Rotulação de indivíduos representativos no aprendizado semissupervisionado baseado em redes: caracterização, realce, ganho e filosofia

Bilzã Marques de Araújo

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Bilzã Marques de Araújo

Rotulação de indivíduos representativos no aprendizado semissupervisionado baseado em redes: caracterização, realce, ganho e filosofia

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Zhao Liang

USP – São Carlos
Junho de 2015

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

A658r

Araújo, Bilzã Marques de
Rotulação de indivíduos representativos no aprendizado semissupervisionado baseado em redes: caracterização, realce, ganho e filosofia / Bilzã Marques de Araújo; orientador Zhao Liang. -- São Carlos, 2015.
149 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

1. Aprendizado semissupervisionado. 2. Redes complexas. 3. Amostragem de dados. I. Liang, Zhao, orient. II. Título.

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Bilzã Marques de Araújo

Representative individuals labeling in the network-based semi-supervised learning: characterization, highlighting, gain and philosophy

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Zhao Liang

USP – São Carlos
June 2015

Para transformar uma realidade, não é necessária uma multidão,
mas umas poucas pessoas engajadas nesse propósito.

. . .

De Antioquia ao Chipre, e de ali à Galácia; da Frígia à Mísia, e então a Filipo, e por toda Macedônia e Acaia, incluindo as cidades de Tessalônica, Atenas e Corinto; de Jerusalém a Rodes, e, após o naufrágio, a Roma; preso enviou cartas confirmando às igrejas; cartas que chegaram até os confins da terra. Foi assim que Paulo, guiado pelo Espírito Santo, se tornou um dos principais elementos na propagação do evangelho de Cristo. Não obstante, o número daqueles que hoje se chamam cristãos supera 2 bilhões de pessoas. Paulo foi a pessoa certa para levar a mensagem do evangelho de Cristo aos lugares certos; E, desse modo, possibilitar a chegada do evangelho aos nossos dias.

In memoriam de meu pai,
Manoel de Araújo.

Posso imaginar o orgulho com que ele diria
‘‘este meu filho é doutor’’.

Agradecimentos

Geralmente termino agradecendo a Deus, mas dessa vez sinto que preciso começar de maneira diferente. Já passam das 5 da manhã, e, o que dizer, senão que o Senhor tem de fato me sustentado? Como há algum tempo acontece, me surpreende a organização e beleza com que ages em minha vida. Em muitos momentos não há e não existirão palavras que possam dizer o que eu sinto no fundo da minha alma. Não dá pra pensar em ti, senão com a certeza de que és um Deus pessoal, que falas conosco da única maneira que podemos de fato ouvir-te, e, se fingimos não te escutar, não te importas em falar novamente. Obrigado Senhor! Somente, obrigado!

Diversas pessoas colaboraram direta e indiretamente para a realização desse trabalho. Alguns o fizeram profissionalmente, outros por um sorriso e um papo ‘‘jogado fora’’, outros para ganhar a minha confiança e amizade, e outros, não sei precisar o porquê. Só pode ser amor!

Agradeço aos meus colegas e amigos Fabiano Berardo, Filipe Neto, Andrés Eduardo, Roberto Gueleri, Fernando Zuher, Murillo Carneiro e Newton Spolaôr pelos momentos além do cotidiano.

Aos professores Ricardo Campello, André Carvalho, Rosane Minghim e Graça Nunes pelas colaborações e pelo cuidado.

Especialmente, ao Prof. Zhao Liang, que, passados quase 7 anos, ainda me surpreende. Muito obrigado pela modo diferenciado com que tratas os seus alunos. Para mim é um privilégio ter tido a sua orientação durante todo esse tempo.

Agradeço também aos meus colegas do Biocom. Acho que poucas vezes pude ser parte da vida de vocês, além do nosso ambiente de trabalho, mas, quando pude, foi muito deleitoso.

Agradeço especialmente a Maria da Graça, minha mãe. Embora a maior parte desse tempo tenhamos estado distantes, você foi meu apoio em muitos momentos. Poder ouvir suas aflições sempre foi um meio de me sentir perto. Quem sabe um dia eu possa retribuir alguma coisa ou pelo menos te fazer orgulhosa.

Agradeço à minha irmã, Acássia. Não consigo imaginar as razões porque duas pessoas que se desentendem tanto, têm tanto prazer em ver, ouvir e pensar uma na outra.

Agradeço especialmente a você, minha esposa, Nelma Araújo. É uma das poucas pessoas que tenho certeza que lerá esses agradecimentos. Obrigado pela disposição em me ajudar. Podem não acreditar, minha esposa foi minha revisora. Obrigado por ser minha amiga em todas as horas. Obrigado por ser mãe de meus ‘‘filhotes’’. Obrigado por ser a rainha do nosso lar e a dona do

meu coração. Saiba que meus dias não seriam os mesmos, em São Carlos, ou em qualquer outro lugar, se você não estivesse ao meu lado. Superamos, e superaremos muitos desafios. E vamos fazer do nosso Deus orgulhoso de ter nos unido.

Agradeço aos amigos que fiz durante esse período e neste lugar, pessoas especiais que contribuíram para a formação desse que hoje eu sou. Juro que tentei fazer uma lista, mas me pareceu infindável, e, com toda certeza, eu acabaria sendo injusto com alguns. Portanto, obrigado aos amigos da IPB Ebenézer, do Futebol CAASO e dos caminhos sentido a Descalvado. Todos (e repito, todos) foram e são muito importantes na minha vida.

Agradeço formalmente à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro direto, por meio de bolsa, durante a realização da maior parte desse trabalho e à FAPESP e CNPq pelo apoio através do financiamento de projetos do nosso grupo de pesquisa que, direta ou indiretamente, me beneficiaram. Também agradeço à Pró-Reitoria de Pós-Graduação da USP pelo apoio financeiro durante a minha participação no IJCNN2013.

Por fim, agradeço a você que lê esse documento. Espero que ele seja influente em sua vida, de diversas maneiras. E, se for do seu interesse, que possamos realizar colaborações.

Resumo

Aprendizado semissupervisionado (ASS) é o nome dado ao paradigma de aprendizado de máquina que considera tanto dados rotulados como dados não rotulados. Embora seja considerado frequentemente como um meio termo entre os paradigmas supervisionado e não supervisionado, esse paradigma é geralmente aplicado a tarefas preditivas ou descritivas. Na tarefa preditiva de classificação, p. ex., o objetivo é rotular dados não rotulados de acordo com os rótulos dos dados rotulados. Nesse caso, enquanto que os dados não rotulados descrevem as distribuições dos dados e mediam a propagação dos rótulos, os itens de dados rotulados semeiam a propagação de rótulos e guiam-na à estabilidade. No entanto, dados são gerados tipicamente não rotulados e sua rotulação requer o envolvimento de especialistas no domínio, rotulando-os manualmente. Dificuldades na visualização de grandes volumes de dados, bem como o custo associado ao envolvimento do especialista, são desafios que podem restringir o desempenho dessa tarefa. Portanto, o destacamento automático de bons candidatos a dados rotulados, doravante denominados indivíduos representativos, é uma tarefa de grande importância, e pode proporcionar uma boa relação entre o custo com especialista e o desempenho do aprendizado. Dentre as abordagens de ASS discriminadas na literatura, nosso interesse de estudo se concentra na abordagem baseada em redes, onde conjuntos de dados são representados relacionalmente, através da abstração gráfica. Logo, o presente trabalho tem como objetivo explorar a influência dos nós rotulados no desempenho do ASS baseado em redes, i.e., estudar a caracterização de nós representativos, como a estrutura da rede pode realçá-los, o ganho de desempenho de ASS proporcionado pela rotulação manual dos mesmos, e aspectos filosóficos relacionados. Em relação à caracterização, critérios de caracterização de nós centrais em redes são estudados considerando-se redes com estruturas modulares bem definidas. Contraintuitivamente, nós bastantes conectados (*hubs*) não são muito representativos. Nós razoavelmente conectados em vizinhanças pouco conectadas, por outro lado, são; estritamente local, esse critério de caracterização é escalável a grandes volumes de dados. Em redes com distribuição de grau homogênea - modelo Girvan-Newman (GN), nós com alto coeficiente de agrupamento também mostram-se representativos. Por outro lado, em redes com distribuição de grau heterogênea - modelo Lancichinetti-Fortunato-Radicchi (LFR), nós com alta intermedialidade se destacam. Nós com alto coeficiente de agrupamento em redes GN estão tipicamente situados em *motifs* do tipo quase-clique; nós com alta intermedialidade em

redes LFR são *hubs* situados na borda das comunidades. Em ambos os casos, os nós destacados são excelentes regularizadores. Além disso, como critérios diversos se destacam em redes com características diversas, abordagens unificadas para a caracterização de nós representativos também foram estudadas. Crítica para o realce de indivíduos representativos e o bom desempenho da classificação semissupervisionada, a construção de redes a partir de bases de dados vetoriais também foi estudada. O método denominado AdaRadius foi proposto, e apresenta vantagens tais como adaptabilidade em bases de dados com densidade variada, baixa dependência da configuração de seus parâmetros, e custo computacional razoável, tanto sobre dados *pool-based* como incrementais. As redes resultantes, por sua vez, são esparsas, porém conectadas, e permitem que a classificação semissupervisionada se favoreça da rotulação prévia de indivíduos representativos. Por fim, também foi estudada a validação de métodos de construção de redes para o ASS, sendo proposta a medida denominada coerência grafo-rótulos de Katz. Em suma, os resultados discutidos apontam para a validade da seleção de indivíduos representativos para semear a classificação semissupervisionada, corroborando a hipótese central da presente tese. Analogias são encontrados em diversos problemas modelados em redes, tais como epidemiologia, propagação de rumores e informações, resiliência, letalidade, *grandmother cells*, e crescimento e auto-organização.

Palavras-chave: aprendizado semissupervisionado; redes complexas; amostragem de dados.

Abstract

Semi-supervised learning (SSL) is the name given to the machine learning paradigm that considers both labeled and unlabeled data. Although often defined as a mid-term between unsupervised and supervised machine learning, this paradigm is usually applied to predictive or descriptive tasks. In the classification task, for example, the goal is to label the unlabeled data according to the labels of the labeled data. In this case, while the unlabeled data describes the data distributions and mediate the label propagation, the labeled data seeds the label propagation and guide it to the stability. However, as a whole, data is generated unlabeled, and to label data requires the involvement of domain specialists, labeling it by hand. Difficulties on visualizing huge amounts of data, as well as the cost of the specialists involvement, are challenges which may constraint the labeling task performance. Therefore, the automatic highlighting of good candidates to label by hand, henceforth called representative individuals, is a high value task, which may result in a good tradeoff between the cost with the specialist and the machine learning performance. Among the SSL approaches in the literature, our study is focused on the network--based approache, where datasets are represented relationally, through the graphic abstraction. Thus, the current study aims to explore and exploit the influence of the labeled data on the SSL performance, that is, the proper characterization of representative nodes, how the network structure may enhance them, the SSL performance gain due to labeling them by hand, and related philosophical aspects. Concerning the characterization, central nodes characterization criteria were studied on networks with well-defined modular structures. Counterintuitively, highly connected nodes (*hubs*) are not much representatives. Not so connected nodes placed in low connectivity neighborhoods are, though. Strictly local, this characterization is scalable to huge volumes of data. In networks with homogeneous degree distribution - Girvan-Newman networks (GN), nodes with high clustering coefficient also figure out as representatives. On the other hand, in networks with inhomogeneous degree distribution - Lancichinetti-Fortunato-Radicchi networks (LFR), nodes with high betweenness stand out. Nodes with high clustering coefficient in GN networks typically lie in almost-cliques motifs; nodes with high betweenness in LFR networks are highly connected nodes, which lie in communities borders. In both cases, the highlighted nodes are outstanding regularizers. Besides that, unified approaches to characterize representative nodes were studied because diverse criteria stand out for diverse networks. Crucial

for highlighting representative nodes and ensure good SSL performance, the graph construction from vector-based datasets was also studied. The method called AdaRadius was introduced and presents advantages such as adaptability to data with variable density, low dependency on parameters settings, and reasonable computational cost on both pool based and incremental data. Yielding networks are sparse but connected and allow the semi-supervised classification to take great advantage of the manual labeling of representative nodes. Lastly, the validation of graph construction methods for SSL was studied, being proposed the validation measure called graph-labels Katz coherence. Summing up, the discussed results give rise to the validity of representative individuals selection to seed the semi-supervised classification, supporting the central assumption of current thesis. Analogies may be found in several real-world network problems, such as epidemiology, rumors and information spreading, resilience, lethality, grandmother cells, and network evolving and self-organization.

Keywords: semi-supervised learning; complex networks; data sampling.

Sumário

Sumário xi
Lista de Figuras xv
Lista de Tabelas xxi
Lista de Algoritmos xxiii

1 Introdução **1**
1.1 Motivação 4
1.2 Objetivo 6
1.3 Trabalhos relacionados 6
1.4 Organização do documento 8

I Fundamentos e Conceitos **9**

2 Aprendizado Semissupervisionado **11**
2.1 Suposição de suavidade, agrupamento e variedade 14
2.2 Significância do aprendizado semissupervisionado 15
2.3 Principais abordagens de aprendizado semissupervisionado 17
2.4 Aprendizado semissupervisionado baseado em redes 18
2.4.1 Fluxo máximo corte mínimo 19
2.4.2 Campos Gaussianos e funções harmônicas 19
2.4.3 Consistência local e global 20
2.4.4 Framework de regularização 21
2.4.5 Competição e cooperação de partículas 21
2.4.6 Construção do grafo 24
2.4.7 Desafios 26
2.5 Validação e referência de desempenho de ASS 28
2.5.1 Índice de Rand ajustado 29
2.5.2 Silhueta 30

3 Amostragem Aleatória **31**
3.1 Limite de Chernoff 31

4 Redes Complexas **35**
4.1 Definições gerais e notações 36
4.2 Caracterização de indivíduos 38
4.2.1 Medidas relacionadas ao diâmetro 38
4.2.2 Medidas relacionadas ao grau 40
4.2.3 Medidas relacionadas ao fluxo 42
4.2.4 Medidas hierárquicas 44
4.3 Propriedades e modelos de redes 44
4.3.1 Redes aleatórias 44
4.3.2 Transitividade e redes de pequeno mundo 45
4.3.3 Distribuição de grau e redes livres de escala 48
4.3.4 Assortatividade e estruturas de comunidades 49
4.4 Resiliência e epidemiologia 54
4.5 Considerações finais 56

II Resultados e discussões **59**

5 Robustez do aprendizado semissupervisionado **63**
5.1 Dados, métodos e simulações 63
5.2 Comparação entre modelos de construção de rede 66
5.3 Comparação entre métodos de propagação de rótulo 67
5.4 Restrições de ligações de acordo com conhecimentos prévios 70
5.5 Considerações finais 72

6 Caracterização de indivíduos representativos em redes **73**
6.1 Variabilidade de desempenho de aprendizado 74
6.2 Caracterização de representativos 77
6.3 Métodos e análise numérica 78
6.4 Resultados e discussões 79
6.5 Composição de medidas 81
6.6 Considerações finais 84

7 Construção de grafos adaptativos **89**
7.1 Método proposto 90
7.1.1 Custo computacional 90
7.2 Benchmarks, métodos e análise numérica 91
7.3 Validade do método proposto 94
7.4 Resultados e discussões sobre representativos 101

7.4.1 Resultados sobre conjuntos de dados em forma de bananas 101
7.4.2 Resultados sobre conjuntos de dados de Chapelle et al. (2006) 107
7.4.3 Comparação com o *benchmark* de Chapelle et al. (2006) 115

8 Validação de métodos de construção de grafos 117
8.1 Erro de reconstrução 118
8.2 Medida proposta 119
8.3 Problemas *toy* 121
8.4 Resultados e discussões 122
8.5 Considerações finais 126

9 Conclusão 129
9.1 Principais contribuições 131
9.2 Trabalhos Futuros 132

Referências 149

Lista de Figuras

1.1 Rede dos clubes de caratê: em (a) colorida de acordo com o rótulo das academias e em (b) de acordo com a representatividade dos indivíduos. Quanto mais clara a cor do nó em (b), maior a representatividade do indivíduo de acordo com um dos critérios propostos no presente trabalho. 5

2.1 Tarefa de classificação: (a) supervisionada, (b) semissupervisionada indutiva, e (c) semissupervisionada transdutiva. As diferenças entre esses paradigmas são a forma como os dados são tratados com fins à modelagem de um classificador e o domínio do modelo. Enquanto que o supervisionado considera apenas dados rotulados, o semissupervisionado considera tanto dados rotulados quanto dados não rotulados. Com relação ao domínio do modelo, por outro lado, o aprendizado transdutivo infere hipóteses que limitam-se ao conjunto de dados disponível durante o treinamento, enquanto que o aprendizado indutivo infere hipóteses que atuam sobre todo o hiperespaço dos dados. 13

2.2 Ganho com a classificação semissupervisionada. Os círculos escuros denotam dados rotulados, enquanto que círculos claros, dados não rotulados. Enquanto que aplicando-se um algoritmo supervisionado, a fronteira de decisão seria estabelecida próximo à linha pontilhada, aplicando-se um algoritmo semissupervisionado, a fronteira de decisão certamente seria estabelecida próximo à linha contínua. Os dados não rotulados permitem construir superfícies de decisão melhor ajustadas à distribuição dos dados. Figura reproduzida de (Zhu et al., 2007), com permissão. Copyright © 2007, Association for the Advancement of Artificial Intelligence. 15

2.3 Informações intrínsecas acrescentadas por dados não rotulado e pelo conhecimento prévio. Em (a) dois objetos de classes distintas estão devidamente rotulados. A superfície de separação induzida por um humano é ilustrada pela cor de fundo, uma separação linear. Em (b) são considerados, adicionalmente, os dados não rotulados. A superfície de separação induzida é modificada. Figura adaptada de (Belkin et al., 2006) com a permissão dos autores. 16

2.4 Exemplos de padrões tridimensionais considerados por Zhu et al. (2007). A forma geométrica varia de acordo com o parâmetro x, apresentado abaixo de cada forma. No experimento alguma formas foram apresentadas rotuladas; a maioria, no entanto, sem rótulos. Figura reproduzida de (Zhu et al., 2007), com permissão. Copyright © 2007, Association for the Advancement of Artificial Intelligence. 17

2.5 Conjunto de dados em forma de bananas com variedades sobrepostas: (a) rede gerada a partir da similaridade entre os objetos; (b) classificação semissupervisionada. Nesse caso, a abordagem baseada em redes mostra-se incapaz de separar adequadamente as classes. Figura reproduzida de (Zhu & Goldberg, 2009) com permissão da Morgan & Claypool Publishers. 27

4.1 Configurações de redes randômicas com 8 vértices: (a) $p = 0$; (b) $p = 0.1$; (c) $p = 0.3$; (d) $p = 0.8$. 45

4.2 Distribuição de grau média de 10 redes randômicas com 1000 nós e probabilidade $p = 0.1$. Uma distribuição de Poisson. 46

4.3 Modelo de redes de pequeno mundo de Watts & Strogatz (1998). Para $p = 0$ tem-se uma rede regular. Quanto maior a probabilidade p, mais o modelo se aproxima de uma rede randômica. Para $0.01 < p < 0.1$ a rede se comporta como uma rede de pequeno mundo. Observe que a distância entre os nós identificados em vermelho (cinza escuro) e em azul (cinza claro) diminui significativamente, enquanto que o coeficiente de agrupamento mantém-se alto. Figura adaptada com permissão da Macmillan Publishers Ltd:Nature (Watts & Strogatz, 1998), copyright ©1998. 47

4.4 Comportamento da distância média, $\langle d \rangle$, e do coeficiente de agrupamento, C, em função de p, segundo o modelo de Watts & Strogatz (1998). Tanto o coeficiente de agrupamento como a distância média são apresentados em razão dos respectivos valores máximos. Para $0.01 < p < 0.1$ tem-se o efeito de pequeno mundo. Figura reproduzida com permissão de (Newman, 2003b). Copyright ©2003 Society for Industrial and Applied Mathematics. Todos os direitos reservados. 48

4.5 Distribuição de grau média de 10 redes geradas segundo o modelo de Barabási & Albert (1999) com 1000 nós e $m_0 = 5$. 50

4.6 Redes com comunidades bem definidas geradas pelos modelos de redes de Girvan-Newman com 1024 nós, $\langle k \rangle = 16$ e 4 comunidades bem definidas ($Z_{\text{out}} = 2$) de igual tamanho. 53

4.7 Redes com comunidades bem definidas geradas pelos modelos de redes de Lancichinette-Fortunato-Radichi com 1024 nós, $\langle k \rangle = 16$, distribuição de grau heterogênea ($\gamma = 2$), comunidade com tamanhos variados porém homogêneos ($\beta = 1$) e comunidades bem definidas ($\mu = 0.2$). 54

4.8 Resiliência da Internet. Quando os nós são removidos aleatoriamente (quadrados) a distância média da rede, $\langle d \rangle$, permanece estável. Entretanto, quando a remoção ocorre preterindo os nós mais conectivos (circunferências), a depreciação da rede como um todo é eminente. Figura adaptada com permissão da Macmillan Publishers Ltd:Nature (Albert et al., 2000), copyright 2000. 55

5.1 Exemplos de conjuntos de dados em forma de bananas com configurações diversas, gerados pela ferramenta MATLAB PrTools Toolbox (Duin et al., 2007). 65

5.2 ARI médio e desvio padrão obtido pelos modelos MinCut, GFHF, LGC, e PCC combinados aos métodos de construção de rede Co, kNN, wkNN, ϵRd e wϵRd sobre dados em forma de bananas com $n = 1000$, $L = 20$ e $S = [0.6 : 0.2 : 1.6]$. 66

5.3 Resultado médio e desvio padrão obtidos pelos modelos wkNN-MinCut, wkNN-GFHF, wϵRd-LGC e wϵRd-PCC na Simulação I. 68

5.4 Resultado médio e desvio padrão obtidos pelos modelos wkNN-MinCut, wkNN-GFHF, wϵRd-LGC e wϵRd-PCC nas Simulações II e III. 69

5.5 Resultado médio e desvio padrão obtidos pelos modelos wkNN-MinCut, wkNN-GFHF, wϵRd-LGC e wϵRd-PCC na Simulação IV. 70

5.6 Resultados obtidos pelos modelos kNNMCL-PCC, ϵRdMCL-PCC e wϵRd-PCC nas simulações I, III e IV. Em (a), (d) e (g) para $n = 1000$, $L = 20$ e $S = [0.6 : 0.2 : 1.6]$; em (b), (e) e (h) para $n = \{100, 200, 400, 600, 1000, 2000\}$, $L = 0.02 \times n$ e $S = 1.6$; e em (c), (f) e (i) para $n = 1000$, $L = \{0.01, 0.02, 0.04, 0.06, 0.1, 0.2\} \times n$ e $S = 1.6$. São apresentados média e desvio padrão de 80 ensaios - 20 ensaios com divisões aleatórias entre rotulados e não rotulados para cada uma de 4 variações do conjunto de dados. 71

6.1 Desempenho dos métodos GFHF (em (a)) e LGC (em (b)) em relação à amostra disponível rotulada e, no caso do LGC, ao ajuste de α. São apresentados ARI médio, desvio padrão sobre a (e abaixo da) média e pior e melhor desempenho observados em 900 ensaios (símbolo '+'). 75

6.2 Desempenho do método PCC em relação à amostra disponível rotulada e o ajuste dos parâmetros p_{grd} e Δ_v. Em (a) é apresentado o ARI médio; em (b) o número de iterações até a convergência do algoritmo; e em (c) e (d) secções transversais da superfície de desempenho para $p_{\text{grd}} = 0.75$ e $\Delta_v = 0.2$, respectivamente. Desempenhos médios consistem na média de 900 ensaios. Em (c) e (d) são apresentados adicionalmente desvio padrão sobre a (e abaixo da) média e pior ($\triangle$) e melhor ($\triangledown$) desempenho para cada configuração. 76

6.3 Rede de associações entre golfinhos com os nós caracterizados por medidas de centralidade em rede (Lusseau et al., 2003). Quanto mais clara a cor do nó, maior o índice atribuído pela medida. O espalhamento visual da rede é realizado através do método de minimização de energia de um sistema de molas entre todos pares de vértices, cuja rigidez da mola entre cada par de vértices é proporcional à distância mínima entre os mesmos (Kamada & Kawai, 1989). . . 77

6.4 Caracterização de nós em redes de acordo com medidas de centralidade. Em (a) temos a rede de associações entre golfinhos (Lusseau et al., 2003); em (b) uma rede de Girvan-Newman com 1024 nós, 4 comunidades, grau médio $\langle k \rangle = 16$ e $Z_{\text{out}} = 2$; em (c) um rede LFR com distribuição de grau heterogênea ($\gamma = 2$), grau médio $\langle k \rangle = 16$ e comunidades de tamanhos diversos, porém não muito distintos. Em cada *boxplot*, as extremidades do retângulo informam os 25º e 75º percentis, Q_1 e Q_3; a mediana é apresentada dentro do retângulo; as suíças são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$; e valores além disso são representados como *outliers* (símbolo '+'). Pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. Os índices são apresentados reescalados no intervalo $[0, 1]$. 82

6.5 Caracterização de nós em redes de acordo com medidas de centralidade simples e agregadas (à esquerda) e correlação das primeiras com as segundas (à direita) para redes Girvan-Newman (acima) e LFR (abaixo). A rede de Girvan-Newman possui 1024 nós, 4 comunidades, grau médio $\langle k \rangle = 16$ e $Z_{\text{out}} = 2$; a LFR, distribuição de grau heterogênea ($\gamma = 2$), grau médio $\langle k \rangle = 16$ e comunidades de tamanhos diversos, porém não muito distintos. Em cada *boxplot*, as extremidades do retângulo informam os 25º e 75º percentis, Q_1 e Q_3; a mediana é apresentada dentro do retângulo; as suíças são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$; e valores além disso são representados como *outliers* (símbolo '+'). Pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. Os índices são apresentados reescalados no intervalo $[0, 1]$. O mapa de cores da matriz de correlação varia em $[-1, 1]$. Quanto maior a correlação, mais claro. 85

6.6 Desempenho dos métodos GFHF (em (a)), LGC (em (b)) e PCC (em (c)) sobre redes GN. São apresentados ARI médio, desvio padrão sobre a (e abaixo da) média e pior e melhor desempenho (símbolo '+') observados em 900 ensaios com sementes escolhidas ao acaso, e ARI médio para critérios propostos de seleção de representativos. 86

6.7 Desempenho dos métodos GFHF (em (a)), LGC (em (b)) e PCC (em (c)) sobre redes LFR. São apresentados ARI médio, desvio padrão sobre a (e abaixo da) média e pior e melhor desempenho (símbolo '+') observados em 900 ensaios com sementes escolhidas ao acaso, e ARI médio para critérios propostos de seleção de representativos. 87

7.1 Grafos construídos pelos métodos kNN ((a) e (c)) e ϵRadius ((b) e (d)) sobre conjuntos de dados em forma de bananas com $n = 256$ e $S = \{0.6, 1.4\}$. k e ϵ são ajustados no sentido de maximizar o índice de modularidade do grafo resultante, sem fragmentá-lo muito. 95

7.2 Índice de modularidade de grafos kNN (a), ϵRadius (b) e AdaRadius (c) sobre conjuntos de dados em forma de bananas com $n = 256$ e $S = 1.4$ variando-se os parâmetros de densidade. Para o ensaio cuja curva é apresentada em preto, o grafo resultante é desconexo para configurações de parâmetros menores que o valor do limiar discriminado no eixo vertical. 96

7.3 Resultado do método AdaRadius sobre conjuntos de dados em forma de bananas com $n = 256$ e $S = 1.4$. Os *motifs* destacados são ampliados e descritos na Figura 7.4. 97

7.4 *Motifs* construídos por AdaRadius: (a) nó *outlier*, (b) nós de borda, e (c) nó *hub* destacados na Figura 7.3. Conectado a j (um *outlier*), as conexões de i à variedade mais próxima são reforçadas em (a). Situados nas bordas de seus grupos, as conexões de i e j aos seus grupos são reforçadas em (b); Já um *hub* na AGM, devido à sua localização privilegiada, as conexões de i são incrementadas em (c). O valor de ϵ_i é ilustrado através das esferas de cobertura, em cinza. . . . 98

7.5 Distribuições de medidas de centralidade ((a) e (c)) e coeficientes de componentes principais sobre composições de medições de centralidade ((b) e (d)) para redes 5NN construídas sobre dados em forma de bananas com $n = 512$ e $S = 0.6$ (em cima) ou $S = 1.4$ (em baixo). As suíças dos *boxplots* são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$. Valores além disso são representados como *outliers* (símbolo '+'). Padronizados, pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. O mapa de cores da matriz de coeficientes varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1. Notavelmente, a composição dos 4 primeiros componentes para $S = 0.6$ e $S = 1.4$ são muito similares. 106

7.6 Distribuições de medidas de centralidade ((a) e (c)) e coeficientes de componentes principais sobre composições de medições de centralidade ((b) e (d)) para redes AdaRadius construídas sobre dados em forma de bananas com $n = 512$ e $S = 0.6$ (em cima) ou $S = 1.4$ (em baixo). As suíças dos *boxplots* são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$. Valores além disso são representados como *outliers* (símbolo '+'). Padronizados, pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. O mapa de cores da matriz de coeficientes varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1. Notavelmente, a composição dos três primeiros componentes e de Z_5 para $S = 0.6$ e $S = 1.4$ são muito similares. 107

7.7 Coeficientes de componentes principais sobre composições de medições de centralidade em redes kNN construídas sobre os conjuntos de dados de Chapelle et al. (2006). O mapa de cores varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1. 113
7.8 Coeficientes de componentes principais sobre composições de medições de centralidade em redes AdaRadius construídas sobre os conjuntos de dados de Chapelle et al. (2006). O mapa de cores varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1. . . . 114

8.1 Aferições da modularidade e coerência de Katz, variando-se α, sobre redes GN variando-se C e Z_{out}. 122
8.2 Aferições da modularidade e coerência de Katz, variando-se α, sobre redes LFR variando-se γ, β e μ. 124
8.3 (Originalmente colorido) Redes de Mark Newman descritas na Tabela 8.4. Os rótulos *ground truth* das comunidades são discriminados pelas cores e formas dos vértices. O espalhamento visual da rede é realizado através do método de minimização de energia de um sistema de molas entre todos pares de vértices, cuja rigidez da mola entre cada par de vértices é proporcional à distância mínima entre os mesmos (Kamada & Kawai, 1989). 125

Lista de Tabelas

2.1 Descrição resumida dos conjuntos de dados utilizados no *benchmark* de Chapelle et al. (2006) 28

2.2 Descrição resumida dos métodos de ASS utilizados avaliados no *benchmark* de Chapelle et al. (2006) 28

6.1 ARI médio e ranque médio obtidos por métodos de propagação de rótulos semeados por nós realçados por critérios diversos de seleção de representativos. 80

6.2 ARI médio e ranque médio obtidos por métodos de propagação de rótulos semeados por nós realçados pelos melhores critérios de seleção de representativos. 83

7.1 Modularidade ótima obtida por AdaRadius e métodos clássicos sobre os conjuntos de dados de Chapelle et al. (2006). 97

7.2 Configurações de parâmetros para soluções ótimas em termos da modularidade. 98

7.3 Escore de densidade (m/n) para soluções ótimas em termos da modularidade. . 98

7.4 ARI obtido por métodos de rotulação semissupervisionados sobre redes 5NN e AdaRadius construídas sobre dados em forma de bananas 99

7.5 Erro de teste obtido por métodos de rotulação semissupervisionados sobre redes 5NN e AdaRadius construídas sobre os conjuntos de dados de Chapelle et al. (2006) 100

7.6 ARI obtido por GFHF semeado por representativos sobre redes 5NN construídas sobre dados em forma de bananas. 102

7.7 ARI obtido por LGC semeado por representativos sobre redes 5NN construídas sobre dados em forma de bananas. 103

7.8 ARI obtido por GFHF semeado por representativos sobre redes AdaRadius construídas sobre dados em forma de bananas. 104

7.9 ARI obtido por LGC semeado por representativos sobre redes AdaRadius construídas sobre dados em forma de bananas. 105

7.10 ARI obtido por GFHF semeado por representativos sobre redes 5NN construídas sobre conjuntos de dados de Chapelle et al. (2006) 109

7.11 Erro de teste obtido por LGC semeado por representativos sobre redes 5NN construídas sobre conjuntos de dados de Chapelle et al. (2006) 110
7.12 Erro de teste obtido por GFHF semeado por representativos sobre redes AdaRadius construídas sobre conjuntos de dados de Chapelle et al. (2006) 111
7.13 Erro de teste obtido por LGC semeado por representativos sobre redes AdaRadius construídas sobre conjuntos de dados de Chapelle et al. (2006) 112
7.14 Erro de teste e ranque médio obtidos por modelos de ASS semeados por representativos e *benchmark* de Chapelle et al. (2006) 116

8.1 Aferições da modularidade e coerência de Katz, variando-se α, sobre problemas *toy* assortativos. 121
8.2 Aferições da modularidade e coerência de Katz, variando-se α, sobre problemas *toy* dissortativos. 121
8.3 Desvios padrões do resultados do ensaios apresentados na Figura 8.2 normalizados pela diferença entre os valores medidos para μ=0.1 e μ=0.6. 126
8.4 Breve descrição de redes reais com comunidades *ground truth*. 127
8.5 Aferições da modularidade e coerência de Katz, variando-se α, sobre redes reais descritas na Tabela 8.4. 127

Lista de Algoritmos

2.1 Auto-Treinamento 18
2.2 Corte mínimo (Blum & Chawla, 2001) 19
2.3 Campos Gaussianos e funções harmônicas (Zhu & Ghahramani, 2002; Zhu et al., 2003) 20
2.4 Consistência local e global (Zhou et al., 2004) 20
2.5 Competição e cooperação de partículas (Breve et al., 2012; Breve, 2010) 24

Capítulo

1

Introdução

No mundo moderno, a todo instante grandes volumes de dados digitais são gerados. Sistemas de vigilância e monitoramento registram horas e horas de fluxos contínuos de dados, sistemas computacionais diversos registram transações em diversos domínios, microarranjos de DNA são utilizados no registro de expressões gênicas, conteúdos diversos são publicados na web, etc. Além das informações manipuladas diretamente, informações intrínsecas, frequentemente muito valiosas, também são produzidas, p. ex., relações de associações, grupos de dados, margens de separação, conhecimentos coletivos, novidades, entre outros. A cada dia que passa é mais difícil a atuação eficiente de especialistas humanos analisando esses dados, de modo que torna-se necessária a aplicação de sistemas automatizados para o processamento dos dados, extração de informações e produção de conhecimentos. Esta é uma das principais motivações para o aprendizado de máquina.

O **aprendizado de máquina** é a área da ciência da computação que tem por objetivo estudar e desenvolver métodos capazes de ''aprender'' através da experiência, isto é, através da interação com dados. Tem-se por objetivo gerar modelos computacionais capazes de organizar o conhecimento inerente aos dados, replicando de forma automática e eficiente o processo de análise e organização que seria realizado por um especialista humano no domínio. O processo de aprendizado induz hipóteses no sentido de aproximar funções equivalentes às informações intrínsecas de interesse (Mitchell, 1997; Bishop, 2006; Faceli et al., 2011).

Tradicionalmente, o processo de aprendizado é classificado de acordo com o grau de envolvimento de um supervisor externo como supervisionado ou não supervisionado. No **aprendizado supervisionado**, o objetivo é induzir hipóteses a partir de dados rotulados por um especialista. O processo de aprendizado irá construir uma função de mapeamento entrada/saída, capaz de predizer rótulos (classes ou valores) para exemplos não rotulados, através de um viés indutivo. Este tipo de aprendizado também é conhecido como **aprendizado preditivo**. Quando o rótulo é composto por valores discretos, o problema é denominado **classificação** e a função de ma-

peamento é denominada classificador; quando os rótulos são valores contínuos, o problema é denominado **regressão**, e a função de mapeamento, regressor (Mitchell, 1997; Bishop, 2006; Faceli et al., 2011).

Para que se possa induzir um bom classificador, no entanto, tipicamente é necessário uma grande quantidade de dados rotulados. A qualidade da função de mapeamento depende do quão bem os dados rotulados representam as classes. Induzir um classificador utilizando apenas poucos exemplos rotulados frequentemente não resulta em boa qualidade de classificação. Contudo, a rotulação manual de dados através de um especialista é frequentemente onerosa, requerendo, muitas vezes, experimentos demorados e caros (Zhu & Goldberg, 2009). Na anotação de voz, por exemplo, para que seja anotada apenas uma hora de fala, são necessárias, em média, 400 horas de análise (Godfrey et al., 1992). Na caracterização de funções proteicas, por sua vez, o trabalho necessário à identificação das funções de uma única proteína pode requerer anos de pesquisa (Mewes et al., 2004). Em alguns casos, inclusive, é necessário se considerar que dificilmente será possível rotular um montante de dados que descreva efetivamente todas as classes do problema. Tomemos como exemplo conteúdos publicados na *World Wide Web*. À medida que algumas páginas são rotuladas e conteúdos são anotados por especialistas humanos, um número muito maior de páginas é publicado. Por isso, muitas vezes conjuntos de dados estão disponíveis apenas parcialmente rotulados.

Nos últimos anos, tem sido considerada a construção de classificadores levando em conta dados não rotulados em adição aos poucos dados disponíveis rotulados. Essa tarefa é denominada **classificação semissupervisionada** (Chapelle et al., 2006; Zhu & Goldberg, 2009; Zhu, 2008). Especificamente, os dados não rotulados descrevem as distribuições dos dados e mediam a propagação dos rótulos, enquanto que os itens de dados rotulados semeiam a propagação de rótulos e guiam-na à estabilidade. A classificação semissupervisionada e outras tarefas similares compreendem o que tem sido denominado **aprendizado semissupervisionado**. Como são empregados conceitos do aprendizado supervisionado e do aprendizado não-supervisionado, o aprendizado semissupervisionado é entendido como um meio termo entre os paradigmas supervisionado e não supervisionado de aprendizado.

A pesquisa apresentada nesse texto se concentra no problema de classificação semissupervisionada. Por isso, o termo aprendizado semissupervisionado será frequentemente utilizado sinônimo ao termo classificação semissupervisionada. O custo associado à rotulação manual de dados e o desempenho superior da classificação semissupervisionada quando apenas poucos itens de dados estão disponíveis rotulados são algumas das principais razões para o estudo do aprendizado semissupervisionado (Chapelle et al., 2006; Zhu & Goldberg, 2009; Zhu, 2008).

Dentre as diversas abordagens de aprendizado semissupervisionado estudadas na atualidade, estamos interessados na abordagem baseada em redes (Zhu, 2005; Belkin et al., 2006). Nesse caso, conjuntos de dados são representados relacionalmente, através da abstração gráfica. Isto é, cada item de dado é representado por um nó que está conectado com outros nós de acordo com regras de afinidade, por exemplo, com intensidade inversamente proporcional à distância entre eles. A partir de conjuntos de dados vetoriais, a construção de um grafo (ou rede) é

uma forma de redução não linear de dimensionalidade e pode corresponder ao aprendizado de variedades (Roweis & Saul, 2000; Tenenbaum et al., 2000; Belkin & Niyogi, 2003; Belkin et al., 2006; Zhu, 2005). A propagação de rótulo sobre essa estrutura, por sua vez, é semelhante ao espalhamento de rumores, informações ou vírus em redes sociais, vírus ou *spans* em redes de e-mails e redes tecnológicas, e à propagação de sinais em redes neurais biológicas (Lloyd & May, 2001; Newman et al., 2002; Moreno et al., 2004; Vogels & Abbott, 2005; Barrat et al., 2008).

Devido à grande disponibilidade de recursos computacionais e à evolução das redes de comunicação, modelagens de dados bastante informativas, p. ex. baseadas em redes, passaram a ser amplamente usadas recentemente. O tópico denominado **redes complexas** alcançou posição de destaque, atraindo o interesse de pesquisadores de diversas áreas, inclusive da computação. Redes complexas é o nome dado às redes de grande escala que possuem padrão de conexões não trivial, tipicamente resultantes da evolução do sistema que modelam (Newman, 2003b). Dentre os sistemas frequentemente representados e estudados sob essa perspectiva, destacam-se a Internet (Faloutsos et al., 1999), a *World Wide Web* (Albert et al., 1999), as redes de distribuição de energia elétrica (Albert et al., 2004), as redes de colaboração entre cientistas (Newman, 2001), as redes de interações sociais entre indivíduos (Scott, 2000), as cadeias de contatos sexuais (Liljeros et al., 2001), as cadeias alimentares (Montoya & Solé, 2002), as cadeias de interações metabólicas (Jeong et al., 2000), e as redes neurais biológicas (Sponrs, 2002; Bullmore & Sporns, 2009), dentre muitos outros. Estudos diversos têm sido reportados relacionando características e propriedades de redes com o funcionamento dos sistemas por elas modelados (Newman, 2010; Cohen & Havlin, 2010; Barrat et al., 2008; Costa et al., 2007; Boccaletti et al., 2006; Newman, 2003b; Bornholdt & Schuster, 2003; Albert & Barabási, 2002; Fortunato, 2010).

Um dos tópicos que alavancaram a área foi a **detecção de comunidades**. Girvan & Newman (2002) mostraram que muitas redes reais possuem conjuntos de nós densamente conectados entre si, cujas ligações com nós de fora do grupo são esparsas. Esses conjuntos de nós são denominados comunidades. Diversas propriedades e comportamentos de sistemas complexos, tais como círculos de amizade, famílias e comunidades em redes sociais (Moody & White, 2003), funções proteicas em redes de interações entre proteínas (Rives & Galitski, 2003), páginas com conteúdos relacionados na web (Flake et al., 2002), entre outros, estão relacionados à sua formação. A identificação dessas estruturas é denominada detecção de comunidades (Newman & Girvan, 2004; Clauset et al., 2004; Newman, 2004a; Danon et al., 2005; Lancichinetti & Fortunato, 2009; Fortunato, 2010).

Recentemente, a detecção de comunidades vem sendo estudada considerando o paradigma semissupervisionado. Breve et al. (2012), p. ex., propuseram uma abordagem que estende a dinâmica de competição entre partículas, proposta em (Quiles et al., 2008) à cooperação entre times de partículas. Realizando caminhadas na rede, ora aleatórias, ora determinísticas, as partículas colaboram entre si de acordo com seus times e possuem vizinhanças de segurança em torno de nós rotulados com o rótulo do seu time, para onde retornam quando enfraquecidas e onde recuperam seu potencial de exploração (vide Seção 2.4.5). Ma et al. (2010), por sua vez, propuseram uma abordagem de detecção de comunidades semissupervisionada combinando o

método de fatoração de matrizes não negativas com agrupamento semissupervisionado (Lee & Seung, 1999, 2001). Allahverdyan et al. (2010), por outro lado, demonstraram que a atribuição de rótulos relativos às comunidades de uma quantidade mínima de nós é suficiente para mudar o limiar entre redes com comunidades detectáveis e redes sem comunidades detectáveis (Reichardt & Leone, 2008; Decelle et al., 2011; Steeg et al., 2014; Zhang et al., 2014).

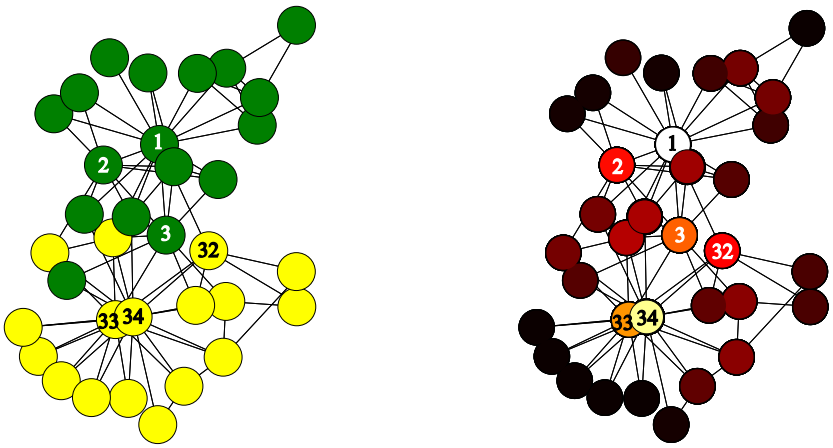
1.1 Motivação

Dado um conjunto de dado com grupos gerados a partir de distribuições arbitrárias, Guha et al. (1998) apresentam um método para estimar o tamanho da amostra aleatória capaz de representar adequadamente todos os grupos em um conjunto de dados, de modo que nenhum grupo seja omitido. Esse resultado é formulado sobre o limite de Chernoff (Motwani & Raghavan, 1995). Segundo Guha et al. (1998), o tamanho da amostra independe do número de objetos do conjunto de dados e está relacionado ao número de grupos de interesse e ao tamanho do menor grupo de interesse, em outras palavras, depende do número de objetos que estimamos ser suficiente para descrever cada grupo arbitrário. Dado um conjunto de dados com 2 grupos, cujo menor grupo de interesse possui metade do tamanho médio dos grupos, e sejam 5 exemplos de cada grupo suficientes para representar satisfatoriamente a distribuição de dados de cada grupo, o tamanho da amostra necessária para se garantir com 99% de certeza que todos os grupos do conjunto de dados são devidamente cobertos é 72 objetos. Aumentando o número de grupos para 10, e o número de elementos necessário à representação da distribuição de cada grupo para 10, o tamanho da amostra necessária chega a cerca de 500 objetos. Rotular manualmente tantos itens de dados é uma tarefa bastante onerosa. A Seção 3 apresenta em maiores detalhes os resultados de Guha et al. (1998).

Para que o aprendizado semissupervisionado ofereça bom desempenho de classificação é necessário que classes e grupos do conjunto de dados sejam consistentes (vide Sec. 2.1), fazendo o resultado acima muito importante para o aprendizado semissupervisionado. Em geral, dados são gerados de forma automatizada não rotulados e sua rotulação requer o envolvimento de especialistas no domínio, rotulando-os manualmente. Porém, dificuldades na visualização de grandes volumes de dados, bem como o custo associado ao envolvimento do especialista, são desafios que podem restringir o desempenho dessa tarefa. Logo, o destacamento automático de bons candidatos a dados rotulados, doravante denominados indivíduos representativos, é uma tarefa de grande importância, e pode proporcionar uma boa relação entre o custo com especialista e o desempenho do aprendizado, possibilitando que o tamanho da amostra suficiente à boa representação das classes de interesse seja minimizado.

Especificamente, concentramos nossos estudos em métodos baseados em redes, pois a abstração gráfica nos oferece diversas vantagens, p. ex., a representação de relações topológicas para além das relações espaciais. Além disso, medidas de redes bem definidas, são capazes de caracterizar os nós de uma rede em diversos níveis. Mais ainda, via de sentido duplo, os critérios propostos podem ser aplicados, não apenas à seleção de indivíduos representativos para semear

o aprendizado semissupervisionado, mas também a problemas em redes complexas, tais como resiliência, epidemiologia, propagação de rumores e informações (Huang et al., 2006), teoria dos jogos (Goyal et al., 2014), crescimento e auto-organização, entre outros (Barrat et al., 2008; Boccaletti et al., 2006).



(a) (b)

Figura 1.1: Rede dos clubes de caratê: em (a) colorida de acordo com o rótulo das academias e em (b) de acordo com a representatividade dos indivíduos. Quanto mais clara a cor do nó em (b), maior a representatividade do indivíduo de acordo com um dos critérios propostos no presente trabalho.

Tomemos como exemplo a rede dos clubes de caratê compilada por Zachary (1977) e ilustrada na Figura 1.1. Essa rede caracteriza as relações de amizade dos indivíduos envolvidos no episódio do rompimento das colaborações entre dois instrutores de caratê de uma academia. Por ocasião, o instrutor B desligou-se desta e tornou-se treinador de outra academia. Na ilustração, os indivíduos são representados pelos nós; as relações de amizade pelas arestas entre eles; e os instrutores pelos nós 1 e 34 (instrutor B). Acredita-se que o desligamento de diversos alunos da primeira academia e matrícula na segunda academia se deve à influência que o instrutor B exercia sobre os mesmos.

Do ponto de vista do aprendizado de máquina, os rótulos das academias associados aos indivíduos é uma informação *ground truth* de suas classes, cuja função de separação entre as mesmas pode ser aprendida. Mesmo que o leitor jamais tenha estudado esse tópico, com base na visualização da rede, ele pode fazer uma conjectura dos grupos inerentes ao problema e até mesmo dos indivíduos representativos para o problema. Em se tratando do aprendizado semissupervisionado, os nós 1 e 34 são considerados boas sementes para a propagação de rótulos e discriminação entre os indivíduos de acordo com suas academias. Adicionalmente, ao considerarmos também a rotulação dos indivíduos 2, 3, 32 e 33, podemos melhorar a qualidade da rotulação semissupervisionada, resultando no melhor ajuste do classificador.

Em cenários análogos, nós representativos podem ser considerados bons vetores de espalhamento de agentes infecciosos, rumores, etc; bons líderes ou reguladores de grupos, até mesmo responsáveis pelo processo de auto organização dos indivíduos no seu entorno; indivíduos promíscuos em redes de contatos sexuais; sistemas neurais centrais; proteínas letais em redes

de interações entre proteínas; etc. Consequentemente, são alvos interessantes, respectivamente, para a imunização eficiente da rede (Albert et al., 2000; Holme et al., 2002; Pastor-Satorras & Vespignani, 2002), para o ataque e fragmentação da rede (Albert et al., 2000; Cohen et al., 2000, 2001; Jeong et al., 2001; Newman et al., 2002), etc.

1.2 Objetivo

O presente trabalho tem como principal objetivo explorar a influência dos dados disponíveis rotulados no desempenho do processo de aprendizado semissupervisionado baseado em redes.

Desconhecidas as características do modelo gerador dos dados, podemos pensar no problema de identificação de um subconjunto representativo dos dados como um problema de otimização combinatória, cuja solução correta, por otimização multiobjetivo, maximiza a acurácia da classificação semissupervisionada e minimiza o tamanho do subconjunto de representativos. Nossa hipótese, no entanto, é a seguinte. Se tivermos os dados representados através de uma rede que reflita devidamente o processo de geração dos dados, itens de dados representativos destacar-se-ão em relação aos demais. Portanto:

I. Propriedades e medidas de caracterização de indivíduos centrais em redes complexas são estudadas a fim de propor critérios para a caracterização e identificação de nós representativos;

II. Sabido que em redes com características diversas critérios diversos de caracterização de representativos se destacam, visamos também a proposição de uma abordagem unificada para o problema;

III. Além disso, abordagens para a construção de rede a partir de dados vetoriais são estudadas no sentido de realçar nós representativos. Os grafos resultantes devem ser capazes de representar a relação entre rótulos e dados apropriadamente. Por questões de eficiência também devem ser esparsos, porém não granulados, de modo que o desempenho da propagação de rótulos não esteja condicionado à identificação de nós representativos em cada componente do grafo;

IV. Medidas de validação de construção de redes para essa finalidade também são estudadas.

1.3 Trabalhos relacionados

A identificação de indivíduos representativos dentro do aprendizado de máquina não é um problema exatamente novo. Quando em métodos clássicos de agrupamento de dados, como k-médias e k-medoides, são estimados os centroides e medoides, respectivamente, busca-se por representativos dos grupos que resultam na minimização do erro empírico,

$$\varepsilon = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - m_i\|^2, \tag{1.1}$$

onde C_i é o conjunto de objetos da partição i e m_i é o centroide (ou medoide) estimado para i (Jain & Dubes, 1988; Kaufman & Rousseeuw, 1990; Jain, 2010). A relação com o objeto do presente trabalho torna-se mais evidente ao considerarmos o método Clustering by Using REpresentatives (CURE), que, com fins à representação apropriada de grupo de dados com formatos arbitrários, considera diversos representativos por grupo (Guha et al., 1998). De outra maneira, o método DBSCAN define os itens de dados como pontos centrais, pontos de borda, ou *outliers*, de acordo com suas disposições em relação aos demais (Ester et al., 1996). Ainda mais relacionado ao presente trabalho, Frey & Dueck (2007) propuseram um método de agrupamento de dados baseado na propagação de mensagens em uma rede de similaridades entre pares de indivíduos, denominado propagação de afinidade. Esse método considera a identificação de um subconjunto de exemplos representativos, através dos quais os grupos emergem. Os demais itens de dados são conectados aos exemplos representativos que melhor lhes representem.

Esses métodos buscam por indivíduos representativos considerando o espalhamento dos dados. Em se tratando de dados em rede, outra abordagem bastante conveniente é primeiro identificar as comunidades da rede, e, em seguida, caracterizar os indivíduos representativos de cada comunidade (Motta et al., 2009). A hipótese estudada no presente trabalho, vai um pouco além, no entanto. Para nós, a representatividade é um fenômeno diretamente relacionado à formação de comunidades em redes. Isto é, as estruturas de comunidades em uma rede são resultado da organização de seus indivíduos em torno de seus representativos. E, por isso, assumimos que estruturas de rede consistentes com o processo de geração dos dados devem realçar indivíduos representativos. Até onde sabemos, abordagens mais semelhantes à abordagem proposta não são conhecidas. Podemos considerá-la no entanto, como uma variação do **aprendizado ativo**, uma vez que os critérios estudados possibilitam a escolha dos dados sobre os quais o processo de aprendizagem é realizado (Settles, 2009; Olsson, 2009).

Aprendizado semissupervisionado e aprendizado ativo são paradigmas de aprendizado de máquina relacionados que atacam a indisponibilidade de dados rotulados de formas opostas. Como destaca Settles (2009), enquanto que o aprendizado semissupervisionado explora aquilo que se pensa saber sobre os dados não rotulados (p. ex., que sua distribuição é consistente com as distribuições das classes), métodos de aprendizado ativo são propostos de forma a explorar aspectos dos dados que por ora são desconhecidos. Seja L o orçamento total previsto para a rotulação de dados, o aprendizado ativo consiste em iterativamente selecionar um subconjunto dos dados não rotulados, $l \ll L$, apresentar os itens de dados a um oráculo capaz de rotulá-los com certo grau de confiança, e utilizá-los para induzir um classificador. Na primeira iteração o subconjunto de dados selecionado para a rotulação manual é escolhido arbitrariamente ou ao acaso. Da segunda iteração em diante, a seleção acontece de acordo com critérios como nível de incerteza do classificar sobre o rótulo predito para o item de dado, divergência entre os rótulos atribuídos por um comitê de classificadores, mudança esperada no modelo, redução esperada do erro do modelo e redução esperada da variância do modelo. Em outras palavras, métodos de aprendizado ativo selecionam seu conjunto de indivíduos representativos de forma adaptativa, de acordo com a resposta do modelo aos incrementos no conjuntos dos dados rotulados. A

abordagem estudada no presente trabalho, difere, portanto, das abordagens tradicionais de aprendizado ativo ao realizar a seleção do conjunto de representativos em uma única iteração, buscando tirar vantagem da representação dos dados. Mesmo abordagens de aprendizado ativo em lote, que maximizam l, têm o ganho de desempenho atrelado à iteratividade do aprendizado ativo (Settles, 2009; Brinker, 2003; Xu et al., 2007; Hoi et al., 2006b,a).

1.4 Organização do documento

Doravante este documento esta organizado em duas partes. A primeira parte compreende uma revisão bibliográfica sobre fundamentos e conceitos relacionados aos tópico aprendizado semissupervisionado, amostragem aleatória e redes complexas e a segunda parte compreende resultados, discussões e conclusões que suportam a presente tese. O leitor que tem certo domínio sobre os tópicos aprendizado semissupervisionado e redes complexas pode passar diretamente à segunda parte desse texto e inteirar-se das principais contribuições desse trabalho. Sempre que tópicos específicos forem abordados, referências às seções que os introduzem são feitas, possibilitando ao leitor obter maiores esclarecimentos sobre o tópico, caso lhe pareça necessário.

Parte I

Fundamentos e Conceitos

Capítulo

2

Aprendizado Semissupervisionado

Aprendizado de máquina é a área da ciência da computação que visa desenvolver métodos capazes de ‘‘aprender’’ através da experiência. Modelos computacionais são construídos com o objetivo de organizar o conhecimento inerente aos dados através da inferência de hipóteses. Esse processo imita de forma automática o processo de análise e organização dos dados que seria realizado por um especialista humano no domínio (Faceli et al., 2011; Bishop, 2006; Mitchell, 1997). Duas das principais motivações para o aprendizado de máquina são: a crescente complexidade dos problemas computacionais; e o grande volume de dados gerados por setores diversos. Nesse contexto, a demanda das tarefas é muito grande, fazendo da atuação de especialistas humanos, além de onerosa, ineficiente. Logo, as técnicas de aprendizado de máquina reduzem a necessidade de intervenção humana e a dependência dos especialistas, gerando hipóteses através da exposição aos dados.

Tradicionalmente, o processo de aprendizado é classificado como supervisionado ou não supervisionado, de acordo com o grau de envolvimento de um supervisor externo. Alguns autores, no entanto, consideram a categorização em relação à tarefa de aprendizagem realizada. De acordo com esse critério o aprendizado de máquina é dividido em **aprendizado preditivo** e **aprendizado descritivo** (Faceli et al., 2011; Mitchell, 1997).

Podemos formular o aprendizado descritivo da seguinte maneira. Seja $X = \{x_1, \ldots, x_n\}$ um conjunto de n itens de dados (também denominados objetos, exemplos, instâncias, padrões, ou registros), onde $x_i \in \mathcal{X}$, um hiperespaço $\mathbb{R}^D$, D-dimensional; X também pode ser definido em termos da matriz $(n \times D)$-dimensional $\mathbf{X}$, onde cada linha compreende um exemplo x_i. Seja cada exemplo x_i identicamente e independentemente distribuído (i.i.d), amostrado a partir de uma distribuição em $\mathcal{X}$, o aprendizado descritivo consiste fundamentalmente em estimar a densidade geradora de X (Chapelle et al., 2006). No entanto, hipóteses mais fracas também são tomadas, tais como o agrupamento dos dados, a detecção de *outliers*, a sumarização, e a associação. Na tarefa de agrupamento de dados, por exemplo, os dados são agrupados de acordo

com algum critério de similaridade entre eles, de modo que sejam formados grupos homogêneos (Jain, 2010; Everitt et al., 2009; Jain & Dubes, 1988). Na detecção de *outliers*, outro exemplo, o objetivo é identificar dados discrepantes, sendo que a avaliação sobre os dados pode se dar: em relação a todo o conjunto de dados, onde os *outliers* são ditos pontuais; em um contexto específico, onde os *outlier* são ditos contextuais; ou no coletivo, onde coleções de objetos são ditos *outliers* coletivos (Araújo, 2010; Chandola et al., 2009). Na sumarização (ou redução de dimensionalidade), por sua vez, o objetivo é encontrar uma representação simples e compacta dos dados. Nesse sentido, são utilizados desde medidas estatísticas a técnicas de visualização. Na associação, por fim, o objetivo é a identificação de padrões de associação frequentes. Cada objeto representa uma transação, e é composto por um conjunto de itens. Subconjuntos de itens que frequentemente aparecem associados em transações são ditos padrões de associação frequentes (Faceli et al., 2011).

