UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LEILA WEITZEL COELHO DA SILVA

Abordagem baseada na Análise de Redes Sociais para estimativa da reputação de fontes de informação em saúde

Tese apresentada como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação

Prof. Dr. José Palazzo M. de Oliveira
Orientador

Porto Alegre, março de 2013.

Leila Coelho da Silva, Weitzel

Abordagem baseada na Análise de Redes Sociais para estimativa da reputação de fontes de informação em saúde / Leila Weitzel Coelho da Silva. -- 2013.

105 f.:il

Orientador: José Palazzo M. de Oliveira.

Tese (doutorado) -- Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação, Porto Alegre, BR - RS, 2013.

1.Redes Sociais. 2.Reputação. 3.Análise de Redes Sociais. 4. Saúde. I. Palazzo, José M. de Oliveira. II. Título.

AGRADECIMENTOS

Sei que poucos irão ler essa parte da tese, mas para mim é uma das partes mais importante, porque se eu não tivesse o apoio destas pessoas, provavelmente nada do que foi feito aqui teria acontecido.

A primeira pessoa da (longa) lista é a minha querida, linda, maravilhosa, compreensiva filha Rachel Weitzel Martins (agora também sra. Sanches). Desde a sua remota infância vem sofrendo a falta da sua mãe que teimou um dia ser "alguém" nesta vida. Estive ausente em momentos importantes para ela, mas sei que dentro do possível fiz de tudo para suprir suas necessidades em todos os sentidos. Que Deus a tenha em conta e que realize todos seus sonhos como realizou o meu.

Agradeço aos meus pais, que infelizmente não puderam participar desta minha conquista, mas que sei que ficariam orgulhosos de mim.

Duas outras pessoas foram muito importantes na minha vida, a primeira foi a minha melhor, maior e mais querida amiga Helena Folly que Deus levou para outro plano. Esteve presente em todos os momentos importantes, na graduação, no mestrado e estará no doutorado tenho certeza.

A segunda pessoa foi e continua sendo, o melhor companheiro que uma mulher poderia pedir aos céus. Gentil e delicado, apesar dos seus 2 metros de altura, a pessoa mais inteligente, culta e brilhante que já conheci. Está hoje do outro lado do mundo, agradeço a você Serguei Filippov por me deixar fazer parte da sua vida. Agradeço pela força, compreensão e carinho.

Agradeço aos meus queridos orientadores professores José Palazzo e Paulo Quaresma. Ambos são de uma doçura e gentileza sem tamanho. O Palazzo por me pegar pelo meio do caminho e confiar que eu teria competência para levar o doutoramento a cabo. O Paulo pelo seu bom humor e sorriso, no auxilio do dia a dia do doutorado sanduiche em Évora, Portugal, fazendo com que a distância fosse menos dolorosa. Agradeço também a todos os professores da Universidade de Évora que contribuíram para o bom andamento da tese.

Agradeço à Universidade Federal do Pará pelo suporte financeiro durante o doutorado. Agradeço à Capes pela bolsa que me foi concedida. Ao CNPq pelo financiamento de parte da pesquisa, no doutoramento sanduiche em Portugal.

A toda a turma que passou pelo Laboratório 213, e aos de fora também. Muitos almoços e cafezinhos para descontrair a maratona. Citando: Ana Pernas, Giseli Lopes, Daniel Lichtnow, Isabela Gasparini e Lucineia Thom e ao pessoal novato da sala que está entrando no ritmo.

Agradeço aos professores da UFRGS pela a atenção e o carinho dedicado a mim, em especial: Renata Galante, Leandro Wives e Viviane Moreira.

Um agradecimento especial ao Professor Roberto da Silva pela auxilio prestimoso da parte matemática da tese.

Agradeço aos meus "síndicos" Roberto e Eliane que sempre me trataram com carinho e atenção, "quebrando os galhos da vida mundana".

Não poderia deixar de agradecer aos meus queridos alunos e pupilos da UFPA sempre mandando bons fluidos para mim entre eles se destacam o Fernando G. Rodrigues e o Jhoseph Araújo com a ajuda providencial na programação.

E a todos que direta ou indiretamente contribuíram para o bom andamento desta jornada.

Fonte: The New Yorker, 1993

"O objetivo da vida é uma vida de objetivo".

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

HON	Health on Net
ARS	Análise de Redes Sociais
SNA	Social Network Analysis
AVR	Ambientes Virtuais de Relacionamento
HITS	Hypertext Induced Topic Search
WWW	World Wide Web
BBS	Bulletin Board Systems
IRC	Internet Relay Chat
ICQ	I Seek You
GPS	Global Positioning System
ESPN	Entertainment and Sports Programming Network
P2P	Peer-to-Peer, redes Ponto a Ponto
NPR news	National Public Radio
CNN	Cable News Network
AIDS	Adquired Imunity Deficient Syndrom
MAP	Mean Average Precision
AP	Average Precision
ORA	Organization Risk Analyzer
RT-NET	Rede Retweet com metodologia de pesos proposta – equação 4.1
RT-BASE	Rede Retweet com pesos binários {0;1} 0 arco ausente, 1 arco presente
p@n	Precisão em n

LISTA DE SÍMBOLOS

G	Grafo (ou rede)
G = {N, E}	Grafo não direcionado
v, w, u	Representam os vértices do grafo G
e	Representam os arcos do grafo G
N	Conjunto de vértices ou nós de G
E	Conjunto de arcos, arestas ou ligações de G
E(G)	Conjunto das arestas de G
N(G)	Conjunto de vértices de G
n(G)	Número de vértices de G
m(G)	Número de arestas de G
Adj|N|	Lista de adjacência de G
B = (b_{ij})	Matriz de Incidência de G
A	Matriz de Adjacência de G
a_{ij}	Elementos da matriz de adjacência A
T	Transitividade
n_{Δ}	Quantidade de cliques
n_{Λ}	Quantidade de tríades presentes na rede.
l	Caminho do grafo ou rede G
D	Diâmetro do Grafo ou rede G
E	Quantidade de conexões: Conjunto de arcos, arestas ou ligações de G
de	Densidade da rede (ou grafo G)
C	Coeficiente de agrupamento ou de aglomeração da rede (ou grafo G)
k	Centralidade do Grau de um vértice *v*

$<k>$	Grau médio de um Grafo G
k^{in} , Dc_{in}	Grau de entrada de um vértice *v*
k^{out} , Dc_{out}	Grau de saída de um vértice *v*
P_k	Probabilidade de distribuição das conexões
Cc	Centralidade de Proximidade
Bc	Centralidade de Intermediação
d, g_{ij}	Distância geodésica ou caminho mais curto
Ec	Centralidade de Autovetor
$Pg(\lambda)$	Polinômio característico da matriz de adjacência A do grafo G
λ	Autovalores da matriz de adjacência A do grafo G
γ	Coeficiente de Escala ou de Potência
μ	Média de uma amostra
σ	Desvio padrão de uma amostra
$G_{RT}^{\rightarrow} = (N, E^{\rightarrow}, \mathcal{W})$	Rede de Re*tweet*
$\sum \mathrm{RT}_{v_j}$	Número total de *retweet* de um usuário alvo v_j
$\mathrm{RT}_{\mathrm{total}}$	Quantidade total de *retweets* da amostra
α	Parâmetro mitigador do efeito celebridade
$w(e_k)$	Peso dos arcos da rede ***RT-net*** – Rede de *Retweets*
$\mathcal{R}\mathrm{a}\mathcal{R}_{v_j}$	Rank Reputation
M	Conjunto de métricas
m_{i}	Elemento do conjunto M de métricas
W	Conjunto de pesos associados às métricas do conjunto M
w_i	Elemento do conjunto de pesos associados às métricas m_i
P@n	Precision at *n* – Precisão em *n*
$\mathrm{L_p}$	Lista ordenada em ordem descendente dos valores de $\mathcal{R}\mathrm{a}\mathcal{R}_{v_j}$
$\mathrm{T_k}$	As linhas desta tabela armazenam os valores calculados de $\mathcal{R}\mathrm{a}\mathcal{R}_{v_j}$

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

A Internet tem sido uma importante fonte para as pessoas que buscam informações de saúde. Isto é particularmente problemático na perspectiva da Web 2.0. A Web 2.0 é a segunda geração da *World Wide Web*, onde os usuários interagem e colaboram uns com os outros como criadores de conteúdo.

A falta de qualidade das informações médicas na Web 2.0 tem suscitado preocupações com os impactos prejudiciais que podem acarretar. São muitos os aspectos relacionados à qualidade da informação que devem ser investigados, como por exemplo, existe alguma evidência de que o autor tem alguma autoridade no domínio da saúde? Há indícios de que os autores são tendenciosos? Como saber se a fonte de informação tem reputação, como separar as fontes de boa qualidade das outras? Esses questionamentos se tornam mais evidentes quando se faz buscas no *Twitter*. O usuário precisa por si só selecionar o conteúdo que acredita que tenha qualidade entre as centenas de resultados.

Nesse contexto, o principal objetivo deste trabalho é propor e avaliar uma abordagem que permita estimar a reputação de fontes de informação no domínio da saúde. Acredita-se que discussões sobre reputação só fazem sentido quando possuem um propósito e estão inseridas em um contexto. Sendo assim, considera-se que reputação é um atributo que um usuário se apropria quando a informação que ele divulga é crível e digna de confiança. As contribuições desta tese incluem uma nova metodologia para estimar a reputação e uma estrutura topológica de rede baseada no grau de interação entre atores sociais.

O estudo permitiu compreender como as métricas afetam o ordenamento da reputação. Escolher a métrica mais apropriada depende basicamente daquilo que se quer representar. No nosso caso, o *Pagerank* funcionou como um "contador de arcos" representando apenas uma medida de popularidade daquele nó. Verificou-se que popularidade (ou uma posição de destaque na rede) não necessariamente se traduz em reputação no domínio médico.

Os resultados obtidos evidenciaram que a metodologia de ordenamento e a topologia da rede obtiveram sucesso em estimar a reputação. Além disso, foi verificado que o ambiente *Twitter* desempenha um papel importante na transmissão da informação e a "cultura" de encaminhar uma mensagem permitiu inferir processos de credibilidade e consequentemente a reputação.

Palavras-Chave: Análise de Redes Sociais, Redes Sociais, reputação, *Twitter*.

ABSTRACT

The Internet is an important source for people who are seeking healthcare information. This is particularly problematic in era of Web 2.0. The Web 2.0 is a second generation of World Wide Web, where users interact and collaborate with each other as creators of content.

Many concerns have arisen about the poor quality of health-care information on the Web 2.0, and the possibility that it leads to detrimental effects. There are many issues related to information quality that users continuously have to ask, for example, is there any evidence that the author has some authority in health domain? Are there clues that the authors are biased? How shall we know what our sources are worth, how shall we be able to separate the bad sources from the good ones? These questions become more obvious when searching for content in *Twitter*. The user then needs to manually pick out high quality content among potentially thousands of results.

In this context, the main goal of this work is to propose an approach to infer the reputation of source information in the medical domain. We take into account that, discussion of reputation is usually not meaningful without a specific purpose and context. Thus, reputation is an attribute that a user comprises, and the information disseminated by him is credible and worthy of belief. Our contributions were to provide a new methodology to Rank Reputation and a new network topological structure based on weighted social interaction.

The study gives us a clear understanding of how measures can affect the reputation rank. Choosing the most appropriate measure depends on what we want to represent. In our case, the PageRank operates look alike "edges counts" as the "popularity" measures. We noticed that popularity (or key position in a graph) does not necessarily refer to reputation in medical domain.

The results shown that our rank methodology and the network topology have succeeded in achieving user reputation. Additionally, we verified that in *Twitter* community, trust plays an important role in spreading information; the culture of "re*tweet*ing" allowed us to infer trust and consequently reputation.

Keywords: Social Network Analysis, Social Network, reputation, *Twitter*.

1 INTRODUÇÃO

O capítulo introdutório tem como objetivo contextualizar esta pesquisa. Nele são discutidos o processo de formulação do tema, a identificação do problema, a natureza da pesquisa e o seu objetivo, ressaltando a sua importância no contexto atual.

Para tanto, foram introduzidas breves definições de termos relacionados que serão discutidos em maior ou menor profundidade nos capítulos posteriores em função dos recortes que são dados ao tema da pesquisa.

1.1 Identificação do problema

Os avanços tecnológicos, a inclusão digital e o barateamento das tecnologias vêm modificando as formas de comunicação e interação entre os indivíduos. Nas últimas duas décadas, a Internet tornou-se tão integrada em nossas vidas como uma importante, se não indispensável, ferramenta de informação, comunicação e interação. É indiscutível a vertiginosa evolução da Internet como fonte de pesquisa e informação. Informações que antes ficam limitadas a formatos impressos, como livros e revistas, hoje podem ser obtidas na Web. Nesse ambiente, e devido à sua abertura, qualquer pessoa pode publicar qualquer tipo de informação. Não existem avaliações prévias do que é disponibilizado, e o acúmulo de informações sem relevância aponta para a necessidade de metodologias que permitam a filtragem e em última análise, uma estimativa da qualidade dessa informação.

Da mesma forma, este fato vem ocorrendo com as informações em saúde. Informações estas que ficavam restritas ao ambiente do consultório médico ou em formatos impressos encontram-se agora dispersas por toda Web. Podendo ser facilmente acessadas em qualquer lugar ou tempo com o uso de dispositivos de comunicação tais como: *smartphones* e *tablets*.

A segunda geração ou uma forma aprimorada da Web é conhecida como Web 2.0[1]. Esse ambiente enfatiza colaboração e partilha de conhecimentos e conteúdos entre os usuários. A ideia da Web 2,0 é tornar o ambiente *on-line* mais dinâmico e fazer com que os usuários colaborem na organização de conteúdo. O contexto é particularmente

[1] O termo Web 2.0 foi cunhado em meados de 2004 por Dale Dougherty da empresa americana O'Reilly Media. A ideia por traz do conceito é a de aprimorar a troca de informações e colaboração dos usuários como meio de criar, agregar, compartilhar, colaborar e publicar a informação digital em qualquer formato (p.ex. música, texto, imagem, vídeo, áudio) (O'REILLY, 2005). O'Reilly Media é uma editora de livros e revistas, e promotora de conferências e serviços on-line: http://radar.oreilly.com/2006/05/controversy-about-our-web-20-s.html.

importante do ponto de vista da criação e distribuição de informações em saúde nesse ambiente.

Na Web 2.0 qualquer um pode publicar informações independentemente da sua formação, qualificação ou intenção (PURCELL; BRENNER; RAINIE, 2012). Há uma infinidade de textos publicados por leigos contendo dados imprecisos, errôneos, que não têm caráter científico. Os textos podem ser inseridos entre as reportagens de jornais e revistas para atrair a atenção do leitor. Não existem padrões universais de publicação nesse ambiente, e qualquer conteúdo publicado pode ser facilmente alterado e deturpado (O'GRADY, 2006; RIEH; DANIELSON, 2007).

A Web 2.0 apresenta assim, um potencial perigoso em difundir informações de qualidade duvidosa. Informações de qualidade duvidosa interferem negativamente na relação do paciente com sua doença, e/ou com seu médico. Se por um lado, ao ter acesso à informação, o usuário pode ser estimulado ao autocuidado, por outro, pode induzi-lo ao erro, ou pior ao óbito (EYSENBACH, 2002; POWELL; DARVELL; GRAY, 2003; WANG; LIU, 2007).

A literatura trata de critérios para avaliar sites na Web, onde é ressaltada a importância de se avaliar principalmente os critérios: autoridade, atualidade das informações, precisão entre outros. Já no domínio médico agências nacionais e internacionais de saúde vêm desenvolvendo instrumentos reguladores para a certificação da qualidade de sites nessa área. Constituídas por profissionais da saúde e grupos multidisciplinares, essas entidades, desde 1996, elaboram e definem ações que se concretizaram em diretrizes respeitadas e geralmente seguidas pela comunidade de médica. As Agências: *National Institutes of Health*[2], *Health Summit Working Group*[3] e *a Health On the Net* (HON) *Foundation*[4], se destacam neste sentido.

Os instrumentos reguladores de certificação da qualidade se baseiam em critérios técnicos e éticos de conduta, listados em categorias e subcategorias. Dentre uma variedade deles se destacam:

1. Credibilidade (fonte, atualização periódica, pertinência/utilidade, e processo de revisão editorial para a informação);
2. Privacidade (confidencialidade dos dados);
3. Apoio financeiro (deve ser identificado claramente, incluindo a identidade das organizações comerciais e não comerciais que tenham contribuído com financiamento, serviços ou materiais);
4. Política de publicidade (propaganda e *marketing* da página);
5. Projeto visual (acessibilidade, organização navegabilidade, e capacidade de pesquisa interna);
6. Interatividade (inclui mecanismos de *feedback* e meios para o intercâmbio de informações entre os usuários- fórum de discussão);
7. Atualidade (indicar últimas atualizações), entre outros. Aos sites que seguem esses critérios técnicos e éticos é concedido um selo de certificação da

[2] http://www.nlm.nih.gov/medlineplus/webeval/webeval.html
[3] http://www.ahrq.gov/
[4] http://www.hon.ch/

qualidade, por exemplo, o HONcode da agência *Health On the Net* (HON) *Foundation*[5].

Contudo, quando não existe o selo de certificação no *site*, essas Agências aconselham que o usuário deva avaliar, por conta própria, a qualidade (baseando-se nos fatores técnicos e éticos de conduta) à medida que faz a sua busca. De acordo com as agências, o usuário que busca a informação deve ter um comportamento reflexivo na construção de um modelo que avalie a reputação tanto da fonte quanto da própria informação. Do ponto de vista de um usuário leigo e iniciante é particularmente necessário ter-se cautela ao analisar a qualidade desses sites, especialmente em um ambiente da Web 2.0. Desta forma, a tarefa de avaliação acaba por tornar-se um processo complexo.

1.2 Objetivo

Assim, em face do exposto acima, esta pesquisa tem como objetivo propor e avaliar uma abordagem que permita estimar a reputação de fontes de informação no domínio da saúde. Para o estudo de caso, foi selecionado um ambiente da Web 2.0 conhecido como *Twitter*. A abordagem proposta inclui uma metodologia para estimar a reputação, um modelo de rede e uma metodologia de ponderação de arcos. Para viabilizar esta abordagem, propõe-se o uso da Análise de Redes Sociais (ARS).

Nesta pesquisa, a reputação é definida como uma avaliação social feita sobre uma fonte de informação que pode ser uma pessoa, um grupo de pessoas ou uma organização (podendo ser chamada também de entidade social). Assim sendo, fonte de informação que exibe reputação é aquela que participa tanto da criação quanto da distribuição de informação considerada relevante (que apresenta credibilidade e confiança) no domínio da saúde.

Acredita-se que discussões sobre reputação só fazem sentido quando inseridas em um contexto (neste caso o contexto da saúde) e com um propósito definido (avaliar fontes de informação). Considera-se que reputação é um atributo que um usuário se apropria quando a informação que ele divulga é crível e digna de confiança.

Foge do escopo desta pesquisa avaliar a qualidade do conteúdo da informação ou qualidade de sites na área de saúde. Tanto do ponto de vista dos instrumentos reguladores citados anteriormente, quanto do ponto de vista do conteúdo médico que está sendo publicado. Essas avaliações têm sido extensivamente pesquisadas e estudadas no âmbito das agências e conselhos na área médica.

1.3 Organização da Tese

A Tese estrutura-se em mais 5 Capítulos, além desta Introdução e as Referências Bibliográficas.

No Capítulo 2 tem-se a fundamentação teórica que norteou a pesquisa. Discute-se os principais conceitos relacionados às Redes Sociais e Analise de Redes Sociais. Aborda-se a representação das redes no âmbito da Teoria de Grafos, as propriedades gerais e estruturais.

[5] http://www.hon.ch/

Ainda neste capitulo são apresentadas as principais medidas de posição (ou de centralidade) que são utilizadas na literatura, além dos Modelos Teóricos de redes complexas. Ao final são apresentados os ambientes virtuais de relacionamento.

No Capitulo 3 são apresentados os trabalhos correlatos que foram utilizados como referência nesta tese. Os trabalhos, aqui apresentados, têm como principal objetivo evidenciar a importância de um ator em função da sua posição em uma estrutura de rede. Quatro deles tratam de averiguar a importância em redes virtuais de relacionamento, neste caso em particular o *Twitter*, um utiliza uma rede do mundo real e o último utiliza uma rede de confiança (baseada na opinião dos usuários sobre determinado assunto ou produto).

No Capítulo 4 são apresentados: o ambiente da pesquisa, neste caso o *Twitter* e suas principais características; o modelo de rede proposto e as propriedades verificadas; a metodologia para estimativa da reputação e a avaliação do desempenho desta metodologia.

O Capitulo 5 aborda como a coleta de dados ocorreu e os resultados verificados.

Ao final é feita uma discussão sobre esses resultados. Além das conclusões, contribuições e trabalhos futuros. Por fim apresenta-se o Referencial Bibliográfico e os Anexos.

1.4 Principais contribuições

As contribuições trazidas por esta tese na área de Sistemas de Informação são:

- ✓ **Análise de Redes Sociais para estimar reputação**: Apesar de o tema ter sido utilizado em outros problemas transversais a este, até onde vai nosso conhecimento, é a primeira vez que é utilizado para estimar a reputação no domínio da saúde.
- ✓ **Algoritmo da estimativa da reputação**: O algoritmo permitiu encontrar o melhor desempenho da metodologia; além de verificar quais são as métricas que melhor representam a reputação de entidades sociais no domínio de saúde;
- ✓ **Modelagem da Rede Social:** até onde vai nosso conhecimento, é a primeira vez que uma Rede (grafo) de *Retweets* é modelada com o propósito de estimar a reputação de uma entidade social no domínio de saúde.
- ✓ **Metodologia de ponderação dos arcos**: a metodologia de ponderação dos pesos dos arcos é baseada no grau de interação entre os atores (ponderação dos *retweets*). A metodologia proposta é empregada pela primeira vez em uma rede de relacionamentos.
- ✓ **Característica topológica da rede de *Retweets***: os experimentos e cálculos evidenciaram que a rede de *Retweets* apresenta topologia de uma rede Livre de Escala com comportamento fractal.

2 FUNDAMENTAÇÃO

Neste capítulo apresentam-se os aspectos históricos da abordagem de Analise de Redes Sociais e os conceitos básicos relacionados ao tema. Apresentam-se as propriedades gerais das Redes Complexas, os modelos clássicos dessas redes (Aleatórias, Mundo Pequeno e Livres de Escala), propriedades topológicas e estruturais.

2.1 Introdução

Durante as últimas duas décadas os sistemas complexos têm despontado como um tema interdisciplinar que conecta áreas tradicionais como física, química, biologia, ecologia e até mesmo ciências sociais tais como economia e sociologia. Sistemas podem ser considerados complexos quando mostram emergência de formas coerentes longe da aleatoriedade. Uma definição mais compreensível é a de que sistemas complexos são compostos de um grande número de componentes ou "agentes", interagindo de tal modo que o comportamento coletivo não é uma simples combinação dos comportamentos individuais. A principal propriedade dos sistemas complexos reside na possibilidade de ocorrer uma ação coletiva e coerente em grande escala, produzindo estruturas (padrões) interessantes (ALBERT; BARABÁSI, 2002; NEWMAN, 2003).

Os grafos são utilizados para representar redes de pequeno porte, consideradas não complexas. Redes Complexas (RC) envolvem não só o formalismo matemático da Teoria dos Grafos, mas também a análise baseada em ferramentas da Física Estatística.

Há certa dificuldade em se encontrar na literatura uma conceituação clara e universalmente aceita aplicável às RC. Uma definição geralmente aceita é a que define RC como sendo um grafo que possui uma topologia não trivial. Em outras palavras, não é um grafo aleatório e nem tão pouco apresenta padrões de conexão entre seus elementos; o grau de conectividade (número de arestas por vértice) não é regular e nem randômico. Propriedades essenciais das RC estão relacionadas na própria topologia não trivial, e na sua descrição física e geométrica (ALBERT; BARABÁSI, 2002; DOROGOVTSEV; MENDES, 2003; NEWMAN, 2003; NEWMAN; GIRVAN, 2004).

As redes complexas devido a sua flexibilidade em representar sistemas discretos tanto do mundo real quanto dos sistemas projetados pelo homem têm sido amplamente utilizadas em diferentes domínios e aplicações. Pode-se classificar as redes em função de características comuns em:

1) Redes Sociais são formadas por pessoas ou grupo de pessoas conectadas por algum tipo de ligação (relacionamento) de amizade, comercial colaboração científica, participação em filmes etc (ALBERT; BARABÁSI, 2002; GLEISER; DANON, 2003; MATIA *et al.*, 2005).

2) Redes de Informação são redes cujos vértices representam informação armazenada, por exemplo a WWW, Wordnet, etc (BARABÁSI; ALBERT; JEONG, 2000; HUBERMAN, 2002).

3) Redes Tecnológicas são redes artificiais construídas pelo homem, por exemplo, redes de distribuição de energia, de transporte, etc (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999).

4) Rede Biológicas são redes do mundo real como por exemplo rede do sistema biológico de proteínas, bactérias, etc (NUNES AMARAL; MEYER, 1999; CAMACHO; STOUFFER; AMARAL, 2007).

Esta pesquisa trata apenas de Ambientes Virtuais de Relacionamento (AVR), também conhecidos como Redes Socais.

2.2 Redes Sociais

As Redes sociais são descritas como um conjunto de vértices (nodos) que são ligados por arestas (conexões, ligações ou links) devido a algum tipo de interação (NEWMAN, 2003b). No caso de uma rede social, o vértice pode ser chamado de ator ou entidade social, que é o elemento básico de uma rede, entende-se como ator qualquer entidade existente no contexto social.

1) **Ator**: O ator é a entidade social que participa de determinada rede e é capaz de agir e formar ligações com outros atores. Pode ser um indivíduo, uma corporação ou um coletivo social, ou qualquer objeto. Quando todos os atores de uma rede são do mesmo tipo, chamamos esta rede de monomodal. Mas, há casos em que tem-se diferentes atores e a rede é chamada multimodal.

2) **Ligação**: Uma conexão entre dois atores em uma rede social é chamada de ligação, relação, aresta, laços, etc. É definida por algum tipo de relação entre esses atores, conforme o tipo de sociedade. Entre empresas, a ligação pode ser um contrato comercial de fornecimento. Entre pessoas, numa empresa, pode ser o laço hierárquico, se considerarmos o organograma, ou pode ser de relações de amizade.

A Figura 2-1 ilustra uma rede que foi modelada como um grafo não direcionado de usuários da rede Facebook, os nós são os usuários e os arcos é a relação de amizade. No Facebook as relações são, por padrão, todas recíprocas e mandatórias, ou seja, se A é amigo de B então B é amigo de A, mas cabe a A aceitar a amizade.

A Figura 2-2 ilustra uma rede modelada como um grafo direcionado da rede de seguidores do *Twitter*. Na rede *Twitter* as relações não são mandatórias. A segue B sem que seja necessária a aprovação por padrão. No Capitulo 4 seção 4.1 tem-se uma descrição mais detalhada do ambiente *Twitter*.

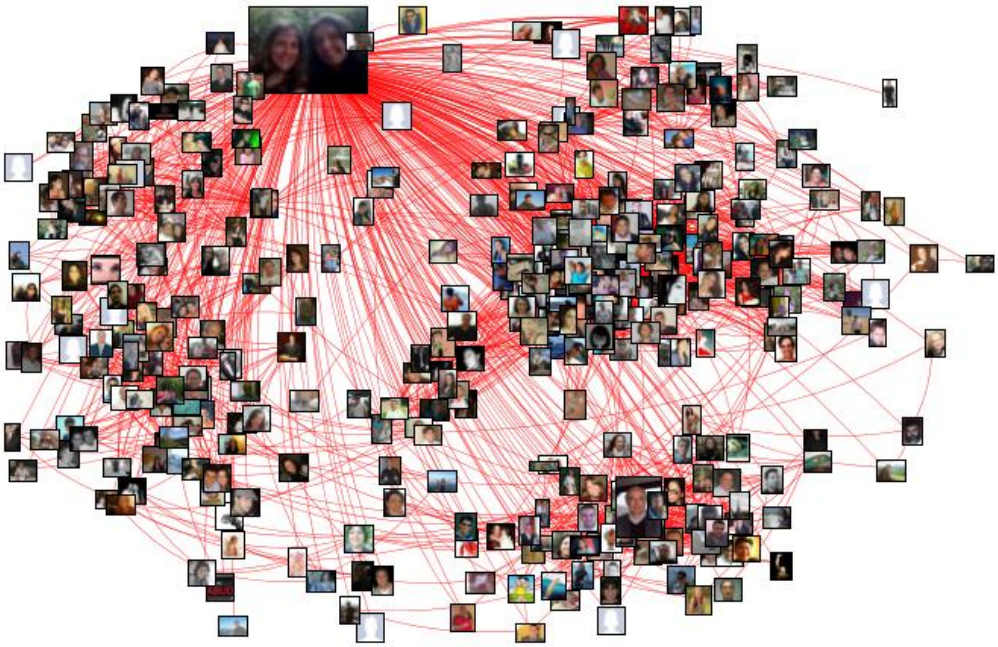


Figura 2-1: Redes de relacionamentos de amizade no Facebook

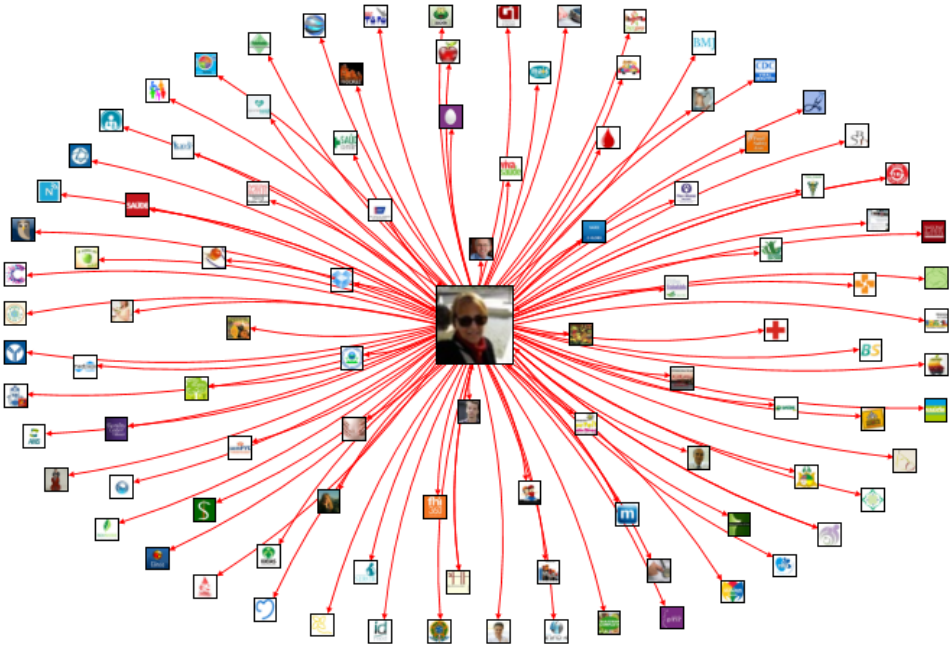


Figura 2-2: Rede de seguidores do *Twitter*

2.3 Redes Sociais na internet ou Ambientes Virtuais de Relacionamento

As Redes Sociais surgem da necessidade do ser humano em compartilhar com o outro, criar laços sociais que são norteados por afinidades entre eles (MORENO, 1934).

Ao contrário do que se possa pensar, falar sobre Redes Sociais não necessariamente significa falar de Internet. Hoje em dia a Internet está integrada no cotidiano dos indivíduos modificando as formas de comunicação e interação, contexto propício para o aparecimento dos AVR. Um AVR é um serviço, plataforma ou site que facilita a construção de Redes Sociais ou relações sociais entre as pessoas distantes geograficamente (na maioria das vezes, mas não necessariamente).

Considera-se como origem destes ambientes o *Bulletin Board Systems* (BBS) criado em 1978. Esse ambiente permitia a troca de dados e mensagens, participação em fóruns, leitura de notícias e utilização de jogos. Em meados dos anos 90 surgem o IRC (Internet Relay Chat) e ICQ (I Seek You) que revolucionaram as trocas instantâneas de mensagens criando grandes comunidades de usuários. Em 1997 é lançada a *Sixdegree* que foi a primeira rede social generalista permitindo a criação de perfis e a manutenção dos contatos, mas foi extinta em 2001.

Em 2002 é lançado o *Friendster* marco importante na história das redes sociais, devido à sua grande audiência. Logo em seguida, tem-se o aparecimento do *MySpace*, o *Flick* que era dedicado à fotografia e o *Orkut*. Em 2005 surge o *Youtube*, dedicado a vídeos; no mesmo ano surge a Rede Social que viria a ser a mais popular de todos os tempos: o *Facebook*.

O ano de 2006 marcou o aparecimento de outra grande rede social, o *Twitter*. Em 2011 surgem dois nomes também sonantes na história das redes sociais, o *Google Plus*, a rede social da *Google*, e o *Pinterest*, caracterizado pela partilha de imagens (Figura 2-3).

Figura 2-3: Ícones de algumas Redes Sociais

Algumas redes foram desenvolvidas especificamente para dispositivos móveis (Celulares e Tablets), como por exemplo:

a. *Foursquare* (rede social e microblogging que permite ao utilizador indicar onde se encontra, e procurar por contactos que estejam próximo desse local);

b. *Instagram* (permite aos usuários tirar uma foto e compartilhá-la em uma variedade de redes sociais, incluindo o próprio *Instagram*);

c. *Waze* (possibilita manter o usuário informado sobre as rodovias, trânsito e até mesmo a localização de radares com usuários cadastrados, sendo também um GPS - Sistema de Posicionamento Global).

