



Pós-Graduação em Ciência da Computação

“Mensurando o Valor de Membros de Redes Sociais Digitais”

By

Raony Mascarenhas de Araújo

M.Sc. Dissertation



Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, AGO/2010



Universidade Federal de Pernambuco
Centro de Informática
Pós-graduação em Ciência da computação

Raony Mascarenhas de Araújo

“Mensurando o Valor de Membros de Redes Sociais Digitais”

Trabalho apresentado ao Programa de Pós-graduação em Ciência da computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obteno do grau de Mestre em Ciência da Computação.

A M.Sc. Dissertation presented to the Federal University of Pernambuco in partial fulfillment of the requirements for the degree of M.Sc. in Computer Science.

Advisor: *Silvio Romero de Lemos Meira*

RECIFE, AGO/2010

*Eu dedico esta dissertação para minha família, amigos e
professores que me deram todo o apoio necessário para
chegar aqui.*

Agradecimentos

Eu gostaria de agradecer ao nosso Pai em primeiro lugar, aos meus pais e família em segundo; a todos os meus amigos, principalmente aqueles que seguraram a barra nos diversos deveres que sacrifiquei para alcançar esse objetivo; ao meu orientador e a todos os que compõem o programa de pós-graduação do CIn por realizarem um excelente trabalho; e a todas as pessoas que de alguma forma contribuíram para essa realização, que o nosso Pai possa somar os meus pequenos agradecimentos às bençãos que já distribui para todos nós.

*Pois eu vos digo que a todo o que tem, mais lhe será dado; mas ao que
não tem, até aquilo que tem ser-lhe-á tirado.*

—LUCAS (19:26)

Resumo

O valor de um membro em uma rede social é a influência que ele tem sobre os processos da rede. Essa influência pode estar fragmentada em múltiplas conexões periféricas ou em poucas conexões centrais, em diferentes contextos ou em uma área de especialização, em toda a rede ou em comunidades específicas. O reconhecimento de atores-chaves nos processos de difusão de inovações em redes sociais pode levar a campanhas de marketing viral mais eficientes, a identificação de especialistas, a reconhecer fraquezas na segurança da rede e a promover a adoção de conhecimentos em redes sociais de aprendizagem. O presente estudo reúne as abordagens utilizadas para a mensuração de redes sociais digitais, apresenta seus desafios e também suas possíveis soluções dentro de uma metodologia e vocabulário até então pouco explorados na literatura do assunto.

Palavras-chave: social network, social influence, SNA, KDDM, tie strength, relationship modeling, viral marketing, information diffusion;

Abstract

Sumário

Lista de Acrônimos	xii
1 Introdução	1
2 Redes Sociais	2
2.1 ⊗ Alguma Formalização	3
2.2 Redes sociais digitais	3
3 Mensurando a rede	5
4 Entendendo o domínio do problema	7
4.1 Dificuldades e critérios na mensuração	7
5 Entendendo os dados	13
5.1 Por uma tipologia das interações digitais	14
5.1.1 Por conteúdo	15
5.1.2 Por abrangência	16
5.1.3 Por intenção	17
5.1.4 Interações passivas	17
5.1.5 ⊗ Notação	18
5.2 Uma teoria da atenção como capital social	18
5.2.1 ⊗ Formalização	20
6 Preparação dos dados	21
6.1 Interações não-textuais	22
6.2 Interações textuais	23
6.2.1 ⊗ Formalização	25
6.3 Critérios de escolha da representação	27
6.3.1 ⊗ Mínima redundância	28
6.3.2 Redundância das conexões	29
6.3.3 ⊗ Redes discretas e/ou esparsas	29
6.3.4 Redes contínuas densas	30
7 Análise da influência	31
7.1 Sobre Proeminência	31

7.1.1	Tipo de envolvimento	33
7.1.2	Propriedade do caminho	33
7.1.3	Redundância na proeminência	34
7.1.4	Relevância na proeminência	34
8	Conclusão e trabalhos futuros	37
	Bibliografia	39
	Appendices	49
A	Estudo de caso: A.M.I.G.O.S.	50
A.1	Entendendo o domínio	50
A.2	Entendendo os Dados	53
A.3	Preparando os dados	55
A.4	Análise da influência	58

Lista de Figuras

4.1	Poder preditivo das sete dimensões da força da conexão (Gilbert and Karahalios, 2009)	10
7.1	Gráfico estrela	32
A.1	<i>Screenshot</i> da interface do A.M.I.G.O.S.	53
A.2	Esforço relativo para cada passo do KDDM (Cios and Kurgan, 2005) . .	55
A.3	Exemplo de modelagem no <i>framework</i>	56
A.4	Distribuição da Probabilidade Cumulativa da Atenção	56
A.5	Visualização da rede de contatos do a.m.i.g.o.s.	57
A.6	Visualização da rede de tópicos do a.m.i.g.o.s.	57
A.7	Visualização da rede de histórias do a.m.i.g.o.s.	57
A.8	Visualização da rede de atenção completa do a.m.i.g.o.s.	58

Lista de Tabelas

4.1	Sumário dos componentes da força da conexão (Petróczi <i>et al.</i> , 2006)	11
A.1	Tamanho e propriedades das redes	58
A.2	<i>QAP</i> aplicado as quatro redes ($p. < 0,001$)	59
A.3	Os 9 membros mais bem posicionados na rede de Contatos	59
A.4	Os 9 membros mais bem posicionados na rede de Atenção	60
A.5	Correlação do índice de centralidade	60

Lista de Acrônimos

A.M.I.G.O.S. Ambiente Multimídia para Integração de Grupos e Organizações Sociais

API Application Programming Interface

KDDM Knowledge Discovery and Data Mining

ORM Object-Relational Mapping

SNA Social Network Analysis

SQL Structured Query Language

TIC Tecnologias da Informação e Comunicação

UFPE Universidade Federal de Pernambuco

1

Introdução

A análise de redes sociais tem sido utilizada para a investigação de temas tão diversos quanto a difusão de inovações (Coleman *et al.*, 1966), oportunidades de emprego (Granovetter, 1995), prevenção contra fraude (Neville *et al.*, 2005) e marketing (Domingos and Richardson, 2001). Muito dessa pesquisa inicial se baseia em redes pequenas em torno de indivíduos escolhidos por amostragem (Wasserman and Faust, 1994)(Newman *et al.*, 2006), porém a recente disponibilidade de informações relacionais na internet através de sites de relacionamentos permitiu o desenvolvimento da análise de redes em larga escala (Boyd and Ellison, 2007). Não obstante, o meio ainda carece de modelos dinâmicos de representação para redes sociais observadas a partir do fenômeno digital (Xiang *et al.*, 2010) e é justamente nesse ponto que o trabalho atual se concentra. Nosso objetivo é avaliar as dificuldades, parâmetros e modelos existentes para a representação e modelagem não-supervisionada de redes sociais digitais em larga escala e tempo real, especificamente para aplicações que façam uso da rede para identificar atores chaves em processos de difusão de conhecimento, inovações e recursos.

No Capítulo 2 introduziremos os conceitos de redes sociais, redes sociais digitais e mensuração. Depois aprofundaremos o conceito de mensuração da rede a análise de influência trazendo métodos da mineração de dados e *knowledge discovery* no Capítulo 3. A partir daí nos Capítulos 4, 5, 6 e 7 visitamos cada passo do processo delineando as dificuldades e soluções existentes no tratamento de dados relacionais, na agregação das interações em uma representação da rede e no análise de influência através de técnicas de cálculo de proeminência tradicionais; também introduziremos uma forma de agregar as interações através do montante de atenção investido. Finalmente encerramos com a conclusão e trabalhos futuros.

2

Redes Sociais

O estudo de rede sociais inicia nas décadas de 40 e 50, inicialmente voltado para o estudo de pequenos grupos de indivíduos e suas interações, a rede era mensurada através de observações, questionários e entrevistas (Wasserman and Faust, 1994). Diferentemente de outras ciências sociais que consideravam apenas os indivíduo e seus atributos, o estudo das redes sociais considera suas relações e os atributos dessas relações. A rede social é um fenômeno complexo envolvendo os relacionamentos de diversos atores em suas particularidades e que, através de um processo que chamamos de mensuração, pode ser traduzido em uma representação. Toda representação da rede social, por ser um modelo, é naturalmente parcial e enviesado. Comumente, as pesquisas de redes sociais trabalham com grafos onde os vértices são os atores e os arcos entre os vértices são as relações mensuradas; e matrizes, onde as linhas e colunas são conjuntos de atores e que a posição (i, j) da matriz representa o arco **do** ator i **para** o ator j . Para economizar repetições, no decorrer deste trabalho quando estivermos nos referindo ao fenômeno observado, utilizaremos o termo **rede social observada**, enquanto que os termos **representação** e **rede social** serão intercambiáveis.

Devido à disponibilidade de ferramentas matemáticas para o tratamento de grafos, a análise de redes sociais desenvolveu-se rapidamente construindo métodos e modelos estatísticos apropriados (Butts, 2009). A partir desse ferramental, o ramo das ciências sociais passou a quantificar diversos fenômenos antes considerados apenas do ponto de vista subjetivo, como a proeminência dos atores, que estaria relacionada com a sua centralidade no grafo.

2.1 \otimes Alguma Formalização

Representamos a rede como um grafo $G(V, E)$ onde V é um conjunto de n vértices (também chamados de atores) e E o conjunto de arestas que os conectam. Quando a representação é não-direcionada o par (i, j) é chamado de aresta e temos que $(i, j) \in E$ implica $(j, i) \in E$. No caso em que há direcionamento, essa afirmação não se sustenta e por isso $(i, j) \in E$ não implica em $(j, i) \in E$ necessariamente, além do que chamamos o par de arco de i para j . De maneira geral, sempre trataremos a representação como sendo direcionada.