O aprendizado preditivo, por sua vez, pode ser formulado como segue. Seja $X = \{x_1, \ldots, x_n\}$ um conjunto de n dados, como descrito no parágrafo anterior, e seja $Y = \{y_1, \ldots, y_n\}$ um conjunto de rótulos, cada elemento $x_i \in X$ possui um elemento y_i associado. Existem duas famílias de algoritmos preditivos: os **algoritmos generativos** e os **algoritmos discriminativos** (Chapelle et al., 2006). Os algoritmos generativos tentam modelar a densidade condicional dos rótulos, $p(x|y)$, utilizando algum procedimento de aprendizado não supervisionado. Um exemplo é a inferência da densidade preditiva através do teorema de Bayes, onde

$$p(y|x) = \frac{p(x|y)p(y)}{\int_{\mathcal{Y}} p(x|y)p(y)dy}. \tag{2.1}$$

$p(x|y)p(y) = p(x, y)$ é a densidade conjunta dos dados a partir da qual cada par (x_i, y_i) pode ter sido gerado. Os algoritmos discriminativos, por sua vez, tentam estimar $p(y|x)$. Sejam (x_i, y_i) pares gerados i.i.d., de acordo com uma distribuição no hiperespaço $\mathcal{X} \times \mathcal{Y}$, o objetivo do aprendizado é aprender uma função de mapeamento $\hat{f} : x \to y$, capaz de aproximar a função f, desconhecida, que de fato mapeia $\mathcal{X} \times \mathcal{Y}$. Alguns métodos, por exemplo, limitam-se a construir um modelo de inferência que avalie apenas se $p(y|x)$ é maior ou menor que certo limiar. É o caso das máquinas de vetor de suporte (SVM). Adicionalmente, se, $\forall x_i$, $\hat{f}(x_i) = y_i \in \{c_1, \ldots, c_\kappa\}$, i.e., $\hat{f}(x_i)$ assume valores em um conjunto finito e discreto de rótulos, a função $\hat{f}$ é dita um **classificador** e a tarefa de aprendizado é denominada **classificação**; se, por outro lado, $\forall x_i$, $\hat{f}(x_i) = y_i \in \mathbb{R}$, i.e., $\hat{f}(x_i)$ assume valores em um conjunto infinito e ordenado de valores, i.e., valores contínuos, $\hat{f}$ é dita um **regressor** e a tarefa é denominada **regressão**.

Outro critério para categorizar o aprendizado diz respeito ao domínio da hipótese ‘‘aprendida’’. Frequentemente se deseja generalizar o conhecimento aprendido para dados além dos dados disponíveis no processo de aprendizado, i.e., além do conjunto de treinamento. Quando a hipótese é capaz de fazê-lo efetivamente, o aprendizado é dito **indutivo**. Porém, em alguns casos é suficiente que a hipótese seja capaz de lidar apenas com o conjunto de treinamento. Nesse caso o aprendizado é dito **transdutivo**. Considere, por exemplo, a tarefa de classificação. Enquanto que um classificador indutivo discriminativo infere uma função $\hat{f} : x \to y$ no hiperespaço $\mathcal{X} \times \mathcal{Y}$ capaz de realizar predições sobre qualquer $x_i \in \mathcal{X}$, um classificador transdutivo discriminativo

infere uma função $\hat{g} : X_u \rightarrow y$ definida apenas no conjunto finito de elemento X_u, conhecido durante o processo de aprendizado, dito conjunto de treinamento. Ainda que inferidas a partir do mesmo conjunto de pares (x_i, y_i), as funções $\hat{f}$ e $\hat{g}$, possuem domínios distintos, diferindo claramente no viés sob o qual foram inferidas. Como considerado por Vapnik (1999), no sentido de resolver um problema, não se deve resolver um problema mais complexo como um passo intermediário. Não se deve inferir uma hipótese indutiva no sentido de resolver apenas um problema transdutivo. Baseados em redes, os classificador utilizados no presente trabalho são em geral transdutivos. Extensões são estudadas a fim de ampliar o domínio de inferência de suas hipóteses.

Como um meio termo entre os paradigmas não supervisionado e supervisionado de aprendizado, recentemente, o paradigma de aprendizado denominado **aprendizado semissupervisionado** tem sido considerado. Nesse caso, poucos dados possuem rótulos associados, enquanto que a maior parte não possui (Chapelle et al., 2006; Zhu, 2008). Tanto tarefas preditivas, como tarefas descritivas podem ser tratadas sob o paradigma semissupervisionado. No entanto, o viés utilizado nessas tarefas é modificado para incorporar informações tanto de dados rotulados como de dados não rotulados.

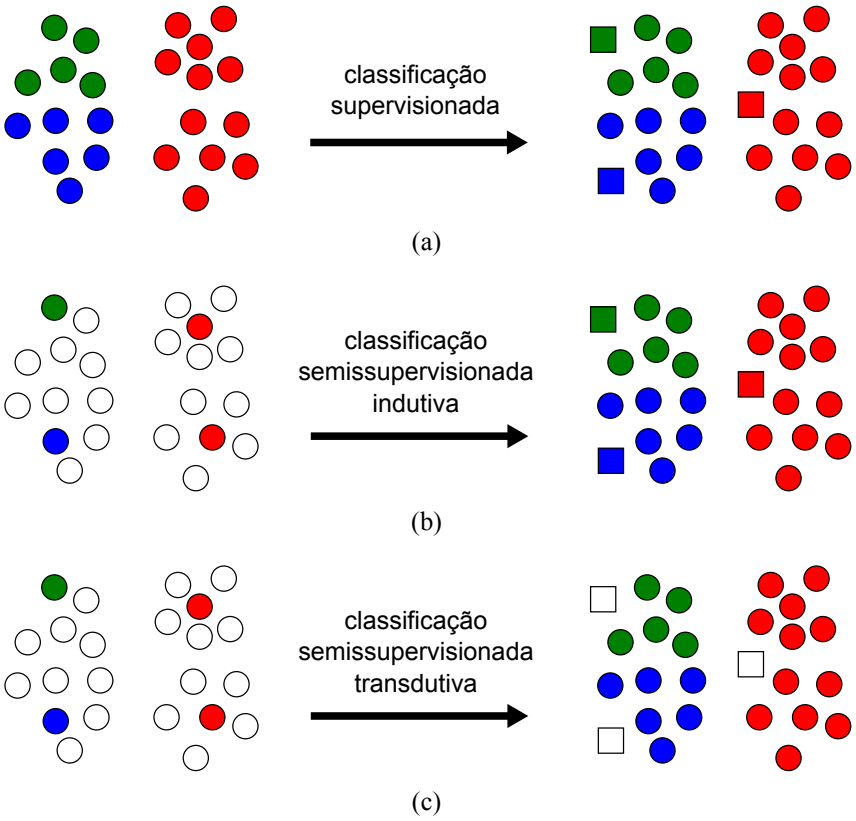


Figura 2.1: Tarefa de classificação: (a) supervisionada, (b) semissupervisionada indutiva, e (c) semissupervisionada transdutiva. As diferenças entre esses paradigmas são a forma como os dados são tratados com fins à modelagem de um classificador e o domínio do modelo. Enquanto que o supervisionado considera apenas dados rotulados, o semissupervisionado considera tanto dados rotulados quanto dados não rotulados. Com relação ao domínio do modelo, por outro lado, o aprendizado transdutivo infere hipóteses que limitam-se ao conjunto de dados disponível durante o treinamento, enquanto que o aprendizado indutivo infere hipóteses que atuam sobre todo o hiperespaço dos dados.

No presente trabalho a tarefa de interesse é a **classificação semissupervisionada**, formalizada da seguinte maneira. Seja o conjunto de dados $X = \{x_1, \ldots, x_n\}$ dividido em dois subconjuntos: $X_l = \{x_1, \ldots, x_l\}$ e $X_u = \{x_{l+1}, \ldots, x_n\}$, onde $l \ll n$. Seja $Y = \{y_1, \ldots, y_l\}$ um conjunto de rótulos, em que todo $y_i \in \{c_1, \ldots, c_\kappa\}$. Todo $x_i \in X_l$ possui um rótulo y_i associado. Os rótulo do subconjunto X_u não são conhecidos a priori. O aprendizado semissupervisionado tem como objetivo inferir um classificador $\hat{f} : x \rightarrow y$, caso indutivo, ou um classificar $\hat{g} : X_u \rightarrow y$, caso transdutivo, capaz de rotular os dados não rotulados de acordo com o conjunto de rótulos, Y, associado a X_l. Tanto dados rotulados como dados não rotulados são levados em consideração na indução da hipótese.

Na classificação supervisionada, apenas dados rotulados são considerados. Nesse caso, para que seja possível induzir uma boa hipótese, que de fato aproxime o mapeamento existente em $\mathcal{X} \times \mathcal{Y}$, é necessário que os dados rotulados representem muito bem as características dos dados em todo hiperespaço. Em outras palavras, a qualidade do modelo induzido está condicionada a quão bem os dados são representados para amostra de dados disponível rotulada. Logo, quando estão disponíveis apenas poucos dados rotulados, a qualidade de classificação supervisionada é frequentemente ruim. Na classificação semissupervisionada, os dados não rotulados possibilitam, na ausência de suficientes dados rotulados, a descrição do espalhamento das classes, suportado pelas hipóteses de suavidade e grupos. A Figura 2.1 ilustra a diferença entre a tarefa de classificação supervisionada, semissupervisionada indutiva e semissupervisionada transdutiva.

2.1 Suposição de suavidade, agrupamento e variedade

O paradigma semissupervisionado torna-se terminantemente interessante quando, para essa tarefa, são satisfeitas duas suposições sobre a consistência dos dados: a **suposição de suavidade** e a **suposição de grupos** (Zhou et al., 2004; Belkin et al., 2006). Para que seja satisfeita a suposição de suavidade, dados próximos no espaço de atributos devem possuir rótulos similares. A hipótese induzida para a tarefa de classificação semissupervisionada deve ser mais suave em regiões de alta densidade do que em regiões de baixa densidade. A suposição de grupos, por sua vez, determina que dados que estão localizados em um mesmo grupo possuam um mesmo rótulo.

Enquanto que a suposição de grupos garante a consistência global entre rótulos e grupos, a suposição de suavidade garante a consistência local entre rótulos e grupos, completando uma à outra. Satisfeitas essas suposições, o paradigma semissupervisionado de classificação consegue melhorar o paradigma supervisionado com poucos dados rotulados. Considere a Figura 2.2 em que os círculos escuros denotam dados rotulados, enquanto que os círculos claros representam dados não rotulados. Aplicando-se um algoritmo supervisionado, a fronteira de decisão é estabelecida próximo à linha pontilhada. No entanto, considerando um algoritmo semissupervisionado, a fronteira de decisão é estabelecida próximo à linha contínua. Ao considerar a distribuição dos dados não rotulados um algoritmo semissupervisionado irá estabelecer sua superfície de decisão satisfazendo às suposições de suavidade e grupos, refletindo mais fielmente a distribuição dos dados.

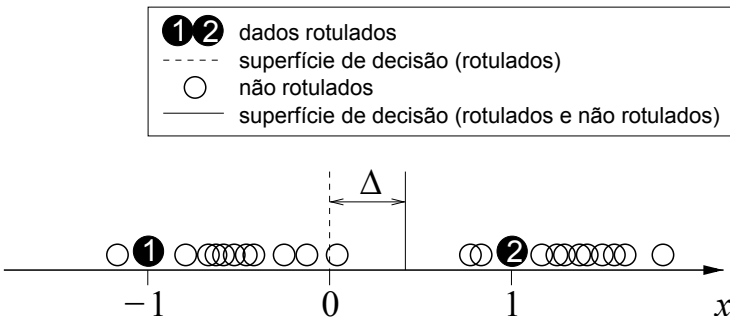


Figura 2.2: Ganho com a classificação semissupervisionada. Os círculos escuros denotam dados rotulados, enquanto que círculos claros, dados não rotulados. Enquanto que aplicando-se um algoritmo supervisionado, a fronteira de decisão seria estabelecida próximo à linha pontilhada, aplicando-se um algoritmo semissupervisionado, a fronteira de decisão certamente seria estabelecida próximo à linha contínua. Os dados não rotulados permitem construir superfícies de decisão melhor ajustadas à distribuição dos dados. Figura reproduzida de (Zhu et al., 2007), com permissão. Copyright © 2007, Association for the Advancement of Artificial Intelligence.

Adicionalmente, temos também a **suposição de variedade**. Sejam os dados representados em um espaço D-dimensional através de hiper-superfícies curvas, ditas variedades[1], assumimos que existe um espaço d-dimensional, $d \ll D$, onde as variedades são mais facilmente separáveis. Dados pertencentes a uma mesma variedade devem possuir um mesmo rótulo.

Resumindo, deve existir consistência entre classes e grupos, variedades e regiões densas. O conhecimento sobre $p(x)$ adquirido através dos dados não rotulados deve consistir em informações úteis à inferência de $p(y|x)$. Caso contrário, o aprendizado semissupervisionado, ao invés de induzir hipóteses melhores que o aprendizado supervisionado, pode induzir hipóteses degeneradas, desvirtuando $p(y|x)$. Esse efeito é investido por Cozman et al. (2003); Cozman & Cohen (2006).

2.2 Significância do aprendizado semissupervisionado

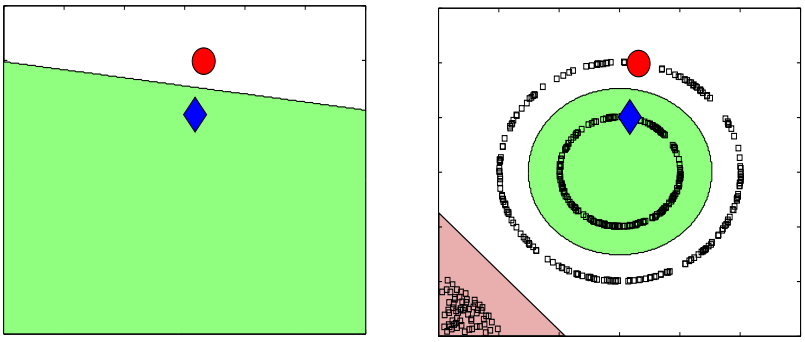
As motivações mais comuns para o aprendizado semissupervisionada estão na dificuldade de obtenção de dados rotulados e, quando apenas poucos dados estão disponíveis rotulados, na melhoria da qualidade de aprendizado ao se considerar adicionalmente os dados não rotulados. No entanto a significância do aprendizado semissupervisionado vai além dessas questões. Como considerado por Belkin et al. (2006), sem dúvidas, a maior parte do **aprendizado natural** ocorre em regime semissupervisionado, pois vivemos em um mundo no qual os seres vivos são constantemente expostos a um fluxo de estímulos naturais. Esses estímulos correspondem a dados não rotulados aos quais temos fácil acesso. Por exemplo, consideremos o problema de aquisição fonológica, onde uma criança é exposta continuamente a diversos estímulos acústicos. Naturalmente, os estímulos não são acompanhados de marcadores fonológicos identificáveis (rótulos). Nesse caso, o *feedback* corretivo (positivo ou negativo) por parte de uma outra pessoa é a principal fonte de rótulos, sendo, uma pequena porção de *feedbacks*, suficiente para permitir

[1]Do inglês: *manifold*.

à criança dominar o mapeamento acústico-fonético de quaisquer linguagens.

Ainda segundo Belkin et al. (2006), a habilidade humana de aprender conceitos não supervisionados, como, por exemplo, grupos e categorias de objetos, sugere que dados não rotulados podem ser utilizados satisfatoriamente para aprender invariâncias, para formar categorias, e para construir classificadores. Na maioria das tarefas de reconhecimento de padrões, os seres humanos têm acesso apenas a uma pequena quantidade de dados rotulados. Logo, o sucesso do aprendizado humano se deve em muito à utilização efetiva dos dados não rotulados para extrair informações necessárias à generalização do aprendizado.

Considere a ilustração apresentada na Figura 2.3. Em (a) são apresentados dois objetos rotulados de acordo com classes distintas. No sentido de separar os objetos da imagem em classes, a superfície de separação induzida por um humano se aproxima da superfície ilustrada pela cor de fundo, uma separação linear. Além disso, o conhecimento prévio sobre classes, também induz o observador a concluir que existem apenas duas classes ilustradas no problema. Considerando adicionalmente dados não rotulados, como em (b), o processo de indução é enviesado. A superfície de separação é ajustada à distribuição observada nos objetos não rotulados. Intuitivamente, os dados não rotulados acrescentam informações ao processo de aprendizagem.



(a) (b)

Figura 2.3: Informações intrínsecas acrescentadas por dados não rotulado e pelo conhecimento prévio. Em (a) dois objetos de classes distintas estão devidamente rotulados. A superfície de separação induzida por um humano é ilustrada pela cor de fundo, uma separação linear. Em (b) são considerados, adicionalmente, os dados não rotulados. A superfície de separação induzida é modificada. Figura adaptada de (Belkin et al., 2006) com a permissão dos autores.

Não por acaso, ao observarmos os padrões tridimensionais apresentados na Figura 2.4 somos capazes de separá-los em duas classes. Zhu et al. (2007) realizaram um estudo utilizando padrões como esses em um experimento comportamental no sentido de verificar se de fato os seres humanos realizam aprendizado semissupervisionado. Para avaliar o aprendizado, novos exemplos foram apresentados, para os quais foram registradas a inferência e o tempo de reação dos participantes. Os resultados obtidos levaram os autores a concluir que um comportamento semissupervisionado tenha sido, de fato, realizado pelos participantes da pesquisa. Aprendendo apenas com padrões rotulados a inferência dos mesmos sobre novos exemplos resultou em uma superfície de separação bastante estável a meia distância entre os padrões rotulados. Quando apresentados, adicionalmente, padrões não rotulados, a superfície de separação foi deslocada

de tal forma a refletir a distribuição através da qual os padrões não rotulados foram gerados. Além disso, também foi observado o deslocamento na distribuição do tempo médio de resposta. Esse resultado também é significativo, pois padrões próximos da superfície real de separação requerem mais tempo para serem categorizados. A observação do deslocamento da distribuição do tempo médio de resposta sugere que de fato padrões não rotulados, modificam não somente as inferências, como também a real superfície de separação criada pelos participantes.

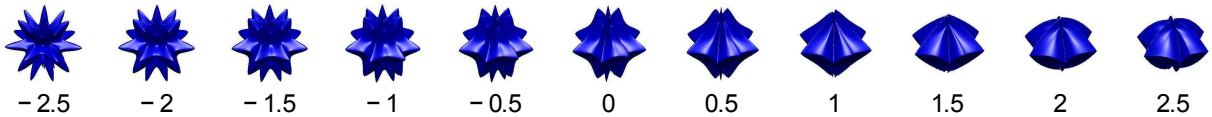


Figura 2.4: Exemplos de padrões tridimensionais considerados por Zhu et al. (2007). A forma geométrica varia de acordo com o parâmetro x, apresentado abaixo de cada forma. No experimento alguma formas foram apresentadas rotuladas; a maioria, no entanto, sem rótulos. Figura reproduzida de (Zhu et al., 2007), com permissão. Copyright © 2007, Association for the Advancement of Artificial Intelligence.

2.3 Principais abordagens de aprendizado semissupervisionado

Vários algoritmos de aprendizado semissupervisionado tem sido propostos na literatura (Zhu, 2008; Chapelle et al., 2006; Zhu & Goldberg, 2009), sendo que grande parte dos mesmos utiliza como base algum algoritmo que foi projetado como técnica de aprendizado supervisionado ou não supervisionado. Esses algoritmos são modificados de modo que possam tratar tanto dados rotulados como dados não rotulados. Temos como exemplo o COP k-médias (Wagstaff et al., 2001), que é uma modificação do k-médias (MacQueen, 1967); máquinas de vetor de suporte transdutivas (TSVM) (Vapnik, 1998), que é uma extensão das máquinas de vetor de suporte (SVM) (Vapnik, 1999; Cortes & Vapnik, 1995); e até mesmo algoritmos baseados em maximização de expectativa (EM) (Nigam et al., 2000). Segundo Chapelle et al. (2006), as técnicas de aprendizado semissupervisionado podem ser classificadas em: modelos geradores, modelos de separação por baixa densidade e modelos baseados em redes. A forma como as abordagens de aprendizado semissupervisionado tratam as suposições de suavidade, grupos e variedades representa uma das diferenças fundamentais entre elas.

Zhu & Goldberg (2009) consideram, adicionalmente, as abordagens baseadas em auto-treinamento e em co-treinamento. O auto-treinamento funciona como um método *wrapper*, onde a escolha da técnica de aprendizado utilizada é totalmente aberta. O classificador escolhido é inicialmente treinado com o pequeno montante de dados rotulados inicialmente disponível, e então é utilizado para classificar os dados não rotulados. Dos dados inicialmente não rotulados, aqueles classificados com a maior certeza e os respectivos rótulos associados são adicionados ao conjunto de treinamento. O processo de indução do classificador se repete iterativamente até que todos os objetos inicialmente não rotulados sejam rotulados com um alto grau de certeza (Algoritmo 2.1). O classificador usa sua própria predição para ensinar a si mesmo.

Algoritmo 2.1 Auto-Treinamento

1: Defina $L \leftarrow \{(x_i, y_i)_{i=1}^{l}\}$ e $U \leftarrow \{(x_j, y_j)_{j=l+1}^{l+u}\}$;
2: **repita**
3: Treine um classificador f no conjunto L utilizando aprendizado supervisionado;
4: Aplique f aos dados não rotulados em U;
5: Remova de U o subconjunto S cuja classificação possui auto índice de certeza;
6: Adicione $\{(x, f(x)) | x \in S\}$ a L;
7: **até que** $U = \emptyset$;

Uma desvantagem nessa abordagem é que eventuais erros de classificação serão propagados para as iterações subsequentes. Porém, alguns algoritmos evitam que erros de classificação sejam reforçados ''desaprendendo'' dados cujo nível de certeza seja depreciado abaixo de um certo limiar. Outras abordagens consideram diversos classificadores atuando de forma cooperativa, como *ensembles* (Zhou & Li, 2005). Considerando dois classificadores, por exemplo, o primeiro classificador rotula exemplos que serão utilizados pelo segundo classificador e vice-versa, no que é conhecido como **co-treinamento** (Blum & Mitchell, 1998; Goldman & Zhou, 2000).

2.4 Aprendizado semissupervisionado baseado em redes

Foco do presente trabalho, dedicamos uma seção especificamente à apresentação da abordagem de aprendizado semissupervisionado baseada em grafos (ou redes). Nessa abordagem itens de dados são representados através de vértices conectados entre si, através de arestas definidas a partir de um função de similaridade, dissimilaridade, ou afinidade. Especificamente, X é representado através de um grafo $G = (V, E)$, $V = \{v_1, \ldots, v_n\}$ e $E = \{e_1, \ldots, e_m\} \subset V \times V$. Cada nó $v_i \in V$ representa uma instância de dados $x_i \in X$ e cada aresta $e_r = (v_i, v_j) \in E$ uma relação de afinidade (similaridade ou dissimilaridade) entre um par de vértices. E é frequentemente representado através de uma matriz de adjacências generalizada, $\mathbf{W}$, onde $w_{ij} > 0$ se a aresta $(v_i, v_j) \in E$, caso contrário $w_{ij} = 0$.

A abstração gráfica apresenta como principal característica a capacidade de abstração dos dados através de uma topologia, mimetizando similaridades topológicas através de arestas sucessivas. Essa característica é particularmente relevante pois itens de dados espacialmente distantes podem tornar-se topologicamente próximos. Consequentemente, pode-se representar dados como distribuições espaciais diversas. Uma vez construída a rede, o problema de aprendizado deixa de depender de uma função de similaridade. Inclusive, muitos problemas de aprendizado têm seus dados originalmente representados através de redes, tais como a *World Wide Web*, a Internet, redes de interações entre proteínas, etc (problemas desse gênero são melhor discutidos na Seção 4).

Uma ampla revisão de algoritmos de aprendizado semissupervisionado baseados em redes pode ser encontrada em (Zhu, 2005, 2008; Zhu & Goldberg, 2009; Chapelle et al., 2006). Suas principais referências são (Blum & Chawla, 2001; Zhu et al., 2003; Zhou et al., 2004; Joachims, 1999; Belkin et al., 2006). A seguir são apresentados os métodos estudados no presente trabalho. Embora métodos de rotulação semissupervisionados sejam frequentemente

propostos juntamente com um métodos específico de construção de rede a partir de bases de dados vetoriais, quaisquer métodos de construção de rede podem ser utilizados para esse fim. Por isso, dedicamos uma Seção especificamente à discussão desses métodos. A interpretação do aprendizado semissupervisionado baseado em redes como um *framework* de regularização também é discutida a seguir.

2.4.1 Fluxo máximo corte mínimo

Considerando o problema de classificação binário, uma das primeiras propostas de aprendizado semissupervisionado baseado em redes foi apresentada por Blum & Chawla (2001). Com dados rotulados de acordo com as classes positivo e negativo, $X_l = X_l^+ \cup X_l^-$, uma analogia é a seguinte. Os elementos em X_l^+ são ligado a uma fonte de um fluido que flui através da rede até um sumidouro ligado aos elementos em X_l^-. O problema de classificação binária equivale a eliminar o fluxo da fonte ao sumidouro através de um corte mínimo que particiona a rede em duas subredes. As arestas que compõem esse corte são aquelas por onde passa o fluxo máximo. Uma vez dividida a rede, os nós conectados às fontes recebem rótulo positivo, enquanto que os nós conectados aos sumidouros recebem rótulo negativo. O algoritmo proposto por Blum & Chawla (2001) é apresentado no Algoritmo 2.2.

Algoritmo 2.2 Corte mínimo (Blum & Chawla, 2001)

1: Construa uma rede $G \leftarrow (V, E)$, onde $V \leftarrow X_l \cup X_u \cup \{v_+, v_-\}$;
2: Defina $w(v, v_+) \leftarrow \infty$ para todo $v \in X_l^+$ e $w(v, v_-) \leftarrow \infty$ para todo $v \in X_l^-$;
3: Defina as demais ligações de acordo com a Eq. 2.13;
4: Calcule o corte mínimo (v_+, v_-) e remova as arestas da partição entre $v_- \in V_-$ e $v_+ \in V_+$;
5: Atribua rótulo positivo aos exemplo em V_+ e rótulo negativo aos exemplos em V_-.

O cálculo do corte mínimo equivale a, fixados os rótulos de X_l, identificar os rótulos de $x_i \in X_u$, $f(x_i) \in \{-1, 1\}$, $i = l + 1, \ldots, n$, tal que $\sum_{i,j:f(x_i) \neq f(x_j)} w_{ij}$ seja minimizado, o que é um problema de otimização combinatória. O algoritmo de corte mínimo e fluxo máximo apresenta uma solução polinomial para esse problema (L. R. Ford & Fulkerson, 1956; Cormen et al., 2001).

2.4.2 Campos Gaussianos e funções harmônicas

Uma das limitações do algoritmo de corte mínimo é que o classificador f é capaz apenas de atribuir rótulos rígidos aos dados, $f(x) \in \{-1, 1\}$. No entanto, especialmente em regiões de bordas de grupos, ou mesmo regiões de sobreposição de classes, os dados podem não ser tão confidentes aos respectivos rótulos. Além disso, sua implementação em problema de classificação multiclasse requer a utilização de uma matriz de códigos. Logo, Zhu & Ghahramani (2002); Zhu et al. (2003) propuseram uma relaxação desse método, utilizando funções harmônicas.

No contexto de redes, uma função harmônica é uma função f que atribui a um nó não rotulado um rótulo que corresponde à média ponderada de seus vizinhos. f é utilizada nessa proposta a fim de realizar classificações suaves, $f(x) \in [0, 1]$. O Algoritmo 2.3 apresenta a

solução proposta por Zhu & Ghahramani (2002); Zhu et al. (2003). Inicialmente, $f_j = y_j$ para todo vértice $j \in X_l$ e para os vértices não rotulados um valor arbitrário qualquer. Interativamente, o valor de f, para todo vértice não rotulado, é atualizado utilizando a função harmônica,

$$f_i \leftarrow \frac{\sum_j w_{ij} f_j}{\sum_j w_{ij}}.$$

f_i, $i \in X_u$, garantidamente converge para uma solução equivalente ao campo gaussiano de f_j, $j \in X_l$, independentemente dos valores iniciais que lhes foram atribuídos.

Algoritmo 2.3 Campos Gaussianos e funções harmônicas (Zhu & Ghahramani, 2002; Zhu et al., 2003)

1: Construa uma matriz de afinidade $\mathbf{W}$, onde $w_{ii} \leftarrow 0$ e w_{ij} é dado pela Eq. 2.13 para $i \neq j$;
2: Calcule a matriz diagonal $\mathbf{D}$, onde $d_{ii} \leftarrow \sum_j w_{ij}$ e $d_{ij} \leftarrow 0$ para $i \neq j$;
3: Inicialize $f_j(0) \leftarrow y_j$ para todo $j \in X_l$ e $f_i(0) \leftarrow 0$ para todo $i \in X_u$;
4: **repita**
5: $\quad f(t+1) \leftarrow \mathbf{D}^{-1}\mathbf{W}f(t)$;
6: $\quad f_j(t+1) \leftarrow f_j(t)$;
7: **até que** f convirja;
8: Rotule cada item de dado usando o sinal de $f_i(\infty)$.

2.4.3 Consistência local e global

Ambos os métodos apresentados acima fixam os rótulos associados às instâncias rotuladas, $f_j = y_j$ para todo $x_j \in X_l$. No entanto, algumas vezes, esses dados podem ter sido incorretamente rotulados, de modo que é muito interessante que o classificador f seja capaz de discordar dos rótulos previamente atribuídos aos dados. O método de consistência local e global (Zhou et al., 2004) trata diretamente esse problema. A cada iteração o rótulo de cada nó é atualizado tanto com base nos seus valores iniciais, como com base nos valores atuais de seus vizinhos. Especificamente, o balanceamento entre as duas contribuições é ponderado pelo parâmetro α. O Algoritmo 2.4 apresenta a solução proposta por Zhou et al. (2004). Podemos dizer que este método favorece à satisfação da suposição de grupos.

Algoritmo 2.4 Consistência local e global (Zhou et al., 2004)

1: Construa uma matriz de afinidade $\mathbf{W}$, onde $w_{ii} \leftarrow 0$ e w_{ij} é dado pela Eq. 2.13 para $i \neq j$;
2: Calcule a matriz diagonal $\mathbf{D}$, onde $d_{ii} \leftarrow \sum_j w_{ij}$ e $d_{ij} \leftarrow 0$ para $i \neq j$;
3: Calcule o Laplaciano normalizado da rede, $\mathcal{L} \leftarrow \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$;
4: Escolha um parâmetro $\alpha \in [0, 1)$;
5: Inicialize $f_j(0) \leftarrow y_j$ para todo $j \in X_l$ e $f_i(0) \leftarrow 0$ para todo $i \in X_u$;
6: **repita**
7: $\quad f(t+1) \leftarrow \alpha\mathcal{L}f(t) + (1-\alpha)Y$;
8: **até que** f convirja;
9: Rotule cada item de dado usando o sinal de $f(\infty)$.

2.4.4 Framework de regularização

O problema de classificação semissupervisionado baseado em redes pode ser definido como um *framework* de regularização. Sejam dados o conjunto de dados, $X = X_l \cup X_u$, e rótulos y_i associados a todo $x_i \in X_l$. Seja $\mathcal{F}$ um conjunto de matrizes $n \times \kappa$ onde $F \in \mathcal{F}$ é uma possível matriz solução para o problema, isto é, $f(x_i) = \arg\max_{c \in \{1,\ldots,\kappa\}} F_{ic}$. A solução do problema de classificação semissupervisionado baseado em redes pode ser definida como:

$$F^* = \arg\min_{F \in \mathcal{F}} Q(F),$$

onde Q é uma função de custo minimizada pela solução F^*. Q deve ser definida de forma a satisfazer duas condições: 1) $f(x_i)$ deve aproximar rótulos $y_i \in Y$ dos dados previamente rotulados; e 2) $f(x_i)$ deve ser suave em toda a rede não apresentando grandes variações em regiões bem definidas. A primeira condição é denominada erro e a segunda regularização (Zhu, 2008). Seguem as formulações de Q para os métodos apresentados acima, considerando, por simplicidade, o problema de classificação binário e $y_i = 0$ para todo $x_i \in X_u$.

A função de custo minimizada pelo método corte mínimo fluxo máximo, definida para $f : f(x) \in \{-1, 1\}$, é dada por:

$$Q = \sum_{i:x_i \in X_l; j:x_j \in X_u} (y_i - f(x_j))^2 + \sum_{i,j=1}^{n} w_{ij}(f(x_i) - f(x_j))^2, \tag{2.2}$$

onde w_{ij} é dado pelo peso da aresta (v_i, v_j) e $y_i = 0$ se $x_i \in X_u$. O primeiro termo corresponde ao erro e o segundo à regularização. A função de custo minimizada pelo método campos Gaussianos e funções harmônicas também é dada pela Eq. 2.2, porém f é relaxado, $f : f(x) \in [-1, 1]$. Também definida para $f : f(x) \in [-1, 1]$, a função de custo minimizada pelo método consistência local e global é dada por:

$$Q = \sum_{i:x_i \in X_l; j:x_j \in X_u} (y_i - f(x_j))^2 + \mu \sum_{i,j=1}^{n} w_{ij} \left(\frac{f(x_i)}{\sqrt{d_i}} - \frac{f(x_j)}{\sqrt{d_j}} \right)^2, \tag{2.3}$$

onde $d_i = \sum_k w_{ik}$ e $\lambda = \alpha/1 - \alpha$. Notavelmente, de acordo com a definição de α, o termo de regularização pode ter uma contribuição muito maior que o termo de erro, dominando a função de custo.

2.4.5 Competição e cooperação de partículas

Todas as técnicas de aprendizado semissupervisionado baseado em redes apresentadas até o presenta momento são transdutivas. Como já considerado, o domínio de sua aplicação são os dados disponíveis durante o treinamento. Não existe nenhuma forma direta de predizer coerentemente os rótulos de dados ainda não vistos. Muitas vezes esse comportamento é indesejado. A presente subseção apresentará a técnica de competição e cooperação de partículas, desenvolvida em nosso grupo de pesquisa (Breve et al., 2012; Breve, 2010; Breve et al., 2010,

2009; Silva & Zhao, 2012). Embora também seja transdutiva, essa técnica oferece a possibilidade de continuar o processo de aprendizado ao introduzir novos dados, rotulados ou não rotulados, na rede, pois implementa uma dinâmica competitiva, que permite a redefinição de rótulos. Assim como os métodos campos Gaussiano e funções harmônicas e consistência local e global, esse método pode ser aplicado ao problema de classificação multiclasse sem a necessidade de utilizar matrizes de códigos.

O mecanismo de cooperação e competição implementado por esse método é inspirado no comportamento social de alguns animais em bando e no comportamento de seres humanos em jogos coletivos. Dada uma rede, possivelmente construída a partir de dados vetoriais como mostra a Seção 2.4.6, l partículas são inseridas na rede, cada uma das quais na posição de um dos l nós previamente rotulados. Cada nó rotulado é dito o nó-casa da respectiva partícula. Partículas associadas a nós que possuem um mesmo rótulo pertencem a um mesmo time. Cada nó da rede, por sua vez, possui um vetor de níveis de dominação, onde cada entrada corresponde ao nível de domínio de um time sobre o nó. Iterativamente, as partículas movem-se na rede de acordo com uma regra ora aleatória, ora gulosa e, ao visitar um nó, o nível de domínio do time da partícula atual é aumentado, enquanto que os níveis de domínio dos outros times é diminuído. Cada time de partículas tenta dominar a maior quantidade possível de nós e ao mesmo tempo evitar a invasão de partículas de outros times. Partículas de um mesmo time cooperam entre si, enquanto que os times de partículas competem entre si. Ao final do processo de competição e cooperação, cada nó é rotulado de acordo com o rótulo do time que possui maior nível de dominação sobre o mesmo.

Existem duas dinâmicas bem definidas no algoritmo: a dinâmica das partículas, e a dinâmica dos nós. Cada partícula ρ_j possui duas variáveis, $\rho_j^{\omega}(t)$ e $\rho_j^{d}(t)$. $\rho_j^{\omega}(t) \in [0,1]$ é o potencial de exploração da partícula. $\rho_j^{d}(t) = \{\rho_j^{d_1}(t), \ldots, \rho_j^{d_n}(t)\}$, por sua vez, é uma tabela de distâncias onde $\rho_j^{d_i}(t) \in [0, n-1]$ é a menor distância registrada do nó-casa da partícula até cada um dos nós da rede, mantida no sentido de guiar o movimento guloso. Cada partícula é inicializada com potencial máximo de exploração, $\rho_j^{\omega}(0) = 1$. Com exceção da distância do nó-casa de cada partícula a ele mesmo, que é zero, todas as entradas da tabela de distância são inicializadas com valor máximo, $n-1$,

$$\rho_j^{d_i}(0) = \begin{cases} 0 & \text{se } i = j \\ n-1 & \text{se } i \neq j \end{cases} . \tag{2.4}$$

$\rho_j^{d_i}(t+1)$ é atualizado iterativamente, à medida que a partícula visita os nós da rede. Esteja ρ_j movendo-se do nó v_i para o nó vizinho v_k,

$$\rho_j^{d_k}(t+1) = \begin{cases} \rho_j^{d_i}(t) + 1 & \text{se } \rho_j^{d_i}(t) + 1 < \rho_j^{d_k}(t) \\ \rho_j^{d_k}(t) & \text{caso contrário} \end{cases} . \tag{2.5}$$

Cada nó v_i, por sua vez, possui uma variável $v_i^{\omega}(t) = \{v_i^{\omega_1}(t), \ldots, v_i^{\omega_\kappa}(t)\}$ dita vetor de dominação, onde κ é o número de times de partículas ou classes do problema, e $v_i^{\omega_c}(t) \in [0,1]$ é o nível de domínio do time c sobre o nó v_i. Caso v_i seja um nó rotulado todos os níveis de

$v_i^{\omega}(0)$ são inicializados com valor nulo, exceto o nível correspondente ao time do rótulo de v_i, que recebe o valor máximo; caso contrário, todos os níveis de $v_i^{\omega}(0)$ são inicializados com valor $\frac{1}{\kappa}$, representando igual domínio de todos os times,

$$v_i^{\omega_c}(0) = \begin{cases} 1 & \text{se } y_i = c \\ 0 & \text{se } y_i \neq c \text{ e } y_i \in L \\ \frac{1}{\kappa} & \text{se } y_i = \emptyset \end{cases} . \tag{2.6}$$

Por padrão, à exceção dos nós previamente rotulados, os níveis de domínio de cada nó são atualizado sempre que o mesmo é visitado por uma partícula. Toda partícula aumenta o nível de domínio de seu time ao tempo que diminui os níveis de domínio dos demais times, como descrito pela seguinte equação:

$$v_i^{\omega_c}(t+1) = \begin{cases} \max\{0, v_i^{\omega_c}(t) - \frac{\Delta_v \rho_j^{\omega}(t)}{\kappa - 1}\} & \text{se } y_i = \emptyset \text{ e } c \neq \rho_j^{\tau} \\ v_i^{\omega_c}(t) + \sum_{q \neq c} v_i^{\omega_q}(t) - v_i^{\omega_q}(t+1) & \text{se } y_i = \emptyset \text{ e } c = \rho_j^{\tau} \\ v_i^{\omega_c}(t) & \text{se } y_i \in L \end{cases} , \tag{2.7}$$

onde $0 < \Delta_v \leq 1$ é um parâmetro externo de controle da agressividade do processo de competição e ρ_j^{τ} representa o time da partícula ρ_j. O ajuste do potencial da partícula, por sua vez, é dado pela seguinte equação:

$$\rho_j^{\omega}(t+1) = v_i^{\omega_c}(t+1), \tag{2.8}$$

onde c é o rótulo do time da partícula ρ_j, isto é, $c = \rho_j^{\tau}$. Quanto o nó visitado é dominado pelo time da partícula que o visita, o potencial da partícula é incrementado; no contrário é decrementado.

A cada iteração cada partícula tem probabilidade p_{grd} de realizar um movimento guloso e probabilidade $1 - p_{\text{grd}}$ de realizar um movimento aleatório, seja $0 \leq p_{\text{grd}} \leq 1$. Enquanto que o movimento aleatório representa o comportamento exploratório da partícula no sentido de dominar novos nós, o movimento guloso é responsável por restringir o movimento da partícula mantendo-a no seu território e garantindo a resistência à invasão de partículas de outros times. Definido o tipo de movimento, no movimento aleatório o nó para onde a partícula se move é escolhido com igual probabilidade dentre um dos vizinhos do nó atual. Esteja ρ_j em v_i,

$$p(v_k|\rho_j) = \frac{w_{ik}}{\sum_{\mu} w_{i\mu}}. \tag{2.9}$$

No movimento guloso, por outro lado, a probabilidade de escolha de um vizinho, v_k, é proporcional ao nível de domínio do time da partícula sobre ele, $v_k^{\omega_c}$, seja $c = \rho_j^{\tau}$, e à distância dos do nó-casa da partícula até ele, $\rho_j^{d_k}$. Esteja ρ_j em v_i,

$$p(v_k|\rho_j) = \frac{w_{ik} v_k^{\omega_c} \frac{1}{(1+\rho_j^{d_k})^2}}{\sum_{\mu} w_{i\mu} v_{\mu}^{\omega_c} \frac{1}{(1+\rho_j^{d_{\mu}})^2}}. \tag{2.10}$$

O Algoritmo 2.5 descreve o método de competição e cooperação de partículas.

Algoritmo 2.5 Competição e cooperação de partículas (Breve et al., 2012; Breve, 2010)

1: Construa uma matriz de afinidade $\mathbf{W}$ entre os itens de dados;
2: Inicialize l partículas com potencial máximo, $\rho_j^{\omega}(0) \leftarrow 1$;
3: Posicione as l partículas nos respectivos nós-casa;
4: Inicialize as tabelas de distâncias conforme a Eq. 2.4;
5: Inicialize os níveis de domínio conforme a Eq. 2.6;
6: **repita**
7: **para** cada partícula **faça**
8: Com probabilidade p_{grd} selecione entre movimento aleatório e guloso;
9: Selecione o nó alvo usando a Eq. 2.9 ou a Eq. 2.10;
10: Atualize os níveis de domínio do nó alvo usando a Eq. 2.7;
11: Atualize o potencial da partícula usando a Eq. 2.8;
12: Atualize a tabela de distância da partícula usando a Eq. 2.5;
13: **fim-para**
14: **até que** que o critério de parada seja satisfeito;
15: Rotule cada item de dado não rotulado com o maior nível de domínio, $\arg\max_c v_i^{\omega_c}(\infty)$.

Segundo Breve et al. (2012), os nós em fronteiras de classes frequentemente possuem níveis de dominação instáveis, alternando entre os times que o dominam. Logo, a convergência dos rótulos pode frequentemente não ser um bom critério de parada para o algoritmo. Os autores sugerem o monitoramento da média dos maiores níveis de domínio dos nós ($v_i^{\omega_c}$, $c = \arg\max_q v_i^{\omega_q}$) e o encerramento da execução quando não houver incremento significativo nesse valor. Os rótulos preditos podem ser apresentados crisp, como no Algoritmo 2.5 ou de forma suave, de acordo com os níveis de dominação, $y_i = v_i^{\omega}$. Além disso, a correção de nós previamente rotulados de forma inconsistente pode ser feita permitindo-se a alteração dos níveis de dominação dos nós-casa.

2.4.6 Construção do grafo

Embora algoritmos de aprendizado semissupervisionado baseado em redes sejam tipicamente propostos associados a um método de construção da rede, em geral, quaisquer métodos de construção de rede podem ser aplicados a quaisquer algoritmos. A boa definição de uma abstração gráfica geralmente é mais crítica do que a propagação de rótulos (Zhu, 2008; Jebara et al., 2009). A partir de dados vetoriais, tanto abordagens supervisionadas como não supervisionadas têm sido estudadas. Métodos de construção de grafos supervisionados assumem que os dados disponíveis rotulados foram confiavelmente rotulados e buscam tirar vantagem disso, tornando-os nós centrais nas redes (Rohban & Rabiee, 2012; Berton & de Andrade Lopes, 2014). Métodos não supervisionados, por outro lado, buscam tirar vantagem da descrição das distribuições dos dados, suportados apenas pelas suposições de suavidade, grupos e variedades. Como um dos objetivos do presente trabalho é tirar vantagem da abstração gráfica para realçar itens de dados representativos, apenas métodos não supervisionados são estudados.

Três métodos de construção de redes são amplamente utilizados na literatura: redes totalmente conectadas com pesos, k-vizinhos-mais-próximos e ϵRadius. Suas formulações são apresentadas a seguir.

Seja $\hat{\mathbf{W}}$ uma matriz de pesos com *kernel* Gaussiano, cujas entradas w_{ij} são dadas por

$$\hat{w}_{ij} = \hat{w}_{ji} = \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right), \tag{2.11}$$

onde σ é um parâmetro de largura de banda e $\|\cdot\|$ é uma função de distância entre vetores, por exemplo,

$$\|x_i - x_j\| = \left(\sum_{k=1}^{D}(x_{ik} - x_{jk})^p\right)^{\frac{1}{p}}, \tag{2.12}$$

onde D define o número de dimensões dos dados (ou seja, o número de atributos) e p é o parâmetro norma. Para $p = 2$, por exemplo, temos a norma l_2 (ou distância Euclidiana). Outras medidas de distância como cosseno e correlação linear também podem ser definidas. A matriz de adjacências generalizada, $\mathbf{W}$, é dada por

$$w_{ij} = w_{ji} = \max(p_{ij}, p_{ji})\hat{w}_{ij}, \tag{2.13}$$

onde, para a rede totalmente conectada com pesos,

$$p_{ij} = p_{ji} = \begin{cases} 1 & \text{se } i \neq j \\ 0 & \text{se } i = j \end{cases}. \tag{2.14}$$

Embora possa parecer interessante manter a informação de afinidade entre todos os pares de itens de dados, o alto custo computacional envolvido no tratamento de redes densamente conectadas pode tornar-se limitante ao desempenho dos métodos de propagação de rótulos. Por isso, abordagens de construção de rede esparsas são muito mais frequentemente usadas. Em geral elas são definidas em dois passos: esparsificação e ponderação das arestas.

Seja $\mathbf{W}$ definido conforme a Eq. 2.13, segundo o método k-vizinhos-mais-próximos

$$p_{ij} = \begin{cases} 1 & \text{if } x_j \in \ k\text{-vizinhos-mais-próximos de } x_i \\ 0 & \text{caso contrário} \end{cases}, \tag{2.15}$$

onde os k-vizinhos-mais-próximos de x_i são ranqueados considerando também uma função de distância entre vetores. Menos usual, a simetria pode ser interessante em alguns casos,

$$w_{ij} = w_{ji} = \min(p_{ij}, p_{ji})\hat{w}_{ij}. \tag{2.16}$$

Nesse caso, o método é denominado k-vizinhos-mais-próximos mútuo. Para o método ϵRadius, por sua vez,

$$p_{ij} = p_{ji} = \begin{cases} 1 & \text{if } \|x_i - x_j\| <= \epsilon \text{ e } i \neq j \\ 0 & \text{if } \|x_i - x_j\| > \epsilon \text{ ou } i = j \end{cases}, \tag{2.17}$$

onde ϵ define o raio de cobertura dentro do qual ligações são definidas.

A escolha dos parâmetros σ, k e ϵ é usualmente arbitrária. Como o parâmetro σ determina a suavidade do kernel Gaussiano, ele deve ser definido de acordo com a distribuição de densidade

dos dados, maior quão esparso forem as distribuições. Notavelmente, alguns autores abordam diretamente essa questão (Kapoor et al., 2005; Zhang & Lee, 2006). Pequenos valores de ϵ ou k, por sua vez, resultam em grafos bastante esparsos, porém com múltiplos componentes, o que também não é desejado em se tratando de abstrações o aprendizado semissupervisionado, pois será requerido maior número de instâncias rotuladas *a priori* com fins à cobertura de todos os componentes. Por outro lado, valores grandes de ϵ ou k resultam em grafos densos que podem não representar bem a distribuição dos dados. Bons métodos de construção de redes devem resultar em redes esparsas porém conectadas e cuja estrutura se ajuste bem a distribuição dos dados e se possível a distribuição das classes.

Métodos de construção de redes do estado-da-arte incluem b-matching (Jebara et al., 2009), vizinhança linear (Wang & Zhang, 2008), combinações de árvores geradoras mínimas perturbadas e disjuntas (Carreira-perpiñán & Zemel, 2004) e métodos baseados em no *single linkage* (Cupertino et al., 2013). Em comparação com o método kNN, b-matching fortalece ainda mais a regularidade da rede restringindo o número de ligações de cada vértice a um valor específico, b, tomando como suporte o fato de que a maioria dos métodos de propagação de rótulos podem ser definidos em termos do *framework* de regularização. Vizinhança Linear, por sua vez, normaliza os pesos das arestas para cada nó de acordo com o erro de reconstrução linear (Roweis & Saul, 2000). Por outro lado, as propostas de Carreira-perpiñán & Zemel (2004) procuram superar a sensibilidade do método de árvores geradoras mínimas a *outliers* combinando árvores geradoras mínimas construídas sobre dados perturbados com algum ruido. As redes resultantes são tão esparsas quanto as árvores geradoras mínimas. Cupertino et al. (2013), por fim, propõem uma abordagem mais elaborada. A partir de grafos nulos (sem arestas) até a obtenção de um grafo conectado, iterativamente, os dois módulos mais próximos são unidos através das k-menores-arestas entre eles cujos comprimentos não são maiores que um limiar, d_{thr}, γ-proporcional à maior entre as distâncias médias intra-módulo de cada módulo.

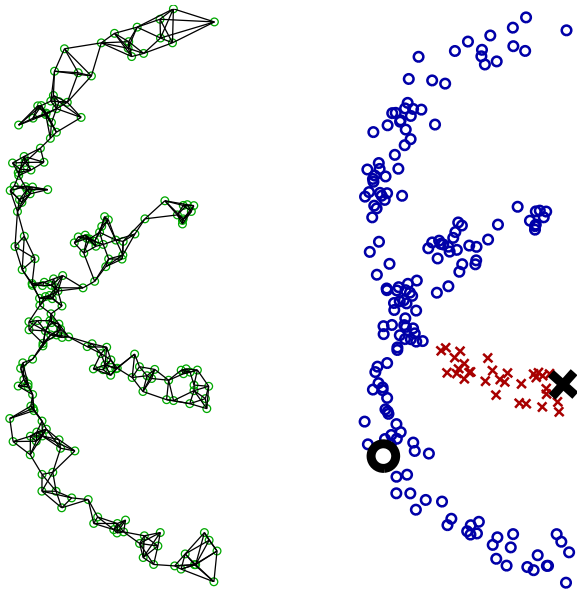
Embora elegantes, esses método ou se apoiam em suposições fortes, como regularidade ou esparsidade, ou procuram modelar propriamente as distribuições dos dados ao custo da definição de diversos parâmetros externos. No presente trabalho propomos um métodos de construção de redes a partir de dados vetoriais baseado no método ϵRadius e em árvores geradoras mínimas, que apresenta diversas vantagens, como baixa dependência em relação à definição de parâmetros e baixo custo computacional tanto sobre dados *pool-based* como dados incrementais. Vide Capítulo 7. Também, uma métrica para a avaliação de métodos de construção de redes é proposta no Capítulo 8.

2.4.7 Desafios

Técnicas de aprendizado semissupervisionado baseados em redes tipicamente são transdutivas. Seus algoritmos realizam a rotulação de dados não rotulados sem induzir um classificador. Logo, a classificação de itens de dados não disponíveis durante a execução do algoritmo não é direta, sendo frequentemente necessária a reexecução do algoritmo sempre que novos dados são adicionados. Isso constitui um desafio para essa abordagem, pois algoritmos baseados em redes

são frequentemente custosos, pelo menos $O(n^2)$ (Zhu et al., 2009).

Recentemente, algumas técnicas de **aprendizado *on-line*** foram introduzidas . Essas técnicas baseiam-se em processos dinâmicos que ocorrem sobre redes, tais como difusão, competição de partículas e sincronização. Não prejudicam o processo de aprendizado pois são capazes de adaptar-se a pertubações incrementais na rede, encontrando novos pontos de estabilidade (baixa entropia). Para que estes métodos funcionem incrementalmente, no entanto, novos vértices e arestas devem ser incluídos/removidos na rede de forma adequada, levando em consideração o modelo gerador da rede, permanecendo inalteradas suas características. Nesse caso, a fase de construção da rede é ainda mais relevante.



(a) (b)

Figura 2.5: Conjunto de dados em forma de bananas com variedades sobrepostas: (a) rede gerada a partir da similaridade entre os objetos; (b) classificação semissupervisionada. Nesse caso, a abordagem baseada em redes mostra-se incapaz de separar adequadamente as classes. Figura reproduzida de (Zhu & Goldberg, 2009) com permissão da Morgan & Claypool Publishers.

Zhu et al. (2009) também apontam outros desafios para a abordagem baseada em redes, tais como: 1) lidar com a **sobreposição de variedades** de classes distintas, e 2) lidar com dados que, embora satisfaçam a suposição de grupos, não satisfazem a suposição de suavidade. Considere o problema ilustrado na Figura 2.5. Nele duas variedades de dados estão parcialmente sobrepostas. A menos que exista um número significativo de dados previamente rotulados, a separação das classes na região de sobreposição é um desafio. Goldberg et al. (2009) propuseram uma solução para esse problema utilizando a distância de *Hellinger* e o método k-Mahalanobis-vizinhos-mais-próximos na construção da rede. Abordagens alternativas podem ser propostas utilizando hipergrafos. Nesse caso, hiperarestas poderiam ser construídas levando em consideração o alinhamento de triplas de objetos, permitindo a representação do alongamento da distribuição dos dados nas regiões de sobreposição (Zhou et al., 2007).

Tanto suas vantagens como seus desafios são motivações para nosso interesse em abordagens baseadas em grafos. Mais justificativas serão apresentadas no Capítulo 4.

2.5 Validação e referência de desempenho de ASS

Como referência de desempenho de aprendizado semissupervisionado, Chapelle et al. (2006) disponibilizam um compêndio de resultados obtidos por métodos de classificação semissupervisionada em bases de dados reais e sintéticas. Um resumo sobre os conjuntos de dados é apresentado na Tabela 2.1. Descrições detalhadas podem ser encontradas em (Chapelle et al., 2006), Capítulo 21. São considerados os métodos listados na Tabela 2.2, cujo desempenho é verificado em termos do erro de predição dos rótulos das amostras não rotuladas, denominado erro de teste.

Tabela 2.1: Descrição resumida dos conjuntos de dados utilizados no *benchmark* de Chapelle et al. (2006)

	# classes	# dimensões	tamanho	características
g241c	2	241	1500	artificial
g241d	2	241	1500	artificial
Digit1	2	241	1500	artificial
USPS	2	241	1500	desbalanceado, $\lambda = 0.8$
COIL	6	241	1500	
BCI	2	117	400	

Tabela 2.2: Descrição resumida dos métodos de ASS utilizados avaliados no *benchmark* de Chapelle et al. (2006)

Abreviação	Técnica	Ref(s).
MVU + 1-NN	Maximum variance unfolding	(Weinberger & Saul, 2006), (Sun et al., 2006)
LEM + 1-NN	Laplacian eigenmaps	(Belkin & Niyogi, 2003)
QC + CMR	Quad. criterion and class mass reg.	(Belkin et al., 2004), (Delalleau et al., 2005)
Discrete reg.	Discrete regularization	(Zhou & Schölkopf, 2006)
TSVM	Transductive support vector machines	(Joachims, 2003; Chapelle & Zien, 2005)
SGT	Spectral graph transducer	(Joachims, 2003)
Cluster-Kernel	Cluster kernels	(Chapelle et al., 2002)
Data-Dep. Reg.	Data-dependent regularization	(Corduneanu & Jaakkola, 2006)
LDS	Low-density separation	(Chapelle & Zien, 2005)
Laplacian RLS	Laplacian regularized least squares	(Smith et al., 2002)
CHM (normed)	Conditional harmonic mixing	(Burges & Platt, 2006)

Considere o problema de classificação binário. Podemos definir uma matriz de confusão entre as classes positivo (P) e negativo (N) preditas e conhecidas antecipadamente,

	P	N
P	VP	FN
N	FP	VN

onde VP, FN, FP e VN registram, respectivamente, a quantidade de instâncias da classe P preditas com o rótulo P, da classe P preditas com o rótulo N, N com o rótulo P, e N com o rótulo N. O erro de teste é definido como

$$\text{ERR} = \frac{\text{FP} + \text{FN}}{n}, \tag{2.18}$$

onde $n = \text{VP} + \text{VN} + \text{FP} + \text{FN}$.

Além do erro de teste, diversas outras medidas de validação são definidas na literatura. A seguir são apresentadas duas métricas de validação frequentemente utilizadas na validação de agrupamento de dados, uma externa e outra relativa. Essas medidas são bastante interessantes pois permitem verificar a satisfação das suposições de suavidade e de agrupamento, respectivamente. Em se tratando da validação do aprendizado semissupervisionado baseado em redes, métricas específicas não são conhecidas.

2.5.1 Índice de Rand ajustado

O **índice de Rand ajustado** (ARI) é originalmente uma medida externa de validação de agrupamento (Hubert & Arabie, 1985; Jain & Dubes, 1988). Trata-se de uma extensão do índice de Rand (Rand, 1971) independente do número de grupos do conjunto de dados. Contudo pode ser utilizada como medida de validação de classificação semissupervisionada tomando como referência rótulos conhecidos *a priori*, conforme a seguinte descrição.