Um dos aspectos mais interessantes das Redes Sociais online é a vasta quantidade de dados que produzem. Mais especificamente, nesses ambientes são criados dados relacionais: informações sobre quem conhece quem ou é amigo de quem, quem fala com quem, quem anda nos mesmos lugares, e que gosta das mesmas coisas. Esses dados podem ser utilizados para compreender melhor os indivíduos, organizações e comunidades em diferentes estudos.

2.4 Representação e caracterização estrutural das Redes Sociais

Nesta seção é apresentado um breve histórico da Teoria de Grafos, suas formas de representação. Para aprofundamento sobre a Teoria de Grafos recomenda-se a leitura de Bollobás (1998) e Diestel (2000).

2.4.1 Teoria de Grafos - Breve histórico

Em vários campos de aplicação, diferentes instrumentos ou objetos interagem ou estão correlacionados, formando um sistema. Uma maneira eficiente de estudar um sistema e entender a sua dinâmica é modelá-lo. A modelagem consiste em representar de uma forma simplificada e abstrata o sistema, levando-nos à compreensão mais fácil de suas inter-relações. Um método de modelagem popular é representar os objetos e seus relacionamentos por um grafo, permitindo utilizar um conjunto de métodos genéricos para o seu estudo. Em um grafo, os objetos são representados por nós e relações entre elas são representadas por ligações. Eles são utilizados como uma ferramenta de modelagem em muitos domínios diferentes (LESKOVEC; KLEINBERG; FALOUTSOS, 2005).

Grafos têm sua origem na matemática, com Euller. O grafo das pontes de Konigsberg (Figura 2-4) foi o marco inicial para a teoria de grafos. Euler provou de modo simples que não é possível fazer uma caminhada pela cidade e passar por cada uma das sete pontes uma única vez. A partir do grafo das Pontes de Königsberg (Figura 2-5), ele provou que não havia uma rota que cruzasse cada ponte apenas uma vez. Era necessário inserir pelo menos mais uma ponte para tornar possível esta solução. Todavia não foi a prova que entrou para a história, mas sim os passos que ele usou para resolver o problema.

A grande ideia foi, sem dúvida alguma, tratar o problema das pontes de Konigsberg como sendo um grafo, uma coleção de vértices conectados por ligações. O aspecto mais importante da prova de Euler é que a existência do caminho é uma propriedade intrínseca do grafo, ou seja, ou o caminho existe ou não existe.

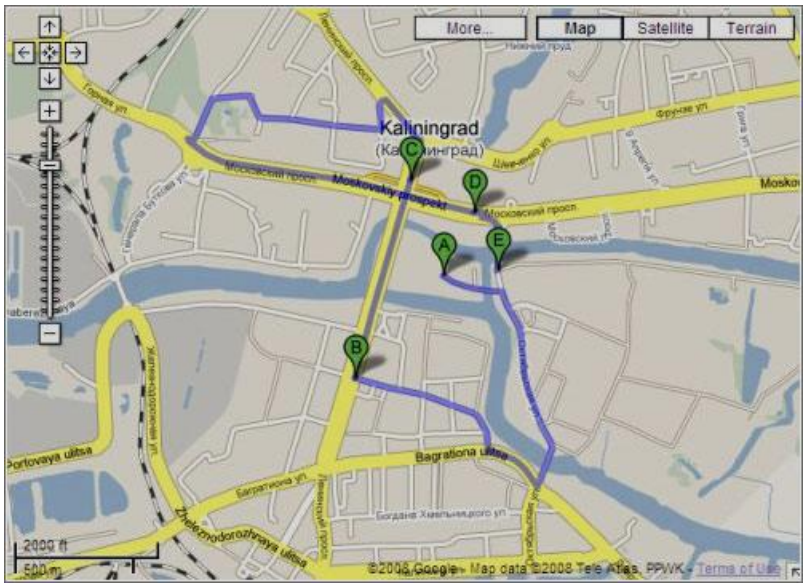


Figura 2-4: As Sete pontes de Konigsberg

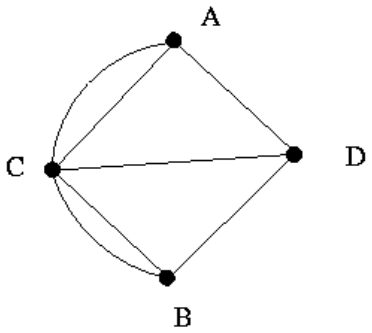


Figura 2-5: Representação gráfica do problema das sete pontes de Konigsberg

A teoria dos grafos estuda objetos combinatórios, os grafos servem de modelo para muitos problemas em vários ramos da matemática, da computação, da engenharia entre outros.

Para qualquer conjunto **G,** denotaremos por $\mathbf{G^{(2)}}$ o conjunto de todos os pares não ordenados de elementos de **G**. Se **G** tem ***n*** elementos então $\mathbf{G^{(2)}}$ tem $\binom{n}{2} := \frac{n(n-1)}{2}$ elementos. Os elementos de $\mathbf{G^{(2)}}$ serão identificados com os subconjuntos de **V** que têm cardinalidade 2. Assim, cada elemento de $\mathbf{G^{(2)}}$ terá a forma ***{v,w}***, sendo ***v*** e ***w*** dois elementos distintos de **G**. Um grafo **G** é um par (**N** , **E**) em que **N** é um conjunto arbitrário e **E** é um subconjunto de $\mathbf{N^{(2)}}$. Os elementos de **N** são chamados vértices (*nodes*) e os de **E** (*edges*) são chamados arestas. Uma aresta como ***{v,w}*** será denotada simplesmente por ***vw*** ou por ***wv***. Diz-se que a aresta ***vw*** incide em ***v*** e em ***w*** e que ***v*** e ***w*** são as pontas da aresta. Se ***vw*** é uma aresta, diz-se que os vértices ***v*** e ***w*** são vizinhos ou adjacentes (BOLLOBAS, 1998).

Assim, se **G** for um grafo, conjunto dos seus vértices será denotado por **N(G)** e o conjunto das suas arestas por **E(G)**. O número de vértices de G é denotado por **n(G)** e o número de arestas por **m(G)**. Na Figura 2-6, tem-se o grafo dos estados do Brasil, onde

é definido que cada vértice é um dos estados da República Federativa do Brasil. Dois estados são ditos adjacentes se possuem fronteira comum.

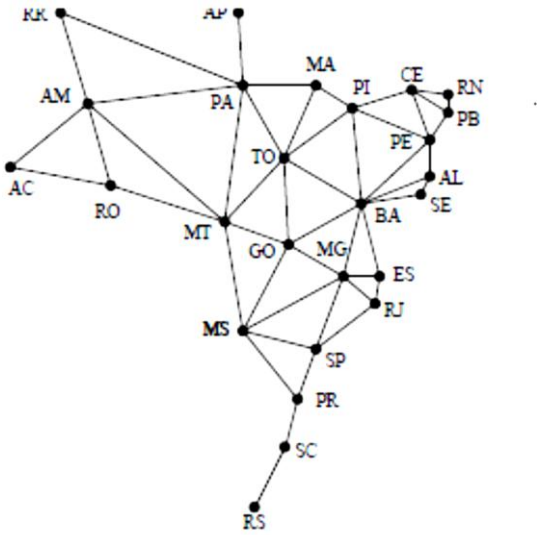


Figura 2-6: Adjacência entre os Estados do Brasil

A representação de um grafo pode ser feita por:

1) Lista de Adjacência: Consiste de um vetor de Adj|N| listas, um para cada vértice de N. Para cada *v* em V, Adj[v] consiste de todos os vértices de G adjacentes a *v*. Vértices armazenados de forma arbitrária na lista. Também pode ser utilizada no caso de grafos dirigidos (Figura 2-7).

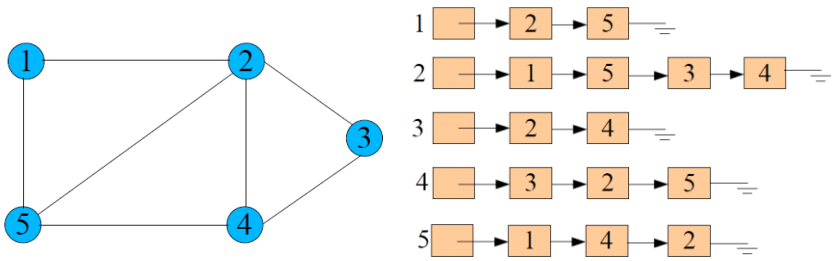
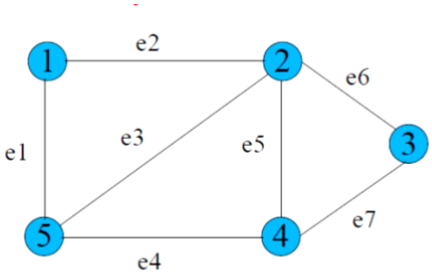


Figura 2-7: Lista de adjacência

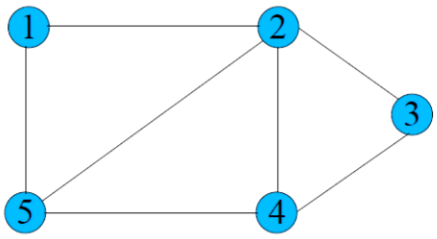
2) Matriz de Incidência: seja a matriz B = (b_{ij}), de ordem |N| x |E|, onde b_{ij} =1 se o vértice v_i e a aresta e_j forem incidentes, ou b_{ij} = 0 caso contrário (Figura 2-8).



	1	2	3	4	5	6	7
1	1	1	0	0	0	0	0
2	0	1	1	0	1	1	0
3	0	0	0	0	0	1	1
4	0	0	0	1	1	0	1
5	1	0	1	1	0	0	0

Figura 2-8: Matriz de Incidência

3) Matriz de Adjacência: o vértice ***i*** é adjacente a ***j*** se ***i*** está conectado a ***j***. Em termos matricial, tem-se que se posição **A (i,j)** é igual a 1 então existe uma aresta direcionada do vértice ***i*** ao ***j***. Caso contrário, **A(i,j)** é igual a zero. Em uma rede não direcionada, uma conexão entre ***i*** e ***j*** implica **a(i,j) = a(j,i) = 1**, ou seja, a matriz **A** é necessariamente simétrica (Figura 2-9).



	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

Figura 2-9: Matriz de Adjacência

Grafos não direcionados são representados por um par ordenado $\boldsymbol{G} = (\boldsymbol{N}, \boldsymbol{E})$, formado por um conjunto $\boldsymbol{N} \equiv \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_N}\}$ de vértices (nós) e por um conjunto de $\boldsymbol{E} \equiv \{\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_E}\}$ de arcos (ou laços). Grafos direcionados são representados por $\boldsymbol{G}^{\rightarrow} = (\boldsymbol{N}, \boldsymbol{E}^{\rightarrow})$ de pares ordenados. Tem-se também grafos ponderado direcionado (ou não direcionados) $\boldsymbol{G}^{\rightarrow} = (\boldsymbol{N}, \boldsymbol{E}^{\rightarrow}, \boldsymbol{\mathcal{W}})$, onde os vértices são $\boldsymbol{N} \equiv \{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots\}$, os arcos são $\boldsymbol{E}^{\rightarrow} \equiv \{\boldsymbol{e_1}, \boldsymbol{e_2}, \dots\}$ **e** $\boldsymbol{\mathcal{W}}$ é o conjunto de pesos associado aos arcos Figura 2-10.

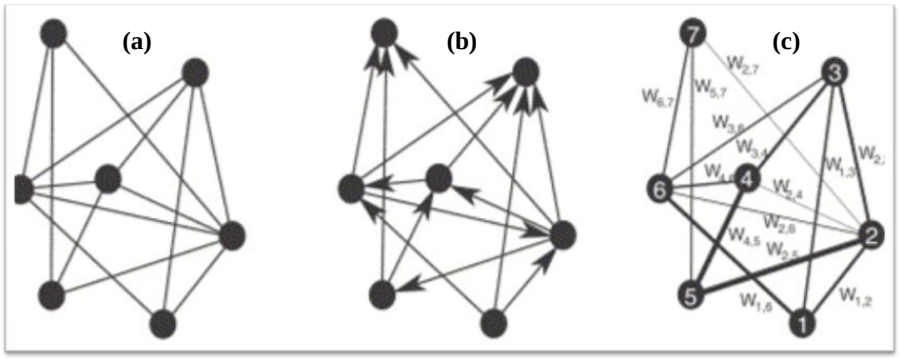


Figura 2-10: (a) grafo não direcionado, (b) grado direcionado, (c) grafo não direcionado ponderado

2.5 Propriedades Gerais das Redes

A seguir são apresentadas apenas as propriedades gerais mais relevantes nesta pesquisa, para estudos mais abrangentes cita-se (ERDÕS; RÉNYI, 1960; MCPHERSON; SMITH-LOVIN; COOK, 2001; ALBERT; BARABÁSI, 2002; NEWMAN, 2003; RAVASZ; BARABÁSI, 2003) esses autores apresentam um referencial teórico amplo sobre o tema.

2.5.1 Homofilia e Heterofilia

Homofilia (antônimo de Heterofilia) é a tendência dos atores se conectarem preferencialmente de acordo com as semelhanças existente entre eles. Esta semelhança pode ser devida à topologia de rede ou os atributos dos vértices.

As pessoas têm uma tendência para interagir com pessoas que apresentam orientações semelhantes ou pela similaridade de atributos como sexo, raça, idade ou classe social. Por exemplo, em uma rede de amizade, as pessoas da mesma idade estão conectadas entre si. A similaridade dos atores facilita a transmissão de informação e conhecimento, aumenta a cooperação e evita potenciais conflitos. Todavia grupos demasiadamente homogênoes perdem as vantagens competitivas da diversidade (IBARRA, 1992; MCPHERSON; SMITH-LOVIN; COOK, 2001).

A Figura 2-11 ilustra a propriedade homofilia na rede social Facebook. Pode-se verificar que os nós dentro dos grupos estão altamente conectados. Por exemplo, o grupo, em destaque na Figura 2-11 por uma linha, estão altamente conectados entre si porque pertencem à mesma Instituição de Ensino.

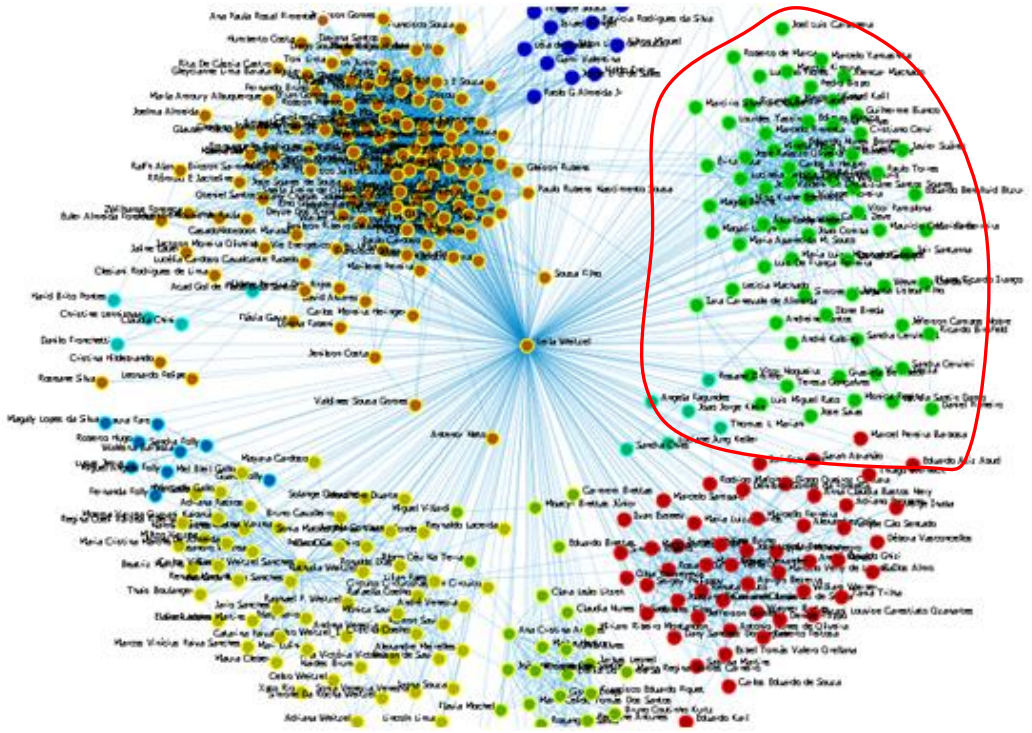


Figura 2-11: Representação da propriedade homofilia em uma rede social Facebbok

2.5.2 Transitividade

A díade é uma unidade relacional básica nos estudos de redes. As díades são formadas pela ligação entre dois atores. As relações entre três atores são tríades, compreende um conjunto de três nós não totalmente conectados. A transitividade corresponde o percentual de cliques presente na rede. Um clique é um conjunto de nós totalmente conectados (completa reciprocidade) sobre todas as possíveis tríades. Quanto mais transitiva é uma rede, mais provável encontrar um nó com vizinhanças comuns. Em outros termos, se **A** está conectado a **B** e **B** está conectado a **C**, então existe uma grande probabilidade de que o nó **A** esteja conectado a **C**. A transitividade é observada em redes reais, biológicas e sociais (BARABÁSI, 2003; NEWMAN, 2003).

$$\mathrm{T} = \frac{n_{\Delta}}{n_{\Lambda}} \qquad \text{Equação 2-1}$$

Onde: n_{Δ} é a quantidade de cliques e n_{Λ} é a quantidade de tríades presentes na rede.

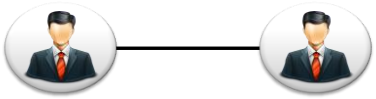


Figura 2-12: Representação de uma díade

Figura 2-13: Representação de uma tríade (à esquerda) e um clique (à direita)

2.5.3 Hierarquia

Os vértices podem ser conectados em uma ordem hierárquica, desta forma, as ligações entre os nós têm uma topologia hierárquica (COSTA *et al.*, 2007). Nestas redes existem mais ligações entre nós do mesmo nível que em níveis diferentes. Algumas redes do mundo real têm hierarquia, como por exemplo, a rede que representa a relação entre os funcionários de uma empresa. Na Figura 2-14, tem-se ilustrada uma rede hieráquica de uma empresa. O Diretor é o nó de número 27, e tem a posição mais central; os Gerentes (33, 18,19,41,44) estão diretamente conectados ao Diretor; o restante são os subordinados de cada departamento.

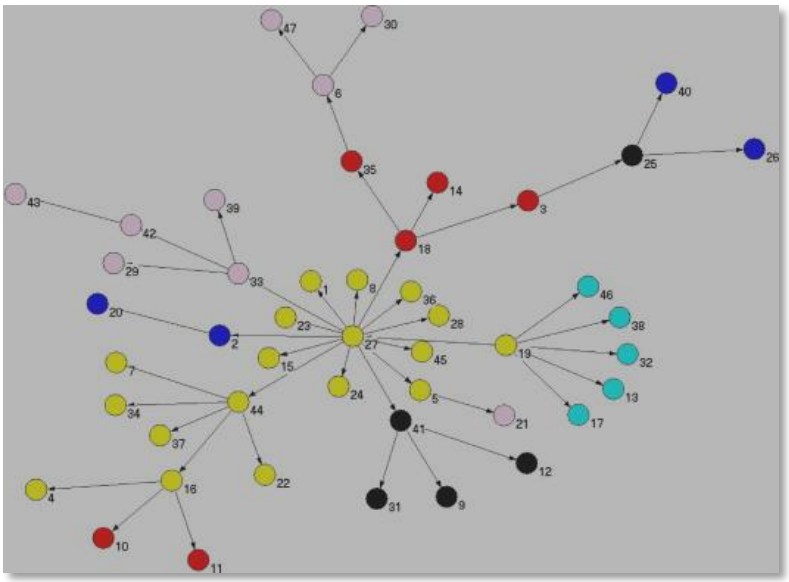


Figura 2-14: Rede Hierárquica em uma empresa

2.5.4 Comunidades

Essa propriedade está relacionada com a estrutura da rede, se a rede apresenta subgrupos, ou seja, apresenta vértices altamente conectados como em uma clique. É uma característica comum em redes sociais e biológicas. A propriedade comunidade pode ser também definida baseada na similaridade dos nós. Nós que pertencem à mesma comunidade devem ser mais similares (propriedade de homofilia) estruturalmente uns com outros que com nós fora da comunidade. Assim como os cliques, as comunidades são subconjuntos altamente conectados. Redes densas não possuem comunidades (NEWMAN, 2003). A Figura 2-11 mostra uma rede onde os nós com maior similaridade foram agrupados. Os grupos estão representados por diferentes cores de seus nós.

2.6 Propriedades Estruturais

Nesta seção é apresentado apenas um recorte das propriedades estruturais das Redes Complexas que serviram de base teórica neste estudo. Para estudos mais aprofundados encontra-se uma vasta bibliografia sobre o assunto em: (FREEMAN, 1979; FREEMAN, 1996; WASSERMAN, 1999; DIESTEL, 2000; ALBERT; BARABÁSI, 2002; DOROGOVTSEV; MENDES, 2003; NEWMAN, 2003; RAVASZ; BARABÁSI, 2003; NEWMAN; GIRVAN, 2004; WATTS, 2004; BOCCALETTI *et al.*, 2006; NEWMAN; BARABASI; WATTS, 2006).

2.6.1 Caminho - *l*

É chamado de Caminho de comprimento $\boldsymbol{l}$ qualquer sequencia de $\boldsymbol{n}$ vértices $\{\boldsymbol{v_1}, \boldsymbol{v_2},....\boldsymbol{v_l}\}$ tal que $\boldsymbol{v_i} \in \mathbf{N}$ e $(\boldsymbol{v}_i, \boldsymbol{v_{i+1}}) \in \mathbf{E}$ e não exista repetição de vértices nesta sequência, nem de arestas entre estes vértices.

Utiliza-se este conceito para definir a métrica de distância topológica em redes, de tal maneira que a distancia entre dois vértices quaisquer $\boldsymbol{i}$ e $\boldsymbol{j}$ é dada pelo comprimento do menor caminho que os une e denotada por $\boldsymbol{l_{ij}}$. A distância média entre um vértice $\boldsymbol{i}$ e todos os outros da rede é chamada comprimento do menor caminho médio e é denotada por $\boldsymbol{l_i}$. Tomando a média de $\boldsymbol{l_i}$ sobre todos os vértices $\boldsymbol{i}$, obtém-se o diâmetro **D** da rede

que fornece uma estimativa de quantas arestas existem em média entre dois vértices quaisquer da rede. Na rede sintética da Figura 2-15, o diâmetro **D** é igual à 7,00. O parâmetro <*l*> é caminho mínimo médio da rede, na Figura 2-15 tem valor igual à 3,13.

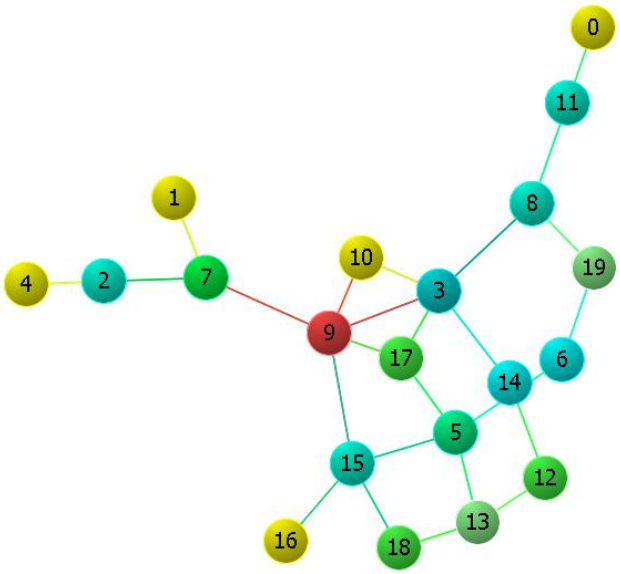


Figura 2-15: Rede sintética com N = 20 vértices e E = 24 arestas e o diâmetro da rede é 7,00. Nesta rede, o menor caminho entre os vértices v_{19} e v_7 possui comprimento $l_{19,7}$ igual à 4 passando pelos vértices {19,8,3,9,7}

2.6.2 Tamanho - E

É a quantidade de conexões existentes entre os vértices da rede (**E).**

2.6.3 Densidade - *de*

A densidade ***de*** é a razão entre o número de conexões existentes pelo número de conexões que são possíveis em uma rede. Em redes não direcionadas ***n*** é o número de vértices (N). O valor ***de*** varia entre 0 e 1 (o valor 1 quer dizer que todos os nós da rede estão conectados entre si). Uma rede perfeitamente conectada é um clique e tem densidade =1. A rede é considerada esparsa se o numero de conexões **E** é da mesma ordem do número de vértices ***n(G)***. Um grafo direcionado terá menor densidade que um grafo não direcionado equivalente, pois existem 2 vezes mais possibilidades de conexões (n(n-1)).

$$de = \frac{\mathrm{E}}{n(n-1)} \qquad \text{Equação 2-2}$$

2.6.4 Coeficiente de Agrupamento (ou Aglomeração) - *C*

O Coeficiente de Agrupamento de *i* é a razão entre o número de arestas existentes entre os vizinhos de *i* e o número máximo de arestas possíveis entre estes vizinhos. O Coeficiente de Agrupamento expressa a probabilidade de dois vértices (e_i) que estão conectados possuírem uma conexão em comum com um terceiro vértice. É utilizado para indicar transitividade e para indicar estrutura de comunidade. Os valores variam no

intervalo [0,1], onde zero implica em uma rede pouco transitiva e com poucas comunidades ou nenhuma. Redes reais tendem a exibir um alto grau de agrupamento.

$$C_i = \frac{2\, e_i}{k_i\,(k_i - 1)} \qquad \text{Equação 2-3}$$

Na Figura 2-16 o coeficiente de agrupamento do vértice V é igual à 2/3 = 0,666 porque existem três possibilidades de conexão existente entre seus vizinhos, mas apenas duas são efetivas. Os outros Coeficientes são w = 2/3, x = 1/3 e y =1/3. O coeficiente da rede é então igual à 1/2 = 0,5.

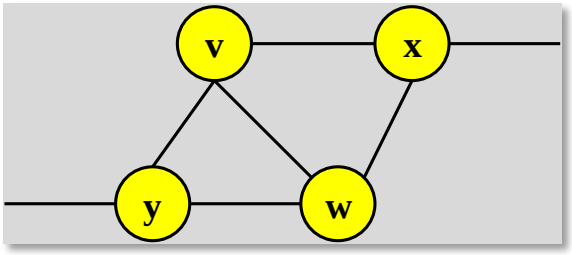


Figura 2-16: Coeficiente de Agrupamento

2.7 Análise de Redes Sociais

As pessoas estão inseridas na sociedade por meio das relações que desenvolvem durante toda sua vida, primeiro no âmbito familiar, em seguida na escola, na comunidade em que vivem, no trabalho, em agremiações esportivas ou religiosas, etc.

Uma Rede Social é uma estrutura social composta por indivíduos ou organizações, conectadas por um ou vários tipos de relações, onde partilham valores, informações, conhecimentos, interesses e esforços em busca de objetivos comuns (WASSERMAN; FAUST, 1994; WASSERMAN, 1999).

A Análise de Redes Sociais (SNA da expressão em inglês Social Network Analysis) é oriunda da sociologia, da psicologia social e da antropologia e a sua origem é inspirada na descoberta feita por J. L. Moreno (MORENO, 1934).

Em outono de 1932, houve uma grande tensão devido as constantes fugas da Escola Hudson (para meninas no estado de Nova Iorque). Em um período de apenas duas semanas, 14 meninas haviam fugido, uma taxa 30 vezes maior que a normal. Moreno (1934), um psiquiatra, sugeriu que a razão para a onda de fugas tinha estreita relação ao sistema social e não aos fatores individuais referentes às meninas, como por exemplo, personalidade e motivação. Moreno modelou a rede social utilizando a abordagem de sociogramas. O autor concluiu que, as fugas não ocorreram de modo consciente, era a posição que elas ocupavam na rede social que determinou quando e se elas fugiriam. A abordagem utilizada pelo psiquiatra ficou conhecida como sociometria. Nesta abordagem as relações interpessoais são representadas graficamente, através de um diagrama que representa as forças de atração, repulsão e indiferença que operam nos grupos.

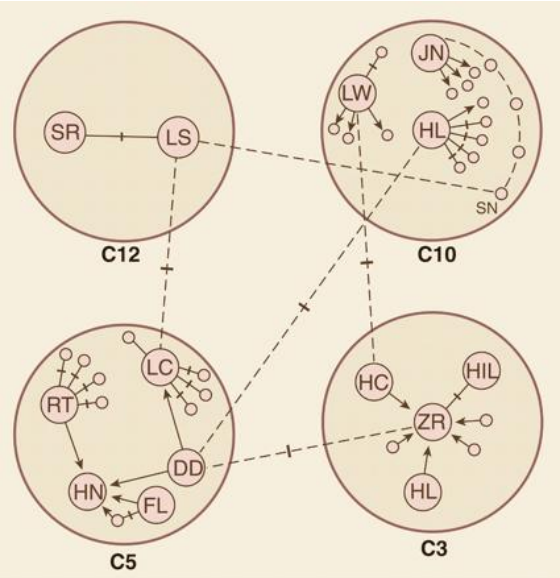


Figura 2-17: Modelo de rede proposto por Moreno (1934). Os quatro maiores círculos (C12, C10, C5, C3) representam as áreas onde as meninas viviam e os círculos menores representam todas as meninas. As 14 meninas fugitivas são identificados pelas iniciais SR, HC, etc. Os arcos não direcionados representam as forças de atração mútuas, e os não direcionados, caso contrário.

Foi Radcliffe Brown & Alfred Reginald (1940) e John Barnes (1954) que começaram a usar o termo "Rede Social". Essas pesquisas se baseavam nas características da estrutura global da sociedade e não nas características das redes pessoais. A ARS interessa a pesquisadores de vários campos do conhecimento que, na tentativa de compreender o seu impacto sobre a vida social, deram origem a diversas metodologias de análise que têm como base as relações entre os indivíduos, em uma estrutura em forma de redes (WATTS, 1999).

Em seu estágio inicial, ARS tinham como objetivo extrair dados de comportamentos sociais e analisá-los. Em geral, essas análises se baseavam no estudo dos participantes e suas ações, com pouca ou nenhuma ênfase aos relacionamentos. Somente mais tarde, com a incorporação de ferramentas matemáticas e, posteriormente, com a computação, que as análises puderam evoluir e alcançar diferentes campos em diferentes aplicações. A grande distinção da ARS hoje é a importância dada às relações entre os seus participantes (DEGENNE; FORSÉ, 1999).

Borgatti (2009) destaca a contribuição de Auguste Comte (1798- 1857), filósofo francês, para a origem das ideias e práticas relacionadas à intuição estrutural que permeiam a ARS atualmente. Segundo o autor, Comte foi o primeiro estudioso que propôs uma maneira de observar a sociedade em termos de interconexões entre atores sociais.

As ferramentas matemáticas, capazes de fundamentar a ARS, provieram da teoria de redes complexas, grafos e teoria estatística. Dessas teorias, especialmente de grafos, derivam muitos dos conceitos de redes amplamente utilizados. Essas ferramentas são utilizadas para avaliar as mudanças nas redes ao longo do tempo. Com o tempo, a forma ou a topologia da rede pode mudar. Um bom exemplo é movimento de uma pessoa em uma estrutura hierárquica. Consequentemente, ao longo do tempo, quem ou o que é

crítico em uma rede pode mudar. A ARS de redes não constitui um fim em si mesma. Ela é o meio para realizar uma análise estrutural cujo objetivo é explicar os fenômenos encontrados (DEGENNE; FORSÉ, 1999).

A ARS destaca um aspecto essencial: a ênfase que se dá nas relações entre entidades, em contraposição à ênfase que é dada às características (ou atributos) em outros métodos de análise (FREEMAN, 1979).

A ARS tem recebido grande visibilidade e despertado interesse da comunidade cientifica em várias áreas. Ela está fundamentada na observação que os atores sociais são interdependentes e que as conexões entre eles possuem importantes consequências para cada indivíduo. Diversos domínios de aplicação têm seus dados modelados como redes complexas, por exemplo, a Internet, World Wide Web (WWW), as redes sociais, de colaboração, biológicas, etc. (ALBERT; BARABÁSI, 2002; NEWMAN, 2003; WATTS, 2004).

A Análise de Redes Sociais (ARS) tem sido utilizada para:

a. Entender como e porque os usuários participam de uma comunidade, como essas redes crescem pelo aumento das relações de amizade (SIMPSON; MARKOVSKY; STEKETEE, 2011);
b. Verificar como a informação é propagada entre amigos (YE; WU, 2010; ZAMAN *et al.*, 2010; ZHOU *et al.*, 2010);
c. Identificar Atores Sociais influentes na rede (JIANWEI; LILI; TIANZHU, 2008; BONGWON *et al.*, 2010; GAYO-AVELLO, 2010; SAKAKI; MATSUO, 2010);
d. Divulgar produtos e ou serviços através de aplicações de marketing viral (HONG; DAN; DAVISON, 2011);
e. Identificar novas comunidades (HAN; LI; WANG, 2011).

Outros estudos incluem também:

a. A extração de propriedades estatísticas, por exemplo, a distribuição dos graus (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999) e o calculo do diâmetro (MILGRAM, 1967);
b. Desenvolvimento de modelos de redes que permitem entender o significado das propriedades estatísticas extraídas e como e porque as redes seguem tais propriedades(NEWMAN, 2003);
c. Predição do comportamento das redes e sua evolução (LESKOVEC; KLEINBERG; FALOUTSOS, 2005).