Definimos a matriz de adjacência da rede X com n linhas e colunas, onde $x_{ij} = 1$, se $(i, j) \in E$ ou 0 de outra forma. Essa matriz de X é uma representação binária da rede. Podemos generalizar essa notação para incluir redes com valores discretos onde $0 \leq x_{ij} \leq m$ ou contínuos dispostos em um intervalo. Ambas as notações, de grafo e de matriz, serão usadas de forma intercalada neste trabalho.

2.2 Redes sociais digitais

Com a popularização da Internet é fato que pessoas se conectam umas às outras virtualmente por seu intermédio. Os mecanismos de interação à disposição vão da simples troca de mensagens, à venda e troca de produtos, à participação conjunta em jogos *multiplayer* massivos. Indo além do que sociólogo algum sonhou realizar no início dos estudos de redes sociais, grande parte dessas interações estão registradas, ou podem ser registradas eletronicamente a baixo custo, fornecendo uma quantidade nunca antes disponível de informações para estudos antropológicos e sociais da rede.

E assim tem sido, desde o nível micro com a análise dos conteúdos trocados entre as interações pontuais de alguns indivíduos (Recuero, 2008), passando por análise de potencial de marketing (Clemons *et al.*, 2007; Domingos and Richardson, 2001; Richardson and Domingos, 2002; Ma *et al.*, 2008), busca de pessoas (ADAMIC and ADAR, 2005), de especialistas (Ehrlich *et al.*, 2007), formação de grupos (Adamic *et al.*, 2003; Backstrom *et al.*, 2006; Kumar *et al.*, 2006), divulgação de notícias (Gruhl *et al.*, 2004), dinâmicas de prestígio (Salganik *et al.*, 2006; Song *et al.*, 2007).

Enquanto nosso objetivo é alcançar resultados similares as pesquisas anteriores de influência em redes sociais digitais, decidimos antes colocar a questão: como mensurar a rede social digital? Cada pesquisa teve seu critério: quantidade de e-mails trocados, recomendações, similaridades de perfil, participação nas mesmas comunidades. Dissemos

no começo que a rede social observada é um fenômeno que pode ser representado, mas que não é a representação em si, por esta razão, toda representação carrega um *bias*. Ora, ao acrescentarmos digital ao termo, queremos dizer que estamos tratando da observação do fenômeno através de mídias digitais; não mais das interações ao vivo e analógicas, mas através de ferramentas eletrônicas que permitem a fácil armazenagem, indexação e recuperação dessa informação.

Para responder essa questão precisamos definir quais ferramentas são essas. Uma resposta óbvia seria sites de relacionamento (ou sites de redes sociais), definido como sendo um espaço (virtual) em que seja possível 1) criar um perfil, 2) relacionar uma lista pública de amigos, 3) navegar por essa rede de perfis interligados (Boyd and Ellison, 2007). Porém tais sites são apenas um dentre muitos tipos de ferramentas que podem ser analisados, como por exemplo: fóruns, listas de discussão, sites de compartilhamento de conteúdo, comércio eletrônico, *blogs*, *microblogs* (e.g., Twitter), salas de bate-papo. Por questões de privacidade deixaremos de lado as formas pessoais de interação, como *instant messengers* e e-mails.

Ou seja, qualquer espaço (virtual) em que se é possível 1) identificar unicamente um ator, 2) mapear atores agentes e receptores a uma determinada interação com suas propriedades, pode ser insumo para a mensuração da rede. Mais adiante veremos que idealmente também será necessário demarcar a posição dessa interação no tempo, para possibilitar uma análise longitudinal da evolução da rede. Chamamos de **medianeiro** qualquer espaço (virtual) que satisfaça a condição acima.

Porque nos parece evidente a impossibilidade aplicar questionários ou entrevistas com centenas de milhares de atores, respondemos a questão de como mensurar a rede assim: através das interações encontradas nos medianeiros. A mensuração é um processo de mineração de dados e, portanto, sujeita a todos os empecilhos típicos do campo como informações incompletas, ruidosas, esparsas, redundantes. A questão que nos aparece agora é como as interações observadas combinam-se para formar tal rede e se ela é significativa para a análise de proeminência.

3

Mensurando a rede

Knowledge discovery é o processo maior “... não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis” (Fayyad *et al.*, 1996) e mineração de dados é uma de suas etapas. Chamamos de mensuração da rede social digital o processo de minerar dados provenientes de meios digitais de interação social para formar uma representação. Na última década alguns métodos norteadores para os projetos de mineração de dados foram propostos, dentre os quais escolhemos o método de 6 passos descritos por Cios and Kurgan (2005) e que consiste em sua forma geral:

- 1. Entendendo o domínio do problema** Neste passo devemos determinar os objetivos do projeto e aprender sobre as possíveis técnicas conhecidas para alcançá-los.
- 2. Entendendo os dados** Este passo inclui coletar os dados, decidir quais serão utilizados, priorizar atributos, verificar sua utilidade em relação aos objetivos. Os dados precisam ser verificados em termos de completude, plausibilidade, etc.
- 3. Preparação dos dados** Este é o passo chave do qual o sucesso de todo o processo depende; Ele geralmente consome metade de todo o esforço da mineração. Aqui, decidimos quais dados serão usados como entradas para quais técnicas de mineração do passo 4. O que pode envolver levantar amostragem de dados, executar testes de correlação e significância, remoção e correção de ruído, etc. Os dados tratados depois poderão serem processados para a seleção de características, algoritmos de extração para redução da dimensionalidade dos dados, derivação de novos atributos (discretização) e agregação dos dados (granularização). O resultado é um novo conjunto de dados que atendem aos requisitos específicos para servir de entrada às ferramentas de mineração selecionadas.
- 4. Mineração dos dados** Aplicação dos métodos de mineração selecionados. Apesar de

ser através das ferramentas de mineração que as novas informações são descobertas, sua utilização normalmente envolve menos esforço do que preparar os dados. Ferramentas de mineração reúnem diversos tipos de algoritmos como conjuntos *fuzzy*, métodos Bayesianos, computação genética, aprendizado de máquina, redes neurais, etc. Para uma visão mais detalhada desses algoritmos, nos referimos a Han and Kamber (2006).

5. Avaliação do conhecimento descoberto Interpretação dos resultados, verificando a relevância da informação encontrada. Somente os modelos aprovados são mantidos, todo o processo pode ser revisitado e ações alternativas que levem à melhoria dos resultados podem ser identificadas.

6. Usando o conhecimento descoberto Entrega do conhecimento produzido. Criação de um plano para monitorar sua utilização, documentação do projeto, extender sua aplicação para outros domínios.

O método de 6 passos para mineração de dados foi escolhido dentre outros possíveis devido ao seu viés acadêmico e modelo iterativo com ciclos de *feedback* explícitos que orientam o retorno a passos anteriores para a melhoria do processo (KURGAN and MUSILEK, 2006). Mostraremos agora, para cada passo do método quais as considerações necessárias, dificuldades e possíveis soluções no contexto da mensuração de redes sociais digitais para a análise da influência.

4

Entendendo o domínio do problema

No caso da análise da influência a mineração se divide em duas etapas a saber: a mensuração da rede propriamente dita e o reconhecimento de atores chave. Da segunda, depende a primeira. Caso a abordagem seja encontrar esses atores por simulação ou por algoritmos de otimização, então a representação deve ter um caráter probabilístico. Essa probabilidade pode ser inferida por um modelo a partir de uma primeira representação bruta das interações ou gerado através de propriedades comuns entre os atores. Por outro lado, existe a possibilidade de utilizar técnicas da análise de redes sociais tradicional como métricas de centralidade e prestígio para aproximar a influência do ator na rede. Nesse caso, a representação utilizada para o cálculo dessas métricas não precisa ser probabilística, mas também deve satisfazer alguns critérios que serão vistos mais adiante.

Em Domingos and Richardson (2001), os autores levantaram a questão: como selecionar o conjunto ótimo de atores para a influência na rede? Demonstraram que a solução dessa questão é *NP* e apresentaram três algoritmos de aproximação considerando uma representação probabilística da rede. Depois deles, outros vieram, mas todos voltados para modelos probabilísticos da rede. Devido a essa dependência, achamos que é necessário uma atenção maior sobre a mensuração da rede através de modelos probabilísticos devido a dificuldades que se apresentam, mas as considerações gerais que serão feitas não deixam de se aplicar também para representações não probabilísticas.

4.1 Dificuldades e critérios na mensuração

Apesar dos algoritmos para aprendizagem de máquina e mineração de dados estarem bem consolidados (Cios and Kurgan, 2005), a maioria deles não foram feitos para dados relacionais. Algoritmos tradicionais de mineração de dados buscam padrões considerando cada entrada de dados como independentes, mas dados relacionais possuem

o que chamamos de autocorrelação relacional.

Dados com autocorrelação relacional são aqueles em que as entradas possuem correlação entre si a depender de uma relação comum. Em ciências sociais, essa autocorrelação pode ser fruto de uma propriedade chamada de homofilia e que pode ser vista de maneira simples como pessoas parecidas tendem a formar laços e vice versa. Recentemente, algumas técnicas de mineração de dados tem sido desenvolvidas para tratar dados relacionais como os modelos relacionais probabilísticos, classificadores Bayesianos de primeira ordem e árvores de probabilidades relacionais (Jensen and Neville, 2003). Técnicas mais antigas incluem programação de lógica indutiva e a análise de redes sociais tradicional.

