas da *world wide web* é possível constatar que páginas descrevendo tópicos relacionados tendem a ser mais densamente conectadas entre si do que com o restante da rede [Flake et al. 2002]. Esta propriedade também é compartilhada por redes reais provenientes de outros domínios, como redes biológicas [Jeong et al. 2000], rotas de transporte aéreo [Guimerà et al. 2003], redes metabólicas [Guimerà & Amaral 2005], dentre outras.

O processo de detecção de comunidades em uma rede não é computacionalmente trivial. Por exemplo, o problema de dividir um grafo em duas partes de mesmo tamanho de tal forma que número de arestas ligando estas partes seja mínimo é definido como um problema NP-Completo [Danon et al. 2005]. Para complicar ainda mais este problema, que pode ser visto como um caso simples da tarefa de detecção de comunidades, as redes reais podem ser compostas por um número não conhecido de comunidades e não apenas duas como no caso anterior. Além disso, as comunidades por si só podem ser definidas por estruturas hierárquicas na qual uma comunidade é formada por outras sub-comunidades aninhadas [Ravasz & Barabasi 2003, Danon et al. 2005]. Devido a importância do problema e a dificuldade computacional em sua solução, diversos autores têm proposto modelos computacionais para realizar de forma automática a detecção de comunidades em redes complexas.

Recentemente, vários métodos para detecção de comunidades têm sido propostos e aplicados em diversos domínios [Danon et al. 2005, Newman & Girvan 2004, Newman 2004b, Reichardt & Bornholdt 2004, Boccaletti et al. 2007, Quiles et al. 2008]. Dada a grande quantidade de modelos propostos e seus distintos mecanismos computacionais, uma forma tradicional de compará-los é através do uso de redes randômicas clusterizadas [Newman 2004b, Danon et al. 2005]. Tais redes são compostas por N vértices divididos em M grupos (comunidades). A rede é formada com base em duas probabilidades definidas a priori, p_{in} e p_{out}, que representam a probabilidade de ligação entre vértices de uma mesma comunidade e entre vértices de comunidades distintas, respectivamente. p_{in} e p_{out} são escolhidos de tal forma a controlar o número de intra-conexões z_{in} e o número de inter-conexões z_{out} para um dado grau médio de conexão definido $<k>$. Com base nestes parâmetros, a fração de intra-conexões $z_{in}/ <k>$ e a fração de inter-conexões $z_{out}/ <k>$ da rede são definidas, onde $(z_{in}/ <k> + z_{out}/ <k>) = 1$. Mais especificamente, para testar e comparar a precisão das técnicas de detecção de comunidade, redes com $N = 128$ vértices divididas em $M = 4$ comunidades iguais com $<k> = 16$ são utilizadas. Assim, partindo-se de redes com $z_{out}/ <k> = 0$, isto é, onde não existem ligações entre comunidades distintas, até $z_{out}/ <k> = 0.5$, no qual em média metade das arestas de um vértice estão ligadas a vértices da mesma comunidade e a outra metade estão ligadas a vértices de outras comunidades, é possível estudar de forma controlada a capacidade dos algoritmos.

Em [Danon et al. 2005], um estudo comparativo de diversas técnicas utilizando a metodologia descrita acima foi realizado. Neste estudo foi observado que, para redes com estrutura de comunidades bem definidas ($z_{out}/ <k> < 0,3$), a precisão de todos os algoritmos considerados é alta. Entretanto, conforme a proporção de ligações entre vértices de comunidades distintas se aproxima do número de ligações intra-comunidade, a precisão dos algoritmos é reduzida, uma vez que apenas alguns algoritmos são capazes de obter um acerto superior a 80%. De uma forma geral, os algoritmos que apresentam uma maior precisão no processo de detecção de comunidades expõem como desvantagem um maior custo computacional. Por outro lado, os algoritmos mais eficientes geralmente não apresentam alta precisão, principalmente quando as comunidades não estão bem definidas.