A ARS envolve estudos sobre as propriedades estruturais, propriedades gerais, identificação da topologia e a identificação dos atores (entidades) importantes (influentes sob algum ponto de vista). As propriedades estruturais e gerias servem para caracterizar e classificar o tipo de rede que se está tratando. Calcular a centralidade de uma entidade social significa identificar a posição em que ele se encontra em relação às trocas e a comunicação da rede. Embora não se trate de uma posição fixa,

hierarquicamente determinada, a centralidade em uma rede traz consigo a ideia de influência.

2.8 Medidas de Posição (ou Medidas de Centralidade)

A centralidade é uma medida de posição de um nó em relação à posição de outros nós da rede. Essa medida dá uma indicação da visibilidade do nó na rede. É neste sentido que as medidas de centralidade tentam descrever as propriedades da localização de um ator na rede. Estas medidas levam em consideração as diferentes maneiras em que um ator interage e se comunica com o restante da rede, sendo mais importantes, ou centrais, àqueles vértices localizados em posições mais estratégicas na rede, dada em função de alguns invariantes do grafo. Medidas de posição são as abordagens utilizadas

Um ator com alto grau de centralidade mantém numerosos contactos com outros atores da rede e podem ganhar o acesso e ou influência sobre os outros. Um ator central ocupa uma posição estrutural estratégica servindo como uma fonte ou um canal para grandes trocas de informações, transações ou outros recursos com outros atores. Um ator periférico mantém poucas relações, portanto, situa-se espacialmente nas margens de um diagrama de rede.

Freeman (1979) abordou o conceito de centralidade revisando um grande número de medidas até então publicadas. Diversas medidas foram criadas com base apenas em ideias intuitivas, sem buscar a formalização necessária para o desenvolvimento da teoria e outras foram introduzidas de maneira tão complexa que é difícil até mesmo descobrir o que está sendo medido, trazendo poucos reflexos no aspecto qualitativo. Freeman (1979) restringiu em apenas três medidas: Centralidade do Grau (*Degree Centrality*), Centralidade de Proximidade (*Closeness Centrality*), e Centralidade de Intermediação (*Betweenness Centrality*).

Em (BROWN; REGINALD, 1940; NIEMINEN, 1974; FREEMAN, 1979; WASSERMAN; FAUST, 1994; FREEMAN, 1996; DEGENNE; FORSÉ, 1999; WASSERMAN, 1999; ALBERT; BARABÁSI, 2002; DANA; LOEWENSTEIN, 2003; DOROGOVTSEV; MENDES, 2003; NEWMAN, 2003; WATTS, 2004; LESKOVEC; KLEINBERG; FALOUTSOS, 2005; BOCCALETTI *et al.*, 2006; JAMALI; ABOLHASSANI, 2006; KUMAR; NOVAK; TOMKINS, 2006; NEWMAN; BARABASI; WATTS, 2006; AHN *et al.*, 2007; COSTA *et al.*, 2007; BORGATTI *et al.*, 2009) tem-se um extenso referencial teórico sobre as medidas de posição e a sua aplicabilidade. As definições de medidas de centralidade que são apresentadas nas seções seguintes são baseadas nos autores acima citados.

2.8.1 Centralidade do Grau ou Grau total (*Degree Centrality*) - *k*

A medida Centralidade do Grau foi proposta inicialmente por Nieminen (1974). O grau de um vértice qualquer em uma rede define o número de arestas que incidem (conectam) aquele vértice. A medida leva em consideração apenas as conexões diretas (relações diretas com os vértices vizinhos). É, portanto, a contagem do número de adjacências de um vértice v_i. Nós isolados são àqueles que possuem zero grau de entrada e zero grau de saída.

Definição: Seja **G** um grafo qualquer (conexo ou não) com n vértices e seja v_i um vértice de **G**. A Centralidade do Grau de v_i, que é denotada por k_i , é o número de arestas incidentes a v_i . Isto é,

$$k_i = \sum_{j=1}^{n} a_{ij} \qquad \text{Equação 2-4}$$

Onde: a_{ij} são os elementos da matriz de adjacência do grafo **G**.

Considere a Figura 2-16, os graus dos vértices desta rede são w = v = x = y = 3.

Em uma rede direcionada os Graus de Entrada (*indegree* - k_i^{in}) e de Saída (*outdegree* - k_i^{out}) são as ligações que entram e saem respectivamente. O grau da conexão ***k*** de um vértice ***i*** em uma matriz de adjacência **A**, é dado pelo somatório dos graus de entrada e de saída deste nó:

$$K_i = \sum_{j=1}^{n} a_{ij} = k_i^{in} + k_i^{out} \qquad \text{Equação 2-5}$$

Na Figura 2-18, Grau de entrada do vértice de número 6 é igual a 4 e grau de saída é igual a 2. O Grau total deste vértice é igual a 6.

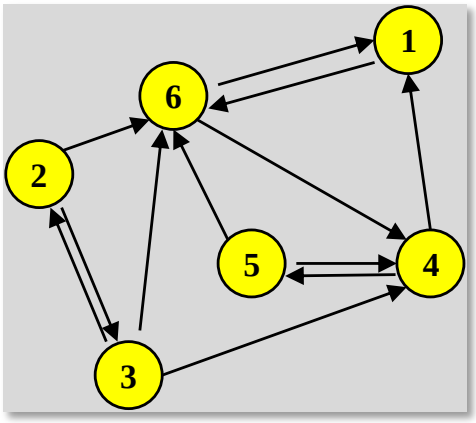


Figura 2-18: Rede direcionada

O grau médio de uma rede $\langle k \rangle$ é obtido a partir dos graus de cada vértice da rede. Essa medida identifica os *hubs* da rede, que são os vértices que têm alto grau de entrada em relação à média e os nós *authority* que são aqueles que apresentam alto grau de saída acima da média.

$$\langle k \rangle = \frac{\mathrm{E}}{\mathrm{N}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} k_i \qquad \text{Equação 2-6}$$

Onde:

E é numero total de conexões e **N** é numero total de vértices. Na Figura 2-18, o grau médio é igual a 12/6 = 2.

A distribuição de graus (ou distribuição de conectividade) é uma função de distribuição probabilística que indica a probabilidade de um determinado vértice ter grau fixo. Uma maneira de quantificar essa distribuição é por meio de uma Função de Distribuição Cumulativa (Equação 2.7), onde ρ_k é a fração de nós da rede com grau ***k*** e ***p(k)*** é a função cumulativa de distribuição de probabilidades.

$$p(k) = \sum_{k'=k}^{\infty} \rho_{k'} \qquad \text{Equação 2-7}$$

Na Figura 2-19 (a) e (b) verifica-se que a distribuição dos graus $p(k)$ segue a Lei de Potência ou Escala, pelo decaimento linear de $p(k)$ em função de *k*. O parâmetro γ é o coeficiente de Escala dado pela equação $p(k) \sim k^{-\gamma}$. Este parâmetro também é

conhecido como ponto crítico (no campo da Termodinâmica). No ponto crítico o sistema é invariante para transformações de escala.

Por exemplo, em um recipiente com água fervente, as bolhas de vapor, crescem, libertam-se, e flutuam até à superfície de onde se escapam para a atmosfera. À temperatura de ebulição, a água existe simultaneamente em duas fases distintas – líquido e gás – e à medida que as bolhas se formam as duas fases separam-se no espaço. Se fecharmos o recipiente a temperatura de ebulição aumenta, como em uma panela de pressão. À medida que a pressão aumenta, o sistema atinge o ponto crítico, onde as propriedades do líquido e do gás se tornam idênticas. Acima desta temperatura, no regime supercrítico, deixam de existir duas fases distintas e existe apenas um fluido homogêneo. Perto do ponto crítico, a matéria flutua sem limites. Bolhas e gotas, umas tão pequenas como uns quantos átomos, outras tão grandes como o recipiente, aparecem e desaparecem, juntam-se e se separam. Exatamente no ponto crítico a escala das maiores flutuações diverge, mas o efeito das flutuações em escalas menores não é desprezível. Neste exemplo, a distribuição das flutuações é invariante para transformações de escala.

Nas seções seguintes tem-se a explicação completa do conceito e onde se aplica.

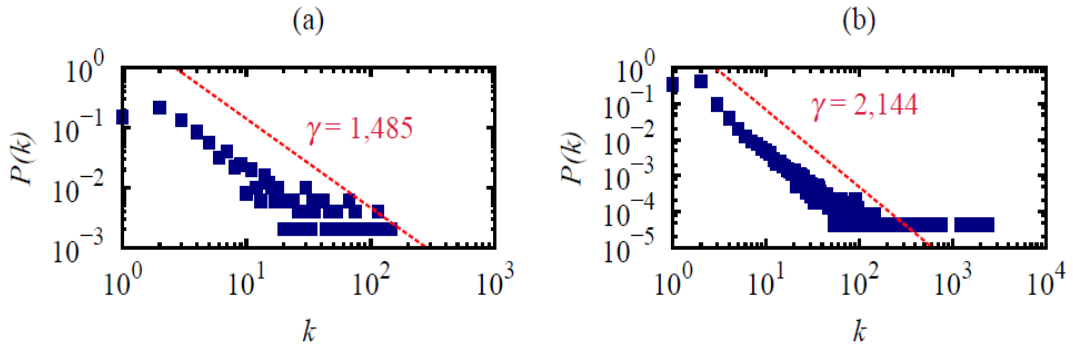


Figura 2-19: Distribuição de graus: (a) Rede de aeroportos (COLIZZA; PASTOR-SATORRAS; VESPIGNANI, 2007) onde **N** = 500, <***k***> = 11,92. (b) Rede da Internet onde **N** = 22963, <***k***> = 4,219 (ZHANG *et al.*, 2005)

2.8.2 Centralidade de Proximidade (*Closeness Centrality*) - *Cc*

Denomina-se de centralidade de proximidade de um ator a sua independência em relação aos outros e ele é tão mais central quanto menor o caminho que ele precisa percorrer para alcançar os outros elos da rede. Este tipo de centralidade depende não apenas das relações diretas, mas das relações indiretas, especialmente quando dois atores não estão adjacentes. A medida é baseada na soma das distâncias de um vértice em relação aos demais vértices do grafo. O distanciamento de um ator é a soma das distâncias geodésicas (caminho mais curto) para todos os outros atores. Entende-se que quanto mais próximo, mais rápida será a interação. Assim, maior a distância do nó em relação ao restante da rede, menor a sua centralidade de proximidade.

Definição: Seja **G** um grafo conexo com ***n*** vértices e seja v_i um vértice de **G**. A centralidade de proximidade de v_i é dada pelo inverso da soma das distâncias de v_i a todos os demais vértices do grafo. Onde ***d*** é o caminho mais curto.

$$Cc(v_i) = \frac{1}{\sum_{j=1}^{n} d\ (v_j, v_i)}$$

Por exemplo, em uma rede de transporte pode ser útil para determinar a localização de um centro de distribuição de mercadorias. Uma vez que o vértice com menor valor de centralidade de proximidade é aquele que poderá realizar com maior rapidez o processo de deslocamento de mercadorias para outras regiões.

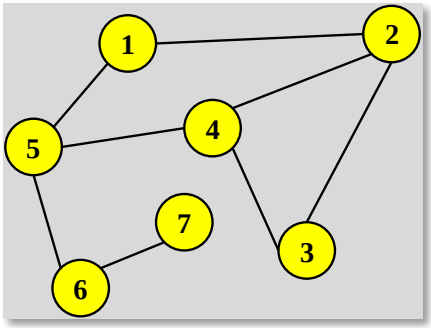


Figura 2-20: Centralidade de Proximidade

Na Figura 2-20, a ***Cc*** (v_1) = 11, *Cc* (v_2) = 12, *Cc* (v_3) = 13, *Cc* (v_4) = 10, *Cc* (v_5) = 9, *Cc* (v_6) = 12 e *Cc* (v_7) = 17. Assim o nó mais central é o nó v_5.

2.8.3 Centralidade de Intermediação (*Betweenness Centrality*) - *Bc*

A Centralidade de Intermediação é o potencial daqueles atores que servem de intermediários. Representa o quanto um nó atua como "ponte", facilitando o fluxo de informação em uma determinada rede.

O vértice com maior valor de ***Bc*** é aquele que participa mais ativamente em um processo de interação, onde os caminhos mais curtos são percorridos. É calculado através do somatório da quantidade de caminhos geodésicos que passam por um determinado vértice.

Definição: Seja **G** um grafo conexo ou não com ***n*** vértices. Seja v_k um vértice de **G** e ***Bc*** (v_k) a Centralidade de Intermediação de v_k . Considere um par de vértices v_i e v_j em **G**, tal que $i \neq j$, $i \neq k$ e $j \neq k$

Então:

$$Bc_{ij}(v_k) = \begin{cases} 0\ se\ não\ existir\ caminho\ entre\ eles \\ \frac{g_{ij}\ (v_k)}{g_{ij}}\ caso\ contrário \end{cases} \qquad \text{Equação 2-8}$$

Onde g_{ij} denota a número de geodésicas entre v_i , v_j e g_{ij} (v_k) denota o número de geodésicas entre v_i e v_j que passam por v_k.

Assim a Centralidade de Intermediação de v_k é dada por:

$$Bc(v_k) = \sum_{\substack{1 \le i < j \le n \\ i,j \neq k}} Bc_{ij}(v_k) \qquad \text{Equação 2-9}$$

Na Figura 2-21, o ***Bc***(1) =0, ***Bc***(2) =8, ***Bc***(3)= 0, ***Bc***(4) = 4, ***Bc***(5) = 0 e ***Bc***(6) = 0.

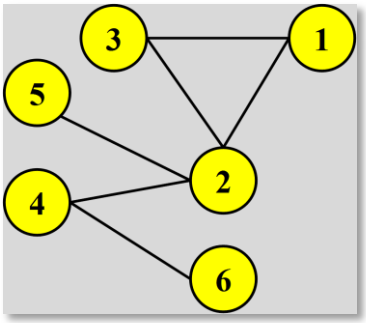


Figura 2-21: Centralidade de Intermediação

2.9 Outras Medidas

Nesta seção é apresentado um conjunto de outras medidas que são utilizadas na literatura, como por exemplo, a Centralidade do Autovetor e o *Pagerank* que serão utilizadas nesta pesquisa.

2.9.1 Centralidade do Autovetor (*Eigen Vector Centrality*) - *Ec*

Se **G** é um grafo e seu polinômio característico $\mathbf{p_g(\lambda) = det\ (A - \lambda I_n)}$ é o polinômio característico da matriz de adjacência de **G**, então as raízes do polinômio são os autovalores do grafo. O espectro de **G** é o conjunto dos autovalores[6], geralmente apresentados em ordem decrescente, associado às suas respectivas multiplicidades algébricas. A multiplicidade algébrica do autovalor **c** de uma matriz **A** é o número de vezes que o fator **(λ-c)** ocorre no polinômio característico de **A. A** soma das multiplicidades algébricas das raízes de um polinômio de grau n é n.

O conjunto de autovalores dessas matrizes é chamado o espectro da rede. Ele pode ser usado para localizar uma partição da rede ou para detectar os nós centrais (BONACICH; LLOYD, 2001; BONACICH, 2007).

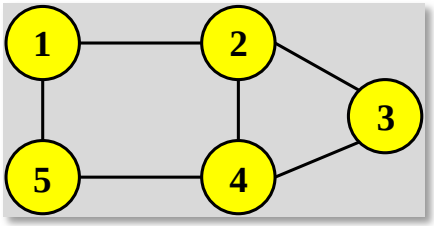


Figura 2-22: Representação do Autovetor

[6] Um autovalor de uma matriz quadrada **A** é um escalar c tal que **Av=cv** é verdadeiro para algum vetor **v** não nulo.

A representação da rede da Figura 2-22 em uma matriz de adjacencia é dada por:

$$A(G) = \begin{vmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{vmatrix}$$

Assim, $\boldsymbol{Ec}(1) = 0{,}36$ $\boldsymbol{Ec}(2) = 0{,}53$, $\boldsymbol{Ec}(3) = 0{,}43$, $\boldsymbol{Ec}(4) = 0{,}53$ e $\boldsymbol{Ec}(5) = 0{,}36$. Os vértices mais centrais são 2 e 4.

2.9.2 PageRank

O algoritmo *Pagerank* (PAGE *et al.*, 1999) assinala um valor de importância a paginas da Web, de forma que uma pagina ***p*** tem um peso proporcional ao número e importância das paginas que apontam (através de *hyperlinks*) para ***p***. Basicamente o *Pagerank* é uma avaliação da relevância da página. Essa relevância é divulgada em uma escala de zero a 10, quanto maior é esse numero maior é a relevância da página. Diferente do algoritmo HITS (KLEINBERG, 1999), o algoritmo de *PageRank* é independente de consulta.

2.9.3 HITS

Outro algoritmo que leva em conta a relação entre as paginas é o HITS (KLEINBERG, 1998) *Hypertext Induced Topic Search*, que utiliza o conceito de autoridades e *hubs*. Intuitivamente, *hubs* são páginas que não são autoridades por si só, mas direcionam os usuários a páginas importantes. Páginas importantes (autoridades), por sua vez, são páginas que são apontadas por vários *hubs* diferentes. O algoritmo é processado em tempo de consulta e não em tempo de indexação.

2.10 Modelos Teóricos de Redes

A análise estatística das propriedades das redes, como distribuição de graus entre os vértices, medida do caminho médio entre dois vértices e tamanho do componente conexo principal do grafo é importante para formalização de modelos de redes. Essas análises permitem entender melhor o processo de formação e também a organização das redes analisadas. Idealmente, se busca desenvolver modelos matemáticos que permitam a reprodução e geração de grafos com características estatísticas semelhantes àquelas da topologia da rede estudada; e de modelos capazes de prever o comportamento desses sistemas, baseados nas características dos vértices individuais ou da rede como um todo (NEWMAN, 2003).

Sendo assim, nesta seção é introduzido o conceito de Redes Regulares, Redes Aleatórias, Redes Livres de Escala e Redes Pequeno Mundo através de suas propriedades gerais e topologia.

2.10.1 Redes Aleatórias (ERDÕS; RÉNYI, 1959; 1960) e Redes Regulares

Em Redes Regulares os N vértices estão dispostos ao longo de um anel, onde todos os vértices apresentam o mesmo grau e a probabilidade de conexão $\boldsymbol{p}$ = 0. Redes Regulares estão associadas a altos valores de coeficiente de agrupamento **C** e de grau médio <$\boldsymbol{k}$> (Figura 2-25(a)). Em Redes Regulares, o grau médio é da ordem do número de nós **N** e o coeficiente de agrupamento **C** = ½.

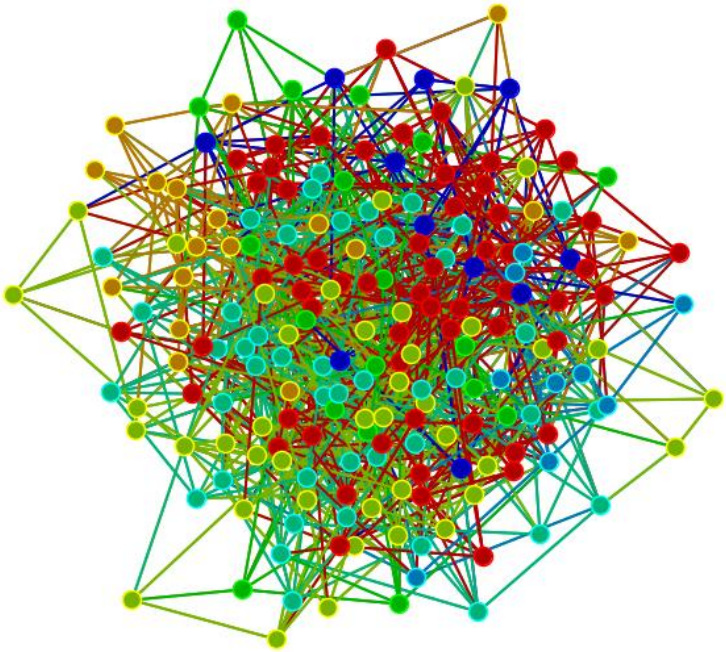


Figura 2-23: Exemplo de Modelo Aleatórias de Rede Erdõs e Rényi, N = 200 e grau médio <k> ≈ 6

As Redes Aleatórias foram analisadas inicialmente pelo matemático Paul Erdõs em conjunto com Alfred Rényi. Uma Rede Aleatória é construída definindo-se um conjunto de vértices V e conectando pares de vértices com probabilidade $\boldsymbol{p}$. As arestas são escolhidas aleatoriamente, dentre as ${}^{k_i(k_i-1)}/_{2}$ combinações possíveis da seguinte forma:

$$< k >= \frac{2n}{\mathrm{N}} = p(\mathrm{N}-1) \cong p\mathrm{N} \qquad \text{Equação 2-1}$$

Onde <k> é grau médio

Portanto, com $\boldsymbol{p}$ = 0 obtém-se uma rede completamente fragmentada e $\boldsymbol{p}$ = 1 tem-se uma rede completamente conectada (Figura 2-25(c)). A Rede Aleatória é referida como uma rede em equilíbrio onde a distribuição dos graus é independente do tempo.

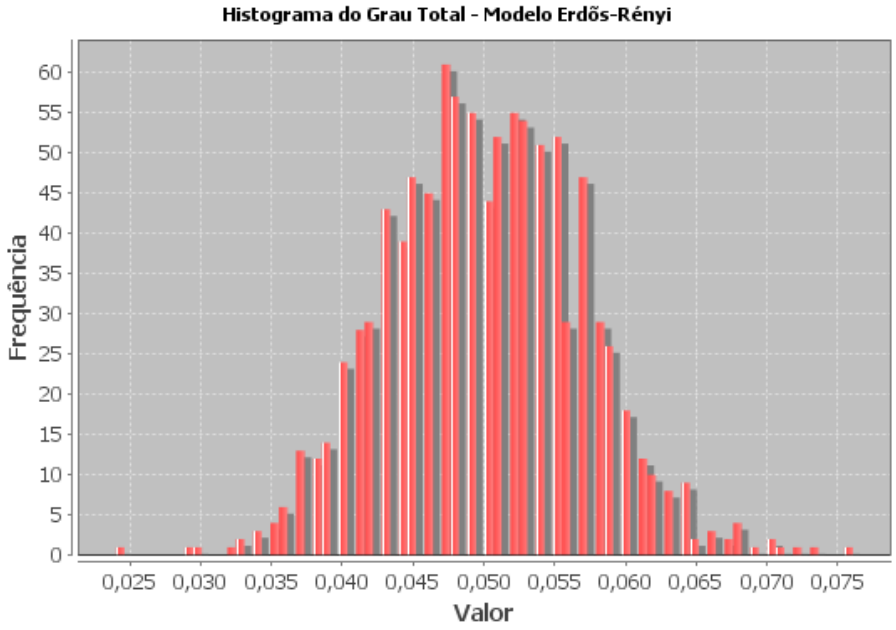


Figura 2-24: Histograma da frequência do Grau Total, N = 10000, grau médio <k> ≈ 14, tem distribuição normal

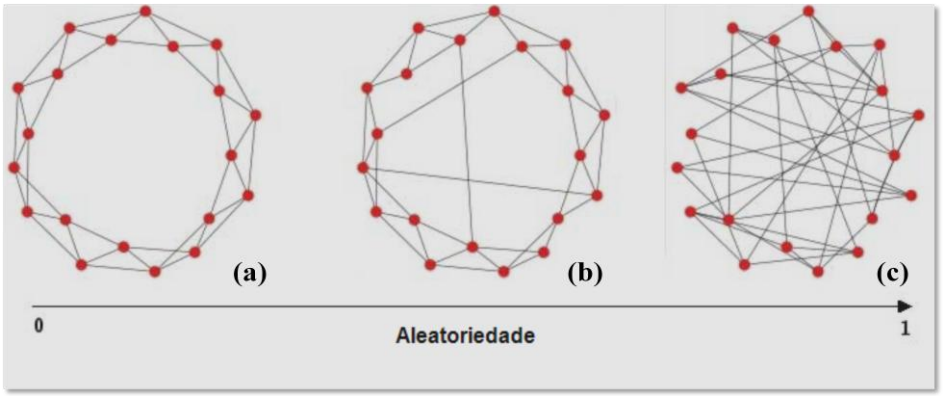


Figura 2-25: (a) Rede Regular, (b) Rede Pequeno Mundo (c) Rede Aleatória.

Em Redes Aleatórias, **N** é grande e ***p*** é mantido constante para todos os vértices, a distribuição do grau tende à distribuição de Poisson (ou Normal ou Gaussiana - Figura 2-26 e Figura 2-27). A curva tem formato de sino, que representa a distribuição Gausiana. Esta curva mostra, por exemplo, a variação nos preços de certo produto durante certo período de tempo. A maioria dos valores discretos dos preços situa-se na parte central da curva, ou seja, na média, enquanto que, nos lados, a curva cai rapidamente, como uma exponencial.

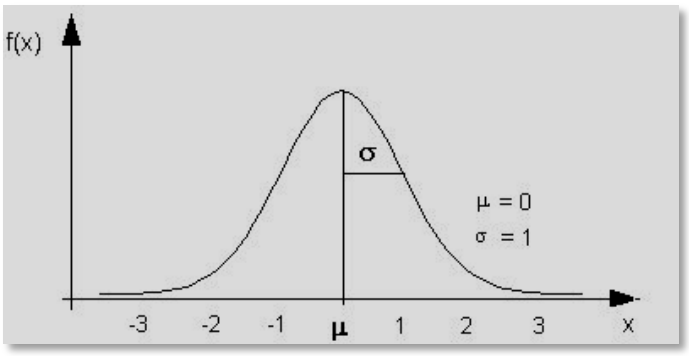


Figura 2-27: Distribuição Normal

2.10.2 Redes Mundo Pequeno

Milgran (1967), psicólogo social da Universidade de Harvard, promoveu um experimento para estudar o chamado Efeito do Pequeno Mundo para avaliar o grau de ligação entre pessoas. O experimento consistia em pedir a indivíduos de algumas cidades que enviassem cartas a conhecidos, com o objetivo de chegar a determinados residentes em Boston. A partir deste experimento, surgiu o conceito de seis graus de separação entre pessoas, mostrando que há uma probabilidade alta de que indivíduos desconhecidos possuam amigos em comum.

Todavia, em 2008, Leskovec e Horvitz (LESKOVEC; HORVITZ, 2008) estudaram a rede Messenger sistema de mensagens instantâneas da Microsoft. A rede foi modelada como um grafo não direcionado possuindo 179.792.538 vértices e 1.342.246.427 arestas. Os vértices são os usuários e os arcos representam as trocas de mensagens entre eles. Neste estudo os autores verificaram que nesta rede os graus de separação são em média 6,6 passos contradizendo os achados de Milgran (1967). A rede não tem características de redes livre de escala, pois, o coeficiente de escala γ é de 0,6 (considerado muito pequeno). Possui alta transitividade, isto é, pessoas com amigos em comum tendem a serem amigos e tem diâmetro igual a 7,8.

Na Figura 2-28 tem-se um exemplo de Rede Pequeno Mundo.

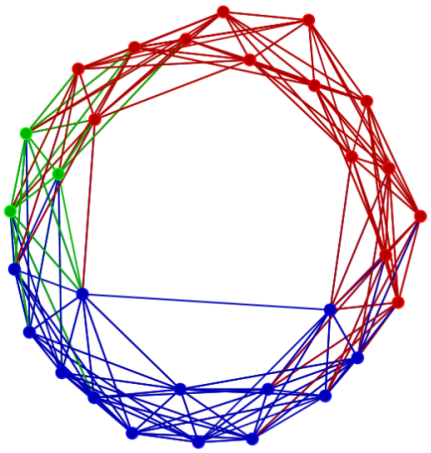


Figura 2-28: Exemplo de Rede Pequeno Mundo. **N** =30, grau médio <***k***> = 2

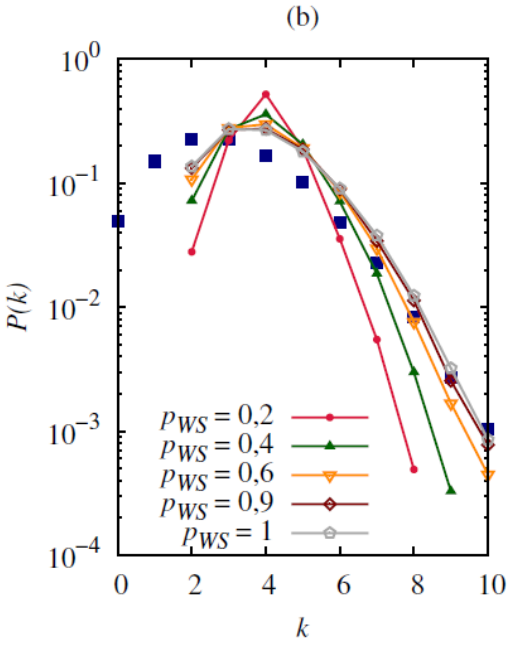


Figura 2-29: Modelo de Rede Pequeno mundo é gerado a partir de uma estrutura regular onde cada nó se conecta aos vizinhos mais próximos. Então, para cada conexão, o nó de uma das extremidades é trocado com probabilidade ***p***. No exemplo a rede possui **N**=1000 e grau médio <***k***> = 4 para diversos valores de ***p***.

Watts e Strogatz (1998), propuseram um algoritmo baseado em Redes Aleatórias, no qual buscavam mimetizar a topologia de interações sociais em um modelo abstrato para tentar estudar o efeito pequeno mundo. Os autores verificaram que este modelo de rede tem como características:

1. A distância média entre quaisquer dois vértices de uma rede muito grande não ultrapassa um número pequeno de vértices;
2. A distância entre os nós, pois a maioria dos vértices se conecta a outros através de um caminho mínimo.

Verificaram que Redes Reais não são inteiramente regulares nem completamente aleatórias (Figura 2-25 (b)).

2.10.3 Redes Livre de Escala

Ao analisar a rede mundial de computadores WWW, Barbási e colaboradores (2000) verificaram que a distribuição dos graus não é aleatória. Na web e em diversas redes reais a distribuição dos graus (***p***) tem um decaimento seguindo uma Lei de Potência (Figura 2-31), tal distribuição é chamada Livre de Escala, reforçando a ideia de que o universo aleatório de Erdõs e Rény tende a não estar presente na natureza.

$$\boldsymbol{p}\,(k) \sim k^{-\gamma} \quad \text{sendo que } 2 < \gamma < 3$$

As Redes Aleatórias servem como, parâmetros de medição e comparação entre modelos de redes. O modelo de Barbasi descarta a aleatoriedade e mostra que existem leis que regem a estrutura das redes naturais. Uma rede “sem escalas” apresenta baixo coeficiente de agrupamento e uma lei facilmente observada: poucos nós apresentam muitas arestas, enquanto muitos apresentam poucas arestas. É conhecido como o

princípio preferencial de conexão, quanto maior o grau do nó, maior a probabilidade de que ele receba mais vizinhos na próxima iteração do processo de crescimento da rede, conhecido como *rich get richer* - ricos ficam mais ricos. Ou seja, quanto mais conexões um nó possui, maiores as chances de ele ter mais novas conexões. Na Figura 2-30, pode-se verificar que existem nós com mais conexões que a média, esses nós são conhecido como *Hubs.*

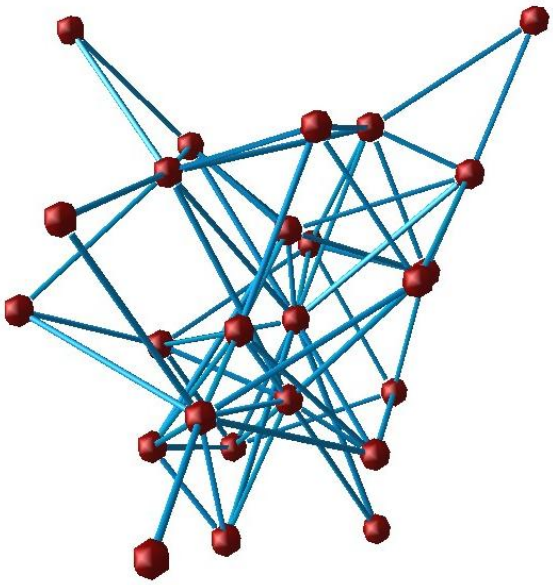


Figura 2-30: Exemplo do Modelo de Rede Livre de Escala, **N** =30 e grau médio <***k***> = 2

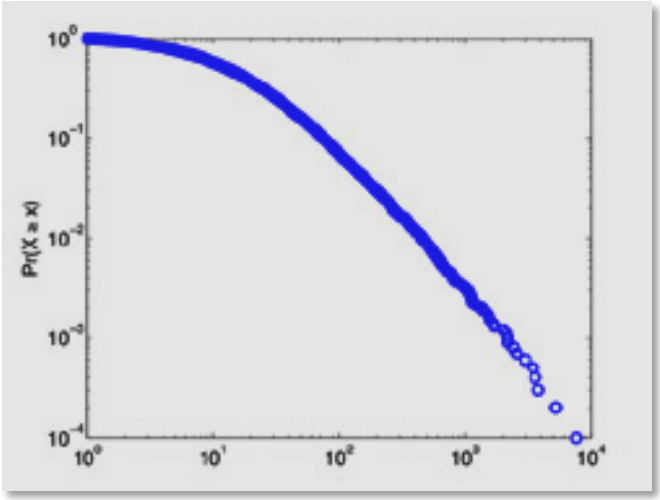


Figura 2-31: Gráfico de distribuição de grau com decaimento segundo a Lei de Potência, N = 1000 e $\boldsymbol{p}(k) \sim k^{-\gamma}$ e $\gamma = 2{,}7$

Redes Livre de Escala são mais tolerantes a falhas. Remoções aleatórias de vértices atingem, em grande parte das vezes, vértices de baixo grau, pois eles são maioria; de forma que, considerando o grau de um vértice como uma medida da sua informação, há baixa probabilidade de uma grande perda de informação em processos de falha numa rede Livre de Escala (COSTA *et al.*, 2007).