Considere uma tabela de contingência entre uma partição dos dados criada de acordo com os rótulos originais dos dados, $P = \{P_1, \ldots, P_k\}$, e uma partição dos dados de acordo com a classificação semissupervisionada, $Q = \{Q_1, Q_2, \ldots, Q_l\}$,

P \ Q	Q_1	Q_2	$\cdots$	Q_l	
P_1	n_{11}	n_{12}	$\cdots$	n_{1l}	a_1
P_2	n_{21}	n_{22}	$\cdots$	n_{2l}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
P_k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	a_k
	b_1	b_2	$\cdots$	b_l	

onde a_i denota o número de objetos no grupo P_i, b_j denota o número de objetos no grupo Q_j, e n_{ij} denota o número de objetos em comum entre os grupos P_i e Q_j, isto é, $n_{ij} = |P_i \cap Q_j|$. O ARI é dado por:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_i}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}\right] / \binom{N}{2}}. \tag{2.19}$$

Essa medida varia no intervalo $[-1, 1]$, aumentando de acordo com a correspondência entre a predição e os rótulos conhecidos antecipadamente. Baseado na expectância estatística, o ARI apresenta uma interpretação bastante justa, tornando desnecessária a realização de testes de significância estatística. A hipótese nula de rótulos aleatórios equiprováveis resulta $\text{ARI} = 0$ e a correspondência exata entre predição e rótulos conhecidos antecipadamente, $\text{ARI} = 1$.

Outras medidas de validação externa são informação mútua normalizada, índice de Jaccard e outras baseadas na matriz de confusão de verdadeiros e falsos positivos e negativos. Uma

revisão mais extensa sobre essas métricas pode ser encontrada em (Faceli et al., 2011).

2.5.2 Silhueta

Em contraste com o ARI, apresentamos também a **silhueta** (Rousseeuw, 1987). Originalmente, a silhueta é uma medida de validação relativa de partições de dados. O rótulo original dos dados não é levado em consideração na avaliação de desempenho. Tomando como referência a matriz de dissimilaridades entre os itens de dados e uma partição dos dados, a silhueta afere quão bem ajustados estão os dados aos respectivos grupos utilizando como viés a distância média aos itens de dados de cada grupo. Especificamente, a medida é enviesada pela medida de distância utilizada para caracterizar os grupos de interesse, sendo necessário que está descreva bem os relacionamentos aferidos.

Considere um conjunto de dados particionado em grupos. Para cada objeto i, $a(i)$ é a distância média de i a cada um dos demais objetos no mesmo grupo. $a(i)$ significa o quão bem situado está i no seu grupo. Quanto menor o valor de $a(i)$ melhor o ajuste. Considerando cada um dos grupos de que i não é membro, é aferida a distância média de i a seus objetos; a menor dessas distâncias médias é denotada por $b(i)$. O grupo com menor dissimilaridade média de que i não é membro, é dito o ''grupo vizinho''. $b(i)$ significa quão bem i poderia estar situado se pertencesse ao grupo vizinho. A silhueta de um item de dado i é dada por:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{2.20}$$

o que também pode ser escrito como:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{se } a(i) < b(i) \\ 0, & \text{se } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{se } a(i) > b(i) \end{cases} . \tag{2.21}$$

$-1 \leq s(i) \leq 1$. Para que o valor de $s(i)$ seja próximo de 1 é necessário que $a(i) \ll b(i)$. Valores de $s(i)$ próximos a zero, por sua vez, significam que o respectivo item de dado i situa-se na borda entre dois grupos. A média de $s(i)$ para todo i em um determinado grupo, informa quão bem definido é o grupo. E logo, a média de $s(i)$ para todos o itens de dados do conjunto de dados informa quão apropriada é a partição dos dados.

Destaque-se ainda que não necessariamente grupos são consistentes com classes, portanto, para que a silhueta seja uma medida apropriada à validação do aprendizado semissupervisionado, é necessário que a suposição de grupos seja satisfeita. Descrições e comparações entre diversas medidas de validação relativa de agrupamento podem ser encontradas em (Vendramin et al., 2010).

Capítulo

3

Amostragem Aleatória

Originalmente o aprendizado semissupervisionado foi concebido com o objetivo de tirar vantagem das informações inerentes aos itens dados não rotulados quando poucos itens de dados estão disponíveis rotulados. Em alguns cenários, conjuntos de dados são de fato encontrados dessa maneira, com pouco itens de dados rotulados e muitos itens de dados não rotulados. Muito mais frequentemente, no entanto, conjuntos de dados precisam ser analisados estando disponíveis sem quaisquer rótulos associados. O processo de seleção e rotulação de itens de dados requer de fato a atuação de especialistas no domínio. Contudo, nem mesmo assim, pode-se garantir que todo o conjunto de dados seja devidamente avaliado no sentido de selecionar uma amostra representativa dos dados para a rotulação manual. Embora a escolha de uma amostra representativa seja de interesse, em geral, o volume dos dados torna essa tarefa proibitiva. Tão logo, apenas uma pequena parcela dos dados é analisada de fato. Nessa seção analisaremos a aptidão do processo de amostragem aleatória na seleção de amostras de dados suficientemente representativas.

3.1 Limite de Chernoff

Guha et al. (1998, 2001) apresentam uma maneira de estimar o número de itens de dados, s, que precisam ser selecionados, através de amostragem aleatória, para representar grupos de interesse em um conjunto de dados. Essa análise é feita sobre o limite de Chernoff (Motwani & Raghavan, 1995).

Sejam $X_1, X_2, \ldots, X_s$ variáveis aleatórias independentes que representam eventos de seleção de itens de dados ao acaso, onde s é o tamanho da amostra, $X_j \in \{0, 1\}$; $X_j = 1$ se o j-ésimo item de dado amostrado pertence ao grupo u; caso contrário, $X_j = 0$. Definimos $\Pr[X_j = 1] = p_j$ e $\Pr[X_j = 0] = 1 - p_j$. A esperança de X, $\mu = E[X] = E[\sum_{j=1}^{s} X_j] = \sum_{j=1}^{s} E[X_j]$, onde

$E[X_j] = X_j \Pr[X_j] = 1 \times \Pr[X_j = 1] + 0 \times \Pr[X_j = 0] = p_j$, logo $\mu = sp_j$. A probabilidade de um item de dado amostrado ao acaso pertencer ao grupo u é dada por $p_j = \frac{|u|}{n}$, onde n é o número total de instâncias do conjunto de dados e $|u|$ o número de instâncias pertencentes ao grupo u. Seja $f|u|$, $0 \leq f \leq 1$, o número de instâncias do grupo u suficientes à sua representação, queremos saber o tamanho da amostra aleatória suficiente à descrição de u. Seja $1 - \delta$ o nível de certeza da inferência, $\Pr[X \geq f|u|] \geq (1 - \delta)$, ou, em termos da incerteza,

$$\Pr[X < f|u|] < \delta. \tag{3.1}$$

Os limites de Chernoff definem para $0 < \epsilon \leq 1$, seja $X_1, X_2, \ldots, X_s$ uma distribuição de Poisson,

$$\Pr[X < (1 - \epsilon)\mu] < e^{-\frac{\mu\epsilon^2}{2}}. \tag{3.2}$$

Reescrevendo a Eq. 3.1 de acordo com a Eq. 3.2, temos

$$\Pr[X < (1 - (1 - \frac{f|u|}{\mu}))\mu] < \delta.$$

Logo,

$$e^{-\frac{\mu(1-\frac{f|u|}{\mu})^2}{2}} \leq \delta.$$

Substituindo $\mu = \frac{s|u|}{n}$ na equação acima, e resolvendo-a para s,

$$s \geq fn + \frac{n}{|u|}\ln(\frac{1}{\delta}) + \frac{n}{|u|}\sqrt{(\ln(\frac{1}{\delta}))^2 + 2f|u|\ln(\frac{1}{\delta})}. \tag{3.3}$$

Concluímos portanto que, para que um amostra aleatória contenha pelo menos $f|u|$ itens de dados pertencentes ao grupo u (com alta probabilidade), precisamos que a amostra tenha tamanho superior à fração f do total de itens de dados. À primeira vista podemos pensar que, em razão do tamanho do conjunto de dados, s varia; quanto maior o tamanho do conjunto de dados, maior a amostra necessária. Contudo, isso não é verdade. Mantidas as distribuições dos grupos, se incrementamos o conjunto de dados, também são aumentadas as densidades dos grupos, e portanto menor f para uma amostra suficiente à descrição do grupo u. O principal fator determinante do tamanho da amostra é o tamanho dos grupos de interesse.

Obviamente, em se tratando de conjuntos de dados reais, grupos pequenos não são de interesse. Estes frequentemente são definidos como *outliers*. Portanto, definimos o tamanho do menor grupo de interesse proporcional ao tamanho médio dos grupos. Seja k o número de grupos do problema e u_{min} o menor grupo de interesse, definimos $|u_{min}| = \frac{n}{k\rho}$, onde $\rho \geq 1$. Além disso, redefinimos f em razão do número de instâncias ξ necessárias à representação do menor grupo de interesse, $f = \frac{\xi}{|u_{min}|}$; uma amostra aleatória, nesse caso, terá $\frac{\xi|u|}{|u_{min}|}$ instâncias de um grupo u qualquer. Logo, podemos definir (com alta probabilidade), em razão do tamanho do menor grupo de interesse, o tamanho da amostra mínima, s_{min}, necessária a cobertura adequada de todos os grupos do conjunto de dados,

$$s_{min} = \xi k\rho + k\rho \ln(\frac{1}{\delta}) + k\rho\sqrt{(\ln(\frac{1}{\delta}))^2 + 2\xi \ln(\frac{1}{\delta})}. \tag{3.4}$$

Seguem alguns exemplos. Dado um conjunto de dados com $k = 2$ grupos, o menor com metade do tamanho médio dos grupos ($\rho = 2$) e satisfatoriamente representável por $\xi = 5$ instâncias. Com confiança $1 - \delta = 0.99$, o tamanho da amostra capaz de representar bem todos os grupos de interesse é dada por $s_{min} = 72$; diminuindo a confiança da inferência, $1 - \delta = 0.90$, $s_{min} = 51$; considerando grupos de tamanhos balanceados ($\rho = 1$), $s_{min} = 26$. Seja $k = 4$, $\rho = 2$, $\xi = 5$ e $\delta = 0.01$, $s_{min} = 143$; $k = 10$, $\rho = 2$, $\xi = 10$ e $\delta = 0.01$, $s_{min} = 505$.

Embora concebidos para o problema da amostragem de grandes quantidades de dados na tarefa de agrupamento de dados, podemos considerar esses resultados factíveis com o problema estudado no presente trabalho. Satisfeita a suposição de grupos, a amostra que semeará o processo de rotulação semissupervisionado pode ser selecionada através de amostragem aleatória. Sua representatividade está estritamente relacionada ao seu tamanho e cobertura apropriada dos dados. Contudo, tenhamos em vista que, diferente do problema de amostragem para o agrupamento de dados, itens de dados não selecionados para rotulação prévia são incluídos no modelo como itens de dados não rotulados, contribuindo ainda assim para o aprendizado semissupervisionado.

Capítulo

4

Redes Complexas

Redes complexas é o nome dado às redes que modelam sistemas interconectados de grande escala com padrões de ligação não trivial. Desde sua concepção por Euler (Biggs et al., 1976), não se via uma mudança tão significativa no estudo da abstração gráfica. Motivado pela grande disponibilidade de recursos computacionais e pela evolução das redes de comunicação, as redes passaram a ser estudadas de uma maneira muito mais extensiva. Com os recursos existentes nos séculos anteriores, era possível apenas a análise de redes da ordem de dezenas, quando muito, centenas de nós. Hoje, por outro lado, frequentemente são analisadas redes da ordem de milhões e até mesmo bilhões de nós. Mais ainda, como considerado por Costa et al. (2007), houve uma mudança do ponto de vista da visualização. Enquanto que em redes com dezenas de nós era relativamente simples desenhar um grafo e responder questões acerca da sua estrutura observando sua representação, em redes da ordem de milhares ou milhões de nós, essa metodologia é impraticável. Por isso, soluções exatas deixaram de ser o alvo, para dar lugar a propriedades e dinâmicas complexas e soluções aproximadas. Modelagens realistas minuciosas têm sido substituídas por abstrações que preservam apenas as características relevantes dos problemas estudados. Acredita-se, inclusive, que muitas das questões que não foram respondidas, ou foram respondidas apenas parcialmente, através da ciência reducionista, possam ser melhor examinadas e eventualmente elucidadas do ponto de vista da ciência do complexo (Costa et al., 2007).

Especificamente, três tópicos protagonizaram os avanços recentes em redes complexas: a descoberta do efeito de pequeno mundo (Watts & Strogatz, 1998), a caracterização de redes livres de escala (Barabási & Albert, 1999), e a identificação de estruturas de comunidades (Girvan & Newman, 2002). Sistemas complexos, como a Internet (Faloutsos et al., 1999), a *World Wide Web* (Albert et al., 1999), redes de distribuição de energia elétrica (Albert et al., 2004), redes de colaboração entre cientistas (Newman, 2001), redes de interações sociais entre indivíduos (Scott, 2000), redes de interações entre companhias e organizações (Mizruchi, 1982), cadeias de

contatos sexuais (Liljeros et al., 2001), cadeias alimentares (Montoya & Solé, 2002), cadeias de interações metabólicas (Jeong et al., 2000), o sistema de distribuição da corrente sanguínea (West et al., 1999), redes neurais biológicas (Sponrs, 2002; Bullmore & Sporns, 2009), entre muitos outros, são alguns exemplos de aplicações que têm sido estudadas sob essa nova abordagem. As redes complexas têm se consolidado como um tópico unificador em sistemas complexos (Bornholdt & Schuster, 2003).

4.1 Definições gerais e notações

Como representado por Euler os principais elementos de um grafo (ou rede) são os vértices e as arestas. Um grafo pode ser formulado como $G = (V, E)$, onde V é o conjunto de vértices e E o conjunto de arestas do grafo. Sempre que um grafo possuir arestas direcionadas, i.e. cuja ligação representada se dá em apenas um sentido, o grafo é dito um **dígrafo**. Quando as arestas possuem pesos associados, por sua vez, o grafo é dito **ponderado**. Considere, então, um dígrafo ponderado, G, definido pelo par ordenado (V, E). V consiste em um conjunto de n vértices e E em um conjunto de m arestas. Cada nó pode ser identificado por um valor inteiro ordenado, $i = 1, 2, \ldots, n$. Cada ligação de um nó i até um nó j, pode ser identificada por um par (i, j), que possui um peso associado, identificado por $\omega(i, j)$, $\omega : E \mapsto \mathbb{R}$. No caso de uma rede não ponderada, as ligações não possuem peso associado, portanto $\omega = 1$ para todas as arestas. No caso de um grafo não direcionado (com ou sem pesos), as ligações se dão em ambos os sentidos. Portanto, se $(i, j) \in E$, significa que existe uma ligação tanto de i até j, como também de j até i.

De outra forma, uma rede também pode ser completamente definida em termos de uma matriz ponderada $\mathbf{W}$, denominada **matriz de adjacências generalizada**. Cada entrada w_{ij} de $\mathbf{W}$ representa o peso da respectiva ligação de i até j, $w_{ij} = \omega(i, j)$. Se G é uma rede não direcionada, então $\mathbf{W}$ é uma matriz simétrica, i.e., $\mathbf{W} = \mathbf{W}^{\intercal}$. Seja G uma rede sem pesos, se existe uma ligação de i até j, então w_{ij} assume valor 1; caso contrário, assume valor 0. Nesse caso $\mathbf{W}$ é dita uma **matriz de adjacências**, tipicamente representada por $\mathbf{A}$.

Seja G uma rede direcionada. Se $a_{ij} \neq 0$, diz-se que j é vizinho de (ou adjacente a) i. Seja G uma rede não direcionada. Se $a_{ij} \neq 0$, diz-se que i e j são **adjacentes**, ou vizinhos. Se todos os nós de G forem adjacentes uns aos outros, $m = n(n-1)/2$ e a rede é dita **totalmente conectada**, ou **completa**. A **vizinhança** de um vértice i é dada pelo conjunto $v(i)$ de todos os nós adjacentes a i. O **grau** de um nó i, k_i, é dado pelo número de vizinhos de i, isto é, a cardinalidade de $v(i)$,

$$k_i = \sum_{j=1}^{n} a_{ij}.$$

O **grau médio** de uma rede,

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij},$$

i.e. a média aritmética de k_i, $\forall i \in V$. Consideremos uma rede direcionada, novamente. O grau de i, k_i, é igual a $k_i^{in} + k_i^{out}$, onde

$$k_i^{out} = \sum_{j=1}^{n} a_{ij},$$

dito o **grau de saída**, contempla arestas saindo de i e incidindo sobre $v(i)$, e

$$k_i^{in} = \sum_{j=1}^{n} a_{ji},$$

dito o **grau de entrada**, contempla arestas saindo de $v(i)$ e incidindo sobre i.

Considere uma **rede esparsa**, i.e., onde apenas uma pequena fração do total de possíveis ligações de fato existe, $m = O(n)$. Sejam i e j dois nós não adjacentes. Embora não sejam vizinhos, i e j podem estar conectados através de uma sequência de l arestas, $(i, k_1), (k_1, k_2), \ldots, (k_{l-1}, j)$. Essa sequência é dita um **passeio** de i até j, cujo **comprimento** é l. Dizemos que dois nós estão **conectados**, se existe pelo menos um passeio conectando-os. Se todos os nós em um passeio são distintos, adicionalmente, o passeio é dito um **caminho**. O comprimento do menor caminho entre dois nós, i e j, é dito **distância geodésica**, d_{ij}. A **distância média** de uma rede, por sua vez, é dada pela média aritmética da distância geodésica entre todos os pares de nós da rede,

$$\langle d \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}, \tag{4.1}$$

$\forall (i, j)$, restrito a $i \neq j$. A maior distância geodésica entre pares de nós em uma rede é dita o **diâmetro** da rede. Se o passeio começa e termina no mesmo nó e passa uma única vez por cada nó, o passeio é dito um **ciclo**. Se o passeio começa e termina no mesmo nó e passa uma única vez por cada aresta, o passeio é dito um **circuito**. Se, adicionalmente, um ciclo passa por todos os n nós da rede, ele é dito um **ciclo Hamiltoniano**. Se um circuito passa por todas as m arestas de uma rede, ele é dito um **circuito Euleriano**.

Sejam $G_1 = (V_1, E_1)$ e $G_2 = (V_2, E_2)$ duas redes. Se $V_2 \subseteq V_1$ e $E_2 \subseteq E_1$, G_2 é dito uma **subrede** (ou subgrafo) de G_1. Se, além disso, todos os nós de G_2 estiverem conectados entre si, G_2 é dito um **componente conexo** de G_1. Uma rede pode possuir um ou mais de um componente conexo, sendo aquele que possui o maior número de vértices dito **componente principal**. Caso, adicionalmente, o número de vértices no componente principal seja $O(n)$, ele é dito um **componente gigante**. Se $V_2 = V_1$ e $E_2 \subseteq E_1$, G_2 é dito um subgrafo gerador de G_1 e G_1 é dito um subgrafo induzido de G_2. Se G_2 é um subgrafo gerado de G_1 e é conexo e não possui ciclos, G_2 é dito uma árvore geradora de G_1. Se, adicionalmente, G_2 é ponderado e

$$w(G_2) = \sum_{(i,j) \in E_2} w(i, j)$$

é mínimo, G_2 é dito uma árvore geradora mínima de G_1.

4.2 Caracterização de indivíduos

O presente trabalho tem como um de seus principais objetivos propor critérios para seleção de nós representativos em uma rede. Uma abordagem utilizada nesse sentido é caracterizar seus indivíduos através de medidas de centralidade. Algumas das medidas propostas a esse fim caracterizam seus indivíduos de forma local, outras em relação à região onde estão situados, e outras em relação a toda a rede. Kang et al. (2011) categorizam-nas em três grupos: medidas relacionadas ao grau, medidas relacionadas ao diâmetro, e medidas relacionadas ao fluxo. Além dessas medidas, medidas hierárquicas também são apresentadas a seguir. Descrições mais extensivas podem ser encontradas em (Koschützki et al., 2005; Costa et al., 2007; Rubinov & Sporns, 2010; Costa & Silva, 2006). Além disso, implementações estão disponíveis para MATLAB nos pacotes Brain Connectivity Toolbox[1] e Matlab Boost Graph Library[2].

4.2.1 Medidas relacionadas ao diâmetro

Nessa seção serão apresentados alguns critérios baseados no problema de localização de estabelecimentos (Koschützki et al., 2005). Esses critérios são divididos de acordo com a sua função objetivo. O primeiro critério, denominado **excentricidade**, visa otimizar a função *min-max*, que corresponde à solução de interesse para a localização de um estabelecimento de emergência, tal como um hospital. Deseja-se que a distância entre quaisquer localizações e o estabelecimento de emergência jamais seja suficientemente grande a ponto de que ocorrências não sejam atendidas em tempo. Nesse caso, o problema é identificar um local que minimize a distância máxima entre a instalação e os locais das ocorrências. O segundo critério, dito **proximidade**, visa otimizar a função *min-sum*, que corresponde à solução do problema de localização de um estabelecimento de serviço, como, por exemplo, um *shopping*. O objetivo, nesse caso, é minimizar o tempo médio de trânsito, de modo que o serviço seja localizado próximo da maior parte das localizações. O terceiro critério, por fim, baseia-se na localização de um estabelecimento comercial que opera em um ambiente competitivo. Nesse caso, o objetivo é localizar o estabelecimento em um local em que seja possível capturar a maior quota de mercado dos estabelecimentos concorrentes. Trata-se de um critério mais sofisticado, baseado na teoria dos jogos, denominado **valores centroides**.

4.2.1.1 Excentricidade

Seja $G = (V, E)$ uma rede, considere um nó $u \in V$. A excentricidade de u, é dada por $e(u) = \max\{d_{uv} : v \in V\}$, seja d_{uv} é a distância geodésica de u a v. Segundo a excentricidade, os nós menos excêntricos em uma rede são os mais centrais. Hage & Harary (1995) definem um índice de centralidade nesse caso,

[1] Em https://sites.google.com/site/bctnet/.
[2] Em https://www.cs.purdue.edu/homes/dgleich/packages/matlab_bgl/.

$$c_E(u) = \frac{1}{\max\{d_{uv} : v \in V\}} = \frac{1}{e(u)}. \tag{4.2}$$

$c_E(u)$ cresce à medida que o valor máximo de d_{uv}, $v \in V$, decresce.

4.2.1.2 Proximidade

Estudada originalmente por Bavelas (1950), a medida de proximidade, também conhecida como a inversa do distanciamento (Freeman, 1978; Sabidussi, 1966), é dada por

$$c_P(v) = \frac{1}{\sum_{t \in V \setminus v} d_{vt}}. \tag{4.3}$$

Normalizada, também pode ser considerada de acordo com a seguinte formulação (Beauchamp, 1965):

$$c_P(v) = \frac{n-1}{\sum_{t \in V \setminus v} d_{vt}}. \tag{4.4}$$

Contudo, são inapropriadas à aplicação em redes desconexas. Sejam v e t nós pertencentes a componentes distintos, $d_{vt} = \infty$, e, portanto, $c_P = 0$. Tão logo, Dangalchev (2006) propôs, alternativamente,

$$c_P(v) = \sum_{t \in V \setminus v} 2^{-d_{vt}}, \tag{4.5}$$

e, mais recentemente, Boldi & Vigna (2014) propuseram a utilização da média harmônica ao invés de média aritmética,

$$c_H(v) = \sum_{t \in V \setminus v} \frac{1}{d_{vt}}, \tag{4.6}$$

adotando como convenção $1/\infty = 0$, tomando como base a proposta de Marchiori & Latora (2000), discutida mais adiante. Segundo Boldi & Vigna (2014), essa medida é também a única dentre aquelas estudadas em seu trabalho (dentre elas grau, ranque de Page, status de Katz, proximidade e intermedialidade) que atende aos critérios axiomáticos de centralidade propostos. Além disso, para redes simples, a proximidade harmônica está fortemente correlacionada à proximidade aritmética. Portanto, dadas as suas características, essa abordagem de proximidade é estudada no presente trabalho.

Considerando a computação de $c_H(v)$ para todo $v \in V$, o custo computacional é limitado inferiormente pelo custo da computação do menor caminho de cada nó a todos os demais, que é $O(nm + n^2 \log n)$ e $O(nm)$ para redes não direcionadas ponderadas e não ponderadas, respectivamente. Kang et al. (2011) têm estudado abordagens de computação massiva para a aplicação da proximidade e outras medidas de centralidade a redes de grande escala. No caso da proximidade, a abordagem denominada proximidade efetiva é $O(dm)$, onde d é o diâmetro da rede.

4.2.1.3 Radialidade

Adicionalmente, (Valente & Foreman, 1998) propuseram uma medida de centralidade dita radialidade de um nó em uma rede,

$$c_R(u) = \frac{\sum_{v \in V}(\Delta_G + 1 - d_{uv})}{n-1}. \tag{4.7}$$

Δ_G denota o diâmetro da rede. Essa medida também pode ser entendida como baseada na proximidade. Ela mede especificamente o quão integrado um nó está a uma rede.

4.2.1.4 Valor centroide

Seja $\gamma_u(v)$ o número de nós em uma rede que localizam-se mais próximos de u que de v, $\gamma_u(v) = |\{w \in V : d(u,w) < d(v,w)\}|$. Se um comerciante seleciona a localização u de seu comercio enquanto que a sua concorrência seleciona a localização v, então ele terá $\gamma_u(v) + \frac{1}{2}(n - \gamma_u(v) - \gamma_v(u)) = \frac{1}{2}n + \frac{1}{2}(\gamma_u(v) - \gamma_v(u))$ clientes. Logo, seja $f(u,v) = \gamma_u(v) - \gamma_v(u)$, a concorrência selecionará, à medida que lhe interesse, o nó v que minimize $f(u,v)$ para instalar o seu estabelecimento. Como o comerciante sabe a estratégia da concorrência, ele deve selecionar a localização u para a qual o pior caso de $f(u,v)$ seja mais vantajoso,

$$c_F(u) = \min\{f(u,v) : v \in V - u\}. \tag{4.8}$$

$c_F(u)$ é dito o valor centroide e mede, especificamente, a vantagem de u em relação a todas as demais localizações, que é a diferença mínima no número de clientes que o comerciante ganha ou perde ao selecionar a localização u e sua concorrência selecionar a localização v mais apropriada diferente de u.

4.2.2 Medidas relacionadas ao grau

Diversas medidas de centralidade estão estritamente relacionadas ao grau. A seguir são apresentadas medidas que oferecem caracterizações para além da localidade, seja por meio de recorrência ou caminhadas aleatórias. Em geral seus custos computacionais estão atrelados a soluções de fórmulas fechadas através do cálculo de matrizes inversas. Em alguns casos são conhecidas algumas manobras que as tornam menos custosas sob certas circunstâncias. Soluções iterativas também são conhecidas.

4.2.2.1 Status de Katz

O status de Katz é conhecidamente uma das primeiras abordagens de medida de centralidade recorrente (Katz, 1953). No sentido de avaliar a relevância de um nó em uma rede, são contados, além dos votos diretos de seus vizinhos, também votos indiretos. Se considerarmos, por exemplo que um nó i está conectado a apenas a um nó j e recebe votos diretos apenas dele, mas j recebe votos de muitos nós na rede, podemos considerar que indiretamente os muitos indivíduos

votam em i. Tão logo, considerando que $(\mathbf{A}^k)_{ji}$ informa o número de caminhos de j a i com comprimento k,

$$c_K(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (\mathbf{A}^k)_{ji}, \tag{4.9}$$

onde α é um fator de amortecimento que garante a convergência da Eq. 8.10 para valores de $\alpha < 1/\lambda_1$, onde λ_i é o autovalor de Perron Frobenius de $\mathbf{A}$, também chamado raio espectral, i.e., o maior autovalor de $\mathbf{A}$. Satisfeita essa condição, a Eq. 8.10 converge de acordo com a fórmula fechada

$$(I - \alpha \mathbf{A}^T) c_K = 1_n, \tag{4.10}$$

onde 1_n é o vetor unitário e c_K o vetor solução. O cálculo de c_K através da fórmula fechada tem complexidade de tempo $O(n^3)$, devido ao cálculo da matriz inversa. No entanto, Foster et al. (2001) apresentam uma manobra que simplifica os cálculos resultando em complexidade de tempo $O(n + m)$.

4.2.2.2 Ranque de Page

Conhecidamente umas das medidas utilizadas pelo Google em sua máquina de busca, o ranque de Page também é uma medida de centralidade recorrente, e caracteriza o nós em relação à sua disposição na topologia da rede (Page et al., 1999). O índice $c_{PR}(i)$ é proporcional ao número de visitas que um nó i recebe de um caminhante realizando uma caminhada aleatória na rede,

$$c_{PR}(i) = \alpha \sum_{j \in v(i)} \frac{c_{PR}(j)}{k_j} + (1 - \alpha), \tag{4.11}$$

onde o fator de amortecimento, α, determina a fração de vezes que o movimento realizado pelo caminhante acontece sentido a um dos vizinhos de i. Nas demais transições, o caminhante é reiniciado aleatoriamente em quaisquer dos nós da rede. Para valores de $\alpha < 1$ a convergência da Eq. 4.11 é garantida.

4.2.2.3 Centralidade de uma caminhada aleatória

Como a sua denominação a descreve, essa medida toma como base a movimentação de um caminhante aleatório sobre a rede (Noh & Rieger, 2004). Dada a matriz de adjacências (generalizada ou binária), $\mathbf{A}$, a cada movimento o caminhante localizado em um nó i escolhe com probabilidade,

$$p_{ij}^{(1)} = \frac{a_{ik}}{\sum_{l=1}^{n} a_{ik}}, \tag{4.12}$$

um dentre os nós adjacentes a i para movimentar-se. Iterativamente,

$$p_{ij}^{(2)} = \sum_{k_1 \neq j} p_{ik_1}^{(1)} p_{k_1 j}^{(1)},$$

$$p_{ij}^{(3)} = \sum_{k_1 \neq j; k_2 \neq j} p_{ik_1}^{(1)} p_{k_1 k_2}^{(1)} p_{k_2 j}^{(1)},$$

e assim por diante, de modo que a probabilidade de um caminhante aleatório mover-se do nó i até o nó j em qualquer número de de passos é dada por

$$p_{ij} = p_{ij}^{(1)} + \sum_{l=1}^{\infty} l \sum_{k_1 \neq j; \ldots; k_l \neq j} p_{ik_1}^{(1)} \ldots p_{k_l j}^{(1)}, \tag{4.13}$$

cuja solução no domínio de l, $\forall l \in \mathbb{N}$, não é trivial. No entanto, considerando a analogia com cadeias de Markov, podemos redefinir o nó alvo, j, como um estado de absorção, removendo a j-ésima linha e coluna de P e denotando a matriz resultante como P_{-j}. Seja

$$p_{ij}^{(l)} = \sum_{k_1 \neq j; \ldots; k_l \neq j} (P_{-j}^{l-1})_{ik_l} p_{k_l j}^{(1)}, \tag{4.14}$$

onde $(P_{-j}^{l-1})_{ik_l}$ informa a probabilidade de começando a caminhada em i estar em k_l após $l-1$ passos. Substituindo Eq. 4.14 na Eq. 4.13, temos

$$p_{ij} = \sum_{l=1}^{\infty} l \sum_{k_1 \neq j; \ldots; k_l \neq j} (P_{-j}^{l-1})_{ik_l} p_{k_l j}^{(1)},$$

que, como uma série geométrica somável, converge para

$$p_{ij} = \sum_{k_1 \neq j; \ldots; k_l \neq j} ((I - P_{-j})^{-2})_{ik_l} p_{k_l j}^{(1)}, \tag{4.15}$$

onde I é dado pela matriz identidade $(n-1)$-dimensional. Por conveniência,

$$P_j = (I - P_{-j})^{-1} \mathbf{1}_(n-1), \tag{4.16}$$

onde P_j é dado pelo vetor de tempos de primeira passagem de um caminhante que terminou sua caminhada em j tendo começado em cada um dos $(n-1)$ demais nós. Logo,

$$c_{RW}(i) = \frac{n}{\sum_{j=1}^{n} p_{ji}}. \tag{4.17}$$

Mais uma vez, o custo de da medida de centralidade é determinado pelo cálculo da matriz inversa, $O(n^3)$.

4.2.3 Medidas relacionadas ao fluxo

Definimos como medidas relacionadas ao fluxo medidas que avaliam a ocorrência de vértice e arestas em caminhos mínimo ou ciclo locais ou de tamanhos diversos. Destacamos dentre elas,

o coeficiente de agrupamento local, a eficiência local e a intermedialidade.

4.2.3.1 Coeficiente de agrupamento local

Watts & Strogatz (1998) apresentam em seu trabalho a definição de coeficiente de agrupamento local,

$$C_i = \frac{N_{\triangle}^i}{N_3^i}, \tag{4.18}$$

onde $N_{\triangle}^i$ é o número de triângulos envolvendo o nó i e N_3^i é o número de triplas que incluem i. Essa medida caracteriza relações de transitividade estritamente locais entre os nós de uma rede. O custo computacional varia em relação à densidade da rede, podendo ser definido como $O(n)$ para redes esparsas a $O(n^2)$ para redes densas.

4.2.3.2 Eficiência local

A eficiência local é uma medida estritamente relacionada ao coeficiente de agrupamento local. Através dela pode-se caracterizar nós atalho, cuja remoção pode depreciar significativamente a robustez da rede. Seja d_{jk} o comprimento do menor caminho de j a k,

$$\text{eff}_i = \frac{1}{k_i(k_i-1)} \sum_{j,k \in v(i)} \frac{1}{d_{jk}}. \tag{4.19}$$

4.2.3.3 Intermedialidade

Um critério ainda mais sofisticado para avaliar a centralidade de um nó é a **intermedialidade**. De acordo com esse critério, um nó central seria frequentemente intermediador entre informações transientes na rede. Para se medir o nível de intermedialidade de um nó em uma rede, Freeman (1977) supôs que nós intermediadores participariam frequentemente dos menores caminho entre pares de nós. Seja $I_{st}(u)$ a fração de menores caminho entre s e t dos quais o nó u participa,

$$I_{st}(u) = \sum \frac{\sigma_{st}(u)}{\sigma_{st}},$$

onde $\sigma_{st}(u)$ é o número de menores caminhos entre s e t que passam por u, e σ_{st} é o número total de menores caminhos entre s e t. $I_{st}(u)$ também pode ser entendido como a probabilidade de um nó u está envolvido na comunicação entre os nós s e t. Logo,

$$c_I(u) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} I_{st}(u), \tag{4.20}$$

é dito o índice de centralidade baseado na intermedialidade. O cálculo de c_I é $O(nm)$ para redes sem pesos e $O(nm + n2\log n)$ para redes ponderadas.

Supor que informações se espalham através de uma rede apenas através do menor caminho é uma hipótese bastante rígida. Por isso, Newman (2005) propôs mais recentemente uma variação

da medida de intermedialidade baseada em caminhadas aleatórias. Assim a intermedialidade de um nó pode ser medida através de transformações da matriz de probabilidade de transição ou iterativamente.

4.2.4 Medidas hierárquicas

Com fins à avaliação de características topológicas de redes, para além das avaliações estritamente locais, Costa & Silva (2006); Costa & da Rocha (2006) propõem diversas medidas hierárquicas de avaliação de nós em redes. Para isso ampliam a vizinhança dos nós definindo anéis de cobertura. Embora, originalmente tenham sido propostas apenas medidas relacionadas com o grau e com o coeficiente de agrupamento, quaisquer medidas podem ser definidas utilizando esse *framework*. Portanto, definimos a seguinte medida hierárquica, denominada variação hierárquica, com fins a avaliação de um nó alvo em relação a sua vizinhança,

$$\mathrm{H_Q}(i) = \frac{\mathrm{Q}(i)}{\sum_{j=1}^{n} w_{ij}\mathrm{Q}(j)}, \tag{4.21}$$

onde Q pode ser qualquer medida de caracterização de nós, como por exemplo, grau, coeficiente de agrupamento, intermedialidade, etc.

4.3 Propriedades e modelos de redes

A fim de possibilitar o estudo mais profundo de propriedades e dinâmicas de rede, diversos modelos de redes têm sido propostos. Esses modelos refletem de forma bem comportada e independente algumas propriedades observadas em redes reais, permitindo o estudo mais preciso de seus efeitos sobre dinâmicas em redes. O modelo de redes de pequeno mundo (Watts & Strogatz, 1998), por exemplo, modela a formação de comunidades decorrente do aumento do coeficiente de agrupamento ao passo que as distâncias entre os indivíduos são diminuídas. O modelo de redes livre de escala (Barabási & Albert, 1999), por sua vez, modela redes heterogêneas, cujos indivíduos diferem entre si significativamente com relação a conectividade. Adicionalmente, o modelo de redes randômicas de Erdös & Rényi (1959) é conhecidamente o modelo mais popular na construção de redes sintéticas. Cada um desses modelos será apresentado a seguir, bem como modelos de redes com estruturas de comunidades (Girvan & Newman, 2002; Lancichinetti et al., 2008).

4.3.1 Redes aleatórias

Um dos primeiros passos no sentido de reproduzir sinteticamente uma rede de grande escala foi dado por Erdös & Rényi (1959). Diante da diversidade de sistemas e de suas particularidades, os matemáticos húngaros propuseram um modelo randômico para construção de redes, denominado grafos randômicos (*Random Graphs*) ou redes aleatórias. Dado inicialmente uma rede com n nós totalmente desconectados, cada par de nós, dentre os $\frac{1}{2}n(n-1)$ possíveis pares, é

recuperado e, com uma probabilidade p constante, é estabelecida uma ligação (não direcionada e sem pesos) entre eles. O resultado é uma rede homogênea com uma dentre as $C^m_{n(n-1)/2}$ possíveis configurações de redes com n nós e m arestas, sem ligações retroativas ou ligações duplicadas. Para $p = 0$, nenhuma ligação é estabelecida. Para $p = 1$, todas as possíveis ligação são estabelecidas; o resultado é uma rede totalmente conectada, com grau médio $\langle k \rangle = n - 1$. Para uma probabilidade p qualquer, o grau médio da rede é dado por $\langle k \rangle = p(n-1)$. A Figura 4.1 mostra algumas configurações de redes randômicas com 8 nós. Uma rede formada segundo esse modelo possui distribuição de grau bastante homogênea, seguindo a distribuição binomial

$$P(k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}.$$

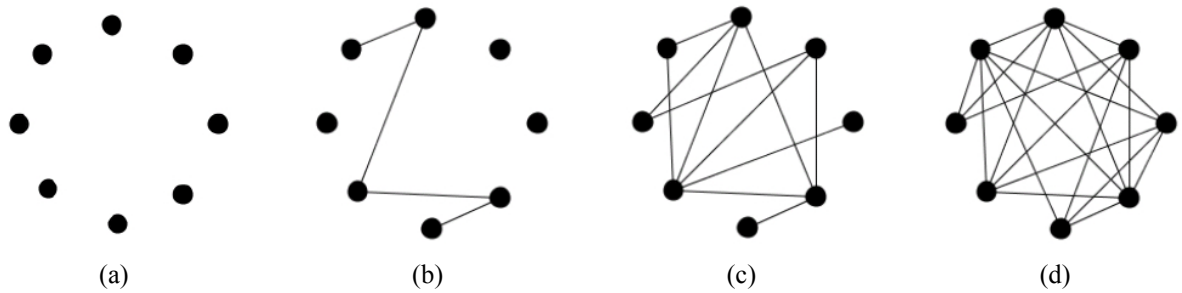


Figura 4.1: Configurações de redes randômicas com 8 vértices: (a) $p = 0$; (b) $p = 0.1$; (c) $p = 0.3$; (d) $p = 0.8$.

Para $n \to \infty$, uma rede de grande escala, essa distribuição é uma distribuição de Poisson (Costa et al., 2007),

$$P(k) = \frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!}.$$

A Figura 4.2 mostra a distribuição de grau média de 10 redes randômica com 10000 nós e probabilidade $p = 0.2$. O coeficiente de agrupamento de uma rede randômica é dado por $C = p$. A distância média da rede é dada por

$$\langle d \rangle \sim \frac{\ln n}{\ln \langle k \rangle}.$$

Embora a distância média de uma rede randômica seja pequena, quanto maior o número de nós, menor a probabilidade de existirem cíclos locais. Uma vez que cíclos locais são comuns em redes reais, essa característica torna as redes randômicas deficientes na representação de redes reais.

4.3.2 Transitividade e redes de pequeno mundo

A propriedade de **transitividade** é caracterizada pela existência de pequenos ciclos locais na rede (Holland & Leinhardt, 1971; Barrat & Weigt, 2000). Essa propriedade pode ser verificada através do **coeficiente de agrupamento**,

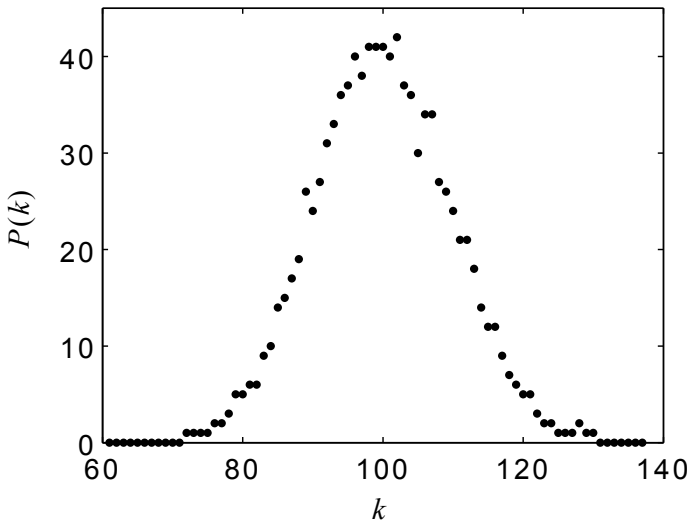


Figura 4.2: Distribuição de grau média de 10 redes randômicas com 1000 nós e probabilidade $p = 0.1$. Uma distribuição de Poisson.

$$C = \frac{3N_{\triangle}}{N_3},$$

onde $N_{\triangle}$ é o número de triângulos na rede e N_3 é o número de triplas de nós conectados. Um triângulo é um conjunto de três nós com ligações entre cada par de nós. O coeficiente de agrupamento é medido em razão do número de triplas de nós, de modo que $0 \leq C \leq 1$. O número de triângulos em uma rede pode ser calculado por

$$N_{\triangle} = \sum_{k>j>i} a_{ij}a_{ik}a_{jk},$$

enquanto que o número de triplas de nós conectados pode ser calculado por

$$N_3 = \sum_{k>j>i} (a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}).$$

Considerando a abordagem baseada no coeficiente de agrupamento local de Watts & Strogatz (1998), o coeficiente de agrupamento da rede é dado por

$$C = \frac{1}{N}\sum_i C_i,$$

onde C_i é dado pela Eq. 4.18.

Por volta de 1967, o psicólogo e cientista social Milgram, conduziu um experimento no intuito de identificar a distância geodésica entre indivíduos nos Estados Unidos (Milgram, 1967), que o levou a concluir de que o menor caminho entre dois indivíduos qualquer é mediado por cerca de 6 indivíduos. Esse fenômeno ficou conhecido como efeito de pequeno mundo, e deu origem ao termo ‘‘seis graus de separação’’ (Watts, 1999, 2004). O efeito de pequeno mundo afeta diretamente a dinâmica dos processos que ocorrem em uma rede. Por exemplo, considere

o espalhamento de informação através de uma rede. Uma rede com propriedade de pequeno mundo certamente irá espalhar informações rapidamente, na rede de relações de amizades entre norte-americanos, por exemplo, em apenas seis iterações todos os indivíduos teriam acesso à informação.

Logo, Watts & Strogatz (1998) propuseram um modelo de rede que reproduz a propriedade de pequeno mundo. Através do restabelecimento aleatório de algumas ligações em uma rede regular, esse modelo constrói redes que são um meio termo entre redes regulares e redes randômicas. São criados atalhos na rede, de modo que a distância média da rede cai significativamente, porém mantendo a maioria dos ciclos locais característicos de uma rede regular e se aproximando mais das redes reais de que as redes aleatórias.

Dado uma rede circular regular com n nós, onde cada nó está conectado aos κ vizinhos mais próximos em cada direção e $\langle k \rangle = 2\kappa$, $n \gg \kappa \gg \log(n) \gg 1$, cada uma de suas ligações é recuperada e restabelecida com probabilidade p. Para $p = 0$ mantém-se uma rede regular, com alto coeficiente de agrupamento e distância média grande. Para $p \to 1$, tem-se uma rede randômica com distância média pequena mas também com coeficiente de agrupamento pequeno. Para uma valor pequeno de p, $0.01 < p < 0.1$, por sua vez, obtemos uma rede com distância média pequena e coeficiente de agrupamento grande, uma rede de pequeno mundo. A Figura 4.3 ilustra o processo de transformação de uma rede regular ($\kappa = 2$) numa rede aleatória, enquanto que a Figura 4.4 mostra como se comportam o coeficiente de agrupamento e a distância média em função de p.

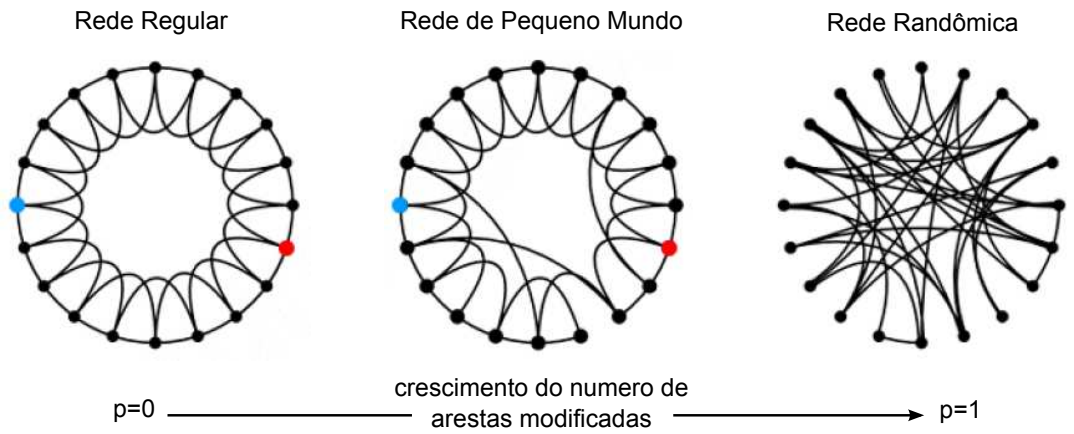


Figura 4.3: Modelo de redes de pequeno mundo de Watts & Strogatz (1998). Para $p = 0$ tem-se uma rede regular. Quanto maior a probabilidade p, mais o modelo se aproxima de uma rede randômica. Para $0.01 < p < 0.1$ a rede se comporta como uma rede de pequeno mundo. Observe que a distância entre os nós identificados em vermelho (cinza escuro) e em azul (cinza claro) diminui significativamente, enquanto que o coeficiente de agrupamento mantém-se alto. Figura adaptada com permissão da Macmillan Publishers Ltd:Nature (Watts & Strogatz, 1998), copyright ©1998.

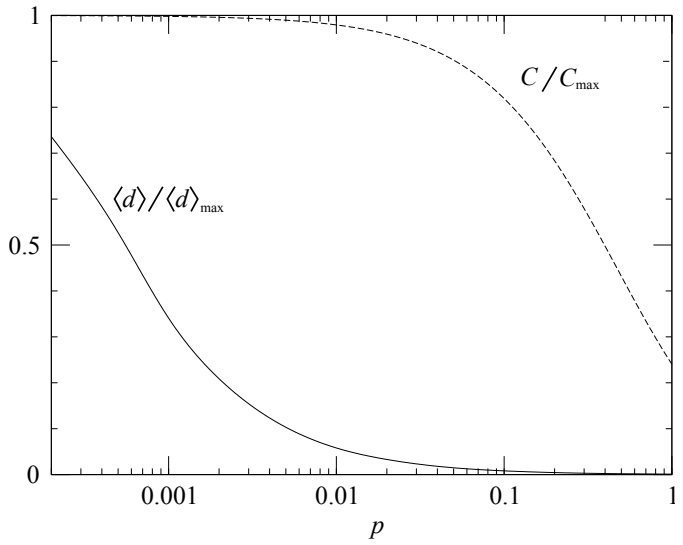


Figura 4.4: Comportamento da distância média, $\langle d \rangle$, e do coeficiente de agrupamento, C, em função de p, segundo o modelo de Watts & Strogatz (1998). Tanto o coeficiente de agrupamento como a distância média são apresentados em razão dos respectivos valores máximos. Para $0.01 < p < 0.1$ tem-se o efeito de pequeno mundo. Figura reproduzida com permissão de (Newman, 2003b). Copyright ©2003 Society for Industrial and Applied Mathematics. Todos os direitos reservados.

4.3.3 Distribuição de grau e redes livres de escala

Embora o grau caracterize individualmente os nós de uma rede e o grau médio a rede como um todo, nenhuma dessas duas medidas é capaz de descrever como os indivíduos da rede conectam-se uns aos outros e interagem entre si. Isso não é possível pois, frequentemente, redes reais possuem padrões de conectividade não uniformes. A caracterização da distribuição cumulativa de grau tem sido uma das principais maneiras de se identificar o padrão de conectividade dos indivíduos de uma rede. Redes randômicas, por exemplo, possuem padrão de conectividade com distribuição cumulativa de Poisson (Figura 4.2).

Dada uma rede $G = (V, E)$, $p(k)$ informa a fração de nós de G que possuem grau k, ou, de outro modo, a probabilidade de um nó escolhido ao acaso possuir grau k. A projeção da distribuição cumulativa na escala log-log permite a caracterização da distribuição de grau da rede. Se a distribuição de grau se ajusta bem à escala log-log, descrevendo uma reta decrescente, podemos dizer que a distribuição de grau segue a **lei de potência**, sendo de fato bastante heterogênea. Se a distribuição de grau se ajusta bem a uma escala log-linear, por outro lado, diz-se que a distribuição de grau segue uma lei exponencial. Distribuições homogêneas (p. ex., de Poisson) são projetadas de forma bastante curva nessas escalas. Clauset et al. (2009), por sua vez, propõem um método de maximização da verossimilhança para estimar o coeficiente de uma distribuição cumulativa de probabilidade a partir de dados empíricos.

Barabási & Albert (1999) mostram que muitas redes reais, por exemplo, a *World Wide Web*, a Internet, e a rede de interações entre proteínas, são caracterizadas por uma distribuição de grau heterogênea, propriedade que não é satisfeita pelos modelos de redes aleatórias ou de pequeno mundo (Erdös & Rényi, 1959; Watts & Strogatz, 1998). Enquanto que alguns nós são altamente

conectivos, a maioria dos nós é pouco conectado. Mais especificamente, Barabási & Albert (1999) mostraram que muitas redes reais possuem uma distribuição de grau que de fato segue a **lei de potência**,

$$p(k) \sim k^{-\gamma},$$

onde $2 \leq \gamma \leq 3$, e, portanto, propuseram um modelo para a construção de redes desse tipo. Esse modelo consiste em dois fundamentos básicos: crescimento e ligação preferencial. Dada uma rede qualquer, inicialmente com n_0 nós, iterativamente são adicionados novos nós, até que se atinja uma configuração com n nós. Para cada nó adicionado, m_0 novas ligações são estabelecidas entre o novo nó e os nós preexistentes. Os nós selecionados para estabelecer ligação com o novo nó, são escolhidos preferencialmente, proporcionalmente ao seu grau. Isto é, dado um novo nó i, i será conectado a um nó preexistente j com probabilidade

$$p(i \to j) = \frac{k_j}{\sum_u k_u}.$$

Inicialmente, os nós com mais ligações na rede tendem a ser preteridos nas ligações. Esses nós tornam-se altamente conectados e, iterativamente, tornam-se ainda mais conectados, efeito conhecido como ‘‘o rico torna-se mais rico’’ (Barabasi, 2003; Barabási & Albert, 1999). Ao final, temos uma rede com $n = n_0 + t$ nós e $m = m_0 t$ ligações, seja t o número de iterações. Especificamente, o modelo gera redes com distribuição de grau,

$$P(k) \sim k^{-3};$$

o grau médio é dado por $\langle k \rangle = 2m_0$. Uma vez que o grau médio não é suficiente à descrição do padrão de conectividade da rede, essas redes são denominadas **redes livres de escala**. A Figura 4.5 mostra a distribuição média de 10 redes geradas segundo o modelo de Barabási & Albert (1999), com 10000 nós e $m_0 = 5$.

4.3.4 Assortatividade e estruturas de comunidades

Algumas redes possuem vértices de diferentes tipos. Um exemplo clássico são as redes sociais. Nelas os nós representam indivíduos e as ligações representam relações sociais entre eles. Em redes cujos indivíduos são de fato heterogêneos, uma questão de interesse é saber como indivíduos de tipos distintos interagem entre si. Por exemplo, quão provável é o relacionamento de amizade entre indivíduos de classes sociais distintas. Uma medida que responde a esse questionamento é o **coeficiente de assortatividade**. Seja $\mathbf{E}$ uma matriz de cujas entradas são dadas por e_{ij}, e_{ij} o número de ligações conectando vértices do tipo i a vértices do tipo j (ou a soma dos pesos das conexões de vértices do tipo i a vértices do tipo j, no caso de redes ponderadas). $\mathbf{E}$ pode ser normalizada como

$$\hat{\mathbf{E}} = \frac{\mathbf{E}}{\|\mathbf{E}\|},$$

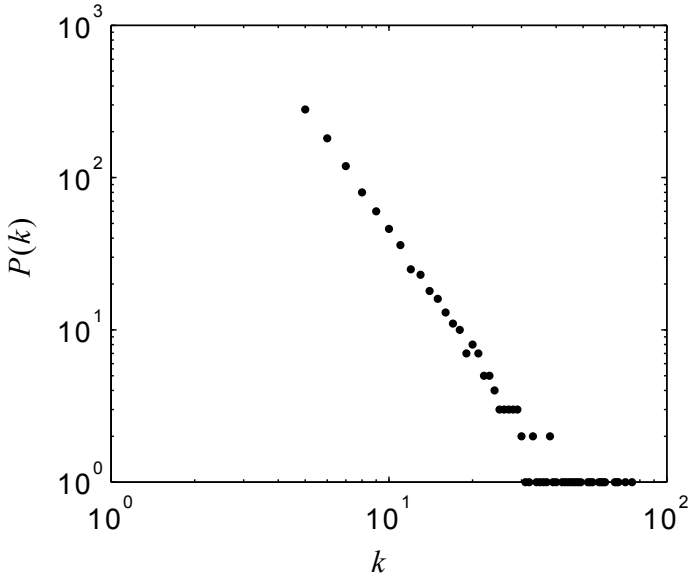


Figura 4.5: Distribuição de grau média de 10 redes geradas segundo o modelo de Barabási & Albert (1999) com 1000 nós e $m_0 = 5$.

onde $\|\mathbf{E}\|$ representa a soma de todas as entradas de $\mathbf{E}$. A probabilidade de um vértice do tipo i possuir um vizinho do tipo j é dada pela probabilidade condicional $P(j|i)$,

$$P(j|i) = \frac{\hat{e}_{ij}}{\sum_k \hat{e}_{ik}}.$$

$\sum_j P(j|i) = 1$. Gupta et al. (1989) sugere que o coeficiente de assortatividade, i.e., a fração de ligações entre indivíduos do mesmo tipo, é dada por

$$\tilde{Q} = \frac{\sum_i P(i|i) - 1}{q - 1},$$

onde q é o número de tipos de vértices existentes na rede. $0 \leq \tilde{Q} \leq 1$. Caso $\tilde{Q} = 1$, significa que as interações se dão exclusivamente entre indivíduos do mesmo tipo; caso $\tilde{Q} = 0$, as interações se dão aleatoriamente, independente do tipo dos vértices. No entanto, $\tilde{Q}$ tanto não trata matrizes assimétricas, como pesa igualmente os tipos, independente do número de nós de cada tipo, isto é, de suas expectativas. Newman (2003a) propôs um coeficiente de assortatividade alternativo,

$$Q = \frac{\mathrm{Tr}\, e - \|e^2\|}{1 - \|e^2\|},$$

onde $\mathrm{Tr}\, e$ é a soma dos elementos da diagonal principal da matriz $\mathbf{E}$, enquanto que $\|e^2\|$ é a soma do quadrado de todos os elementos em $\mathbf{E}$. Q retorna o valor 0 no caso de ligações aleatórias, enquanto que 1 no caso de assortatividade máxima. Porém, os resultados não se alteram no caso de uma matriz assimétrica transposta e o nível de contribuição do tipo para o coeficiente é ponderado pelo número de indivíduos do tipo.

Um caso particular de análise de assortatividade se dá avaliando os nós de acordo com a sua conectividade (Newman, 2002), no que é algumas vezes denominado correlação de grau. Nesse caso, os indivíduos são avaliados com relação à tendência de estarem conectados a outros

indivíduos que possuem o mesmo grau, por exemplo, nós bastante conectivos estarem conectados a nós bastante conectivos, ou nós pouco conectivos estarem conectados a nós bastante conectivos. Frequentemente observado em redes sociais, o primeiro caso é dito assortativo. Frequentemente observado em redes tecnológicas, redes de informação e redes biológicas, o segundo caso é dito dissortativo, para o qual Q retorna valores negativos (Newman, 2002, 2003b).

De um modo geral, o fenômeno da assortatividade sugere que indivíduos com experiências em comum estejam conectados entre si. Em redes sociais, por exemplo, muitas das relações entre os indivíduos ocorrem em razão de interesses em comum, ocupações em comum, faixa etária, entre outros critérios que os fazem semelhantes. Esse tipo de organização é denominada **estrutura de comunidade**, e estão presentes em diversos tipos de sistemas interconectados. Radicchi et al. (2004) apresentam duas definições para comunidades, uma do ponto de vista qualitativo e outra do ponto de vista quantitativo. Do ponto de vista qualitativo, se as ligações entre vértices de uma subrede são mais densas que as ligações com o restante da rede, a subrede é dita uma comunidade. Do ponto de vista quantitativo, por outro lado, uma subrede é dita uma comunidade simplesmente se a soma das ligações de suas vértices para dentro da subrede for maior que a soma das ligações de seus vértices para fora da subrede. Allém dos motivos supracitados, a identificação e caracterização dessas estruturas é importante pois, indivíduos em uma mesma comunidade tendem a compartilhar propriedades e dinâmicas (Clauset et al., 2004; Holme et al., 2003).