Recentes trabalhos na área utilizam de diversos tipos de interação para a mensuração da rede como por exemplo a similaridade entre classificação de produtos (Richardson and Domingos, 2002), similaridade em termos extraídos de mecanismos de busca na internet (MATSUO *et al.*, 2007) e co-autoria em artigos científicos (Kempe *et al.*, 2003). Em Xiang *et al.* (2010) temos uma mensuração combinando diversas interações observadas nos sites de relacionamento Facebook e LinkedIn, que possuem propostas diferentes de utilização. A existência ou ausência de cada interação, como a recomendação, troca de mensagens, marcação em fotos, é reunida em um vetor binário. Da mesma forma, a similaridade entre os atores é reunido em outro vetor binário, onde cada dimensão representa uma propriedade comparada e recebe 0 caso seja diferente, 1 se igual. A partir de então, os autores propõem um modelo de força da relação como variável escondida com causa na similaridade e com consequência na interação.

Para a relação força e interação os autores propõe um modelo discriminativo que possibilita o calculo da probabilidade condicional da interação a partir da força da relação, isso é feito por amostragem da população. Essa estimativa é usada para calibrar um modelo generativo que relaciona a similaridade à força da conexão, que é utilizado por sua vez para calcular a força entre dois atores quaisquer. Essa combinação de modelo discriminativo e generativo possibilita a inferência não supervisionada da força do laço, o que é um ponto positivo para esta abordagem, porém ela subestima outros fatores igualmente importantes.

A similaridade é apenas uma das possíveis componentes da força da conexão, propriedade conhecida como homofilia. Para um entendimento de um processo geral de mensuração da rede digital, precisamos remontar a um dos trabalhos fundantes da análise de redes sociais onde Granovetter lança, pela primeira vez, a hipótese:

(Granovetter, 1973) *A força da conexão [entre dois atores] é uma combinação (provavelmente linear) da quantidade de tempo, intensidade emocional,*

intimidade (confidência mútua) e serviços recíprocos que a caracterizam.

Granovetter também separa indicadores de prescritores, indicadores são componentes de fato da força da conexão, enquanto que prescritores são contingências contextuais que influenciam, mas não são componentes. Dentre os prescritores está a homofilia, o que quer dizer que a similaridade entre os atores influencia a conexão, mas é a intensidade, duração, intimidade e reciprocidade que indicam a sua força. Dessas quatro dimensões iniciais (tempo, intensidade, intimidade, reciprocidade), acrescentaram-se nove outras no decorrer de três décadas de pesquisa. Na Tabela 4.1 temos um apanhado dos indicadores propostos na literatura.

Em Gilbert and Karahalios (2009) encontramos uma análise da composição desses indicadores na força da conexão. Através de um modelo discriminativo os autores encontram a correlação linear entre 74 variáveis em cima das interações coletadas no site de relacionamento Facebook e a força da relação coletada através de questionário com uma amostra de 32 indivíduos. Sua análise leva em consideração sete indicadores de força, separando as variáveis em dimensões do tipo intensidade, intimidade, duração, reciprocidade, estrutural, emocional, similaridade. É interessante notar que duas dimensões escolhidas são consideradas como prescitoras da força e não indicadoras, que são a estrutura e a similaridade. Sobre similaridade já falamos, já as variáveis estruturais referem-se a padrões da rede como a transitividade, i.e., amigo de amigo seu tem maior chance de ser seu amigo também.

Os resultados mostram uma partição da força por dimensão que pode ser vista na Figura 4.1. As dimensões de intimidade, intensidade e duração são as três maiores componentes da força. As variáveis de estrutura, sozinhas, são as que menos utilidade tem para a previsão da força, porém quando analisada em par com outras dimensões demonstram alto de grau de interação. Nas palavras dos autores:

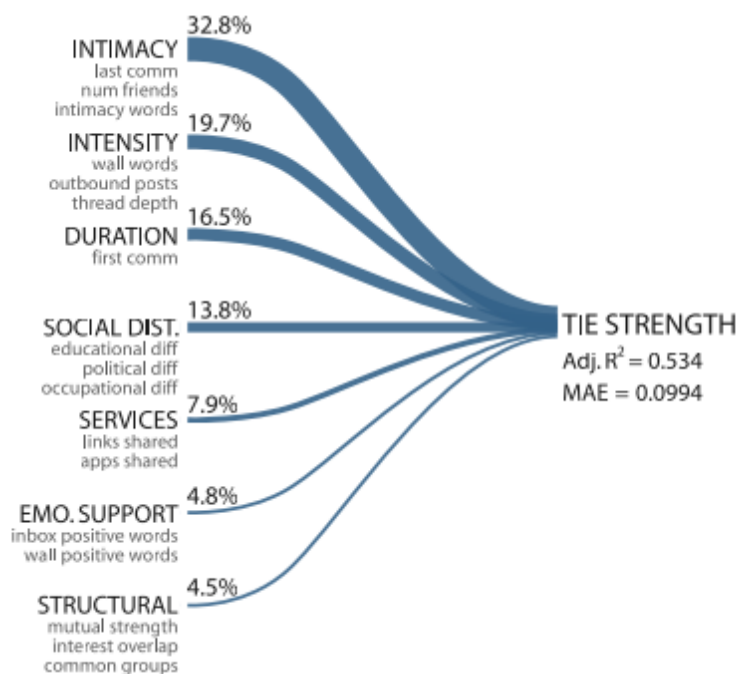
(Gilbert and Karahalios, 2009) A dimensão estrutural possui um papel menor como fator linear. Entretanto, ela possui um papel regulador importante através dessas interações [com as outras dimensões]. Uma forma de interpretar este resultado é que interações individuais importam, porém elas são filtradas através do clique¹ de amigos antes de impactar a força da relação.

Esse resultado confirma o papel prescritivo das variáveis estruturais. Não obstante os resultados, o método de Gilbert and Karahalios (2009) possui restrições quanto a utilidade

¹ clique pode ser traduzido como grupo de amigos com interesses similares e que são vistos por outros como exclusivistas; panelinha.

4.1. DIFICULDADES E CRITÉRIOS NA MENSURAÇÃO

Figura 4.1 Poder preditivo das sete dimensões da força da conexão (Gilbert and Karahalios, 2009)



em nosso cenário digital. Em primeiro lugar, sua abordagem supervisionada não escala com a rede e não há estudo sobre o erro induzido pela amostragem. Em segundo lugar, o tipo de conexão não é direcionado, isto é, ambos atores conferem o mesmo valor para a relação. Essa simplificação é nociva para o nosso objetivo de análise de influência, já que a assimetria de influência é bastante comum, por exemplo, um aluno pode ter grande consideração por um professor, mas o inverso em pé de igualdade raramente é verdade.

Isso nos leva a reflexão de que, para análise de influência, a utilização de todas as dimensões da força da conexão pode ser muito restritiva. Uma delas, como no exemplo já citado, é a dimensão de reciprocidade: um ator com prestígio exerce grande influência sobre seus admiradores, mas cada admirador individualmente não consegue exercer influência de igual intensidade sobre ele. Outro exemplo são os indivíduos de fronteira, que se conectam à periferia de dois grupos; as suas conexões são fracas, mas exerce grande influência na transferência de conhecimento entre os grupos. Daí temos que apesar da noção de força de Granovetter está relacionada com a influência, ela não é condição necessária (Brown *et al.*, 2007).

Finalmente, nenhuma das soluções estudadas considera a evolução da rede no tempo. Nesse aspecto algumas pesquisas utilizam fotografias na rede em momentos diferentes ou agregações das interações por período. A partir daí utiliza-se modelos probabilísti-

Tabela 4.1 Sumário dos componentes da força da conexão (Petróczi *et al.*, 2006)

Dimensão	Referências
Frequência	Benassi <i>et al.</i> (1999); Blumstein and Kollock (1988); Granovetter (1995); Marsden and Campbell (1984); MATHEWS <i>et al.</i> (1998); Mitchell (1987); Perlman and Fehr (1987)
Intimidade	Blumstein and Kollock (1988); Marsden and Campbell (1984); MATHEWS <i>et al.</i> (1998); Mitchell (1987); Perlman and Fehr (1987)
Investimento voluntário na relação	Blumstein and Kollock (1988); Perlman and Fehr (1987)
Aconselhamento dado e recebido	MATHEWS <i>et al.</i> (1998)
Desejo de companhia	Blumstein and Kollock (1988); Perlman and Fehr (1987)
Multiplicade dos contextos em que há interação	Blumstein and Kollock (1988); Granovetter (1973); Marsden and Campbell (1984); Perlman and Fehr (1987)
Duração	Blumstein and Kollock (1988); Granovetter (1973); Marsden and Campbell (1984); Perlman and Fehr (1987)
Reciprocidade	Blumstein and Kollock (1988); Friedkin (1980); Granovetter (1973); MATHEWS <i>et al.</i> (1998); Perlman and Fehr (1987)
Suporte emocional (intensidade)	Blumstein and Kollock (1988); Granovetter (1973); Mitchell (1987); Perlman and Fehr (1987); Wellman (1982); Wellman and Wortley (1990)
Confiança	Granovetter (1973); Marsden and Campbell (1984); MATHEWS <i>et al.</i> (1998)
Sociabilidade	Mitchell (1987)

4.1. DIFICULDADES E CRITÉRIOS NA MENSURAÇÃO

cos (Sarkar and Moore, 2005), meta-grupos (Berger-Wolf and Saia, 2006) ou modelos provenientes da psicologia comportamental e da administração (Brelger *et al.*, 2004). Nenhum deles, entretanto, é combinado com a teoria da força das relações para permitir sua aplicação em processos de mensuração de redes sociais digitais.