3 TRABALHOS CORRELATOS

Nesta seção são apresentados os trabalhos que orientaram nossa pesquisa. O objetivo em comum destas investigações é a avaliação da importância de um usuário, considerado "influenciador" em uma Rede Social ou em Redes do mundo real. Esses estudos, de um modo geral, fazem uso da estrutura e da topologia das Redes Sociais para a definição de metodologias de ordenamento da influência.

Algumas pesquisas consideram a influência de um usuário como sendo a capacidade de disseminar informação, comparável aos formadores de opinião. Outras consideram que os influenciadores são àqueles que propagam informações relevantes. A maioria desses estudos está voltada para aplicações estratégicas de marketing no endosso de produtos e serviços, campanhas políticas, estudos sociais entre outras aplicações.

Em geral, os "influenciadores" são definidos como indivíduos que produzem efeitos sobre as ações, comportamentos ou opinião de outras pessoas. Todavia, essa definição apresenta ambiguidades em relação à natureza da influência em questão, e consequentemente, sobre o indivíduo que pode ser considerado influente.

Usuários comuns (ou ordinários) e as mídias, celebridades e políticos poderiam ser igualmente considerados influentes? O endosso de um produto/serviço/mensagem que é dado por uma celebridade, ou por um amigo de confiança, ou por um perito influenciam de diferentes formas um indivíduo? Por exemplo, no *Twitter*, usuários com o mesmo número de seguidores não necessariamente exercem a mesma influência. Em ambientes virtuais de relacionamento (redes sociais, *blogs*, *wikis* etc) é indispensável ter em mente que influência pode não necessariamente ser sinônimo de popularidade, e popularidade nem sempre é sinônimo de reputação. Discussões sobre reputação e popularidade são feitas ao final deste capitulo.

3.1 Kwak e colaboradores (2010)

Kwak e colaboradores (KWAK *et al.*, 2010) estudaram as características topológicas da Rede *Twitter* e sua capacidade como nova plataforma para trocas de informação. Na época da publicação foi considerado o trabalho mais amplo em termos de base de dados já feito com a rede *Twitter*.

O objetivo da pesquisa foi avaliar a influência de usuários, medida em termos de popularidade na rede. Com este propósito foram extraídas 41.7 milhões de contas de usuários representando os vértices e 1.47 bilhões de relações, representando as ligações ou arcos no ano de 2009. Foram coletados também 106 milhões de *tweet* para avaliar quais eram os tópicos mais recorrentes. A avaliação foi feita em função de palavras

chave e de palavras precedidas do símbolo # (*hash*). Foi utilizado um suplemento do *Fire-Fox*[7] chamado *Clean Tweets* para o tratamento de *spam*.

Para a avaliação da influência, a rede foi modelada em um tido grafo direcionado onde os vértices representam os usuários e os arcos as relações do tipo follower (seguidor). A rede modelada apresenta as seguintes propriedades:

a. A distribuição de graus não segue a Lei de Potência;
b. O diâmetro da rede (d = 4) é considerado pequeno se comparado com outros tipos de redes;
c. A média dos caminhos médios é igual a 4,12, também considerado pequeno para o tamanho da rede;
d. A rede tem baixa reciprocidade, ou seja, existem poucas ligações bidirecionais (usuário A segue usuário B, mas B não segue usuário A).

Para o ordenamento da influência foram utilizadas as métricas: Quantidade de follower, *Pagerank* e Quantidade de Re*tweet*.

Tabela 3-1: Ordenação pela quantidade de seguidores (followers)

Ranking by # of followers		
ID	Name	Remark
aplusk	ashton kutcher	actor
britneyspears	Britney Spears	musician
TheEllenShow	Ellen DeGeneres	show host
cnnbrk	CNN Breaking News	news
Oprah	Oprah Winfrey	show host
twitter	Twitter	subject of this paper
BarackObama	Barack Obama	president of U.S.
RyanSeacrest	Ryan Seacrest	show host
THE_REAL_SHAQ	THE_REAL_SHAQ	sports star
KimKardashian	Kim Kardashian	model
johncmayer	John Mayer	musician
mrskutcher	Demi Moore	actress
iamdiddy	iamdiddy	musician
jimmyfallon	Jimmy Fallon	actor
lancearmstrong	Lance Armstrong	sports star
algore	Al Gore	politician
mileycyrus	Miley Cyrus	actress / musician
nytimes	The New York Times	news
coldplay	Coldplay	musician
TheOnion	The Onion	news

Fonte: (KWAK *et al.*, 2010)

[7] É um navegador livre e multi-plataforma desenvolvido pela Mozilla Foundation. Fonte: http://pt.wikipedia.org/wiki/Mozilla_Firefox

Tabela 3-2: Ordenação pela métrica *PageRank*

Ranking by PageRank in the following/follower network		
ID	Name	Remark
aplusk	ashton kutcher	actor
BarackObama	Barack Obama	president of U.S.
cnnbrk	CNN Breaking News	news
TheEllenShow	Ellen DeGeneres	show host
britneyspears	Britney Spears	musician
Oprah	Oprah Winfrey	show host
THE_REAL_SHAQ	THE_REAL_SHAQ	sports star
johncmayer	John Mayer	musician
twitter	Twitter	subject of this paper
RyanSeacrest	Ryan Seacrest	show host
lancearmstrong	Lance Armstrong	sports star
jimmyfallon	Jimmy Fallon	actor
iamdiddy	iamdiddy	musician
mrskutcher	Demi Moore	actress
PerezHilton	Perez Hilton	power blogger
nytimes	The New York Times	news
mileycyrus	Miley Cyrus	actress / musician
stephenfry	Stephen Fry	actor
TheOnion	The Onion	news
KimKardashian	Kim Kardashian	model

Fonte: (KWAK *et al.*, 2010)

Tabela 3-3: Ordenação pela quantidade de *retweet*.

Ranking by # of retweet in the diffusion network		
ID	Name	Remark
mashable	Pete Cashmore	news on social media
BreakingNews	BNO News	news
tweetmeme	TweetMeme	news on Twitter
oxfordgirl	oxfordgirl	journalist
cnnbrk	CNN Breaking News	news
TechCrunch	Michael Arrington	news on technology
myfabolouslife	Fabolous	musician
nytimes	The New York Times	news
lilduval	lil duval	comedian
IranRiggedElect	Iran	about Iran
espn	ESPN Sports News	news
persiankiwi	persiankiwi	about Iran
aplusk	ashton kutcher	actor
StopAhmadi	Raymond Jahan	about Iran
Alyssa_Milano	Alyssa Milano	actress
huffingtonpost	HuffingtonPost.com	news
iamdiddy	iamdiddy	musician
iranbaan	Fershteh Ghazi	about Iran
nprnews	NPR News	news
PerezHilton	Perez Hilton	power blogger

Fonte: (KWAK *et al.*, 2010)

3.1.1 Principais contribuições

Além da ampla análise da estrutura, topologia e propriedades da rede, os autores verificaram que:

a. A maioria dos usuários que têm menos de 10 seguidores nunca postou um *tweet* ou o fez apenas uma única vez. A média foi de apenas um *tweet* por usuário;

b. As celebridades, como por exemplo, atores (Ashton Kuster), músicos (Britney Spears, Cold Play), apresentadores (Ellen Degeneres, Oprah Winfrey), políticos (Barack Obama) mídias (CNN, The New York Times), etc apareceram nas primeiras posições do *rank* que utilizou a métrica quantidade de follower (Tabela 3-1);

c. O mesmo ocorreu com a metodologia que utilizou o *Pagerank*. As celebridades aparecem no topo da listagem, com pequena ou nenhuma modificação no posicionamento (Tabela 3-2);

d. O *rank* utilizando o quantidade de *retweet* apresentou variação com relação aos dois *rank* apresentados acima. A maioria dos usuários ordenado no topo são mídias, como por exemplo, *The Breaking NewsWire*, *ESPN Sports News*, *the Huffington Post* e *NPR News* (Tabela 3-3);

e. A Rede *Twitter* é diferente de outras redes (ponto-a-ponto – P2P, WWW, biológica e outras redes sociais como o Facebook);

f. O número de seguidores sozinho não reflete a influência de um usuário;

g. A ascensão das mídias no *rank* por *retweet* evidencia a confiança no conteúdo publicado por estes meios de comunicação.

Esta pesquisa servirá de base para estudos comparativos com a metodologia proposta nesta tese. Serão utilizadas as métricas: quantidade de seguidores, o *Pagerank* e quantidade de retweet.

3.2 Jianwei e colaboradores (2008)

Em JIANWEI; LILI; TIANZHU (2008) é avaliada a importância de um vértice (usuário) através de um conjunto específico de métricas. A metodologia de ordenamento utiliza as métricas:

a. Centralidade de Proximidade (***Cc***);

b. Centralidade de Intermediação (***Bc***);

c. Grau Total (***Dc***)

Os autores consideram que a importância de um vértice está relacionada não só aos relacionamentos diretos, mas também aos relacionamentos indiretos. Os relacionamentos diretos dizem respeito ao Grau Total e os relacionamentos indiretos à Centralidade de Intermediação e de Proximidade. Os autores acreditam que um conjunto de métricas com pesos associados é capaz de representar adequadamente a importância de um vértice. Representa-se assim de forma mais global e não apenas local a importância dos vértices.

$$C(v) = (\alpha * Dc) + (\beta * Cc) + (\gamma * Bc) \qquad \text{Equação 3-1}$$

Onde:

$C(v)$ é a importância do vértice e α , β e γ são os pesos associados e seguem a regra: α > β > γ e α + β + γ = 1.

A base de dados utilizada é uma rede real sobre AIDS (*Acquired Immune Deficiency Syndrome*). Os vértices representam os indivíduos portadores da síndrome e os arcos binários representam a relação sexual (existente ou não). A rede possui ao total 40 nós e 41 arcos.

Foram originadas cinco listas ordenadas, três listas com as métricas isoladamente e duas listas com a metodologia proposta (Equação 3-1). Os autores verificaram que pelo menos 13 vértices são recorrentes no topo da listagem com as métricas ***Bc*** e na $C(v)$, considerando-se apenas as quarenta primeiras posições. Observaram também que os três ranking que utilizavam as métricas isoladamente estão fortemente correlacionados. Os pesos utilizados são ilustrados nas equações abaixo:

$$C(v) = (0{,}6 * Dc) + (0{,}3 * Cc) + (0{,}1 * Bc) \quad \text{Equação 3-2}$$

$$C(v) = (0{,}5 * Dc) + (0{,}3 * Cc) + (0{,}2 * Bc) \quad \text{Equação 3-3}$$

3.2.1 Principais críticas

Algumas considerações sobre o trabalho de Jianwei e colaboradores (2008) devem ser destacadas. Em primeiro lugar os autores propõem avaliar a importância de um vértice, mas não descrevem o contexto ou propósito da importância deste nó na rede. Os resultados dos ordenamentos não foram explicados e não apresentaram significado claro, porque não ficou claro na pesquisa o objetivo adjacente da importância do ordenamento dos vértices. Por exemplo, o vértice classificado no topo é importante como transmissor ou portador da síndrome. Não foi feito um estudo da topologia ou da estrutura da rede.

A pesquisa de Jianweii e colaboradores (2008) serviu de inspiração à nossa pesquisa. Em especial na utilização de uma metodologia que possa avaliar de forma mais global e não apenas local um vértice na rede. Os autores utilizaram uma estratégia de pesos associados às métricas em uma espécie de combinação linear, e é essa estratégia que irá nortear a nossa abordagem de estimar a reputação.

3.3 Cha e colaboradores (2010)

Em Cha e colaboradores (CHA *et al.*, 2010) estuda-se padrões de influência de um usuário no ambiente *Twitter*. A influência é definida como sendo aquele usuário que é um importante instrumento na propagação da informação. Procuram estabelecer a dinâmica dessa importância em função dos tópicos que são publicados ao longo do tempo. Como objetivo adjacente, buscam entender o fenômeno que torna usuários corriqueiros em indivíduos com grande influência em um curto espaço de tempo. A rede foi modelada como um grafo onde os nós representam os usuários e os arcos a relação padrão do *Twitter*, ou seja, seguidor.

A base de dados conta com dois bilhões de arcos e 54 milhões de usuários e 1,7 bilhões de *tweets*. Para a avaliação da influência são utilizadas as métricas: (i) - Grau de Entrada (***Dci$_n$***); (ii) – Quantidade de *retweet* e (iii) – Quantidade de *mention*. O *mention*

é uma ferramenta de monitoramento de conversas no *Twitter*. Acontece quando há uma resposta a um *tweet* (ou quando se comenta) citando determinado usuário marcado pelo caractere arroba - @ e inserido ao conteúdo do *tweet*.

Exemplo:

@***l_weitzel*** você viu o que ***@nature*** escreveu?

Neste exemplo a mensagem é direcionada ao usuário ***l_weitzel*** e o usuário ***nature*** está sendo citado (mencionado) no *tweet*.

3.3.1 Principais contribuições

a. As análises das três métricas proporcionaram um melhor entendimento dos diferentes papeis que os usuários podem assumir em redes sociais;

b. Verificaram que o Grau de Entrada (Dc_{in}) representa a popularidade de um usuário na rede;

c. Verificaram que os *retweets* agregam valor a um *tweet*;

d. Verificaram que o *mention* agrega valor ao usuário que é citado;

e. Os achados podem auxiliar em campanhas de marketing na rede, uma vez que os usuários mais influentes possuem também influencia significativa em uma variedade de tópicos;

f. Verificaram que, quanto mais ativo um usuário for, maior será a influência exercida, ou seja, quanto mais *retweet* ou *mention* um usuário tiver da sua rede, maior será sua influência.

Esta pesquisa também irá servir de base para um estudo comparativo. Serão utilizadas as métricas Grau de Entrada e a Quantidade de *retweet*. Tem-se como objetivo verificar se a relação entre o Grau de Entrada e a popularidade se mantém caso fosse utilizada uma estrutura e topologia de redes diferentes.

3.4 Yang e colaboradores (2012)

Yang et al (2012) propõem uma rede onde os nós representam tanto os usuários quanto os *tweets* e os arcos são os *retweet*. O objetivo da modelagem é encontrar *tweets* interessantes.

A rede é modelada como um grafo direcionado **G = (N , E)** onde **N** são os nós e **E** conjunto de vértices. Os nós são do tipo **U = {u_1,....u_n}** de usuários e **T =**{t_1,.....t_n**)** de *tweets*. As relações de *retweet* entre esses nós, por exemplo se o usuário u_a re*tuitou* u_b é dada por: e_{u_a,u_b}. Da mesma forma se um *tweet* t_a criado por um usuário u_a foi retuitado por um usuário u_b, tem-se um arco e_{t_a,t_b}.

Os autores propuseram uma variante dos algoritmos **HITS** - *Hyperlink Induced Topic Search* com as medidas *Hubs* e *Authority* (KLEINBERG, 1999). Originalmente este algoritmo permite inferir autoridades dentro de um conjunto, baseado na relação entre páginas que são autoridades e páginas que interligam essas autoridades, isto é, os hubs. O algoritmo **HITS** modificado atualiza de maneira iterativa os pesos, utilizando as equações a seguir. Essas equações calculam $A(u_i)$ e $H(u_i)$ para a rede de usuários.

$$A(u_i) = \sum_{\forall j : e_{u_j,u_i} \in E} \frac{|\{u_k \in U : e_{u_j,u_k} \in E\}|}{|\{k : e_{u_j,u_k} \in E\}|} \times H(u_j)$$

$$H(u_j) = \sum_{\forall i : e_{u_j,u_i} \in E} \frac{|\{u_k \in U : e_{u_k,u_i} \in E\}|}{|\{k : e_{u_k,u_i} \in E\}|} \times A(u_i)$$

As equações abaixo calculam $A(t_i)$ e $H(t_i)$ para a rede de *tweets*.

$$A(t_i) = S_{U_A}(C(t_i)) + \alpha \sum_{\forall j : e_{t_j,t_i} \in E} F(e_{t_j,t_i}) \times H(t_j)$$

$$H(t_j) = S_{U_H}(C(t_j)) + \alpha \sum_{\forall i : e_{t_j,t_i} \in E} F(e_{t_j,t_i}) \times A(t_i)$$

O desempenho da metodologia foi calculado usando as medidas: P@10, P@20, R-Precision e MAP (Mean Average Precision). Os autores não especificaram a escolha dos diferentes níveis da precisão. Os resultados dos diferentes desempenhos oscilaram entre 85% e 77% aproximadamente.

3.4.1 Principais contribuições

a. Os autores observaram que apenas a contagem de *retweet* não é um indicador suficiente para medir significância de um *tweet*. Os autores discutem que deve haver outros fatores cognitivos e de comportamento envolvidos no processo de *retweet*, uma vez que nem todo *tweet* é retuitado;

b. Observaram que os usuários com maior autoridade são àqueles que têm *tweet* mais interessantes;

c. Utilizaram um fator para mitigar a característica que alguns usuários têm de atrair centenas de milhões de usuários, como por exemplo, as celebridades. Usuários com milhares de seguidores têm maior tendência de serem retuitados.

Até onde vai nosso conhecimento, a pesquisa de Yang et al (2012) é a única que também utiliza um grafo onde as relações são representadas pelo processo de *retweet*. Ressalta-se ainda que, esta pesquisa é posterior à nossa pesquisa.

3.5 Anger e Kittl (2011)

A pesquisa de Anger e Kittl (2011) buscou avaliar a influência de usuários na rede *Twitter*. Os autores propõem duas metodologias para ordenação da influência:

a. Taxa de *Retweet* que é calculada pela razão entre a quantidade de *retweet* e o número de seguidores;

b. Taxa de Interação que é calculada pela razão entre a quantidade de *retweet* e a quantidade de *tweet*;

c. Taxa de seguidores que é calculada pela razão entre quantidade de seguidores e quantidade de seguidos

Os resultados encontrados foram comparados com a medida *Klout*. A medida *Klout*[8] é um indicador de influência que está disponível na web. A medida varia de 0 à 100, quanto maior, maior é a influência. São utilizadas 25 variáveis para o calculo da medida e agrupadas em três classes:

a. *True Reach (*Alcance Real);

b. *Amplification probablity* (Probabilidade de Amplificação);

c. *Network Influence* (Influência na Rede).

O Alcance Real mede o tamanho da rede considerada ativa, dando prioridade aos usuários mais ativos. Os principais fatores avaliados são: seguidores, amigos (friendship), total de *retweet*, entre outros.

A probabilidade de amplificação mede o potencial que suas mensagens têm de impactar terceiros (seguidores dos seus seguidores) e até onde essa mensagem pode chegar. Privilegiam quem tem "bons" seguidores, ou seja, pessoas relevantes lhe seguindo também. Os principais fatores levados em consideração nesse critério são: *retweet* únicos, mensagens únicas *retuitadas*, percentual de *retweet* dos seguidores, menções únicas, porcentagem de menções, número de mensagens enviadas entre outras.

Na medida Influência na Rede a pontuação varia de acordo com a influência das pessoas que você se relaciona, qual o perfil das pessoas que lhe seguem, se são importantes ou ativas e os valores da medida *Klout* delas. Alguns fatores avaliados são: inclusões em listas, diferenças entre seguidores e seguidos, porcentagem de seguidos seguidores, envios únicos, *retweet* únicos, influência dos seguidores, influência dos *retuitadores* e dos *mentions.*

É também utilizada a ferramenta *Twitter Grader*[9] como instrumento de comparação. Essa ferramenta estabelece um valor percentual de 0 a 100. Uma nota 87, por exemplo, significa que sua conta é mais bem graduada que 87% das contas já avaliadas pela ferramenta. Calcula a medida de acordo com um conjunto de fatores por eles definidos. Alguns desses que são divulgados incluem: O número de seguidores, a influência da sua rede de seguidores, a frequência de suas atualizações, a completude da sua conta do *Twitter* e outros que não são divulgados. Os dados foram extraídos dos 10 maiores usuários em função da quantidade de seguidores, apenas de usuários que declararam ser natural da Áustria.

3.5.1 Principais contribuições

a. A quantidade de seguidores não reflete a influência de um usuário.

b. A revisão da literatura revelou que existem diferentes métodos de ordenamento da influência em diferentes contextos resultando em diferentes *ranks*;

c. Os autores ainda chamam a atenção para o fato de que não existe um consenso sobre o que é influência no ambiente *Twitter*.

[8] http://www.klout.com

[9] http://*tweet*.grader.com/

3.6 Mishra e Bhattacharya (2011)

Os autores discorrem sobre redes baseadas em credibilidade, que estas são diferentes de outros tipos de redes. Uma ligação em uma rede como o Facebook ou Youtube significa que dois nós estão conectados, isso por si só não garante que, só por estarem conectados, existe confiança entre eles. Por outro lado, em uma rede baseada em credibilidade, tal como o *Slashdot*[10] (http://www.slashdot.org) e *Epinions*[11] (http://www.epinions.com), uma opinião neutra não quer dizer que não exista ligação entre os usuários. Considere o seguinte exemplo: O nó A tem 1.000 opiniões neutras e 10 negativas e, o nó B tem apenas 10 opiniões negativas. É intuitivo achar que o nó A tem maior notoriedade. No entanto, se a opinião neutra é modelada pela ausência de uma ligação, os dois nós têm a mesma pontuação, o que não é verdade. Sendo assim, uma opinião neutra não é o mesmo que uma conexão ausente. Em outras palavras, uma conexão com peso zero é diferente de uma conexão ausente.

Tradicionalmente, técnicas como HITS (KLEINBERG, 1998) e *Pagerank* (PAGE *et al.*, 1999) classificam prioritariamente os vértices com maior conectividade. Todavia em redes baseadas em confiança, o prestígio de um vértice depende da opinião de outros vértices. Todavia, a confiabilidade de um vértice depende de o quão verdadeiro é esta opinião frente a outras opiniões. Por exemplo, se A tem uma opinião negativa sobre B, e B tem todas as outras opiniões positivas, pode haver um viés na opinião de A.

Os autores propõem uma metodologia para calcular o prestígio e confiabilidade dos vértices de uma rede baseada em confiança. Os nós são os usuários e os arcos são as opiniões. Consideram que a opinião de nós confiáveis deve pesar mais que a opinião de outros nós menos confiáveis. Algumas restrições são impostas à metodologia:

a. Se um usuário não opinar sobre alguma coisa ou sobre alguém isso que dizer que não existe conexão;

b. Se a opinião for neutra o peso do arco é zero;

c. Se um usuário sempre opina positivamente em qualquer ocasião, o peso assume valores não negativos, diminuindo a confiabilidade do nó.

A rede é modelada como um grafo G = {V, E} onde os arcos $e_{ij} \in E$ (do nó i para nó j) tem os pesos associados $w_{ij} \in [-1, 1]$.

As redes utilizadas no estudo são: *Slashdot*[12] (www.slashdot.org) e *Epinions*[13] (www.epinions.com) redes onde é possível dar opiniões sobre um determinado usuário ou conteúdo.

---

[10] É um site de notícias. A maior parte dos artigos é de sumários de notícias publicadas em outros sites, com espaço aberto ao comentário dos leitores.

[11] O *Epinions* é um serviço online que oferece avaliações de produtos feitas pelos próprios consumidores. É uma espécie de rede social para compartilhamento de opiniões dadas pelos usuários acerca de determinado item.

[12] É um site de notícias. A maior parte dos artigos é de sumários de notícias publicadas em outros sites, com espaço aberto ao comentário dos leitores.

3.6.1 Principais contribuições

a. Os resultados evidenciaram que redes sociais e redes ponto-a-ponto podem ser avaliadas através da ponderação dos seus arcos;

b. Os vértices podem ser avaliados em termos de confiabilidade e prestígio em redes sociais e redes ponto-a-ponto.

Resumo da seção:

Destaca-se aqui os trabalhos de KWAK *et al.* (2010), JIANWEI; LILI; TIANZHU (2008) e CHA *et al.* (2010). Estas pesquisas irão servir de base para estudos comparativos. Pretende-se utilizar as métricas encontradas nestes trabalhos como um conjunto teste da nossa metodologia.

3.7 Reputação, Popularidade e Autoridade

Os valores, como por exemplo, Reputação, popularidade e autoridade envolvem mudanças bastante expressivas nos modos através dos quais esses valores são construídos e moldados em AVR.

3.7.1 Reputação

A reputação é compreendida como a percepção construída de alguém pelos demais atores e, portanto, envolvem três elementos: o eu e o outro e a relação entre ambos. O conceito de reputação, portanto, implica diretamente no fato de que há informações sobre quem somos e o que pensamos, que auxiliam outros a construir, por sua vez, suas impressões sobre nós.

De uma forma geral, a reputação de alguém seria uma consequência de todas as impressões dadas e emitidas deste indivíduo. Reputação de uma entidade social é uma opinião sobre essa entidade, é um resultado da avaliação social de um conjunto de critérios. Reputação não é simplesmente o número de leitores de um blog, o número de seguidores do *Twitter* ou o número de opiniões positivas.

No sistema de reputação do site *eBay* (http://www.ebay.com/), compradores e vendedores podem avaliar uns aos outros após cada transação, e a reputação global de um participante é a soma destas classificações ao longo dos últimos 6 meses. As pontuações são dadas em função da experiência direta entre compradores e vendedores (HOOKER, 1912; RESNICK *et al.*, 2000; GUHA *et al.*, 2004; JØSANG; ISMAIL; BOYD, 2007; VARLAMIS; EIRINAKI; LOUTA, 2010; O'DOHERTY; JOUILI; ROY, 2012).

3.7.2 Popularidade

A popularidade é um valor relacionado à audiência, que é também facilitada nas redes sociais na Internet. Como a audiência é mais facilmente medida na rede, é possível visualizar as conexões e as referências a um indivíduo, assim a popularidade é mais facilmente percebida.

---

[13] O *Epinions* é um serviço online que oferece avaliações de produtos feitas pelos próprios consumidores. É uma espécie de rede social para compartilhamento de opiniões dadas pelos usuários acerca de determinado item.

A popularidade nada mais é que um valor relativo à posição de um ator dentro de sua rede social. Um nó mais centralizado na rede é mais popular, porque há mais pessoas conectadas a ele, mas isso não quer dizer necessariamente que este nó terá uma capacidade de influência mais forte que outros nós na mesma rede. A avaliação da popularidade de um blog pode ser feita pela quantidade de comentários recebidos, ou pelo número de seguidores. Da mesma forma, no *Twitter* a popularidade de um perfil pode ser dada pela quantidade de seguidores, pela quantidade de *tweet*, ou outras métricas comumente utilizadas (HOOKER, 1912; RESNICK *et al.*, 2000; GUHA *et al.*, 2004; JØSANG; ISMAIL; BOYD, 2007; VARLAMIS; EIRINAKI; LOUTA, 2010; O'DOHERTY; JOUILI; ROY, 2012).

3.7.3 Autoridade

Pode haver um terceiro valor, a autoridade. A autoridade refere-se ao poder de influência de um nó na rede social. Não é a simples posição do nó na rede, ou mesmo, a avaliação de sua centralidade. É uma medida da efetiva influência de um ator com relação à sua rede, juntamente com a percepção dos demais atores da reputação dele. Autoridade, portanto, compreende também reputação, mas não se resume a ela. Autoridade é uma medida de influência, da qual se depreende a reputação. Autoridade é uma medida que só pode ser percebida através dos processos de difusão de informações nas redes sociais e da percepção dos atores dos valores contidos nessas informações (HOOKER, 1912; RESNICK *et al.*, 2000; GUHA *et al.*, 2004; JØSANG; ISMAIL; BOYD, 2007; VARLAMIS; EIRINAKI; LOUTA, 2010; O'DOHERTY; JOUILI; ROY, 2012).

.

4 AMBIENTE E METODOLOGIA DA PESQUISA

Neste Capítulo, em usa primeira seção, é descrito o ambiente - a rede Twitter - no qual a pesquisa se insere. Inicia-se pela descrição de suas características e modelo de rede que é utilizado. Na segunda seção é descrito em detalhes o modelo de rede proposto, as características da topologia e as suas propriedades. Na terceira seção é descrita a metodologia para estimar a reputação, o algoritmo desenvolvido e, por fim, discute-se como será avaliada essa metodologia.

4.1 Principais características do Twitter

O *Twitter* surgiu em março de 2006. A proposta inicial do serviço era permitir que seus usuários interagissem com seus amigos, colegas e parentes, contando o que faziam enquanto navegavam na internet. Conforme a ferramenta se popularizou, os usuários se apropriaram da ideia para várias funções - como mural de recados, difusor de informações, etc, e hoje é considerado um bloco virtual de anotações.

A característica principal, no entanto, permanece, o *Twitter* é uma Rede Social de envio de mensagens (*post*) curtas e instantâneas (de até 140 caracteres) chamadas de *tweet*. O laço que é estabelecido neste ambiente é o de *follower* ou seguidor, que são as pessoas que seguem determinado usuário. O objetivo da relação é manter-se atualizado através do recebimento dos *tweet* dos indivíduos que você segue.

Um usuário no *Twitter* não necessariamente é uma pessoa, pode ser uma organização (jornais, revistas televisão), órgão governamental, entre outros. O relacionamento constituído é do tipo direcionado, o sentido padrão da direção vai do seguido (*following*) para o seguidor (*follower*) a seta aponta para o seguidor. Todavia é possível que haja reciprocidade, isto é, o sentido seguido-seguidor e seguidor-seguido formando os laços de amizade (*friendship*).

É interessante notar que as relações recíprocas no *Twitter* não tem o mesmo significado como em redes tal como o *Facebook*, *Orkut* etc. As relações de reciprocidade não são estabelecidas automaticamente. Por exemplo, se o usuário **A** segue o usuário **B**, isso não quer dizer que o usuário **B** automaticamente segue o usuário **A**. A relação de reciprocidade não é mandatória, mas se assim o fizer é porque **B** compartilha preferências de **A** sobre um tema, ou por qualquer outro fator.

A Figura 4-1 exemplifica uma visão reduzida da topologia da rede *Twitter*, ou rede de seguidores como é conhecida. O sentido das setas ilustra o laço padrão apontando para o seguidor. O ator B tem como seguidores A e D. O ator C tem como seguidores B, F, G e H. O ator A tem como seguidor B e E, os restantes, D, E, F, G e H não possuem seguidores. Note que o relacionamento de B com A e A com B destacado nas cores verde e azul formam laços de amizade.

Ao se referir a uma rede social pelo ponto de vista de um ator em particular a sua rede é referida como *ego-network* e qualquer outro ator fora de sua *ego* é referido como *alter* e a sua rede como *alter-network*. Por exemplo, a *ego-network* do ator B são os atores A, C e D e a sua *alter-network* são os atores E, F, G e H.

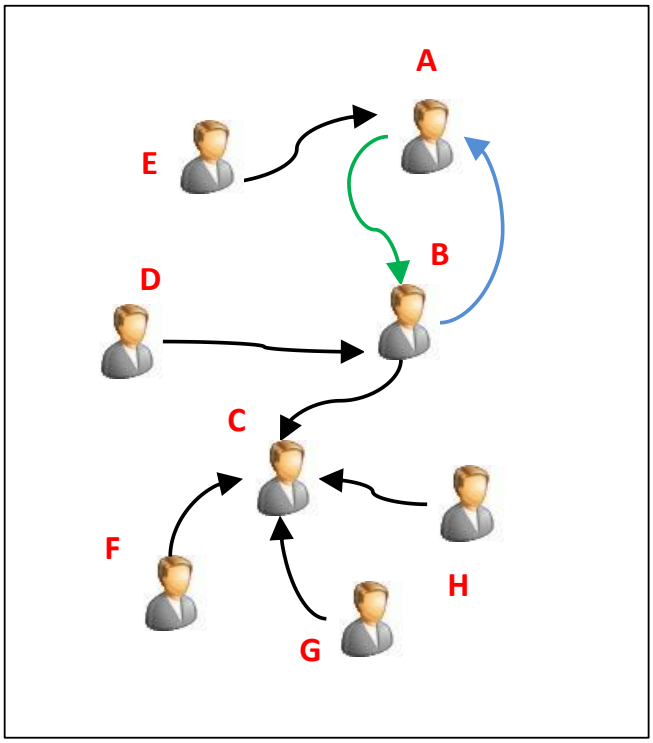


Figura 4-1: Topologia da rede de seguidores do *Twitter*

O *Twitter* possui pelo menos dois instrumentos para coordenar a comunicação, um é o mecanismo de *retweet* e o outro é a *hashtag* (etiqueta ou rótulo).

Quando um usuário publica um *tweet*, e outros usuários acham-no interessante ou relevante, eles retransmitem esse *tweet*, e o mecanismo é conhecido como *retweet*. O *retweet* é desta forma, uma mensagem que foi transmitida anteriormente e repassada a rede.

O processo de *retweet*, até pouco tempo, era feito de forma manual, e o usuários para distinguirem um *retweet* de um *tweet* precediam as mensagem pelos caracteres **Via** ou **RT**. Hoje o ambiente oferece um mecanismo automático de encaminhamento que tem o seguinte formato: Caracteres RT+simbolo arroba @+nome do usuário (usuário que publicou o *tweet* – o usuário que foi retuitado)+corpo da mensagem, que é exemplificado a seguir.

```
User: Very cool!! RT@TheNaturalNews:#Alzheimer's patients
treated by playing internet games: http://t.co/dSAmzTv.
```

A estrutura do *tweet* originário seria:

```
TheNaturalNews:#Alzheimer's patients treated by playing
internet games: http://t.co/dSAmzTv.
```

A mensagem ainda pode ter uma etiqueta (*hashtag*) precedida do caractér # (*jogo da velha*). Essas etiquetas designam um assunto abordado no *tweet* e assim auxiliam no

encontro de mensagens sobre determinado assunto, além da aproximação de pessoas que tenham interesses em comum. No exemplo acima a etiqueta é a palavra Alzheimer.