No sentido de validar uma partição de uma rede em comunidades, Newman & Girvan (2004) propuseram uma medida denominada **modularidade**. Considere uma rede representada através da matriz de adjacências generalizada $\mathbf{W}$, $w_{ij} > 0$ se existe uma aresta (v_i, v_j), caso contrário, $w_{ij} = 0$, e uma partição disjunta da rede em C comunidades. Compomos uma matriz $\mathbf{E}$, $C \times C$, entre as C comunidades, onde as entradas na diagonal principal, e_{ii}, informam a fração de pesos de arestas entre nós de uma mesma comunidade e as entradas no triângulo superior e inferior, e_{ij}, $i \neq j$, informam a fração de pesos de arestas entre nós em comunidades distintas [3]. O índice de modularidade da partição é dado por

$$Q = \sum_i (e_{ii} - a_i^2), \tag{4.22}$$

onde $a_i = \sum_j e_{ij}$. Essa medida varia no intervalo $[-1, 1]$. Quanto maior Q melhor a partição de comunidades. Como pode-se presumir, a modularidade é uma reinterpretação da medida de assortatividade, onde as comunidades são os tipos de nós (Clauset et al., 2004; Newman, 2006b; Fortunato, 2010).

De outra maneira, a Eq. 4.22 pode ser definida explicitamente em função das arestas,

$$e_{ij} = \frac{1}{2m} \sum_{u,v} w_{uv} \delta(c_u, i) \delta(c_v, j),$$

[3] Ao calcular a fração de pesos de arestas cada aresta deve ser considerada apenas uma vez, isto é, em se tratando de uma matriz de misturas simétrica, os pesos de arestas conectando nós de comunidades distintas deve ser dividido entre as entradas e_{ij} e e_{ji}.

$$a_i = \frac{1}{2m} \sum_u k_u \delta(c_u, i),$$

$$\delta(c_u, c_v) = \sum_i \delta(c_u, i)\delta(c_v, i),$$

onde $k_u = \sum_v w_{uv}$ e $m = \sum_u \sum_v w_{uv}$. No caso de partições rígidas, temos o δ de Kronecker, $\delta(c_u, c_v) = 1$ se $c_u = c_v$, caso contrário $\delta(c_u, c_v) = 0$. Logo,

$$\begin{aligned} Q &= \textstyle\sum_i \left(\frac{1}{2m} \sum_{u,v} w_{uv}\delta(c_u, i)\delta(c_v, i) - \frac{1}{2m} \sum_u k_u \delta(c_u, i) \frac{1}{2m} \sum_v k_v \delta(c_v, i) \right) \\ &= \textstyle\frac{1}{2m} \sum_{uv} \left(w_{uv} - \frac{k_u k_v}{2m} \right) \sum_i \delta(c_u, i)\delta(c_v, i) \\ &= \textstyle\frac{1}{2m} \sum_{uv} \left(w_{uv} - \frac{k_u k_v}{2m} \right) \delta(c_u, c_v). \end{aligned}$$

O primeiro termo da equação à direita,

$$\frac{1}{2m} \sum_{uv} w_{uv} \delta(c_u, c_v),$$

corresponde à fração de pesos de arestas entre nós do mesmo tipo, isto é, da mesma comunidade; e o segundo termo,

$$\frac{1}{2m} \sum_{uv} \frac{k_u k_v}{2m} \delta(c_u, c_v),$$

corresponde à mesma quantia, porém se as arestas fossem estabelecidas ao acaso, sem qualquer denominação de tipo, como uma hipótese nula ou amostra de controle. Portanto, $Q \approx 0$, caracteriza a inexistência de comunidades e $Q \approx 1 - \frac{1}{C}$ caracteriza a existência de comunidades totalmente independentes. Em redes reais, tipicamente Q assume valores no intervalo $[0.3, 0.7]$ para comunidades bem definidas.

A medida de modularidade tem sido utilizada em diversos trabalhos para a identificação da melhor partição de uma rede em comunidades, i.e., detecção de comunidades (Newman, 2004b; Clauset et al., 2004; Guimerà et al., 2004; Duch & Arenas, 2005; Guimerà & Amaral, 2005; Pujol et al., 2006; Ye et al., 2008). Outras abordagens para a detecção de comunidades também têm sido consideradas na literatura: métodos hierárquicos (Girvan & Newman, 2002; Zhou, 2003b,a; Newman & Girvan, 2004), métodos que mimetizam dinâmicas em redes, tais como sincronização (Reichardt & Bornholdt, 2004; Arenas et al., 2006) e competição (Quiles et al., 2008), métodos baseados em análise espectral (Newman, 2006a) e até mesmo em teoria de informação (Ziv et al., 2005). Mais recentemente, até mesmo abordagens semissupervisionadas foram propostas (Breve et al., 2012; Ma et al., 2010; Allahverdyan et al., 2010; Decelle et al., 2011; Steeg et al., 2014; Zhang et al., 2014). Análises comparativas entre diversos métodos podem ser encontradas em (Newman, 2004a; Danon et al., 2005; Lancichinetti & Fortunato, 2009). Uma revisão mais extensa por ser encontrada em (Fortunato, 2010).

Como *benchmark* para a avaliação do desempenho de métodos de detecção de comunidades,

Girvan & Newman (2002) propuseram um modelo para a construção de redes aleatórias como comunidades bem definidas. Originalmente, são geradas redes com $n = 128$ nós divididos em $C = 4$ comunidades, cada uma com 32 vértices; e grau médio $\langle k \rangle = 16$. Ligações são estabelecidas entre pares de vértices aleatoriamente, com probabilidades p_{in} entre nós definidos em uma mesma comunidade e p_{out} entre nós definidos em comunidades distintas. Contudo, configurações diversas têm sido utilizados na literatura. Seja $p_{\text{out}} = 1 - p_{\text{in}}$, o número de ligações de um nó para outra comunidades é dado por $Z_{\text{out}} = p_{\text{out}}\langle k \rangle$. Esse parâmetro é tipicamente definido como parâmetro de mistura entre as comunidades. A Figura 4.6 apresenta um exemplo construído de acordo com esse modelo. Duas coleções de *benchmarks* de detecção de comunidades bastante abrangentes considerando esse modelo são encontradas em (Danon et al., 2005; Lancichinetti & Fortunato, 2009).

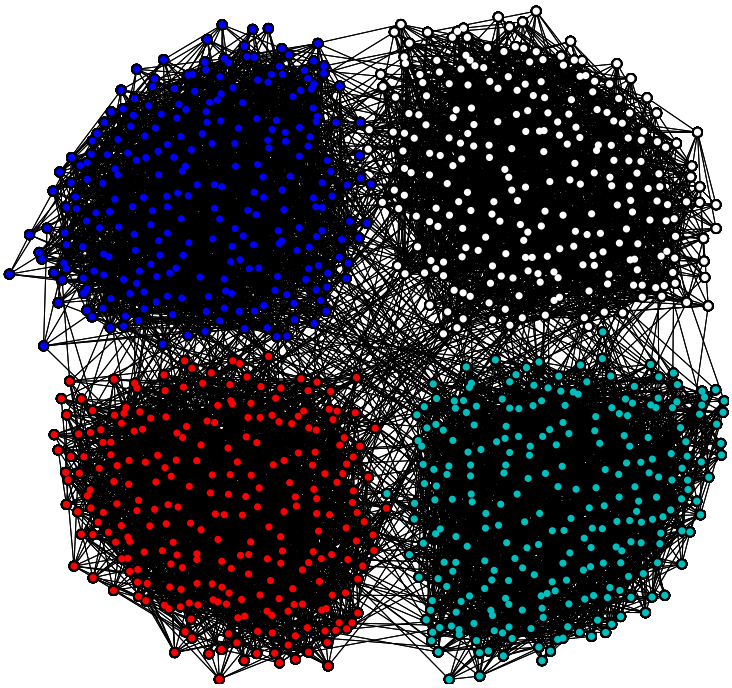


Figura 4.6: Redes com comunidades bem definidas geradas pelos modelos de redes de Girvan-Newman com 1024 nós, $\langle k \rangle = 16$ e 4 comunidades bem definidas ($Z_{\text{out}} = 2$) de igual tamanho.

No entanto, como discutido anteriormente, redes reais são frequentemente caracterizadas por distribuições de grau heterogêneas e comunidades de tamanhos variados. Lancichinetti et al. (2008) propuseram, portanto, um modelo de rede que define ambas a distribuições de grau e de tamanhos de comunidades como leis de potência, respectivamente, com expoente $\gamma \in \{2, 3\}$ e $\beta \in \{1, 2\}$[4]. O modelo de configuração proposto por Molloy & Reed (1995) é utilizado para construir redes com o grau de seus nós definido a partir da distribuição com expoente γ; os extremos da distribuição, $k_{\min}$ e $k_{\max}$, são definidos de modo que grau médio $\langle k \rangle$ seja satisfeito. Cada nó compartilha uma fração $1 - \mu$ de suas arestas com outros nós da mesma comunidade e uma fração μ de suas arestas com nós de outras comunidades, $\mu \in [0, 1]$. O tamanho das comunidades é obtido da distribuição com expoente β de modo que a soma de seus tamanhos

[4]Código disponíveis em `https://sites.google.com/site/andrealancichinetti/files`.

é igual a n e as restrições de tamanho, $s_{\min}$ e $s_{\max}$, são impostas de acordo com a definição das comunidades, $s_{\min} > k_{\min}$ e $s_{\max} > k_{\max}$. O número de nós é dado pelo parâmetros n. A Figura 4.7 apresenta um exemplo construído de acordo com esse modelo.

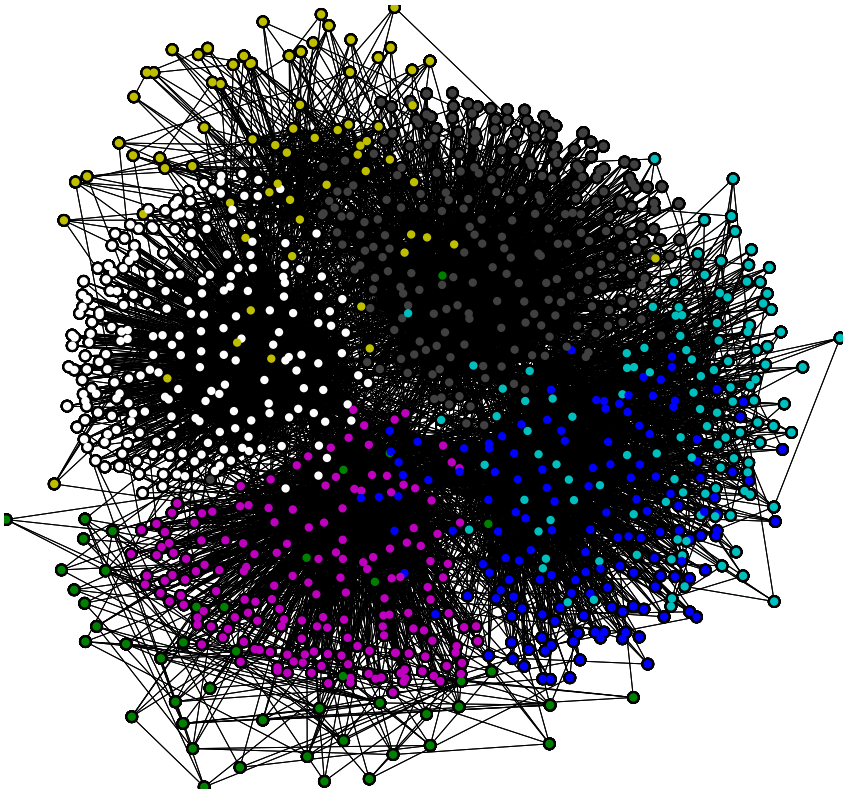


Figura 4.7: Redes com comunidades bem definidas geradas pelos modelos de redes de Lancichinette-Fortunato-Radichi com 1024 nós, $\langle k \rangle = 16$, distribuição de grau heterogênea ($\gamma = 2$), comunidade com tamanhos variados porém homogêneos ($\beta = 1$) e comunidades bem definidas ($\mu = 0.2$).

4.4 Resiliência e epidemiologia

Um dos principais resultados acrescentados através da descoberta das redes livres de escala diz respeito à **resiliência** da rede. Anteriormente, pensava-se que a remoção de nós de uma rede acarretaria a degeneração da rede, independentemente de sua disposição na rede. No entanto, como destacado em (Cohen et al., 2000), redes livres de escala são particularmente resistentes à remoção aleatória de nós, porém são grandemente afetadas pela remoção direcionada dos nós mais conectivos da rede, i.e., *hubs*, como destaca (Cohen et al., 2001). Seja p o valor percentual de nós eliminados de uma rede livre de escala, como mostra a Figura 4.8, se os nós removidos forem *hubs*, um pequeno percentual p de nós removidos acarreta no crescimento significativo da distância média da rede, $\langle d \rangle$, e consequente degeneração da rede. Esses resultados foram verificados em redes de interação de proteínas, na WWW e na Internet (Albert et al., 2000; Jeong et al., 2001; Newman et al., 2002), por exemplo.

Outro resultado diz respeito à compreensão dos mecanismos através dos quais doenças, informações, vírus (biológicos, ou computacionais), rumores, riquezas, entre outros, se espalham

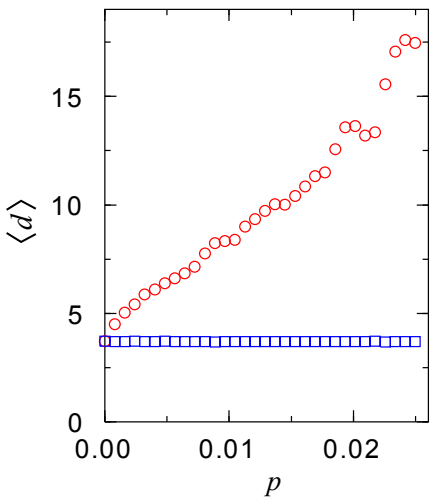


Figura 4.8: Resiliência da Internet. Quando os nós são removidos aleatoriamente (quadrados) a distância média da rede, $\langle d \rangle$, permanece estável. Entretanto, quando a remoção ocorre preterindo os nós mais conectivos (circunferências), a depreciação da rede como um todo é eminente. Figura adaptada com permissão da Macmillan Publishers Ltd:Nature (Albert et al., 2000), copyright 2000.

através das redes. Como considerado por Newman (2003b), por exemplo, a principal razão para a modelagem de redes de contatos sexuais é a sua análise e talvez controle do espalhamento de doenças sexualmente transmissíveis. Do mesmo modo, redes de email nos permitem analisar o espalhamento de vírus de computadores, e redes sociais nos permitir compreender como rumores e informações se espalham. Através da análise do **processo epidemiológico**, e das características das redes, pode-se, por fim, prever se o processo de epidemiológico retrocederá e a informação/vírus/etc se extinguirá ou se o processo epidemiológico evoluirá a uma epidemia, ou mesmo, uma pandemia.

Um dos primeiros trabalhos a modelar a evolução de um processo epidemiológico em uma rede complexa foi realizado por Pastor-Satorras & Vespignani (2001). Os autores analisaram dados reais de infecções de vírus computacionais a fim de identificar o tempo médio de vida e a eventual persistência do vírus na Internet. Com base nos dados, os autores desenvolveram um modelo de espalhamento de infecções suscetível-infectado-suscetível (SIS), através do qual puderam concluir que, nesse tipo de rede, uma infecção pode se proliferar independente da taxa de espalhamento. Por outro lado, na maioria das variações da taxa de espalhamento, a epidemia não persiste, regredindo após algum tempo. Esses resultados tornam evidente a influência da topologia da rede no processo epidemiológico de uma infecção e contrariam muitas das conclusões padrão em modelagem epidêmicas que, até então, consideravam a taxa de espalhamento um fator crítico.

Barthélemy et al. (2004), por sua vez, acrescentam que, em populações heterogêneas, embora os hubs sejam pouco numerosos, eles são os principais responsáveis pela proliferação de infecções, pois um efeito em cascata é iniciado após a sua contaminação; em se tratando de redes sociais (assortativas), dos *hubs* para indivíduos com conectividade intermediária, daí para os indivíduos menos conectivos, resultando na infecção de toda rede. Esses resultados são particularmente importantes para a criação de estratégias de controle epidêmico. Estratégias de

imunização direcionadas que se adaptam ao longo do tempo à população de risco podem ser particularmente efetivas. Em infecções tipicamente iniciadas em indivíduos pouco conectivos, a imunização de *hubs* pode ser uma excelente estratégia, restringindo a infecção a comunidades isoladas da rede.

Mais recentemente, Yang et al. (2008) analisaram o processo de infecção ótimo. Os autores modelaram uma estratégia universal para o processo de infecção em redes complexas a fim de identificar a estratégia ótima para o espalhamento. Para isso enviesaram o espalhamento utilizando apenas a informação local acerca do grau dos nós vizinhos, em alguns casos dando prioridade à infecção de indivíduos mais conectivos, noutros aos indivíduos menos conectivos. Dentre as estratégias adotadas, a estratégia ótima, i.e. cujo maior número de indivíduos foi infectado até a extinção da epidemia, contra-intuitivamente, atribuiu a nós relativamente pouco conectivos maior relevância.

Outros resultados destacam nós com alta intermedialidade e nós pertencentes a núcleos conectivos das redes (k-*cores*) como espalhadores influentes (Holme et al., 2002; Kitsak et al., 2010; Chen et al., 2012). Em se tratando de dinâmicas competitivas, como propagação de rumores e disputas por consumidores, outros comportamentos são observados (Borge-Holthoefer & Moreno, 2012; Borge-Holthoefer et al., 2012, 2013; Goyal et al., 2014; Galeotti & Goyal, 2009). Vide (Barrat et al., 2008; Boccaletti et al., 2006) para discussões mais extensas.

4.5 Considerações finais

O presente trabalho tem como um de seus principais objetivos propor critérios para seleção de nós representativos em uma rede. Uma abordagem utilizada nesse sentido é caracterizar seus indivíduos através de medidas de centralidade. Algumas dessas medidas são descritas na Seção 4.2 e seus resultados são apresentados e discutidos na segunda parte desse texto. Destaca-se dentre elas a variação hierárquica de grau (Eq. 4.21).

Além disso, o estudo de distribuições cumulativas, modelos de redes, validação de partições de redes, bem como aplicações diversas é bastante relevante para o desenvolvimento do presente trabalho. Por exemplo, em adição à distribuição de grau, apresentada na Seção 4.3.3, outras medidas podem ser estudadas. Nos Capítulos 6 e 7, analisamos diversas distribuições de medidas de caracterização de nós em redes.

No Capítulo 7, por sua vez, propomos a utilização do índice de modularidade como uma medida de validação de métodos de construção de redes a partir de dados vetoriais para o aprendizado semissupervisionado baseado em redes. Contudo, identificamos alguns inconvenientes da medida, e portanto, propomos uma medida mais adequada para este fim no Capítulo 8.

Os modelos de redes modulares de Girvan-Newman e Lancichinette-Fortunato-Radichi são utilizados como *benchmark* de validação de desempenho de rotulação semissupervisionado no Capítulo 6. No caso, a comunidade dos nós é utilizada como um rótulo *ground truth*, de modo que, para valores moderados de mistura entre comunidades, μ e Z_{out}, a suposição de grupos é satisfeita.

Por fim, os problemas discutidos na Seção 4.4 são particularmente importantes, pois caracterizam analogias à identificação de indivíduos representativos em uma rede. Como é discutido nos Capítulos 6 e 7, enquanto que métodos difusivos apresentam ganho de desempenho significativo quando semeados por representativos, métodos que implementam dinâmicas de competição apresentam outros comportamentos.

Parte II

Resultados e discussões

Reiterando, o presente trabalho tem como principais objetivos:

- Explorar a influência dos objetos previamente rotulados no desempenho do processo de aprendizado semissupervisionado baseado em redes; e

- Estudar e propor critérios e estratégias para a caracterização e realce de itens de dados representativos.

Nesse sentido, estudamos primeiramente a **robustez dos algoritmos de aprendizado semissupervisionado** baseados em redes utilizando uma abordagem sistemática (Capítulo 5). Os resultados mostram que o tamanho da amostra de dados disponível rotulada é determinante para o desempenho do aprendizado, indício de que o aprendizado semissupervisionado depende da qualidade da amostra disponível rotulada. Se pequenas, variações de desempenho significativas são observadas; se suficientemente grandes, os desempenhos nos ensaios se aproximam do desempenho médio. Além disso, o desempenho do aprendizado varia de acordo com a consistência entre grupos e classes dos itens de dados e de acordo com a conveniência do método de construção da rede para o processo de rotulação realizado pelo algoritmo de aprendizado semissupervisionado. A abstração gráfica oferece diversas vantagens, dentre elas a representação de similaridades topológicas. Em contrapartida, no entanto, ao utilizá-la abrimos mão da descrição completa dos dados, o que pode limitar o desempenho do aprendizado. Considerando adicionalmente restrições de ligações de acordo com os rótulos dos itens de dados disponíveis rotulados na construção da rede, o desempenho de classificação semissupervisionada é melhorado significativamente. Esses resultados dão razão aos estudos apresentados nos Capítulos 6 e 7. Reiterando a hipótese proposta no presente trabalho, sejam os dados representados através de uma rede que reflita consistentemente o processo de geração dos dados, itens de dados representativos destacar-se-ão em relação aos demais.

Logo, estudamos em seguida **abordagens para a caracterização de indivíduos representativos** em redes modulares (Capítulo 6), evitando-se, por ora, o viés da conveniência do método de construção da rede para o método de rotulação semissupervisionado. Foram considerados os modelos de Girvan-Newman (GN) e Lancichinetti-Fortunato-Radicchi (LFR) para gerar redes com estruturas modulares bem definidas. Contraintuitivamente, nós bastantes conectados (*hubs*) não são muito representativos. Nós razoavelmente conectados em vizinhanças pouco conectadas, por outro lado, são; estritamente local, esse critério de caracterização é escalável a grandes volumes de dados. Em redes com distribuição de grau homogênea - modelo Girvan-Newman (GN), nós com alto coeficiente de agrupamento também mostram-se representativos. Por outro lado, em redes com distribuição de grau heterogênea - modelo Lancichinetti-Fortunato-Radicchi (LFR), nós com alta intermedialidade se destacam. Nós com alto coeficiente de agrupamento em redes GN estão tipicamente situados em *motifs* do tipo quase-clique; nós com alta intermedialidade em redes LFR são *hubs* situados na borda das comunidades, interligando-as, e funcionando como portas de entrada para as mesmas. Em ambos os casos, os nós identificados são excelentes regularizadores.

Em suma, esses resultados apontam para a possibilidade de se obter ganho significativo de desempenho de aprendizado semissupervisionado semeando-o com indivíduos representativos. No entanto, em modelos de redes distintos, critérios distintos caracterizam bem os indivíduos mais representativos. Por isso estudamos a combinação desses critérios no sentido de prover uma **abordagem unificada para a caracterização de indivíduos representativos** (Seção 6.5). Nos cenários destacados acima, os critérios que melhor caracterizam indivíduos representativos são aqueles cujas distribuições de suas aferições são mais heterogênea. Logo, propomos a utilização da análise de componentes principais (PCA) para a composição de componentes ortogonais ordenadas de acordo com suas variâncias. Mais ''informativas'', acredita-se que as primeiras componentes principais sejam composições das distribuições mais heterogêneas. Dentre elas, a segunda componente principal se destaca.

A fim de possibilitar a extensão desses resultados a cenários diversos, estudamos então a **construção da rede a partir de bases de dados vetoriais**. Como destaca o Capítulo 5, o desempenho do aprendizado semissupervisionado é dependente da conveniência do método de construção da rede para o processo de rotulação realizado pelo algoritmo de aprendizado semissupervisionado. Métodos tradicionais de construção de redes ou são dependentes do ajuste de seus parâmetros de densidade ou se sustentam em suposições fortes. Contudo, a construção da rede a partir de dados vetoriais é um processo que não está subjugado à conveniência dos processos de rotulação do aprendizado semissupervisionado. Este problema equivale à redução não linear de dimensionalidade e ao aprendizado de variedades. Por isso, estudamos a construção da rede em sí mesma. O Capítulo 7 apresenta resultados obtidos nesse sentido. Um método denominado AdaRadius é proposto e apresenta diversas propriedades interessantes, tais como adaptabilidade em dados com densidade variada, baixa dependência da configuração de seus parâmetros e custo computacional razoável. Além disso, as redes resultantes são esparsas e permitem que o método de rotulação tire grande vantagem da rotulação prévia de indivíduos representativos. De certo modo, indivíduos representativos são realçados. Para além disso, no sentido de avaliar a qualidade de métodos de construção de redes, são estudadas medidas de validação (Capítulo 8).

As principais contribuições do presente trabalho são compiladas no Capítulo 9, bem como são discutidos as questões em aberto e os desdobramentos desse trabalho como perspectivas de trabalhos futuros.

Capítulo

5

Robustez do aprendizado semissupervisionado

Como primeiro estudo e também no sentido de desenvolver uma metodologia consistente para a avaliação de desempenho de métodos de aprendizado semissupervisionado baseados em redes, o seguinte trabalho foi realizado. Em resumo, métodos de aprendizado semissupervisionado baseado em redes foram aplicados ao problema de classificação binária. Foi avaliada a adequação de métodos de construção de rede a métodos de rotulação; o comportamento dos modelos resultantes quando não são satisfeitas as suposições de suavidade e grupos; sua dependência da qualidade da amostra disponível rotulada; e o ganho de desempenho com a adição de restrições de ligação baseadas em conhecimentos prévios sobre os dados.

5.1 Dados, métodos e simulações

Embora métodos de aprendizado semissupervisionado baseado em redes sejam habitualmente propostos associados a um método de construção de rede específico, geralmente, quaisquer métodos de construção de redes podem ser aplicados a quaisquer métodos de aprendizado semissupervisionado. Portanto, estudamos os seguintes métodos de construção de redes a partir de dados vetoriais: rede completa ponderada (Co), k-vizinhos-mais-próximos (kNN), kNN ponderado (wkNN), ϵRadius (ϵRd), e ϵRd ponderado (wϵRd). Todos são apresentados na Seção 2.4.6. O parâmetro de largura de banda dos métodos ponderados, σ, é definido segundo valor de referência em (Breve et al., 2012), $\sigma = 1$. Os parâmetros de densidade, k, dos métodos kNN e wkNN, e ϵ, dos métodos ϵRd e wϵRd, foram ajustados de modo a se obter redes esparsas porém com um único componente; k, incrementalmente a partir de 1; e ϵ decrementalmente a partir do valor correspondente à maior distância entre pares de itens de dados. Adicionalmente,

os métodos kNN e ϵRd também foram combinados às restrições *must/cannot-link* como no método de agrupamento e rotulação semissupervisionado em (Wagstaff et al., 2001), resultando nos modelos kNNMCL e ϵRdMCL.

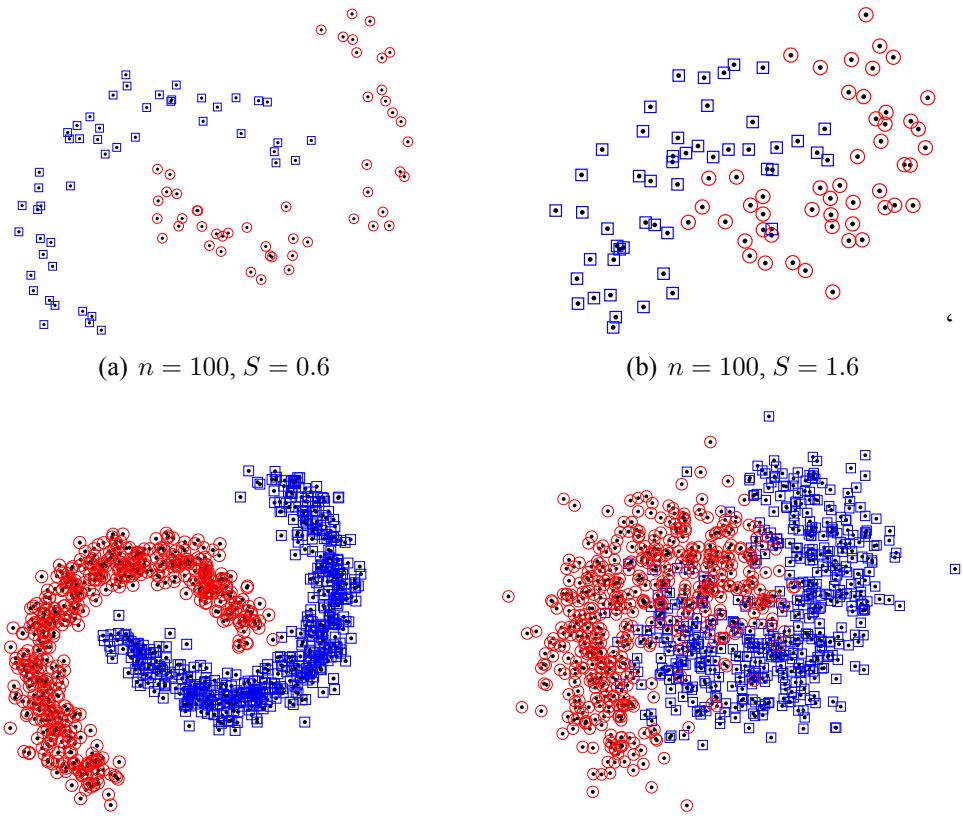
Como método de rotulação, estudamos os métodos fluxo-máximo corte-mínimo (MinCut), campos Gaussianos e funções harmônicas (GFHF), consistência local e global (LGC), e competição e cooperação de partículas (PCC), apresentados na Seção 2.4. Como discutido na Seção 2.4.4, MinCut, GFHF e LGC podem ser facilmente interpretados de acordo com o *framework* de regularização. Em razão da definição de α, LGC, especificamente, possui certa propensão à correção de nós rotulados de forma inconsistente em relação de seus vizinhos. O parâmetro externo de LGC, α, e os parâmetros externos de PCC, p_{grd} e Δ_v, foram definidos de acordo com valores de referência em (Zhou et al., 2004; Breve et al., 2012), $\alpha = 0.99$, $p_{\text{grd}} = 0.70$ e $\Delta_v = 0.35$.

No sentido de avaliar o desempenho do aprendizado, três medidas de validação distintas foram consideradas: o índice de Rand ajustado (ARI), a silhueta, e a modularidade; as duas primeiras definida na Seção 2.5 e a última na Seção 4.3.4. É importante que se reitere que cada uma dessas medidas possui um viés diferente. Quanto maior a consistência entre o resultado do aprendizado e os rótulos conhecidos a priori, maior o ARI; quanto maior a consistência do resultado com os grupos do conjunto de dados, maior a silhueta; e, maior a modularidade, quanto maior a consistência com as comunidades da rede. Destaque-se adicionalmente que a modularidade avalia apenas o desempenho do método de rotulação, sendo incapaz de avaliar a construção da rede.

Por simplicidade, consideramos apenas conjuntos de dados sintéticos em forma de bananas, gerados pela ferramenta MATLAB PrTools Toolbox (Duin et al., 2007). Apenas duas classes são consideradas. O número de instâncias e a dispersão dos dados são definidos de acordo com os parâmetros n e S, respectivamente. Para quaisquer configurações de n e S, 4 instâncias do conjunto de dados são geradas. A Figura 5.1 apresenta exemplos de redes resultantes. Fixado n, quanto maior S, maior a sobreposição entre os grupos. Fixado S, quanto menor n, menor a densidade dos grupos.

Com o objetivo de garantir que o processo de rotulação semissupervisionado seja semeado com amostras rotuladas de qualidades diversas, para cada configuração e instância do conjunto de dados, 20 ensaios são realizados com divisões aleatórias do conjunto de dados em itens de dados rotulados e não rotulados. Garante-se apenas que pelo menos um exemplo de cada classe esteja disponível rotulado. Média e desvio padrão de 80 ensaios para cada configuração de n e S são exibidos como resultado. Em algumas simulações o tamanho da amostra rotulada, L, é constante, em outras variado. Para valores grandes de L a amostra cobre-se razoavelmente ambas as classes do problema, como propõe Guha et al. (1998) (vide Capítulo 3). No entanto, para valores pequenos de L não se pode garantir a cobertura apropriada de ambas as classes. Esse comportamento é de particular interesse, pois nos permite avaliar a robustez dos métodos de aprendizado semissupervisionado em relação à ‘‘qualidade da amostra’’.

Adotamos a seguinte metodologia para a realização das simulações. Primeiro, os métodos MinCut, GFHF, LGC, e PCC são aplicados sobre os grafos gerados pelos métodos de construção



(a) $n = 100$, $S = 0.6$ (b) $n = 100$, $S = 1.6$

(c) $n = 1000$, $S = 0.6$ (d) $n = 1000$, $S = 1.6$

Figura 5.1: Exemplos de conjuntos de dados em forma de bananas com configurações diversas, gerados pela ferramenta MATLAB PrTools Toolbox (Duin et al., 2007).

da rede Co, kNN, wkNN, ϵRd e wϵRd sobre conjuntos de dados em forma de bananas com $n = 1000$, $L = 20$, e misturas entre classes variadas, $S = [0.6 : 0.2 : 1.6]$. Os modelos mais robustos são elencados e selecionados para as seguintes simulações:

I. Sejam $n = 1000$ e $L = 20$, variamos a mistura entre as classes, $S = [0.6 : 0.2 : 1.6]$;

II. Sejam $S = 1$ e $L = 0.02 \times n$, variamos o número de objetos do conjunto de dados, $n = \{100, 200, 400, 600, 800, 1000, 1500, 2000\}$;

III. Sejam $S = 1.6$ e $L = 0.02 \times n$, variamos $n = \{100, 200, 400, 600, 800, 1000, 1500, 2000\}$;

IV. Sejam $n = 1000$ e $S = 1.6$, variamos a fração de objetos rotulados, $L = \{0.01, 0.02, 0.04, 0.06, 0.1, 0.2\} \times n$.

Como referência de desempenho, para cada medida de validação também são exibidos o desempenho do classificador de rótulos aleatórios equiprováveis (H_0) e do classificador exato (H_1). Destaque-se adicionalmente que H_1 é consistente com as classes, e não com os grupos ou comunidades.

5.2 Comparação entre modelos de construção de rede

Os métodos MinCut, GFHF, LGC, e PCC combinados às variações de métodos de construção de rede Co, kNN, wkNN, ϵRd e wϵRd foram aplicados primeiramente aos conjuntos de dados em forma de bananas variando-se a mistura entre as classes, $n = 1000$, $L = 20$, $S = [0.6 : 0.2 : 1.6]$. No caso do MinCut, especificamente, devido a limitações do algoritmo, apenas os métodos wkNN e wϵRd foram considerados. Em razão dos empates, caminhos de aumento não podem ser computados para redes com arestas não ponderadas (vide Seção 2.4). Além disso, o algoritmo é bastante custoso, impossibilitando a utilização de redes completas (Co). Foi utilizado apenas o ARI como critério para a eleição dos melhores modelos.

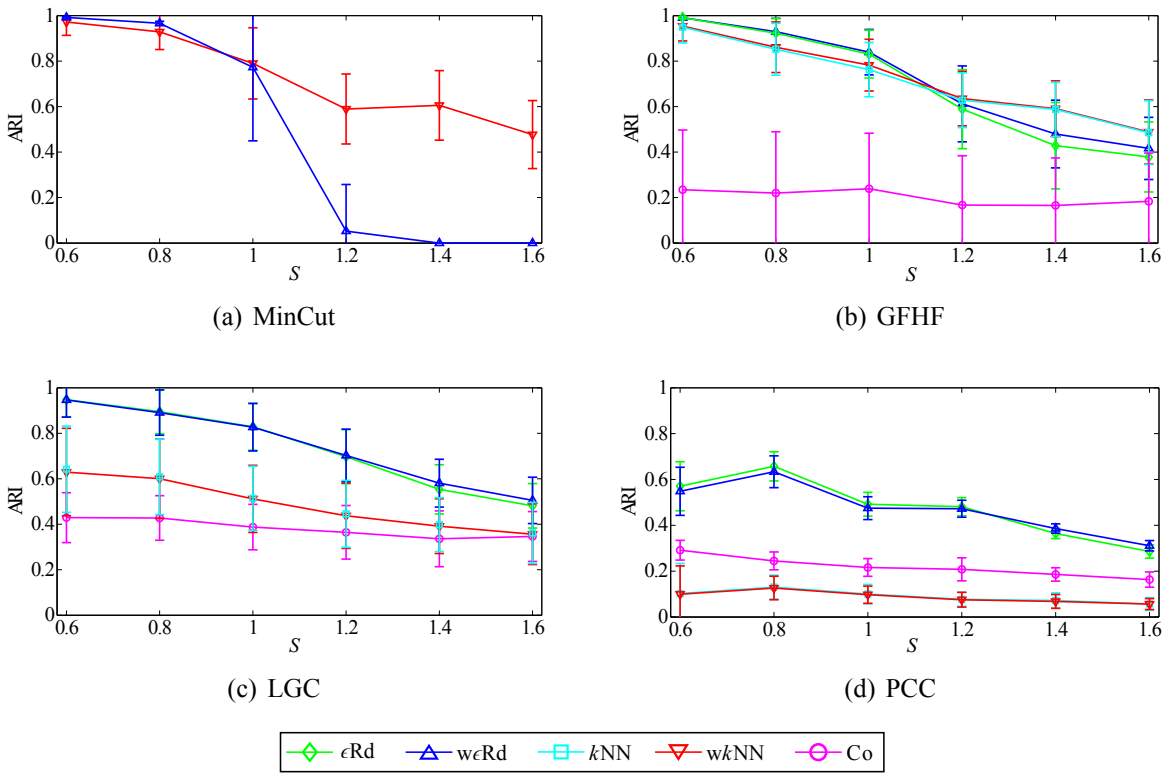


Figura 5.2: ARI médio e desvio padrão obtido pelos modelos MinCut, GFHF, LGC, e PCC combinados aos métodos de construção de rede Co, kNN, wkNN, ϵRd e wϵRd sobre dados em forma de bananas com $n = 1000$, $L = 20$ e $S = [0.6 : 0.2 : 1.6]$.

A Figura 5.2 apresenta os resultados obtidos pelos modelos. Como pode ser observado, à medida que é aumentada a mistura entre as classes, os pares que se destacam positivamente são wkNN-MinCut, wkNN-GFHF, kNN-GFHF, wϵRd-LGC, ϵRd-LGC, wϵRd-PCC e ϵRd-PCC. Métodos de construção de redes diversos melhor se ajustam aos métodos de rotulação. MinCut e GFHF funcionam melhor com os métodos baseados em vizinhos mais próximos, enquanto que os métodos LGC e PCC funcionam melhor com os métodos radiais. Apenas com o método PCC o método de construção de redes completas (Co) não apresentou o pior desempenho.

A pequena superioridade de desempenho dos métodos com pesos em relação aos métodos sem pesos evidencia que, embora mais custosas, redes ponderadas podem não oferecer ganhos

significativos de desempenho. Contudo, ainda assim, os pares ponderados foram selecionados para as simulações a seguir.

A variação de desempenho de acordo com o método de construção de rede, por sua vez, evidencia a existência de vieses internos nos métodos de rotulação, que se adequam melhor a tipos distintos de redes.

5.3 Comparação entre métodos de propagação de rótulo

Dentre os pares de métodos destacados na seção anterior, foram selecionados os pares wkNN-MinCut, wkNN-GFHF, wϵRd-LGC e wϵRd-PCC para serem analisados comparativamente. Nesta seção eles serão avaliados de forma mais extensa, considerando, além do ARI, as medidas de silhueta e modularidade. Adicionalmente, os desempenhos de referência dos classificadores H_0 e H_1 também são apresentados.

A Simulação I tem a finalidade de avaliar o desempenho dos métodos de aprendizado em relação à dificuldade do problema de aprendizado. Quanto maior S, maior a dispersão das classes e maior a região de sobreposição entre as classes. Como mostra a Figura 5.3, os pares wϵRd-LGC, wkNN-GFHF e wkNN-MinCut apresentam desempenho similar nesse caso. Para $S = 1.6$, de acordo com o ARI, o modelo wϵRd-LGC apresenta desempenho um pouco melhor que wkNN-GFHF e wkNN-MinCut. De acordo com a medida de silhueta, no entanto, os modelos wkNN-MinCut e wϵRd-LGC apresentam melhor desempenho que o modelo wkNN-GFHF. De acordo com a modularidade, por sua vez, as três primeiras posições do ranque são quase que indistinguivelmente ocupadas pelos mesmos 3 modelos. Em todos os casos o modelo wϵRd-PCC apresenta o pior desempenho. Considerando o viés de cada medida de validação, podemos dizer que os modelos wϵRd-LGC e wkNN-MinCut preterem a consistência com os grupos do conjunto de dados, enquanto que o modelo wkNN-GFHF pretere a consistência com as classes.

Consideremos então as Simulações II e III: fixados o parâmetro de mistura, respectivamente, $S = 1$, e $S = 1.6$, e a fração de objetos previamente rotulados, $L = 0.02 \times n$, variemos apenas o número de objetos, $n = \{100, 200, 400, 600, 800, 1000, 1500, 2000\}$, de modo que a densidade interna dos grupos é variada. Quanto menor o número de objetos, menor a densidade intragrupo, e consequentemente menor a proporção da densidade intragrupo em relação à densidade nas regiões entre os grupos. Em outras palavras, os dados não satisfazem a suposição de suavidade. Na Simulação III, onde $S = 1.6$, também existe alta sobreposição entre as classes, não sendo satisfeita também a suposição de grupos (vide Figura 5.1).

Como mostra a Figura 5.4, tanto para $S = 1$ como para $S = 1.6$, o desempenho dos pares de métodos é melhor quanto maior o número de objetos do conjunto de dados. Para conjuntos de dados pequenos ($n < 1000$), tanto o desempenho médio é depreciado, como também é aumentado o desvio padrão, validando portanto a necessidade de que os dados satisfaçam a suposição de suavidade. Além disso, como na Simulação I, os resultados para $S = 1.6$ também são inferiores aos resultados para $S = 1$. Em todo o caso, no entanto, o número de objetos não influi no ranque de desempenho. O modelo wϵRd-LGC continua a apresentar o melhor

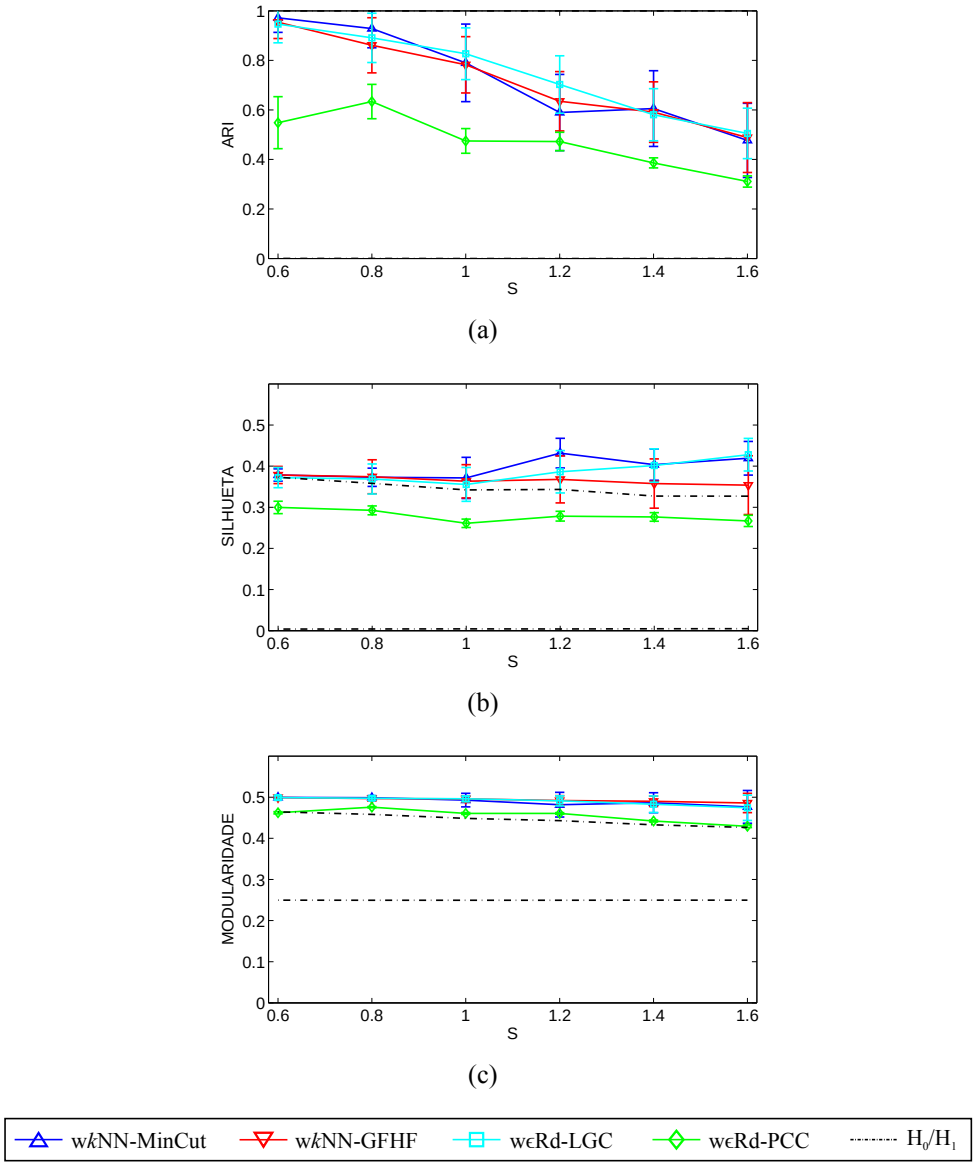


Figura 5.3: Resultado médio e desvio padrão obtidos pelos modelos wkNN-MinCut, wkNN-GFHF, wϵRd-LGC e wϵRd-PCC na Simulação I.

desempenho em termos de ARI, enquanto que o modelo wϵRd-PCC continua a apresentar o pior desempenho.

Mais relevante para os propósitos da presente pesquisa, avaliemos o desempenho dos modelos de aprendizado em função do tamanho da amostra conforme a Simulação IV. Essa discussão é particularmente importante no sentido de avaliar a robustez dos modelos de aprendizado em função da qualidade da amostra. Em se tratando do desempenho médio com amostras diversas, o tamanho da amostra é um indicador de sua qualidade. Quanto menor L, menos confiável a amostra.

Como mostra a Figura 5.5, quanto maior o tamanho da amostra, melhor o desempenho do aprendizado. A diferença entre os desempenhos de wϵRd-PCC e dos demais modelos também diminui significativamente. Tomando como referência a medida de silhueta, notavelmente os desempenhos de wkNN-MinCut, wkNN-GFHF e wϵRd-PCC aproximam-se do desempenho de referência (H_1), reafirmando a maior consistência dos respectivos métodos de rotulação com as

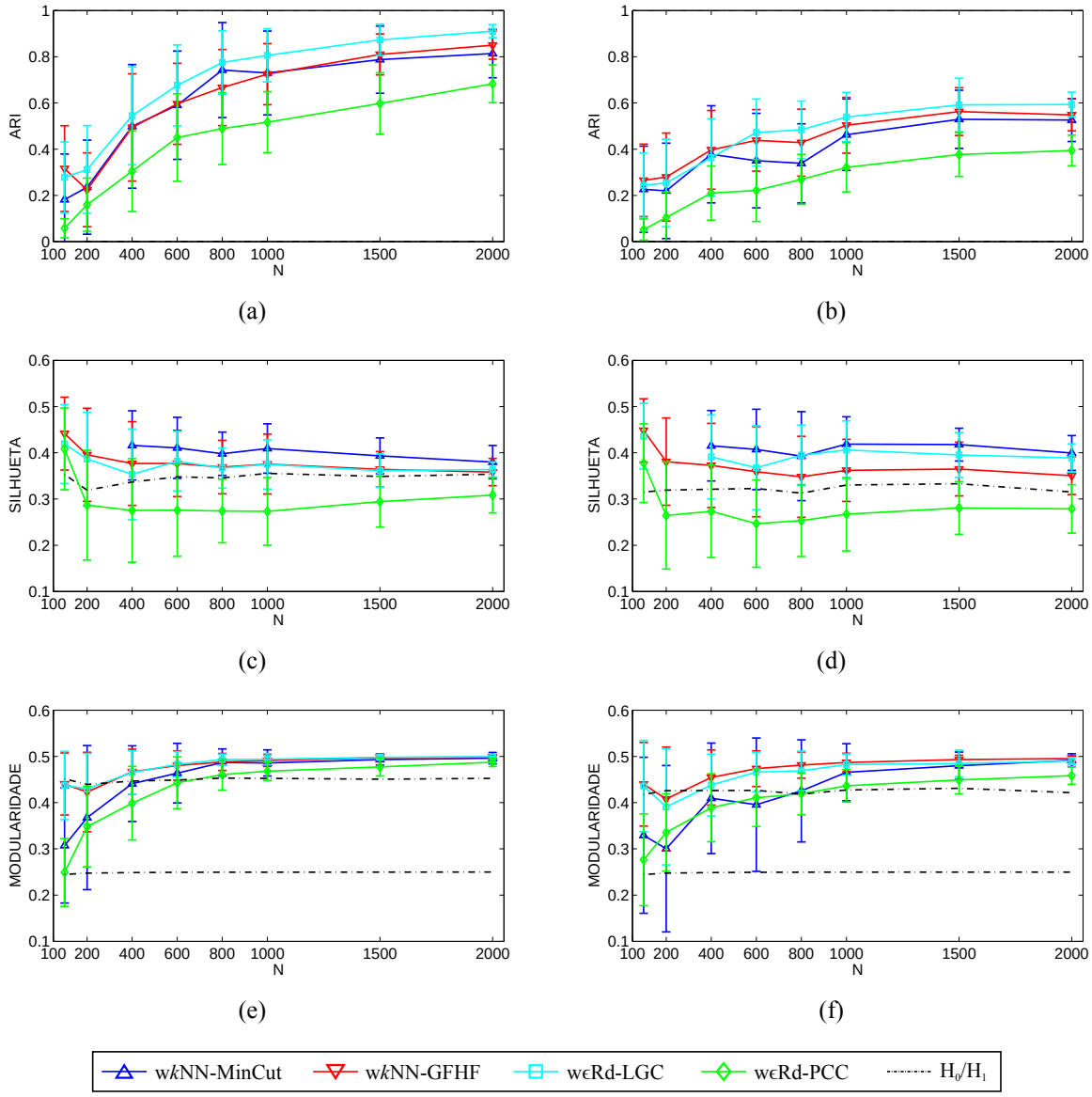


Figura 5.4: Resultado médio e desvio padrão obtidos pelos modelos wkNN-MinCut, wkNN-GFHF, wϵRd-LGC e wϵRd-PCC nas Simulações II e III.

classes, em detrimento da consistência com os grupos.

Novamente, o desempenho de wϵRd-LGC é diferenciado. Nesse caso, no entanto, para $L > 40$ o desempenho do modelo é significativamente depreciado em termos do ARI, porém, em termo da silhueta, o seu desempenho é superior aos desempenhos dos demais modelos. Em outras palavras, wϵRd-LGC pretere a consistência com os grupos em detrimento da consistência com as classes, resultado que se justifica pelo valor atribuído ao parâmetro de balanceamento entre o termo de regularização e a função de erro do método LGC, $\lambda = \frac{0.99}{1-0.99}$.

Para valores pequenos de L, o desempenho médio dos modelos é depreciado e maior é a variação de desempenho entre os ensaios, destacando assim a dependência que o aprendizado semissupervisionado tem da qualidade da amostra rotulada, especialmente quando poucos itens de dados estão disponíveis rotulados. Esse ponto será discutido mais a fundo nos próximos capítulos.

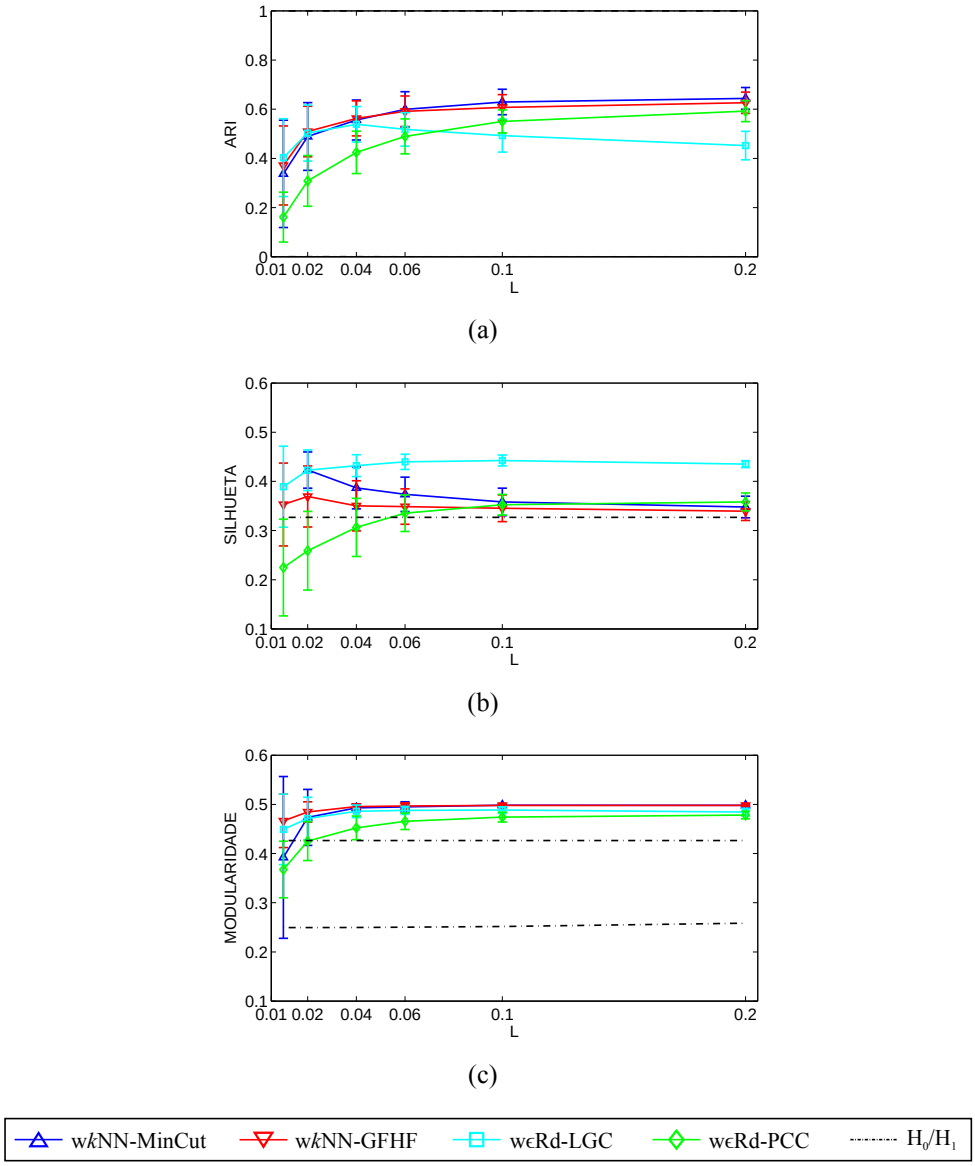


Figura 5.5: Resultado médio e desvio padrão obtidos pelos modelos wkNN-MinCut, wkNN-GFHF, wϵRd-LGC e wϵRd-PCC na Simulação IV.

5.4 Restrições de ligações de acordo com conhecimentos prévios

Em todas as simulações, diferente do que se esperava a princípio, o método PCC obteve desempenho inferior aos demais métodos estudados. Definidos de acordo com os valores sugeridos por Breve et al. (2012), certamente o melhor ajuste de p_{grd} e Δ_v não melhoraria em muito o desempenho do método. Contudo, outro parâmetro externo foi definido arbitrariamente, a tolerância do critério de parada. Como sugerido pelos autores, uma abordagem para a verificação da convergência do método é avaliar a taxa de crescimento da pertinência dos nós aos times de partículas, que tende a diminuir à medida que os times de partículas dominam comunidades da rede. Em relação a isso, observamos que mesmo depois de muitas iterações esse índice oscila um pouco, caracterizando a mudança do comportamento predominante de pelo menos um dos times

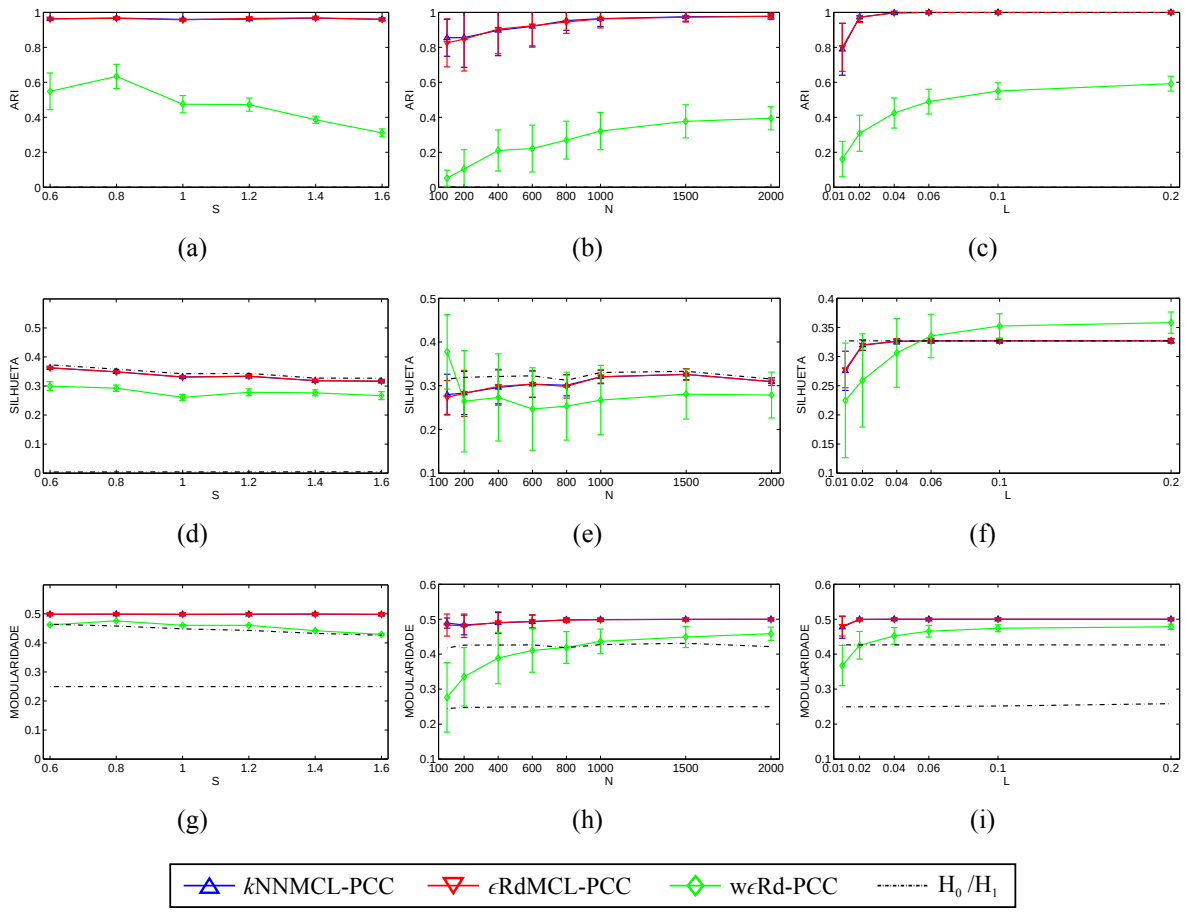


Figura 5.6: Resultados obtidos pelos modelos kNNMCL-PCC, ϵRdMCL-PCC e wϵRd-PCC nas simulações I, III e IV. Em (a), (d) e (g) para $n = 1000$, $L = 20$ e $S = [0.6 : 0.2 : 1.6]$; em (b), (e) e (h) para $n = \{100, 200, 400, 600, 1000, 2000\}$, $L = 0.02 \times n$ e $S = 1.6$; e em (c), (f) e (i) para $n = 1000$, $L = \{0.01, 0.02, 0.04, 0.06, 0.1, 0.2\} \times n$ e $S = 1.6$. São apresentados média e desvio padrão de 80 ensaios - 20 ensaios com divisões aleatórias entre rotulados e não rotulados para cada uma de 4 variações do conjunto de dados.

de partículas, do comportamento defensivo para o comportamento explorativo. Esse fenômeno é bastante interessante no sentido de permitir que a solução obtida fuja de mínimos locais. No entanto, se o parâmetro de convergência não é definido de forma bastante específica, ao atingir um mínimo local (equilíbrio instável), o algoritmo pode ser encerrado, o que parece ter acontecido com razoável frequência nas simulações anteriores. Como as discussões realizadas até o presente momento foram suficientemente produtivas, optamos por não reexecutar as simulações, mas, ao invés disso, realizar as simulações I, III e IV com os métodos de construção de rede ϵRdMCL e kNNMCL combinados a PCC.