A partir de então sugerimos alguns critérios para as ferramentas de mineração que serão usadas no processo de análise de influência:

1. Deve ser não supervisionada para aumentar a chance de escalar com o tamanho da rede;
2. Deve considerar a força das relações na maior quantidade possível de seu espectro de componentes. Por tanto, a representação resultante não deve ser binária;
3. Deve também levar em consideração aspectos estruturais da rede como transitividade e *brokerage*;
4. Deve ser longitudinal, isto é, incorporar o elemento tempo na análise, levando em conta a formação e dissolução de relações, entrada e saída de atores;

Uma vez escolhida as ferramentas de mensuração e análise, devemos agora avançar para o passo seguinte e selecionar quais dados serão usados e se são válidos.

5

Entendendo os dados

A primeira crítica que deve ser avaliada é a de que interações digitais não são um indicador confiável da relação entre dois indivíduos (Clemons *et al.*, 2007). Que se um amigo virtual não é mais do que um conhecido, não há relação de influência significativa entre eles. Porém, as pesquisas em redes sociais digitais tem demonstrado que em sua maior parte, os indivíduos utilizam os meios digitais para continuar uma relação já existente no mundo *off-line* (Haythornthwaite, 2005; Recuero, 2008; Sassen, 2002). Outra vantagem de ater-se a interação observada é que ela não carrega alguns pontos fracos da abordagem direta (questionário, entrevista) que é o esquecimento e a omissão das relações nas respostas (a taxa de erro quando interrogados sobre as interações que mantiveram chega a 50% em comparação com a observação) (Mislove *et al.*, 2007). Mas a vantagem mais óbvia e determinante é o custo, coletar e processar os dados advindo das interações dos atores no meio digital é muito mais barato que fazer o mesmo para interações não digitais, na medida em que o tamanho da rede aumenta, ou mesmo em comparação com entrevistas e questionários. Na casa dos milhões de membros na rede, qualquer coisa que não seja a primeira opção é atualmente inviável.

Nessa etapa da mineração, uma vez delineada as técnicas de mensuração, é necessário escolher quais medianeiros serão utilizados. Recapitulando a definição de medianeiro: é todo espaço (virtual) em que seja possível 1) definir unicamente um ator, 2) mapear atores agentes e receptores a uma interação e suas propriedades. Alguns exemplos são: salas de batepapo, fóruns, lista de discussão, sites de relacionamentos, sites de compartilhamento de fotos e vídeos. Para cada medianeiro, o pesquisador pode ter maior ou menor acesso à informação.

Quando a única informação disponível são as publicadas nas páginas da internet ou outros meios de acesso ao medianeiro, dizemos que sua análise é **extrínseca**. A grande maioria dos trabalhos em redes sociais digitais hoje é feito dessa forma com a confecção

de *crawlers*, programas de computador que percorrem conteúdos *on-line* retirando informações estruturadas de dados semi-estruturados próprios para o uso humano, como as *web pages*. Esse tipo de análise é limitada muitas vezes à interação presumida, já que geralmente através dela é possível saber que um ator publicou conteúdo na comunidade, mas não quem da comunidade parou para vê-lo.

Quando as informações são coletadas diretamente dos banco de dados do medianeiro, chamamos essa análise de **intrínseca**. A análise intrínseca encerra diversas vantagens, principalmente por ter acesso ao funcionamento da rede, podendo coletar dados sobre seu uso. São informações como mensagem lidas e não lidas, tempo usado em cada uma, e que formam um tipo de interação que chamaremos de **passiva**.

Porém muitos são os tipos de interação digitais encontradas e enquanto não tivermos um maior entendimento das suas semelhanças e diferenças, não nos será possível construir uma abordagem integrada de mensuração. Por uma tipologia das interações digitais é então que nossa atenção deve agora se voltar.

5.1 Por uma tipologia das interações digitais

O estudo das interações humanas é o objeto das ciências sociais, notadamente, no contexto micro, da antropologia e etnografia. Somente com a popularização da internet é que as interações digitais (mediadas por computador) ocuparam maior destaque nesse meio (Wellman *et al.*, 1996; Herring, 2002). Uma tipologia inicial foi proposta por Burnett (2000) e revisada por Burnett and Buerkle (2004), porém o seu foco é na distinção entre interações direcionadas e as não-direcionadas para a aquisição de informação. Nossa tipologia começa dela, mas expande incorporando conceitos de capital social (Recuero, 2008) e abrangência (MARTÍNEZ and Figueroa, 2000).

Podemos as interações digitais dividir por tipo do **conteúdo**, **abrangência** e **intenção**. Os tipos de conteúdos podem ser divididos em dois grandes grupos: textuais e não textuais. Abrangência consiste se a interação é individual básica, individual desenvolvida, individual generalizada ou comum. Finalmente, quanto a intenção podemos classificar a interação em afirmativa, negativa, conversacional, informativa, conectiva. É importante ter em mente que as categorias apresentadas não são, de forma alguma, mutuamente exclusivas, podendo mesmo numa interação singular ser combinadas em diferentes formas.

5.1.1 Por conteúdo

O tipo do conteúdo da comunicação diz muito sobre o capital social que carrega (Kim *et al.*, 2007), por exemplo, mensagens síncronas possuem vocabulário mais limitado, são mais informais e possuem maior carga social fática (Danet *et al.*, 1998; Ko, 1996; WERRY, 1996) enquanto que mensagens assíncronas tendem a ser maiores, mais multifuncionais e linguisticamente complexas (Herring, 1999). De acordo com essas diferenças, mensagens síncronas parecem ser mais apropriadas para a interação social enquanto que mensagens assíncronas o são para discussões mais complexas e resolução de problemas.

Também não é nosso objetivo ainda nos aprofundarmos numa análise do discurso mediado por computador (Herring, 2001). A classificação que utilizaremos aqui usa o conceito de protótipo e por tanto tem expressividade reduzida diante do surgimento de formas interação inovadoras (Herring, 2007). Porém ela será suficiente nesse estudo inicial sobre mensuração de redes para a combinação de diferentes interações. Dividimos inicialmente em duas grandes categorias: textual e não textual; para cada uma então são listados seus principais protótipos.

conteúdo textual Já foi dito o quanto o texto é importante para a comunicação mediada por computador, mesmo numa época em que o compartilhamento de vídeos está na moda, o texto, na forma de comentários, continua sendo o principal móvel das trocas sociais (Herring, 2002). São desse tipo:

- comentários;
- mensagens;
- tópicos;
- *blogs* e similares;
- *microblogging* e similares;
- descrições.

conteúdo não-textual Nesse grupo estão relacionados não só as interações audiovisuais, mas também as interações “mudas” como a formação de laços explícito entre os atores, a classificação mútua dicotômica ou graduada e a recomendação de conteúdos de terceiros. Exemplos:

- fotos;
- vídeos;

- *links*;
- laços explícitos entre os atores;
- classificação (*rating*, *ranking*, favoritos);
- compartilhamento.

5.1.2 Por abrangência

A abrangência define quais são os atores influenciados pela interação, em outras palavras, os participantes do “discurso”(Dooley and Levinsohn, 2001). É através da abrangência da interação que o processo de mensuração recupera os atores participantes na conexão mensurada e por tanto representa a forma como os agentes buscam se posicionar na rede como um todo. Essa classificação foi apresentada uma primeira vez em (MARTÍNEZ and Figueroa, 2000).

individual básica Classificam-se neste grupo as interações de caráter privativo que partem de um ator específico para outro ator. As interações que tipicamente pertencem a este grupo são as mensagens, os pedidos de conexão e o compartilhamento de conteúdo do tipo “*forward*”.

individual desenvolvida Quando a interação se inicia em um ator e envolve sua rede imediata de contatos sem estar publicamente disponível para qualquer membro da rede. São desse grupo, em sua maioria: tópicos de fóruns, fotos, vídeos, compartilhamentos, microblogging.

individual generalizada Quando a interação se inicia em um ator e torna-se pública para todos os atores da rede. Todos os tipos de interação podem se classificar nesse grupo, depende do grau de livre acesso que a rede proporciona aos seus membros.

comum Quando a interação se dá num espaço de igualdade, quer dizer, que todos podem interagir com todos no mesmo nível, chamamos de interação comum. Um exemplo claro dessa categoria é as salas de batepapo onde todos podem publicar mensagens para todos lerem. Listas de discussão também seguem esse modelo. Um *Blog* em particular não é uma interação comum, na maioria dos casos, porque só o mantenedor do *blog* pode publicar artigos nele, já o espaço de comentários do artigo possa ser considerado um espaço comum.

5.1.3 Por intenção

Por último, mas não por menos, temos a intenção sobre a qual o ator reveste sua interação. A diferença de intenção não só pode representar mudança significativa na força da influência, como necessariamente define os possíveis resultados da interação. Em Recuero (2008) encontramos uma classificação da intenção das interações, que para o escopo proposto por esse trabalho é suficiente e as divide em cinco grandes grupos:

afirmativa Trata da afirmação de suporte social entre um ator e outro. Pode ser um comentário positivo relacionado a uma interação prévia do ator elogiado, uma avaliação positiva, a recomendação do seu conteúdo para outros.

negativa Quando a interação se dá para depreciar o outro. Pode ser por comentário, tópicos, mensagens, avaliação negativa.

conversacional Quando a interação tem caráter pessoal, relacionado a uma conversação que se inicia ou que está em andamento.

informativo Quando a intenção é informar um grupo de atores sobre determinado assunto. Avisos, artigos, propaganda, críticas.

conectivos Quando a intenção é formar laços explícitos através da rede. Pedidos para conexão como adicionar à lista de contatos/amigos/etc.