4.2 Modelo e Estrutura da Rede

O processo de *retweet* é uma prática muito comum entre atores no ambiente *Twitter*. Entretanto, esse processo é efetuado com certa parcimônia, pois nem todos os *tweet* são encaminhados. Os usuários são mais sistemáticos com relação ao conteúdo a ser encaminhado (KWAK *et al.*, 2010).

Quando um usuário repassa um *tweet*, ele está dando um endosso tanto à mensagem quanto ao usuário, na realidade, está creditando uma espécie de reputação, ao partilhar o *tweet* de terceiros com os seus próprios seguidores (ZAMAN *et al.*, 2010; HONG; DAN; DAVISON, 2011; WEITZEL; QUARESMA; DE OLIVEIRA, 2012).

O encaminhamento de *tweet*, ou seja, o processo de *retweet* faz com que exista uma "rede emergente e virtual" de curto e médio prazo sobreposta à rede formal de seguidores. Imagine o seguinte fluxo: ator B encaminha o *tweet* postado pelo ator A para a sua propria rede. A rede emergente está representada na Figura 4-2 agora com setas na cor preta, e a rede *Twitter* de seguidores por setas pontilhadas na cor vermelha. Observe que essas duas redes podem ter arcos coincidentes (por exemplo as ligações do ator C com B, F,G e H) ou não coincidentes (Ator E com ator B).

Baseado nas observações feitas nos paragrafos acima, se descreve aqui a estrutura e o modelo da rede. Daqui por diante chamaremos a "rede emergente" de "Rede de *Retweet*" ou simplesmente ***RT-net***.

A ***RT-net*** é modelada como um grafo direcionado onde os vértices são os atores ou usuários e os arcos são formadas pelo processo de *retweet*, isto é, um arco partindo do ator B para o ator A significa que ator B "*retuitou*" o ator A. Nomeou-se o ator B de usuário fonte e o ator A de usuário alvo.

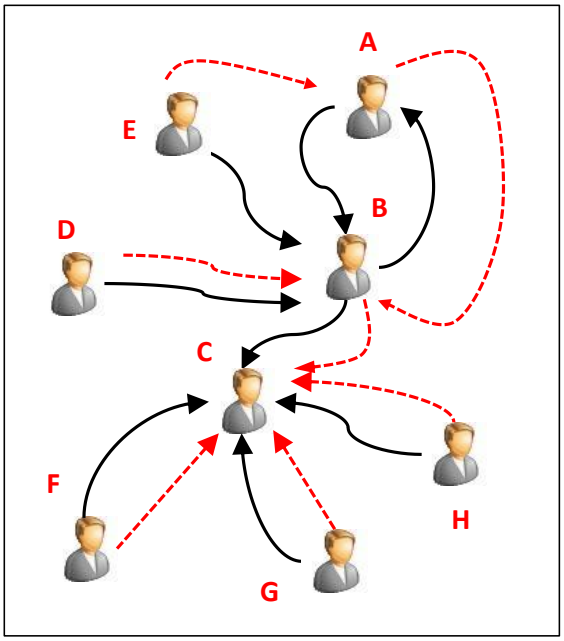


Figura 4-2: Rede de relacionamentos de *retweet* e de seguidores

Sabe-se que no português (do Brasil) não existe o verbo "retuitar", mas para efeitos de praticidade e melhor entendimento do texto, optou-se por conjugar conforme norma dos verbos terminados em "ar" da língua portuguesa.

Na Figura 4-2 tem-se que os atores B, F, G e H retuitaram o ator C, e os arcos formados são coincidentes com a rede de seguidores. O ator D retuitou B e os arcos são coincidentes e assim por diante. O ator E retuitou B, mas ele não é seguidor de B, ou seja, não faz parte da sua *ego-network*.

Na Tabela 4-1 tem-se a representação esquemática da rede ***RT-net***. O ator C se destaca nesta rede pois foi retuitado mais vezes (quatro vezes) em segundo lugar está o ator B com três *retweet* e o ator A com um *retweet* apenas.

Tabela 4-1: Relação de *retweet* por usuário

Usuário Fonte	Usuário Alvo	Relação
A	B	Seguidor
B	A	Seguidor
E	B	Sem relação
D	B	Seguidor
B	C	Seguidor
H	C	Seguidor
G	C	Seguidor
F	C	Seguidor

É comum pensar que quanto mais seguidores um ator tiver, maior é a tendência de que seu *tweet* seja repassado. Conforme dito anteriormente, deve haver outros fatores cognitivos e de comportamento envolvidos no processo de *retweet*, uma vez que nem todo *tweet* é retuitado (YANG *et al.*, 2012)

No ambiente do *Twitter* ocorre, muito frequentemente, o efeito "popularidade". Algumas celebridades, como atores, cantores, políticos, organizações e mídias em geral atraem naturalmente centenas de milhares de seguidores. Por exemplo, Lady Gaga, Britney Spears, Ashton Kutcher, Barack Obama, CNN (Cable News Network) entre outros possuem quase cinco milhões de seguidores cada um. Por consequência, a quantidade de *retweet* destas pessoas é maior que a de um usuário com poucos seguidores. Se utilizássemos apenas a quantidade de *retweet* para inferir a reputação de um ator, o fenômeno popularidade poderia ser gerar resultados falso-positivos.

Algumas investigações chamam a atenção para o fato de que quando o processo de *retweet* ocorre fora da *ego-network*, ao usuário alvo é creditada influência e ao *tweet* é creditado relevância (BONGWON *et al.*, 2010; CASTILLO; MENDOZA; POBLETE, 2011; ROMERO *et al.*, 2011). A ***RT-net*** é construída levando-se em consideração as observações descritas acima.

A ***RT-net*** é modelada como um grafo ponderado direcionado $\boldsymbol{G_{RT}}^{\rightarrow} = (\boldsymbol{N}, \boldsymbol{E}^{\rightarrow}, \boldsymbol{\mathcal{W}})$, com as seguintes propriedades:

1. Conjunto de vértices ou nós representando os usuários alvo e fonte

$$\boldsymbol{N} \equiv \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots\}$$

2. Conjunto de arcos representando o processo de *retweet*

$$\boldsymbol{E^{\rightarrow}} \equiv \{\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots\}$$

Se $\exists$ arco $\boldsymbol{e_k} = (\boldsymbol{v_i}, \boldsymbol{v_j}) \in \boldsymbol{E}$ de $\boldsymbol{v_i}$ para $\boldsymbol{v_j}$ significa que $\boldsymbol{v_i}$ "re*tweet*ou" $\boldsymbol{v_j}$. Sendo $\boldsymbol{v_i}$ o usuário fonte e $\boldsymbol{v_j}$ o usuário alvo.

3. Conjunto de pesos representando o grau ou a intensidade da relação entre usuário fonte e usuário alvo:

$$\boldsymbol{W}(\boldsymbol{e_k}) \equiv \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots\}$$

A função $\boldsymbol{W}(\boldsymbol{e_k})$ é definida como:

$$\boldsymbol{W}(\boldsymbol{e_k}) = \left(\frac{\sum \mathbf{RT}_{\boldsymbol{v_j}}}{\mathbf{RT_{total}}}\right) + \boldsymbol{\alpha} \qquad \text{Equação 4-1}$$

Onde o numerador $\sum \mathbf{RT}_{\boldsymbol{v_j}}$ é o número total de *retweet* de um usuário alvo $\boldsymbol{v_j}$ e o denominador $\mathbf{RT_{total}}$ é a quantidade total de re*tweet* da amostra.

O primeiro parâmetro da Equação 4-1 foi inspirado na Taxa de *Retweet* proposta por Anger e Kittl (2011) descrita no Capitulo 3 (seção 3.5).

O segundo parâmetro é um fator mitigador do efeito "popularidade" que algumas celebridades, políticos etc têm de atrair centenas de milhares de seguidores (YANG *et al.*, 2012). É natural que um usuário com maior número de seguidores tenha maior número de re*tweet* e esse fenômeno pode superestimar o calculo dos pesos. Desta forma, o parâmetro $\boldsymbol{\alpha}$ tem como objetivo descontar o fenômeno "popularidade" no computo geral do peso dos arcos e valorizar os *retweet* provenientes das redes *alter* dos usuários alvo.

O parâmetro $\boldsymbol{\alpha}$ é estimado de acordo com:

- $\boldsymbol{\alpha}$ = 0.9 se **não** existe relação entre usuário alvo e fonte, neste caso $\boldsymbol{\alpha}$ assume o maior valor,
- $\boldsymbol{\alpha}$ = 0.1 em todos os outros casos e $\boldsymbol{\alpha}$ assume o menor valor.

Os valores do parâmetro $\boldsymbol{\alpha}$ foram estimados levando-se em consideração o percentual de cada relação observado na amostra. Esses percentuais encontram-se descritos na Seção 5.2 do Capitulo 5 ilustrados na Figura 5-2.

Os pesos $\boldsymbol{W}(\boldsymbol{e_k})$, em ultima instância, representam a intensidade do relacionamento. A intensidade está diretamente relacionada à relevância ou reputação que é atribuída ao usuário alvo. Considera-se que o mecanismo de *retweet* pode ser interpretado como uma forma de endosso da reputação tanto da mensagem quanto do usuário-alvo.

Na Tabela 4-2 tem-se exemplificado a rede descrita acima (Tabela 4-1) com os pesos calculados de acordo com a Equação 4.1. Verifica-se que apesar do usuário alvo C ter sido retuitado mais vezes, o fenômeno popularidade foi minimizado pelo parâmetro $\boldsymbol{\alpha}$.

Tabela 4-2: Tabela de arcos e pesos

Usuário Fonte	Usuário Alvo	Relação	Pesos
A	B	Seguidor	1/8 + 0,1 = 0,225
B	A	Seguidor	3/8 + 0,1 = 0,475
E	B	Sem relação	3/8 + 0,9 = 1,275
D	B	Seguidor	3/8 + 0,1 = 0,475
B	C	Seguidor	4/8 + 0,1 = 0,6
H	C	Seguidor	4/8 + 0,1 = 0,6
G	C	Seguidor	4/8 + 0,1 = 0,6
F	C	Seguidor	4/8 + 0,1 = 0,6

Para avaliação da topologia da rede ***RT-net***, foi construída uma rede ***RT-base*** cujos pesos dos arcos são binários, ou seja, $\boldsymbol{w}\ (\boldsymbol{e_k}) = 1$.

4.3 Metodologia para estimativa da Reputação

Baseado nas discussões feitas no Capítulo 3, seção 3.7, avaliação da reputação não tem significado se não se define um propósito e um contexto no qual esteja inserida. Além disso, avaliar a reputação de um vértice baseado apenas na posição que ocupa na rede comporta uma visão global. Nesta pesquisa, busca-se também a avaliar a influência local deste vértice por meio da análise do grau de interação entre os usuários.

Neste sentido a estimativa da reputação é baseada tanto nos aspectos globais quanto locais dentro de um contexto. Assim, a Reputação $\mathcal{R}\mathrm{a}\mathcal{R}_{\mathcal{v}_j}$ (**Ra**nk **R**eputation) é estimada interativamente pela Equação 4-2, como uma média ponderada de um conjunto de métricas.

$$\mathcal{R}\mathrm{a}\mathcal{R}_{\mathcal{v}_j} = \frac{\sum_{i=1}^{n} m_{ij} w_i}{Max\ \{m_{ij} w_i\}_{i=1}^{n}} \qquad \text{Equação 4-2}$$

$$0 < \mathcal{R}\mathrm{a}\mathcal{R}_{\mathcal{v}_j} < 1, \quad \sum_{i=1}^{\mathrm{n}} \mathrm{w}_i = 1, \quad j = 1, .., \mathrm{m}$$

A Reputação $\mathcal{R}\mathrm{a}\mathcal{R}_{\mathcal{v}_j}$ é função de **(** M, W) com as seguintes propriedades:

1. $M = \{\, \mathrm{m}_{1\mathrm{j}}, .., \mathrm{m}_{\mathrm{n}} \}$ um conjunto de medidas de centralidade $\{\, \mathcal{D}c, \mathcal{B}c, \mathcal{C}c, \mathcal{E}c, \mathrm{Prank}, d_{\mathrm{in}}, d_{\mathrm{out}} \}$, Onde:

$\mathcal{D}c$ = Grau,
$\mathcal{B}c$ = Centralidade de Intermediação,
Cc = Centralidade de Proximidade,
$\mathcal{E}c$ = Centralidade de Autovalor,
Prank = Page Rank,
d_{in} = Grau de Entrada,
d_{out} = Grau de Saída.

2. $W = \{ w_1 , \ldots , w_n \}$ é um conjunto de pesos, onde:

   Restrição: $\forall w_i \in W, \sum_{i=1}^{n} w_i = 1$ e $w_i \subset \{0.0, 0.1, 0.2, \ldots 1.0\}$,

   É possível que $\exists w_k = 0 \mid m_k * w_k = 0$.

4.3.1 Avaliação do desempenho da metodologia

Buscou-se no domínio da Recuperação de Informação medidas que pudessem avaliar o desempenho da metodologia proposta. A precisão em **n** (P@n) mede a relevância dos ***n*** primeiros documentos de uma lista ordenada, calculado por:

$$P@n = \frac{r}{n} \qquad \text{Equação 4-3}$$

Onde, ***n*** é o número de documentos retornados e ***r*** é o número de documentos considerados relevantes e retornados até a posição ***n*** da lista ordenada. Por exemplo, se os 10 primeiros documentos retornados por uma consulta são:

{relevante, irrelevante, irrelevante, relevante, relevante, relevante, irrelevante, irrelevante, relevante, relevante}

Então os valores de P@1 até P@10 são:

{1, 1/2, 1/3, 2/4, 3/5, 4/6, 4/7, 4/8, 5/9, 6/10}

A segunda medida de desempenho utilizada é a Precisão Média (AP – Average Precision) que é definida como uma média para todos os valores de *P@n* para todos os documentos relevantes, que é dada pela equação:

$$AP = \frac{\sum_{n=1}^{m} P@n * rel\ (n)}{r} \qquad \text{Equação 4-4}$$

Onde, ***r*** é o número total de documentos considerados relevantes; ***m*** é o número de documentos recuperados, e ***rel(n)*** é uma função binária sobre a relevância do ***n^{ésimo}*** documento, ou seja,os resultados possíveis são relevante ou não relevante.

Para estimar o desempenho é necessário que os documentos sejam classificados quanto a sua relevância. Nesta pesquisa é empregada a relevância binária (relevante e irrelevante). O parâmetro de relevância segue a seguinte estratégia: usuários relevantes são aqueles que pertencem estritamente à área de saúde e ao mesmo tempo tenham suporte financeiro público. De acordo com a HON[14] a presença de informações enviesadas é devida, em parte, ao financiamento privado ou à presença de propaganda de laboratórios e correlatos. Essa classificação foi feita de forma manual analisando-se o perfil e o sítio associado a este perfil em um total de 1232 usuários, foram classificados 212 usuários relevantes.

4.3.2 Algoritmo para cálculo do $\mathcal{R}a\mathcal{R}$

Nesta seção é descrito o algoritmo para o cálculo da reputação $\mathcal{R}a\mathcal{R}$ de todos os vértices. O algoritmo leva em consideração o melhor desempenho, ou seja, que o valor da AP é aproximadamente 100%, ou algum valor próximo. Os passos do algoritmo são descritos nos parágrafos posteriores.

[14] http://www.hon.ch

No primeiro Passo, é selecionado um subconjunto $\boldsymbol{W'}$ (de 7 elementos) do conjunto $\boldsymbol{W}$ de pesos, a seleção sendo feita de forma randômica. A única condição imposta é que a soma de todos os pesos de $\boldsymbol{W'}$ seja igual à unidade (descrita pela equação 4-2). Cada peso que foi selecionado é associado a uma métrica do conjunto **M**. E assim, são formados os pares de pesos e métricas. O conjunto destes pares (sete ao todo) são armazenados em uma matriz $\boldsymbol{T_{k,j,l}}$. Cada linha desta matriz representa uma estratégia de associação (peso-métrica) a ser testada.

Um exemplo de estratégia poderia ser conforme ilustrado na Equação 4-5. Neste caso, as métricas Centralidade de Proximidade (Cc) e o Grau de Entrada (D_{in}) e Saída (D_{out}) e seus respectivos pesos (0,7; 0,2 e 0,1) foram testados.

$$((0{,}7 * Cc) + (0{,}2 * D_{in}) + (0{,}1 * D_{out})) \quad \text{Equação 4-5}$$

No Passo 2 é verificado se a estratégia de associação já foi utilizada. Caso seja verdadeiro, volta-se ao passo 1. Caso contrário, calcula-se a reputação - $\mathcal{R}a\mathcal{R}$ (pela equação 4-2) de todos os vértices, e gera-se uma lista $\mathbb{L}_{\boldsymbol{p}}$ ordenada descendentemente dos valores $\mathcal{R}a\mathcal{R}$ encontrados.

No inicio do Passo 3 verifica-se se o melhor desempenho foi encontrado. Caso a sentença seja verdadeira, o algoritmo chega ao fim e mostra o resultado final: AP (melhor), P@n ($\mathcal{R}a\mathcal{R}$), $\boldsymbol{L_p}$ (lista ordenada) e a melhor estratégia de associação - T_k (os valores dos pesos e as métricas que foram utilizadas) fechando o passo 4. Caso contrário, voltar ao passo 1.

O maior valor de AP retornado significa que a listagem correspondente apresenta os 212 usuários relevantes próximos ao topo da listagem. Um valor de AP = 100% que dizer que todos os usuários relevantes ocupam as primeiras 212 posições da lista.

Passo 1	**Enquanto** AP(melhor)não encontrada faça **Para cada** métrica $m_{ij} \in M$ **Selecionar** randomicamente um subconjunto $\boldsymbol{W} \subset \boldsymbol{W}$ tal que $\sum_{i=1}^{n} \mathbf{w}_i = 1$ onde $\boldsymbol{n} = 7$ **Selecionar** $m_{ij} \in \boldsymbol{M}$ **Associar e Armazenar** os pares $(\boldsymbol{m_i}, \boldsymbol{w_i})$ em uma matriz $T_{k,j,l} = \{(\boldsymbol{m_i}, \boldsymbol{w_i}); \ldots; (\boldsymbol{m_n}, \boldsymbol{w_n})\}$ **Fim para**
Passo 2	**Se** $T_k \subset T_{k,j,l}$ **Voltar** Passo 1 **Senão** **Para cada** vértice **Calcular** $\mathcal{R}a\mathcal{R}$ **Fim para** **Criar** $\boldsymbol{L_p}$ **Calcular** AP($\mathcal{R}a\mathcal{R}$) e P@n($\mathcal{R}a\mathcal{R}$) de $\boldsymbol{L_p}$ **Fim se**
Passo 3	**Se** AP($\mathcal{R}a\mathcal{R}$) > AP(melhor) AP(melhor) ← AP($\mathcal{R}a\mathcal{R}$) **Senão** **Voltar** ao passo 1 **Fim se** **Fim enquanto**
Passo 4	**Retorne** AP(melhor), P@n($\mathcal{R}a\mathcal{R}$), $\boldsymbol{L_p}$ e T_k

Serão abordados três estudos comparativos, o primeiro estudo é metodologia proposta nesta pesquisa, o segundo estudo consideram as métricas na literatura (baseadas nos trabalhos correlatos discutidos ao final do Capítulo 3) e por fim um estudo utilizando-se as métricas isoladamente para estimar a reputação. Assim, estes estudos serão chamados de:

a) Metodologia proposta
b) Métricas da literatura
c) Métricas isoladamente

5 COLETA DE DADOS E PRINCIPAIS RESULTADOS

Neste capitulo apresenta-se a coleta dos dados em sua primeira seção, Na segunda seção as análise estatísticas dos resultados, em seguida é apresentada a análise da estrutura, da topologia, analise dos resultados. Ao final do capitulo tem-se uma discussão sobre os achados verificados.

5.1 Coleta dos dados

Para a coleta dos dados foi desenvolvido um *crawler*[15] que pudesse extrair as dados necessários a essa proposta. A extração dos dados ocorreu em março-abril de 2011.

A metodologia da coleta obedeceu aos critérios discutidos a seguir. Primeiro foram selecionados randomicamente 152 usuários "sementes" que possuíam interesse na área de saúde conforme declarado em seus perfis. Destes usuários "sementes" foram coletados os *retweet* postados durante o período de coleta mencionado. Esse método de amostragem é uma adaptação do método *snow-ball* proposto por Goodman (GOODMAN, 1961). O método de amostragem *snow-ball*, originariamente aplicado à sociologia é utilizado em populações desconhecidas ou raras. A obtenção de uma amostra de uma população normalmente não permite o uso de metodologias tradicionais de amostragem que exigem que toda a população seja conhecida (por exemplo, a população de estudantes de uma universidade). A amostragem *snow-ball* pode proporcionar abrangentes caracterizações (embora não generalizáveis) de populações desconhecidas. O processo de amostragem é relativamente simples, opera da mesma maneira que uma bola de neve que vai agregando mais e mais flocos de neve, é um processo multi-passo em que cada vez mais pessoas são adicionadas à amostra a cada passo. Normalmente, o passo inicial envolve a identificação de um grupo de indivíduos que são conhecidos membros da população para criar uma "semente".

5.2 Análise estatística dos dados

Foram extraídos apenas os *retweet* em seu formato padrão, conforme descrito e apresentado na seção 4.1 do Capitulo 4, em um total de 4662 *retweet*. O tamanho da amostra é considerado adequado de acordo com os achados de Kwak e colaboradores (2011). Os autores observaram que o processo de *retweet* é feito com parcimônia, ou seja, nem todos os *tweet* são encaminhados. Os autores observaram ainda que parece haver critérios cognitivos ou de confiança na seleção dos *tweet* que serão passados adiante. Observaram também que em média tem-se que um usuário é retuitado de 100 a

[15] Desenvolvido pelo Grupo de Modelagem Computacional da Faculdade de Computação (ARAUJO; RODRIGUEZ, 2011).

1000/mês, esse valor é dependente do perfil do usuário. Por exemplo, quanto maior o numero de seguidores (celebridades, jornais, revistas entre outros) maior é a tendência de ser mais vezes retuitado.

Do corpo de texto dos *retweet* foram extraídos novos usuários ao total de 1080 que somados aos 152 soam um total de 1232 usuários, sendo 1171 usuários alvo (aqueles que foram retuitados) e 61 usuários fonte. Observou-se que maioria (63%) dos usuários alvo tem apenas 1 (um) *retweet*. Na Tabela 5-1 tem-se a estatística da média, desvio padrão, valores mínimo e máximo da amostra. Em média, houve 3,0 (três) *retweet* por usuário alvo. O Coeficiente Gini da distribuição dos re*tweet* apresenta certo grau de desigualdade, com valor em torno de 64%. Esse achado representa que o processo de retweet não é um mecanismo comumente utilizado pelos usuários.

Tabela 5-1: Estatísticas da amostra de *retweet*

RT	
Média	3,7
Desvio Padrão	17,4
Mínimo	1
Máximo	527

A média de *retweet* por usuário fonte (aqueles que encaminharam os *tweet*) foi de 76,4 e desvio padrão de 141,6, sendo o mínimo valor igual a 2 e o máximo igual a 834 *retweet* (Tabela 5-1). O Coeficiente Gini é igual à 63% mostrando a desigualdade na sua distribuição.

Por questões de privacidade, e seguindo as recomendações do ambiente *Twitter*, os nomes de todos os usuários foram ocultados. Na Figura 5-1 tem-se a distribuição de re*tweet* por usuário fonte, pode-se ver que maioria publicou menos de 100 *retweet*.

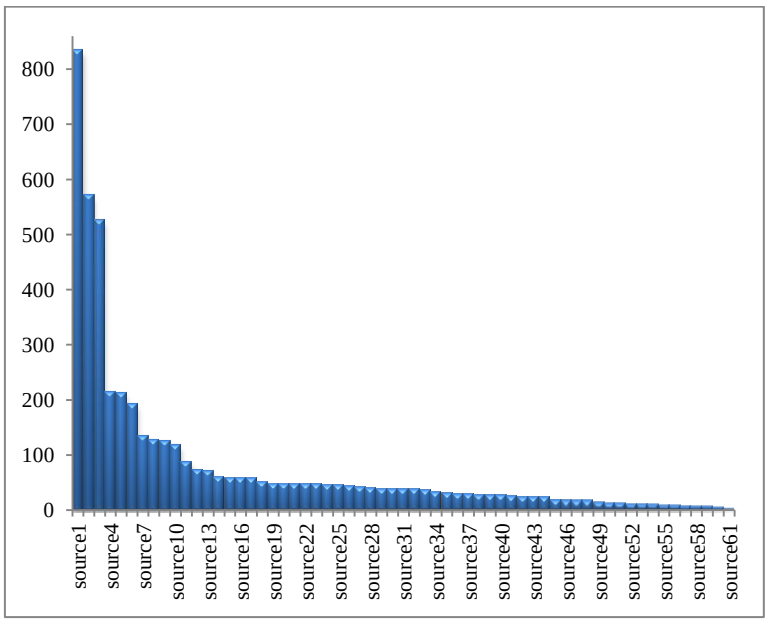


Figura 5-1: Distribuição de *retweet* por usuário fonte

A Figura 5-2 mostra o gráfico de porcentagens das relações entre usuários *alvo* e *fonte* na amostra. A maioria das relações é do tipo seguido ou *follower* (64%), conforme esperado, pois é relação estabelecida por padrão no ambiente *Twitter*. Existe um equilíbrio nas distribuições das porcentagens das relações *following* ou seguidos (15%) e da relação *friendship* (14%) de reciprocidade. O menor percentual (≈ 7%) foi encontrado entre aqueles que não possuem relação direta.

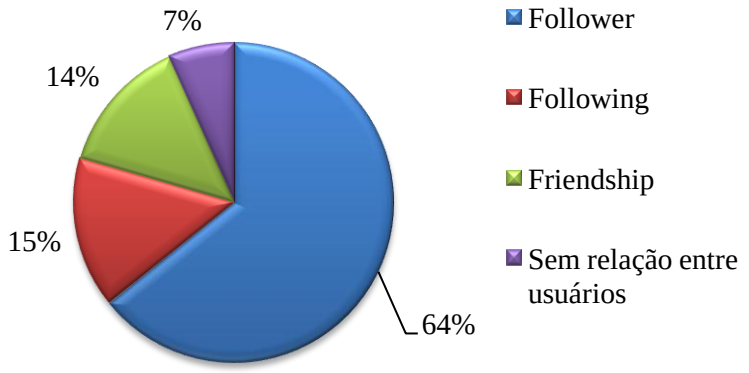


Figura 5-2: Gráfico de porcentagens de relações no conjunto de dados

De acordo com o observado, a maioria das relações é do tipo seguidor. As celebridades, mídias têm como característica atrair grande quantidade de seguidores. Essa particularidade pode enviesar (ou pior superestimar) a estimativa da reputação devido ao efeito popularidade. Sendo assim, para mitigar este efeito foi proposto um parâmetro mitigador (α). As relações seguidores, seguidos e de relações de amizade juntas têm um percentual de aproximadamente 90% e os que não apresentam relação representam 10%. Desta forma, com intuito de dar maior peso aos usuários que *retuitaram* e que não possuem relação com usuários alvo, o valor de α será de 0,9 e em todos os casos contrários $\alpha = 0{,}1$.

Com objetivo de validar a metodologia de ordenamento proposta, analisou-se o perfil de cada usuário, e a partir dos achados verificados, estes foram separados em grupo:

a. Saúde Privada: são as clinicas, hospitais, associações médicas, organizações não governamentais, associação de classes, grupos de apoio, e que possuem financiamento privado;
b. Saúde Pública: são órgãos públicos no domínio de saúde em geral, são hospitais, clínicas que possuem financiamento público; Esses usuários são os usuários considerados relevantes para a avaliação;
c. *Blogs* e outras mídias sociais sobre saúde: que disseminam informações na área de saúde, ou mantém sítios na área de saúde;
d. Congressos: algumas contas de usuários foram criadas para divulgar eventos na área de saúde;

e. Não relacionados à saúde: são órgãos públicos que não estão relacionados à área de saúde, mas que publicaram ou disseminaram informações na área de saúde;

f. Mídias: do tipo jornais, revistas, televisão e *blogs* que publicam assuntos de saúde dentre outros assuntos;

g. Celebridades: como atores, cantores e etc que podem eventualmente publicar ou disseminar assuntos na área de saúde;

Em alguns casos, por falta de informação detalhada do perfil, não foi possível classificação. Optou-se desta forma, por criar um terceiro grupo denominado não-categorizado. Verificou-se também que havia três usuários com a conta encerrada além de dezesseis erros na coleta. Considerou-se como erro de coleta porque não foi possível encontrar o perfil do usuário no ambiente *Twitter*. Como não foi utilizado nenhum mecanismo para tratamento de spam, este fator pode ter favorecido o aparecimento de contas "fantasmas".

A Figura 5-3 ilustra o percentual das categorias verificadas na amostra. Observou-se que 32% das contas de usuários não estão relacionados à área de saúde. O segundo maior percentual (26%) está relacionado aos *blogs*. O terceiro maior percentual (17%) são usuários que pertencem à área de saúde e possuem financiamento público (categoria Saúde Pública). As mídias apresentaram um percentual de aproximadamente 13%. A categoria Saúde Privada representa 8% da amostra. O restante está distribuído entre os percentuais 0,24% e 1,14%. Os não-categorizados representam apenas 0,41% e os erros cometidos representam 0,24%.

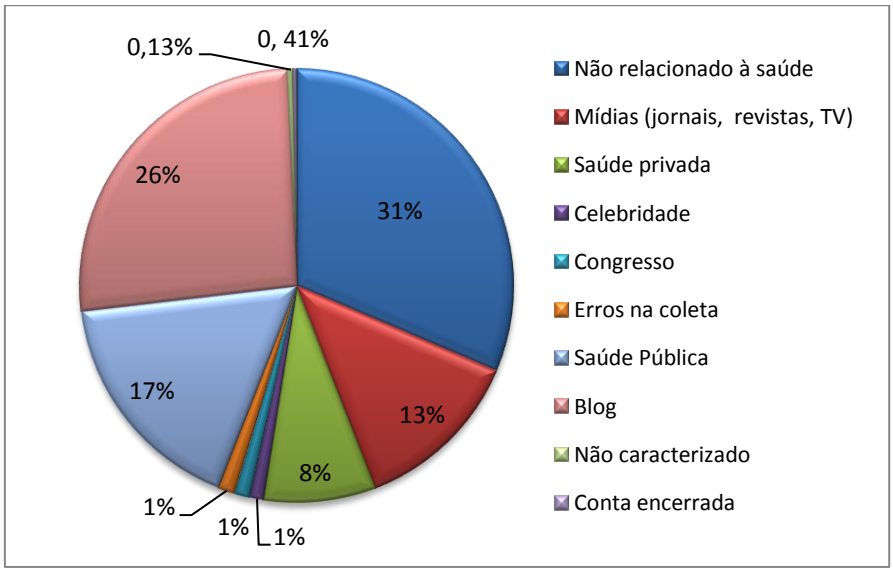


Figura 5-3: Analise do perfil de todos os usuários

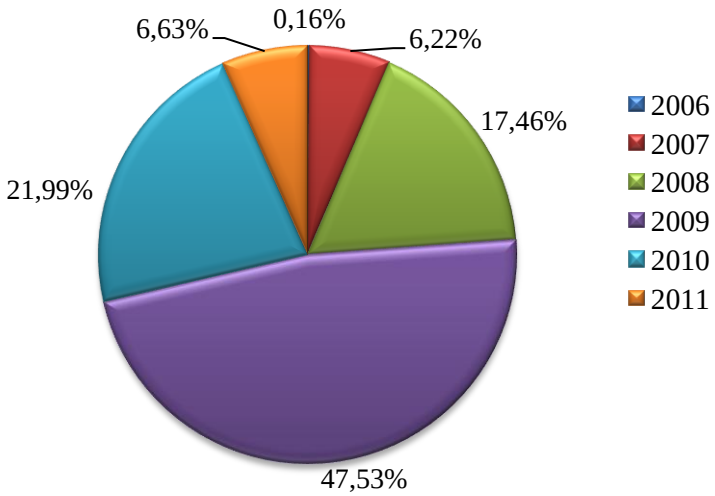


Figura 5-4: Percentual de data de ingresso no Ambiente *Twitter*

Na, Figura 5-4 tem-se o gráfico das porcentagens referente às datas de ingresso no ambiente *Twitter*. A maioria (~ 48%) dos usuários ingressou no ano de 2009. O menor percentual está relacionado ao ano de 2006 com apenas 0,16%, ano de inauguração da plataforma *Twitter*.

Tabela 5-2: Estatística da amostra

	Following	***Follower***	***Tweet***
Média	2592	135463	8219
Desvio Padrão	12204	1016490	15913
Máximo	221077	22439192	187264
Mínimo	0	0	0

A Tabela 5-2 ilustra os valores das médias, mínimo e máximo para todos os usuários da amostra. Percebe-se o fenômeno "popularidade", isto é, a característica que alguns usuários têm de atrair centenas de milhares de seguidores, e a grande maioria (89,94%) com até 135.463 seguidores, e 50% tem menos de 2.100 seguidores, conforme ilustra a Tabela 5-3. Na Tabela 5-4 tem-se o perfil dos 10 primeiros posicionados em uma lista ordenada (*rank*) descendente, isto é, os 10 maiores valores verificados para seguidos, seguidores, *tweet e retweet* (na perspectiva do usuário alvo). É interessante perceber que no parâmetro seguidores e seguidos as categorias Não-categorizado e *Blog* ocupam posição de destaque, isto é, representam 40% da amostra. Por outro lado, a categoria que obteve maior quantidade de *tweet* foi justamente àquela relacionada à categoria Saúde Publica. Verificou-se que o perfil com maior porcentagem (50%) de *retweet* (em relação ao usuário alvo) é da categoria Saúde-pública.