Considerando o dilema precisão-eficiência observado em diversas técnicas, neste trabalho uma nova técnica para detecção de comunidades baseada na teoria da correlação temporal, mais especificamente na teoria da correlação oscilatória, é proposta. Segundo

[von der Malsburg 1981] investigações das funções cerebrais e da organização perceptual indicam um mecanismo de correlação temporal como uma estrutura de representação no cérebro. A teoria de correlação temporal define que um objeto é representado pela correlação temporal dos disparos (potenciais de ação) de células neurais espacialmente distribuídas que representam diferentes características de um mesmo objeto, enquanto que neurônios codificando características de objetos distintos não possuem suas atividades correlacionadas. Uma forma natural de realizar a correlação temporal é dada através do uso de osciladores. Desta forma, cada segmento (objeto) é representado por um conjunto de osciladores com atividades síncronas, enquanto que segmentos distintos são representados por grupos de osciladores fora de sincronia. Esta forma especial da correlação temporal é denominada teoria da *Correlação Oscilatória* [Wang 2005]. No modelo aqui proposto, cada oscilador representa um vértice da rede de tal forma que grupos de osciladores densamente conectados, representando comunidades, têm seus atividades de disparo sincronizadas, enquanto que comunidades distintas permanecem com trajetórias não correlacionadas.

Além da busca por um modelo que apresente uma alta precisão a um baixo custo computacional, uma outra grande motivação deste trabalho está no fato de que a detecção de comunidades em redes complexas é tarefa importante por revelar estruturas topológicas na rede. Tais técnicas são importantes em aprendizado de máquina, como em agrupamento de dados [Karypis et al. 1999, Cook & Holder 2000, Schaeffer 2007]. De forma geral, a estrutura de comunidades revela similaridade por meio de conexões entre os vértices pertencentes a um mesmo grupo. Estas similaridades, por sua vez, podem revelar agrupamentos nos dados e, de maneira análoga, evidenciar classes em problemas de classificação. Além disso, por representar os dados em uma rede, classes ou agrupamentos de formatos não triviais podem ser obtidos. Como consequência, o desenvolvimento de novas técnicas de detecção de comunidades pode contribuir para o desenvolvimento de novos algoritmos para aprendizado de máquina.

A seguir, na Seção 2 o modelo proposto é apresentado. Os experimentos computacionais são descritos na Seção 3. Por fim, algumas conclusões são apresentadas na Seção 4.

2. O Modelo

Neste modelo, cada vértice da rede é representado por um neurônio do tipo Integra e Dispara (I&D) [Izhikevich 2004] acoplados por dois tipos de conexões: conexões excitatórias e conexões inibitórias. As conexões excitatórias formam um mecanismo cooperativo responsável por sincronizar grupos de neurônios que representam uma mesma comunidade. Por outro lado, as conexões de inibição, definida por um termo de inibição global, têm por finalidade segregar as diversas comunidades que compõem a rede de tal forma que cada comunidade seja representada por um único trem de pulso.

Cada vértice da rede é modelado por um neurônio do tipo I&D representado pela seguinte equação:

$$\frac{dv_i}{dt} = -v_i + I_i(t) + E_i(t) - Y_i(t) \tag{1}$$

onde v_i representa o potencial do neurônio, I_i define a estimulação externa, $E_i(t)$ define

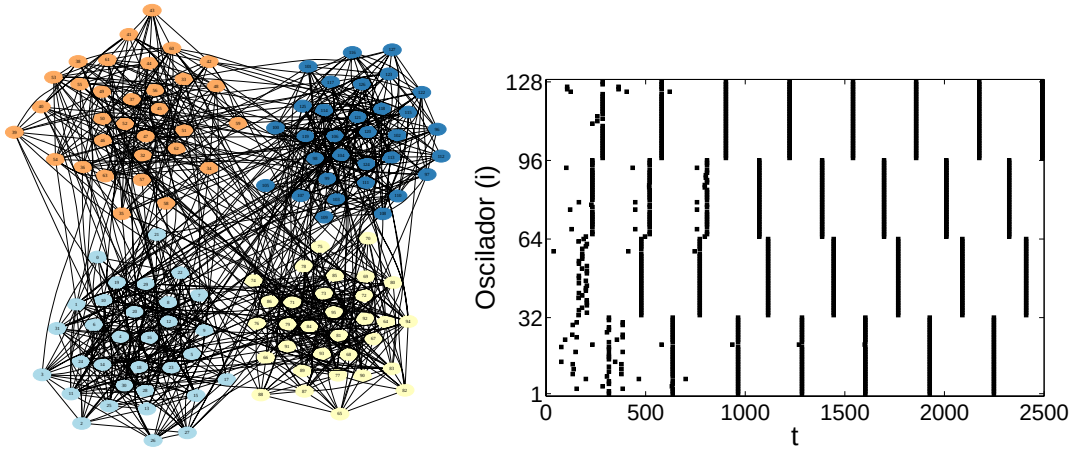

Figura 1. Ilustração do processo de detecção de comunidades usando o modelo proposto. Nesta simulação, $N = 128$**,** $M = 4$**,** $<k>= 16$**,** $z_{out}/ <k>= 0,2$ **e** $c = 0,1$**. (a) Rede randômica clusterizada. (b) Série temporal de disparo dos 128 neurônios.**