Das três dimensões de classificação da interação, a intenção é certamente a mais difícil de aferir computacionalmente. Isso porque sua característica subjetiva o que a torna também objeto ideal para a pesquisa qualitativa. Porém recente evolução da mineração de opinião e análise de sentimentos pode trazer opções para a análise de intenção no contexto da mensuração das redes sociais digitais. Para uma compilação dos principais avanços e desafios na área, nos referimos a Wilson *et al.* (2005); Ding and Liu (2007); Pang and Lee (2008). Por sua importância e dificuldade, teremos o cuidado de anotar mensurações de redes sociais digitais que não levem em consideração a intenção, como **análises ingênuas**.

5.1.4 Interações passivas

Para finalizar essa seção sobre interações, falaremos aqui de interações passivas e comportamentais que podem levar a um maior entendimento de como o ator investe sua atenção na rede e, por isso mesmo, como se posiciona na rede em relação a outros atores.

Podemos considerar como interação passiva como aquela que não é necessariamente percebida pelos outros atores além do interagente, nesse tipo encontram-se todas as interações do ator com o sistema como: mensagens lidas, não lidas, tempo usado para cada mensagem, perfis visitados, tempo usado na leitura de outros conteúdos. Essas informações poderiam ser utilizadas numa análise intrínseca da rede para a concepção de um modelo mais real do interagente quanto ao dispêndio da atenção.

5.1.5 \otimes Notação

Seja \mathcal{X} um medianeiro, existirá um conjunto \mathcal{L} de interações associado a \mathcal{X} . Considere o conjunto \mathcal{N} de atores envolvidos através do medianeiro \mathcal{X} . Temos que para cada interação $l_k \in \mathcal{L}$ existe pelo menos um ator $n_i \in \mathcal{N}$ que provocou a interação, chamamos esses atores de agentes ou autores. Também existe pelos menos um ator $n_j \in \mathcal{N}$ que recebeu a interação, chamamos esses atores de receptores, audiência ou abrangência. Assim para fins de notação, sendo $\mathcal{P}(\cdot)$ o conjunto das partes, considere $A : \mathcal{L} \rightarrow \mathcal{P}(\mathcal{N})$ onde $A(l_k)$ é conjunto de autores de l_k . Da mesma forma, faça $R : \mathcal{L} \rightarrow \mathcal{P}(\mathcal{N})$ onde $R(l_k)$ é conjunto de receptores de l_k . Nenhuma das duas funções são inversíveis, porém para economizar a notação definimos $A^{-1} : \mathcal{N} \rightarrow \mathcal{P}(\mathcal{L})$ onde $A^{-1}(n_i)$ é o conjunto de interações l tal que $n_i \in A(l)$. O mesmo deve ser considerado para R^{-1} .

5.2 Uma teoria da atenção como capital social

Antes de prosseguir para a etapa seguinte na mensuração da rede, gostaríamos de discutir uma aproximação para a influência entre os atores: atenção como capital social. Capital social é todo recurso que mantém a rede social (Coleman, 1988) e pode ser mobilizado através das conexões (Gyarmati and KYTE, 2004). Exemplos de capital social são a confiança e o suporte emocional. Por esta razão uma outra forma de chamar as dimensões de força da conexão de Granovetter é de capital social, i.e., quanto maior o fluxo de capital social que a conexão suporte, maior é a sua força. Mostraremos que a atenção não só atende a definição de capital social, como se aproxima mais do conceito de influência e, certamente, é mais fácil de medir.

Foi o vencedor do prêmio nobel, o economista Herbert Simon, que disse:

(Simon, 1996) *What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention*

Vivemos em uma era de riqueza de informação e por esta razão estamos em escassez de atenção (Goldhaber, 1997). Atenção é vista atualmente como o mais importante recurso para as organizações (Davenport and Beck, 2001). Ela também assume papel crucial na formação e manutenção de relacionamentos e como é um recurso limitado, o mais comum é que cada um invista nas relações que percebem ser mais importantes (Dindia and Canary, 1993). Por esta razão a atenção satisfaz os critérios de capital social, a saber 1) contribuir na manutenção da rede, 2) pode ser mobilizada através das conexões. Vários estudos consideram a atenção como a nova economia na internet, através de comentários, classificações positivas e *profiles* (Humphreys and Kozinets, 2009; Wu and Huberman, 2009; Skå geby, 2009). Nessa economia nós temos “celebridades” e “fãs”, muito similar ao que encontramos no contexto da influência.

Na verdade somente o fato de prestar atenção em si já é uma influência per si, independente das futuras escolhas do receptor. Em uma pesquisa em Parma, na Itália, por volta de 1991, neurocientistas conectaram cabos ao cérebro de macacos de forma que o computador reconhecia pelo padrão de ativação dos neurônios quando ele levava um amendoim à boca. Não obstante a ativação fosse registrada eletronicamente, eles também ligaram o aparato a um auto-falante de forma que mesmo estando distantes soubessem quando o evento acontecia. Quando um aluno passou pelo laboratório e levou uma banana à boca, o alarme soou. Porém o macaco não estava fazendo nada além de observar o estudante. Estava claro que a atenção no gesto do estudante disparava o cérebro a reproduzir (em sua mente) o movimento como se fosse seu (Rizzolatti *et al.* 1996; apud Goldhaber 2006).

Resta-nos saber como mediremos a atenção. Ora, em qualquer interação o que é trocado em primeiro lugar entre o agente e os receptores é atenção. A atenção flui não só do receptor para o agente ao receber a “mensagem”, no que chamamos de atenção **direta**, mas também do agente para os receptores, por a ter preparado e comunicado, que chamamos de atenção **residual** (Goldhaber, 1997). Apesar de que em quantias bem menores proporcionalmente, por que o que cada receptor recebe é uma atenção ilusória, é uma fração da atenção do agente e que o satisfaz de alguma forma na interação de modo que para ele parece um bom negócio continuar retribuindo-a com a sua. Mas como essa atenção se relaciona com as dimensões da força de Granovetter? Para responder essa questão precisaremos usar nossa tipologia da interação.

Em primeiro lugar, frequência e duração são dimensões temporais que são recuperadas a partir do tratamento longitudinal da rede. A frequência das interações e data do começo delas são métricas simples de colher e que contribuem para força total. Quanto

a intimidade, podemos verificar que a depender da abrangência teremos espaço para maior ou menor intimidade, ou seja, quando mais restrito a abrangência mais íntima a interação, sendo a mensagem de pessoa a pessoa a forma mais íntima possível. Quanto a carga emocional, mesmo não tendo à disposição meios de análise de sentimento, sabemos que as interações que mais nos chamam atenção são as que possuem carga emocional (Davenport and Beck, 2001). Daí podemos aproximar a dimensão de suporte emocional por uma combinação da de duração, frequência e intensidade.

Para finalizar nossa consideração sobre a atenção como capital social, é importante mencionar que a atenção não é direcionada apenas sobre um foco, mas sobre o contexto. Isto quer dizer que alguns indivíduos podem receber atenção indiretamente terem sido citados, por intermediarem a interação ou por simplesmente fazer parte do contexto de alguma forma. Essa atenção chamamos de **transitiva**.

5.2.1 \otimes Formalização

Considere dois atores i e $j \in \mathcal{N}$; se \mathcal{X} é um medianeiro para \mathcal{N} com um conjunto de interações \mathcal{L} , considere também l uma interação de \mathcal{L} . Se $i \in A(l)$ e $j \in R(l)$ então existe uma atenção direta $v_{ijl} \neq 0$ de i para j e uma atenção residual de j para i expressa como $\dot{v}_{jil} \neq 0$. Usaremos o símbolo $+$ em um índice quando queremos dizer que trata-se do somatório sobre todo o seu conjunto, alguns exemplos: $v_{i++} = \sum_{j \in \mathcal{N}} v_{ijl}$ representa toda a atenção direta investida por i na interação l ; $v_{ij+} = \sum_{l \in \mathcal{L}} v_{ijl}$, toda atenção direta cedida de i para j considerando todas as interações; e $v_{++l} = \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} v_{ijl}$, toda a atenção direta cedida na rede através da interação l .

Algumas restrições são óbvias, como o total de atenção investida um ator qualquer é: $v_{i++} + \dot{v}_{i++} = c, \forall i \in \mathcal{N}$. Essa restrição, no entanto, não se sustenta quando só temos as interações da rede para avaliar já que representam apenas uma fração da atenção total do ator.

Para definir a atenção transitiva é necessário primeiro estabelecer a relação $T : \mathcal{L} \rightarrow \mathcal{L}$ tal que se l e $q \in \mathcal{L}$ e uma faz parte do contexto da outra, isto é, estão encadeadas, então $q \in T(l)$ se q veio antes. Dizemos que há atenção transitiva \hat{v}_{ijl} se há um caminho entre i e j através de v , isto é, se $i \in R(l)$ e $j \in A(q)$ para algum $q \in T(l)$. Exemplos de interações encadeadas são comentários de *posts*, *threads* em fóruns e recomendação de conteúdo.

6

Preparação dos dados

Essa etapa confunde-se com a próxima que trata da aplicação de técnicas de mineração. Isso se dá porque, dependendo do objetivo do estudo, a mensuração da rede pode ser o fim ou o meio para a aplicação de um outro processo de análise. Por isso nessa seção falaremos um pouco de algumas técnicas de mineração quando usadas no contexto da preparação dos dados. Também estaremos atentos para algumas dificuldades que se apresentam quando utilizamos aprendizagem de máquina e faremos algumas considerações sobre técnicas para a escolha dos dados mais relevantes.