A Figura 5.6 reapresenta os resultados obtidos por wϵRd-PCC nas simulações I, III e IV em contraste com os resultados obtidos por ϵRdMCL-PCC e kNNMCL-PCC. ϵRdMCL-PCC e kNNMCL-PCC obtém resultados muito bons, próximo do desempenho de referência do classificador exato (H_1). As modificações resultantes das restrições de ligação simplificam significativamente o problema de classificação. Arestas entre nós com rótulos distintos são eliminadas e arestas entre nós com o mesmo rótulo são estabelecidas, em alguns casos ligando nós amostrados das regiões de sobreposição entre as classes com nós centrais. Essa solução

se ajusta muito bem à dinâmica de método PCC. As partículas passam a dispor de atalhos até a região de segurança de seu time e por isso conseguem reforçar sua energia e dominar seu território de forma mais eficiente. E para tanto, nem mesmo é necessário o ajuste do parâmetro de convergência. Atalhos até regiões dominadas por partículas adversárias também são eliminados analogamente.

5.5 Considerações finais

Nesse capítulo, métodos diversos de construção de redes e de rotulação foram estudados tomando como critério de avaliação medidas de validação externa, validação relativa de agrupamento, e validação de detecção de comunidades. Casamentos diversos de métodos de construção de redes e métodos de rotulação apresentaram bom desempenho. No entanto, em se tratando da tarefa de classificação, a construção da rede considerando restrições para o estabelecimento de ligações entre nós rotulados, possibilitou ganho significativo de desempenho, tornando a rotulação semissupervisionada muito mais eficiente. Além disso, pudemos observar a degeneração do aprendizado semissupervisionado quando não são satisfeitas as suposições de suavidade e de agrupamento, bem como quando poucos itens de dados estão disponíveis rotulados. O desvio padrão dos resultados das simulações é significativo nesse caso.

Um aspecto que precisa ser destacado é o seguinte. Em função da definição de seus parâmetros e de seu viés interno, métodos de rotulação podem preterir a consistência com as classes ou a consistência com os grupos, como mostram os resultados acima. Se classes e grupos não são consistentes, isto é, a suposição de suavidade não é satisfeita, os resultados oferecidos pelos métodos podem não atender à necessidade específica da aplicação. Em se tratando da um tarefa de classificação, obviamente, é interessante que exista consistência com as classes; agrupamento, por sua vez, com os grupos. No entanto, de um ponto de vista mais amplo, diríamos, do ponto de vista do aprendizado semissupervisionado, faz-se necessária uma análise reflexiva sobre os propósitos desse paradigma de aprendizado. Não seria o caso de se manter um meio termo entra a consistência com ambos, classes e grupos? Essa é uma questão que precisa ser melhor estudada e discutida.

Capítulo

6

Caracterização de indivíduos representativos em redes

No Capítulo 5, vimos como a qualidade da amostra disponível rotulada influencia no desempenho do aprendizado semissupervisionado (ASS) baseado em redes. Esse ponto será estudado mais a fundo nas próximas seções.

Uma abordagem conveniente para a identificação de boas sementes para o aprendizado semissupervisionado baseado em redes é primeiro identificar as comunidades da rede, e, em seguida, caracterizar os representativos de cada comunidade (Motta et al., 2009). No entanto, apresentamos uma hipótese que vai um pouco além desse tipo de abordagem. Para nós a representatividade é um fenômeno diretamente relacionado à formação de comunidades em redes. Isto é, as estruturas de comunidades em uma rede são resultado da organização de seus indivíduos em torno de seus representativos. E por isso, assumimos que boas estruturas de rede devem realçar nós representativos. Logo, propomos critérios de caracterização de nós representativos em redes, tomando como base índices de centralidade, desconsiderando qualquer informação adicional sobre as estruturas de grupos ou comunidades.

Apesar das diferentes denominações, indivíduos representativos são estudados em diversos problemas de redes complexas, tais como resiliência e epidemiologia (Albert et al., 2000; Holme et al., 2002; Pastor-Satorras & Vespignani, 2002). Redes livres de escala, por exemplo, são conhecidamente resilientes a ataques randômicos, porém bastante sensíveis a ataques direcionados sobre os *hubs* (Albert et al., 2000; Cohen et al., 2000, 2001; Jeong et al., 2001; Newman et al., 2002). *Hubs*, nesse caso, podem ser ditos representativos. Outros resultados destacam nós com alta intermedialidade e nós pertencentes a núcleos conectivos das redes (k-*cores*) como espalhadores influentes (Holme et al., 2002; Kitsak et al., 2010; Chen et al., 2012). Em se tratando de dinâmicas competitivas, como propagação de rumores e disputas por consumidores,

outros comportamentos são observados (Borge-Holthoefer & Moreno, 2012; Borge-Holthoefer et al., 2012, 2013; Goyal et al., 2014; Galeotti & Goyal, 2009).

Nesse capítulo estudamos fenômenos semelhantes em redes com estruturas de comunidades bem definidas, a saber, redes de Girvan-Newman e Lancichinetti-Fortunato-Radicchi (vide Seção 4.3.4). A propagação de rótulos pode ser entendida como uma dinâmica de rede, considerando, no caso, os métodos campos Gaussianos e funções harmônicas, consistência local e global, e competição e cooperação de partículas (vide 2.4). Os dois primeiros métodos realizam dinâmicas difusivas, enquanto que o último, dinâmica competitiva. Além do desempenho do ponto de vista do aprendizado de máquina, *insights* com relação a dinâmicas em redes podem ser extraídos. Critérios de caracterização de nós centrais diversos são estudados no sentido de caracterizar boas sementes para o processo de rotulação semissupervisionado, destacando-se os critérios que caracterizam os nós da rede de forma mais heterogênea. Adicionalmente, a técnica análise de componentes principais é utilizada para agregar índices de centralidade no sentido de propor critérios unificados de caracterização de representativos, pois critérios diversos se destacam em redes de modelos diversos.

6.1 Variabilidade de desempenho de aprendizado

Tanto em métodos clássicos de aprendizado de máquina como na natureza podemos observar como processos de aprendizagem dependem da qualidade de suas amostras. Como um exemplo, a má inicialização dos centroides no k-médias pode levar a uma solução pobre no hiperespaço de soluções (Jain, 2010; MacQueen, 1967; Jain & Dubes, 1988). Estendemos um pouco mais o estudo dessa questão nessa seção.

Consideremos os métodos de rotulação semissupervisionada campos Gaussianos e funções harmônicas (GFHF), consistência local e global (LGC) e competição e cooperação de partículas (PCC). Além da rede e itens de dados rotulados, LGC recebe como entrada um parâmetro externo, α, utilizado para balancear o termos de regularização e erro do método. Consideremos um nó rotulado qualquer da rede. Quanto maior α, maior a influência de seus vizinhos no seu rótulo final. GFHF, por outro lado, não permite a correção de rótulos de dados rotulados *a priori* inconsistentes em relação à sua vizinhança. Portanto suas únicas entradas são a rede e os nós rotulados. PCC, por sua vez, realiza uma dinâmica mais complexa, difícil de ser modelada em termos do *framework* de regularização. É realizado um processo de competição e cooperação entre partículas que caminham na rede fazendo movimentos ora aleatórios, ora gulosos. Movimentos gulosos caracterizam comportamento defensivo, enquanto que movimentos aleatórios, comportamento explorativo, sendo o parâmetro p_{grd} responsável pelo balanço entre os comportamentos. Quanto maior p_{grd}, mais defensivo o comportamento das partículas. Adicionalmente, o parâmetro Δ_v controla a intensidade da competição entre as partículas. Quanto maior Δ_v, mais agressivo o efeito da visita de uma partícula a um nó. α, p_{grd} e Δ_v são definidos no intervalo [0, 1].

Dadas como entrada redes de Girvan-Newman com 1024 nós, 4 comunidades bem definidas, grau médio $\langle k \rangle = 16$ e $\text{Z}_{\text{out}} = 2$ e os rótulos de amostras de dados selecionadas randomicamente,

cobrindo 2% dos nós (em média 20 nós), para cada configuração de algoritmo de rotulação e parâmetros externos, realizamos 900 ensaios (30 amostras de dados rotulados em 30 realizações da rede). Os resultados são apresentados em termos do índice de Rand ajustado, conforme apresenta a Seção 2.5. Como os rótulos *ground truth* de referência correspondem às comunidades, o índice de modularidade é bastante correlacionado ao ARI.

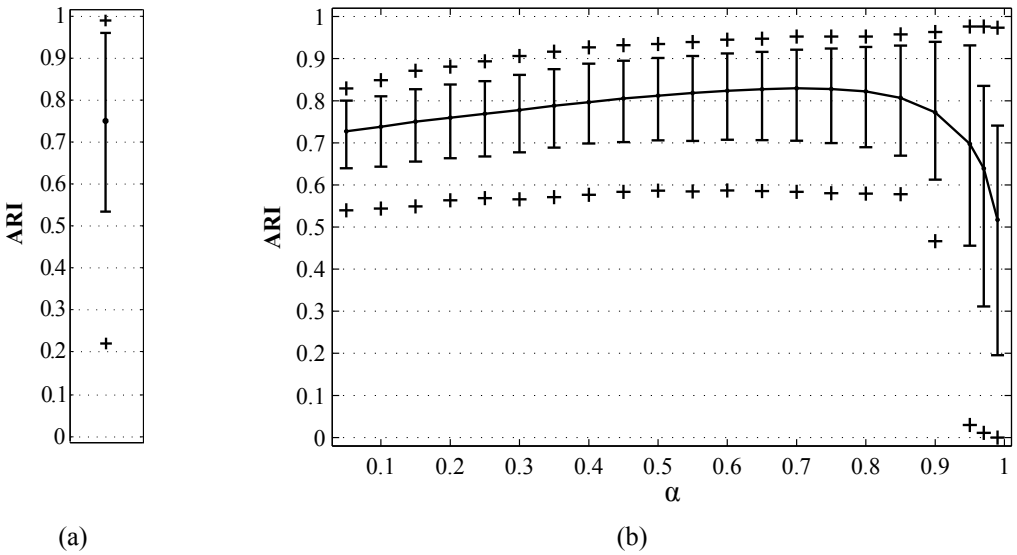
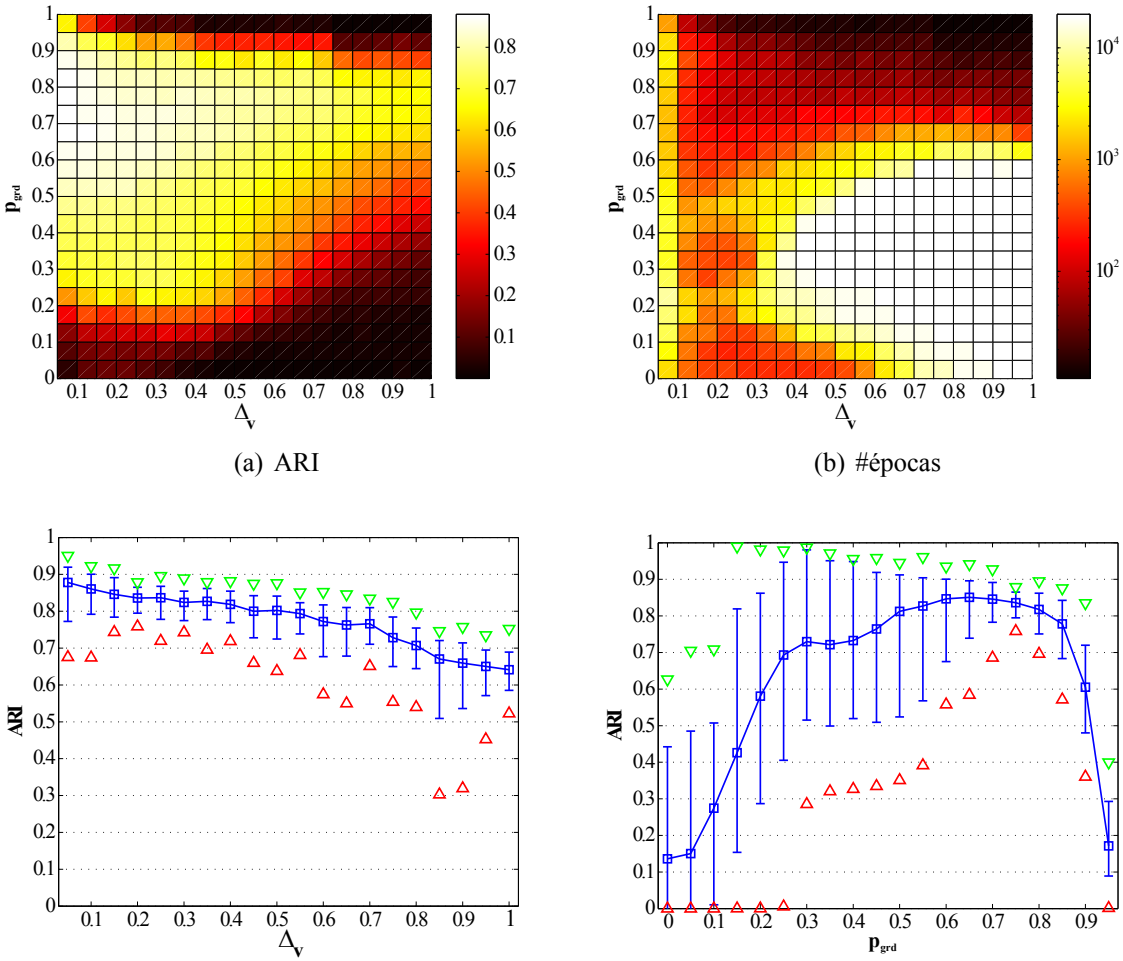


Figura 6.1: Desempenho dos métodos GFHF (em (a)) e LGC (em (b)) em relação à amostra disponível rotulada e, no caso do LGC, ao ajuste de α. São apresentados ARI médio, desvio padrão sobre a (e abaixo da) média e pior e melhor desempenho observados em 900 ensaios (símbolo '+').

Como mostra a Figura 6.1, o desempenho de GFHF varia significativamente em razão da qualidade das amostras. Enquanto que o desempenho médio resulta ARI $= 0.7146$, o melhor desempenho resulta ARI $= 0.9920$ e o pior, ARI $= 0.2213$. Variando a amostra rotulada e α, o desempenho de LGC também varia significativamente. Bons desempenhos médios são observado para $0.6 < \alpha < 0.8$. Para $\alpha > 9$ os melhores e piores desempenhos são observados; também o desvio padrão observado é muito maior que para $\alpha < 0.8$. Reiterando os resultados apresentados no Capítulo 5, valores altos de α favorecem à consistência com grupos, em detrimento das classes, resultando, no caso, no alagamento de comunidades com pouco dados disponíveis rotulados pelos rótulos difundidos a partir de outras comunidades. Em termos de desempenho de classificação, podemos dizer que, embora menos arriscado, definir $0.6 < \alpha < 0.8$ restringe o desempenho de LGC. Notavelmente, amostras podem ser extraídas da rede e rotuladas de modo a possibilitar que LGC apresente desempenho consistente tanto com os grupos como com as classes, inclusive para $\alpha > 9$.

A Figura 6.2 apresenta o ARI médio e desvio padrão obtidos pelo método PCC, bem como o número médio de iterações necessário até a convergência do algoritmo. Variando-se a amostra disponível rotulada, p_{grd} e Δ_v, temos variações significativas de desempenho. Quanto menor Δ_v, melhor o desempenho médio, caracterizando a maior suavidade do processo competitivo; p_{grd}, por outro lado, resulta em bom desempenho médio no intervalo [0.6, 0.8]. Diferente do



(a) ARI

(b) #épocas

(c) $p_{\text{grd}} = 0.75$

(d) $\Delta_v = 0.2$

Figura 6.2: Desempenho do método PCC em relação à amostra disponível rotulada e o ajuste dos parâmetros p_{grd} e Δ_v. Em (a) é apresentado o ARI médio; em (b) o número de iterações até a convergência do algoritmo; e em (c) e (d) secções transversais da superfície de desempenho para $p_{\text{grd}} = 0.75$ e $\Delta_v = 0.2$, respectivamente. Desempenhos médios consistem na média de 900 ensaios. Em (c) e (d) são apresentados adicionalmente desvio padrão sobre a (e abaixo da) média e pior ($\triangle$) e melhor ($\triangledown$) desempenho para cada configuração.

que pode-se pensar, a convergência do algoritmo é mais demorada para valores grandes de Δ_v e valores pequenos de p_{grd}, pois a competição entre as partículas é mais agressiva, resultando em reiteradas batalhas pelo domínio de territórios fronteiriços. Garantias de bom desempenho, no entanto, são oferecidas apenas para $p_{\text{grd}} = 0.75$ e $\Delta_v = 0.2$, onde o desvio padrão do desempenho é reduzido. Fixado $\Delta_v = 0.2$, desempenhos melhores podem ser obtidos quando dada maior capacidade de exploração às partículas, especificamente, para valores de p_{grd} entre 0.15 e 0.3. Nesse caso faz-se necessário, contudo, que uma amostra representativa esteja disponível rotulada.

Podemos concluir portanto que, além da definição dos parâmetros externos, a qualidade da amostra disponível rotulada enviesa o desempenho do aprendizado semissupervisionado baseado em redes, e, portanto, não apenas parâmetros externos devem ser ajustados, mas boas amostras dos dados também precisam ser selecionadas e rotuladas *a priori*. Esses resultados dão razão ao estudo apresentado nas próximas seções no sentido de propor critérios para a seleção de

representativos para semear o aprendizado semissupervisionado.

6.2 Caracterização de representativos

Medidas de centralidade são capazes de caracterizar nós localmente, em relação aos seus vizinhos, regionalmente e até mesmo em relação à rede como um todo. Com fins à caracterização de nós representativos, estudamos as seguintes medidas de centralidade em redes: grau (deg), eficiência local (eff), coeficiente de agrupamento (CC), intermedialidade (btw), ranque de Page (PgRC), variação hierárquica de grau (H_{deg}) e variação hierárquica de coeficiente de agrupamento (H_{CC}). Suas descrições podem ser encontradas também em (Araújo & Zhao, 2013a,b; Rubinov & Sporns, 2010; Brandes & Erlebach, 2005; Costa & Silva, 2006), bem como na Seção 4.2. A seguir apresentamos a caracterização de nós em uma rede de associações entre golfinhos nariz de garrafa compilada por Lusseau et al. (2003).

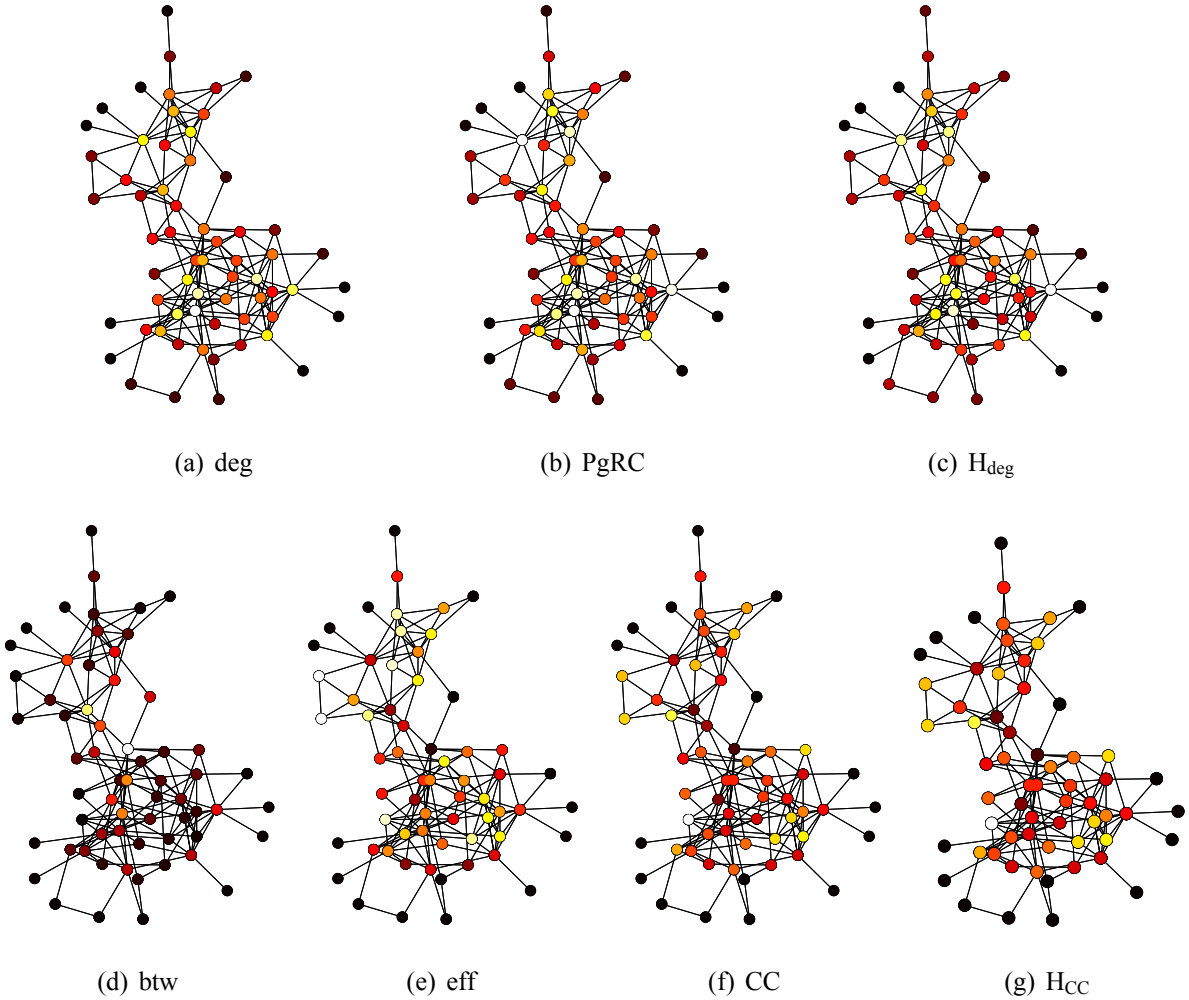


Figura 6.3: Rede de associações entre golfinhos com os nós caracterizados por medidas de centralidade em rede (Lusseau et al., 2003). Quanto mais clara a cor do nó, maior o índice atribuído pela medida. O espalhamento visual da rede é realizado através do método de minimização de energia de um sistema de molas entre todos pares de vértices, cuja rigidez da mola entre cada par de vértices é proporcional à distância mínima entre os mesmos (Kamada & Kawai, 1989).

Composta por 62 golfinhos que vivem em Doubtful Sound, Nova Zelândia, essa rede é tipicamente descrita como possuindo duas comunidades, uma delas com 3 subcomunidades, que caracterizam homofilia (Lusseau & Newman, 2004; Lusseau, 2003). A Figura 6.3 mostra instâncias dessa rede onde os nós são coloridos de acordo com os índices de centralidade atribuídos por deg, eff, CC, btw, PgRC, H_{deg}, H_{CC}. Quanto mais clara a cor do nó, maior o índice atribuído pela medida. Pode-se observar a distribuição das duas principais comunidades através da representação visual. Destaque-se a alta correlação entre os índices das medidas baseadas no grau (deg, PgRC e H_{deg}) e a disposição dos nós com alta intermedialidade na região entre as duas principais comunidades. A Figura 6.4 (a) mostra adicionalmente a distribuição cumulativa de cada medida. Notavelmente, btw provê uma caracterização mais heterogênea dos nós. Esses aspectos serão discutidos mais à frente.

6.3 Métodos e análise numérica

No sentido de avaliar a performance da rotulação semissupervisionada semeada por representativos, primeiramente consideramos as medidas apresentadas acima, deg, eff, CC, btw, PgRC, H_{deg}, H_{CC}. Cada medida é considerada um critério para a caracterização de representativos. Quanto maior o índice recebido pelo nó, maior a sua representatividade de acordo com o respectivo critério.

Novamente, consideramos os métodos de propagação de rótulos GFHF, LGC e PCC. Com base nos resultados apresentados na Seção 6.1, definimos os parâmetros externos, $\alpha = 0.7$, $\Delta_v = 0.2$ e $p_{\text{grd}} = 0.75$, de modo a favorecer o desempenho da abordagem de seleção de sementes através de amostragem aleatória. Bastante correlacionados, a modularidade e o índice de Rand ajustado, optamos por apresentar os resultado considerando apenas o ARI (vide Seção 2.5) conforme apresentado na Seção 2.5. Destaque-se que um classificador aleatório equiprovável obtém valores de ARI próximos de 0, enquanto que o classificador exato, ARI=1. Além de redes de Girvan-Newman, utilizamos também redes Lancichinetti-Fortunato-Radicchi como *benchmark*. Por ora, evitamos o viés dos métodos de construção de rede.

Foram geradas redes randômicas com 1024 nós, 4 comunidades, grau médio $\langle k \rangle = 16$, e $Z_{out} = \{1, 2, \ldots, 8\}$ através do modelo GN. A distribuição de grau das redes resultantes é bastante homogênea. Por outro lado, foram geradas redes modulares com 1024 nós, grau médio $\langle k \rangle = 16$, distribuição de grau heterogênea que segue a lei de potência com expoente $\gamma = 2$, e comunidade de tamanhos variados considerando distribuição homogênea ($\beta = 1$), e $\mu = [0.1 : 0.05 : 0.5]$ como o modelo LFR. Os parâmetros Z_{out} e μ controlam o nível de mistura entre as comunidades. Quanto maior Z_{out} ou μ maior a sobreposição entre as mesmas (vide 4.3.4).

Para cada configuração de rede são geradas 30 redes. Não determinísticos, a cada realização os modelos produzem uma rede distinta, com pequenas variações em relação às definições de parâmetros. Determinísticos, cada critério de seleção de representativos é aplicado uma única vez em cada rede. Os resultados apresentados são a média dos desempenhos obtidos no conjunto

de redes gerado para cada configuração. Para comparação, amostras também são selecionadas ao acaso. Trinta ensaios são realizados em cada rede, totalizando 900 ensaios. O ARI médio é reportado como desempenho equivalente a um *ensemble* de classificadores semeados por amostras randômicas.

Os ensaios foram realizados considerando apenas 2% dos dados disponíveis rotulados (em média 20 itens de dados) e não garantem que nós de todas as classes sejam selecionados. Esse cenário é equivalente àquele encontrado no mundo real, onde nem rótulos nem distribuições de classes são conhecidos *a priori*. Depois de selecionados, no entanto, ambos dados selecionados aleatoriamente e segundo os critérios propostos são rotulados confidentemente. Contudo, como são consideradas desde redes com comunidades bem definidas até redes sobreposição significativa entre as classes, a consistência de grupos nem sempre é satisfeita.

6.4 Resultados e discussões

Denominamos a propagação dos rótulos de uma classe para além das comunidades correspondentes em redes com comunidades bem definidas, inundação ou intrusão. Em casos extremos observamos até todos os nós a priori não rotulados sendo rotulados com rótulos de uma única classe. Uma vez que as amostras estudadas não são estratificadas, quando existe um desbalanço significativo entre o número de nós rotulados de cada classe, observamos a ocorrência desse fenômeno para os métodos LGC e GFHF. Esta é a principal razão para LGC e GFHF apresentarem desempenho significativamente inferior a PCC (vide Tabela 6.1). Com um bom ajuste de seus parâmetros externos, PCC se mostra bastante robusto. Isso se justifica na dinâmica de competição entre as partículas, que suprimem intrusões/inundações mesmo quando seu time possui número de partículas (nós rotulados) inferior aos demais times (classes).

Como mostra a Tabela 6.1, contraintuitivamente, rotular nós com alto grau não é uma estratégia confiável para semear a propagação de rótulos em redes de Girvan-Newman e Lancichinetti-Fortunato-Radicchi. Apenas LGC e GFHF conseguem obter melhores desempenhos utilizando esse critério de que semeados por amostras aleatórias e apenas em redes LFR. O critério PgRC, da mesma forma, apenas se destaca em redes LFR semeando os métodos LGC e GFHF.

A variação hierárquica de grau (H_{deg}), por outro lado, destaca-se significativamente entre os melhores critérios para a seleção de representativos em redes de ambos os tipos. Esse resultado mostra que mais importante do que ser muito conectivo é ser razoavelmente conectivo em vizinhanças pouco conectivas. Os critérios baseados no coeficiente de agrupamento (CC e H_{CC}) e na intermedialidade (btw), por sua vez, proporcionam os melhores desempenhos dentre os critérios estudados, respectivamente, em redes GN e LFR. Nós com alto coeficiente de agrupamento em redes GN estão tipicamente situados em *motifs* do tipo quase-clique, emergentes da pseudo-aleatoriedade computacional das ligações estabelecidas pelo modelo. Nós com alta intermedialidade em redes LFR são quase-*hubs* situados na borda das comunidades, interligando-as. Em ambos os casos, os nós identificados são excelentes regularizadores. Nós com alta eficiência local, por

Tabela 6.1: ARI médio e ranque médio obtidos por métodos de propagação de rótulos semeados por nós realçados por critérios diversos de seleção de representativos.

	GN N=1024, C=4, $\langle k \rangle$=16, Z_{out}								Ranque	LFR N=1024, $\langle k \rangle$=16, τ_1=2, τ_2=1, μ									Ranque
	1	2	3	4	5	6	7	8	Médio	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	Médio
GFHF sobre																			
ensemble	*0.90*	*0.68*	*0.50*	*0.36*	*0.24*	*0.14*	*0.08*	*0.05*	***3.75***	*0.67*	*0.57*	*0.44*	*0.35*	*0.29*	*0.23*	*0.15*	*0.11*	*0.06*	***4.33***
top deg	0.84	0.62	0.39	0.23	0.19	0.09	0.07	0.04	**6**	0.63	0.54	0.55	0.50	0.33	0.26	0.27	0.20	0.15	**3.33**
top eff	0.85	0.58	0.43	0.35	0.21	0.16	0.11	0.03	**4.62**	0.18	0.05	0.11	0.12	0.05	0.04	0.09	0.09	0.05	**6.11**
top CC	0.90	0.69	0.54	0.45	0.31	0.24	0.09	0.06	**2.37**	0.18	0.05	0.11	0.12	0.05	0.04	0.09	0.09	0.05	**6.11**
top btw	0.93	0.76	0.46	0.34	0.26	0.11	0.06	0.03	**4.12**	0.77	0.66	0.60	0.55	0.38	0.30	0.28	0.23	0.15	**1.11**
top PgRC	0.92	0.75	0.50	0.30	0.23	0.13	0.06	0.05	**4.25**	0.67	0.58	0.57	0.51	0.35	0.28	0.27	0.20	0.16	**2.44**
top H_{deg}	0.95	0.75	0.55	0.36	0.26	0.16	0.12	0.03	**2.5**	0.74	0.61	0.60	0.53	0.36	0.28	0.27	0.21	0.16	**1.78**
top H_{CC}	0.92	0.82	0.50	0.45	0.27	0.23	0.16	0.08	**1.75**	0.03	0.01	0	0.03	0.01	0.01	0.04	0.04	0.04	**7.11**
LGC sobre																			
ensemble	*0.94*	*0.85*	*0.72*	*0.57*	*0.40*	*0.25*	*0.15*	*0.08*	***3.87***	*0.70*	*0.63*	*0.46*	*0.36*	*0.29*	*0.23*	*0.15*	*0.11*	*0.06*	***4.89***
top deg	0.89	0.79	0.61	0.50	0.37	0.25	0.15	0.08	**5.62**	0.66	0.61	0.62	0.58	0.43	0.38	0.32	0.28	0.18	**3.67**
top eff	0.96	0.84	0.73	0.62	0.42	0.28	0.18	0.07	**2.75**	0.46	0.33	0.30	0.22	0.15	0.17	0.13	0.13	0.06	**5.89**
top CC	0.96	0.87	0.79	0.63	0.46	0.30	0.16	0.09	**1.5**	0.46	0.33	0.30	0.22	0.15	0.17	0.13	0.13	0.06	**5.89**
top btw	0.89	0.80	0.66	0.54	0.41	0.24	0.15	0.07	**5.25**	0.83	0.76	0.72	0.68	0.52	0.46	0.38	0.32	0.19	**1**
top PgRC	0.92	0.86	0.67	0.55	0.40	0.26	0.17	0.08	**4**	0.73	0.64	0.66	0.60	0.47	0.40	0.33	0.28	0.18	**2.67**
top H_{deg}	0.94	0.86	0.71	0.56	0.43	0.28	0.18	0.08	**3**	0.80	0.69	0.70	0.62	0.47	0.42	0.33	0.29	0.19	**1.89**
top H_{CC}	0.97	0.92	0.76	0.60	0.42	0.29	0.18	0.09	**1.75**	0.12	0.05	0.09	0.10	0.07	0.03	0.06	0.07	0.05	**7.22**
PCC sobre																			
ensemble	*0.97*	*0.93*	*0.85*	*0.73*	*0.56*	*0.36*	*0.21*	*0.10*	***3.25***	*0.88*	*0.85*	*0.78*	*0.68*	*0.57*	*0.45*	*0.31*	*0.22*	*0.13*	***2.22***
top deg	0.93	0.89	0.76	0.68	0.52	0.35	0.21	0.10	**5.25**	0.68	0.64	0.63	0.58	0.49	0.40	0.32	0.25	0.15	**4.33**
top eff	0.97	0.90	0.84	0.77	0.57	0.40	0.25	0.12	**2.37**	0.54	0.40	0.39	0.34	0.28	0.24	0.19	0.16	0.09	**6.44**
top CC	0.97	0.92	0.89	0.77	0.62	0.43	0.23	0.12	**1.62**	0.54	0.39	0.40	0.34	0.30	0.23	0.17	0.15	0.09	**6.67**
top btw	0.97	0.93	0.82	0.72	0.54	0.36	0.20	0.10	**4**	0.84	0.78	0.72	0.69	0.56	0.48	0.38	0.28	0.18	**1.44**
top PgRC	0.96	0.95	0.81	0.71	0.55	0.37	0.23	0.11	**3.5**	0.74	0.66	0.66	0.61	0.52	0.43	0.34	0.26	0.15	**3.33**
top H_{deg}	0.98	0.95	0.84	0.73	0.57	0.37	0.22	0.11	**2.62**	0.81	0.71	0.71	0.63	0.51	0.44	0.34	0.27	0.18	**2.67**
top H_{CC}	0.99	0.95	0.88	0.75	0.59	0.43	0.25	0.12	**1.37**	0.14	0.08	0.15	0.17	0.14	0.11	0.12	0.10	0.05	**8.11**

sua vez, são boas sementes apenas em redes de GN.

Esses resultados mostram que é possível se obter ganho significativo de desempenho no aprendizado semissupervisionado semeando o processo de rotulação com itens de dados representativos. Mais robusto que LGC e GFHF, mesmo PCC em redes LFR consegue tirar vantagem de algum dos critérios propostos. A intermedialidade se destaca quando existe alguma sobreposição entre as comunidades. Considerando apenas LGC e GFHF, os ganhos de desempenho são bem mais significativos. Em cada modelo de rede, pelo menos três critérios se destacam. Entretanto, a exceção de H_{deg}, critérios diversos se destacam em tipos diversos de redes. E, por isso, propomos a seguir a agregação das medidas no sentido de propor um critério suficientemente geral.

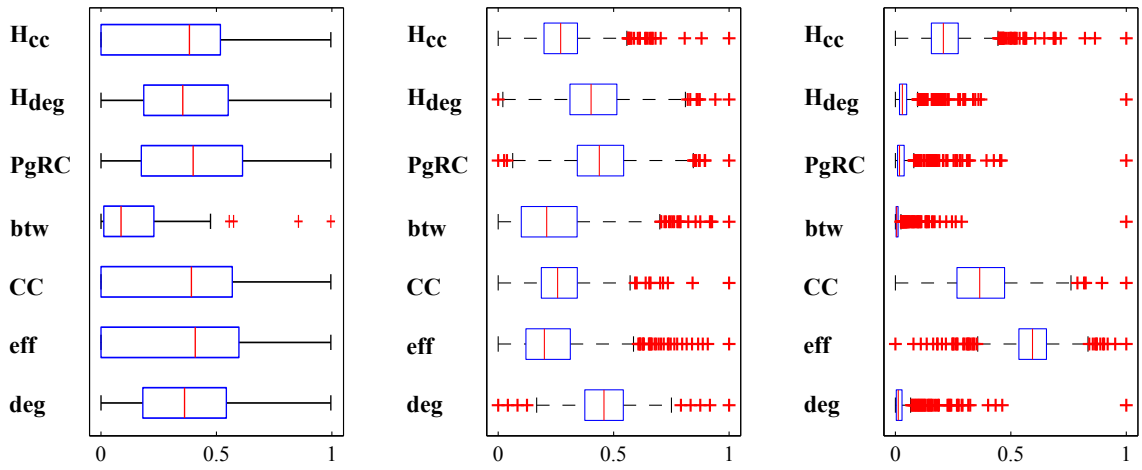
6.5 Composição de medidas

A Figura 6.4 mostra a distribuição cumulativa das medidas estudadas em instâncias de redes GN e LFR, além da rede de golfinhos. Podemos distinguir que alguns nós possuem perfil diferenciado em relação aos demais nós da rede. No *boxplot*, quanto menor a distância entre $Q1$ e $Q3$, mais *outliers* são destacados. Especificamente, nos interessam *outliers* à direita da distribuição, como destaca a medida btw na rede de golfinhos, que apresentam perfis destacado em termos de centralidade. Em redes GN, as medidas CC, H_{CC} e eff destacam melhor esse tipo de nó; em redes LFR, por sua vez, as medidas btw, deg, H_{deg} e PgRC. Essas medidas apresentam distribuições cumulativas significativamente heterogêneas nas respectivas redes.

A técnica análise de componentes principais é frequentemente utilizada na literatura para a redução de dimensionalidade. Dado um conjunto de dados, calcula-se a covariância entre as suas variáveis. Em seguida, decompõe-se a matriz de covariância em autovetores e autovalores associados. Os autovetores são utilizados como base para a transformação das variáveis originais em componentes ortogonais, o que equivale, geometricamente, à rotação das bases do conjunto de dados no sentido de maximizar a variância dos componentes e minimizar a covariância entre os componentes. Correspondentes às variâncias das componentes, os autovalores associados são utilizados como critério de ordenação das variáveis. Com fins à redução de dimensionalidade, componentes com menor variância são descartadas mantendo-se as ‘‘componentes principais’’, ditas de maior ‘‘relevância’’ ou informação agregada (Jolliffe, 2002).

Dados o índices de caracterização de representativos de acordo com diversos critérios, aplicamos a técnica análise de componentes principais para a obtenção de caracterizações ortogonais. As variáveis originais são reescaladas em [0, 1] para prevenir a dominância de variáveis de acordo com suas escalas. Estamos interessados em caracterizações que relevem nós com perfil diferenciado em termos de centralidade.

Como mostra a Figura 6.5, em redes de Girvan-Newman, a primeira componente principal (Z_1) é altamente correlacionada com a intermedialidade e com as medidas baseadas no grau (deg, PgRC e H_{deg}); Z_2, com a eficiência local (eff) e com as medidas de coeficientes de agrupamento (CC e H_{CC}). Em redes LFR, por outro, acontece o oposto. Z_1 é altamente correlacionada com eff, CC e H_{CC}, enquanto que Z_2 é altamente correlacionada com deg, PgRC e H_{deg}. Em ambos



(a) Golfinhos (b) GN (c) LFR

Figura 6.4: Caracterização de nós em redes de acordo com medidas de centralidade. Em (a) temos a rede de associações entre golfinhos (Lusseau et al., 2003); em (b) uma rede de Girvan-Newman com 1024 nós, 4 comunidades, grau médio $\langle k \rangle = 16$ e $Z_{out} = 2$; em (c) um rede LFR com distribuição de grau heterogênea ($\gamma = 2$), grau médio $\langle k \rangle = 16$ e comunidades de tamanhos diversos, porém não muito distintos. Em cada *boxplot*, as extremidades do retângulo informam os 25º e 75º percentis, Q_1 e Q_3; a mediana é apresentada dentro do retângulo; as suíças são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$; e valores além disso são representados como *outliers* (símbolo '+'). Pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. Os índices são apresentados reescalados no intervalo $[0, 1]$.

os casos, a distribuição de Z_2 é a mais heterogênea dentre as distribuições cumulativas das componentes principais, composta pelas medidas que ofereceram os índices mais heterogêneos de caracterização de representativos, que também são aqueles que obtiveram os melhores desempenhos dentre os resultados apresentados acima. Portanto acreditamos que Z_2 possui propriedade inerentes bastante promissoras, e será utilizada como critério para a caracterização de representativos. Para as simulações foram consideradas as mesmas configurações utilizadas nas simulações com critérios simples.

Embora a Tabela 6.2 e as Figuras 6.6 e 6.7 apresentem apenas os resultados com Z_2 em adição aos resultados com critérios simples de caracterização de representativos, também foram estudados os critérios baseados nas demais componentes principais (Z_2-Z_7), bem como critérios com composições das três primeiras componentes principais, tanto em porções iguais (Z_1+Z_2 e Z_1+Z_2+Z_3), como em porções proporcionais às suas variâncias. Contudo, apenas nós destacados pelo critério Z_2 proporcionou ganhos de desempenho na propagação de rótulos.

Estritamente correlacionado com os critérios simples que apresentaram bom desempenho, em redes GN, Z_2 mostra-se capaz de extrair as características mais discriminativas das medidas de centralidade permitindo a detecção de nós representativos nas redes. Como as distribuições de eff, CC e H_{CC}, a distribuição cumulativa de Z_2 é a mais heterogênea dentre as distribuições de componentes principais. Em redes LFR com comunidades bem definidas, nem Z_2, nem nenhum dos demais critérios baseados em componentes principais, possibilita que métodos de

Tabela 6.2: ARI médio e ranque médio obtidos por métodos de propagação de rótulos semeados por nós realçados pelos melhores critérios de seleção de representativos.

	GN N=1024, C=4, $\langle k \rangle$=16, Z_{out}								Ranque	LFR N=1024, $\langle k \rangle$=16, τ_1=2, τ_2=1, μ									Ranque
	1	2	3	4	5	6	7	8	Médio	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	Médio
GFHF sobre																			
ensemble	*0.90*	*0.68*	*0.50*	*0.36*	*0.24*	*0.14*	*0.08*	*0.05*	***3.87***	*0.67*	*0.57*	*0.44*	*0.35*	*0.29*	*0.23*	*0.15*	*0.11*	*0.06*	***4.67***
top deg	-	-	-	-	-	-	-	-	-	0.63	0.54	0.55	0.50	0.33	0.26	0.27	0.20	0.15	**3.33**
top CC	0.90	0.69	0.54	0.45	0.31	0.24	0.09	0.06	**2.12**	-	-	-	-	-	-	-	-	-	-
top btw	-	-	-	-	-	-	-	-	-	0.77	0.66	0.60	0.55	0.38	0.30	0.28	0.23	0.15	**1.11**
top PgRC	-	-	-	-	-	-	-	-	-	0.67	0.58	0.57	0.51	0.35	0.28	0.27	0.20	0.16	**2.44**
top H_{deg}	0.95	0.75	0.55	0.36	0.26	0.16	0.12	0.03	**2.75**	0.74	0.61	0.60	0.53	0.36	0.28	0.27	0.21	0.16	**1.78**
top H_{CC}	0.92	0.82	0.50	0.45	0.27	0.23	0.16	0.08	**1.75**	-	-	-	-	-	-	-	-	-	-
top Z_2	0.89	0.58	0.55	0.42	0.34	0.17	0.12	0.05	**2.75**	0.58	0.47	0.50	0.45	0.30	0.25	0.26	0.19	0.15	**4.22**
LGC sobre																			
ensemble	*0.94*	*0.85*	*0.72*	*0.57*	*0.40*	*0.25*	*0.15*	*0.08*	***3.75***	*0.70*	*0.63*	*0.46*	*0.36*	*0.29*	*0.23*	*0.15*	*0.11*	*0.06*	***5.78***
top deg	-	-	-	-	-	-	-	-	-	0.66	0.61	0.62	0.58	0.43	0.38	0.32	0.28	0.18	**3.67**
top eff	0.96	0.84	0.73	0.62	0.42	0.28	0.18	0.07	**3**	-	-	-	-	-	-	-	-	-	-
top CC	0.96	0.87	0.79	0.63	0.46	0.30	0.16	0.09	**1.5**	-	-	-	-	-	-	-	-	-	-
top btw	-	-	-	-	-	-	-	-	-	0.83	0.76	0.72	0.68	0.52	0.46	0.38	0.32	0.19	**1**
top PgRC	-	-	-	-	-	-	-	-	-	0.73	0.64	0.66	0.60	0.47	0.40	0.33	0.28	0.18	**2.67**
top H_{deg}	0.94	0.86	0.71	0.56	0.43	0.28	0.18	0.08	**3.25**	0.80	0.69	0.70	0.62	0.47	0.42	0.33	0.29	0.19	**1.89**
top H_{CC}	0.97	0.92	0.76	0.60	0.42	0.29	0.18	0.09	**2**	-	-	-	-	-	-	-	-	-	-
top Z_2	0.97	0.86	0.77	0.63	0.48	0.28	0.18	0.09	**1.62**	0.61	0.53	0.57	0.52	0.38	0.35	0.31	0.27	0.17	**4.67**
PCC sobre																			
ensemble	*0.97*	*0.93*	*0.85*	*0.73*	*0.56*	*0.36*	*0.21*	*0.10*	***3.87***	*0.88*	*0.85*	*0.78*	*0.68*	*0.57*	*0.45*	*0.31*	*0.22*	*0.13*	***2.22***
top deg	-	-	-	-	-	-	-	-	-	0.68	0.64	0.63	0.58	0.49	0.40	0.32	0.25	0.15	**4.33**
top eff	0.97	0.90	0.84	0.77	0.57	0.40	0.25	0.12	**3.12**	-	-	-	-	-	-	-	-	-	-
top CC	0.97	0.92	0.89	0.77	0.62	0.43	0.23	0.12	**2**	-	-	-	-	-	-	-	-	-	-
top btw	-	-	-	-	-	-	-	-	-	0.84	0.78	0.72	0.69	0.56	0.48	0.38	0.28	0.18	**1.44**
top PgRC	-	-	-	-	-	-	-	-	-	0.74	0.66	0.66	0.61	0.52	0.43	0.34	0.26	0.15	**3.33**
top H_{deg}	0.98	0.95	0.84	0.73	0.57	0.37	0.22	0.11	**3.25**	0.81	0.71	0.71	0.63	0.51	0.44	0.34	0.27	0.18	**2.67**
top H_{CC}	0.99	0.95	0.88	0.75	0.59	0.43	0.25	0.12	**1.75**	-	-	-	-	-	-	-	-	-	-
top Z_2	0.98	0.92	0.86	0.78	0.63	0.44	0.25	0.12	**1.62**	0.61	0.55	0.58	0.53	0.44	0.38	0.31	0.25	0.15	**5.11**

propagação de rótulos desempenhem melhor que um *ensemble* semeado por amostras aleatórias. No entanto, à medida que é aumentada a mistura entre as comunidades, Z_2 destaca-se superando a abordagem com *ensemble* (vide Fig. 6.7). Essa é uma evidência negativa no sentido de que proposições unificadas baseadas em componentes principais sejam incapazes de caracterizar propriamente representativos em comunidades como as das redes LFR, com poucos acessos de entrada.

6.6 Considerações finais

Nesse capítulo, estudamos a dependência que métodos de rotulação semissupervisionados possuem em relação à amostra de dados disponível rotulada. Foram propostos critérios baseados em índices de centralidade para a seleção de representativos. Em modelos de redes diversos, critérios distintos se destacaram, e portanto buscamos por um critério unificado. O critério agregado proposto baseia-se na técnica análise de componentes principais, dentre os quais se destacou Z_2, critério baseado na segunda componente principal.

Bons resultados foram obtidos para redes Girvan-Newman e redes LFR com algum nível de sobreposição entre as comunidades. No entanto, os resultados em redes LFR com comunidades não foram tão bons quanto os obtidos por critérios simples. Precisamos considerar que a técnica de análise de componentes principais utiliza a covariância entre as variáveis como entrada, de modo que são bem caracterizadas distribuições de Poisson. No entanto, não sabemos muito sobre a caracterização de variáveis com distribuições bastante heterogêneas, como distribuições exponenciais ou distribuições que seguem a lei de potência, de modo que acreditamos ser necessária o estudo de kernels diversos nesse sentido.

Contudo, bons resultados podem ser obtidos com os critérios simples, desde que as redes estudadas sejam consideradas suficientemente semelhantes aos modelos de estudados. Tão logo, esses resultados são estendidos no Capítulo 7 sobre dados vetoriais considerando-se para isso métodos de construção de redes para a abstração gráfica.

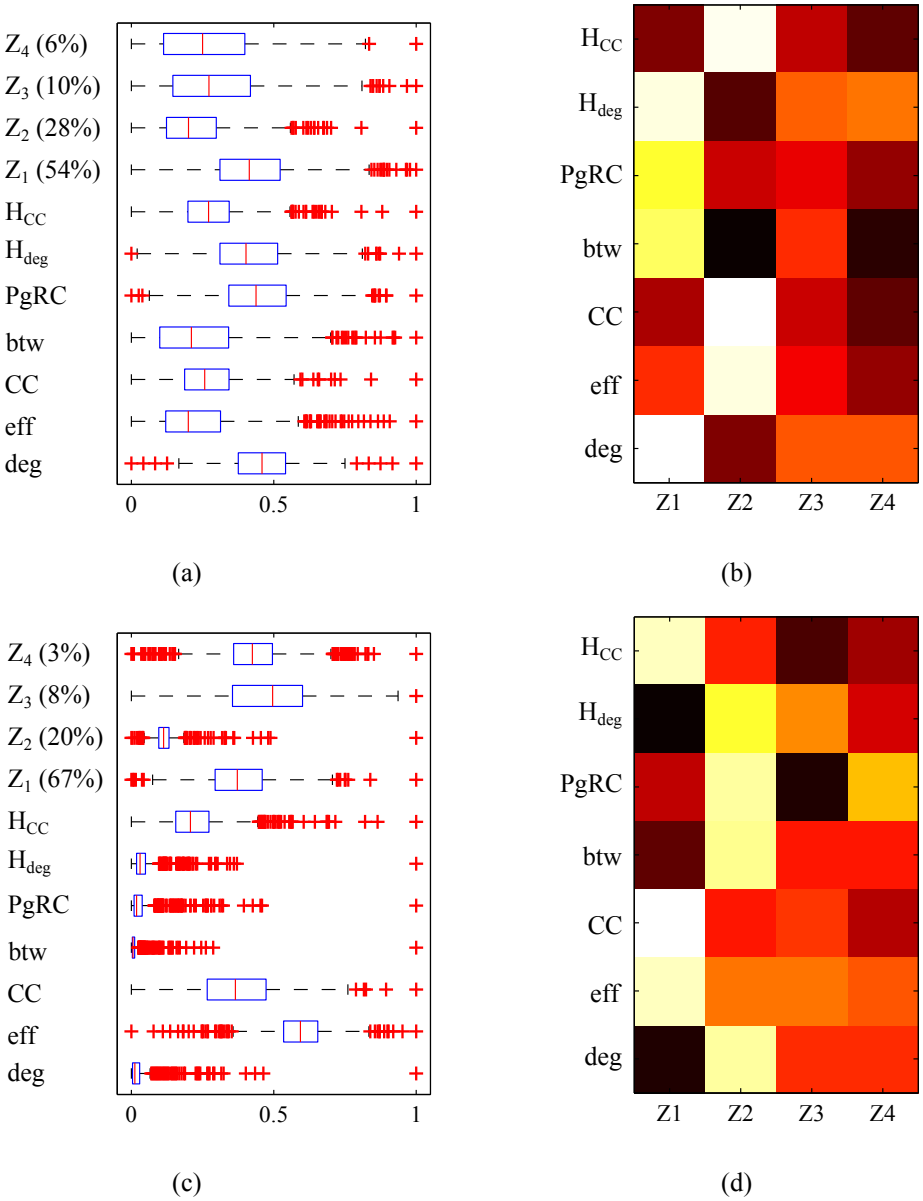


Figura 6.5: Caracterização de nós em redes de acordo com medidas de centralidade simples e agregadas (à esquerda) e correlação das primeiras com as segundas (à direita) para redes Girvan-Newman (acima) e LFR (abaixo). A rede de Girvan-Newman possui 1024 nós, 4 comunidades, grau médio $\langle k \rangle = 16$ e $Z_{\text{out}} = 2$; a LFR, distribuição de grau heterogênea ($\gamma = 2$), grau médio $\langle k \rangle = 16$ e comunidades de tamanhos diversos, porém não muito distintos. Em cada *boxplot*, as extremidades do retângulo informam os 25º e 75º percentis, Q_1 e Q_3; a mediana é apresentada dentro do retângulo; as suíças são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$; e valores além disso são representados como *outliers* (símbolo '+'). Pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. Os índices são apresentados reescalados no intervalo $[0, 1]$. O mapa de cores da matriz de correlação varia em $[-1, 1]$. Quanto maior a correlação, mais claro.

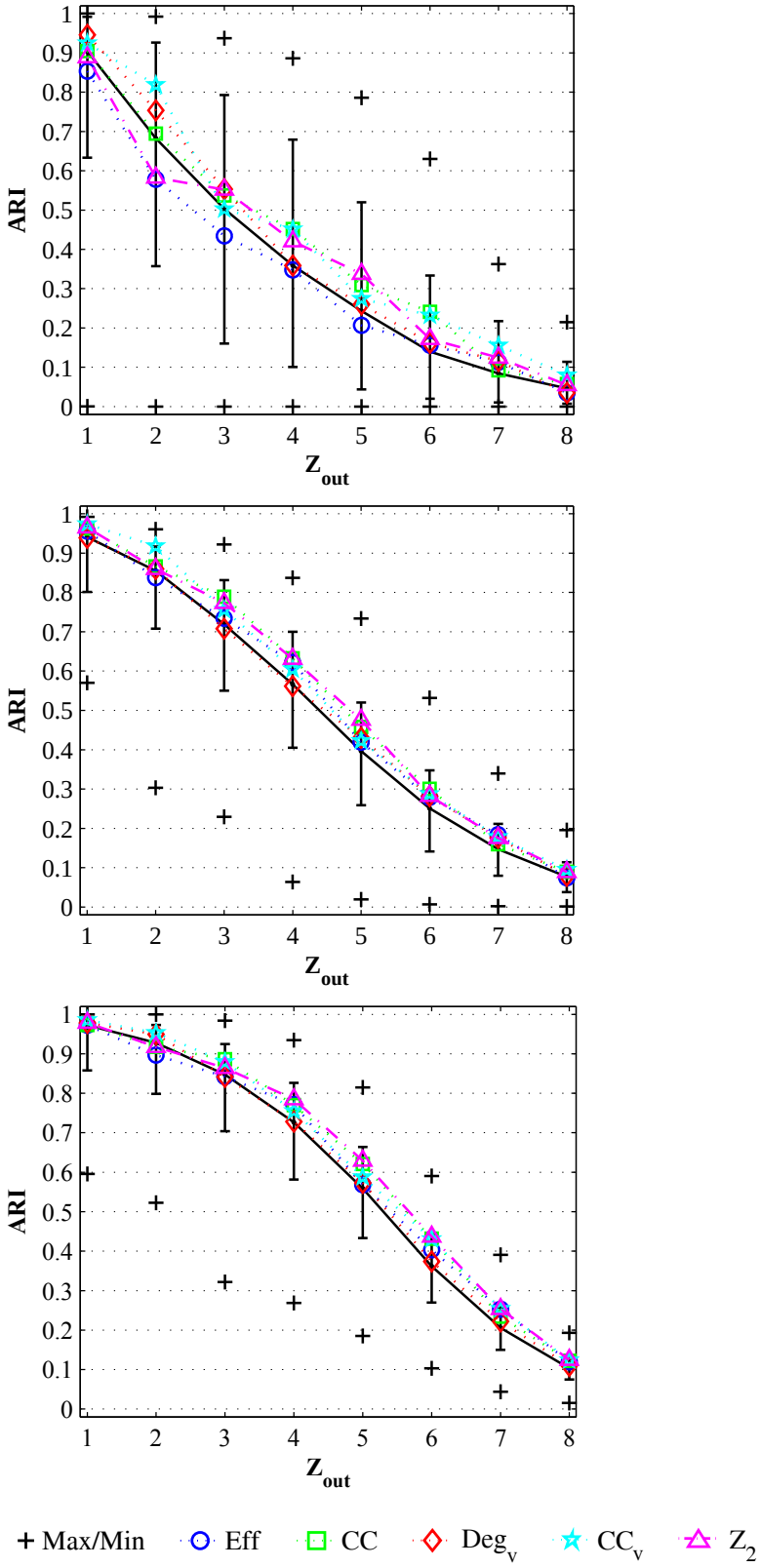


Figura 6.6: Desempenho dos métodos GFHF (em (a)), LGC (em (b)) e PCC (em (c)) sobre redes GN. São apresentados ARI médio, desvio padrão sobre a (e abaixo da) média e pior e melhor desempenho (símbolo '+') observados em 900 ensaios com sementes escolhidas ao acaso, e ARI médio para critérios propostos de seleção de representativos.