o termo de acoplamento excitatório e $Y_i(t)$ o termo de acoplamento inibitório entre os neurônios. O neurônio i dispara sempre que o potencial $v_i \geq \theta_v$, onde θ_v representa o limiar de disparo do neurônio.

O termo de acoplamento excitatório $E_i(t)$ é definido por:

$$E_i(t) = \sum_{j \in \Delta_i} \omega_{ij} \delta(t - t_j) \tag{2}$$

onde δ é a função delta de Dirac, t_j representa o instante em que o neurônio j dispara, Δ_i define a vizinhança de cooperação excitatória do neurônio i definida com base nas ligações (arestas) presentes na rede. ω_{ij} define a força de acoplamento excitatório entre os neurônios i e j e é definida por:

$$\omega_{ij} = \frac{c_E}{|\Delta_i|} \tag{3}$$

no qual $c_E \in [0, 1]$ é uma constante e $|\Delta_i|$ representa o grau do vértice i.

O termo de acoplamento inibitório é definido por:

$$Y_i(t) = \frac{c_Y}{N} \sum_{j=1; j \neq i}^{N} \delta(t - t_j) \tag{4}$$

onde $c_Y \in [0, 1]$ é uma constante que define a força de inibição e N representa o número de vértices na rede.

A dinâmica do modelo apresentado acima pode ser descrita da seguinte maneira. Devido as conexões excitatórias, modeladas pela Equação (2), grupos de neurônios (vértices) densamente conectados, representando comunidades, têm suas atividades de

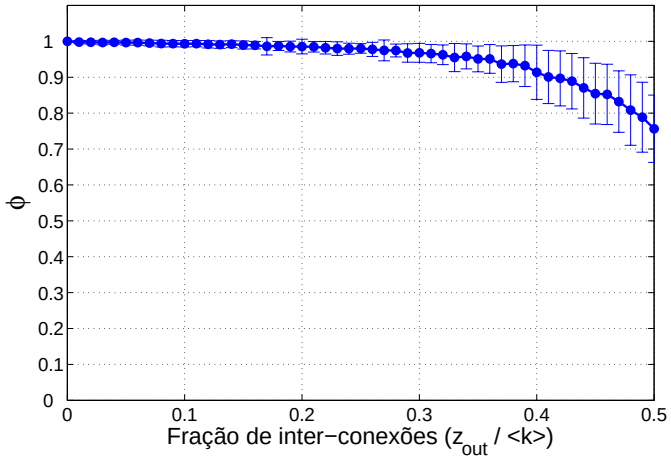

Figura 2. Precisão do processo de detecção de comunidades ϕ versus fração de inter-conexões $z_{out}/ <k>$. Nesta simulação, $N = 128$, $M = 4$, $<k>= 16$ e $c = 0,1$. Cada ponto da curva é obtido através da média de 200 execuções. A barra de erro representa o desvio padrão médio.

disparo sincronizadas. Por outro lado, devido a presença de um inibidor global (Equação (4)) aliada a menor probabilidade de conexões inter-comunidades, as atividades de disparos de comunidades distintas são dessincronizadas. Por esta razão, o modelo proposto é capaz de detectar comunidades em redes no qual cada comunidade tem sua atividade temporal de disparo (trem de pulso) segregada temporalmente das demais comunidades. Além disso, uma característica importante desta abordagem está na simplicidade do algoritmo e na rápida sincronização entre os grupos de neurônios resultando em uma grande eficiência computacional.

Com relação as constantes c_E e c_Y, quanto maior for a primeira, maior é a probabilidade de dois vértices vizinhos estarem agrupados em uma mesma comunidade (sincronizados). Por outro lado, conforme a constante c_Y é aumentada, maior é a segregação entre as comunidades, isto é, um maior número de comunidades menores é obtido.

A seguir, um conjunto de simulações computacionais utilizando redes sintéticas e reais é apresentado.