Trataremos aqui principalmente de mensurações da rede, que é a primeira parte do processo de análise de influência. A mensuração, por sua vez, se divide em dois momentos: dados **brutos** e dados **compilados**. No seu estado bruto, a rede guarda os dados como se lhe apresentam através dos medianeiros como por exemplo: quantidade de interações, tamanho das interações, classificações usadas, data da interação, etc. Esse processo por si só é trabalhoso, pois devemos ter o cuidado de guardar não só as interações, mas as propriedades relacionadas a elas. Também é comum a formação de hipergrafos, isto é, grafos onde os atores são agrupados por alguma relação (Breiger, 1974; Seidman and Foster, 1978). Um exemplo claro de hipergrafo é a representação formada a partir da afiliação dos atores à comunidades, contudo, a rigor, toda interação particiona o grafo em um subconjunto.

⊗ **Um exemplo de hipergrafo:** Considere dois medianeiros \mathcal{X}_1 e \mathcal{X}_2 com seus respectivos conjuntos de interações \mathcal{L}_1 e \mathcal{L}_2 . Ambos referem-se ao mesmo conjunto de atores $\mathcal{N} = \{n_1, n_2, n_3\}$. Para \mathcal{L}_1 temos duas interações l_1 e l_2 , onde n_1 é o autor de ambas, ou seja $A(l_1) = A(l_2) = \{n_1\}$. O ator n_2 é o receptor de l_1 e n_3 é o de l_2 , assim temos $R(l_1) = \{n_2\}$ e $R(l_2) = \{n_3\}$. Se construirmos uma representação de \mathcal{X}_1 chamada de X_1 , onde $x_{ij} = 1$ se existir uma interação l onde $i \in A(l)$ e $j \in R(l)$ então teremos:

	n_1	n_2	n_3
n_1	0	1	1
n_2	0	0	0
n_3	0	0	0

Agora para \mathcal{L}_2 temos apenas uma interação q onde $A(q) = \{n_1\}$ e $R(q) = \{n_2, n_3\}$. Então em apenas uma interação n_1 conseguiu a atenção de n_2 e n_3 . Se utilizarmos a mesma regra para construirmos X_2 também teremos:

	n_1	n_2	n_3
n_1	0	1	1
n_2	0	0	0
n_3	0	0	0

O que é no mínimo inconveniente, já que os dois medianeiros possuem padrões bem diferentes, o primeiro possui dois subgrafos e o segundo apenas um.

Nenhum das abordagens citadas no Capítulo 4 consideram o aspecto de hipergrafo das interações. Seria necessário adaptar as técnicas de análise de influência se quisermos considerar um ator que influencia muitos em uma única interação de outro que o faz através de muitas interações separadas. Tendo isso em mente, partimos para as técnicas de mensuração em si.

6.1 Interações não-textuais

Uma possível abordagem foi a citada no exemplo acima, a da ausência ou presença de interação. Ela foi usada por Xiang *et al.* (2010) e simplifica a representação em uma rede dicotômica. Outras possibilidades está em contar a quantidade de interações ocorridas, a frequência por período, a idade da primeira interação, a presença em comunidades comuns, similaridade de perfil, etc. De forma geral podemos criar a rede por:

- Intensidade/Quantidade
- Frequência
- Presença/Ausência

- Propriedade única (e.g., Longevidade)
- Similaridade

Cada uma dessas apresenta apenas uma dimensão da relação, de forma que para o mesmo medianeiro mais de uma representação pode ser mensurada a partir de diferentes formas de agregação. Outras representações ainda podem ser criadas a partir da combinação de mensurações diferentes, por exemplo: intensidade x frequência, longevidade x similaridade. Deve-se ter um cuidado extra no uso de agregações porque podem introduzir correlações espúrias no conjunto de dados, levando a erros do tipo I (falso positivo) e tipo II (falso negativo), pois a agregação mascara quais elementos estão correlacionados de fato (Jensen and Neville, 2003). Não existe forma correta de escolher as agregações, porém veremos mais adiante como poderemos diminuir o impacto disso no processo como um todo.

No final, teremos uma ou mais representações da rede para cada tipo de medianeiro, que por diferença de natureza não são facilmente integráveis. Contando, porém, com uma análise intrínseca dos medianeiros, poríamos apostar numa integração pela teoria da atenção. Apesar de que tempo gasto não se traduz diretamente em atenção dispendida, poderíamos mensurar o tempo que o usuário do medianeiro empregou em cada interação e daí inferir um valor em unidade única: atenção (Davenport and Beck, 2001). Com isso teríamos o tempo empregado na visualização de um vídeo, de uma foto como estimativa da atenção. Essa análise nem sempre é possível e mesmo que seja, há interações que são naturalmente difíceis de integrar, como as avaliações (*ratings*, *rankings*). As avaliações tomam sempre o mesmo tempo (o do clique do mouse) e seu conteúdo é totalmente “emocional”, isto é, polarizado entre positivo e negativo. Caso seja utilizado técnicas de análise de sentimentos para as outras interações, então talvez as avaliações possam ser integradas pela dimensão emocional da força da relação. Por outro lado, avaliações são excelentes opções para a utilização como grupo de teste para algoritmos de aprendizagem de máquina “supervisionados”.

6.2 Interações textuais

Na seção anterior investigamos as relações não-textuais, agora vamos nos deter nas textuais. Primeiramente, intentamos com essa divisão alcançar uma consequência prática: a integração de diversas interações numa só rede. A rede que vamos construir é valorada, isto é, as conexões possuem um peso na forma de um número real dentro de um intervalo.

Ao contrário de redes binárias, as redes valoradas proporcionam um indicativo da **força** do laço entre os atores.

Dentro de uma teoria da atenção como capital social, devemos então nos voltar para a afirmação de que laços por onde trafegam muitos recursos são laços fortes e o contrário também é verdade. Em cada interação na rede, como vimos, tem um ator-agente que inicia a interação e um grupo de abrangência, que chamaremos de atores-receptores, que é afetado por ela. Com esse modelo, podemos estimar que os receptores cedem atenção para o agente e a recíproca também é verdadeira, pois o agente escolheu interagir com este grupo e essa escolha já é um indicativo de atenção cedida.

Idealmente, essa atenção poderia ser mensurada a partir do tempo empregado pelos atores na interação, mas quase nunca essa informação está disponível. Por esta razão, sugerimos a quantidade de palavras na interação textual como uma estimativa para a quantidade de atenção trocada na interação. A sugestão advém naturalmente do fato de que quanto maior o texto, mais tempo é empregado na sua leitura, possivelmente mais recursos cognitivos também serão empregados para a sua apreensão. Isso nem sempre é verdade para todos os textos, ou tipos de textos, mas restringindo nosso escopo para comunidades virtuais de amigos e/ou comunidades de prática (Lave, 1991; Lave and Wenger, 1991), acreditamos ser esse um bom ponto de partida.

Outro cuidado necessário é na determinação se de fato a interação foi percebida pelos receptores. Na maioria das vezes essa determinação é inviável, aumentando a incerteza do modelo. Entretanto, um recurso simples que pode ser utilizado é procurar por interações encadeadas, isto é, quando uma interação posterior é reação a outra anterior. Por exemplo, quando uma mensagem é respondida, um texto é comentado, um conteúdo é recomendado, em todos esses casos podemos avaliar que o agente que reagiu não só percebeu a interação do agente anterior, como se deu o trabalho de responder. Em verdade, para interações que normalmente se encadeiam como *threads* de fóruns ou listas de discussão, podemos considerar a resposta como um sinal de atenção não só para com o participante imediatamente anterior, mas para vários antes dele que de certa forma influenciaram na resposta atual.

Assim, podemos construir uma rede onde os nós são os atores e os laços é o somatório da atenção trocada pelos mesmos. Quanto mais textos de um ator *a* lidos, respondidos, recomendados, avaliados por um ator *b*, maior será a força da relação direcionada de *b* para *a*. Quanto mais conteúdo um ator *a* publica para um determinado público, mais forte também será sua relação direcionada com o mesmo. Tal rede, por tanto, integraria insumos de diversos tipos de interação mensurados, desde que sejam de conteúdo textual.