Tabela 5-3: Tabela de frequência e percentual acumulado de Seguidores

Bloco	Freqüência	% Cumulativo
529	348	28,25%
2.117	256	49,03%
8.466	266	70,62%
33.866	158	83,44%
135.463	80	89,94%
1.151.954	96	97,73%
2.168.444	14	98,86%
3.184.934	3	99,11%
4.201.424	4	99,43%
5.217.914	2	99,59%
6.234.404	1	99,68%
7.250.894	0	99,68%
8.267.384	1	99,76%
9.283.874	0	99,76%
10.300.364	0	99,76%
11.316.854	0	99,76%
12.333.344	1	99,84%
13.349.835	0	99,84%
14.366.325	0	99,84%
15.382.815	0	99,84%
16.399.305	0	99,84%
17.415.795	0	99,84%
18.432.285	0	99,84%
19.448.775	0	99,84%
20.465.265	1	99,92%
22.439.192	1	100,00%
Total	1232	100,00%

Tabela 5-4: Perfil das 10 primeiras posições de cada parâmetro

Seguidos	**%**
Não-Caracterizado	**40**
Saúde-Publica	**20**
Saúde-Privada	**10**
Mídia	**10**
Não-Caracterizado	**10**
Blog	**10**
Ao total	**100**

Seguidores	**%**
Blog	**40**
Não-Categorizado	**20**
Mídia	**10**
Saúde-Publica	**10**
Celebridade	**10**
Saúde-Privada	**10**
Ao total	**100**

Tweet	**%**
Saúde-Publica	**40**
Mídia	**20**
Blog	**20**
Não-Categorizado	**10**
Saúde-Privada	**10**
Ao total	100

***Retweet* alvo**	**%**
Saúde-Publica	**50**
Blog	**20**
Mídia	**20**
Não-relacionado à saúde	**10**
Ao total	100

Em estatística é comum a realização de testes de ajuste de curvas de sistemas reais. Os resultados são obtidos na forma de pontos cujo comportamento demonstra o relacionamento de uma variável independente (ou explicativa) com uma, ou mais, variáveis dependentes. O gráfico destes pontos é chamado de diagrama de dispersão.

Uma das vantagens de se obter uma curva que se ajusta adequadamente a estes pontos é a possibilidade de prever os valores da função (variável dependente) para valores da variável explicativa que estão fora do intervalo fornecido. Ou seja, é possível fazer uma extrapolação com uma aproximação razoável.

A Figura 5-5 mostra o diagrama de dispersão dos *retweet vs tweet*. Pode-se verificar que não existe uma relação linear entre as variáveis e a correlação entre estas variáveis é de ≈ 7%.

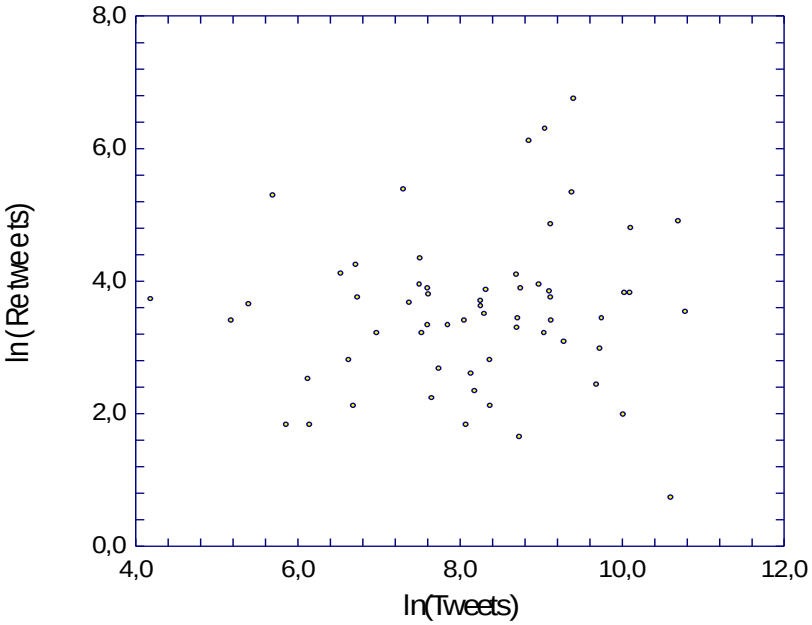


Figura 5-5: Gráfico de dispersão dos parâmetros *Retweet vs Tweet*

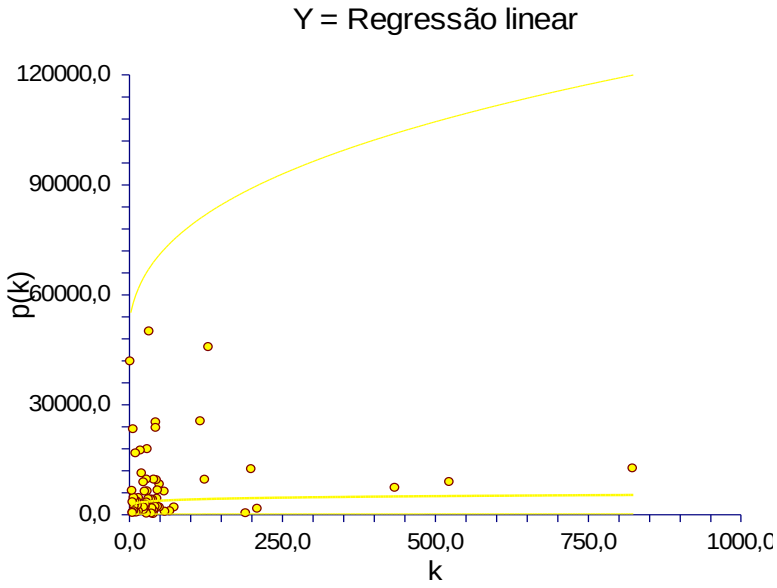


Figura 5-6: Modelo de regressão linear e a quadrática: LN(*tweet*)=A+B*(LN(re*tweet*))

Modelo estimado

Parâmetro Nome	**Parâmetro Estimado**	**Erro**	**Inferior 95% FC**	**Superior 95% FC**
A	7,77761	0,59937	6,57784	8,97738
B	0,12169	0,16329	-0,20518	0,44855

R^2 = 0,009484, Modelo: LN(*tweet*)=A+B*(LN(re*tweet*)); FC = Fator de Confiança do modelo.

Não foi possível encontrar um ajuste através do modelo de regressão linear ou quadrática (teste de ajuste de curva), observando os valores dos resíduos na Figura 5-7 e o pequeno valor do R^2. Um valor de R^2 próximo da unidade indica uma forte relação

entre as duas variáveis. A análise do gráfico dos resíduos com os valores ajustados deve apresentar pontos dispersos aleatoriamente, sem nenhum padrão definido. Todavia, os resíduos seguem uma tendência linear, ou seja, a variância do erro ε_i cresce com os valores da variável *retweet*. Essa característica indica que os parâmetros não se ajustam em um modelo linear.

Nas Figura 5-8 e Figura 5-9 (para melhor visualização, ambos os eixos estão em potencia de 10) tem-se os gráficos de dispersão e de ajustes dos parâmetros seguidor e seguido *vs retweet*. Não foi possível estimar uma curva de ajuste que representasse a distribuição, o valor do R^2 ficou em torno de 4,4E-4 e 0,009 respectivamente.

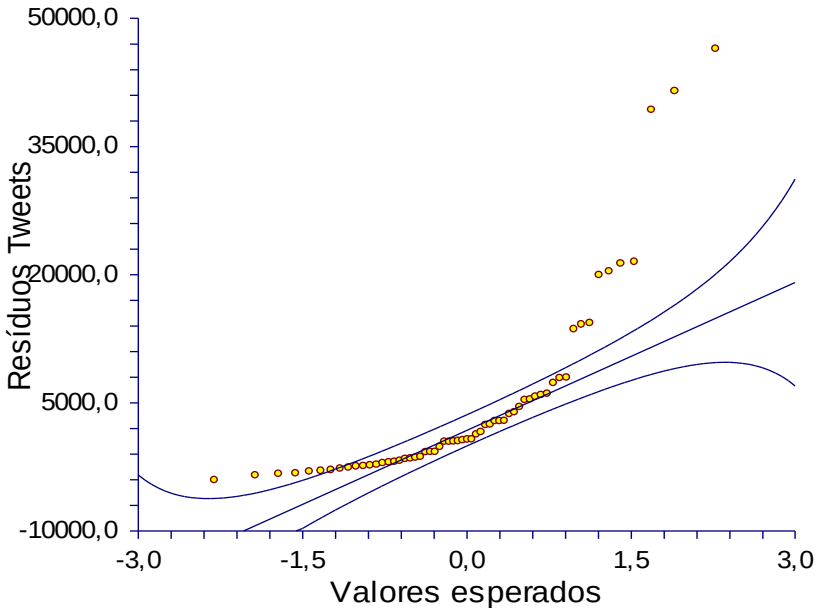


Figura 5-7: Gráfico de dispersão dos resíduos (*tweet*) *vs* valores esperados (*Retweet*)

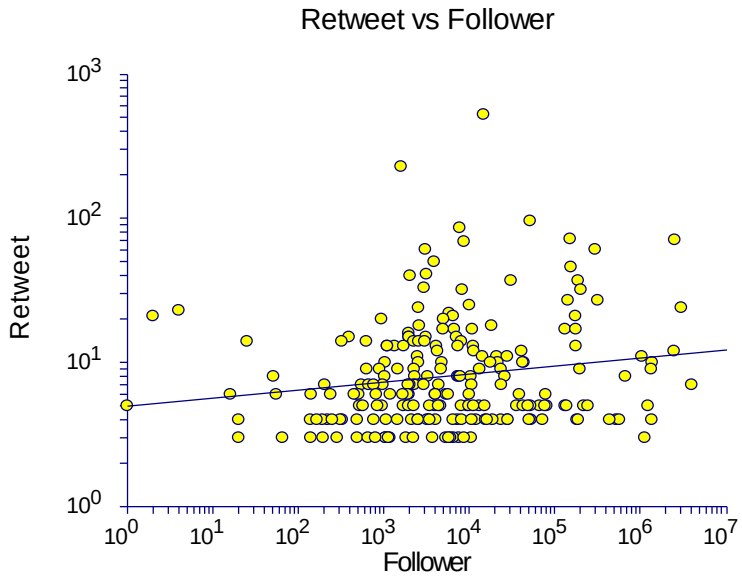


Figura 5-8: Gráfico de dispersão dos parâmetros seguidores vs *retweet*

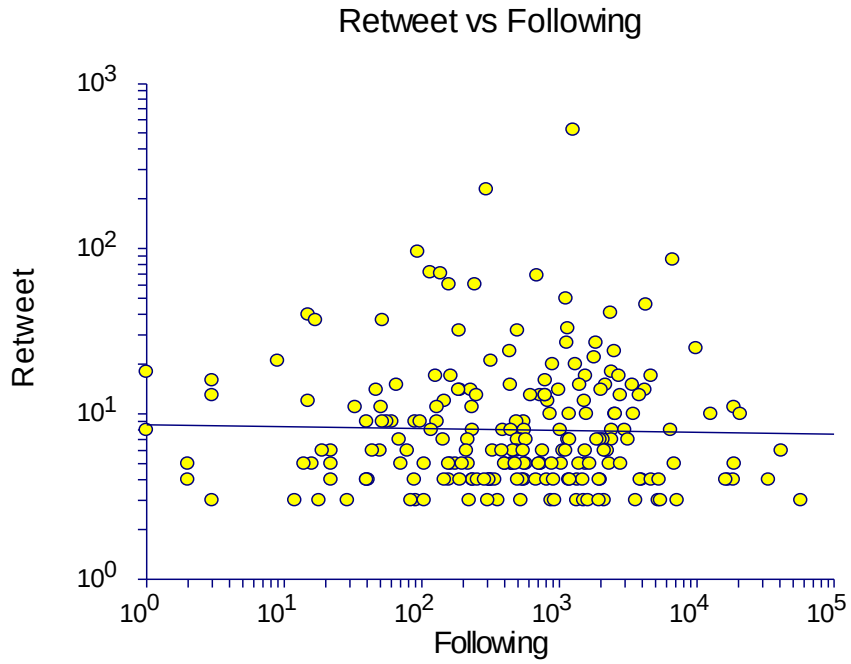


Figura 5-9: Gráfico de dispersão dos parâmetros seguidos vs *retweet*

Desta forma, foi verificado que os parâmetros *retweet*, seguidores, seguidos e *tweet*, não apresentam uma correlação, ou uma dependência entre si. Isso pode indicar que o Mecanismo de *retweet* obedece a outros fatores, além dos estabelecidos nas relações entre os usuários.

Para o cálculo do conjunto das métricas $M = \{ m_1 , .. , m_n \}$ foi utilizado o *software* Gephi desenvolvido pelo Consórcio Gephi (https://gephi.org/). É um programa de código aberto com livre distribuição. Para a visualização da rede ilustrada na Figura 5-10 utilizou-se o ORA[16] - *Organization Risk Analyzer* desenvolvido pela Universidade Carnegie Mellon (Pittsburgh - EUA). É também um programa de código aberto com livre distribuição. A Figura 5-10 ilustra apenas uma parte da estrutura da rede após ter sido submetida ao algoritmo de Girvan-Newman (GIRVAN; NEWMAN, 2002; NEWMAN; GIRVAN, 2004) que detecta comunidades pelo método de agrupamento hierárquico.

[16] http://www.casos.cs.cmu.edu/projects/ora/

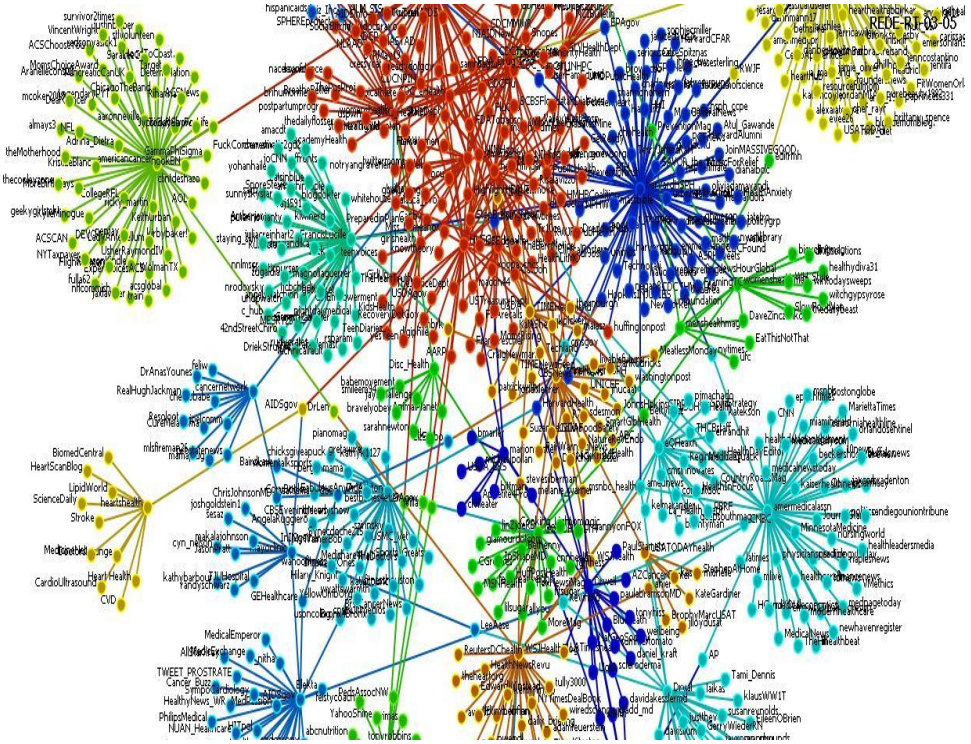


Figura 5-10: *RT-net* modelada pelo algoritmo hierárquico de agrupamento

A rede ***RT-net*** possui ao total 1.232 vértices e 1.411 arcos. Os pesos dos arcos foram calculados de acordo com:

$$\boldsymbol{W}\,(\boldsymbol{e_k}) \;=\; \left(\frac{\sum \mathbf{RT}_{\boldsymbol{V_j}}}{\mathbf{RT_{total}}}\right) + \;\boldsymbol{\alpha} \qquad \text{Equação 4-1}$$

E os valores variaram conforme ilustrado na Equação 5-1 ilustrada abaixo:

$\boldsymbol{W}\,(\boldsymbol{e_k})$ **é** [a , b] = { $\boldsymbol{W}\,(\boldsymbol{e_k}) \in \mathbb{R} \;/\; a \leq \boldsymbol{W}\,(\boldsymbol{e_k}) \leq b$} = [1,140 ; 0,071] Equação 5-1

5.3 Análise da estrutura da rede *RT-net*

Nesta seção são apresentados os resultados da análise da estrutura da rede ***RT-net***. Primeiramente apresenta-se as medidas, propriedades gerais e as propriedades estruturais. Em seguida faz-se uma analise estatística dos parâmetros, e por fim os resultados da metodologia de ordenação.

Na Tabela 5-5, tem-se os principais valores computados da rede ***RT-net***. Percebe-se que o diâmetro da rede **D** é pequeno, característica de Redes Aleatórias. A densidade ***de*** é a razão entre o número de conexões existentes pelo número de possíveis conexões. O valor encontrado é pequeno, indicando que a rede é pouco densa. Isso que dizer que os nós não estão altamente conectados tal como em uma Rede Regular. Essa característica pode ser observada também pelo pequeno valor encontrado na métrica grau médio <***k*** >. Cabe ressaltar que redes densas não possuem comunidades.

O coeficiente de agrupamento (ou coeficiente de aglomeração) médio < C > expressa a probabilidade de dois vértices que estão conectados possuírem uma conexão em comum com um terceiro vértice. O valor 0,017 indica que existem poucos vértices

vizinhos altamente conectados. É uma rede pouco fragmentada com estrutura hierárquica, possuindo 430 tríades e 260 cliques.

Tabela 5-5: Valores das medidas e propriedades da rede ***RT-net*** e ***RT-base***

Medidas e propriedades	**Valores**
Diâmetro ***D***	8
Tamanho **E** (número de arcos)	1409
Vértices **N**	1232
Densidade ***de*** [0,1]	0,001
Grau máximo - mínimo de entrada	13 - 0
Grau máximo - mínimo de saída	129 - 0
Coeficiente de Agrupamento Médio < **C** >	0,017
Grau Médio < ***k*** >	1,139
Caminho Mínimo Médio < ***l*** > *average shortest path*	3,573
Fragmentação [0,1]	0,2598
Hierarquia [0,1]	0,9884
Cliques	260
Tríade	430

Na Tabela 5-6, tem-se os valores do Coeficiente de Gini calculados para as métricas Centralidade de Intermediação, de Proximidade, de Autovalor e Grau Total. Este coeficiente foi proposto inicialmente para medir a desigualdade de uma distribuição de renda, mas pode ser utilizada para qualquer outro tipo de distribuição. Seus valores variam de 0 (completa igualdade) a 1 (completa desigualdade).

Observa-se que a Centralidade de Intermediação (***Bc***) e ***Prank*** estão próximos à completa igualdade, ou seja, apresentam distribuição uniforme. O contrário ocorre com as métricas C. de Proximidade (***Cc***) e C. de Autovalor (***Ec***), e nada pode-se afirmar sobre a distribuição do Grau Total na avaliação da rede ***RT-net***.

Para a rede ***RT-base***, os valores do Coeficiente Gini para as métricas ***Bc***, ***Prank*** e ***Ec*** apresentam completa desigualdade enquanto que a ***Cc*** e ***Prank*** têm parcial igualdade.

A correlação de Spearman um teste não paramétrico utilizado para calcular correlação em amostras ordenadas (listas). Dada duas amostras de observação ordenáveis, substitui-se cada um dos seus valores pela sua ordem na listagem. Por exemplo, seja uma lista ordenada na seguinte forma: x = 0,2, y =0,4 e z = 0,3, a substituição feita é x=1, y=3 e z=2 e com estes valores são calculadas as correlações. Da mesma forma, o coeficiente de Kendall-tau é um método (não paramétrico) aplicado no calculo da correlação em listas ordenadas.

Tabela 5-6: Coeficiente Gini das medidas de posição da rede ***RT-net*** e ***RT-base***

Métrica	**Coeficiente Gini** ***RT-net***	**Coeficiente Gini** ***RT-base***
Cc	99%	33%
Bc	1%	98%
Ec	85%	97%
Dc	54%	54%
Prank	3%	65%

O coeficiente de Kendall-Tau é muitas vezes descrito como uma medida de concordância entre dois conjuntos de classificações relativas a um conjunto de objetos ou experiências. Basicamente este coeficiente mede a diferença entre a probabilidade de as classificações estarem na mesma ordem e a probabilidade de estarem em ordens diferentes. Do ponto de vista amostral estas probabilidades são dadas através das frequências relativas respectivas (SIEGEL; CASTELLAN, 1988). As tabelas a seguir mostram a matriz de correlação Spearman-Rho e Kendall-Tau das redes ***RT-net*** e ***RT-base***. Os pares de métricas fortemente correlacionadas são: *Bc-Cc; Bc-*Dc_{out}*; Cc-*Dc_{out}*; CC-Dc; *Dc_{out}*-Dc; *Dc_{in}*-Dc* conforme ilustra a Tabela 5-7. A

Tabela 5-8 mostra que na a correlação Kendall-Tau, apesar dos valores variarem um pouco, os pares que estão fortemente correlacionados se mantiveram. Estas correlações indicam que alguns vértices atuam como "ponte", facilitando o fluxo de informação, e que são independentes em relação aos outros vértices, isto é, o caminho que ele precisa percorrer para alcançar os outros elos da rede é considerado o menor.

Em ambas as tabelas,

Tabela 5-9 e Tabela 5-10, verificou-se forte correlação entre a métrica Page Rank (***Prank***) e Centralidade de Autovalor (***Ec***) diferentemente do verificado na rede ***RT-net.***

Tabela 5-7: Matriz de Correlação Spearman-Rho da rede ***RT-net***

	Bc	***Cc***	***Ec***	***Prank***	***Dc_{in}***	***Dc_{out}***	***Dc***
Bc	-						
Cc	79%	-					
Ec	15%	-3%	-				
Prank	21%	3%	20%	-			
Dc_{in}	33%	3%	40%	44%	-		
Dc_{out}	78%	97%	-2%	4%	1%	-	
Dc	53%	67%	21%	27%	59%	68%	-

Tabela 5-8: Matriz de Correlação Kendall-Tau da rede ***RT-net***

	Bc	***Cc***	***Ec***	***Prank***	***Dc_{in}***	***Dc_{out}***	***Dc***
Bc	-						
Cc	78%	-					
Ec	13%	-2%	-				
Prank	16%	1%	12%	-			
Dc_{in}	30%	-2%	35%	37%	-		
Dc_{out}	95%	77%	12%	16%	27%	-	
Dc	52%	63%	19%	22%	57%	52%	-

Tabela 5-9: Matriz de Correlação Kendall-Tau da rede ***RT-base***

	Dc_{in}	***Dc_{out}***	***Dc***	***Cc***	***Bc***	***Prank***	***Ec***
Dc_{in}	-						
Dc_{out}	-1 %	-					
Dc	66%	57%	-				
Cc	-1%	14%	-2%	-			
Bc	77%	31%	52%	1%	-		
Prank	98%	-1%	65%	-1%	78%	-	
Ec	98%	-0,09%	65%	-1%	78%	99%	-

Tabela 5-10: Matriz de Correlação Spearman-Rho da rede ***RT-base***

	Dc_{in}	***Dc_{out}***	***Dc***	***Cc***	***Bc***	***Prank***	***Ec***
Dc_{in}	-						
Dc_{out}	-2%	-					
Dc	67%	58%	-				
Cc	-13%	18%	-3%	-			
Bc	79%	32%	53%	2%	-		
Prank	100%	-1%	67%	-13%	79%	-	
Ec	100%	-1%	67%	-13%	79%	100%	-

5.4 Fractalidade

Kitsak e colaboradores (KITSAK *et al.*, 2007) verificaram que a Centralidade de Intermediação (***Bc***) e o Grau Total (***Dc***) são fracamente correlacionadas (na correlação de Pearson) em Redes do tipo Fractal se comparados com redes não Fractais. Os autores ainda constataram que pequenos valores de Grau Total estão associados à altos valores de Centralidade de Intermediação.

A Tabela 5-11 ilustra os valores da correlação calculados para as métricas ***Bc*** e ***Dc***. Verifica-se a fraca correlação (54%) entre elas. Assim, ***RT-net*** e ***RT-base*** podem apresentar comportamento fractal. Na Figura 5-11, tem-se o gráfico de dispersão entre ***Bc*** e ***Dc***. A concentração dos pontos está localizada no quadrante esquerdo inferior. De acordo com os achados de (KITSAK *et al.*, 2007) a concentração dos pontos nesta região pode caracterizar rede com comportamento Fractal. A fraca correlação também foi verificada nas correlações de Spearman-rho e Kendall-Tau entre essas métricas.

Tabela 5-11: Matriz de correlação Pearson entre ***Dc*** e ***Bc*** da rede ***RT-net*** e ***RT-base***

	***Dc* (*RT-net*)**	***Dc* (*RT-base*)**
Bc	54%	53%

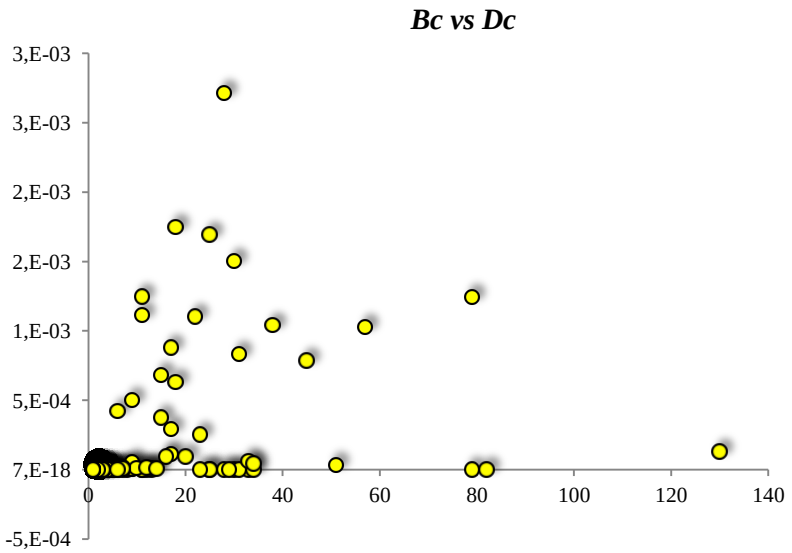


Figura 5-11: Gráfico de dispersão do *Dc* e *Cc*.

Uma distribuição que segue uma lei de potencia é uma distribuição na forma:

$$p(x) = c.x^{-\gamma} \quad \text{Equação 5-2}$$

Onde p(***x***) é a probabilidade de ***x*** ocorrer, sendo c uma constante de proporcionalidade e γ o expoente da potência. Distribuições que seguem uma lei de potencia ocorrem em muitos problemas e são importantes para o entendimento de fenômenos naturais e humanos. Em redes complexas, é comum que a distribuição do grau dos vértices tende a seguir as Leis de Potência. A Lei de Potencia são muitas vezes chamada de distribuições livre de escala, significando que uma distribuição se parece

com ela mesma independente da escala, implícita de invariante de escala ou autosemelhança. Na geométrica fundamental, à medida que um objeto aumenta de tamanho, seu volume aumenta, mudam-se as propriedades que dependem do volume também como por exemplo, o peso. Todavia, a propriedade de auto similaridade faz com que alguns sistemas mantenham as mesmas propriedades seja qual for a escala[17] utilizada. A propriedade livre de escala significa que aumentando a escala ou unidade pela qual $\boldsymbol{x}$ é medido por um fator $\boldsymbol{b}$, o formato da distribuição de $\boldsymbol{p}(x)$ não é alterada. A distribuição Gaussiana (Normal) não são livres de escala (KIM *et al.*, 2007; ZHOU; JIANG; SORNETTE, 2007; LONG; XU, 2009).

Fractal segundo Mandelbrot (1977) é uma entidade caracterizada por irregularidades que governam a sua forma e complexidade Figura 5-12. Em suma, fractal, em geral são objetos gerados pela repetição de um processo recursivo, apresentando auto similaridade, complexidade e dimensão fracionada. A auto similaridade é a característica de apresentar cópias de si mesmo em seu interior em diferentes tamanhos. Uma pequena parte é semelhante ao todo, uma fração de um fractal é uma replica do todo, isto é, pequenas partes de um fractal em diferentes escalas é semelhante ao todo.

A complexidade significa que não é possível representa-los completamente, pois os detalhes são infinitos, justamente por sua característica de auto semelhança.

Qualquer conjunto matemático pode ser caracterizado por sua dimensão, que representa a quantidade de parâmetros necessária para se descrever qualquer ponto deste conjunto. A dimensão no espaço Euclidiano é inteira, isto é, um ponto tem dimensão zero, uma linha é unidimensional, o plano bidimensional, o sólido tridimensional, já os fractais tem dimensão fracionada que está relacionada ao grau de irregularidade dos mesmos.

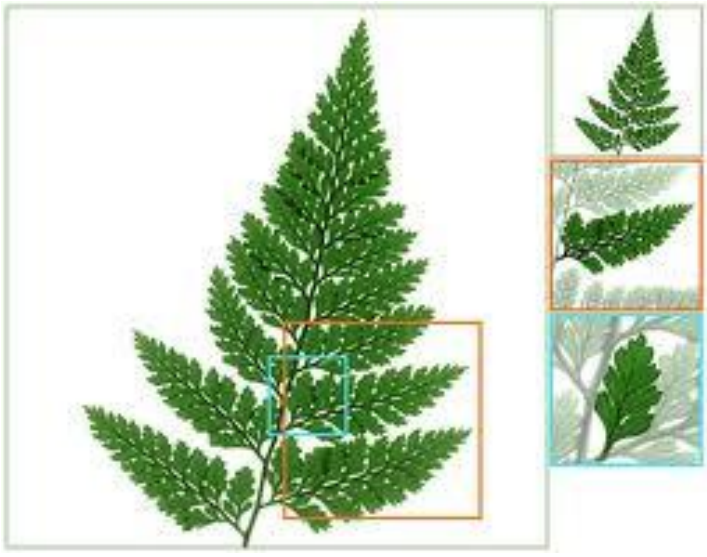


Figura 5-12: Exemplos de fractais na natureza. As folhas da planta Samambaia crescem de acordo com o padrão fractal, onde cada ramo é semelhante ao todo.

[17] Escala é a dimensão espacial ou temporária de um fenômeno

5.5 Análise da topologia da rede *RT-net*

São basicamente dois os tipos de estudos em Lei de Escala, o primeiro abrange as redes complexas no auxilio aos estudos da topologia, o segundo dentro da geometria fractal, é utilizada para calculo da dimensão fractal.

Foram analisadas as características da Lei de Escala da rede ***RT-net***. Em primeiro lugar, foi verificado o histograma de graus e foi realizado um ajuste não-linear Levenberg-Marquardt (MORÉ, 1978) por uma função potência do tipo $p(x) = c.x^{-\gamma}$ ao invés de tomar o método que utiliza o coeficiente de inclinação em um gráfico log-log. A estratégia justifica-se pos, a amostra possui apenas 1.232 usuários e o método de determinação por gráfico log-log não seria muito apropriado. Cabe salientar que o expoente não é realmente importante, mas em avaliar se a Rede ***RT-net*** tem comportamento Livre de Escala com um valor de $\gamma < 3$.

Se tomarmos o logaritmo da Equação 5-2 em ambos os lados, tem-se uma equação de uma reta onde o parâmetro γ é o coeficiente angular.

$$p(x) = c.x^{-\gamma} \quad \rightarrow \quad \log\big(p(x)\big) = \log(c) + \gamma \log(x) \quad \text{Equação 5-3}$$

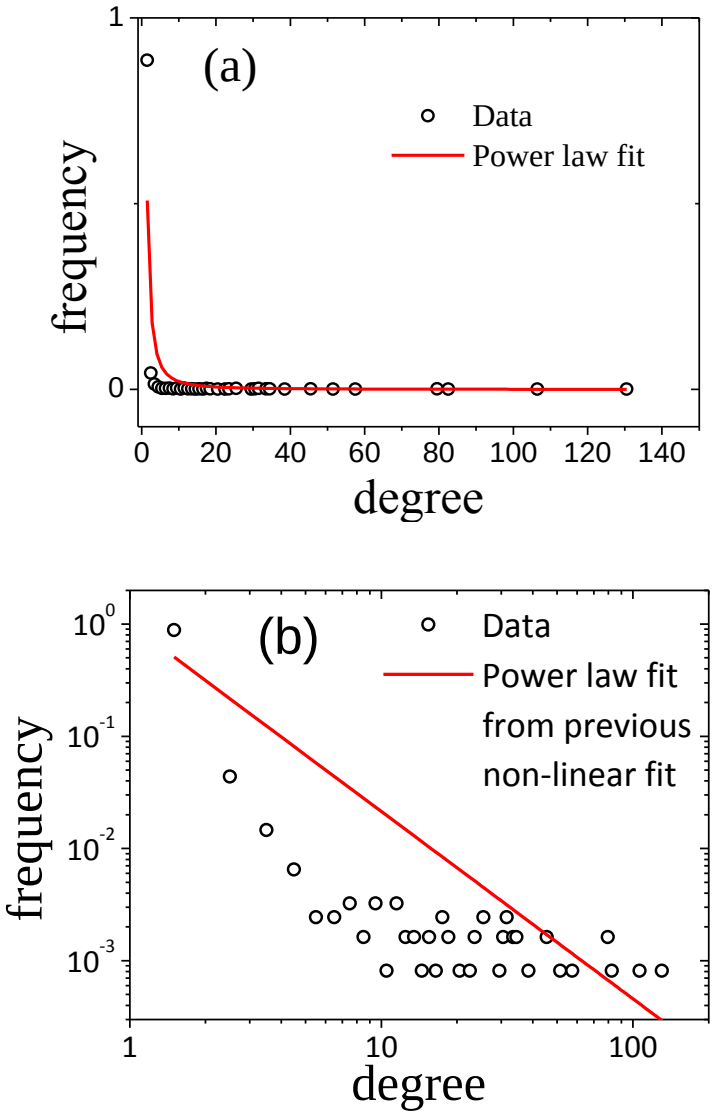


Figura 5-13: Ajuste dos pontos: (a) Ajuste direto com o método Marquardt method (não linear) (b) Os mesmos pontos em um gráfico log-log com ajuste não linear.