3. Experimentos Computacionais

Esta seção apresenta um conjunto de simulações com o objetivo de testar a capacidade do modelo como uma ferramenta computacional para detecção de comunidade. Em todas as simulações apresentadas a seguir, as constantes $c_E = c_Y = c$.

A Figura 1 apresenta uma ilustração do processo de detecção em uma rede randômica clusterizada composta por quatro comunidades. A partir da Figura 1(b) é possível observar o processo de sincronização entre os grupos de neurônios representando as comunidades. Uma vez estabelecida a sincronização entre estes, as comunidades podem ser facilmente identificadas através da atividade de disparo de cada grupo. Isto se deve ao mecanismo da correlação temporal no qual cada grupo é representado por uma trajetória temporal distinta.

A Figura 2 apresenta a fração dos vértices corretamente classificados como função

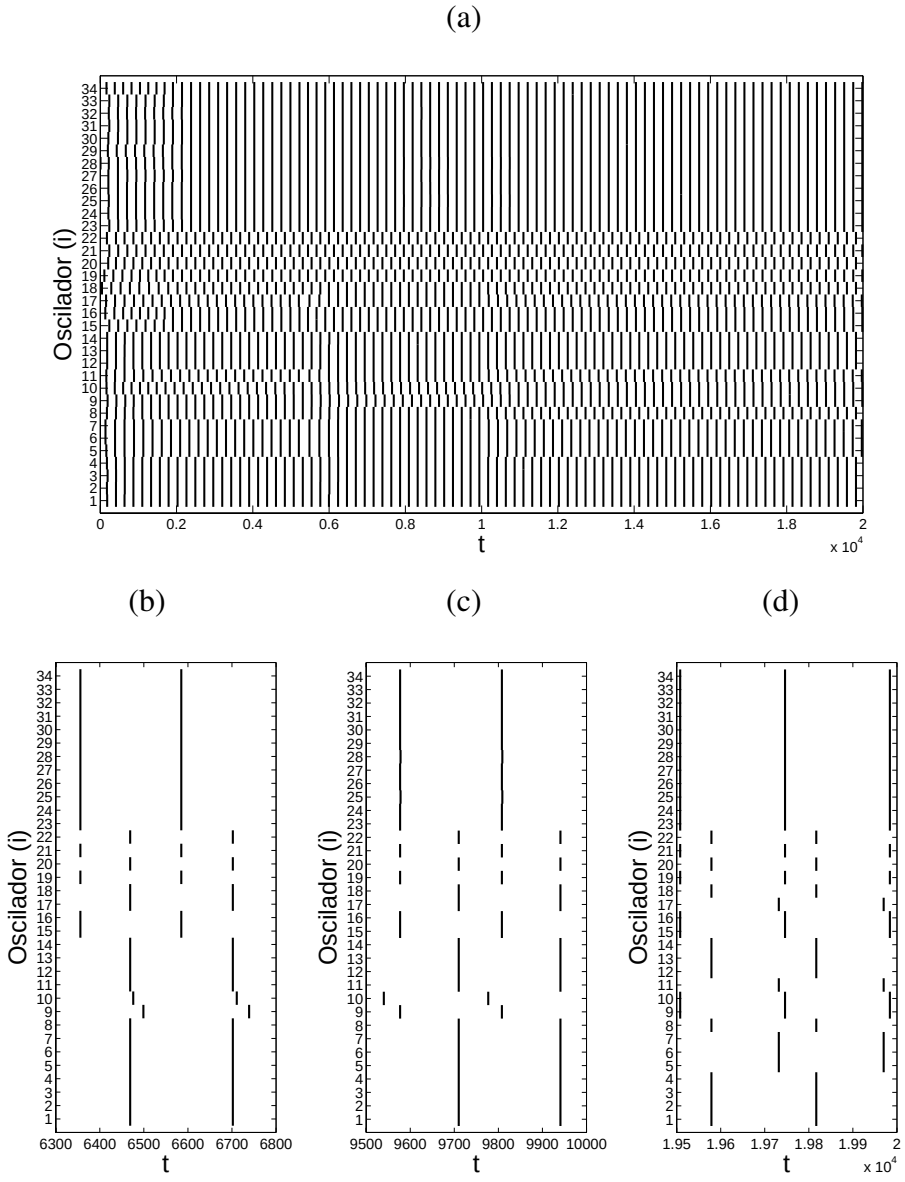

Figura 3. Séries temporais dos osciladores representando os vértices da rede de interação social entre indivíduos pertencentes a um clube de Karate [Zachary 1977]. $c = 0,1$**. (a) Série temporal completa. (b)-(d) Representação em maior resolução temporal para algumas faixas de** t**.**