6.2.1 \otimes Formalização

Para a formalização dessa métrica, vamos usar um exemplo prático. Em um fórum qualquer, vamos tomar um tópico para nosso exemplo. O fórum é o **medianeiro**, os membros assinantes do fórum os **atores**. O tipo de interação é **textual** quanto ao conteúdo, **comum** quanto a abrangência e nossa análise será **ingênua**, isto é, não consideraremos a intenção. Revisando:

- \mathcal{N} o conjunto de N atores e o i -ésimo ator é n_i , $i = 1, 2, \dots, N$;
- \mathcal{L} o conjunto de L tópicos do fórum e o k -ésimo tópico é l_k , $k = 1, 2, \dots, L$;
- $A : \mathcal{L} \rightarrow \mathcal{N}$ tal que $A(l_k)$ seja o autor do tópico l_k , a função A não é inversível, dado que um ator pode ser autor de mais de um tópico, porém considere a função $A^{-1} : \mathcal{N} \rightarrow \mathcal{P}(\mathcal{L})$ tal que $A^{-1}(n_i) = \{x \in \mathcal{L} | A(x) = n_i\}$ representando o conjunto de tópicos escritos pelo ator;
- $R : \mathcal{L} \rightarrow \mathcal{P}(\mathcal{N})$ tal que todo $x \in R(l_k)$ é um receptor de l_k , considere também o conjunto de tópicos recebidos por um ator como sendo $R^{-1} : \mathcal{N} \rightarrow \mathcal{P}(\mathcal{L})$ tal que $R^{-1}(n_i) = \{x \in \mathcal{L} | n_i \in R(x)\}$;
- Se $x \in \mathcal{L}$, $|x|$ é a quantidade de palavras de x , para todos os outros casos, quando aplicável, é a cardinalidade do conjunto;

Definimos a atenção total cedida de n_i para n_j , \vec{v}_{ij+} como a soma de suas atenções diretas v_{ij+} , residuais \dot{v}_{ij+} e transitivas \hat{v}_{ij+} . A atenção direta é aquela formada quando n_i é receptor de n_j numa interação l_k e definimos como:

$$v_{ijk} = \alpha |l_k| \quad (6.1)$$

A atenção residual é cedida quando n_j esta na abrangência da interação l_k cujo autor é n_i e definimos como:

$$\dot{v}_{ijk} = \beta |l_k| \quad (6.2)$$

A atenção transitiva é aquela n_j recebe de n_i se este é o autor de algum tópico l_k que responde a um anterior de autoria do primeiro. Para definirmos \hat{v} devemos antes modelar o encadeamento dos tópicos. Seja $T : \mathcal{L} \rightarrow \mathcal{L} \cup \{\emptyset\}$, se $l_h = T(l_k)$ então l_k é uma resposta a l_h , se l_k é a mensagem que iniciou o tópico então $T(l_k) = \emptyset$. Agora faça $T^* : \mathcal{L} \rightarrow \mathcal{P}(\mathcal{L})$:

$$T^*(l_k) = \begin{cases} \emptyset & \text{se } T(l_k) = \emptyset \\ \{T(l_k)\} \cup T^*(T(l_k)) & \text{em todo outro caso} \end{cases} \quad (6.3)$$

Daí temos que $T^*(l_k)$ é conjunto de todas as m mensagens anteriores na *thread* de l_k e podemos definir uma sequência t_{kq}^* , onde $q = 1, 2, \dots, m$, sobre esse conjunto considerando a ordenação trivial: $a > b \iff a \in T^*(b)$. Dessa forma l_k é uma resposta a t_{k1}^* , que por sua vez responde a t_{k2}^* e assim sucessivamente até a mensagem inicial. Levando em consideração a definição na Definição 6.3, definimos a atenção transitiva como:

$$\hat{v}_{ijk} = \sum_q (1 - \alpha) \gamma^q |t_{kq}^*|, \quad q \in \{x | A(t_{kx}^*) = n_j\} \quad (6.4)$$

A partir das Definições em 6.1, 6.2 e 6.4 podemos calcular a atenção total entre n_i e n_j como:

$$\vec{v}_{ij+} = E(n_i) \frac{v_{ij+} + \dot{v}_{ij+} + \hat{v}_{ij+}}{v_{i++} + \dot{v}_{i++} + \hat{v}_{i++}} \quad (6.5)$$

Evidentemente, este é um modelo exploratório para a mensuração de redes a partir de interações textuais sob um ponto de vista generalista. Nesse sentido há muito espaço para aperfeiçoamento dentro do campo experimental. Para finalizar, devemos fazer mais algumas considerações:

- \vec{v}_{ij+} tem sua soma normalizada devido ao limite de atenção natural de cada um;
 - Porém, cada pessoa dedica quantidades diferentes de atenção total aos medianeiros analisados, de forma que seria simplório normalizar todos em 1. Por esta razão, cada ator tem a soma de suas atenções distribuídas na rede normalizada em um fator $E(n_i)$ que indica o seu grau de atividade na rede. Esse fator, *a priori* será sempre relativo e caso o escolhamos como sendo a razão do total de atenção cedido pelo ator sobre a maior atenção medida, teremos uma normalização absoluta para toda a rede baseada no seu máximo. O que pode não ser a melhor escolha, pois idealmente a restrição de atenção deve servir para reduzir a variância das forças dos laços na rede;
 - O parâmetro γ representa a força da influência de mensagens antigas na *thread* sobre um determinado ator em sua participação atual. Por esta razão $0 \leq \gamma \leq 1$, sendo que $\gamma = 0$ anula toda a atenção transitiva e $\gamma = 1$ considera todas as mensagens anteriores como sendo igualmente influentes na mensagem atual;
-

- O parâmetro β representa a atenção recíproca do agente para os seus receptores, também varia de 0 a 1, sendo que $\beta = 0$ anula a atenção residual da composição e $\beta = 1$ indica que o autor da mensagem cedeu individualmente a cada receptor tanto quanto este para com ele. Um valor interessante para β é o inverso da quantidade de atores na abrangência da interação em contexto, isto é, $\beta = 1 / |A(I_k)|$. Assim, o autor cede uma fração igual para cada receptor que é tão menor quanto maior for seu público. Essa proporção também satisfaz a dimensão de intimidade da força do laço, já que interações mais íntimas (menor abrangência) mobilizam maior atenção;
- O parâmetro α representa o quão certos estamos de que a interação foi recebida integralmente pelo receptor. A sua escolha depende de diversos fatores da análise empregada sobre os medianeiros, pois que para cada caso poderemos (ou não) averiguar a atenção média cedida por um ator a uma interação que lhe chega qualquer. Se $\alpha = 0$ indica que caso o receptor não tenha respondido à interação de alguma forma, ele não foi influenciado por ela. $\alpha = 1$ retira qualquer valor do fato do ator ter respondido ou não determinada mensagem, igualando os dois casos. Em geral, mesmo que tenhamos o indicativo de que o receptor recebeu a interação, é interessante manter um valor de α diferente dos extremos para premiar a relação entre atores que se respondem. O parâmetro α também pode ser visto no contexto da dimensão de reciprocidade da força da conexão, já que quanto menor o α mais forte será proporcionalmente as conexões em que houveram reciprocidade.

6.3 Critérios de escolha da representação

Qual a rede ideal para a análise de influência? Não há resposta objetiva para essa pergunta. Cada rede é uma representação de um fenômeno social que vai além das ferramentas, mesmo sendo capaz de coletar e integrar todas as interações digitais, as pessoas ainda vão ser capazes de tomar café juntas e nossa visão será apenas parcial. Sendo assim, é evidente que cada representação mensurada é uma parte da informação e, portanto, capaz de descrever aspectos diferentes ou não do fenômeno. São essas variações entre as representações de redes que chamaremos de **critérios de escolha**.

Para entender sua utilidade, um exemplo: imaginemos que ao final da análise tenhamos duas ou mais representações da rede: uma textual e algumas não textuais. Para cada uma encontramos valores diferentes de proeminência, então qual usar? Colocando de outra forma, quanto de informação estarei perdendo caso considere apenas uma delas?

A seleção de característica (*feature selection*), na aprendizagem de máquina, é a tarefa

de escolher quais variáveis serão consideradas no treinamento do sistema de forma a obter a melhor aproximação do modelo (Jain and Zongker, 1997; Blum, 1997; Jain and Mao, 2000). De forma similar devemos ser capazes de selecionar redes que maximizem nossa análise de influência. Recentemente Peng *et al.* (2005) sugere que as variáveis sejam escolhidas de forma a maximizar a relevância e reduzir a redundância, mRMR (*minimal-redundancy-maximal-relevance*). No nosso caso, não conhecemos o modelo da rede *a priori* e a inferência é não-supervisionada. Por isso, não temos como avaliar a relevância de uma rede em relação a outra, todas são relevantes. Poderíamos considerar que redes que não demonstrem características de redes sociais poderiam ser vistas como fortemente enviesadas e por isso de baixa relevância, a saber: diâmetro pequeno (*small world*) (Milgram, 1967; Watts and Strogatz, 1998), poucos muito conectados e muitos pouco conectados (*power-law*) (Liljeros *et al.*, 2001; Mitzenmacher, 2004) e atores com muitas conexões tendem a estar conectados uns aos outros (*scale-free*) (Li *et al.*, 2005).

Por outro lado, a redundância da informação nos ajudará a separar as representações importantes das que não acrescentam informação substancial, ou seja, são redundantes. Sendo assim, nosso objetivo é formular alguns critérios de escolha, de modo que tenhamos em mãos o conjunto de representações minimamente redundante.

6.3.1 \otimes Mínima redundância

Dada uma entrada de dados D , composta de N amostras e M características $X = \{x_i, i = 1, \dots, M\}$, dizemos que um subconjunto $S \subseteq X$ de m características $\{x_i\}$ é minimamente redundante se atende a condição:

$$\min R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \quad (6.6)$$

$I(\cdot)$ é a função de informação mútua e que mede uma forma de dependência entre duas variáveis aleatórias. Dado duas variáveis aleatórias discretas, X e Y definimos a informação mútua de ambas como:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (6.7)$$

Um dos inconvenientes de utilizar a função de mútua informação como definida é que no caso de redes sociais, cada conexão pode ser considerada como uma variável aleatória de resultados possíveis $\{0, 1\}$ quando binária, ou um intervalo definido. Se assim for, não podemos dizer muito da probabilidade da relação existir ou não (por enquanto).

Se por outro lado considerarmos a variável como sendo a possibilidade da presença da relação (para simplificar) e cada par de atores uma amostra dessa variável, então estaremos encontrando dependência ou correlação sobre a densidade da rede apenas, deixando de lado importantes padrões estruturais como a transitividade e a reciprocidade. Por esta razão, precisamos adaptar nossa função de mútua informação para considerar as peculiaridades das redes sociais.