Do gráfico pode-se retirar o valor de $\gamma \approx 1.66$ (Figura 5-13 (a)) sendo interpretado como um valor baixo em comparação com a média. Foi feito também um ajuste linear, que pode ser visto na Figura 5-14, e novamente o valor foi considerado pequeno $\gamma \approx 0.95$.

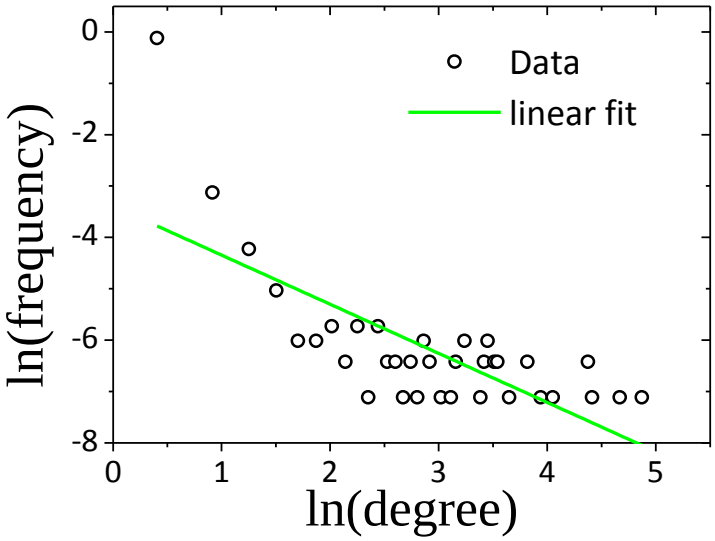


Figura 5-14: Gráfico de Ajuste Linear da frequência do Grau total

Um terceiro método foi aplicado. O Método de Cálculo de Momentos, comparando momentos experimentais e teóricos onde o $k^{\text{ésimo}}$ momento experimental é dado por:

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} {x_i}^k \quad \text{Equação 5-4}$$

Onde x_i é o grau do $i^{\text{ésimo}}$ nó e o $k^{\text{ésimo}}$ momento teórico é calculado por:

$$\langle x^k \rangle = (\gamma - 1)\int_1^{\infty} x^{k-\gamma} dx = \frac{(\gamma-1)}{(\gamma-k-1)} \quad \text{Equação 5-5}$$

E assim é feita a comparação entre os momentos $\bar{x}$ e $\langle x^k \rangle$ pela razão entre eles.

$$e(k) = \bar{x}_k / x^k \text{ e } t(k) = \langle x^k \rangle / \langle x \rangle^k = \frac{(\gamma-2)^k}{(\gamma-k-1)(\gamma-1)^{k-1}}. \quad \text{Equação 5-6}$$

O método de momentos se fez necessário, pois a amostra foi considerada pequena para o cálculo do coeficiente de escala (ou de potência) γ.

Conforme pode ser observado na Figura 5-15, o valor da potencia $\gamma \approx 2{,}37$, sugerindo assim, que a rede ***RT-net*** tem a distribuição de grau que segue a Lei de Potência, e como consequência é uma rede do tipo Livre de Escala. Pode-se inferir assim, que os resultados verificados para a Rede são invariantes na escala, ou seja, se o tamanho da rede aumentar, suas propriedades gerais serão preservadas.

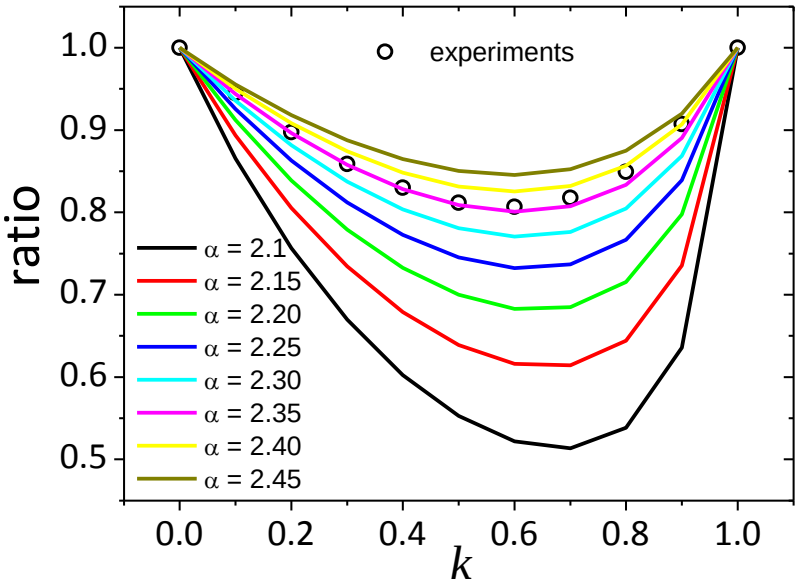


Figura 5-15: Razão entre os momentos experimentais e teóricos $\gamma \approx 2.37$

5.6 Resultados do algoritmo de ordenamento

Para avaliar a metodologia de ordenamento proposta nesta pesquisa, utilizou-se como *baseline* diferentes metodologias de ordenamento propostas pelos autores: (JIANWEI; LILI; TIANZHU, 2008; CHA *et al.*, 2010; KWAK *et al.*, 2010). Além disso, foi construída uma rede ***RT-base*** cujos pesos dos arcos são binários, ou seja, $\boldsymbol{w}\ (\boldsymbol{e_k}) = 1$ com o mesmo propósito.

A Tabela 5-12 mostra os desempenhos do ordenamento para a rede ***RT-net*** das seguintes estratégias:

1) Metodologia Proposta: mostra-se apenas os melhores desempenhos de $\mathcal{R}a\mathcal{R}^1$ à $\mathcal{R}a\mathcal{R}^{10}$
2) Métricas da literatura (baseline): de $\mathcal{R}a\mathcal{R}^{11}$ à $\mathcal{R}a\mathcal{R}^{17}$
3) Métricas isoladamente: de $\mathcal{R}a\mathcal{R}^{18}$ à $\mathcal{R}a\mathcal{R}^{21}$

Se considerarmos apenas o maior valor de AP, as equações $\boldsymbol{\mathcal{R}a\mathcal{R}^5}$, $\boldsymbol{\mathcal{R}a\mathcal{R}^6}$ e $\boldsymbol{\mathcal{R}a\mathcal{R}^7}$ obtiveram o maior desempenho. Se considerarmos apenas o valor de P@212, as equações $\boldsymbol{\mathcal{R}a\mathcal{R}^7}$ e $\boldsymbol{\mathcal{R}a\mathcal{R}^8}$ obtiveram melhor desempenho. Considerando os dois valores tem-se a equação que representa $\boldsymbol{\mathcal{R}a\mathcal{R}^7}$:

$$\frac{((\mathbf{0.7}*\boldsymbol{Cc})+(\mathbf{0.2}*\boldsymbol{D_{in}})+(\mathbf{0.1}*\boldsymbol{D_{out}}))}{\boldsymbol{Max}\{(\boldsymbol{m_i}*\boldsymbol{w_i}),..,(\boldsymbol{m_n}*\boldsymbol{w_n})\}} \qquad \text{Equação 5-7}$$

Observou-se que a métrica presente nos melhores desempenhos para a rede ***RT-net*** é a Centralidade de Proximidade (*Cc*).

A Tabela 5-13 mostra os desempenhos do ordenamento para a rede ***RT-base*** das seguintes estratégias:

1) Metodologia Proposta: de $\mathcal{R}a\mathcal{R}^{1}$ à $\mathcal{R}a\mathcal{R}^{10}$, são as mesmas equações, de modo a comparar com os resultados da rede ***RT-net***, e os melhores desempenhos de ordenamento para a rede ***RT-base***, de $\mathcal{R}a\mathcal{R}^{22}$ à $\mathcal{R}a\mathcal{R}^{25}$
2) Métricas da literatura (baseline): de $\mathcal{R}a\mathcal{R}^{11}$ à $\mathcal{R}a\mathcal{R}^{17}$
3) Métricas isoladamente: de $\mathcal{R}a\mathcal{R}^{18}$ à $\mathcal{R}a\mathcal{R}^{21}$

Destaca-se que, em princípio, a métrica follower (seguidores) é o mesmo que métrica Grau de Entrada na rede *Twitter*. Todavia os valores da métrica Grau de Entrada foram calculados em função da rede ***RT-net*** e a follower foi retirada da rede *Twitter* como um todo. Por isso os valores resultantes das duas métricas são diferentes nas tabelas a seguir.

Tabela 5-12: Desempenhos verificados rede ***RT-net***

$\mathcal{R}a\mathcal{R}^{\varphi}$	Equação	AP (%)	p@212
$\mathcal{R}a\mathcal{R}^{1}$	$\frac{((Bc) + (0.5 * Cc) + (0.2 * Dc) + (0.3 * Ec))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	54	51
$\mathcal{R}a\mathcal{R}^{2}$	$\frac{((Bc) + (0.6 * Cc) + (0.2 * Dc) + (0.2 * Ec))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	56	54
$\mathcal{R}a\mathcal{R}^{3}$	$\frac{((Bc) + (0.6 * Cc) + (0.1 * Prank) + (0.3 * Dc))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	56	50
$\mathcal{R}a\mathcal{R}^{4}$	$\frac{((0.3 * Cc) + (0.2 * D_{out}) + (0.4 * Ec) + (0.1 * D_{in}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	56	50
$\mathcal{R}a\mathcal{R}^{5}$	$\frac{((06 * Cc) + (0.2 * Ec) + (0.2 * Dc))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	58	55
$\mathcal{R}a\mathcal{R}^{6}$	$\frac{((0.6 * Cc) + (0.1 * Ec) + (0.2 * D_{in}) + (0.1 * D_{out}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	58	55
$\boldsymbol{\mathcal{R}a\mathcal{R}^{7}}$	$\frac{((\mathbf{0.7} * \boldsymbol{Cc}) + (\mathbf{0.2} * \boldsymbol{D_{in}}) + (\mathbf{0.1} * \boldsymbol{D_{out}}))}{\boldsymbol{Max}\{(\boldsymbol{m_i} * \boldsymbol{w_i}), .., (\boldsymbol{m_n} * \boldsymbol{w_n})\}}$	**58**	**56**
$\mathcal{R}a\mathcal{R}^{8}$	$\frac{((0.7 * Cc) + (0.1 * D_{in}) + (0.2 * D_{out}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	53	56
$\mathcal{R}a\mathcal{R}^{9}$	$\frac{((0.1 * Bc) + (0.7 * Cc) + (0.1 * Ec) + (0.1 * Prank))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	49	44
$\mathcal{R}a\mathcal{R}^{10}$	$\frac{((0.25 * Bc) + (0.25 * Cc) + (0.25 * Dc) + (0.25 Prank))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	48	43
$\mathcal{R}a\mathcal{R}^{11}$	Re*tweet*	31	34
$\mathcal{R}a\mathcal{R}^{12}$	Follower	18	20

$\mathcal{R}a\mathcal{R}^{13}$	Dc_{in} (CHA *et al.*, 2010)	39	37
$\mathcal{R}a\mathcal{R}^{14}$	*Prank*	37	37
$\boldsymbol{\mathcal{R}a\mathcal{R}^{15}}$	$C(v) = (0{,}6 * Dc) + (0{,}3 * Cc) + (0{,}1 * Bc)$ Equação 5-8 (JIANWEI; LILI; TIANZHU, 2008)	**53**	**50**
$\boldsymbol{\mathcal{R}a\mathcal{R}^{16}}$	$C(v) = (0{,}5 * Dc) + (0{,}3 * Cc) + (0{,}2 * Bc)$ Equação 5-9 (JIANWEI; LILI; TIANZHU, 2008)	**53**	**50**
$\boldsymbol{\mathcal{R}a\mathcal{R}^{17}}$	***Dc***	**53**	**50**
$\mathcal{R}a\mathcal{R}^{18}$	*Cc*	42	33
$\mathcal{R}a\mathcal{R}^{19}$	*Bc*	33	25
$\mathcal{R}a\mathcal{R}^{20}$	Dc_{out}	42	33
$\mathcal{R}a\mathcal{R}^{21}$	*Ec*	44	42

Tabela 5-13: Desempenhos verificados rede ***RT-base***

$\boldsymbol{\mathcal{R}a\mathcal{R}^{\varphi}}$	**Equação**	**AP (%)**	**p@212**
$\mathcal{R}a\mathcal{R}^{1}$	$\frac{((Bc) + (0.5 * Cc) + (0.2 * Dc) + (0.3 * Ec))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	41	30
$\mathcal{R}a\mathcal{R}^{2}$	$\frac{((Bc) + (0.6 * Cc) + (0.2 * Dc) + (0.2 * Ec))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	41	27
$\mathcal{R}a\mathcal{R}^{3}$	$\frac{((Bc) + (0.6 * Cc) + (0.1 * Prank) + (0.3 * Dc))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	43	33
$\mathcal{R}a\mathcal{R}^{4}$	$\frac{((0.3 * Cc) + (0.2 * D_{out}) + (0.4 * Cc) + (0.1 * D_{in}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	39	25
$\mathcal{R}a\mathcal{R}^{5}$	$\frac{((06 * Cc) + (0.2 * Ec) + (0.2 * Dc))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	43	25
$\mathcal{R}a\mathcal{R}^{6}$	$\frac{((0.6 * Cc) + (0.1 * Ec) + (0.2 * D_{in}) + (0.1 * D_{out}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	40	25
$\mathcal{R}a\mathcal{R}^{7}$	$\frac{((0.7 * Cc) + (0.2 * D_{in}) + (0.1 * D_{out}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	38	24

$\mathcal{R}a\mathcal{R}^{8}$	$\frac{((0.7 * Cc) + (0.1 * D_{in}) + (0.2 * D_{out}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	32	22
$\mathcal{R}a\mathcal{R}^{9}$	$\frac{\left((0.1 * Bc) + (0.7 * Cc\right) + (0.1 * Ec) + (0.1 * Prank))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	32	22
$\mathcal{R}a\mathcal{R}^{10}$	$\frac{\left((0.25 * Bc) + (0.25 * Cc\right) + (0.25 * Dc) + (0.25Prank))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	35	22
$\mathcal{R}a\mathcal{R}^{11}$	Re*tweet*	31	34
$\mathcal{R}a\mathcal{R}^{12}$	Follower	18	20
$\mathcal{R}a\mathcal{R}^{13}$	Dc_{in} (CHA *et al.*, 2010)	39	37
$\mathcal{R}a\mathcal{R}^{14}$	*Prank*	55	49
$\mathcal{R}a\mathcal{R}^{15}$	$C(v) = (0{,}6 * Dc) + (0{,}3 * Cc) + (0{,}1 * Bc)$ Equação 5-10 (JIANWEI; LILI; TIANZHU, 2008)	53	50
$\mathcal{R}a\mathcal{R}^{16}$	$C(v) = (0{,}5 * Dc) + (0{,}3 * Cc) + (0{,}2 * Bc)$ Equação 5-11 (JIANWEI; LILI; TIANZHU, 2008)	54	51
$\mathcal{R}a\mathcal{R}^{17}$	*Dc*	53	50
$\mathcal{R}a\mathcal{R}^{18}$	*Cc*	19	15
$\mathcal{R}a\mathcal{R}^{19}$	*Bc*	53	49
$\mathcal{R}a\mathcal{R}^{20}$	Dc_{out}	42	33
$\mathcal{R}a\mathcal{R}^{21}$	*Ec*	55	49
$\mathcal{R}a\mathcal{R}^{22}$	$\frac{\left((0.4 * Prank) + (0.2 * Ec\right) + (0.4 * Dc_{in}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	54	54
$\mathcal{R}a\mathcal{R}^{23}$	$\frac{((0.3 * Prank) + (0.5 * Dc_{in}) + (0.2 * Dc_{out}))}{Max\{(m_i * w_i), .., (m_n * w_n)\}}$	54	54
$\boldsymbol{\mathcal{R}a\mathcal{R}^{24}}$	$\frac{((\mathbf{0.3} * \boldsymbol{Prank}) + (\mathbf{0.3} * \boldsymbol{Dc_{in}}) + (\mathbf{0.3} * \boldsymbol{Dc_{out}}) + (\mathbf{0.1} * \boldsymbol{Ec}))}{\boldsymbol{Max}\{(\boldsymbol{m_i} * \boldsymbol{w_i}), .., (\boldsymbol{m_n} * \boldsymbol{w_n})\}}$	**55**	**54**
$\boldsymbol{\mathcal{R}a\mathcal{R}^{25}}$	$\frac{((\mathbf{0.2} * \boldsymbol{Prank}) + (\mathbf{0.3} * \boldsymbol{Dc_{in}}) + (\mathbf{0.1} * \boldsymbol{Dc_{out}}) + (\mathbf{0.3} * \boldsymbol{Bc}))}{\boldsymbol{Max}\{(\boldsymbol{m_i} * \boldsymbol{w_i}), .., (\boldsymbol{m_n} * \boldsymbol{w_n})\}}$	**55**	**54**

Se considerarmos apenas o maior valor de AP, as equações $\boldsymbol{\mathcal{R}a\mathcal{R}^{14}}$, $\boldsymbol{\mathcal{R}a\mathcal{R}^{21}}$, $\boldsymbol{\mathcal{R}a\mathcal{R}^{24}}$ e $\boldsymbol{\mathcal{R}a\mathcal{R}^{25}}$ obtiveram o maior desempenho. Se considerarmos apenas o valor de P@212, as equações $\boldsymbol{\mathcal{R}a\mathcal{R}^{22}}$, $\boldsymbol{\mathcal{R}a\mathcal{R}^{23}}$, $\boldsymbol{\mathcal{R}a\mathcal{R}^{24}}$ e $\boldsymbol{\mathcal{R}a\mathcal{R}^{25}}$ obtiveram melhor desempenho. Considerando os dois valores tem-se a equação que representa $\boldsymbol{\mathcal{R}a\mathcal{R}^{24}}$ e $\boldsymbol{\mathcal{R}a\mathcal{R}^{25}}$. É interessante notar que as métricas presentes nos melhores resultados são o Page Rank (***Prank***) e Grau de Entrada ($\boldsymbol{Dc_{in}}$) e Saída ($\boldsymbol{Dc_{out}}$).

Na Tabela 5-14 tem-se a avaliação em função da medida P@n. Pode-se verificar que para os níveis até a centésima posição, a medida está em torno de 80%.

Ao se fazer uma busca simples na ferramenta Google, pelo assunto Alzheimer (ilustrado no Anexo II), foram retornadas 14 páginas. As páginas recuperadas foram: 07 agências de saúde não governamentais, 01 agência de notícia (mídia), 01 ambiente colaborativo de construção de texto (Wikipédia) e 05 agências governamentais de saúde. Verifica-se que as agências de saúde foram recuperadas ao final do *rank* (na posição de número 9 à 13). Utilizando a metodologia de ordenamento proposta aqui nesta pesquisa, tem-se que a P@10 foi de 20%, e P@14 igual a 29% pois as primeiras páginas recuperaram apenas agências de saúde privada, mídia além da Wikipédia que podem conter informações enviesadas.

Tabela 5-14: Diferentes níveis de P@n

P@n	% níveis
P@10	80%
P@20	80%
P@30	87%
P@40	83%
P@50	86%
P@60	83%
P@70	84%
P@80	83%
P@100	80%

5.7 Discussão geral dos resultados observados

Faz-se agora uma discussão dos achados observados na amostra e nas redes ***RT-net*** e ***RT-base***.

Observou-se que a maioria dos usuários alvo foram retuitados apenas uma vez e o foi praticado por usuários da sua própria rede de seguidores (followers). A relação follower foi justamente a relação com maior percentual na amostra. Observou-se ainda que o fenômeno "celebridade" está presente na amostra, ou seja, usuários com pouco ou nenhum seguidor e outros com centenas de milhares.

A distribuição dos *retweet* por usuário alvo e por usuário fonte apresentou certo grau de desigualdade na sua distribuição, considerando-se os valores calculados do Coeficiente Gini. Verificou-se que existem usuários que desempenham papel importante na propagação de informações e outros que são mais atuantes como fonte de informações.

As propriedades gerais calculadas e ilustradas na Tabela 5-5 para a rede ***RT-net*** se verificaram também na rede ***RT-base***, mesmo porque as redes apresentam a mesma estrutura, sendo diferenciadas apenas pelo valor do peso dos seus arcos. Ambas as redes não são densas, isso quer dizer que os usuários não estão perfeitamente conectados, tal como em uma rede Regular. Ambas também apresentam estrutura de comunidade (grupos) além de topologia hierárquica.

Apesar das propriedades gerais se preservarem em ambas as redes, elas obtiveram valores diferentes no computo do Coeficiente Gini. Por exemplo, ainda que a metodologia de cálculo do ***Ec*** seja semelhante à do ***Prank***, os valores desta ultima passaram da completa igualdade (na ***RT-net***) à completa desigualdade na ***RT-base*** e se mantiveram praticamente o mesmo na métrica ***Ec***. Uma razoável explicação para tal fato é que os melhores desempenhos de ordenamento da rede ***RT-base*** foram aqueles que justamente utilizavam a métrica ***Prank*** na metodologia de ordenamento. De um modo geral, o ***Prank*** calcula a importância de um vértice atribuindo pontuações relativas de forma que conexões com altas pontuações contribuem mais para a pontuação deste vértice que vértices com baixa pontuação. Ou seja, quanto mais arestas com nós de alta pontuação apontam para um vértice, mais alta será a sua pontuação. Assim, na rede ***RT-net*** que tem pesos diferenciados nas arestas, apresentou uma distribuição mais homogênea dos valores do ***Prank***, e assim essa métrica pouco influenciou a metodologia de cálculo da reputação proposta.

A situação inversa ocorreu com a métrica ***Bc*** na análise do Coeficiente Gini. A métrica ***Bc*** que representa os usuários que controlam o fluxo de informações, a distribuição passou de completa igualdade na rede ***RT-net*** para completa desigualdade na ***RT-base***. A distribuição desigual destacou os usuários que atuam como fonte de disseminação da informação, e em ultima análise, influenciou a metodologia de cálculo da reputação. Já a métrica ***Cc*** passou de completa desigualdade na rede ***RT-net*** para completa igualdade na rede ***RT-base***, as mesmas observações podem ser feitas. A ***Cc*** na rede ***RT-base*** tem pouco ou nenhuma influência na metodologia.

Os pares de métricas {***Bc*** ; ***Cc***}, {$\boldsymbol{Dc_{out}}$; ***Bc***} e {$\boldsymbol{Dc_{out}}$; ***Cc***} estão positivamente correlacionados na rede ***RT-net***, o mesmo não se verificou na rede ***RT-base***. Nesta ultima, os pares correlacionados foram {$\boldsymbol{Dc_{in}}$; ***Ec***}, { $\boldsymbol{Dc_{in}}$; ***Prank***}, {$\boldsymbol{Dc_{in}}$; ***Bc***} e {***Prank*** ; ***Ec***}. É interessante notar que o grau de saída ($\boldsymbol{Dc_{out}}$) tem forte correlação com as métricas ***Cc*** e ***Bc*** na rede ***RT-net***, que estão justamente presente na equação do cálculo reputação do melhor desempenho. O mesmo poder ser verificado na rede ***RT-base***, com as métricas $\boldsymbol{Dc_{in}}$ e ***Prank***.

Não foi possível encontrar uma curva de ajuste dos parâmetros *retweet* vs *tweet*, *follower* ou *following* nesta a amostra de dados. Esse fato pode indicar que o processo de *retweet* obedece a outros fatores além daqueles que são estabelecidos nas relações intrínsecas do *Twitter* (seguidores e seguidos). Isto é, o processo de *retweet* envolve não só as relações que são estabelecidas como também processos cognitivos na escolha dos *tweet* que são encaminhados. Esperava-se que as relações amizade fossem talvez mais fortes no ambiente *Twitter* que em outras plataformas sociais. No entanto, o que se verificou foi que estas relações parecem influenciar apenas em parte essas escolhas.

5.8 Avaliação da metodologia proposta

Nesta seção avalia-se a metodologia de cálculo da reputação e ordenamento proposto em relação ao *baseline* da literatura e aos resultados de ordenamento da reputação da rede ***RT-base***.

O melhor desempenho tanto do ponto de vista da medida AP quanto da P@212 no ordenamento foi a equação que possuía como parcelas as métricas Grau de Entrada, Grau de Saída e Centralidade de Proximidade, essa última com maior peso. Algumas variações na equação também alcançaram bons valores no desempenho, todas utilizando a Centralidade de Proximidade. Essa métrica, conforme dito anteriormente, é baseada na distância entre os vértices, o quão próximo um nó se encontra em relação aos outros na rede. Levando em consideração não só as relações diretas, mas das relações indiretas, especialmente quando dois atores não estão adjacentes. Desta forma, quanto maior é o valor desta métrica, maior é a capacidade de um vértice interagir com o restante da rede e assim, tornam-se instrumentos poderosos no compartilhamento de informações.

Primeiramente, faz-se a análise para os achados verificados com a rede ***RT-base***, e em seguida os achados verificados com a rede ***RT-base***.

Ao se comparar os resultados do ordenamento proposto na literatura e os valores encontrados com a metodologia proposta, verifica-se o melhor desempenho desta em relação aos da literatura. A diferença é significativa se for utilizado apenas a quantidade de *follower*, ou a quantidade de *retweet*, usuários não relevantes aparecem no topo da listagem. Até mesmo, a reconhecida métrica *PageRank*, amplamente utilizada em avaliações de "popularidade" e "importância", desempenhou papel insignificante no ordenamento dos relevantes. Cabe salientar que apenas a proposta da literatura que utilizava o Grau Total em associação com a Centralidade de Proximidade e Centralidade de Intermediação obteve um desempenho plausível. Contudo, o mesmo desempenho pode ser verificado utilizando-se apenas a métrica Grau Total na equação.

Desta forma, a metodologia de ordenamento aqui proposta, mostrou-se adequada no ordenamento dos relevantes, colocando-os o mais próximo possível do topo da listagem.

O Anexo I apresenta os 50 primeiros usuários que foram recuperados pelo algoritmo de ordenamento. Cabe salientar que os piores resultados, ou seja, os usuários que não foram bem ranqueados são justamente àqueles que ocupam a posição terminal, isto é, são os nós de ponta da rede. A posição ocupada (nó de ponta) na topologia da rede parece interferir na metodologia de cálculo da reputação.

Já na análise dos achados com a rede ***RT-base***, destaca-se a participação da métrica ***Prank***. A métrica está presente na maioria dos melhores desempenhos, porém inferior ao verificado na rede ***RT-net***. Não foi possível obter os mesmos desempenhos utilizando as mesmas equações nas duas redes. Cada uma delas possui propriedades topológicas diferentes, que influenciaram no cálculo da reputação. Os resultados obtidos com a rede ***RT-base*** foram superiores aos resultados obtidos com as propostas da literatura, entretanto foram inferiores aos resultados verificados com a rede ***RT-net***.

Em resumo, o melhor desempenho na ordenação da reputação foi verificado com a rede ***RT-net*** que utiliza a Centralidade de Proximidade juntamente com o Grau de Entrada e Grau de Saída. Em segundo lugar ficou a Rede ***RT-base*** que utiliza a métrica Page Rank, juntamente com o Grau de entrada e Grau de Saída, mas com desempenho inferior. Por último tem-se os desempenhos de ambas as redes que utilizam as

metodologias de ordenamento da reputação proposta na literatura. O que ficou evidente foi a contribuição das métricas Grau de Entrada e Grau de Saída, quando utilizadas nas equações de cálculo da reputação em ambas as redes.

Cabe por fim ressaltar que de acordo com os achados, a rede ***RT-net*** apresenta do comportamento Fractal, pois a correlação entre o Grau Total e a Centralidade de Intermediação é que pequena se comparada com a rede de seguidores. E ainda, que a distribuição dos graus segue uma Lei de Potencia, com características das redes Livre de Escala.

6 CONCLUSÃO E TRABALHOS FUTUROS

O objetivo desta pesquisa foi propor e avaliar uma abordagem que permitisse estimar a reputação de fontes de informação na área de saúde. Como caso de estudo, explorou-se o ambiente *Twitter*, e o paradigma da Análise das Redes Sociais aplicado como ferramenta de abordagem do problema.

O aumento da produção de informação virtual de todo o tipo, não necessariamente científica, vem ocorrendo de maneira acelerada, não ordenada, não revisada e irrestrita. Por um lado, este crescimento é devido, em parte, às novas tecnologias de transmissão e ao barateamento de dispositivos móveis, tornando-os integrados e ubíquos no cotidiano das pessoas. Por outro, é devido ao aumento das contribuições individuais e coletivas, pela proliferação de Serviços Sociais Online na Web 2.0. Todos os dias são produzidos uma quantidade enorme de textos, fotos e vídeos com possibilidades de publicação e replicação instantâneas na *web*. Neste contexto, o sujeito é visto agora como criador e validador.

Análise de Redes Sociais se consolidou nos últimos anos, devido ao seu potencial como ferramenta versátil na representação de vários problemas. É possível enumerar aplicações em áreas tão distintas como engenharia, sociologia, computação, linguística e biologia.

Os resultados verificados permitiram avançar na compreensão da dinâmica dos processos de interação entre os atores. O desempenho da ordenação dos usuários relevantes (aqueles que exibem reputação) mostrou-se satisfatória considerando-se o percentual de acertos encontrados.

6.1 Principais Publicações Relacionadas

Nesta seção são apresentadas as principais publicações da autora relacionadas com esta tese. No primeiro artigo, citado a seguir, tem-se o estudo inicial que serviu como norteador da pesquisa. O objetivo era desenvolver um framework de um sistema de recomendação baseado no perfil do usuário e na qualidade da informação recomendada. A partir deste estudo, o escopo do problema foi se delineando. Neste estudo inicial percebeu-se que os Sistemas de Recomendação atuais devem trazer para o plano frontal não só a relevância do documento recuperado de acordo com as necessidades dos usuários, mas também a qualidade do documento recuperado. A importância está em deixar os usuários mais aptos a encontrar a "melhor" opção em termos de qualidade.

O primeiro artigo é publicação em evento nacional indexado pela Capes como B4. É um evento que vem se consolidando ao longo dos anos organizado pela Sociedade Brasileira de Informática em Saúde – SBIS, só existem dois eventos nacionais indexados que tratam do assunto. O segundo artigo foi publicado no "Cadernos de Informática" da UFRGS.

WEITZEL, L.; Palazzo, M. de Oliveira J. Sistemas de Recomendação baseado no perfil do usuário. In: XII Congresso Brasileiro de Informática em Saúde (CBIS2010), Porto de Galinhas, PB, 2010.

WEITZEL, L.; Palazzo, M. de Oliveira J.; et al. Proposta de métricas de avaliação da qualidade da informação médica para Sistemas de Recomendação baseados no perfil do usuário. Cadernos de Informática (UFRGS), v. 5, p. 23-48, 2010.

A partir dos estudos citados acima, a pesquisa foi se desenvolvendo no sentido a materializar os requisitos para a avaliação da qualidade. Nos artigos citados abaixo, são descritas a proposta da modelagem da rede e sua topologia, além da metodologia de cálculo dos pesos dos arcos. Nesta etapa, foi decidido que o ambiente da pesquisa seria a Rede Social *Twitter*, por apresentar características que propiciavam o estudo. O primeiro artigo foi apresentado no evento internacional indexado pela Capes em B2. O segundo artigo não é indexado pela Capes.

WEITZEL, L.; Quaresma, P.; Palazzo, M. de Oliveira J. Measuring node importance: A Multi-criteria Approach. Proceedings of the International Conference on WWW/Internet Interaction, 2011, Rio de Janeiro, RJ: IADIS Press, 2011. 415-420 p..

WEITZEL, L.; Quaresma, P.; Palazzo, M. de Oliveira J. Analyzing the strength of ties of Re*tweet* in health domain. Proceedings of the 2ª Jornadas de Informática da Universidade de Évora (JIUE'2011), Évora, PT: Universidade de Évora, 2011.

Nos dois últimos artigos, é aperfeiçoada a metodologia de ordenamento da relevância dos autores. E é proposta uma metodologia de avaliação do desempenho dos resultados verificados o que culminou no desenvolvimento de um algoritmo. O primeiro artigo foi publicado em um evento internacional indexado pela Capes como A2, é o encontro anual IEEE Computer Society sobre o tema e indexado também pela DBLP (Digital Bibliography & Library Project).

O segundo artigo foi publicado em um evento internacional, mas que ainda não foi avaliado pela Capes e tem indexação pela DBLP.