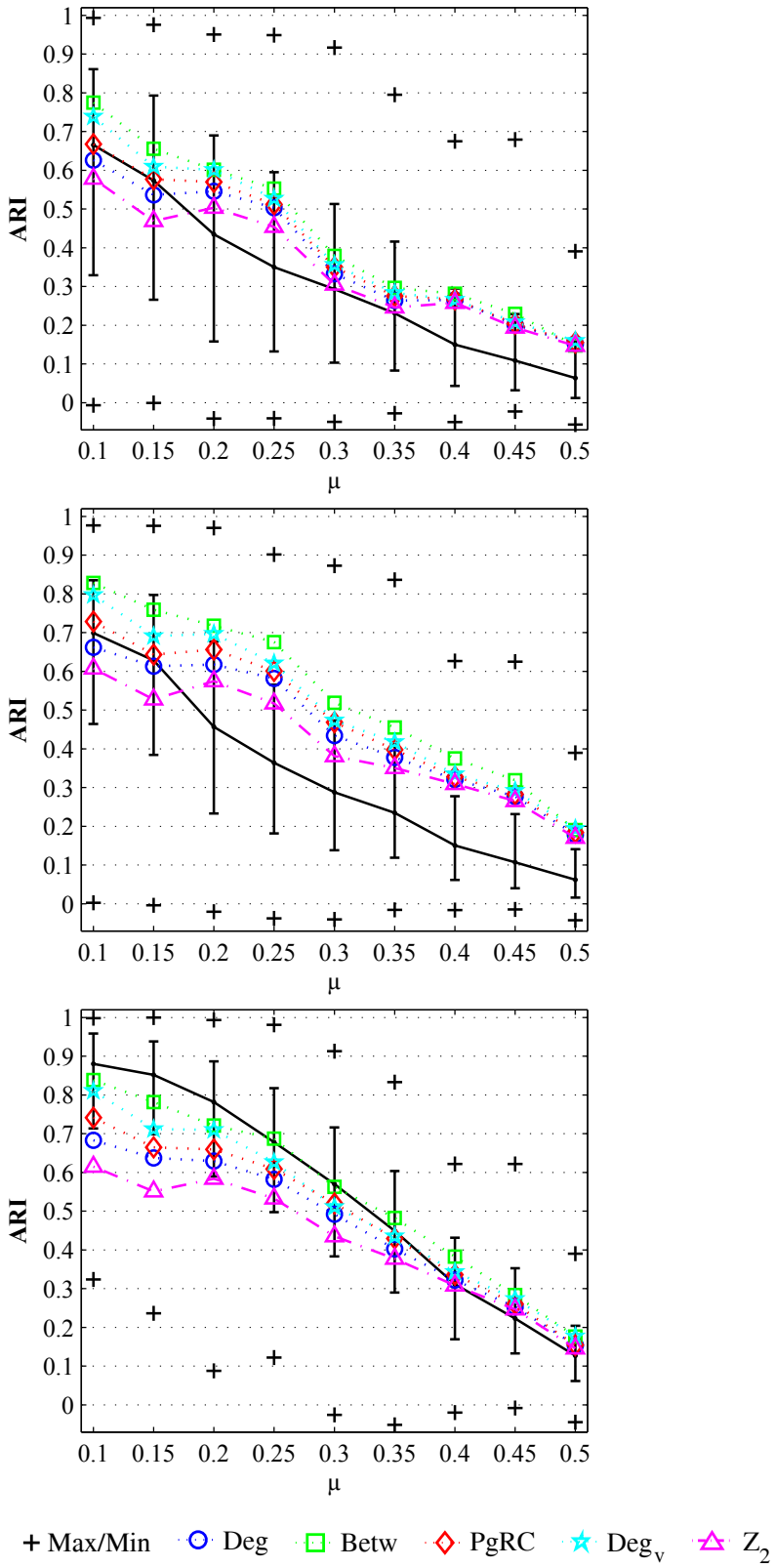


Figura 6.7: Desempenho dos métodos GFHF (em (a)), LGC (em (b)) e PCC (em (c)) sobre redes LFR. São apresentados ARI médio, desvio padrão sobre a (e abaixo da) média e pior e melhor desempenho (símbolo '+') observados em 900 ensaios com sementes escolhidas ao acaso, e ARI médio para critérios propostos de seleção de representativos.

Capítulo

7

Construção de grafos adaptativos

Nesse capítulo estudamos a construção do grafo para o aprendizado semissupervisionado baseado em redes. Como discute o Capítulo 5, bons métodos de construção de grafos podem ser acoplados a quaisquer métodos de rotulação. Crucial para o aprendizado baseado em redes, essa etapa equivale à redução não linear de dimensionalidade e ao aprendizado de suavidade (Roweis & Saul, 2000; Belkin & Niyogi, 2003; Zhu, 2005; Belkin et al., 2006). No entanto, muitas vezes essa etapa tem sido negligenciada. Mesmo hoje, não existe consenso com relação a quais abordagem são mais vantajosas ou em qual cenário alguma se destaca (Rohban & Rabiee, 2012; de Sousa et al., 2013; Jebara et al., 2009; Carreira-perpiñán & Zemel, 2004; Zhu, 2008). Métodos do estado-da-arte, por exemplo, frequentemente apresentam algum inconveniente. Ou se apoiam em suposições fortes, como regularidade ou esparsidade (Jebara et al., 2009; Carreira-perpiñán & Zemel, 2004), ou procuram modelar precisamente a distribuição dos dados ao custo da definição de diversos parâmetros (Cupertino et al., 2013).

Ao construir uma rede a partir de uma base de dados vetorial, buscamos por uma abstração capaz de sintetizar eficazmente as relações de afinidade entre suas instâncias. Logo, nesse capítulo propomos uma nova abordagem para a construção de redes baseada nas árvores geradoras mínimas e no método ϵRadius. Esse método apresenta diversas vantagens, tais como adaptação a dados com distribuições de densidade variadas, baixa dependência da configuração de parâmetros, e custo computacional competitivo tanto em dados *pool based* como em dados incrementais. Além disso, os grafos resultantes são esparsos porém conectados e permitem que métodos de rotulação semissupervisionada tirem grande vantagens da rotulação de nós representativos realçados pela sua estrutura. Grafos construídos sobre conjuntos de dados com grupos bem definidos assemelham-se a redes do modelo Lancichinetti-Fortunato-Radicchi, estudadas no Capítulo 6. Para além disso, o estudo relativo à proposição de uma abordagem unificada de seleção de representativos é estendido.

7.1 Método proposto

O método de construção de redes proposto é formulado sobre a árvore geradora mínima da matriz de distâncias entre pares de instâncias e o método ϵRadius. A árvore geradora mínima (AGM) funciona como o esqueleto da rede resultante. Sobre ela são estimados raios de cobertura, ϵ_i, para cada nó da rede. A descrição detalhada do método, doravante chamado AdaRadius, é apresentada a seguir.

Dado um conjunto de dados vetorial, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \subset \mathbb{R}^d$,

1. Computamos a matriz de distâncias entre todos os pares de vértices,

$$\mathbf{\Lambda}_{ji} = \mathbf{\Lambda}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|, \tag{7.1}$$

   onde $\| \cdot \|$ é uma função de distância tal como a norma l_2 (distância Euclidiana) ou a distância cosseno;

2. Seja $\hat{G} = (V, \hat{E})$ um grafo completo onde cada vértice, $v_i \in V$, representa um item de dados $\mathbf{x}_i \in \mathbf{X}$ e cujos pesos das arestas em $\hat{E}$ são dados por $\mathbf{\Lambda}$, computamos a árvore geradora mínima de $\hat{G}$, $\check{T} = (V, \check{E})$, onde $\check{E} \subseteq \hat{E}$ é um subconjunto acíclico de arestas que conecta todos os nós em V,

$$\text{sujeito a} \qquad \min_{\check{\mathbf{\Lambda}}_{i,j}} \sum_{(i,j)} \check{\mathbf{\Lambda}}_{ij},$$

   onde $\check{\mathbf{\Lambda}}$ é a matriz de pesos de $\check{E}$.

3. Dado $\check{T}$, para cada nó $v_i \in V$, nós calculamos um raio de cobertura, ϵ_i,

$$\epsilon_i = \alpha \max_j \check{\mathbf{\Lambda}}_{ij}; \tag{7.2}$$

4. Dado $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$, estabelecemos as demais arestas,

$$w_{ij} = w_{ji} = \max(p_{ij}, p_{ji}), \tag{7.3}$$

   onde

$$p_{ij} = \begin{cases} 1 & \text{if } \|x_i - x_j\| <= \epsilon_i \text{ e } i \neq j \\ 0 & \text{if } \|x_i - x_j\| > \epsilon_i \text{ ou } i = j \end{cases} . \tag{7.4}$$

7.1.1 Custo computacional

O custo computacional base para métodos de construção de grafos é $O(n^2)$, pois uma tabela de distâncias entre todos os pares de vértices precisa ser computada. Além disso, AdaRadius

é limitado inferiormente pelo custo da descoberta da AGM, que para grafos densos (no caso, completo) é $O(n^2)$, isto é, $O(m+n\log n)$ considerando o algoritmo de Dijkstra-Prim-Jarník com heaps de Fibonacci, que, para $m = O(n^2)$, totaliza $O(n^2)$ (Fredman & Tarjan, 1987). Soluções com soft heaps, $O(m\ \alpha(m,n))$ - onde α é a função de Ackermann que cresce de forma bem lenta e para fins práticos pode ser considerada uma constante não maior que 4 - e soluções de custo ótimo - cuja complexidade computacional é desconhecida - também são conhecidas, mas são pouco utilizadas na prática uma vez que suas implementações, embora difíceis, não oferecem ganhos de desempenho muito significativos (Chazelle, 2000a,b; Karger et al., 1995). Soluções com computação paralela que de custo linear também são conhecidas mas requerem computação distribuída escalável (Chong et al., 2000; Pettie & Ramachandran, 2002a,b). O cálculo de ϵ_i através da Eq. 7.2, por sua vez, é quase linear pois a AGM é extremamente esparsa, $\sum_i k_i = 2m$, $m = n - 1$, de modo que apenas umas poucas arestas são avaliada para cada nó. Finalmente, a computação da Eq. 7.3 é $O(n^2)$ pois a tabela de distância entre pares de instâncias precisa ser reavaliada. Logo, definimos o custo computacional de AdaRadius $O(n^2)$.

A maioria dos métodos apresentados na Sec. 2.4.6 precisam ter seus parâmetros de densidade (σ, k, ϵ, b, d_{thr}, γ, etc) definidos arbitrariamente ou estimados por busca e otimização, requerendo ou o envolvimento de especialistas no domínio dos dados e/ou a realização de diversos ensaios. Além disso, ao lidar com novas instâncias de dados, em geral esses métodos precisam ser reexecutados. AdaRadius, por outro lado, estima a densidade dos dados diretamente através da AGM da tabela de distâncias. E a AGM é conhecidamente ajustável à inclusão de novos nós em $O(n)$ e à remoção de nós em tempo constante, mantido no último caso, uma tabela de substituição que pode ser construída em $O(n^2)$ (Chin & Houck, 1978). Soluções paralelas polilogarítmicas também são conhecidas (Johnson & Metaxas, 1992, 1996; Pawagi & Ramakrishnan, 1986; Varman & Doshi, 1988). A estimação de ϵ_i para cada nó que teve pelo menos uma de suas arestas atualizada pode ser realizada em tempo linear. A habilidade de AdaRadius para lidar com fluxos contínuos de dados será estudada no futuro.

7.2 Benchmarks, métodos e análise numérica

Simulações foram realizadas em três dimensões: 1) no sentido de afirmar a validade de AdaRadius; 2) no sentido de caracterizar os ganhos com representativos; e 3) no sentido de oferecer análises comparativas com outro métodos de aprendizado semissupervisionado (ASS).

Duas fontes de dados de referência foram consideradas: os modelos disponíveis na ferramenta MATLAB PRTools (Duin et al., 2007); e dados utilizados no *benchmark* de (Chapelle et al., 2006) (vide Seção 2.5).

Conjuntos de dados em forma de duas bananas são produzidos pelo MATLAB PRTools configurando-se o tamanho dos conjuntos de dados, n, e a dispersão das distribuições geradas, S. No corrente estudo, S varia em $[0.6 : 1.4]$. 30 ensaios são realizados para cada configuração sendo, assim, permitidas pequenas variações no tamanho das variedades geradas. Chapelle et al. (2006), por outro lado, consideram 8 conjuntos de dados, 6 dos quais estudados aqui: g241c,

g241d, Digit1, USPS, COIL e BCI.

Uma abordagem no sentido de avaliar a qualidade de um método de construção de redes é considerar a medida de modularidade proposta por Newman & Girvan (2004) para confrontar grafos resultantes com dados de referência, cujos rótulos são conhecidos *a priori*. Sendo bastante claro, nesse caso $\mathbf{y}$, $y_i = \{1, \ldots, C\}$, não é predito pelo método de aprendizagem, mas *groung truth* conhecido antecipadamente. Como discutido na Seção 4.3.4, essa medida se mostra adequada para a validação interna e relativa de detecção de comunidades, isto é, para quantificar quão boa é uma partição de uma rede ou até contrastar quantitativamente diversas partições da rede. Entretanto, considerando redes diversas, esta pode não ser muito apropriada, tendo em vista o viés da medida. É o caso da avaliação de métodos de construção de rede.

Tomemos novamente a equação da modularidade,

$$Q = \frac{1}{2m} \sum_{ij} \left(\mathbf{A}_{ij} - \frac{k_i k_j}{2m} \right) \delta(y_i, y_j), \tag{7.5}$$

onde $\delta(y_i, y_j)$ é 1 se $y_i = y_j$ e 0 caso contrário e $2m = \sum_i k_i$. O primeiro termo da equação à direita,

$$\sum_{ij} \frac{\mathbf{A}_{ij}}{2m} \delta(y_i, y_j),$$

computa a fração de arestas entre vértices com o mesmo rótulo. O segundo termo,

$$\frac{k_i k_j}{4m^2} \delta(y_i, y_j),$$

por sua vez, computa o valor da mesma quantidade, porém levando em consideração que a rede tenha emergido aleatoriamente, sendo considerado portanto a chance de cada aresta acontecer ao acaso, sabido apenas que os nós terão grau k_i e k_j, respectivamente, para todo i, j (Clauset et al., 2004; Newman, 2006b; Good et al., 2010). Logo, a estrutura topológica da rede não é avaliada, nem mesmo ligações preferenciais. Como são pontuadas as arestas e não os nós, nós com alto grau são superestimados em detrimento de nós com baixo grau. Consideremos o seguinte exemplo. Sejam i e j dois vértices conectados e com mesmo rótulo,$\mathbf{A}_{ij} = 1$, $\delta(y_i, y_j) = 1$, porém com graus distintos $k_i = 100$ e $k_j = 1$, $1 - \frac{50}{m}$ é apenas uma das 100 possíveis contribuições de i no somatório, porém toda a contribuição de j no somatório. Portanto a avaliação da medida de modularidade, em redes cuja dispersão de grau é ampla, é tendenciosa. Além disso, o índice de modularidade obtido é determinado pelo número de módulos da rede. De simples análise, consideremos uma rede com C componentes totalmente conectados e nenhuma ligação entre eles. Nesse caso, $Q = 1 - 1/C$. Portanto, quanto maior C, $Q \to 1$.

Esse viés não é adequado à presente avaliação porque: 1) redes resultantes podem não possuir distribuição de grau uniforme; e, 2) redes bastante granuladas, cujos módulos, embora pequenos, sejam bem definidos, são superestimadas em detrimento de redes conectadas. Tão logo, a modularidade será utilizada apenas no estudo da validade do método proposto garantindo-se para os demais métodos analisados que a rede resultante possua não mais que C componentes e que

o maior componente possua pelo menos n/C nós, onde C é o número de módulos, conhecido $apriori$ para bases de referência. No Capítulo 8 discutimos a proposição de uma medida de validação de métodos de construção de rede para o ASS, baseada em um outro *framework*.

Os resultados de ASS sobre as redes estudadas são produzidos pelos métodos de rotulação semissupervisionados campos Gaussianos e funções harmônicas (GFHF) e consistência local e global (LGC). $\alpha = 0.95$ é definido para o LGC de acordo com nossa experiência com o método.

Ambos os métodos, GFHF e LGC, podem ser definidos em termos de um *framework* de regularização. O rótulo de x_i, $f(x_i)$, pode ser estimado para o método GFHF ao se minimizar a seguinte função de custo,

$$Q = \sum_{i:x_i \in X_l; j:x_j \in X_u} (y_i - f(x_j))^2 + \sum_{i,j=1}^{n} w_{ij}(f(x_i) - f(x_j))^2; \tag{7.6}$$

e para LGC,

$$Q = \sum_{i:x_i \in X_l; j:x_j \in X_u} (y_i - f(x_j))^2 + \mu \sum_{i,j=1}^{n} w_{ij} \left(\frac{f(x_i)}{\sqrt{k_i}} - \frac{f(x_j)}{\sqrt{k_j}} \right)^2, \tag{7.7}$$

onde $\mu = \alpha/(1-\alpha)$, $0 < \alpha < 1$. Para valores grandes de α, a regularização de LGC tem peso muito grande sobre Q, de modo que favorece à correção das instâncias rotuladas *a priori*, o que é bastante interessante quando os dados foram rotulados de forma pouco confiável. No entanto, é necessário que as variedades de fato conformem-se com as classes, caso contrário os resultados podem ser piorados ao invés de melhorados - GFHF certamente funcionará melhor que LGC nesse caso. Além disso, a capacidade de correção do método LGC é realçada por redes kNN, que são também propensas à regularização.

Para os conjuntos de dados de Chapelle et al. (2006), o parâmetro k do método kNN foi definido de acordo com a otimização da modularidade nos estudos da Sec. 7.3. Por simplicidade, os grafos resultantes, em todos os casos, são sempre binários.

Além do erro de teste utilizado no *benchmark* de Chapelle et al. (2006), também utilizamos o índice de Rand ajustado como critério de validação externa sobre dados em forma de bananas. Ambos, o erro de teste e o ARI são apresentados na Seção 2.5.

Como apresentado no Capítulo 6, utilizamos a caracterização de nós representativos para a seleção de sementes para o ASS. Nesse sentido, algumas medidas de centralidade foram utilizadas como critério de caracterização: grau (deg), status de Katz (Katz), ranque de Page (PgRC), proximidade de uma caminhada aleatória (RWC), proximidade harmônica (C_h), coeficiente de agrupamento (CC), intermedialidade (btw), variação hierárquica de grau (H_{deg}), variação hierárquica de CC (H_{CC}), e variação hierárquica da intermedialidade (H_{btw}) (vide Seção 4.2). deg, Katz, PgRC, RWC e H_{deg} são ditas medidas relacionadas ao grau; btw e H_{btw}, medidas relacionadas ao fluxo; e CC e H_{CC}, medidas relacionadas ao agrupamento. Em relação aos demais critério de caracterização de representativos, nós denominamos ainda deg, Katz, PgRC, RWC, C_h, CC e btw, critérios de primeira ordem; e H_{deg}, H_{CC} e H_{btw}, critérios de segunda orderm.

Sobre a suposição de que a informatividade de uma medida de caracterização de representativos está associada à sua variância, combinamos medidas de caracterização através da técnica de análise de componentes principais. No Capítulo 6 essa técnica é aplicada sobre a matriz de covariância das medições reescaladas. Em modelos de redes bem comportados, a segunda principal componente, denominada Z_2, se destacou. Reescaladas as medições em [0, 1], distribuições dispersas em forma de sinos são alargadas em detrimento de distribuições heterogêneas, compondo a primeira componente principal, Z_1. Z_2, por sua vez, é uma composição de distribuições de cauda longa, no caso, aquelas de interesse.

Ao considerarmos a matriz de correlação ao invés da matriz de covariância das medições reescaladas, distribuições em forma de sino não são mais superestimadas. Portanto Z_1 deve ser capaz de agregar bem as medidas, como um critério de terceira ordem.

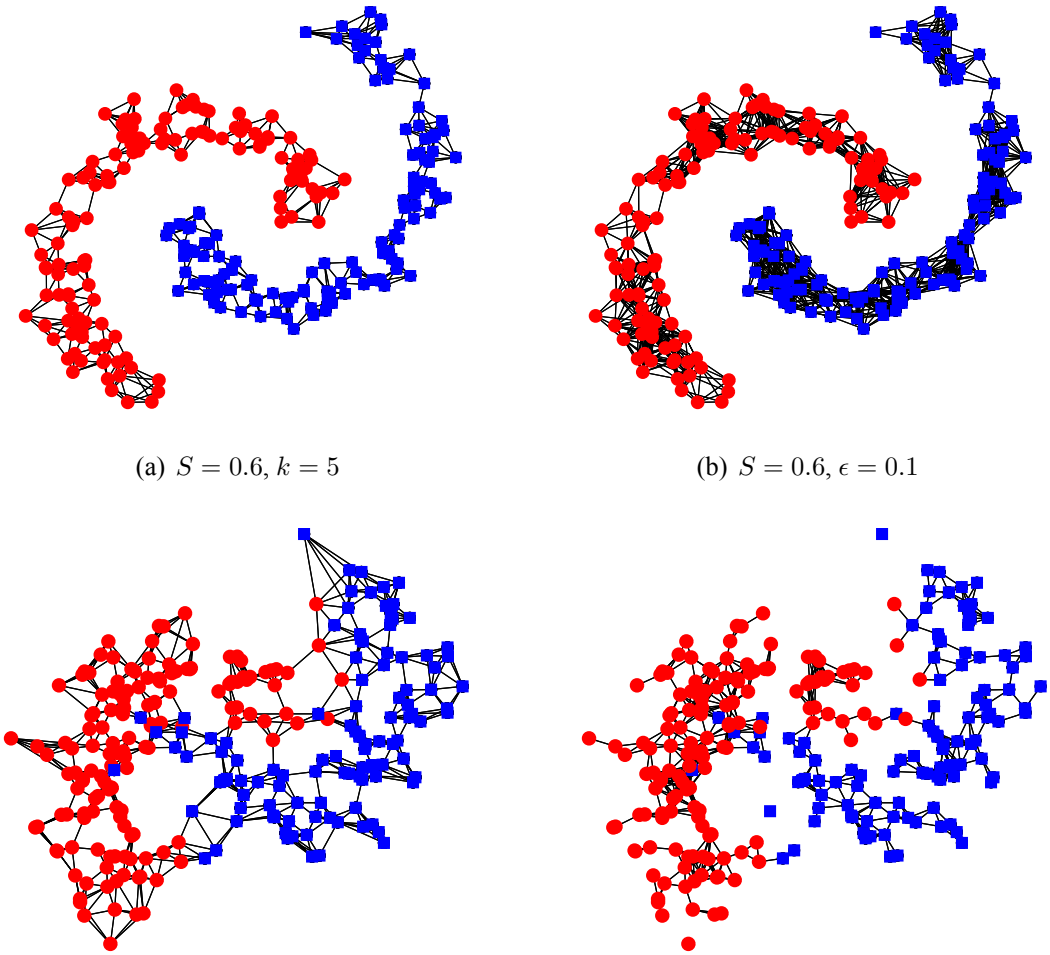
Resultados de ASS são apresentados primeiramente para sementes selecionadas aleatoriamente. 30 amostragens aleatórias são realizados nesse caso, de modo que o ARI médio e desvio padrão reportados dizem respeito ao resultado de 900 ensaios. Para a bases de dados de Chapelle et al. (2006), erro médio de teste e desvio padrão dizem respeito a 90 amostragens aleatórias. Em seguida resultados são apresentados considerando a seleção de representativos. Determinístico, o erro de teste reportado para os conjuntos de dados de Chapelle et al. (2006) é obtido em um único ensaio. Por fim, os melhores resultados com representativos são contrastados com os resultados de referência compilados nas Tabelas 21.9 e 21.11 em (Chapelle et al., 2006). Uma breve descrição dos métodos de ASS utilizados no *benchmark* é apresentada na Tabela 2.2. Descrições detalhadas podem ser encontradas nas referências citadas e em (Chapelle et al., 2006).

7.3 Validade do método proposto

Em geral, métodos de construção de grafos para o ASS apoiam-se fortemente nas suposições de variedade, agrupamento, e suavidades de modo que conjuntos de dados com grupos realmente bem definidos que se conformam com as distribuições de classes tipicamente resultam em redes com comunidades bem definidas. Entretanto, à medida que as distribuições de classes se sobrepõem suas deficiências entram em cena.

Consideremos conjuntos de dados com forma de bananas, $n = 256$ e $S = 0.6$. Os grafos 5NN e 0.1-Radius resultantes possuem dois componentes conexos correspondentes à distribuições de classes, como mostram as Figuras. 7.1 (a) e (b), respectivamente. Ao incrementarmos S, $S = 1.4$, as suposições de variedade, suavidade e agrupamento deixam de ser completamente satisfeitas e, por isso, redes resultantes dos métodos kNN ou ϵRadius não são mais perfeitamente modulares (veja a Figura. 7.1 (c) e (d)). Ajustando k e ϵ no sentido de se maximizar a modularidade, o melhor escore obtido por grafos não muito fragmentados é $Q = 0.4197$ e $Q = 0.4018$, respectivamente, para $k = 5$ e mais ou menos $\epsilon = 0.065$ (veja a Figura 7.2 (a) e (b)). Isto acontece porque os parâmetros de densidade (k e ϵ) são definidos unicamente para todos os itens de dados do conjunto, assumindo que as distribuições dos dados são uniformes e bem comportadas. Melhores escores são obtidos somente ao se validar a construção de redes bastante

fragmentadas para as quais a medida de modularidade é degenerada.



(a) $S = 0.6$, $k = 5$

(b) $S = 0.6$, $\epsilon = 0.1$

(c) $S = 1.4$, $k = 5$

(d) $S = 1.4$, $\epsilon = 0.065$

Figura 7.1: Grafos construídos pelos métodos kNN ((a) e (c)) e ϵRadius ((b) e (d)) sobre conjuntos de dados em forma de bananas com $n = 256$ e $S = \{0.6, 1.4\}$. k e ϵ são ajustados no sentido de maximizar o índice de modularidade do grafo resultante, sem fragmentá-lo muito.

Por outro lado, o método proposto é capaz de construir redes com melhores índices mesmo sem ajustar o parâmetro α, isto é, para $\alpha = 1$. Como mostram as Figuras 7.3 e 7.4, as redes resultantes são tão esparsas quanto aquelas construídas pelos métodos kNN e ϵRadius. Novas arestas são estabelecidas em adição à AGM quando aresta relativamente longas ocorrem na AGM, de modo que a pertinência dos nós extremidades a seus grupos mais próximos é reforçada. Assim, AdaRadius mostra-se robusto à ocorrência de *outliers* pontuais e, embora módulos bem separados sejam conectados, as bordas dos grupos são bem definidas. *Hubs*, por sua vez, podem resultar de localizações privilegiadas na AGM, conectando diversas vias (veja exemplos de *motifs* na Figura. 7.4). Melhores índices de modularidade podem ser obtidos ajustando-se o parâmetro α, no entanto pode não ser conveniente, como no exemplo destacado na Figura 7.2 (c) com $\alpha < 1$, resultando em redes fragmentadas.

Além disso, a Tabela 7.1 contrasta resultados obtidos pelo método proposto com resultados de alguns métodos clássicos sobre conjuntos de dados de Chapelle et al. (2006). Os parâmetros

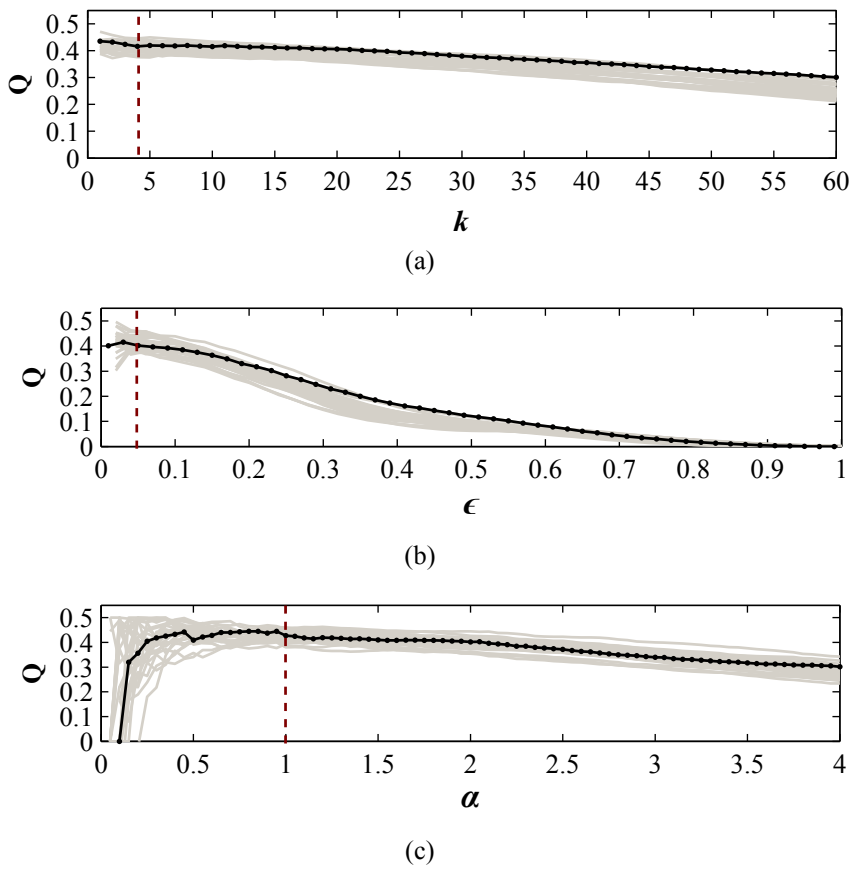


Figura 7.2: Índice de modularidade de grafos kNN (a), ϵRadius (b) e AdaRadius (c) sobre conjuntos de dados em forma de bananas com $n = 256$ e $S = 1.4$ variando-se os parâmetros de densidade. Para o ensaio cuja curva é apresentada em preto, o grafo resultante é desconexo para configurações de parâmetros menores que o valor do limiar discriminado no eixo vertical.

de densidade são ajustados a fim de otimizar a modularidade, garantindo, no entanto, que os grafos resultantes não sejam muito fragmentados, i.e., que tenha não mais componentes de que o número de classes do problema e que o maior componente seja composto de pelo menos n/C nós. α é avaliado no intervalo $[0.5 : 0.01 : 2]$, ϵ no intervalo entre a menor e a maior distância entre pares de nós, porém discretizado em 100 valores, e k no intervalo $[1 : 400]$. Resultados para $\alpha = 1$ também são reportados. As configurações ótimas dos parâmetros são reportadas na Tabela 7.2. Tomando como base o índice de modularidade, redes AdaRadius sempre superam correspondentes produzidas por ϵRadius. Para os conjuntos de dados Digit1, USPS, COIL e BCI, AdaRadius também apresenta resultados competitivos com os demais métodos. Em adição, a medida m/n é reportada como um índice de densidade na Tabela 7.3.

Maiores ganhos em termos da modularidade podem ser obtidos ajustando-se o parâmetro α. Notavelmente, no entanto, o α ótimo não se desvia muito de $\alpha = 1$ em todos conjuntos de dados. Variações de AdaRadius também podem ser estudadas, por exemplo, ajustando α para cada região modular dos dados. Contudo, nós acreditamos que a AGM é capaz de prover boas estimativas de ϵ_i para todos os nós.

Por fim, nós contrastamos o método proposto com o método 5NN em termos de ASS sobre conjuntos de dados em forma de bananas com $n = 512$ e $S = \{0.6, 1, 1.4\}$. Divisões aleatórias

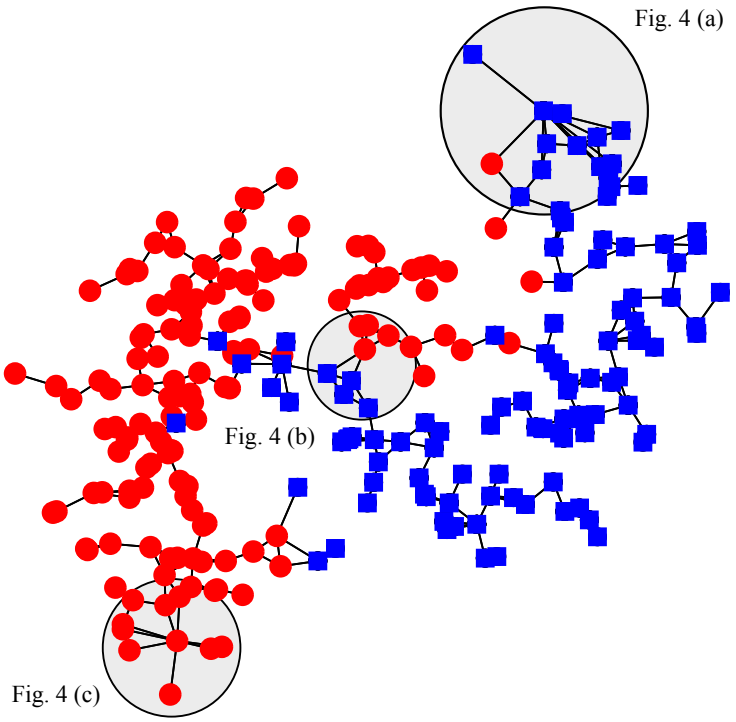


Figura 7.3: Resultado do método AdaRadius sobre conjuntos de dados em forma de bananas com $n = 256$ e $S = 1.4$. Os *motifs* destacados são ampliados e descritos na Figura 7.4.

Tabela 7.1: Modularidade ótima obtida por AdaRadius e métodos clássicos sobre os conjuntos de dados de Chapelle et al. (2006).

	g241c	g241d	Digit1	USPS	COIL	BCI
Opt kNN	0.1158	0.1439	0.4766	0.2639	0.6886	0.0493
Opt mu$k$NN	0.0685	0.1122	0.4752	0.2446	0.649	0.0277
Opt ϵRadius	0.0476	0.0654	0.4456	0.0339	0.1437	0.0024
Opt AdaRadius	0.0611	0.0725	0.4748	0.1713	0.7648	0.0317
AdaRadius	0.0611	0.0725	0.4730	0.1713	0.7648	0.0279

do conjunto de dados entre dados rotulados e não rotulados são consideradas para 1%, 2%, 5% e 10% dos dados rotulados. A Tabela 7.4 mostra o ARI obtido pelos métodos GFHF e LGC sobre as respectivas redes. Os resultados mostram que quando os grupos se conformam com as classes ($S = 0.6$), 5NN funciona melhor que AdaRadius, mas, para $S = \{1, 1.4\}$ AdaRadius sobrepõe-se a 5NN. A única exceção se configura para amostras grandes de dados rotulados (10%) sobre conjuntos de dados com módulos bastante sobrepostos ($S = 1.4$), onde o 5NN funciona um pouco melhor que AdaRadius.

Certamente alguém poderá argumentar que isto acontece porque k não é ajustado para cada conjunto de dados de acordo com a sobreposição entre os módulos. De fato este é o principal desafio enfrentado pela maioria dos métodos de construção de rede amplamente usados, dependência da configuração de parâmetros. AdaRadius supera esse desafio estimando ϵ_i para cada nó, através da AGM da matriz de distâncias. No entanto, mais discussões sobre conjuntos de dados de Chapelle et al. (2006) para os quais valores ótimos de k são conhecidos (Tabela 7.2)

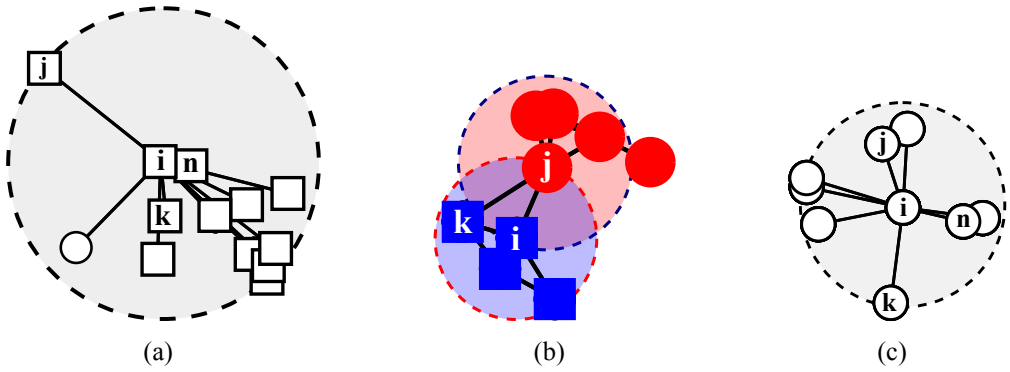


Figura 7.4: *Motifs* construídos por AdaRadius: (a) nó *outlier*, (b) nós de borda, e (c) nó *hub* destacados na Figura 7.3. Conectado a j (um *outlier*), as conexões de i à variedade mais próxima são reforçadas em (a). Situados nas bordas de seus grupos, as conexões de i e j aos seus grupos são reforçadas em (b); Já um *hub* na AGM, devido à sua localização privilegiada, as conexões de i são incrementadas em (c). O valor de ϵ_i é ilustrado através das esferas de cobertura, em cinza.

Tabela 7.2: Configurações de parâmetros para soluções ótimas em termos da modularidade.

	g241c	g241d	Digit1	USPS	COIL	BCI
Opt kNN	3	2	3	3	9	3
Opt mu$k$NN	322	113	9	36	63	19
Opt ϵRadius	21.06	20.52	1.8	6.9	624.24	12.04
Opt AdaRadius	1	1	1.05	1	1	1.08

são apresentados a seguir.

A Tabela 7.5 reporta o erro de teste obtido pelós métodos GFHF e LGC sobre redes construídas por kNN e AdaRadius sobre os conjuntos de dados de Chapelle et al. (2006). Em resumo, para conjuntos de dados cujas classes são facilmente separáveis, kNN supera AdaRadius. No entanto, em problemas mais difíceis AdaRadius funciona tão bem quanto kNN ou melhor. Com 10 itens de dados rotulados, AdaRadius supera kNN nos conjuntos de dados BCI e COIL tanto acoplado ao GFHF como ao LGC e no conjunto USPS acoplado com o LGC apenas. Sobre os conjuntos de dados g241c e g241d, AdaRadius funciona tão bem quanto o kNN. Na verdade, ambos AdaRadius e kNN funcionam não muito bem sobre g241d nessa configuração. Com 100 itens de dados rotulados, AdaRadius supera kNN apenas sobre o conjunto COIL e funciona tão

Tabela 7.3: Escore de densidade (m/n) para soluções ótimas em termos da modularidade.

	g241c	g241d	Digit1	USPS	COIL	BCI
Opt kNN	2.75	1.83	1.98	2.29	5.59	2.18
Opt mu$k$NN	91.12	26.29	3.40	10.60	21.96	6.08
Opt ϵRadius	142.14	72.67	8.36	177.71	95.58	65.71
Opt AdaRadius	13.87	11.15	2.12	4.23	2.21	4.56
AdaRadius	13.87	11.15	1.36	4.23	2.21	2.50

Tabela 7.4: ARI obtido por métodos de rotulação semissupervisionados sobre redes 5NN e AdaRadius construídas sobre dados em forma de bananas

	S=0.6				S=1				S=1.4			
	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%
5NN												
+ GFHF	69.02	86.72	96.77	98.96	33.09	56.83	78.98	85.16	23.08	40.94	60.16	67.86
	±4.14	±2.70	±1.27	±0.44	±4.39	±3.58	±2.07	±1.85	±3.43	±3.10	±2.18	±1.14
+ LGC	65.73	83.03	95.19	98.21	35.83	56.31	77.66	85.80	25.93	41.10	59.95	67.76
	±4.80	±3.91	±1.15	±0.54	±3.83	±4.23	±2.26	±0.92	±3.54	±3.75	±1.86	±1.07
AdaRadius												
+ GFHF	60.01	82.17	94.77	97.79	58.60	73.86	84.20	86.90	36.78	49.42	62.59	66.20
	±6.02	±3.82	±1.43	±0.56	±5.59	±3.91	±1.81	±0.69	±3.94	±3.62	±2.13	±0.94
+ LGC	62.29	82.69	94.43	97.45	56.36	70.87	82.45	86.29	36.42	47.46	60.98	65.99
	±5.53	±4.50	±1.19	±0.61	±5.09	±4.20	±1.80	±0.86	±4.52	±4.20	±2.31	±1.01

Tabela 7.5: Erro de teste obtido por métodos de rotulação semissupervisionados sobre redes 5NN e AdaRadius construídas sobre os conjuntos de dados de Chapelle et al. (2006)

	10 nós rotulados						100 nós rotulados					
	g241c	g241d	Digit1	USPS	COIL	BCI	g241c	g241d	Digit1	USPS	COIL	BCI
kNN												
+ GFHF	49.9	49.76	18.33	18.32	67.46	50.12	45.94	44.81	2.68	6.72	21.28	46.46
	±0.50	±0.58	±13.03	±2.60	±5.72	±1.88	±2.62	±2.27	±0.53	±2.13	±2.63	±2.55
+ LGC	48.95	49.04	12.02	16.92	65.85	50.42	45.29	44.15	3.25	6.38	21.93	48.08
	±1.48	±1.67	±5.27	±5.38	±4.64	±1.72	±2.92	±2.14	±0.78	±1.96	±2.69	±2.80
AdaRadius												
+ GFHF	50.01	50.01	20.23	19.26	64.98	50.09	48.72	48.13	4.71	11.12	17.76	48.15
	±0.23	±0.20	±11.89	±1.62	±6.20	±1.44	±2.32	±2.29	±1.21	±3.70	±3.54	±2.70
+ LGC	49.52	49.44	16.64	15.68	62	50.14	47.87	47.44	4.87	9.04	15.67	48.75
	±1.44	±1.07	±6.93	±5.19	±5.40	±1.53	±3.37	±3.01	±1.26	±3.03	±3.40	±2.77

bem quando kNN sobre o conjunto BCI. Os benefícios de AdaRadius tornam-se mais relevantes quando dados representativos semeiam a rotulação semissupervisionada, o que é estudado na seção a seguir.

7.4 Resultados e discussões sobre representativos

Primeiramente, precisamos esclarecer que representativos podem ser selecionados ao acaso. Dentre 30 amostragens aleatórias para cada conjunto de dados em forma de bananas (30 conjunto de dados de cada configuração são estudados) e dentre 90 amostragens aleatórias sobre conjuntos de dados de Chapelle et al. (2006), algumas amostras certamente são compostas por bons representativos. O desvio padrão do desempenho de ASS sobre divisões aleatórias dos dados oferecem indícios nesse sentido. Para um montante específico de dados rotulados, grandes desvios evidenciam que, embora pouco prováveis, representativos são amostrados ao acaso. Pequenos desvios, por outro lado, evidenciam que ou representativos são difíceis de serem selecionados ao acaso ou que a maioria das amostras aleatórias são representativas. O primeiro caso é recorrente para amostras pequenas e o segundo para amostras suficientemente grandes, por exemplo. Desse modo, determinamos o ARI (erro de teste) médio adicionado (subtraído) pelo desvio padrão como limiar para a enfase (em negrito) de resultados ditos significativos. Resultados próximos mas não superiores (inferiores) a este limiar também serão discutidos, especialmente para grandes desvios padrões de referência.

7.4.1 Resultados sobre conjuntos de dados em forma de bananas

As Tabelas 7.6 e 7.7 mostram o ARI de métodos de rotulação semissupervisionados semeados por representativos em redes kNN. PgRC, H_{deg} e H_{btw} se destacam como critérios de destacamento de representativos quando suficientes dados rotulados são considerados (pelo menos 5% dos itens de dados). PgRC também se destaca significantemente sobre dados bastante sobrepostos ($S = 1.4$). H_{CC}, por outro lado, se destaca quando poucos dados rotulados estão disponíveis, mas não sobre dados muito sobrepostos ($S = 1.4$). Z_1 se destaca apenas para 5% ou mais dados rotulados, mas sempre supera os demais critérios de terceira ordem. Z_1 é principalmente composto de medições de deg, Katz, PgRC e H_{deg} (veja a Figura 7.5). Em geral, resultados obtidos com o LGC são mais estritos que os resultados obtidos como GFHF sobre dados com alta sobreposição ($S = 1, 1.4$) pois a suposição de grupos, sobre a qual a correção do LGC se sustenta, não é satisfeita.

As Tabelas 7.8 e 7.9 mostram o ARI de ASS semeado por representativos para AdaRadius. Além de PgRC, H_{deg} e H_{btw}, RWC e H_{CC} também se destacam na maioria das configurações avaliadas. PgRC somente não se destaca significativamente quando muitos itens de dados estão disponíveis rotulados (5% e 10%) sobre dados bastante sobrepostos. Nesse caso amostras aleatórias funcionam tão bem quanto. Além disso, btw se destaca quando as variedades são bem definidas ($S = 0.6$) e apenas poucos itens de dados rotulados estão disponíveis. Nós com alto btw tipicamente estão situados na borda dos módulos ligando-os a outros módulos. Exceto

Tabela 7.6: ARI obtido por GFHF semeado por representativos sobre redes 5NN construídas sobre dados em forma de bananas.

	S=0.6				S=1				S=1.4			
	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%
amostra rand	69.02 ±4.14	86.72 ±2.70	96.77 ±1.27	98.96 ±0.44	33.09 ±4.39	56.83 ±3.58	78.98 ±2.07	85.16 ±1.85	23.08 ±3.43	40.94 ±3.10	60.16 ±2.18	67.86 ±1.14
deg top	56.26	78.74	94.38	98.67	22.04	40.50	73.55	76.32	17.27	28.30	57.41	65.48
Katz top	56.34	76.60	93.53	98.72	23.07	44.99	70.87	79.37	22.21	31.57	54.85	63.69
PgRC top	68.69	87.72	**98.73**	99.39	33.60	55.59	76.98	83.65	**26.94**	**47.91**	**66.07**	**69.24**
RWC top	45.13	65.72	75.48	90.47	15.73	14.37	25.35	48.88	5.79	6.25	16.39	29.69
C_h top	52.26	68.84	94.28	96.50	13.37	17.34	14.46	27.78	7.83	8.27	8.77	27.60
CC top	42.83	64.84	92.75	98.55	19.04	34.93	70.79	82.56	9.62	23.82	46.23	61.84
btw top	55.69	65.96	81.23	94.77	22.41	25.75	61.32	82.57	8.77	14.02	36.97	65.09
H_{deg} top	71.61	85.11	**98.65**	99.39	31.55	55.60	79.08	83.56	**27.25**	43.86	**64.88**	68.72
H_{CC} top	71.01	87.76	97.27	98.89	**42.64**	**62.35**	77.08	84.08	13.97	38.50	61.05	68.39
H_{btw} top	67.83	86.54	**98.28**	99.28	31.29	**62.60**	80.82	83.64	25.72	39.58	**63.62**	67.85
Z_1 top	64.39	88.85	**98.43**	99.39	35.06	**60.84**	**82.12**	85.24	19.30	38.60	**67.13**	**71.09**
Z_2 top	66.02	73.59	94.14	98.10	15.91	20.32	56.87	81.91	14.19	14.48	38.29	58.92
Z_3 top	42.64	67.52	86.54	93.60	8.94	13.18	24.79	48.59	7.76	6.57	13.33	29.71
Z_4 top	46.76	75.27	88.18	97.46	21.42	34.13	70.84	82.86	9.53	15.93	54.96	68.21
Z_5 top	49.14	78.33	92	95.27	31.02	56.78	69.45	78.74	14.44	30.17	45.09	57.58
Z_6 top	62.58	79.27	94.97	97.61	19.33	43.72	65.86	77.23	22.98	28.92	53.72	63.04
Z_7 top	56.24	63.01	82.37	90.03	30.78	36.11	60.23	72.34	22.11	27.02	44.47	57.83
Z_8 top	56.93	66.16	80.94	85.99	13.91	23.82	51.78	70.87	10.50	25.80	46.80	57.38
Z_9 top	47.49	62.14	94.35	96.09	27.21	33.56	60.96	73.28	16.15	30.88	49.59	60.83
Z_{10} top	36.85	50.71	81.91	88.44	22.43	29.84	52.28	67.20	10.04	17.97	31.15	48.72

Tabela 7.7: ARI obtido por LGC semeado por representativos sobre redes 5NN construídas sobre dados em forma de bananas.

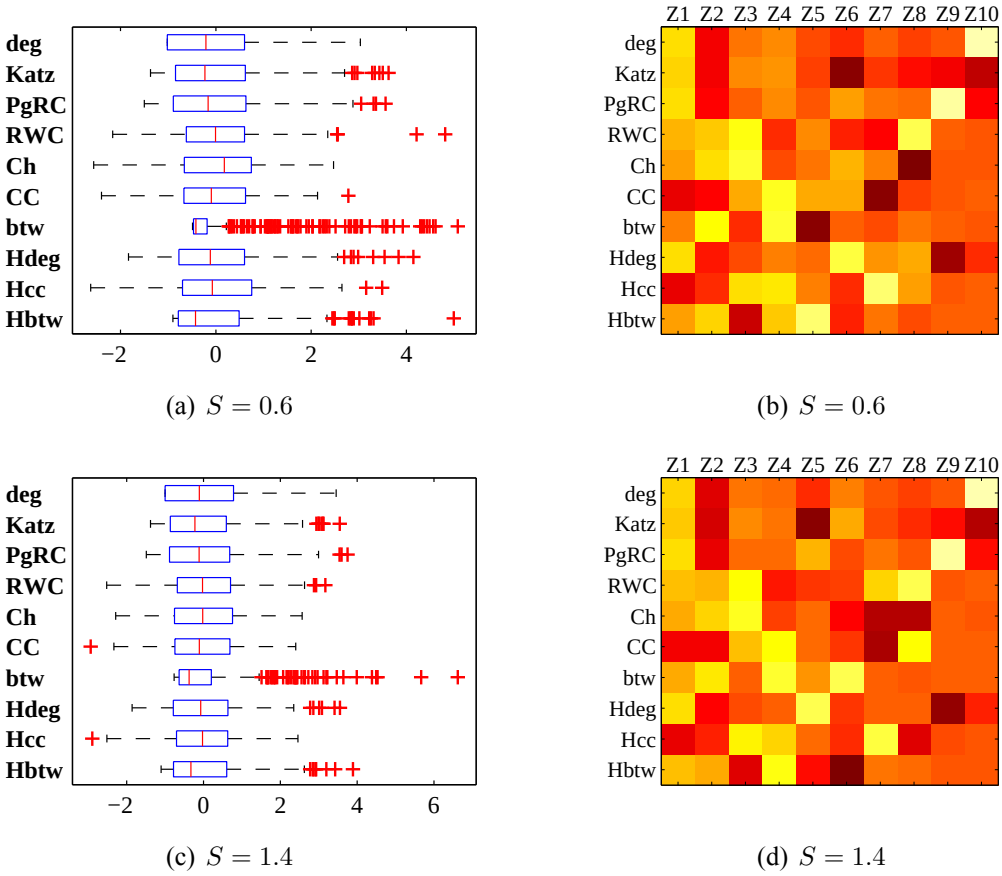
	S=0.6				S=1				S=1.4			
	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%
amostra rand	65.73 ±4.80	83.03 ±3.91	95.19 ±1.15	98.21 ±0.54	35.83 ±3.83	56.31 ±4.23	77.66 ±2.26	85.80 ±0.92	25.93 ±3.54	41.10 ±3.75	59.95 ±1.86	67.76 ±1.07
deg top	50.79	70.03	89.66	95.43	24.33	44.06	76.28	75.43	19.08	27.08	54.69	64.30
Katz top	56.75	68.69	88.73	97.18	25.32	44.57	72.33	80.92	21.55	33.59	54.41	62.13
PgRC top	64.46	79.92	**96.74**	**99.08**	33.61	53.75	78.57	85.71	**29.86**	**48.71**	**64.33**	**69.21**
RWC top	43.37	56.22	66.11	76.56	13.33	11.32	14.22	25.29	7.10	5.25	8.35	14.28
C_h top	48.86	61.19	76.07	93.74	9.05	11.89	12.29	19.82	5.97	4.19	7.24	12.78
CC top	41.35	61.78	87.48	96.13	19.73	34.62	69.17	82.30	11.15	24.17	45.77	59.97
btw top	54.86	62.69	77.91	91.35	18.87	19.96	42.55	72	8.52	11.63	24.78	49.94
H_{deg} top	69.22	80.30	**97.65**	**99.05**	34.56	55.56	**80.82**	86.12	28.07	41.30	**63.63**	**69.39**
H_{CC} top	**70.68**	83.97	95.04	**98.80**	36.90	58.36	79.28	86.49	18.23	39.73	60	67.91
H_{btw} top	64.93	85.68	93.66	**99.28**	35.28	59.72	**82.59**	87.33	28.39	40.05	**63.43**	68.16
Z_1 top	61.96	80.11	**97.15**	**99.05**	34.61	52.27	79.81	**88.07**	19.28	36.24	**63.73**	**71.77**
Z_2 top	64.16	71.04	94.53	97.87	15	16.39	45.71	73.17	11.51	11.56	31.52	50.09
Z_3 top	40.25	61.30	76.93	89.23	6.88	9.90	17.72	33.97	6.85	7.34	9.01	18.23
Z_4 top	46.66	70.21	83.29	94.50	21.11	28.94	60.24	80.61	10.02	15.08	46.27	63.77
Z_5 top	50.98	74.85	89.19	93.36	31.95	58.21	68.19	77.29	15.49	26.64	43.84	56.72
Z_6 top	60.35	75.15	90.29	94.37	21.30	44.11	59.97	76.85	25.71	29.89	51.09	59.77
Z_7 top	57.20	63.61	77.20	85.86	32.62	37.16	58.02	66.76	22.40	29.05	42.33	55.43
Z_8 top	56.46	61.04	75.26	79.75	17.83	24.32	40.38	59.42	14.76	27.96	44.05	52.72
Z_9 top	50.94	61.76	88.09	91.12	26.67	31.66	54.96	68.13	19.69	31.28	47.65	59.21
Z_{10} top	38.09	48.39	80.36	85.19	22.09	31.54	52.81	67.99	10.71	20.77	34.31	45.13

Tabela 7.8: ARI obtido por GFHF semeado por representativos sobre redes AdaRadius construídas sobre dados em forma de bananas.

	S=0.6				S=1				S=1.4			
	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%
amostra rand	60.01 ±6.02	82.17 ±3.82	94.77 ±1.43	97.79 ±0.56	58.60 ±5.59	73.86 ±3.91	84.20 ±1.81	86.90 ±0.69	36.78 ±3.94	49.42 ±3.62	62.59 ±2.13	66.20 ±0.94
deg top	**66.98**	79.16	93.12	95.65	59.16	70.94	81	83.42	38.25	51.61	59.97	61.28
Katz top	65.52	67.58	88.89	95.14	55.31	63.28	75.20	82.43	38.29	45.46	55.05	61.26
PgRC top	**71.82**	**87.27**	**96.68**	**98.91**	**67.87**	76.25	83.84	**87.92**	**45.93**	**54.19**	62.93	66.24
RWC top	51.88	84.62	94.61	**98.52**	50.16	**78.85**	83.59	**88.06**	27.91	46.53	62.83	67
C_h top	63.20	77.65	91.33	**99.01**	39.14	50.85	69	76.93	19.62	25.08	41.30	45.85
CC top	0	0	0	74.96	0	0	0	69.79	0	0	0	37.05
btw top	**89.41**	**96.26**	**99.58**	**99.61**	62.93	74.66	76.89	83.70	25.43	29.22	45.36	54.53
H_{deg} top	**66.47**	**86.33**	**96.72**	**98.89**	62.79	76.36	84.30	**88.42**	**42.72**	52.59	62.84	66.34
H_{CC} top	63.25	83.48	95.15	97.48	57.84	70.98	84.34	87.58	38.52	50.56	60.97	66.50
H_{btw} top	**73.08**	84.91	**96.04**	**98.63**	**67.32**	72.42	85.12	**87.61**	**44.25**	**54.60**	63.13	64.56
Z_1 top	**72.39**	**86.26**	**98.19**	**99.05**	**69.85**	74.11	85.23	**87.98**	**49.38**	**55.51**	61	65.69
Z_2 top	**76.35**	**86.40**	**97.49**	98.11	58.33	71.37	81.26	87.23	36.47	41.97	64.34	63.87
Z_3 top	**79.09**	**87.80**	**99.53**	**99.54**	**69.97**	73.52	81.79	86.10	35.67	36.85	44.41	57.19
Z_4 top	40.54	66.24	90.20	94	50.73	62.74	82.16	86.31	27.47	41.24	54.55	63.43
Z_5 top	58.91	**90.57**	**98.56**	**99.38**	47.85	70.41	83.36	87.11	29.51	46.55	59.98	66.18
Z_6 top	64.47	74.13	89.11	92.47	58.93	69.82	78.53	85.16	33.59	46.56	54.46	62.66
Z_7 top	62.84	79.12	90.53	93.54	49.69	65.20	80.92	86.06	29.27	44.38	58.60	65.15
Z_8 top	49.64	74.60	93.37	96.97	43.79	61.64	79.37	86.57	30.10	36.39	57.43	63.29
Z_9 top	53.06	67.16	95.75	96.23	54.72	64.35	82.13	**88.05**	34.76	52.91	56.97	65.24
Z_{10} top	56.40	73.98	91.62	95.06	49.66	65.03	78.95	80.84	31.39	45.68	57.59	61.88

Tabela 7.9: ARI obtido por LGC semeado por representativos sobre redes AdaRadius construídas sobre dados em forma de bananas.