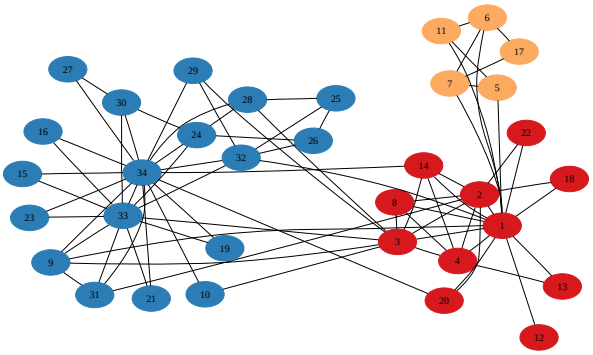

Figura 4. Resultado do processo de detecção de comunidades sobre a rede de interação social entre indivíduos pertencentes a um clube de Karate [Zachary 1977]. $c = 0,1$**.**

da proporção de inter-conexões $z_{out}/ <k>$. Estes resultados foram gerados pela média de 200 execuções do modelo em redes randômicas clusterizadas com $N = 128$, $M = 4$ e $<k>= 16$, geradas conforme descrito na Seção 1. A partir deste resultado pode ser constatado que o modelo apresenta bons resultados de detecção de comunidades para uma ampla faixa de $z_{out}/ <k>$, onde, na média, a precisão obtida para $z_{out}/ <k>= 0,4$ é de aproximadamente 90%. Quando comparado a outros modelos encontrados na litetura, como por exemplo o modelo (GN) proposto em [Girvan & Newman 2002], o modelo baseado em correlação oscilatória apresenta uma precisão superior. Por exemplo, para a rede randômica clusterizada descrita acima, quando $z_{out}/ <k>= 0,4$, o modelo GN apresenta uma precisão de aproximadamente 80% [Girvan & Newman 2002, Danon et al. 2005] contra aproximadamente 90% do modelo de correlação oscilatória. Resultados ainda superiores são obtidos quando $z_{out}/ <k>= 0,5$, neste caso, o modelo GN é capaz de obter uma precisão de aproximadamente 40% enquanto o nosso modelo apresenta uma precisão de $76 \pm 10\%$.

Além disso, ao comparar os resultados obtidos com aqueles apresentados em [Danon et al. 2005], pode-se constatar que o modelo aqui proposto se encontra entre aqueles que apresentam melhor precisão de detecção.

A seguir, duas simulações utilizando redes reais são apresentadas. Na Figura 3(a), as séries temporais de cada um dos osciladores representando os vértices da rede interação social entre indivíduos do clube de Karate [Zachary 1977] são apresentadas. Nesta figura pode ser observado que, após um certo número de ciclos, as comunidades são formadas. Para auxiliar a visualização do processo de detecção de comunidades, na Figura 3(b)-(d), as séries temporais para algumas faixas de t são apresentados em maior resolução temporal. Na Figura 3(b), pode ser observado que, com exceção dos osciladores número 9 e 10, o restante da rede se encontra divida em duas comunidades. Este resultado é coerente com aquele obtido em [Newman 2004a]. Nas Figuras 3(c) e (d) outros dois instantes da simulação são apresentados. No item (c), em especial, três comunidades são obtidas no qual os vértices numerados por 5, 6, 7, 11 e 17 são agrupados em uma terceira comunidade (vértices em cor laranja). A Figura 4 apresenta o resultado real desta divisão. Este resultado também foi observado no estudo apresentado em [Girvan & Newman 2002, Newman 2004a].