WEITZEL, L.; Quaresma, P.; PALAZZO, M. de Oliveira J. Evaluating quality of health information sources. Proceedings of the The 26th IEEE International Conference on Advanced Information Networking, AINA2012, Fukuoka-shi, JP:IEEE Computer Society, 655-661 p. 2012.

WEITZEL, L.; Quaresma, P.; PALAZZO, M. de Oliveira J. Measuring node importance on *Twitter* microblogging. Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS2012), Cracovia,RM, 2012.

6.2 Trabalhos futuros

Nesta seção são apontados alguns pontos que foram identificados como oportunidades de evolução da pesquisa. Dentre eles destacam-se:

- Verificar se a metodologia se aplica em outras áreas de conhecimento e não apenas na área médica. Quer se avaliar se a metodologia para estimar a reputação é apropriada, por exemplo, na área de Educação.
- Outro tema de interesse é verificar se a metodologia proposta se aplica também a outras Redes Sociais que utilizam o mesmo mecanismo de endosso de conteúdo, como por exemplo, a Rede *Pinterest*. A *Pinterest* é uma rede social baseada na publicação, organização e compartilhamento de imagens. A rede se tornou a quarta mais acessada (em novembro de 2012) nos Estados Unidos, segundo a pesquisa da *Experian Hitwise*[18]. Os conteúdos compartilhados são chamados de *pins*, os usuários podem dar *like*, (semelhante ao curtir do Facebook), pode comentar ou dar um *repin* (compartilhar com os seus seguidores). As imagens são organizadas por categorias em um *board*, por exemplo, arquitetura, roupas, alimentos, etc.
- E por fim, seria interessante fazer a análise textual dos *tweet* com objetivo de verificar se existem opiniões embutidas, considerando a polaridade dos conteúdos publicados. Classificando-se as palavras em positivas e negativas com base em uma lista previamente selecionada. Com isso, é possível executar algoritmos para classificação do conteúdo como positivo, negativo ou neutro. A estratégia de análise da polaridade permitiria identificar usuários maliciosos que disseminam mensagens de conteúdo duvidoso ou falso. Por exemplo, mensagens que emitem opiniões falsas, podendo ser falsas opiniões positivas (para promover) ou falsas opiniões negativas (para prejudicar). E assim poderiam ser atribuídos valores que expressassem o quão útil um *tweet* pode ser considerado, e desta forma a metodologia de cálculo da reputação poderia ser aprimorada.

[18] http://www.experian.com/hitwise/online-trends-social-media.html, acesso em 24.11.2012

REFERÊNCIAS

AHN, Y.-Y., *et al.* Analysis of topological characteristics of huge online social networking services. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. 16., 2007. Banff, Alberta, Canada. **Proceedings...** New York, USA: ACM, 2007. p. 835-844.

ALBERT, R.; BARABÁSI, A. L. Statistical mechanics of complex networks. **Reviews of modern physics**, v.74, n.1, p.47-47. 2002.

ANGER, I.; KITTL, C. Measuring influence on Twitter. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE MANAGEMENT AND KNOWLEDGE TECHNOLOGIES. 11., 2011. Graz, Austria. **Proceedings...** New York, USA: ACM, p. 1-4.

ARAUJO, J. D. S. D.; RODRIGUEZ, F. G. JCrawler4T. Pará: Universidade Federal do Pará 2011.

BARABÁSI, A.-L.; ALBERT, R.; JEONG, H. Scale-free characteristics of random networks: the topology of the world-wide web. **Physica A: Statistical Mechanics and its Applications**, v.281, n.1–4, p.69-77. 2000.

BARABÁSI, A.-L. **Linked : How everything is connected to everything else and what it means for business, science, and everyday life**. New York: Plume. 2003

BARNES, J. A. Class and committees in a Norwegian island parish. **Human Relations**, v.7, p.39-58. 1954.

BOCCALETTI, S., *et al.* Complex networks: Structure and dynamics. **Physics Reports**, v.424, n.4–5, p.175-308. 2006.

BOLLOBAS, B. **Modern Graph Theory**: Springer, v.184 of Graduate Texts in Mathematics. 1998

BONACICH, P.; LLOYD, P. Eigenvector-like measures of centrality for asymmetric relations. **Social Networks**, v.23, n.3, p.191-201. 2001.

BONACICH, P. Some unique properties of eigenvector centrality. **Social Networks**, v.29, n.4, p.555-564. 2007.

BONGWON, S., *et al.* Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In: INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING. 2., 2010. Palo Alto Res. Center, Inc., Palo Alto, CA, USA **Proceedings...** IEEE, p. 177-184.

BORGATTI, S. P., *et al.* Network Analysis in the Social Sciences. **Science**, v.323, n.5916, p.892-895. 2009.

BROWN, R.; REGINALD, A. On social structure. **Journal of the Royal Anthropological Institute**, v.70, n.1, p.1-12. 1940.

CAMACHO, J.; STOUFFER, D. B.; AMARAL, L. A. N. Quantitative analysis of the local structure of food webs. **Journal of Theoretical Biology**, v.246, n.2, p.260-268. 2007.

CASTILLO, C.; MENDOZA, M.; POBLETE, B. Information credibility on twitter. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. 20., 2011. Hyderabad, India. **Proceedings...** New York, USA: ACM, p. 675-684.

CHA, M., *et al.* Measuring User Influence in Twitter: The Million Follower Fallacy. In: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA. 4., 2010. Washington, D.C. **Proceedings...** New York, USA: AAAI Press, Menlo Park, California., p.

COLIZZA, V.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Reaction–diffusion processes and metapopulation models in heterogeneous networks. **Nature Physics**, v.3, n.4, p.276-282. 2007.

COSTA, L. D. F., *et al.* Characterization of complex networks: A survey of measurements. **Advances in Physics**, v.56, n.1, p.167-242. 2007.

DANA, J.; LOEWENSTEIN, G. A social science perspective on gifts to physicians from industry. **JAMA**, v.290, p.252-255. 2003.

DEGENNE, A.; FORSÉ, M. **Introducing social networks**. London: SAGE. 1999

DIESTEL, R. **Graph Theory**: Springer-Verlag. 2000

DOROGOVTSEV, S. N.; MENDES, J. F. F. **Evolution of networks : from biological nets to the Internet and WWW**. Oxford: Oxford University Press. 2003

ERDÕS, P.; RÉNYI, A. On Random Graphs. I. **Publicationes Mathematicae**, v.6, p.290-297. 1959.

_____. On the evolution of random graphs. **Publications of the Mathematical Institute of the Hungarian Academy of Sciences**, v.5, p.17-61. 1960.

EYSENBACH, G. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review. **JAMA: The Journal of the American Medical Association**, v.287, p.2691-2700. 2002.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the Internet topology. **SIGCOMM Comput. Commun. Rev.**, v.29, n.4, p.251-262. 1999.

FREEMAN, L. Centrality in social networks: Conceptual clarification. **Social Networks**, v.1, n.3, p.215-239. 1979.

FREEMAN, L. C. Cliques, Galois lattices, and the structure of human social groups. **Social Networks**, v.18, 1996, p.173-187. 1996.

GAYO-AVELLO, D. Detecting Important Nodes to Community Structure Using the Spectrum of the Graph. 2010. Disponível em: <http://arxiv.org/abs/1101.1703>, Acessado em: 16/06/2011.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. **National Academy of Sciences**, v.99, n.12, p.7821-7826. 2002.

GLEISER, P. M.; DANON, L. COMMUNITY STRUCTURE IN JAZZ. **Advances in Complex Systems**, v.06, n.04, p.565-573. 2003.

GOODMAN, L. A. Snowball Sampling. **The Annals of Mathematical Statistics**, v.32, p.148-170. 1961.

GUHA, R., *et al.* Propagation of trust and distrust. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. 13., 2004. New York, NY, USA. **Proceedings...** New York, USA: ACM, p. 403-412.

HAN, Y.; LI, D.; WANG, T. Identifying different community members in complex networks based on topology potential. **Frontiers of Computer Science in China**, v.5, n.1, p.87-99. 2011.

HONG, L.; DAN, O.; DAVISON, B. D. Predicting popular messages in Twitter. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. 20., 2011. Hyderabad, India. **Proceedings...** New york, USA: ACM, p. 57-58.

HOOKER, B. Reputation and Popularity. **The North American Review**, v.195, n.676, p.404-413. 1912.

HUBERMAN, B. A. The laws of the web: Patterns in the ecology of information. **J. Am. Soc. Inf. Sci. Technol.**, v.53, n.11, p.969-970. 2002.

IBARRA, H. Homophily and Differential Returns: Sex Differences in Network Structure and Access in an Advertising Firm. **Administrative Science Quarterly**, v.37, n.3, p.422-422. 1992.

JAMALI, M.; ABOLHASSANI, H. Different Aspects of Social Network Analysis. In: INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE. 4., 2006. Hong Cong - CH. **Proceedings...** New York, USA: IEEE/WIC/ACM p. 66-72.

JIANWEI, W.; LILI, R.; TIANZHU, G. A New Measure of Node Importance in Complex Networks with Tunable Parameters. In: INTERNATIONAL CONFERENCE ON WIRELESS COMUNICATIONS. 4., 2008. Dalian, CH. **Proceedings...** IEEE, p. 1-4.

JØSANG, A.; ISMAIL, R.; BOYD, C. A survey of trust and reputation systems for online service provision. **Decision Support Systems**, v.43, n.2, p.618-644. 2007.

KIM, J. S., *et al.* Fractality in complex networks: Critical and supercritical skeletons. **Physical Review E**, v.75, n.1, p.016110. 2007.

KITSAK, M., *et al.* Betweenness centrality of fractal and nonfractal scale-free model networks and tests on real networks. **Physical Review E**, v.75, n.5, p.056115. 2007.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. In: PROCEEDINGS OF THE NINTH ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS. 1998. San Francisco, California, United States. 315045: Society for Industrial and Applied Mathematics, p. 668-677.

_____. Authoritative sources in a hyperlinked environment. **J. ACM**, v.46, n.5, p.604-632. 1999.

KUMAR, R.; NOVAK, J.; TOMKINS, A. Structure and evolution of online social networks. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. 12., 2006. Philadelphia, PA, USA. **Proceeedings...** New York, USA: ACM, p. 611-617.

KWAK, H., *et al.* What is Twitter, a social network or a news media? In: INTERNATIONAL CONFENRENCE ON WORLD WIDE WEB. 19., 2010. Raleigh, North Carolina, USA. **Proceedings...** New York, USA: ACM, p. 591-600.

LESKOVEC, J.; KLEINBERG, J.; FALOUTSOS, C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In: INTERNETIONAL CONFENRENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. 11., 2005. Chicago, Illinois, USA. **Proceedings...** New York, USA: ACM, p. 177-187.

LESKOVEC, J.; HORVITZ, E. Planetary-scale views on a large instant-messaging network. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. 17., 2008. Beijing, China. **Proceedings...** New York, USA: ACM, p. 915-924.

LONG, G.; XU, C. The Fractal Dimensions of Complex Networks. **Chinese Physics Letters**, v.26, n.8, p.088901. 2009.

MANDELBROT, B. **The Fractal Geometry of Nature**: W. H. Freeman. 1977

MATIA, K., *et al.* Scaling phenomena in the growth dynamics of scientific output: Research Articles. **J. Am. Soc. Inf. Sci. Technol.**, v.56, n.9, p.893-902. 2005.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. **Annual review of sociology**, v.27, p.415-444. 2001.

MILGRAM, S. The Small World Problem. **Psychology Today**, v.2, p.60-67. 1967.

MORÉ, J. The Levenberg-Marquardt algorithm: Implementation and theory. In: G. A. Watson (Ed.). Numerical Analysis: Springer Berlin Heidelberg, v.630, 1978. The Levenberg-Marquardt algorithm: Implementation and theory, p.105-116. (Lecture Notes in Mathematics)

MORENO, J. L. **Who Shall Survive: A New Approach to the Problem of Human Interrelations**: Nervous and Mental Disease Publishing Co. 1934

NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. **The Structure and Dynamics of Networks**. Princeton: Princeton University Press. 2006

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, v.45, n.2, p.167-256. 2003.

NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E**, v.69, n.2, p.026113. 2004.

NIEMINEN, J. On the centrality in a graph. **Scandinavian Journal of Psychology**, v.15, n.1, p.332-336. 1974.

NUNES AMARAL, L. A.; MEYER, M. Environmental Changes, Coextinction, and Patterns in the Fossil Record. **Physical Review Letters**, v.82, n.3, p.652-655. 1999.

O'DOHERTY, D.; JOUILI, S.; ROY, P. V. Towards trust inference from bipartite social networks. In: WORKSHOP ON DATABASES AND SOCIAL NETWORKS. 2., 2012. Scottsdale, Arizona. **Proceedings...** New York, USA: ACM, p. 13-18.

O'GRADY, L. Future directions for depicting credibility in health care web sites. **Int J Med Inform**, v.75, n.1, Jan, p.58-65. 2006.

O'REILLY, T. What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software 2005.

PAGE, L., *et al.* The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab. 1999. p., Disponível em: http://ilpubs.stanford.edu:8090/422/>. Acessado em: 15/10/2010.

POWELL, J. A.; DARVELL, M.; GRAY, J. A. M. The doctor, the patient and the world-wide web: how the internet is changing healthcare. **J R Soc Med**, v.96, n.2, p.74-76. 2003.

PURCELL, K.; BRENNER, J.; RAINIE, L. Search Engine Use 2012. Pew Research Center. Washington, D.C.: March, 9. 2012. 42 p., Disponível em: http://www.pewinternet.org/~/media//Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf>. Acessado em: 05/abril/2012.

RAVASZ, E.; BARABÁSI, A.-L. Hierarchical organization in complex networks. **Physical Review E**, v.67, n.2, p.026112. 2003.

RESNICK, P., *et al.* Reputation systems. **Commun. ACM**, v.43, n.12, p.45-48. 2000.

RIEH, S. Y.; DANIELSON, D. R. Credibility: A multidisciplinary framework. **Annual Review of Information Science and Technology**, v.41, n.1, p.307-364. 2007.

ROMERO, D. M., *et al.* Influence and passivity in social media. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. 20., 2011. Hyderabad, India. **Proceedings...** New York, USA: ACM, p. 113-114.

SAKAKI, T.; MATSUO, Y. How to Become Famous in the Microblog World. In: PROCEEDINGS OF THE FOURTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA (ICWSM2010). 2010. Washington, DC. AAAI p. 324-326.

SIEGEL, S.; CASTELLAN, N. J. **Nonparametric statistics for the behavioral sciences**. New York; London: McGraw-Hill. 1988

SIMPSON, B.; MARKOVSKY, B.; STEKETEE, M. Power and the perception of social networks. **Social Networks**, v.33, n.2, p.166-171. 2011.

VARLAMIS, I.; EIRINAKI, M.; LOUTA, M. A Study on Social Network Metrics and Their Application in Trust Networks. In: INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETOWRKS ANALYSIS AND MINING. 4., 2010. Odense, Denmark. **Proceedings...** New York, USA: IEEE, p. 168-175.

WANG, Y.; LIU, Z. Automatic detecting indicators for quality of health information on the Web. **International Journal of Medical Informatics**, v.76, n.8, p.575-582. 2007.

WASSERMAN, S.; FAUST, K. **Social network analysis**. Cambridge, MA: Cambridge: University Press. 1994

WASSERMAN, S. **Social network analysis : methods and applications**. Cambridge: Cambridge University Press. 1999

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of "small-world" networks. **Nature**, v.393, n.6684, p.440-442. 1998.

WATTS, D. J. The “New” Science of Networks. **Annu. Rev. Sociol.**, v.30, n.1, p.243-270. 2004.

WEITZEL, L.; QUARESMA, P.; DE OLIVEIRA, J. P. M. Evaluating Quality of Health Information Sources. In: INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION NETWORKING AND APPLICATIONS. 26., 2012. Fukuoka, JP. **Proceedings...** IEEE, p. 655-662.

YANG, M.-C., *et al.* Finding interesting posts in Twitter based on retweet graph analysis. In: INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. 35., 2012. Portland, Oregon, USA. **Proceedings...** New York, USA: ACM, p. 1073-1074.

YE, S.; WU, S. Measuring Message Propagation and Social Influence on Twitter.com. In: L. Bolc, M. Makowski, *et al* (Ed.). Social Informatics: Springer Berlin / Heidelberg, v.6430, 2010. Measuring Message Propagation and Social Influence on Twitter.com, p.216-231. (Lecture Notes in Computer Science)

ZAMAN, T. R., *et al.* Predicting Information Spreading in Twitter. In: WORKSHOP ON COMPUTATIONAL SCIENCE AND WISDOM OF CROWDS. 7., 2010. Whistler, Canada. **Proceedings...** New York, USA: ACM, p. 17599-17601.

ZHANG, B., *et al.* Collecting the internet AS-level topology. **SIGCOMM Comput. Commun. Rev.**, v.35, n.1, p.53-61. 2005.

ZHOU, W.-X.; JIANG, Z.-Q.; SORNETTE, D. Exploring self-similarity of complex cellular networks: The edge-covering method with simulated annealing and log-periodic sampling. **Physica A: Statistical Mechanics and its Applications**, v.375, n.2, p.741-752. 2007.

ZHOU, Z., *et al.* Information resonance on Twitter: watching Iran. In: WORKSHOP ON SOCIAL MEDIA ANALYTICS. 1., 2010. Washington D.C., District of Columbia. **Proceedings...** New York, USA: ACM, p. 123-131.

ANEXO I: TOP-50 USUÁRIOS RELEVANTES RECORRENTES

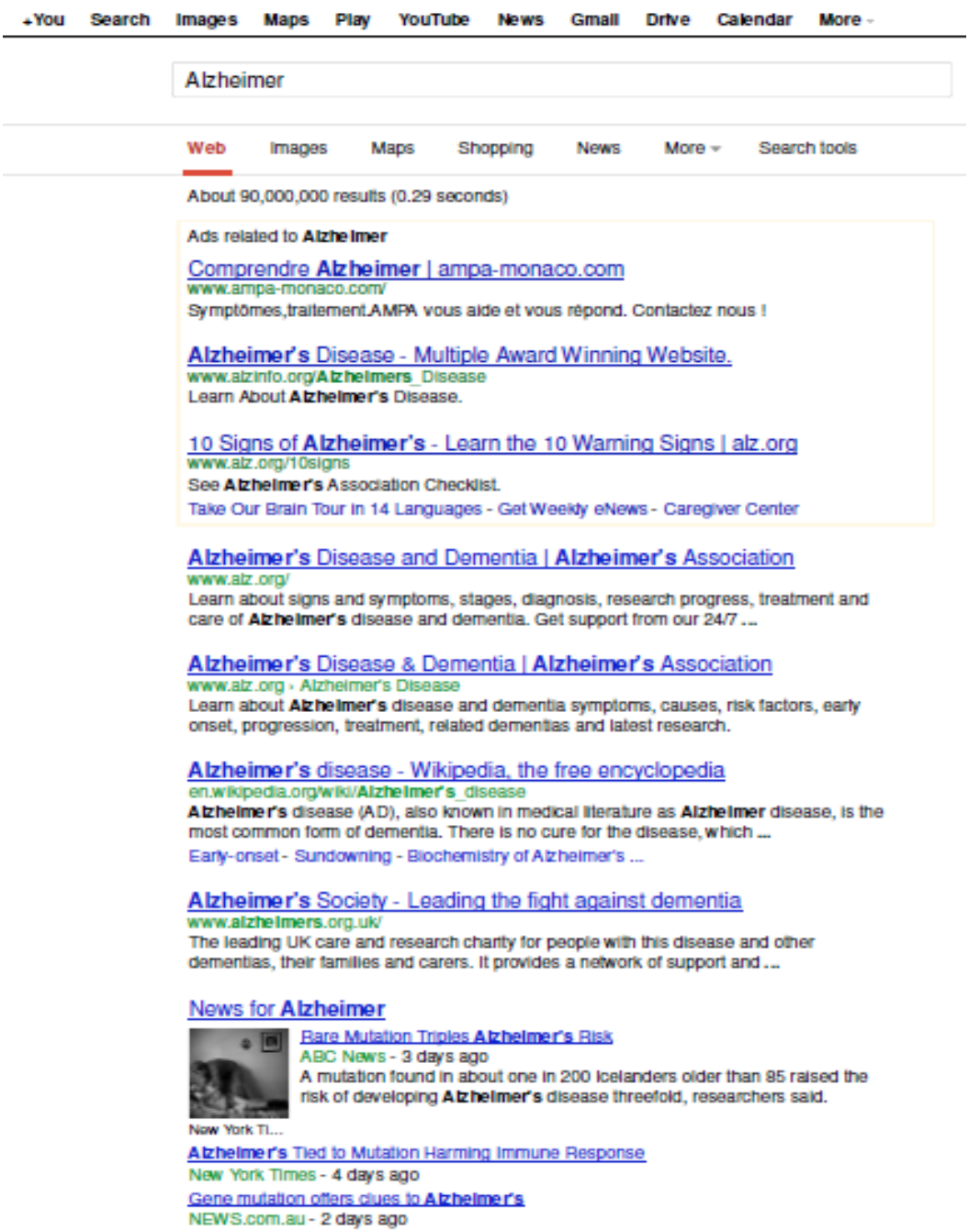
Tabela 0-1: Listagem de classificação final dos usuários. Top -50 recorrentes

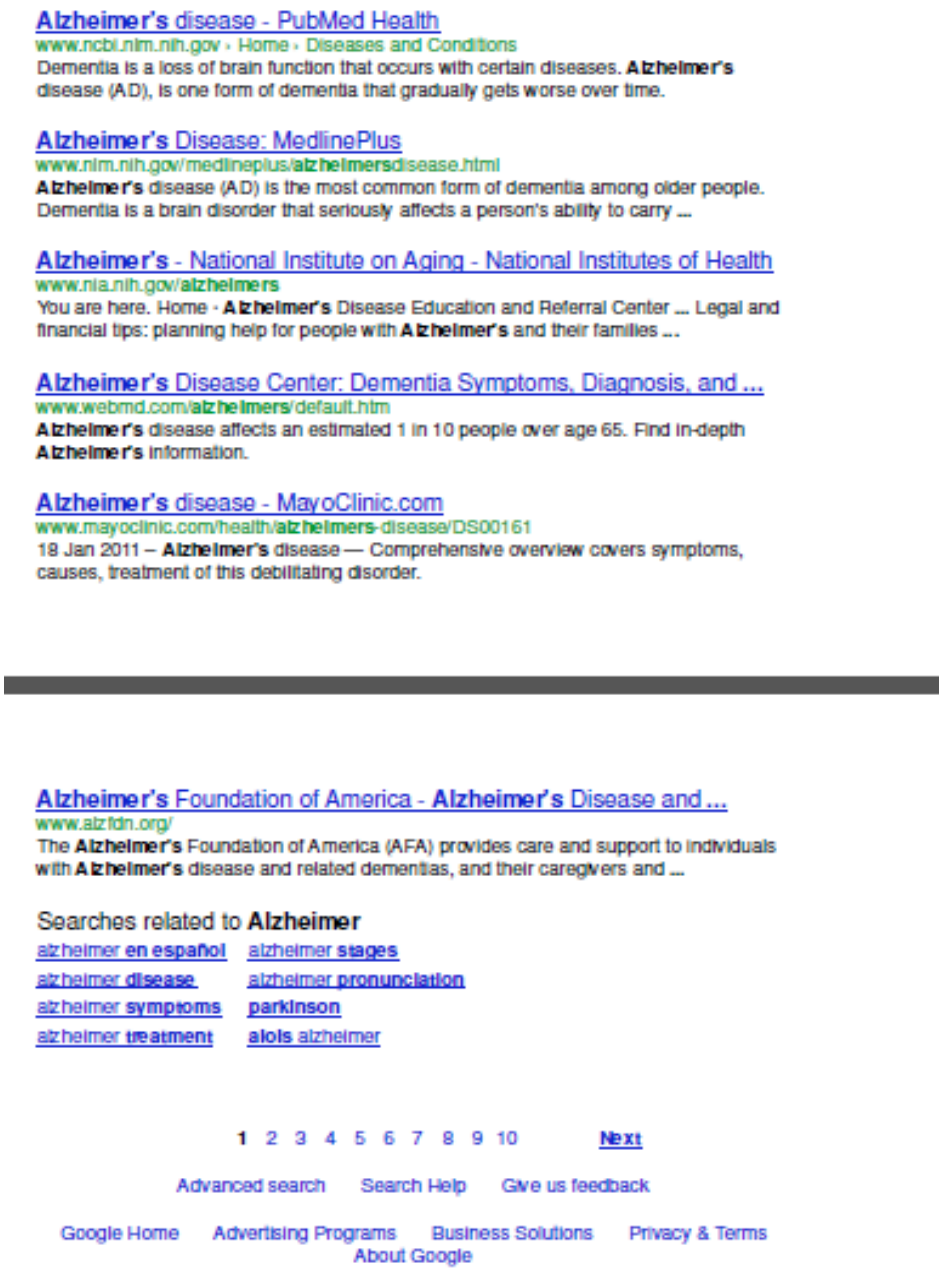
Tipo	Rank	Nome	Following	Follower	*Tweet*	Local	Time zone	Data registro
Saúde Governo EUA	1	AIDSgov	1757	164893	3209	Washington, DC	Eastern Time (US & Canada)	03/12/2008 18:09
Saúde Governo EUA	2	CDCgov	192	92268	3421	Menlo Park, CA | Washington DC	Eastern Time (US & Canada)	23/10/2009 15:17
Saúde Governo EUA	3	HHSGov	121	184586	1104	US, the Americas, South Africa	Pacific Time (US & Canada)	10/02/2009 21:44
Saúde Governo EUA	4	healthfinder	5650	164223	4576	Oakland, California	Pacific Time (US & Canada)	06/07/2011 17:18
Saúde Governo EUA	5	CDCNPIN	711	11102	5970	Atlanta, GA	Central Time (US & Canada)	15/11/2010 20:46
Saúde Governo EUA	6	CDC_eHealth	141	237850	1112	New York	Eastern Time (US & Canada)	01/10/2008 14:33
Saúde Governo EUA	7	womenshealth	1204	425156	5167	Washington, DC	Eastern Time (US & Canada)	24/11/2008 19:36
Saúde Governo EUA	8	talkhiv	877	5944	6488	Boston, MA	Quito	03/09/2008 15:25
Saúde Governo EUA	9	HealthCareGov	17	47663	327	Washington, DC	Eastern Time (US & Canada)	22/12/2008 13:48
Saúde Governo EUA	10	PublicHealth	2307	178870	2424	Washington, DC	Eastern Time (US & Canada)	13/05/2010 13:00
Saúde Governo EUA	11	CDCSTD	432	9405	883	Washington, DC USA	Eastern Time (US & Canada)	21/03/2009 20:03

Governo EUA	12	whitehouse	162	2895037	6317	worldwide	Eastern Time (US & Canada)	23/02/2009 18:17
Saúde Governo EUA	13	NIAIDNews	529	11575	1135	Chicago	Mountain Time (US & Canada)	16/02/2010 19:21
Saúde Governo EUA	14	CDC_cancer	36	20728	1089	Boston	Eastern Time (US & Canada)	26/01/2009 19:36
Saúde Governo EUA	15	NIOSH	2870	157230	7335	Worldwide	New Delhi	24/09/2007 03:28
Saúde Governo EUA	16	GoHealthyPeople	6280	16571	1929	SPHERE@healthimperatives.org	Eastern Time (US & Canada)	04/09/2009 11:49
Governo EUA	17	StateDept	318	282361	19857	Atlanta, GA	Eastern Time (US & Canada)	29/09/2009 02:20
Saúde Governo EUA	18	DesertAIDS	631	1288	3796	USA	Eastern Time (US & Canada)	15/01/2010 20:06
Saúde Governo EUA	19	FDATobacco	497	9749	1088	Princeton, N.J.	Eastern Time (US & Canada)	23/06/2009 14:37
Saúde Governo EUA	20	MinorityHealth	62	8509	1679	Washington, DC	Eastern Time (US & Canada)	26/01/2010 15:48
Saúde Governo EUA	21	AIDSinfo	27	3212	735	Não disponivel	Eastern Time (US & Canada)	24/04/2008 14:07
Saúde Governo EUA	22	Pew_Internet	0	69	4	United States of America	Eastern Time (US & Canada)	10/09/2009 18:16
Saúde Governo EUA	23	NIHforHealth	141	229086	2393	Columbia, MD	Eastern Time (US & Canada)	30/01/2009 15:24
Saúde Governo EUA	24	cdcemergency	106	1345286	551	Não disponivel	Eastern Time (US & Canada)	03/04/2008 13:54
Saúde Governo EUA	25	FDA_Drug_Info	15	60278	948	Washington, DC	Central Time (US & Canada)	05/03/2010 18:03
Saúde Governo EUA	26	CDCFlu	43	194651	419	United States of America	Eastern Time (US & Canada)	23/04/2009 12:40
Saúde Governo EUA	27	NASTAD	176	1056	1034	U.S.	Atlantic Time (Canada)	28/08/2008 15:33
Saúde Governo EUA	28	SLVHealthDept	4239	4768	3895	Atlanta, GA	Eastern Time (US & Canada)	20/04/2010 20:30
Saúde Governo EUA	29	CDCMMWR	18	6132	202	Portland, OR	Eastern Time (US & Canada)	30/03/2007 23:19
Saúde Governo EUA	30	americancancer	197974	275943	2399	Não disponivel		23/08/2011 07:43
Saúde Governo	31	NDEP	140	11655	1012	Não disponivel	Alaska	04/01/2010 15:43

EUA								
Saúde Governo EUA	32	BTPMay19	919	770	5782	New York	Eastern Time (US & Canada)	24/06/2008 22:37
Saúde Governo EUA	33	blackaids	43	1889	1675	San Francisco, California	Pacific Time (US & Canada)	03/11/2009 18:39
Saúde Governo EUA	34	aids2010	130	2381	411	San Francisco	Pacific Time (US & Canada)	16/07/2008 19:10
Saúde Governo EUA	35	redpumpproj	1147	5243	4711	Não disponivel	Quito	29/06/2011 13:11
Saúde Governo EUA	37	samhsagov	176	15999	2727	Washington, DC	Eastern Time (US & Canada)	20/08/2008 21:05
Saúde Governo EUA	38	harvardHSPH	1996	22959	2872	New York City	Central Time (US & Canada)	03/11/2009 16:11
Saúde Governo EUA	39	girlshealth	314	280477	3121	New York, NY	Eastern Time (US & Canada)	18/02/2009 18:54
Saúde Governo EUA	40	Texting4Health	1001	3500	883	Vienna		15/06/2010 08:21
Saúde Governo EUA	41	NLAAD	1134	887	570	Boston, MA	Eastern Time (US & Canada)	30/11/2007 19:57
Saúde Governo EUA	42	NAPWAUS	1370	1382	438	Washington DC	Eastern Time (US & Canada)	09/07/2008 18:06
saude privado	44	SusannahFox	670	9317	12732	Atlanta, GA	Eastern Time (US & Canada)	21/05/2010 19:40
saude privado	45	KaiserFamFound	3	15609	1977	Washington, D.C.	Quito	05/06/2009 01:14
saude privado	46	HelpEndHIV	1923	3618	1052	Washington, DC	Eastern Time (US & Canada)	26/03/2009 18:20
Saúde Governo EUA	47	HUDNews	247	27876	3052	Atlanta, GA	Eastern Time (US & Canada)	01/06/2009 16:01
saude privado	48	SexTech	0	68	2	Atlanta, GA	Eastern Time (US & Canada)	24/07/2008 19:35
Congresso	49	2011NHPC	93	520	211	US	Eastern Time (US & Canada)	30/05/2007 18:39
saude privado	50	amfAR	569	12942	1860	Atlanta	Eastern Time (US & Canada)	02/03/2010 15:12

ANEXO II: PÁGINAS RECUPERADAS





Alzheimer's disease - PubMed Health
www.ncbi.nlm.nih.gov › Home › Diseases and Conditions
Dementia is a loss of brain function that occurs with certain diseases. **Alzheimer's** disease (AD), is one form of dementia that gradually gets worse over time.

Alzheimer's Disease: MedlinePlus
www.nlm.nih.gov/medlineplus/**alzheimers**disease.html
Alzheimer's disease (AD) is the most common form of dementia among older people. Dementia is a brain disorder that seriously affects a person's ability to carry ...

Alzheimer's - National Institute on Aging - National Institutes of Health
www.nia.nih.gov/**alzheimers**
You are here. Home · **Alzheimer's** Disease Education and Referral Center ... Legal and financial tips: planning help for people with **Alzheimer's** and their families ...

Alzheimer's Disease Center: Dementia Symptoms, Diagnosis, and ...
www.webmd.com/**alzheimers**/default.htm
Alzheimer's disease affects an estimated 1 in 10 people over age 65. Find in-depth **Alzheimer's** information.

Alzheimer's disease - MayoClinic.com
www.mayoclinic.com/health/**alzheimers**-disease/DS00161
19 Jan 2011 – **Alzheimer's** disease — Comprehensive overview covers symptoms, causes, treatment of this debilitating disorder.

Alzheimer's Foundation of America - Alzheimer's Disease and ...
www.alzfdn.org/
The **Alzheimer's** Foundation of America (AFA) provides care and support to individuals with **Alzheimer's** disease and related dementias, and their caregivers and ...

Searches related to **Alzheimer**

alzheimer **en español** | alzheimer **stages**
alzheimer **disease** | alzheimer **pronunciation**
alzheimer **symptoms** | **parkinson**
alzheimer **treatment** | **aiols** alzheimer

1 2 3 4 5 6 7 8 9 10 **Next**

Advanced search Search Help Give us feedback

Google Home Advertising Programs Business Solutions Privacy & Terms
About Google

Resultados retornados da busca feita na ferramenta Google no dia 19.11.2012 pela palavra chave Alzheimer. A busca foi feita no idioma inglês.