	S=0.6				S=1				S=1.4			
	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%
amostra rand	62.29 ±5.53	82.69 ±4.50	94.43 ±1.19	97.45 ±0.61	56.36 ±5.09	70.87 ±4.20	82.45 ±1.80	86.29 ±0.86	36.42 ±4.52	47.46 ±4.20	60.98 ±2.31	65.99 ±1.01
deg top	67.29	79.15	92.64	95.31	59.99	70.76	79.15	82.75	37.17	50.03	58.18	59.08
Katz top	65.20	70	88.76	93.70	55.47	65.89	73.82	81.38	38.36	45.16	53.33	59.40
PgRC top	**74.39**	**87.34**	**96.16**	**98.74**	**68.54**	**75.58**	83.27	**87.67**	**44.80**	**53.91**	62.09	66.91
RWC top	54.64	83.39	93.77	**98.88**	48.99	74.22	83.13	**87.31**	26.73	42.30	61.61	**67.09**
C_h top	62.02	70.69	88.25	95.39	38.05	48.47	67.23	74.28	14.24	24.90	39.86	46.37
CC top	0	0	0	71.79	0	0	0	66.59	0	0	0	34.08
btw top	**76.17**	73.07	89.70	**99.71**	61.25	70.56	76.17	83.68	20.84	24.56	46.30	55.93
H_{deg} top	65.36	85.67	95.41	**98.62**	60.63	**75.44**	83.01	**88.42**	**41.39**	49.46	61.36	**68.06**
H_{CC} top	66.85	85.89	95.13	97.48	57.15	68.11	83.56	**87.17**	39.68	48.15	59.27	66.45
H_{btw} top	**73.72**	84.62	96.13	**98.43**	**66.51**	72.64	84.18	**87.67**	**45.24**	**54.58**	62.80	66.78
Z_1 top	**72.16**	85.70	**97.62**	**98.61**	**69.71**	73.98	84.24	**88.25**	**46.66**	**55.58**	60.42	**67.23**
Z_2 top	**76.93**	86.90	**96.82**	97.60	59.62	66.24	80.33	86.24	**41.44**	41.56	61.31	63.93
Z_3 top	**73.71**	84.29	**98.97**	**99.62**	**68.32**	73.95	82.08	86.42	37.48	33.03	43.67	59
Z_4 top	46.70	66.01	88.50	91.12	50.05	58.19	79.32	85.91	27.53	38.08	55.30	65.23
Z_5 top	59.72	**91.93**	**98.40**	**98.97**	47.14	68.35	81.48	86.28	25.14	43.94	58.15	64.60
Z_6 top	64.11	76.19	89.72	92.67	55.37	70.45	78.48	84.98	32.92	45.90	55.30	62.64
Z_7 top	60.49	77.22	88.17	92.26	44.97	65.10	81.29	86.45	31.64	39.57	55.72	64.43
Z_8 top	47.77	73.64	92.41	96.46	41.39	60.80	77.71	85.44	27.65	39.35	54.09	61.96
Z_9 top	50.23	71.07	92.71	95.73	54.96	63.34	79.42	85.83	36.42	**52.73**	54.02	65.10
Z_{10} top	57.15	73.33	89.09	94.49	51.42	66.13	79.40	80.41	33.42	45.25	54.72	61.64



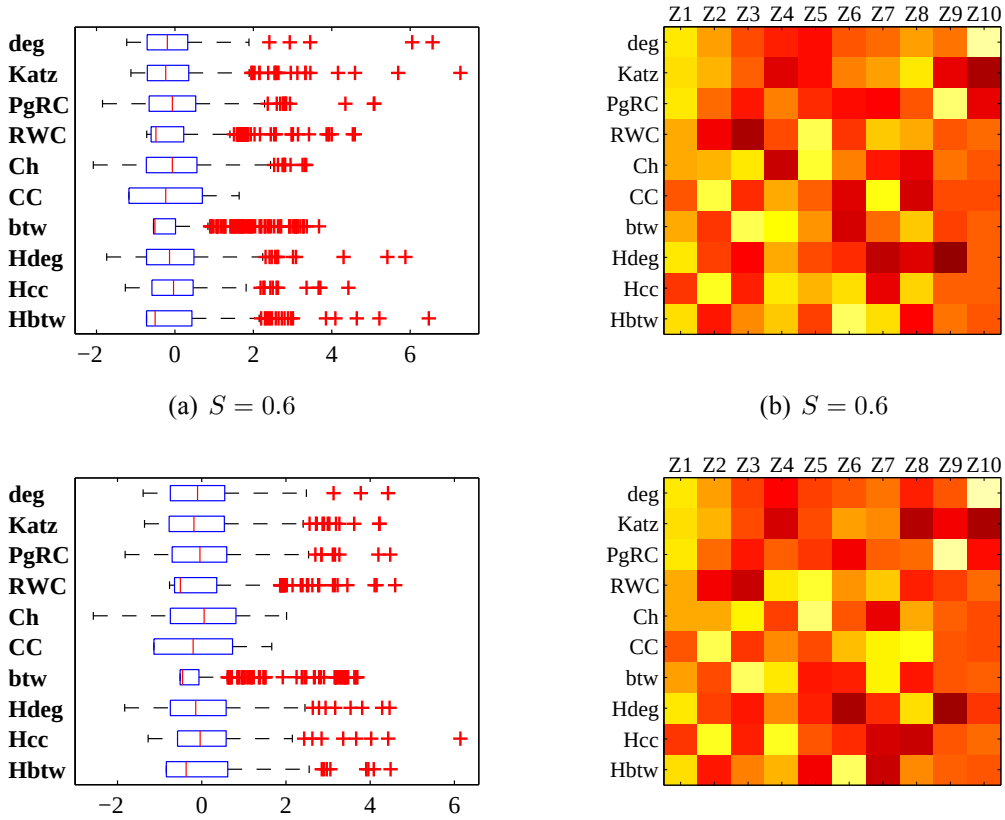
(a) $S = 0.6$ (b) $S = 0.6$

(c) $S = 1.4$ (d) $S = 1.4$

Figura 7.5: Distribuições de medidas de centralidade ((a) e (c)) e coeficientes de componentes principais sobre composições de medições de centralidade ((b) e (d)) para redes 5NN construídas sobre dados em forma de bananas com $n = 512$ e $S = 0.6$ (em cima) ou $S = 1.4$ (em baixo). As suíças dos *boxplots* são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$. Valores além disso são representados como *outliers* (símbolo '+'). Padronizados, pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. O mapa de cores da matriz de coeficientes varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1. Notavelmente, a composição dos 4 primeiros componentes para $S = 0.6$ e $S = 1.4$ são muito similares.

com 5% e 10% dos dados rotulados para $S = 1.4$, quando amostras aleatória apresentam bom desempenho, em geral, o critério Z_1 se destaca significantemente. Além da possibilidade de se compor um critério unificado de seleção de representativos, a principal vantagem oferecida pelo critério Z_1 é destacada por esses resultados. Z_1 funciona bem mesmo quando nenhum dos critérios de primeira e segunda ordem funciona bem. A medidas que compõem Z_1 (deg, Katz, PgRC, H_{deg} e H_{btw}) possuem distribuições bastante heterogêneas (veja a Figura 7.6). Para $S = 0.6$ outros critérios de terceira ordem também se destacam. Z_2 é essencialmente uma composição de CC e H_{CC}, Z_3 de btw e C_h, e Z_5 de RWC e C_h.

Sobre dados em forma de bananas, os resultados obtidos por AdaRadius são mais significativos de que os resultados de 5NN. Representativos são bem caracterizados por critérios de primeira e segunda ordem com distribuições heterogêneas bem como por suas composições utilizando análise de componentes principais. PgRC e Z_1 parecem prover um critério confiável para a seleção de representativos nesse domínio.



(a) $S = 0.6$ (b) $S = 0.6$

(c) $S = 1.4$ (d) $S = 1.4$

Figura 7.6: Distribuições de medidas de centralidade ((a) e (c)) e coeficientes de componentes principais sobre composições de medições de centralidade ((b) e (d)) para redes AdaRadius construídas sobre dados em forma de bananas com $n = 512$ e $S = 0.6$ (em cima) ou $S = 1.4$ (em baixo). As suíças dos *boxplots* são definidas para $Q_3 + 1.5(Q_3 - Q_1)$ e $Q_1 - 1.5(Q_3 - Q_1)$. Valores além disso são representados como *outliers* (símbolo '+'). Padronizados, pequenas diferenças entre Q_1 e Q_3 e muitos *outliers* à direita caracterizam distribuições cumulativas heterogêneas. O mapa de cores da matriz de coeficientes varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1. Notavelmente, a composição dos três primeiros componentes e de Z_5 para $S = 0.6$ e $S = 1.4$ são muito similares.

Novamente, alguém pode argumentar que a performance ruim do kNN se deve a má escolha de k. Tão logo, reiteramos uma das principais vantagens oferecidas por AdaRadius - baixa dependência à definição dos parâmetros de densidade. Contudo, realizamos mais discussões empregando AdaRadius e kNN sobre os conjuntos de dados de Chapelle et al. (2006), para os quais o valor ótimo de k é conhecido (veja a Tabela 7.1).

7.4.2 Resultados sobre conjuntos de dados de Chapelle et al. (2006)

As Tabelas 7.10-7.13 mostram o erro de teste de métodos de propagação de rótulo sobre redes kNN e AdaRadius construídas sobre os conjuntos de dados de Chapelle et al. (2006). Em geral, LGC funciona melhor que GFHF sobre amostras aleatórias e, presumivelmente, a correção de LGC é mais efetiva quando poucos dados estão disponíveis rotulados. Os resultados com seleção de representativos são muito mais estritos para LGC e especialmente sobre 10 itens de

dados rotulados. Por isso, os resultados dos critérios mais bem sucedidos é contrastado com os resultados com amostragem aleatória com menos rigor para LGC de que para GFHF.

Considerando os resultados apresentados na Sec. 7.3 e nossa expertise nesse domínio, nós discriminamos os conjuntos de dados de Chapelle et al. (2006) em 3 grupos de acordo com o tipo de desafio esperado: g241c e g241d são problemas de classificação desafiadores sob o viés errado; Digit1 e USPS são problemas de classificação fáceis, onde os ganho esperado é restrito se suficientes itens de dados rotulados estão disponíveis; e COIL e BCI são problemas desafiadores por natureza. COIL é um problema de classificação multiclasse com alguma sobreposição, enquanto que as variáveis de BCI são pouco informativas - coeficientes de regressões polinomiais de terceira ordem de series temporais de EEG. Os conjuntos de dados g241c e g241d, por sua vez, são projeções em alta dimensão de dados de baixa dimensão com distribuição Gaussiana: g241c com sobreposição parcial entre as classes, satisfazendo propriamente a suposição de grupos; e g241d com duas variedades ambas compostas por itens de dados de ambas as classes, não satisfazendo a suposição de grupos. Logo, quando apenas grafos binários são considerados certamente incorremos na maldição da dimensionalidade. Os itens de dados estão espalhados de forma tão esparsa no espaço hiperdimensional que a distância entre pares de dados é quase a mesma, sendo pequenas variações nos parâmetros de densidade grandemente sentidas. A distância normal l_2 entre pares de dados em g241c varia em $[16.93, 26.98]$. Grafos resultantes de ϵRadius são bastante fragmentados (62 componentes) para $\epsilon_i = 20$; conectados para $\epsilon_i = 20.9$; ou muito densos (53% de todas as possíveis conexões são estabelecidas) para $\epsilon_i = 22$. Assim, exceto pela redução de dimensionalidade ou apropriada ponderação das arestas métodos de ASS baseados em redes enfrentarão grandes desafios.

De um modo geral, os critérios que se destacaram com 100 instâncias rotuladas não se destacaram necessariamente com 10. Com 10 instâncias rotuladas, o desempenho de ASS semeado pelos critérios estudados sobre ambos AdaRadius e kNN foram pouco regulares. Critérios diversos se destacaram sobre conjuntos de dados diversos. Exceto sobre g241c e g241d, critérios diferentes se destacam sobre redes kNN e AdaRadius. Sobre o conjunto de dados Digit1 os critérios RWC, Katz, e deg se destacaram em redes kNN, enquanto que os critérios RWC, btw, Katz, PgRC e CC se destacaram em redes AdaRadius. Sobre o conjunto de dados USPS, apenas H_{btw} e apenas considerando GFHF se destacou sobre redes kNN. Sobre redes AdaRadius, os critérios deg e Z_1 se destacaram considerando ambos GFHF e LGC, enquanto que H_{btw}, H_{deg}, PgRC e RWC se destacaram apenas ao considerarmos LGC. Sobre o conjunto COIL, apenas H_{btw} se destaca sobre redes kNN, contudo não significativamente. RWC e CC se destacaram significativamente sobre redes AdaRadius. Sobre o conjunto BCI, os todos critérios relacionados ao grau e adicionalmente btw e Z_1 se destacaram sobre redes kNN, enquanto que apenas o critério H_{CC} se destacou sobre redes AdaRadius. Por fim, a maioria dos critérios se destacaram significativamente sobre g241c e g241d considerando-se GFHF, mas não considerando-se LGC. Logo, concluímos que 10 instâncias rotuladas em 1500 são poucos exemplos para se prover critérios suficientemente gerais para a seleção de representativos. Os critérios estudados não podem nem mesmo garantir que exemplos de todas as classes são amostrados.

Tabela 7.10: ARI obtido por GFHF semeado por representativos sobre redes 5NN construídas sobre conjuntos de dados de Chapelle et al. (2006)

	10 nós rotulados						100 nós rotulados					
	g241c	g241d	Digit1	USPS	COIL	BCI	g241c	g241d	Digit1	USPS	COIL	BCI
90 amostra rand	*49.9* ±0.50	*49.76* ±0.58	*18.33* ±13.03	*18.32* ±2.60	*67.46* ±5.72	*50.12* ±1.88	*45.94* ±2.62	*44.81* ±2.27	*2.68* ±0.53	*6.72* ±2.13	*21.28* ±2.63	*46.46* ±2.55
deg top	**43.02**	**49.06**	5.97	20.13	71.14	**47.95**	**38.36**	42.86	2.36	7.43	54.71	**42.67**
Katz top	**46.91**	49.46	**2.82**	20.13	75.37	49.49	**38.43**	43.71	**2.14**	7.43	60.43	**43.67**
PgRC top	**43.02**	**49.06**	13.69	20.13	71.07	**47.69**	**38.43**	43.07	2.29	6.14	41.86	44.33
RWC top	**45.57**	49.87	**2.82**	19.53	82.48	49.49	**38.64**	42.86	**2.14**	5.64	68.29	48
C_h top	**46.91**	**47.25**	47.52	19.66	82.21	50.51	**37.43**	43.29	2.29	11.07	65.5	49.33
CC top	50.2	49.93	7.92	19.87	83.96	50.51	52.79	44	2.36	9.57	37.93	50
btw top	**45.3**	**46.58**	49.8	23.22	78.99	49.23	**39**	43.5	2.43	**4.21**	57.64	44.33
H_{deg} top	**46.64**	**48.86**	16.64	18.99	66.85	48.97	**36.79**	43.36	2.21	5.07	28.93	**43.33**
H_{CC} top	49.87	**48.32**	16.85	18.72	65.03	52.05	44.29	45.07	**2.14**	11.36	83.86	50.33
H_{btw} top	**46.64**	**48.59**	9.53	**14.56**	62.89	50.26	**37.21**	43.07	2.57	4.71	21.64	**43**
Z_1 top	**43.02**	**49.06**	12.89	19.26	66.98	48.46	**38.43**	42.86	**2.21**	4.93	36.50	**42**
Z_2 top	**48.26**	**48.32**	20.13	20.13	78.99	50.77	**43.07**	**42.43**	3	13.57	62.57	48.33
Z_3 top	**48.79**	49.26	**4.16**	18.12	87.45	54.62	45.93	47	2.36	7.21	83.57	52
Z_4 top	49.60	49.93	29.73	19.80	78.66	51.03	47.57	**41.71**	2.64	**4.29**	34.14	53.33
Z_5 top	**47.52**	**48.86**	26.11	20.13	71.41	49.49	**39.36**	44.21	3	18.14	82.71	46.33
Z_6 top	**45.50**	**47.32**	20.34	16.98	63.15	**45.38**	**39.79**	44.50	2.79	8.64	**16.07**	44.67
Z_7 top	**49.06**	49.60	9.40	16.78	83.22	49.23	47.36	47.86	2.36	10.64	59.36	44.67
Z_8 top	**48.19**	**49.13**	42.28	19.53	75.17	48.46	**41.14**	43	2.36	9.50	52	49
Z_9 top	50	**48.99**	49.26	**8.19**	75.37	50.26	44.86	**41.64**	**2.21**	**4.21**	50.07	46
Z_{10} top	**41.41**	**48.19**	12.15	20.13	79.60	52.56	**40.36**	45.43	6.86	21.36	64.29	49.67

Tabela 7.11: Erro de teste obtido por LGC semeado por representativos sobre redes 5NN construídas sobre conjuntos de dados de Chapelle et al. (2006)

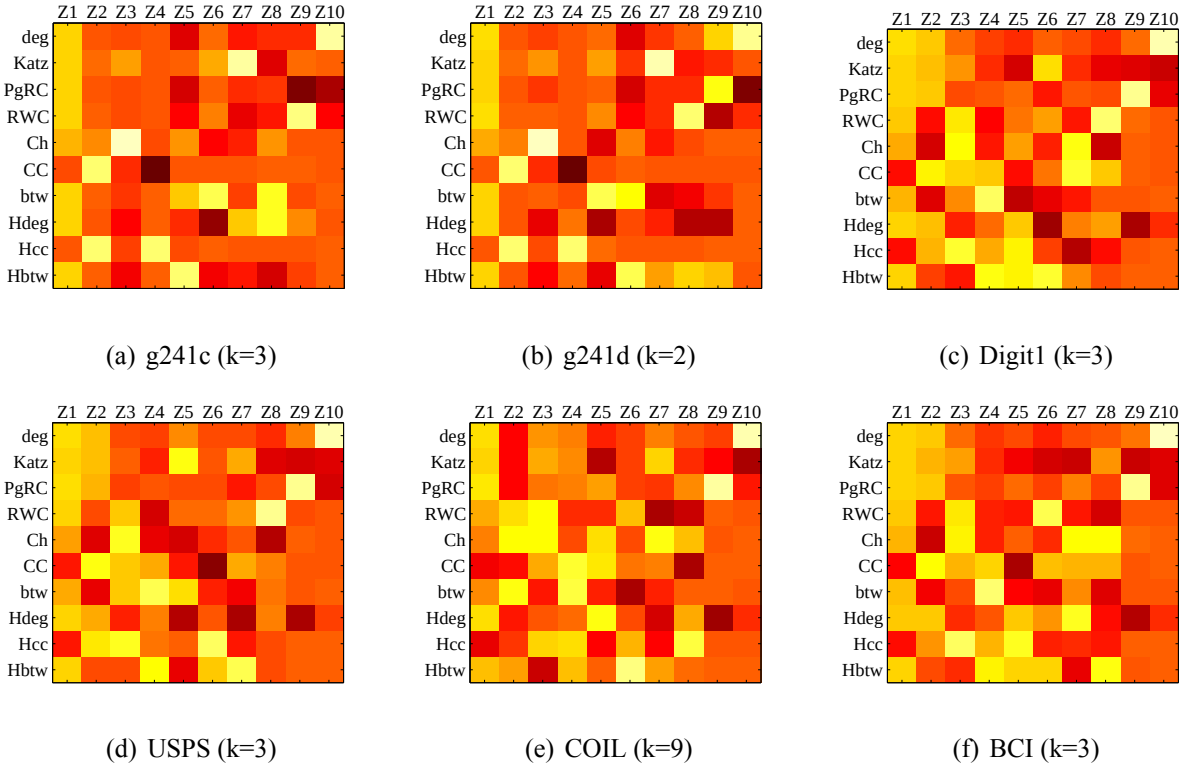
	10 nós rotulados						100 nós rotulados					
	g241c	g241d	Digit1	USPS	COIL	BCI	g241c	g241d	Digit1	USPS	COIL	BCI
90 amostra rand	*48.95* ±1.48	*49.04* ±1.67	*12.02* ±5.27	*16.92* ±5.38	*65.85* ±4.64	*50.42* ±1.72	*45.29* ±2.92	*44.15* ±2.14	*3.25* ±0.78	*6.38* ±1.96	*21.93* ±2.69	*48.08* ±2.80
deg top	**46.38**	50.13	8.72	20.13	71.14	48.72	**38.64**	**40.79**	**2.14**	7.71	55.29	47.67
Katz top	**46.17**	49.93	**4.7**	20.13	75.37	**48.46**	**39.79**	**39**	**2.5**	8.43	60.5	**45**
PgRC top	**46.38**	50.13	10.74	20.13	71.07	**47.18**	**40**	42.21	**2.29**	6.93	44.43	**44.33**
RWC top	**45.64**	49.93	**3.89**	16.64	81.74	48.72	**41.14**	**39.29**	2.64	9.29	75.14	50.33
C_h top	**46.17**	47.72	38.05	19.6	84.16	50	45.21	**41.36**	2.64	14.71	75.29	51
CC top	50.2	49.33	13.62	20.54	83.96	48.97	52.71	45.21	3.29	6.36	39.07	48.67
btw top	**44.77**	**46.51**	44.83	27.92	81.68	**44.62**	**39.86**	**39.71**	**2.36**	4.64	59.14	51.33
H_{deg} top	**45.57**	49.66	19.87	22.75	66.85	48.72	**37.79**	**41.21**	**2.29**	5.86	29.79	47.33
H_{CC} top	49.6	49.87	14.77	23.49	65.91	53.59	43.86	43.86	2.79	10.71	38.43	52.33
H_{btw} top	**45.57**	48.86	13.42	20	63.02	52.56	**38.07**	**41.64**	2.79	5.57	24.29	49.33
Z_1 top	**46.38**	50.13	15.84	17.11	66.98	**45.90**	**39.79**	**40.21**	2.64	6.14	41.36	**44.33**
Z_2 top	**46.85**	49.87	8.99	20.13	85.23	52.82	**42.07**	44.71	2.71	8.29	64.57	48.67
Z_3 top	48.19	48.66	7.72	24.56	81.21	51.54	48.21	44.79	4.43	14.43	74.57	51
Z_4 top	**46.24**	50.47	25.10	24.77	81.81	**47.69**	44.93	42.21	**2.43**	7	34.43	50.33
Z_5 top	47.65	51.14	11.54	20.13	71.68	**47.95**	47.07	43.07	3.21	17.64	52.57	52.67
Z_6 top	47.52	**46.51**	13.62	30.87	62.48	**45.90**	43.14	46.57	4	5.93	**16.86**	48.67
Z_7 top	47.92	49.66	24.56	12.28	75.77	**47.95**	46.64	47.79	5.71	10.71	60.21	48
Z_8 top	**45.70**	48.59	21.41	16.17	75.97	51.54	**42.21**	45.50	3	11.79	53.43	49
Z_9 top	50.34	47.45	49.26	40.47	79.40	48.97	46.43	**40.79**	3.50	8.21	51.57	47.33
Z_{10} top	**42.68**	**47.11**	21.61	20.13	79.53	51.54	43.93	42.79	10.21	21.36	64.64	50.33

Tabela 7.12: Erro de teste obtido por GFHF semeado por representativos sobre redes AdaRadius construídas sobre conjuntos de dados de Chapelle et al. (2006)

	10 nós rotulados						100 nós rotulados					
	g241c	g241d	Digit1	USPS	COIL	BCI	g241c	g241d	Digit1	USPS	COIL	BCI
90 amostra rand	*50.01* ±0.23	*50.01* ±0.20	*20.23* ±11.89	*19.26* ±1.62	*64.98* ±6.20	*50.09* ±1.44	*48.72* ±2.32	*48.13* ±2.29	*4.71* ±1.21	*11.12* ±3.70	*17.76* ±3.54	*48.15* ±2.70
deg top	**47.18**	**49.13**	16.31	**14.09**	71.68	51.79	**36.57**	**39.5**	5.64	7.79	**13.79**	**44**
Katz top	**47.18**	**49.13**	15.64	36.31	71.68	51.54	**36.43**	**39.86**	5.5	11.36	23.36	48.33
PgRC top	**47.18**	**49.13**	8.79	18.19	60.94	51.28	**36.64**	**39.07**	**3.21**	**5**	**11.36**	46
RWC top	**43.62**	**49.13**	9.13	37.65	**45.5**	52.05	**36**	**36.86**	**3.14**	**5.43**	**11.64**	**43**
C_h top	**47.18**	**49.13**	25.97	37.45	67.11	50.26	**36.29**	**39.29**	6.14	12.57	55.36	51
CC top	50.34	50.2	13.96	20.13	**51.34**	**48.97**	53.07	50.36	3.64	15.07	22.07	49.67
btw top	**47.18**	**45.91**	**5.7**	24.7	62.15	52.82	**35.93**	**37.64**	8.57	**4.71**	17.21	46.67
H_{deg} top	**47.18**	**48.99**	26.58	19.73	61.07	53.33	**36.79**	**37.79**	**2.93**	**4.07**	**12.57**	45.67
H_{CC} top	50	49.87	28.86	19.87	70	**48.72**	49.79	46.57	**2.93**	16	29.71	50
H_{btw} top	**47.18**	**49.13**	47.45	19.66	62.89	52.05	**36**	**36.21**	3.86	**4.79**	**10.57**	51.33
Z_1 top	**47.18**	**49.13**	27.58	**16.31**	61.48	51.03	**36.71**	**38.43**	4.93	**4.93**	**12.64**	**43.67**
Z_2 top	50	50.20	32.68	20	71.68	50.51	48.21	50.50	4.64	20.21	52.86	50.67
Z_3 top	**47.58**	50.20	**7.65**	20.13	70.34	52.05	**38.50**	**40.71**	6	21	34.79	49.33
Z_4 top	50	50.07	35.37	52.95	67.99	50.51	**35.43**	48.93	4.43	19.86	26.71	46.33
Z_5 top	50.13	50.13	36.51	59.40	59.87	**47.44**	**38.71**	50.14	5	10.14	33.21	48.33
Z_6 top	49.93	50.34	43.83	71.21	**52.21**	50.77	47.14	46.79	4.79	16.29	16.07	48.67
Z_7 top	**49.46**	50.07	32.89	25.10	74.30	51.79	**43.64**	**36.86**	4.21	20	48.07	47
Z_8 top	**44.30**	**45.37**	30.13	19.80	64.50	50	**38.71**	**42.36**	5.64	15.64	47.57	**44.00**
Z_9 top	49.93	**48.39**	23.56	18.66	61.88	49.23	47.21	**35.07**	4.71	13.14	41	52
Z_{10} top	**49.73**	**49.06**	10.81	20.13	62.15	50.26	**37.86**	47.07	6.29	20.71	33.14	49.33

Tabela 7.13: Erro de teste obtido por LGC semeado por representativos sobre redes AdaRadius construídas sobre conjuntos de dados de Chapelle et al. (2006)

	10 nós rotulados						100 nós rotulados					
	g241c	g241d	Digit1	USPS	COIL	BCI	g241c	g241d	Digit1	USPS	COIL	BCI
90 amostra rand	*49.52* ±1.44	*49.44* ±1.07	*16.64* ±6.93	*15.68* ±5.19	*62* ±5.40	*50.14* ±1.53	*47.87* ±3.37	*47.44* ±3.01	*4.87* ±1.26	*9.04* ±3.03	*15.67* ±3.40	*48.75* ±2.77
deg top	49.87	49.93	17.18	14.7	71.95	52.05	**43.43**	50.36	5.36	21	17.29	**45.33**
Katz top	49.87	49.93	12.01	19.93	71.95	50.26	**40.29**	50.5	5.57	21.14	24.64	**44.33**
PgRC top	49.87	49.93	14.83	**12.21**	60.34	49.23	**43.5**	50.36	**2.93**	9.14	12.36	49
RWC top	**44.23**	49.93	**9.06**	14.83	**45.5**	52.31	**35.57**	49.64	**3.36**	13	**11.14**	**43.67**
C_h top	49.87	49.93	22.21	13.49	67.05	48.72	**40.29**	50.93	8.14	21.14	51.57	48.67
CC top	50.34	49.53	14.23	20.13	**52.82**	53.08	53.07	50.07	4.29	6.14	21.29	50.33
btw top	49.87	**46.04**	11.14	22.28	62.55	50.77	**34.71**	49.14	8.43	7	18.43	47.33
H_{deg} top	49.87	50	25.17	**12.08**	61.14	50.51	**39.29**	50	**2.64**	**5.93**	**11.64**	**44.67**
H_{CC} top	49.93	50.6	25.1	31.07	68.52	**47.95**	49	49.71	**3.43**	9.5	25.71	49.67
H_{btw} top	49.87	49.93	30.13	**11.95**	62.75	51.28	**34.5**	48.79	**3.57**	**4.14**	**11.21**	49.67
Z_1 top	49.87	49.93	22.15	12.21	61.41	49.23	**39.36**	50.21	5.64	14.71	13.93	46
Z_2 top	50	50.20	27.11	22.01	71.68	50.26	49.64	50.50	5.07	19.36	61.36	53
Z_3 top	45.64	50.13	12.82	20.13	70.74	51.28	50.07	50.50	6.21	21.07	33.71	48.67
Z_4 top	49.87	49.26	31.01	48.46	71.95	49.23	48.21	45.64	4.64	7	21.93	48
Z_5 top	50.13	49.13	34.63	53.36	**54.63**	47.18	**41.71**	49.43	8.07	16.14	37.43	49
Z_6 top	50.13	50.34	27.25	65.84	**52.28**	50.77	50.29	49.93	4.50	**4.79**	**9.64**	47.67
Z_7 top	49.93	50.07	28.12	20.07	71.61	50.77	48.64	**41.14**	**3.43**	11.64	46.07	49.33
Z_8 top	48.79	**42.95**	24.50	14.50	65.10	51.79	46.07	50.07	6.21	**3.71**	47.43	**45.67**
Z_9 top	50	**47.99**	12.42	14.36	65.10	54.36	50.43	44.79	5.71	11.14	37.36	47.33
Z_{10} top	49.87	49.46	26.64	20.13	62.21	50.51	47	50.57	6.29	21.21	32.14	52



(a) g241c (k=3) (b) g241d (k=2) (c) Digit1 (k=3)

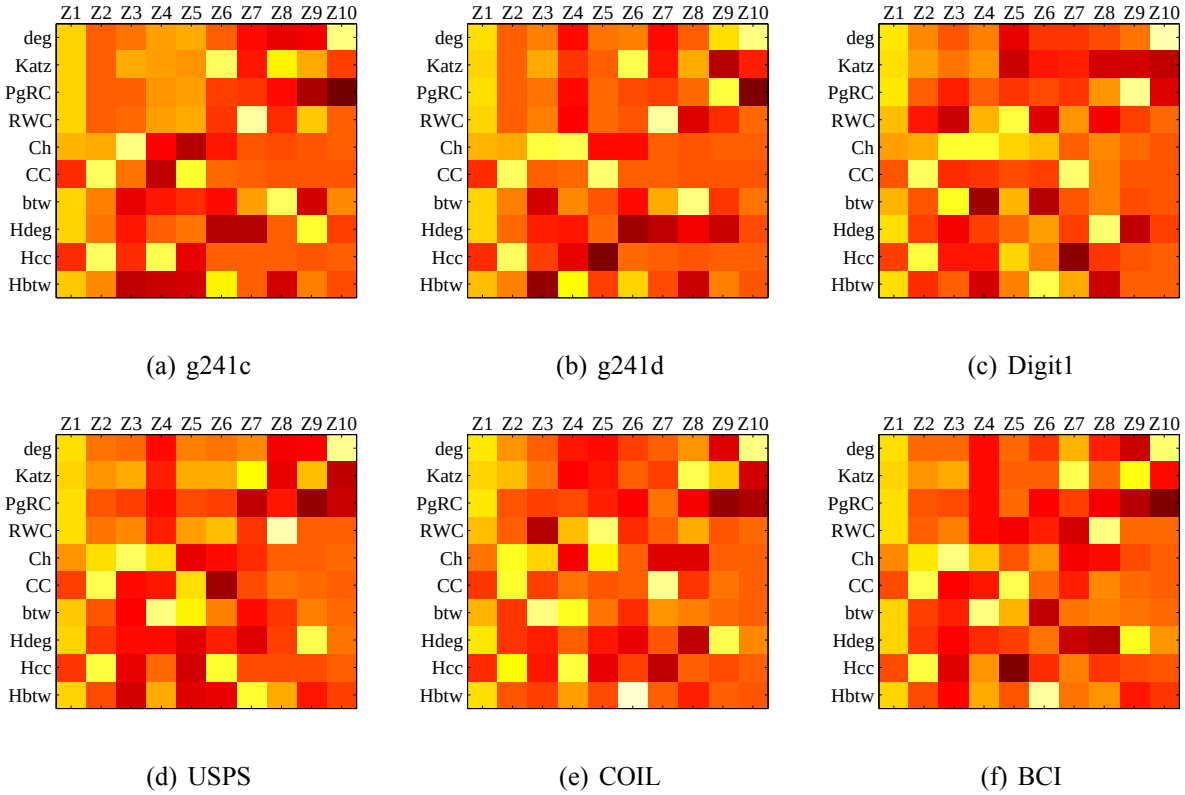
(d) USPS (k=3) (e) COIL (k=9) (f) BCI (k=3)

Figura 7.7: Coeficientes de componentes principais sobre composições de medições de centralidade em redes kNN construídas sobre os conjuntos de dados de Chapelle et al. (2006). O mapa de cores varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1.

Dispondo-se de 100 instâncias rotuladas, por outro lado, os resultados são mais significativos. Exceto sobre COIL, os critérios H_{deg} e H_{btw} se destacam sobre redes kNN. E, exceto com o método LGC sobre g241d, os critérios H_{deg}, RWC e PgRC se destacam. deg também se destaca, mas apenas ao considerarmos GFHF, e H_{btw} se destaca somente ao considerarmos LGC. Dentre esses critérios, H_{deg} merece atenção especial. Estritamente local, além de favorecer ao bom desempenho de ASS, essa abordagem se destaca como um critério de baixo custo, o que pode ser extremamente interessante ao considerarmos conjuntos de dados grandes e fluxos de dados contínuos.

Além disso, em geral critérios de primeira e segunda ordem não funcionam bem em redes kNN construídas sobre o conjunto de dados COIL. A amostragem aleatória funciona tão bem quanto ou melhor em média. Bastante denso e com alguma sobreposição entre suas 6 classes, representativos são difíceis de se encontrar considerando os critérios estudos aqui. Apenas o critério Z_6 funciona bem nesse domínio. Como mostra a Figura 7.7 (e), Z_6 é uma composição bem específica de H_{btw}, btw invertido, e um pouco de RWC e H_{CC}. Contudo, AdaRadius supera os problemas enfrentados por kNN de tal forma que este pode ser considerado apropriado para o tratamento de problemas multiclasse desafiadores. Notavelmente, AdaRadius também supera kNN quando o ASS é semeado por amostragem aleatória.

Exceto sobre o conjunto de dados COIL, Z_1 funciona bem como critério de seleção de representativos em redes kNN. Em redes AdaRadius, Z_1 funciona tão bem quanto a amostragem



(a) g241c (b) g241d (c) Digit1

(d) USPS (e) COIL (f) BCI

Figura 7.8: Coeficientes de componentes principais sobre composições de medições de centralidade em redes AdaRadius construídas sobre os conjuntos de dados de Chapelle et al. (2006). O mapa de cores varia em $[-1, 1]$. Quanto maior o coeficiente, mais claro; vermelho (cinza escuro) significa'zero'; e preto -1.

aleatória ou um pouco melhor considerando-se o método LGC, contudo se destaca significantemente ao considerarmos GFHF. Como mostram as Figuras 7.7 e 7.8, Z_1 é frequentemente composto de medidas relacionadas ao grau e ao fluxo, bem como C_h, cujas distribuições se destacam em termos de variância. Em geral se espera que os primeiros principais componentes (que possuem maior variância) possuam maior informação agregada, mas esse não parece ser o caso aqui.

Além de Z_1, outros componentes principais também se destacam, incluindo alguns de pequena variância, que algumas vezes funcionam melhor que Z_1 como critério de seleção de representativos. É o caso de Z_6 na rede kNN construídas sobre o conjunto COIL. Como Jolliffe (2002) argumenta, os últimos componentes principais não são apenas meras sobras não estruturadas dos componentes principais conhecidamente relevantes. A medida H_{btw}, que essencialmente compõe Z_6 (veja a Figura 7.7 (e)), não é muito correlacionada com as medidas que compõem os primeiros componentes principais, contribuindo com apenas um pouco de sua variância para as suas composições, sendo deixadas as suas especificidades para compor um componente específico, Z_6. Logo, deduzimos que o critério Z_1 apenas se destaca se uma parte razoável dos critérios de primeira e segunda ordem se destacam, de modo que o componente principal resultante de sua combinação possua grande variância. Forte, a suposição de que informação está atrelada à variância, muitas vezes não é satisfeita, especialmente quando apenas o topo dos ranques é de interesse. E concluímos que Z_1 apenas pode ser considerado um critérios de

seleção de representativos confiável em domínios bem conhecidos onde parte considerável dos critérios de primeira e segunda ordem são conhecidamente critérios razoáveis de seleção de representativos. Caso contrário, ao invés de realçar representativos, estes podem se suprimidos.

7.4.3 Comparação com o benchmark de Chapelle et al. (2006)

Por fim, contrastamos o ASS semeado com representativos e o *benchmark* de Chapelle et al. (2006). Como nenhum dos critérios de seleção de representativos estudado pode ser considerado suficientemente geral, consideramos aqueles mais bem sucedidos em cada configuração de conjunto de dados, método de construção de redes e método de rotulação. Além disso, também reportamos o erro de teste médio para as 12 divisões providas por Chapelle et al. (2006) como linha base de performance. Para cada conjunto de dados ordenamos os modelos em ordem ascendente de desempenho e apresentamos sua posição média considerando no cenário com 10 e 100 instâncias disponíveis rotuladas. Os resultados são compilados na Tabela 7.14.

A despeito de otimização de parâmetros ou preprocessamento dos dados, a seleção de representativos se destaca como uma questão bastante importante para o aprendizado semissupervisionado. A principal evidência disso é que ela tira os métodos GFHF e LGC das últimas posições do ranque médio para as primeiras posições. Embora tenhamos estudado apenas métodos de ASS da abordagem baseada em redes, estamos certos que melhoras de desempenho também podem ser obtidas em métodos de outras abordagens ao se rotular manualmente representativos.

Concluímos também que, quando a suposição de variedade é satisfeita, AdaRadius não funciona melhor que os métodos amplamente usados de construção de grafos para o ASS. Redes kNN permitem que métodos de rotulação semissupervisionados tirem grande vantagem mesmo de sementes pouco representativas nesse caso. O método LGC, por exemplo, recorrige rótulos inconsistentes nesse sentido. No entanto, AdaRadius se destaca grandemente quando representativos são selecionados para semear a rotulação semissupervisionada, e especialmente em problemas difíceis de classificação. Por isso, critérios suficientemente gerais de seleção de caracterização de representativos precisam ser estudados e propostos.

Tabela 7.14: Erro de teste e ranque médio obtidos por modelos de ASS semeados por representativos e *benchmark* de Chapelle et al. (2006)

	10 nós rotulados						Ranque	100 nós rotulados						Ranque
	g241c	g241d	Digit1	USPS	COIL	BCI	Médio	g241c	g241d	Digit1	USPS	COIL	BCI	Médio
1-NN	47.88	46.72	13.65	16.66	63.36	49	**11.5**	43.93	42.45	3.89	5.81	17.35	48.67	**15.17**
SVM	47.32	46.66	30.60	20.03	68.36	49.85	**16.5**	23.11	24.64	5.53	9.75	22.93	34.31	**11.5**
MVU + 1-NN	47.15	45.56	14.42	23.34	62.62	47.95	**11**	43.01	38.20	2.83	6.50	28.71	47.89	**14.17**
LEM + 1-NN	44.05	43.22	23.47	19.82	65.91	48.74	**12.17**	40.28	37.49	6.12	7.64	23.27	44.83	**14.67**
QC + CMR	39.96	46.55	9.80	13.61	59.63	50.36	**8.67**	22.05	28.20	3.15	6.36	10.03	46.22	**9.17**
Discrete Reg.	49.59	49.05	12.64	16.07	63.38	49.51	**12.83**	43.65	41.65	2.77	4.68	9.61	47.67	**10.33**
TSVM	24.71	50.08	17.77	25.20	67.50	49.15	**14.83**	18.46	22.42	6.15	9.77	25.80	33.25	**11.33**
SGT	22.76	18.64	8.92	25.36	-	49.59	**10.5**	17.41	9.11	2.61	6.80	-	45.03	**9.17**
Cluster-Kernel	48.28	42.05	18.73	19.41	67.32	48.31	**12.5**	13.49	4.95	3.79	9.68	21.99	35.17	**8.83**
Data-Dep. Reg.	41.25	45.89	12.49	17.96	63.65	50.21	**11**	20.31	32.82	2.44	5.10	11.46	47.47	**8**
LDS	28.85	50.63	15.63	17.57	61.90	49.27	**11.33**	18.04	23.74	3.46	4.96	13.72	43.97	**7.33**
Laplacian RLS	43.95	45.68	5.44	18.99	54.54	48.97	**7.83**	24.36	26.46	2.92	4.68	11.92	31.36	**6.33**
CHM (normed)	39.03	43.01	14.86	20.53	-	46.90	**11**	24.82	25.67	3.79	7.65	-	36.03	**12**
*k*NN														
+ GFHF	50.08	49.98	12.17	18.19	67.24	49.76	**15.17**	46.92	44.89	2.7	6.33	21.02	45.89	**13.5**
+ LGC	49.47	48.53	12.42	13.98	66.45	50	**13.00**	45.69	43.81	3	6.33	21.68	47.94	**15.17**
AdaRadius														
+ GFHF	50.04	49.88	13.65	20.45	64.44	50.09	**16.50**	47.91	47.98	4.17	11.02	17.47	46.69	**17.83**
+ LGC	49.66	48.94	16.12	14.55	63.68	50.49	**15.33**	46.69	46.63	4.59	8.36	15.74	48	**17.17**
*k*NN + representativos														
+ GFHF	41.41	46.58	2.82	8.19	62.89	45.38	**5.17**	36.79	41.64	2.14	4.21	16.07	42	**8**
+ LGC	42.68	46.04	7.72	11.95	45.5	45.90	**4.83**	34.5	40.21	2.43	4.14	11.14	43.67	**7.17**
AdaRadius + representativos														
+ GFHF	43.62	45.37	5.7	14.09	45.5	47.44	**5.33**	35.43	35.07	2.93	4.07	10.57	43	**7.33**
+ LGC	44.77	42.95	3.89	12.21	52.28	44.62	**4**	37.79	39	2.29	3.71	9.64	44.33	**6.83**

Capítulo

8

Validação de métodos de construção de grafos

Ao construir uma rede a partir de uma base de dados vetorial buscamos por uma abstração capaz de sintetizar as relações de afinidade entre suas instâncias eficazmente. Esse problema é equivalente à redução não linear de dimensionalidade e ao aprendizado de suavidade (Roweis & Saul, 2000; Belkin & Niyogi, 2003; Zhu, 2005; Belkin et al., 2006). Nesse sentido, podemos fazer uso de suposições diversas, levando ou não os itens de dados disponíveis rotulados em consideração (Rohban & Rabiee, 2012; Cupertino et al., 2013; Jebara et al., 2009; Carreira-perpiñán & Zemel, 2004).

Amplamente utilizada na validação de métodos de detecção de comunidades e como função objetivo de métodos de detecção de comunidades baseados em otimização (Newman & Girvan, 2004; Danon et al., 2005; Clauset et al., 2004; Newman, 2004b, 2006b; Duch & Arenas, 2005; Fortunato, 2010), propusemos a utilização do índice de modularidade normalizado como critério de validação de métodos de construção de redes (Capítulo 7). Tomando como referência bases de dados de classificação cujos rótulos dos itens de dados são conhecidos *a priori*, contrastamos o número de arestas intra variedade com o número total de arestas estimado se as arestas fossem estabelecidas ao acaso. Embora essa medida possa ser utilizada consistentemente na avaliação relativa de partições para uma determinada rede, a sua utilização na comparação de redes geradas por diversos métodos pode não ser apropriada. Isso acontece, pois, ao avaliarmos arestas em relação à discrepância entre os rótulos de seus vértices extremidades, damos maior peso a vértices bastante conectivos em detrimento de vértices pouco conectivos. Além disso, a variação do número de estruturas modulares dos grafos avaliado resulta no deslocamento da faixa de valores possíveis aferidos pela modularidade. Como discutido no Capítulo 7, grafos com dispersão significativa de grau e desbalanço entre o tamanho de suas estruturas modulares configuram-se,

portanto, um desafio, de modo que a proposição de abordagens melhor fundamentadas é um tópico de interesse.

8.1 Erro de reconstrução

Com fins à redução de dimensionalidade, Roweis & Saul (2000) propuseram o erro de reconstrução das instâncias de uma base de dados,

$$\varepsilon = \sum_i \left\| \mathbf{x}_i - \sum_j w_{ji}\mathbf{x}_j \right\|^2, \tag{8.1}$$

onde $\mathbf{W}$ é dita a matriz de pesos de reconstrução, cujas entradas, w_{ij}, $\forall i, j$, satisfazendo à propriedade de simetria, invariantes a rotações, reescalagem, e translações dos itens de dados e respectivos vizinhos, são estimadas por mínimos quadrados através da minimização de ε. E, portanto, podem ser fixados e utilizados na reconstrução de uma base de dados p-dimensional, $\mathbf{X}$, $n \times p$, em um espaço q-dimensional, $q \ll p$.

A Eq. 8.1 é utilizada semelhantemente como função objetivo em (Belkin & Niyogi, 2003), na construção de auto-mapeamentos. Wang & Zhang (2008), por sua vez, adaptam-na de modo a restringir a vizinhança de cada item de dados,

$$\begin{aligned} \varepsilon_i &= \left\| \mathbf{x}_i - \textstyle\sum_{j:\mathbf{x}_j \in \text{Viz}(\mathbf{x}_i)} w_{ji}\mathbf{x}_j \right\|^2 \\ &= \left\| \textstyle\sum_{j:\mathbf{x}_j \in \text{Viz}(\mathbf{x}_i)} w_{ji}(\mathbf{x}_i - \mathbf{x}_j) \right\|^2 \\ &= \textstyle\sum_{j,k:\mathbf{x}_j,\mathbf{x}_k \in \text{Viz}(\mathbf{x}_i)} w_{ij}w_{ik}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_k) \\ &= \textstyle\sum_{j,k:\mathbf{x}_j,\mathbf{x}_k \in \text{Viz}(\mathbf{x}_i)} w_{ij}G^i_{jk}w_{ik}, \end{aligned} \tag{8.2}$$

onde $\text{Viz}(\mathbf{x}_i)$ é o conjunto dos r vizinhos mais próximos de $\mathbf{x}_i$ e G^i_{jk} representa a (j, k)-ésima entrada da matriz de todos os produtos internos entre eles, dita matriz de Gram local.

$$\mathbf{G}^i = \begin{vmatrix} (\mathbf{x}_i - \mathbf{x}_1)^T(\mathbf{x}_i - \mathbf{x}_1) & \cdots & (\mathbf{x}_i - \mathbf{x}_1)^T(\mathbf{x}_i - \mathbf{x}_r) \\ (\mathbf{x}_i - \mathbf{x}_2)^T(\mathbf{x}_i - \mathbf{x}_1) & \cdots & (\mathbf{x}_i - \mathbf{x}_2)^T(\mathbf{x}_i - \mathbf{x}_r) \\ \vdots & \ddots & \vdots \\ (\mathbf{x}_i - \mathbf{x}_r)^T(\mathbf{x}_i - \mathbf{x}_1) & \cdots & (\mathbf{x}_i - \mathbf{x}_r)^T(\mathbf{x}_i - \mathbf{x}_r) \end{vmatrix}, \tag{8.3}$$

onde $r = |\text{Viz}(\mathbf{x}_i)|$ é o tamanho da vizinhança de $\mathbf{x}_i$. A solução de

$$\begin{aligned} \min_{w_{ij}} \quad & \textstyle\sum_{j,k:\mathbf{x}_j,\mathbf{x}_k \in \text{Viz}(\mathbf{x}_i)} w_{ij}G^i_{jk}w_{ik} \\ \text{sujeito a} \quad & \textstyle\sum_j w_{ij} = 1; w_{ij} \geq 0, \end{aligned} \tag{8.4}$$

por programação quadrática, por exemplo, resulta na matriz $\mathbf{W}$ cujas entradas equivalem aos pesos do grafo normalizado ótimo em que cada vértice é combinação linear de seus vizinhos, e pode ser reconstruído por eles.

8.2 Medida proposta

Fazendo uso do *framework* apresentado acima, propomos a utilização da seguinte medida como critério de validação de métodos de construção de grafos. Dado um conjunto de redes construídas a partir de base de dados de referência, cujos os rótulo dos nós é conhecido *a priori*, avaliamos a capacidade de reconstrução do rótulo y_i, $\forall i$, a partir da combinação linear do rótulo de seus vizinhos, y_j,

$$\varepsilon^* = \sum_i \left\| \sum_{j:\mathbf{x}_j \in \text{Viz}(\mathbf{x}_i)} w_{ji}(y_i - y_j) \right\|^2 . \tag{8.5}$$

Conhecidos $\text{Viz}(\mathbf{x}_i)$, y_i, y_j e w_{ij} para todo i, j, podemos definir o índice de coerência local grafo-rótulo,

$$Q = \frac{1}{n} \sum_{i \in V} Q_i, \tag{8.6}$$

onde

$$Q_i = \frac{\sum_{j \in \mathbf{V}} w_{ji} \delta_{ij}}{\sum_{j \in \mathbf{V}} w_{ji}}, \tag{8.7}$$

onde Δ é dado por uma função de similaridade, no caso,

$$\delta_{ij} = \delta_{ji} = \begin{cases} 1 & \text{se } y_i = y_j \\ -1 & \text{se } y_i \neq y_j \end{cases} . \tag{8.8}$$

A Eq. 8.6 corresponde à função inversa da Eq. 8.5, sendo avaliada a similaridade entre y_i e y_j, ao invés da dissimilaridade. Cada vértice i é avaliado com relação à coerência entre seu rótulo e o rótulo de seus vizinhos ponderados pelos respectivos pesos de reconstrução. Além disso, normalizados os pesos de reconstrução, igual relevância é dada a todos os vértices. Medidas de similaridade contínuas também podem ser consideradas como função Δ, bem como descritores multirótulos, difusos ou probabilísticos. Por simplicidade, no entanto, com fins à discussão objetiva das vantagens dessa abordagem, consideraremos por ora apenas a rotulação rígida $y_i \in \{1, 2, ..., C\}$ e apenas grafos não ponderados, onde $w_{ij} \in \{0, 1\}$. Se vértices adjacentes, i e j, possuem o mesmo rotulo, $y_i = y_j$, Q_i é incrementado. Caso contrário, Q_i é decrementado. Cada nó i é igualmente relevante para o cálculo de Q, somando $1/n$ de Q. Como ilustram os problemas *toy* apresentados nas Tabelas 8.1 e 8.2, nós desconexos não contribuem para a soma e nós de borda de comunidades contribuem apenas parcialmente. Q varia no intervalo $[-1, 1]$, tendendo a -1 no caso de redes dissortativas e a 1 no caso de grafos com estruturas modulares bem definidas que se conformam com as classes.

No aprendizado semissupervisionado baseado em redes, rótulos são propagados não apenas diretamente entre vértices adjacentes, mas também em cascata através da topologia da rede. Essa é uma das principais vantagens do aprendizado baseado em redes, a valorização da similaridade

topológica para além da similaridade geométrica. Portanto, para a avaliação mais ampla da qualidade de um grafo, é de grande valia à computação de votos indiretos além dos votos diretos. Assim, propomos também a generalização da medida de coerência local, utilizando na Eq. 8.7 medidas de influência indireta ao invés do peso de reconstrução. Dentre diversas abordagens possíveis, consideramos o status de Katz (Katz, 1953), reformulando a Eq. 8.7 da seguinte maneira:

$$Q_i = \frac{\sum_{j \in \mathbf{V} \setminus \{i\}} S_{ji} \delta_{ij}}{\sum_{j \in \mathbf{V} \setminus \{i\}} S_{ji}}, \tag{8.9}$$

onde

$$S_{ji} = \sum_{k=1}^{\infty} \alpha^k (\mathbf{W}^k)_{ji}, \tag{8.10}$$

onde α é um fator de amortização das contribuição dos vértices na soma, em razão da distância topológica ao nó avaliado. Em notação matricial,

$$\mathbf{S} = \sum_{k=1}^{\infty} (\alpha \mathbf{W}^T)^k. \tag{8.11}$$

Seja $\mathbf{A}$ uma matriz quadrada não negativa, dita matriz primitiva, cujos os autovalores são dados por $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\lim_{k \to \infty} \mathbf{A}^k = 0 \iff \rho(\mathbf{A}) < 1, \tag{8.12}$$

onde $\rho(\mathbf{A})$, $\rho(\mathbf{A}) = \max_i(\lambda_i)$, é dito o autovalor de Perron Frobenious, também denominado raio espectral (van Mieghem, 2010).

Portanto, a fim de garantir a convergência de $\mathbf{S}$, $(\alpha \mathbf{W}^T)$ deve ser transformado em uma matriz primitiva. E assim, como uma série somável de Cesàro, $\mathbf{S}$ convergirá de acordo com a seguinte fórmula fechada:

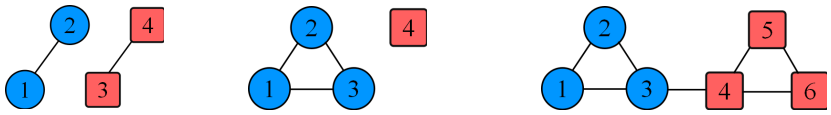
$$\mathbf{S} = (\mathbf{I} - \alpha \mathbf{W}^T)^{-1} \iff 0 < \alpha < \hat{\alpha}, \tag{8.13}$$

onde $\hat{\alpha} = 1/\rho(\alpha \mathbf{W}^T)$ e $\mathbf{I}$ é uma matriz identidade $n \times n$. Quanto menor α, maior o amortecimento de $\mathbf{S}$, de modo que a Eq. 8.9 comporta-se semelhante à Eq. 8.7 para $\alpha \to 0$, computando apenas votos diretos. Por outro lado, à medida que $\alpha \to \hat{\alpha}$, mais suave o amortecimento e portanto vizinhanças mais amplas são consideradas. Em outras palavras, Q pode ser computado localmente, regionalmente, e até mesmo globalmente, de acordo com o valor atribuído a α. E, por isso, α deve ser escolhido tendo em mente quão grande são as estruturas modulares consideradas relevantes. De todo modo, no entanto, as contribuições de cada nó em $\mathbf{S}$ são amortecidas de acordo com a distância topológica dos mesmos ao nós de referência.

8.3 Problemas toy

As Tabelas 8.1 e 8.2 mostram como a medida proposta, doravante denominada coerência grafo-rótulos de Katz (KC), se comporta em contraste com a medida de modularidade (NM) para valores diversos de α. São considerados, $\alpha = [0.1 : 0.2 : 0.9] \times \hat{\alpha}$, bem como o caso específico para $\alpha \to 0$, denominado coerência grafo-rótulos local (LC). Além do score atribuído à rede, são apresentados scores de cada nó individualmente.

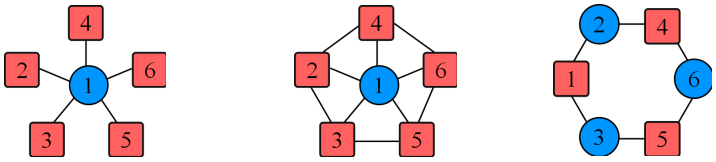
Tabela 8.1: Aferições da modularidade e coerência de Katz, variando-se α, sobre problemas *toy* assortativos.



Toy A Toy B Toy C

	Q_1-Q_4, Q	Q_1-Q_3	Q_4	Q	Q_1-Q_2, Q_5-Q_6	Q_3-Q_4	Q
NM	0.50	-	-	0	-	-	0.36
LC	1	1	0	0.75	1	0.33	0.78
KC, 0.1	1	1	0	0.75	0.96	0.31	0.74
KC, 0.3	1	1	0	0.75	0.85	0.26	0.65
KC, 0.5	1	1	0	0.75	0.69	0.19	0.53
KC, 0.7	1	1	0	0.75	0.47	0.08	0.34
KC, 0.9	1	1	0	0.75	0.11	-0.10	0.04

Tabela 8.2: Aferições da modularidade e coerência de Katz, variando-se α, sobre problemas *toy* dissortativos.



Toy D Toy E Toy F

	Q_1	Q_2-Q_6	Q	Q_1	Q_2-Q_6	Q	Q_1-Q_6, Q
NM	-	-	-0.5	-	-	-0.13	-0.5
LC	-1	-1	-1	-1	0.33	0.11	-1
KC, 0.1	-1	-0.70	-0.75	-1	0.35	0.12	-0.9
KC, 0.3	-1	-0.30	-0.42	-1	0.37	0.14	-0.73
KC, 0.5	-1	-0.06	-0.21	-1	0.40	0.17	-0.58
KC, 0.7	-1	0.11	-0.07	-1	0.43	0.19	-0.43
KC, 0.9	-1	0.23	0.03	-1	0.45	0.21	-0.28

Notavelmente, a medida proposta se distingue da modularidade ao levar em consideração votos indiretos (tome o Toy C, por exemplo). Além disso, suas medições variam linearmente no intervalo $[-1, 1]$ e a boa disposição de todo nó é avaliada igualitariamente (tome os Toys A e B,

por exemplo). A dissortatividade de uma rede (Toys D-F), por sua vez, pode ser avaliada em um contexto estritamente local, regional, ou global de acordo com o valor de α. O Toy F distingui-se dos demais como uma rede estritamente dissortativa; o Toy D caracteriza a má disposição de um *hub*; e o Toy E a disposição de um *outlier* contextual. Ao considerarmos votos indiretos além dos votos diretos, esses casos podem ser discriminados pela coerência de Katz.

8.4 Resultados e discussões

O modelo de rede Girvan-Newman é utilizado na construção de redes randômicas com distribuição de grau homogênea e estruturas modulares balanceadas. O rótulo dos nós está associado à comunidade atribuída ao nó pelo modelo. Número de nós, n; número de comunidades, C; grau médio, $\langle k \rangle$; e números ligações para fora dos módulos, Z_{out}, são definidos. Quanto maior Z_{out} maior a sobreposição entre as comunidades. Fixado Z_{out}, quanto maior $\langle k \rangle$ melhor definidas as comunidades. Fixados $\langle k \rangle$ e Z_{out}, quanto maior C, menores, mais densas, e mais bem definidas as comunidades.