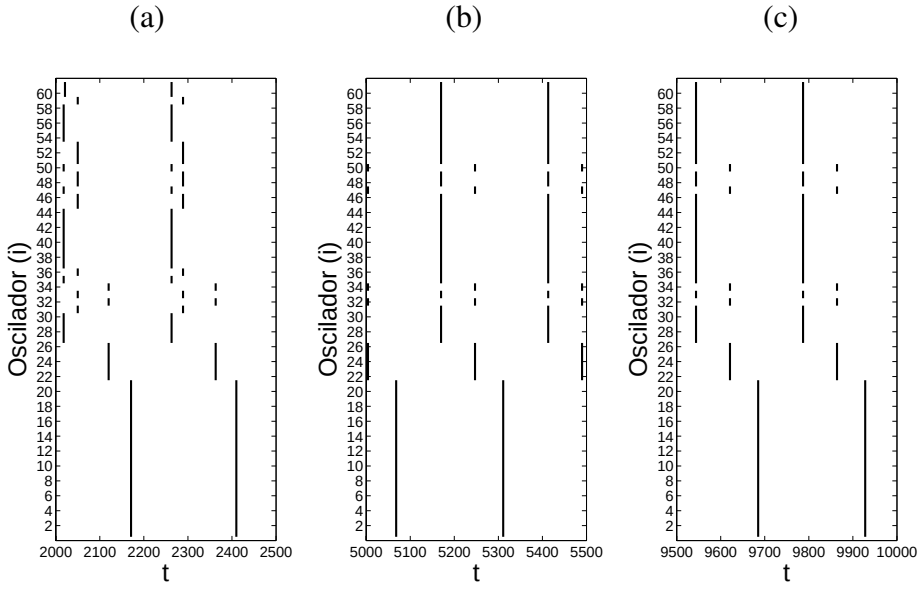

Figura 5. Séries temporais dos osciladores representando os vértices da rede de interação social social entre golfinhos ([Lusseau et al. 2003]). $c = 0,3$**. (a) Série temporal completa. (b)-(d) Representação em maior resolução temporal para algumas faixas de** t**.**

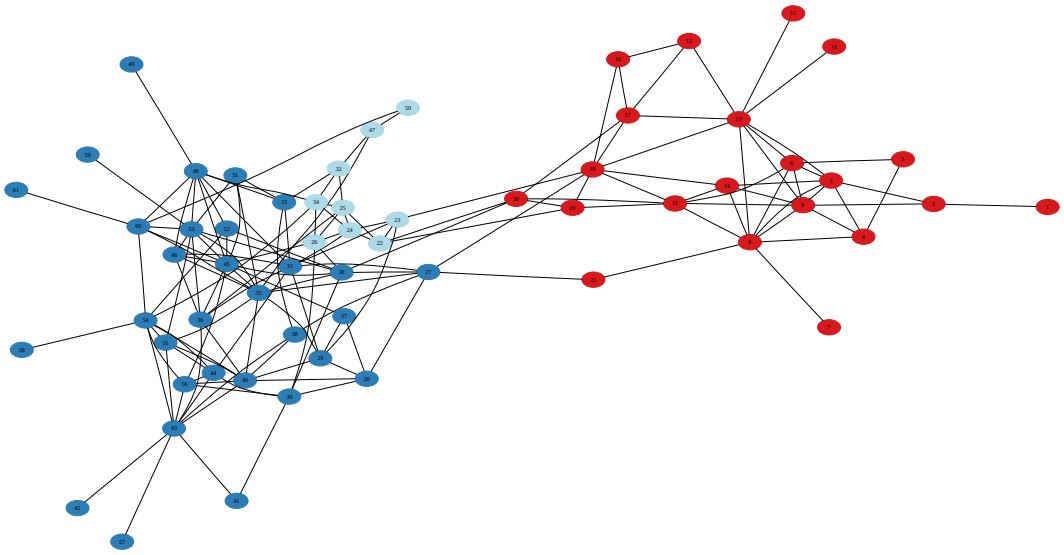

Figura 6. Resultado do processo de detecção comunidades sobre a rede interação social entre golfinhos ([Lusseau et al. 2003]). $c = 0,3$**.**

Nas Figuras 5 e 6, uma simulação utilizando a rede de interação social entre golfinhos [Lusseau et al. 2003] é apresentada. Na Figura 5, três instantes do tempo são apresentados com uma maior resolução temporal (Figura 5(a)-(c)) e o resultado final, obtido pelo modelo proposto, é apresentado na Figura 6. Nesta simulação também foi observado a detecção de uma terceira comunidade (vértices número 22, 23, 24, 25, 26, 32, 34, 47 e 50 - cor azul claro). Resultados análogos foram obtidos em estudos utilizando esta rede[Newman & Girvan 2004, Newman 2004a], o que demonstra a coerência dos resultados obtidos por nosso modelo.