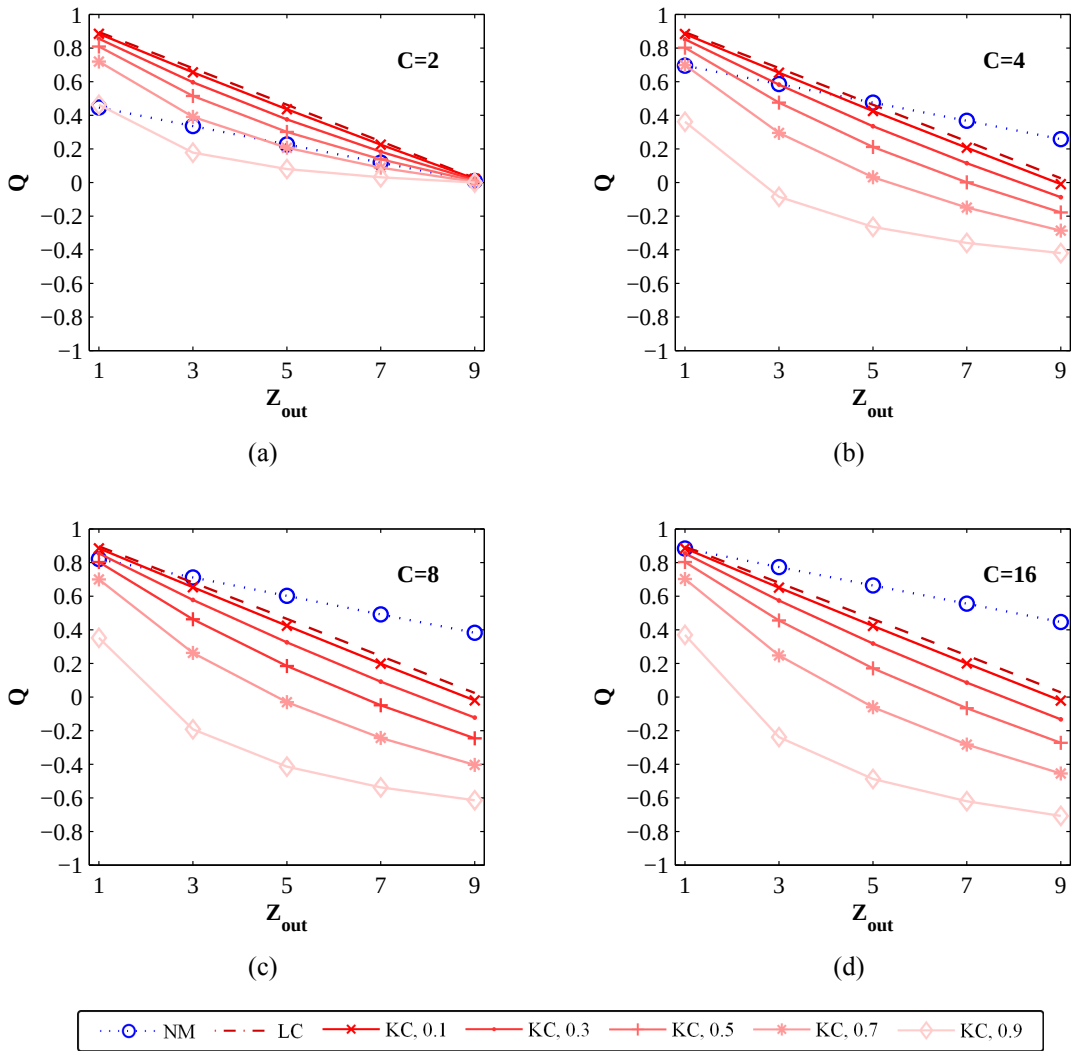


Figura 8.1: Aferições da modularidade e coerência de Katz, variando-se α, sobre redes GN variando-se C e Z_{out}.

A Fig. 8.1 mostra o comportamento da coerência grafo-rótulos de Katz em contraste com a modularidade e a coerência grafo-rótulos local ao variarmos C e Z_{out} e fixarmos os demais parâmetros do modelo. Os valores exibidos são a média de 30 ensaios. O desvio padrão observado é irrelevante. Para $C = 2$, a avaliação da modularidade se estende no intervalo $[0, 0.5]$ de acordo com a sobreposição entre as comunidades. A coerência de Katz, por sua vez, entre $[0, 1]$, variando-se a curva de acordo com α. As medições são bastante similares para Z_{out}=9, quando a rede torna-se efetivamente dissortativa. Para valores maiores de Z_{out} as aferições são simétricas, no intervalo $[-1, 0]$.

Para valores maiores de C, o número de arestas dentro de cada comunidade é maior, enquanto que o número de ligações para fora da comunidade é parcelado entre mais comunidades. Portanto, o ponto de dissortatividade configura-se apenas para valores maiores de Z_{out}, não exibidos na figura. Por outro lado, embora os valores de coerência de Katz sejam sempre 1 para comunidades mutuamente exclusivas, independente de C, a aferição da modularidade varia em razão do número de comunidades. Em todos os casos a coerência local se estende no intervalo $[0, 1]$ seja $Z_{out} \in [0, 9]$. Quanto maior α, maior a cobertura da coerência de Katz e, portanto, α precisa ser ajustado de acordo o tamanho estimado das comunidades de interesse, diga-se de acordo com a distância geodésica.

O modelo de Lancichinetti-Fortunato-Radicchi, por outro lado, é utilizado na construção de redes livres de escala com distribuição de grau heterogênea e estruturas modulares desbalanceadas. Novamente, o rótulo dos nós está associado à comunidade atribuída ao nó pelo modelo. Número de nós, n; grau médio, $\langle k \rangle$; mistura entre comunidades, μ; e coeficientes de escala das distribuições de grau e de tamanhos de comunidades, γ e β, são definidos. Ao considerarmos distribuições heterogêneas de grau e comunidades desbalanceadas, caracterizamos duas dificuldades conhecidamente enfrentadas pela medida de modularidade: redes com desvio padrão de grau significativo e variações significativas no tamanho das comunidades. O primeiro caso, se configura para $\gamma = 2$, enquanto que o segundo para $\beta = 2$. Para $\gamma = 3$, a distribuição de grau é bastante heterogênea e portanto muitos nós possuem grau semelhante, apenas alguns nós com grau dispersando-se significativamente. Além disso, para $\mu > 0.5$ temos redes dissortativas.

A Fig. 8.2 mostra o comportamento da coerência grafo-rótulos de Katz (KC) em contraste com a modularidade e a coerência grafo-rótulos local (LC) ao variarmos γ, β e μ e fixarmos os demais parâmetros do modelo. Novamente é exibida a média de 30 ensaios. Significativos, os desvios padrões são apresentados na Tabela 8.3 normalizados pela diferença entre os valores médio de Q aferidos para μ=0.1 e μ=0.6.

Notavelmente, as curvas das medidas são menos comportadas de que para redes do modelo Girvan-Newman. Contudo, em todo caso, avaliações da coerência local se estendem no intervalo $[0, 1]$ para $\mu = [0, 0.5]$. Incrementando-se o valor de α, pode-se observar a sensibilidade da medida à variação de β. Para $\gamma = 2$ temos dispersões menores de que para $\gamma = 3$.

A medida de modularidade, por outro lado, apresenta comportamento bastante restrito. Com pequena dispersão média para $\mu = [0.1, 0.6]$, apresenta variação significativa entre ensaios distintos. Para $\gamma = 2$, $\beta = 2$ o desvios das aferições é bastante expressivo, em média $\pm 29\%$ da

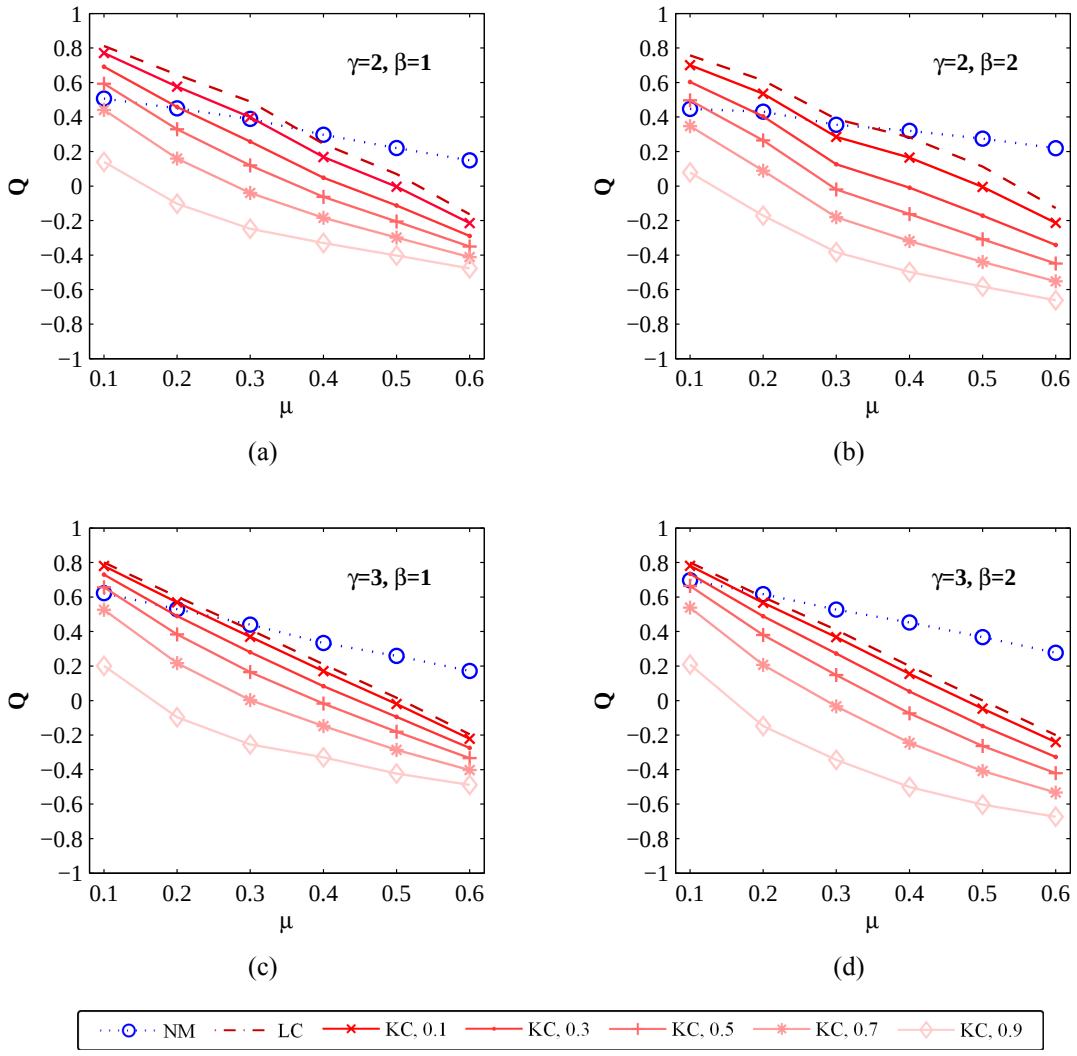
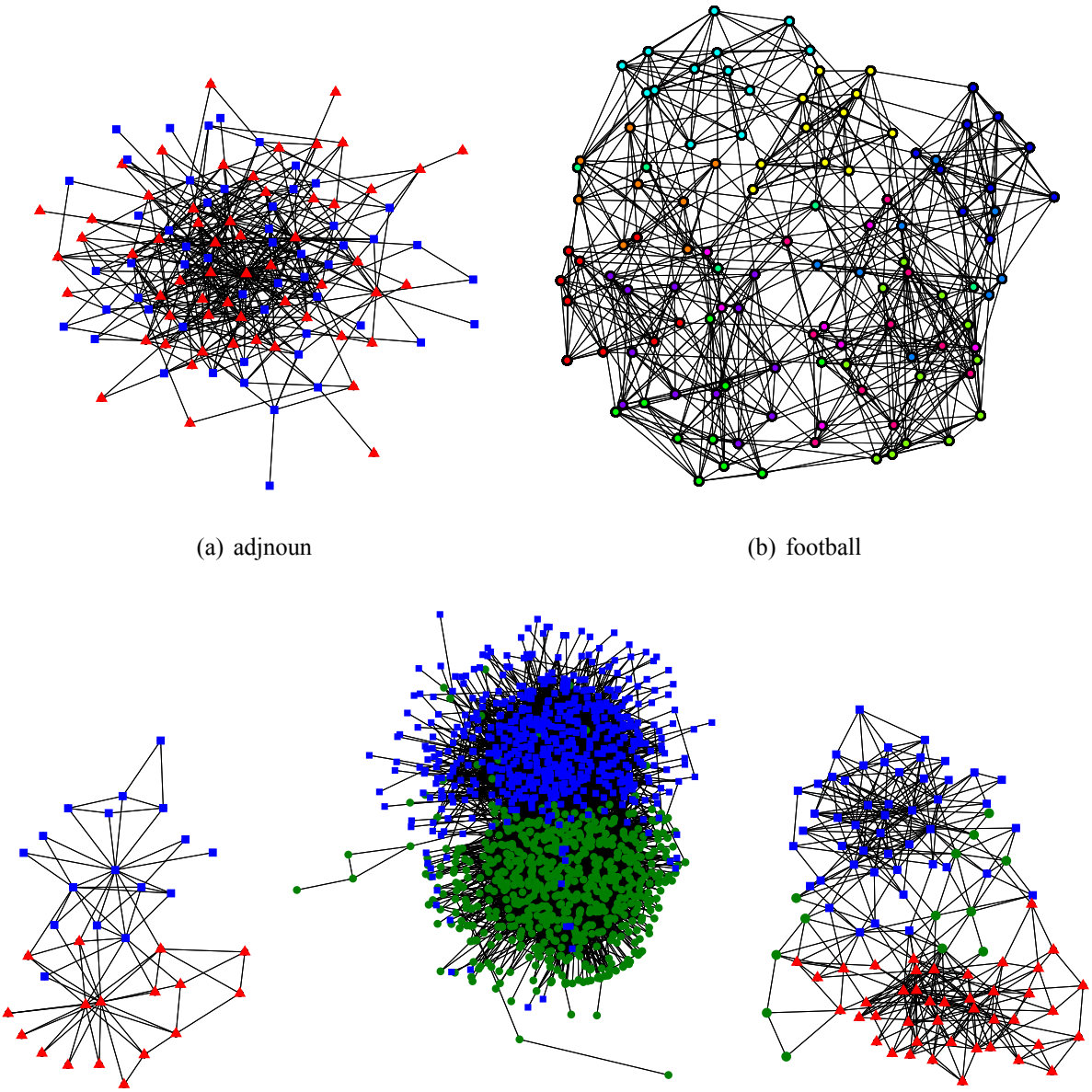


Figura 8.2: Aferições da modularidade e coerência de Katz, variando-se α, sobre redes LFR variando-se γ, β e μ.

diferença entre os valores médios para μ=0.1 e μ=0.6. Por isso, redes com níveis de sobreposição significativamente diferentes podem não ser razoavelmente discriminadas por essa medida, resultando em más avaliações. Bem menos expressivos, resultados pouco discriminativos da coerência de Katz são obtidos apenas para $\alpha = 0.9\hat{\alpha}$, quando a medida avalia a rede como um todo e não apenas seus módulos.

Por fim, aplicamos a medida proposta a redes reais nas quais os rótulos dos nós são *ground truth* conhecidos. Uma breve descrição dessas redes é apresentada na Tabela 8.4. Representações visuais são apresentadas na Figura 8.3. A Tabela 8.5, apresenta os índices obtidos pela coerência de Katz em comparação com a medida de modularidade e sua variação local.

Em geral, adjetivos estão associados a substantivos, e, por tanto, a rede adjnoun é significativamente dissortativa, sendo satisfatoriamente discriminado pela coerência de Katz. Com 12 comunidades *ground truth*, caracterizando os jogos regionais, a rede football é melhor avaliada pela medida de modularidade de que a rede dos clubes de caratê, que possui apenas duas comunidades. Avaliações mais razoáveis são obtidas pela medida de coerência de Katz para



(a) adjnoun (b) football

(c) karate (d) polblogs (e) polbooks

Figura 8.3: (Originalmente colorido) Redes de Mark Newman descritas na Tabela 8.4. Os rótulos *ground truth* das comunidades são discriminados pelas cores e formas dos vértices. O espalhamento visual da rede é realizado através do método de minimização de energia de um sistema de molas entre todos pares de vértices, cuja rigidez da mola entre cada par de vértices é proporcional à distância mínima entre os mesmos (Kamada & Kawai, 1989).

Tabela 8.3: Desvios padrões do resultados do ensaios apresentados na Figura 8.2 normalizados pela diferença entre os valores medidos para μ=0.1 e μ=0.6.

	γ=2, β=1, μ						γ=2, β=2, μ					
	0.1	0.2	0.3	0.4	0.5	0.6	0.1	0.2	0.3	0.4	0.5	0.6
NM	**0.19**	**0.22**	**0.21**	**0.20**	**0.23**	**0.18**	**0.40**	**0.26**	**0.40**	**0.23**	**0.21**	**0.24**
LC	0.04	0.04	0.09	0.08	0.11	0.05	0.10	0.08	0.15	0.10	0.09	0.07
KC, 0.1	0.04	0.04	0.06	0.06	0.10	0.03	0.12	0.08	0.15	0.08	0.07	0.04
KC, 0.3	0.06	0.05	0.06	0.06	0.10	0.04	0.14	0.09	0.15	0.07	0.05	0.02
KC, 0.5	0.08	0.07	0.08	0.07	0.12	0.07	0.15	0.10	0.14	0.07	0.06	0.03
KC, 0.7	0.10	0.10	0.11	0.10	0.15	0.11	0.16	0.11	0.12	0.08	0.07	0.06
KC, 0.9	0.14	0.18	**0.21**	**0.19**	**0.25**	**0.19**	0.14	0.12	0.09	0.10	0.11	0.11
	γ=3, β=1, μ						γ=3, β=2, μ					
	0.1	0.2	0.3	0.4	0.5	0.6	0.1	0.2	0.3	0.4	0.5	0.6
NM	**0.19**	**0.19**	**0.22**	**0.20**	**0.20**	**0.21**	**0.13**	**0.16**	**0.23**	**0.14**	**0.13**	**0.13**
LC	0.00	0.01	0.04	0.04	0.06	0.03	0.00	0.00	0.05	0.00	0.00	0.00
KC, 0.1	0.00	0.01	0.05	0.05	0.06	0.04	0.00	0.00	0.05	0.00	0.01	0.01
KC, 0.3	0.01	0.01	0.05	0.06	0.08	0.06	0.01	0.01	0.06	0.01	0.01	0.02
KC, 0.5	0.01	0.02	0.07	0.08	0.10	0.09	0.01	0.01	0.06	0.02	0.03	0.04
KC, 0.7	0.02	0.04	0.10	0.11	0.13	0.13	0.02	0.02	0.08	0.04	0.04	0.06
KC, 0.9	0.08	0.12	**0.20**	**0.22**	**0.24**	**0.24**	0.04	0.05	0.15	0.09	0.09	0.10

$\alpha \leq 0.1\hat{\alpha}$. Ambas as redes possuem comunidades balanceadas, com 5 a 13 times (mediana 10) e 17 indivíduos cada, respectivamente, sendo que football possui maior sobreposição entre as comunidades, caracterizando os jogos das finais. A despeito da densidade das comunidades em polblogs, polbooks é tão bem avaliado quanto. Avaliações mais razoáveis são oferecidas pela coerência de Katz para $\alpha \leq 0.5\hat{\alpha}$.

8.5 Considerações finais

Concluímos, portanto, que a medida proposta, denominada coerência grafo-rótulos de Katz, apresenta maior aptidão que a modularidade para a avaliação de métodos de construção de redes considerando bases de dados diversas de referência. Mesmo em problemas desafiadores, com dispersão de grau significativa, comunidades desbalanceadas, e sobreposição entre comunidades, a coerência de Katz apresentou descrição apropriada da coerência entre a rede e os rótulos. Nos casos estudados, boas estimações foram observadas para valores de $\alpha \in [0.1, 0.5] \times \hat{\alpha}$. Contudo, uma boa estimação do tamanho da comunidades de interesse (distância geodésica) pode ser necessária. Em se tratando de bases de dados de *benchmark*, não é difícil fazê-lo.

Bastante informativa, a coerência grafo-rótulos de Katz pode ser utilizada não apenas na validação de métodos de construção de redes, mas também na validação de métodos de identificação de comunidades em redes e até mesmo como função objetivo para a proposição de novos métodos. Além do status de Katz, outras medidas de influência indireta podem ser consideradas na definição de $\mathbf{S}$, por exemplo, a inversa do tempo médio de primeira passagem de uma caminhada aleatória (Brandes & Erlebach, 2005; Noh & Rieger, 2004; Tejedor et al., 2009). A Eq. 8.7 também pode ser adaptada de modo a ponderar a relevância de cada nó de acordo com

Tabela 8.4: Breve descrição de redes reais com comunidades *ground truth*.

rede	# nós	# arestas	# rótulos	descrição
adjnoun	112	425	2 {adjective, noun}	Rede de adjetivos e substantivos frequentes em uma mesma oração no romance de Charles Dickens - ‘‘David Copperfield’’ (Newman, 2006a).
football	115	613	12 {Atlantic Coast, Big East, Big Ten, Big Twelve, Conference USA, Independents, Mid-American, Mountain West, Pacific Ten, Southeastern, Sun Belt, Western Athletic}	Rede de jogos de futebol americano entre universidades da Divisão IA durante a temporada regular no outono de 2000 (Girvan & Newman, 2002).
karate	34	78	2	Rede de amizades entre 34 membros de um clube de caratê em uma universidade americana (Zachary, 1977).
polblogs	1222	16716	2 {liberal, conservative}	Blogoesfera das eleições americanas de 2004 (Adamic & Glance, 2005).
polbooks	105	441	3 {liberal, neutral, conservative}	Rede de livros sobre política americana frequentemente adquiridos conjuntamente na Amazon.com (Krebs, 2004).

Tabela 8.5: Aferições da modularidade e coerência de Katz, variando-se α, sobre redes reais descritas na Tabela 8.4.

rede	NM	LC	KC, 0.1	KC, 0.3	KC, 0.5	KC, 0.7	KC, 0.9
adjnoun	0.0340	-0.4956	-0.4282	-0.3132	-0.2152	-0.1272	-0.0440
football	0.5047	0.2724	0.2359	0.1459	0.0208	-0.1697	-0.5134
karate	0.3718	0.8042	0.7714	0.7032	0.6157	0.4796	0.2187
polblogs	0.2490	0.8077	0.7915	0.7549	0.7027	0.6090	0.3698
polbooks	0.2341	0.6375	0.6285	0.6103	0.5894	0.5595	0.4799

qualquer critério, por exemplo seu grau, considerando-se para isso a média harmônica; outras medidas de similaridade podem ser consideradas na Eq. 8.8: contínuas, não lineares, suaves e multirótulos.

Capítulo

9

Conclusão

O presente trabalho teve como principal objetivo explorar a influência dos objetos previamente rotulados no desempenho do aprendizado semissupervisionado. Tipicamente, a rotulação manual de dados requer o envolvimento de especialistas. Contudo, dificuldades, como por exemplo a visualização dos dados, podem prejudicar sua análise, de modo que o realce de itens de dados representativos para a obtenção de uma boa solução é bastante interessante. Em razão das diversas analogias encontradas em redes complexas, o escopo do presente trabalho foi definido no aprendizado semissupervisionado baseado em redes, onde conjuntos de dados são representados relacionalmente através da abstração gráfica. Critérios para a caracterização de nós representativos foram propostos, tomando como base medidas de centralidade em redes. Essa abordagem é considerada pois soluções corretas por otimização multiobjetivo são proibitivas. Resultados significativos nesse problema e em problemas relacionados foram alcançados e são reiterados a seguir juntamente com as principais contribuições do presente trabalho e perspectivas de trabalhos futuros.

Primeiramente (Capítulo 5 e Seção 6.1), a robustez dos métodos de aprendizado semissupervisionado baseados em redes foi estudada utilizando uma abordagem sistemática, onde indícios de que o aprendizado semissupervisionado depende da qualidade da amostra disponível rotulada foram obtidos, mesmo com a rotulação semissupervisionada sendo realizada por métodos bastante robustos. Além disso, o desempenho do aprendizado também varia de acordo com a consistência entre grupos e classes e de acordo com a conveniência do método de construção da rede para o método de rotulação. A robustez do aprendizado é melhorada significativamente ao se considerar restrições de ligações entre nós em razão de seus rótulos.

Em seguida (Capítulo 6), critérios para a caracterização de indivíduos representativos em redes com comunidades foram estudados. Contraintuitivamente, nós bastante conectivos (*hubs*) frequentemente não são muito representativos. Nós razoavelmente conectivos em vizinhanças

pouco conectivas, por outro lado, o são. Em redes do modelo Girvan-Newman, nós com alto coeficiente de agrupamento também mostram-se representativos. Por outro lado, em redes do modelo Lancichinetti-Fortunato-Radicchi, nós com alta intermedialidade se destacam. Nós com alto coeficiente de agrupamento em redes GN estão tipicamente situados em *motifs* do tipo quase-clique, emergentes da pseudo-aleatoriedade computacional das ligações estabelecidas pelo modelo. Nós com alta intermedialidade em redes LFR são *hubs* situados na borda das comunidades, interligando-as. Em ambos os casos, os nós identificados são excelentes regularizadores.

Contudo, como critérios distintos de caracterização de indivíduos representativos se destacam em modelos de rede distintos, faz-se necessária a combinação dos mesmos no sentido de propor critérios suficientemente gerais. Em geral, as medidas que melhor caracterizam indivíduos representativos são aquelas cujas distribuições são bastante heterogêneas. Tomando como princípio essa característica, a análise de componentes principais foi utilizada para a composição de componentes ortogonais. Ordenadas de acordo com a variância, as primeiras componentes principais apresentam maior ''informatividade''. E, dentre elas, a segunda componente principal se destaca. Bons resultados são obtidos em redes GN e em redes LFR com algum nível de sobreposição entre as comunidades. No entanto, resultados em redes LFR com comunidades bem definidas não são tão bons quanto os resultados obtidos pelos critérios simples. A técnica de análise de componentes principais utiliza a covariância entre as variáveis como entrada, de modo que a decomposição em autovalores e autovetores é apropriada apenas sobre variáveis com distribuições de Poisson. Não se sabe muito sobre a caracterização de variáveis com distribuições bastante heterogêneas, como as caracterizações de nós em redes LFR com pouca sobreposição entre comunidades, no entanto.

Esses resultados foram estendidos (Capítulo 7) para conjuntos de dados vetoriais. E para isso, abordagens de construção de redes foram estudadas. Embora o método de construção rede seja crucial para o bom desempenho do aprendizado, mesmo hoje, não existe consenso sobre qual abordagem é a mais benéfica ou em quais situações essa ou aquela abordagem se destaca. Fato é que a construção da rede é usualmente estudada juntamente com a rotulação e, por isso, é frequentemente subvalorizada. Portanto, foi proposto um método de construção de redes denominado AdaRadius. Baseado nas árvores geradoras mínimas e no método ϵRadius, este método apresenta diversas propriedades interessantes, tais como adaptabilidade a dados com densidade variada, baixa dependência da configuração de seus parâmetros e custo computacional razoável tanto sobre dados *pool based* como em dados incrementais. Além disso, as redes resultantes são esparsas e permitem que o método de rotulação tire grande vantagem da rotulação prévia de representativos.

A despeito de otimização de parâmetros ou preprocessamento dos dados, a seleção de indivíduos representativos se destaca como uma questão bastante importante para o aprendizado semissupervisionado. Como evidência disso, a semeadura de métodos de propagação de rótulos pouco robustos com indivíduos representativos os trás das últimas para as primeiras posições do raque comparativo com métodos bem estabelecidos. AdaRadius, por sua vez, proporciona

desempenho diferenciado de classificação quando representativos são selecionados para semear a rotulação semissupervisionada, e especialmente em problemas difíceis de classificação. Ser bastante conectado em vizinhanças compostas por nós pouco conectados faz desses nós representativos. Novamente, no entanto, critérios diversos de caracterização de representativos se destacaram sobre bases de dados diversas. Com amostras de dado suficientemente grandes, no entanto, critérios específicos se destacam, em especial a variação hierárquica de grau. Novamente, foi utilizada a análise de componentes principais, porém sobre a matriz de correlação, ao invés da matriz de covariância, de modo que a primeira componente principal se destacou como critério para a caracterização de representativos.

Amplamente utilizada na validação de métodos de detecção de comunidades e como função objetivo de métodos de detecção de comunidades baseados em otimização, o índice de modularidade normalizado foi utilizado como critério de validação de métodos de construção de redes em si mesmos, independente da etapa de rotulação semissupervisionada (Seção 7.3). Contudo, a sua utilização na comparação de redes geradas por métodos diversos mostrou-se inapropriada, pois em sua soma a modularidade dá maior peso a vértices bastante conectivos em detrimento de vértices pouco conectivos e valoriza redes granuladas em detrimento de redes conectadas, dando razão à proposição de novas abordagens.

Denominada coerência grafo-rótulos de Katz, uma medida de validação foi proposta baseada no erro de reconstrução linear e no status de Katz. Essa medida mostra-se bastante justa na avaliação de métodos de construção de redes considerando diversas bases de dados de referência. Mesmo em problemas desafiadores, com dispersão de grau significativa, comunidades desbalanceadas e sobreposição entre comunidades, avaliações apropriadas da coerência entre rede e rótulos foram obtidas. Bastante informativa, a coerência grafo-rótulos de Katz pode ser utilizada não apenas na validação de métodos de construção de redes, mas também na validação de métodos de identificação de comunidades em redes e até mesmo como função objetivo para a proposição de novos métodos de detecção de comunidades. Além disso, essa medida pode ser adaptada de modo a ponderar a relevância de cada nó de acordo com quaisquer critérios e de modo a considerar quaisquer métricas de similaridade, sejam contínuas, não lineares, suaves, e mesmo multirótulos.

9.1 Principais contribuições

Dentre as principais contribuições do presente trabalho, destacam-se:

- **A proposição de critérios para a seleção de representativos baseados em medidas de centralidade.** Especificamente, não são consideradas quaisquer informações adicionais sobre a definição de grupos ou comunidades. A hipótese levantada vai um pouco além desse tipo de abordagem. Para nós, a representatividade é um fenômeno diretamente relacionado à formação de comunidades em redes. Isto é, as estruturas de comunidades em uma rede são resultado da organização de seus indivíduos em torno de seus representativos. E, por isso, assumimos que estruturas de rede que reflitam devidamente o processo

de geração dos dados devem realçar seus representativos. Até onde sabemos, não são conhecidas abordagens semelhantes aplicadas ao aprendizado semissupervisionado.

- **A proposição de um método adaptativo de construção de rede a partir de dados vetoriais.** O método proposto consegue adaptar-se aos dados, considerando para isso a caracterização das regiões da rede, independente da definição de parâmetros de densidade. Consequência disso é que redes geradas por esse método podem adaptar-se à inserção/remoção incremental de itens de dados, sem a necessidade de sua reconstrução. Métodos tradicionais, por outro lado, são dependentes da definição arbitrária de parâmetros de densidade.

- **A proposição de abordagens para validação de métodos de construção da rede.** Ao utilizar a medida de modularidade para esse fim e propor a coerência grafo-rótulos de Katz, o estudo de métodos de abstração gráfica é enfatizado como um problema crucial para o bom desempenho do aprendizado semissupervisionado baseado em redes. Além disso, com boas medidas de validação pode-se propor, por exemplo, métodos de construção de redes baseados em otimização.

9.2 Trabalhos Futuros

Os avanços obtidos através do desenvolvimento desse trabalho permitiram, além da obtenção de resultados significativos, desdobramentos diversos. Destacamos a seguir alguns deles, sobre os quais são especulados resultados em curto e médio prazo.

A proposição de abordagens unificadas para a seleção de representativos é um tópico bastante desafiador. É natural concluir que em redes com características diversas, medidas de centralidade diversas serão capazes de caracterizar seus nós representativos. Por razões legítimas, segmentos inteiros de aprendizado de máquina são dedicados especificamente a esse problema.

Como desdobramento do presente trabalho,a agregação de ranques (meta-busca) vem sendo estudada aplicada ao cenário em que poucos ranques caracterizam bem indivíduos representativos. A suposição que fundamenta a agregação de ranques é a seguinte. Os ranques agregados são pertubações do ranque *ground truth* procurado (Adali et al., 2006; Dwork et al., 2001). Por tanto, se a maioria dos ranques concorda mas, como observado em nosso estudo, não são bons ranques de caracterização de representativos, o ranque agregado também não será. Contudo, fazendo uso da técnica análise de fatores comuns para a composição de variáveis comuns não observadas (Thompson, 2004), bons resultados têm sido obtidos.

Também desdobramento desse trabalho, um ponto que precisa ser estudado mais a fundo é a proposição de medidas apropriadas para a validação de métodos de construção de rede. Como destacado acima, medidas bem fundadas podem servir de função objetivo para métodos baseados em otimização, não apenas no domínio do aprendizado semissupervisionado, mas também no estudo de comunidades em redes, por exemplo.

Também destacado acima, o método AdaRadius apresenta diversas vantagens, dentre elas

custo razoável em ambientes incrementais. Diferente de muitos métodos tradicionais, não é necessária a reexecução do método quando do incremento dos dados. Combinado a métodos iterativos de rotulação semissupervisionada, como por exemplo, o método de competição e cooperação de partículas, portanto, o modelo proposto pode ser aplicado, por exemplo, a fluxos contínuos de dados. No entanto, para isso, é necessário o estudo do impacto causado por mudanças de conceito, e a razoabilidade da rotina de adaptação do grafo de acordo com o volume de dados fluindo a cada instante de tempo. Esse é um tópico de estudo de bastante interesse na atualidade.

Estritamente relacionado a esse tópico, temos também o estudo de grande volumes de dados, isto é, redes de grande escala e Big data. A maioria das medidas de centralidade aqui estudadas não é escalável a volumes de dados da ordem dezenas de milhares de itens de dados, mesmo utilizando técnicas de processamento massivos como Hadoop e MapReduce (Kang et al., 2011). Portanto faz-se necessário, nesse sentido, a proposição de medidas de centralidade bem sucedidas na caracterização de representativos, porém escaláveis, como, por exemplo, a variação hierárquica de grau.

Outro problema associado é a classificação multirótulo. Embora não tenha sido estudada no presente trabalho, é importante que esta seja considerada, tanto em abordagens *crisp* como em abordagens difusas. Especialmente em se tratando de redes, é muito comum que nós pertençam a mais de uma comunidade/grupo, algumas vezes com níveis de pertinência difusos, outras com pertinência não mensurável. Tanto medidas de validação de rede como a rotulação manual multirótulo precisam ser consideradas no domínio do aprendizado semissupervisionado semeado por indivíduos representativos. A coerência grafo-rótulos de Katz, por exemplo, pode ser facilmente adaptada para esse fim.

Como as abordagens de seleção de representativos aqui estudas não são adaptativas, isto é, seus resultados são conhecidos em uma única realização dos critérios de caracterização de representativos, optamos em não associar diretamente as proposições de seleção de representativos ao paradigma de aprendizado ativo. Contudo, alguns aspectos estudados no aprendizado ativo são de vital importância para a aplicação prática de proposições de seleção de representativos, dentre eles o custo variável da rotulação e a atuação de oráculos não confiáveis.

No presente estudo, avaliamos a qualidade da amostra selecionada para a rotulação manual apenas em função do ganho de desempenho do aprendizado semissupervisionado considerando custo constante. No entanto, em alguns cenários o custo com a rotulação manual pode ser variável. Exemplo disso é a rotulação de proteínas de acordo com suas funções. Se um método de realce de representativos destaca proteínas cujas funções são conhecidas, podemos assumir custo constante, mas se proteínas cujas funções estão em estudo são selecionadas, seus rótulos podem ser erroneamente definidos ou pode-se ter que esperar alguns anos de pesquisa até que se estes estejam rotulados confiavelmente. No primeiro caso, ao invés de melhorar o desempenho do aprendizado, podemos prejudicá-lo significativamente. Esses cenários precisam ser estudados tendo em vista a aplicação prática de soluções nesse domínio.

Por fim, além dos problemas específicos destacados acima, aplicações diversas são prospec-

tadas, especialmente no domínio dos sistemas complexos. Analogias diversas foram discutidas em diversos trechos do presente documento e serão estudadas mais a fundo.

Referências Bibliográficas

Adali, S., Hill, B., & Magdon-Ismail, M. (2006). The impact of ranker quality on rank aggregation algorithms: Information vs. robustness. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on* (pp. 37--37).: IEEE.

Adamic, L. A. & Glance, N. (2005). The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 36--43). New York, NY, USA: ACM.

Albert, R., Albert, I., & Nakarado, G. L. (2004). Structural vulnerability of the north american power grid. *Phys. Rev. E*, 69(2), 025103.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47--97.

Albert, R., Jeong, H., & Barabási, A.-L. (1999). Internet: Diameter of the world wide web. *Nature*, 401, 130--131.

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Attack and error tolerance of complex networks. *Nature*, 406, 378--382.

Allahverdyan, A. E., Steeg, G. V., & Galstyan, A. (2010). Community detection with and without prior information. *Euro Physics Letters*, 90, 18002.

Araújo, B. (2010). *Identificação de outliers em redes complexas baseado em caminhada aleatória*. Dissertação de mestrado, Universidade de São Paulo.

Araújo, B. & Zhao, L. (2013a). Detecting and labeling representative nodes for network-based semi-supervised learning. In *Neural Networks, The 2013 International Joint Conference on* (pp. 1729--1736). Los Alamitos, CA, USA: IEEE Computer Society.

Araújo, B. & Zhao, L. (2013b). Selecting nodes with inhomogeneous profile for labeling for network-based semi-supervised learning. In *Proceedings of the 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence* (pp. 428--434). Washington, DC, USA: IEEE Computer Society.

Arenas, A., Díaz-Guilera, A., & Pérez-Vicente, C. J. (2006). Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.*, 96, 114102.

Barabasi, A.-L. (2003). *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509--512.

Barrat, A., Barthélemy, M., & Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge University Press.

Barrat, A. & Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3), 547--560.

Barthélemy, M., Barrat, A., Pastor-Satorras, R., & Vespignani, A. (2004). Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Phys. Rev. Letters*, 92(17), 178701.

Bavelas, A. (1950). Communication patterns in task orientated groups. *Journal of the Acoustical Society of America*, 22, 271--288.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science*, 10, 161--163.

Belkin, M., Matveeva, I., & Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In *Conference on Learning Theory* (pp. 624--638). Banff, Canada.

Belkin, M. & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 1373--1396.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7, 2399--2434.

Berton, L. & de Andrade Lopes, A. (2014). Graph construction based on labeled instances for semi-supervised learning. In *Pattern Recognition, 22nd International Conference on* (pp. 2477--2482). Stockholm.

Biggs, N., Lloyd, E. K., & Wilson, R. J. (1976). *Graph Theory, 1736-1936*. Oxford University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blum, A. & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 19--26). San Francisco, CA, USA.

Blum, A. & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98 (pp. 92--100). New York, NY, USA: ACM.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424, 175--308.

Boldi, P. & Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, 10(3-4), 222--262.

Borge-Holthoefer, J., Meloni, S., Gonçalves, B., & Moreno, Y. (2013). Emergence of influential spreaders in modified rumor models. *Journal of Statistical Physics*, 151(1-2), 383--393.

Borge-Holthoefer, J. & Moreno, Y. (2012). Absence of influential spreaders in rumor dynamics. *Physical Review E*, 85(2), 026116.

Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2012). Locating privileged spreaders on an online social network. *Physical Review E*, 85(6), 066123.

Bornholdt, S. & Schuster, H. G. (2003). *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH.

Brandes, U. & Erlebach, T. (2005). *Network Analysis: Methodological Foundations*. Springer--Verlag.

Breve, F. (2010). *Aprendizado de máquina em redes complexas*. PhD thesis, Universidade de São Paulo.

Breve, F., Zhao, L., & Quiles, M. (2009). Particle competition in complex networks for semi-supervised classification. In *Complex Sciences* (pp. 163--174). Springer.

Breve, F., Zhao, L., Quiles, M., Pedrycz, W., & Liu, J. (2012). Particle competition and cooperation in networks for semi-supervised learning. *Knowledge and Data Engineering, IEEE Transactions on*, 24(9), 1686--1698.

Breve, F. A., Zhao, L., & Quiles, M. G. (2010). Semi-supervised learning from imperfect data through particle cooperation and competition. In *Neural Networks (IJCNN), The 2010 International Joint Conference on* (pp. 1--8).: IEEE.

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 59--66).

Bullmore, E. & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3), 186--198.

Burges, C. J. C. & Platt, J. C. (2006). Semi-supervised learning with conditional harmonic mixing. In O. Chapelle, A. Z. Schölkopf, & A. Zien (Eds.), *Semi-Supervised Learning* (pp. 251--273). The MIT Press.

Carreira-perpiñán, M. A. & Zemel, R. S. (2004). Proximity graphs for clustering and manifold learning. In *Advances in Neural Information Processing Systems 17* (pp. 225--232).

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(15), 1--57.

Chapelle, O., Schölkopf, A. Z., & Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press.

Chapelle, O., Weston, J., & Schölkopf, B. (2002). Cluster kernels for semi-supervised learning. In *Neural Information Processing Systems* (pp. 585--592).

Chapelle, O. & Zien, A. (2005). Semi-supervised classification by low density separation. In *Tenth International Workshop on Artificial Intelligence and Statistics* (pp. 57--64). Heinemann Medical, London.

Chazelle, B. (2000a). A minimum spanning tree algorithm with inverse-ackermann type complexity. *Journal of the Association for Computing Machinery*, 47(6), 1028--1047.

Chazelle, B. (2000b). The soft heap: an approximate priority queue with optimal error rate. *Journal of the Association for Computing Machinery*, 47(6), 1012--1027.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications*, 391(4), 1777--1787.

Chin, F. & Houck, D. (1978). Algorithms for updating minimal spanning trees. *Journal of Computer and System Sciences*, 16(12), 333--344.

Chong, K. W., Han, Y., & Lam, T. W. (2000). Concurrent threads and optimal parallel minimum spanning trees algorithm. *Journal of the Association for Computing Machinery*, 48(2), 297--323.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70, 066111.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4), 661--703.

Cohen, R., Erez, K., ben Avraham, D., & Havlin, S. (2000). Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85, 4626--4628.

Cohen, R., Erez, K., ben Avraham, D., & Havlin, S. (2001). Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86, 3682--3685.

Cohen, R. & Havlin, S. (2010). *Complex networks: structure, robustness and function*. Cambridge University Press.

Corduneanu, A. & Jaakkola, T. (2006). Data-dependent regularization. In O. Chapelle, A. Z. Schölkopf, & A. Zien (Eds.), *Semi-Supervised Learning* (pp. 16--190). The MIT Press.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Maximum Flow*, chapter 26. The MIT Press, second edition edition.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273--297.

Costa, L. d. F. & da Rocha, L. E. C. (2006). A generalized approach to complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 50(1), 237--242.

Costa, L. d. F., Rodrigues, F. A., Travieso, G., & Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1), 167--242.

Costa, L. d. F. & Silva, F. N. (2006). Hierarchical characterization of complex networks. *Journal of Statistical Physics*, 125(4), 845--876.

Cozman, F. & Cohen, I. (2006). Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. In *Semi-Supervised Learning* (pp. 57--72). MIT Press.

Cozman, F., Cohen, I., & Cirelo, M. C. (2003). Semi-supervised learning and model search. In *Proceedings of the ICML-2003 Workshop: The Continuun from Labeled to Unlabeled Data in Machine Learning and Data Mining* (pp. 111--112). Washington, DC.

Cupertino, T. H., Huertas, J., & Zhao, L. (2013). Data clustering using controlled consensus in complex networks. *Neurocomputing*, 118(22), 132--140.

Dangalchev, C. (2006). Residual closeness in networks. *Physica A: Statistical Mechanics and its Applications*, 365(2), 556--564.

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

de Sousa, C. R., Rezende, S. O., & Batista, G. E. A. P. A. (2013). Influence of graph construction on semi-supervised learning. In *Machine Learning and Knowledge Discovery in Databases* (pp. 160--175).: Springer Berlin Heidelberg.

Decelle, A., Krzakala, F., Moore, C., & Zdeborová, L. (2011). Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Letters*, 107, 065701.

Delalleau, O., Bengio, Y., & Roux, N. L. (2005). Efficient non-parametric function induction in semi-supervised learning. In *10th Int. Workshop on Artificial Intell. and Statistics* (pp. 96--103).

Duch, J. & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72, 027104.

Duin, R., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D., & Verzakov, S. (2007). *Prtools4.1, a matlab toolbox for pattern recognition*. Technical report, Delft University of Technology.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *10th International Conference on World Wide Web* (pp. 613--622). Hong Kong.

Erdös, P. & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6, 290--297.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226--231). Portland, USA: AAAI Press.

Everitt, B. S., Landau, S., & Leese, M. (2009). *Cluster Analysis*. Wiley, 4th edition.

Faceli, K., Lorena, A. C., Gama, J., & de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. LTC.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationship of the internet topology. *Computer Communication Review*, 29, 251--262.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of web communities. *IEEE Computer*, 35(3), 66--71.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486, 75--174.

Foster, K. C., Muth, S. Q., Potterat, J. J., & Rothenberg, R. B. (2001). A faster katz status score algorithm. *Computational & Mathematical Organization Theory*, 7(4), 275--285.

Fredman, M. L. & Tarjan, R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM*, 34(3), 596--615.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40, 35--41.

Freeman, L. C. (1978). Centrality in social networks: conceptual clarification. *Social Networks*, 1, 215--239.

Frey, B. J. & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814), 972--976.

Galeotti, A. & Goyal, S. (2009). Influencing the influencers: a theory of strategic diffusion. *The RAND Journal of Economics*, 40(3), 509--532.

Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *PNAS*, 99(12), 7821--7826.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). Switchboard: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP (pp. 517--520).

Goldberg, A. B., Zhu, X., Singh, A., Xu, Z., & Nowak, R. (2009). Multi-manifold semi-supervised learning. In *Twelfth International Conference on Artificial Intelligence and Statistics*, AISTATS'09.

Goldman, S. & Zhou, Y. (2000). Enhancing supervised learning with unlabeled data. In *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 327--334). San Francisco: Morgan Kaufmann.

Good, B. H., de Montjoye, Y.-A., & Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81, 046106.

Goyal, S., Heidari, H., & Kearns, M. (2014). Competitive contagion in networks. *Games and Economic Behavior*. to appear in.

Guha, S., Rastogi, R., & Shim, K. (1998). Cure: an efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data* (pp. 73--84). New York, NY, USA: ACM.

Guha, S., Rastogi, R., & Shim, K. (2001). Cure: an efficient clustering algorithm for large databases. *Information Systems*, 26(1), 35--58.

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

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70, 025101.

Gupta, S., Anderson, R. M., & May, R. M. (1989). Networks of sexual contacts: Implications for the pattern of spread of hiv. *AIDS*, 3, 807--817.

Hage, P. & Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, 17, 57--63.

Hoi, S. C., Jin, R., & Lyu, M. R. (2006a). Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web* (pp. 633--642).: ACM.

Hoi, S. C., Jin, R., Zhu, J., & Lyu, M. R. (2006b). Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning* (pp. 417--424).: ACM.

Holland, P. W. & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies*, 2(3), 107--124.

Holme, P., Huss, M., & Jeong, H. (2002). Attack vulnerability of complex networks. *Phys. Rev. E*, 65, 056109.

Holme, P., Huss, M., & Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4), 532--538.

Huang, L., Park, K., & Lai, Y.-C. (2006). Information propagation on modular networks. *Physical Review E*, 73(3), 035103.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193--218.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31, 651--666.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.

Jebara, T., Wang, J., & Chang, S.-F. (2009). Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 441--448). New York, NY, USA: ACM.

Jeong, H., Mason, S., Barabási, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411, 41--42.

Jeong, H., Tombor, B., Albert1, R., Oltvai, Z. N., & Barabási, A.-L. (2000). The large scale organization of metabolic networks. *Nature*, 407, 651--654.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99 (pp. 200--209). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Joachims, T. (2003). Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning* (pp. 290--297).

Johnson, D. B. & Metaxas, P. (1996). Optimal algorithms for the single and multiple vertex updating problems of a minimum spanning tree. *Algorithmica*, 16(6), 633--648.

Johnson, D. B. & Metaxas, P. T. (1992). Optimal algorithms for the vertex updating problem of a minimum spanning tree. In *Proceedings of the 6th International Parallel Processing Symposium* (pp. 306--314). Washington, DC, USA: IEEE Computer Society.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Kamada, T. & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31, 7--15.

Kang, U., Papadimitriou, S., Sun, J., & Tong, H. (2011). Centralities in large networks: Algorithms and observations. In *Proceedings of the 2011 SIAM International Conference on Data Mining* (pp. 119--130).

Kapoor, A., (alan Qi, Y., Ahn, H., & Picard, R. W. (2005). Hyperparameter and kernel learning for graph based semi-supervised classification. In *Advances in Neural Information Processing Systems 18* (pp. 627--634).

Karger, D. R., Klein, P. N., & Tarjan, R. E. (1995). A randomized linear-time algorithm to find minimum spanning trees. *Journal of the Association for Computing Machinery*, 42(2), 321--328.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39--43.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data, an Introduction to Cluster Analysis*. Brussels, Belgium: John Wiley & Sons.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, 6(11), 888--893.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). *Network Analysis: Methodological Foundations*, chapter 3, (pp. 16--61). Springer--Verlag.

Krebs, V. (2004). Books about us politics. http://www.orgnet.com/divided.html.

L. R. Ford, J. & Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, 8, 399--404.

Lancichinetti, A. & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80, 056117.

Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78, 046110.

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788--791.

Lee, D. D. & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems* (pp. 556--562).

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., & Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411, 907--908.

Lloyd, A. L. & May, R. M. (2001). How viruses spread among computers and people. *Science*, 292(5520), 1316--1317.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 2), S186--S188.

Lusseau, D. & Newman, M. E. (2004). Identifying the role that animals play in their social networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(Suppl 6), S477--S481.

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. *Behavioral Ecology and Sociobiology*, 54, 396--405.

Ma, X., Gao, L., Yong, X., & Fu, L. (2010). Semi-supervised clustering algorithm for community structure detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 389, 187--197.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. Cam & J. Neyman (Eds.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1 (pp. 281--297). Berkeley: University of California Press.

Marchiori, M. & Latora, V. (2000). Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, 285(3-4), 539--546.

Mewes, H., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., et al. (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic acids research*, 32(Database Issue), D41--D44.

Milgram, S. (1967). The small-world problem. *Psychology Today*, 1, 61--67.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mizruchi, M. S. (1982). *The American Corporate Network, 1904-1974*. Sage.

Molloy, M. & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3), 161--180.

Montoya, J. M. & Solé, R. V. (2002). Small world patterns in food webs. *Journal of Theoretical Biololgy*, 214(3), 405--412.

Moody, J. & White, D. R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, (pp. 103--127).

Moreno, Y., Nekovee, M., & Pacheco, A. F. (2004). Dynamics of rumor spreading in complex networks. *Physical Review E*, 69(6), 066130.

Motta, R., de Andrade Lopes, A., & de Oliveira, M. C. F. (2009). Centrality measures from complex networks in active learning. In *Discovery Science* (pp. 184--196).: Springer.

Motwani, R. & Raghavan, P. (1995). *Randomized Algorithms*. Cambridge University Press.

Newman, M. (2010). *Networks: an introduction*. Oxford University Press.

Newman, M. E., Forrest, S., & Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, 66(3), 035101.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS*, 98(2), 404--409.

Newman, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, 89, 208701.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Phys. Rev. E*, 67, 026126.

Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Review*, 45(2), 167--256.

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. *Phys. Rev. E*, 69, 066133.

Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27, 39--54.

Newman, M. E. J. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74, 036104.

Newman, M. E. J. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577--8582.

Newman, M. E. J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69, 026113.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3), 103--134.

Noh, J. D. & Rieger, H. (2004). Random walks on complex networks. *Phys. Rev. Letters*, 99(11), 118701.

Olsson, F. (2009). *A literature survey of active machine learning in the context of natural language processing*. Technical Report T2009:06, Swedish Institute of Computer Science.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab.

Pastor-Satorras, R. & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Letters*, 86, 3200--3203.

Pastor-Satorras, R. & Vespignani, A. (2002). Immunization of complex networks. *Phys. Rev. E*, 65, 036104.

Pawagi, S. & Ramakrishnan, I. V. (1986). An o(log n) algorithm for parallel update of minimum spanning trees. *Information Processing Letters*, 22(5), 223--229.

Pettie, S. & Ramachandran, V. (2002a). Minimizing randomness in minimum spanning tree, parallel connectivity, and set maxima algorithms. In *ACM-SIAM Symposium on Discrete Algorithms* (pp. 713--722). San Francisco, California: ACM Press.

Pettie, S. & Ramachandran, V. (2002b). A randomized time-work optimal parallel algorithm for finding a minimum spanning forest. *SIAM Journal on Computing*, 31(6), 1879--1895.

Pujol, J. M., Béjar, J., & Delgado, J. (2006). Clustering algorithm for determining community structure in large networks. *Phys. Rev. E*, 74, 016107.

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

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *PNAS*, 101, 2658--2663.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846--850.

Reichardt, J. & Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a potts model. *Phys. Rev. Letters*, 93(21), 218701.

Reichardt, J. & Leone, M. (2008). (un)detectable cluster structure in sparse networks. *Phys. Rev. Letters*, 101, 078701.

Rives, A. W. & Galitski, T. (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, 100(3), 1128--1133.

Rohban, M. H. & Rabiee, H. R. (2012). Supervised neighborhood graph construction for semi-supervised classification. *Pattern Recognition*, 45(4), 1363--1372.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53 -- 65.

Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2323.

Rubinov, M. & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52, 1059--1069.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31, 581--603.

Scott, J. (2000). *Social network analysis: a handbook.* Sage, second edition.

Settles, B. (2009). *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

Silva, T. C. & Zhao, L. (2012). Network-based stochastic semisupervised learning. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(3), 451--466.

Smith, R., Bivens, A., Embrechts, M., Palagiri, C., & Szymanski, B. (2002). Clustering approaches for anomaly-based intrusion detection. In *Intelligent Engineering Systems through Artificial Neural Networks* (pp. 579--584). New York: ASME Press.

Sponrs, O. (2002). Networks analysis, complexity, and brain function. *Complexity*, 8, 56--60.

Steeg, G. V., Moore, C., Galstyan, A., & Allahverdyan, A. (2014). Phase transitions in community detection: A solvable toy model. *EPL (Europhysics Letters)*, 106(4), 48004.

Sun, J., Boyd, S., Xiao, L., & Diaconis, P. (2006). The fastest mixing markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Rev.*, 48(4), 681--699.

Tejedor, V., Bénichou, O., & Voituriez, R. (2009). Global mean first-passage times of random walks on complex networks. *Phys. Rev. E*, 80, 065104.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319--2323.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association.

Valente, T. W. & Foreman, R. K. (1998). Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20, 89--105.

van Mieghem, P. (2010). *Graph Spectra for Complex Networks*. Cambridge University Press.

Vapnik, V. N. (1998). *Statitical learning theory*. New York: Wiley.

Vapnik, V. N. (1999). *The nature of statitical learning theory*. New York: Springer, segunda ed. edition.

Varman, P. & Doshi, K. (1988). An efficient parallel algorithm for updating minimum spanning trees. *Theoretical Computer Science*, 58(1-3), 379--397.

Vendramin, L., Campello, R. J., & Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(4), 209--235.

Vogels, T. P. & Abbott, L. F. (2005). Signal propagation and logic gating in networks of integrate-and-fire neurons. *The Journal of neuroscience*, 25(46), 10786--10795.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 577--584).

Wang, F. & Zhang, C. (2008). Label propagation through linear neighborhoods. In *IEEE Transactions on Knowledge and Data Engineering*, volume 20 (pp. 55--67).

Watts, D. J. (1999). *Small Worlds*. Princeton University Press.

Watts, D. J. (2004). *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of ''small-world'' networks. *Nature*, 393, 440--442.

Weinberger, K. Q. & Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *Int. J. Comput. Vision*, 70(1), 77--90.

West, G. B., Brown, J. H., & Enquist, B. J. (1999). A general model for the structure and allometry of plant vascular systems. *Nature*, 400, 664--667.

Xu, Z., Akella, R., & Zhang, Y. (2007). Incorporating diversity and density in active learning for relevance feedback. *Advances in Information Retrieval*, 4425, 246--257.

Yang, R., Zhou, T., Xie, Y.-B., Lai, Y.-C., & Wang, B.-H. (2008). Optimal contact process on complex networks. *Phys. Rev. E*, 78, 066109.

Ye, Z., Hu, S., & Yu, J. (2008). Adaptive clustering algorithm for community detection in complex networks. *Phys. Rev. E*, 78, 046115.

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

Zhang, P., Moore, C., & Zdeborová, L. (2014). Phase transitions in semisupervised clustering of sparse networks. *Physical Review E*, 90(5), 052802.

Zhang, X. & Lee, W. S. (2006). Hyperparameter learning for graph based semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems 19* (pp. 1585--1592).: MIT Press.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. In *NIPS* (pp. 321--328).: The MIT Press.

Zhou, D., Huang, J., & Schölkopf, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding. In *NIPS*, volume 19 (pp. 1601--1608).: The MIT Press.

Zhou, D. & Schölkopf, B. (2006). Discrete regularization. In O. Chapelle, A. Z. Schölkopf, & A. Zien (Eds.), *Semi-Supervised Learning* (pp. 237--250). The MIT Press.

Zhou, H. (2003a). Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67, 061901.

Zhou, H. (2003b). Network landscape from a brownian particle's perspective. *Phys. Rev. E*, 67, 041908.

Zhou, Z.-H. & Li, M. (2005). Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11), 1529--1541.

Zhu, X. (2005). *Semi-Supervised Learning with Graphs*. Doctoral thesis, Carnegie Mellon University.

Zhu, X. (2008). *Semi-Supervised Learning Literature Survey*. Technical Report 1530, University of Wisconsin-Madison.

Zhu, X. & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation*. Technical report, Carnegie Mellon University.

Zhu, X., Ghahramani, Z., & Lafferty, J. D. (2003). Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML* (pp. 912--919).: AAAI Press.

Zhu, X. & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*, volume 3. Morgan & Claypool.

Zhu, X., Goldberg, A. B., & Khot, T. (2009). Some new directions in graph-based semi-supervised learning. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo* (pp. 1504--1507). Piscataway, NJ, USA: IEEE Press.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans Perform Semi-Supervised Classification Too. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1* (pp. 864--869).: AAAI Press.

Ziv, E., Middendorf, M., & Wiggins, C. H. (2005). Information-theoretic approach to network modularity. *Phys. Rev. E*, 71, 046117.