4. Conclusões

Neste trabalho foi apresentado uma nova técnica para detecção de comunidades em redes baseada na teoria da correlação oscilatória. Dentre as principais características do modelo proposto as seguintes devem ser destacadas. Em primeiro lugar, a simplicidade do modelo no qual a detecção das comunidades na rede é obtida pela sincronização/dessincronização de neurônios simples e não através de um algoritmo complexo; Em segundo, a alta eficiência computacional devido a dinâmica simples de cada oscilador e a rápida sincronização dos grupos de osciladores; Em terceiro, a alta precisão na detecção das comunidades; Por fim, o modelo resultante apresenta poucos parâmetros. Desta forma, acredita-se que a extensão deste modelo como uma ferramenta para mineração de dados seja bastante promissora.

Além disso, como principal conclusão deste estudo, acredita-se que a abordagem unificando *dinâmica+estrutura*, no qual a dinâmica se refere aos fenômenos produzidos por elementos dinâmicos acoplados e a estrutura representando a organização desses elementos dinâmicos, apresentou-se como uma interessante ferramenta computacional alternativa a técnicas clássicas baseadas na teoria de grafos.

Como trabalhos futuros, pretende-se ampliar o domínio das simulações para outros modelos de redes, como redes livres de escala, redes com propriedade de mundo pequeno e redes com comunidades hierárquicas, e também realizar uma análise quantitativa dos parâmetros c_E e c_Y.

Agradecimentos

Este trabalho foi realizado com apoio financeiro da FAPESP e do CNPq.

Referências

Boccaletti, S., Ivanchenko, M., Latora, V., Pluchino, A., & Rapisarda, A. (2007). "Detecting complex network modularity by dynamical clustering". *Physical Review E*, 75:045102(1–4).

Clauset, A. (2005). "Finding local community structure in networks". *Physical Review E*, 72:026132(1–6).

Cook, D. J. & Holder, L. B. (2000). "Graph-based data mining". *IEEE Intelligent Systems*, 15:32–41.

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

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). "Self-organization and identification of web communities". *IEEE Computer*, 35(3):66–70.

Girvan, M. & Newman, M. E. J. (2002). "Community structure in social and biological networks". *Proceedings of the National Academy of Sciences of the USA*, 99(2):7821–7826.

Guimerà, R. & Amaral, L. A. N. (2005). "Functional cartography of complex metabolic networks". *Nature*, 433:895–900.

Guimerà, R., Mossa, S., Turtschi, A., & Amaral, L. A. N. (2003). "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles". *Proceedings of the National Academy of Sciences of the USA*, 102(22):7704–7709.

Izhikevich, E. M. (2004). "Which model to use for cortical spiking neurons?" *IEEE Transactions on Neural Networks*, 15(5):1063–1070.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A.-L. (2000). "The large scale organization of metabolic networks". *Nature*, 407:651–654.

Karypis, G., Han, E.-H., & Kumar, V. (1999). "Chameleon: hierarchical clustering using dynamic modeling". *IEEE Computer*, 32:68–75.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., & Dawson, S. M. (2003). "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. can geographic isolation explain this unique trait?" *Behavioral Ecology and Sociobiology*, 54:396–405.

Newman, M. E. J. (2004a). "Detecting community structure in networks". *The European Physical Journal B*, 38:321–330.

Newman, M. E. J. (2004b). "Fast algorithm for detecting community structure in networks". *Physical Review E*, 69:066133(1–5).

Newman, M. E. J. & Girvan, M. (2004). "Finding and evaluating community structure in networks". *Physical Review E*, 69:026113(1–15).

Quiles, M. G., Zhao, L., Alonso, R. L., & Romero, R. A. F. (2008). "Particle competition for complex network community detection". *Chaos (Woodbury)*, 18:033107(1–10).

Ravasz, E. & Barabasi, A.-L. (2003). "Hierarchical organization in complex networks". *Physical Review E*, 67:026112(1–7).

Reichardt, J. & Bornholdt, S. (2004). "Detecting fuzzy community structure in complex networks with a potts model". *Physical Review Letters*, 93:218701(1–4).

Schaeffer, S. E. (2007). "Graph clustering". *Computer Science Review*, 1:27–34.

von der Malsburg, C. (1981). The correlation theory of brain function. Technical report, Internal report 81-2: Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany.

Wang, D. L. (2005). "The time dimension for scene analysis". *IEEE Transactions on Neural Networks*, 16(6):1401–1426.

Zachary, W. W. (1977). "An information flow model for conflict and fission in small groups". *Journal of Anthropological Research*, 33:452–473.