6.3.2 Redundância das conexões

Uma conexão é redundante quando pode ser encontrada em mais de uma rede. Mais do que isso, duas redes são redundantes quando além de compartilhar conexões também compartilham determinados padrões como triângulos e subgrupos. As duas famílias de ferramentas para acessar essa similaridade estrutural mais desenvolvidas na literatura são: gráficos aleatórios exponenciais e procedimento de atribuição quadrática, respectivamente p^* (*exponential random graphs*) e QAP (*quadratic assignment procedure*).

6.3.3 \otimes Redes discretas e/ou esparsas

O primeiro é mais apropriado para redes binárias ou com valores discretos (Dekker *et al.*, 2007). Consiste em criar um modelo exponencial para a criação de grafos aleatórios a partir de um conjunto de parâmetros relacionados a **configurações** de interesse (ROBINS *et al.*, 2007a). Uma configuração pode ser desde a presença de uma conexão, até a quantidade de k -triângulos e outros padrões mais complexos. Cada configuração tem um parâmetro relacionado que pode ser negativo, indicando que a rede é tem tendência inversa à presença daquela configuração, nula representando a indiferença e positiva para uma tendência de mesmo modo. Assim sendo, cada parâmetro pode ser visto como tendências da rede em relação a, por exemplo: densidade, reciprocidade e transitividade da rede quando suas configurações relativas são respectivamente a presença de conexões, a mutualidade das relações e a presença de triângulos.

Modelos exponenciais de gráficos aleatórios tem a seguinte forma geral:

$$\Pr(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{k}\right) \exp \left\{ \sum_A \eta_A g_A(\mathbf{y}) \right\} \quad (6.8)$$

Onde (i) o somatório é sobre todas as configurações procuradas em \mathbf{y} ; (ii) η_A é o parâmetro relacionado à configuração A ; (iii) $g_A(\mathbf{y}) = 1$ se a configuração é observada em \mathbf{y} , ou 0 de outra forma; (iv) k é uma constante de normalização que garante que

a Definição 6.8 seja uma distribuição de probabilidades. O vetor de parâmetros η é estimado para o grafo a ser modelado, procurando maximizar a sua probabilidade, iterativamente a partir de simulações com o método Monte Carlo (*Markov chain Monte Carlo maximum likelihood estimation*) (ROBINS *et al.*, 2007b; Snijders *et al.*, 2006).

Voltando ao problema de minimizar a redundância, podemos considerar o vetor η no cálculo de “informação mútua” entre as redes, no sentido de que redes que possuem a mesma informação compartilham conexões e tendências similares ao aparecimento de padrões estruturais. Derivamos I_D para redes discretas:

$$I_D(x, y) = \text{sim}(x, y) \rho(\eta_x, \eta_y) \quad (6.9)$$

Onde (i) $\text{sim}(\cdot)$ é a similaridade de Jaccard (Jaccard 1912; apud Berger-Wolf and Saia 2006), utilizada para comparar redes sociais e definida como sendo $\frac{2|x \cap y|}{|x| + |y|}$; (ii) ρ é a correlação de Pearson para os vetores de parâmetros η estimados para x e os estimados para y . Substituindo a equação (6.9) na equação (6.6) temos um modelo para o conjunto de redes discretas com mínima redundância.

6.3.4 Redes contínuas densas

Para rede contínuas, um segundo método pode ser utilizado para calcular a correlação diretamente. O procedimento de atribuição quadrática (*QAP*) recomenda um modelo linear para a correlação das redes, assim, temos que:

$$Y = \alpha X + \varepsilon \quad (6.10)$$

A probabilidade de que a correlação encontrada não seja apenas coincidência é acesada através da permuta das colunas da matriz seguindo algoritmo apropriado (Anderson and Robinson, 2001; Dekker *et al.*, 2007). Não é nosso objetivo nos aprofundar na especificidades do teste, apenas é pertinente considerarmos a utilização da correlação linear α como valor para a informação mútua na equação (6.6) para redes de valores contínuos densos.

Análise da influência

Após as representações terem sido mensuradas, chegamos no objetivo de todo o processo: a análise da influência. Como já foi dito em seções anteriores, o entendimento da dinâmica da influência na rede pode facilitar a criação de campanhas publicitárias virais baseadas no boca-a-boca; a difusão de novas informações para a comunidade, por exemplo, comunicação interna corporativa; reconhecimento de especialistas e atores-chaves na organização; otimizar buscas (Kirchhoff *et al.*, 2009). Abordagens recentes focam em modelos probabilísticos e comportamentais para estimar a dinâmica da influência. Métodos mais tradicionais de redes sociais voltaram-se para métricas individuais de proeminência para reconhecer atores-chave. Em Kempe *et al.* (2003) encontramos uma comparação mostrando que apesar de encontrar um subconjunto ótimo de n atores tal que maximize a dispersão da influência seja um problema *NP*, algoritmos que aproximam a solução para modelos dinâmicos superam soluções baseadas na simples escolha de atores centrais em 18% (para a métrica de grau) e 40% (para a métrica de proximidade). Essas, porém, são métricas tidas na literatura como sendo de pouca expressividade, o que deixa aberta ainda a possibilidade real da utilização de métricas de proeminência na análise de influência de redes sociais digitais.

7.1 Sobre Proeminência

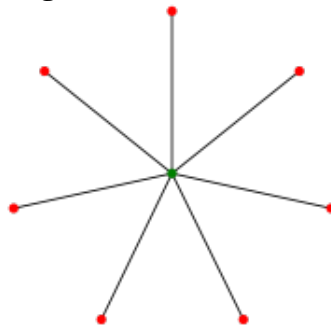
Proeminência é a característica dos relacionamentos de um ator que o põe em evidência para outros atores na rede (Wasserman and Faust, 1994). A proeminência, academicamente, foi separada em dois aspectos: centralidade e prestígio. Centralidade é como o ator se posiciona na rede, sua interação e envolvimento com os outros. Prestígio é como os outros atores se posicionam em relação a ele, como interagem e envolvem-se com ele (Knoke and Burt, 1983).

Nosso objetivo, identificar atores chaves para processos de influência da rede, está relacionado diretamente com o conceito de proeminência de forma que podemos dizer que a usaremos como uma aproximação da influência do ator na transmissão de conhecimento, inovações e opiniões pela rede. Por estes motivos, cuidado especial deve ser tomado sobre a forma como calcularemos a proeminência e que características são esperadas da representação da rede mensurada.

Durante décadas pesquisadores de redes sociais se debruçaram sobre o estudo da proeminência, desenvolvendo inúmeras métricas que conjuntamente são chamadas de métricas de centralidade. Quando calculadas considerando os caminhos que **saem** do ator, referem-se a sua centralidade propriamente dita. Quando considera-se os caminhos que **chegam** no ator, referem-se ao seu prestígio. Não obstante tenha o comum senso de que essas métricas referem-se de alguma forma à proeminência do ator, a forma específica tem ficado a cargo de cada pesquisador que já atribuiu ao seu resultado interpretações como autonomia, controle, risco, influência, corretagem (*brokerage*), independência, etc.

Freeman (1979) revisou as métricas existentes à época e as compilou em três principais centralidades: por grau, por proximidade (*closeness*) e por intermediação (*betweenness*). Interpretando seus resultados em relação ao quanto o grafo se aproxima de uma estrutura de estrela (Figura 7.1), sua visão de centralização ideal, onde suas métricas alcançam o máximo.

Figura 7.1 Gráfico estrela



Recentemente, Borgatti and Everett (2006) nos ofereceram uma outra resposta para a questão: O que as métricas de centralidade medem? E chegaram a conclusão que todas elas investigam o relacionamento do nó com os possíveis caminhos do grafo. Em sua tipologia para a classificação das métricas de centralidade, propõem quatro dimensões: tipo do caminho, propriedade do caminho, tipo do envolvimento e forma de agregação. Em outras palavras, toda métrica de centralidade é uma forma de agregar uma propriedade dos caminhos de determinado tipo em que o nó tem um determinado envolvimento. Das

quatro dimensões, duas são de maior importância: propriedade do caminho e tipo de envolvimento.

7.1.1 Tipo de envolvimento

O tipo de envolvimento é a dimensão de classificação proposta por Borgatti que se traduz na maior variância nos resultados. Pode ser de dois tipos: radial ou medial.

Métricas radiais são as que analisam os caminhos em que o nó está numa das pontas, ou seja, que saem ou chegam nele. Exemplos de métricas radiais são as centralidades de grau e proximidade de Freeman. A análise de Borgatti et al veio confirmar o que as evidências empíricas já sugeriam, que as métricas radiais representam a parcela de responsabilidade que o nó tem na coesão total da rede e que, por este motivo, se enfraquecem a medida que a rede não se adequa ao padrão de núcleo-periferia (Nakao, 1990).

Já as ditas mediais analisam os caminhos que passam pelo nó, classificando-o em seu papel de *gateway* entre partes da rede. O *betweenness* de Freeman é um exemplo de métrica medial. São mais estáveis em relação à característica de núcleo-periferia, por outro lado tendem a dar maior importância a atores que conectam grupos distintos, mas que se encontram na periferia dos mesmos.

Os dois tipos são complementares e mais de uma métrica devem ser combinadas (Stephenson and Zelen 1989; apud Wasserman and Faust 1994). Pelo fato de que usaremos métricas radiais de centralidade para estimar a influência do ator na rede, daí concluímos que:

Quanto mais próximo a representação estiver do modelo de núcleo-periferia, melhor será a aplicabilidade das métricas radiais de centralidade.

7.1.2 Propriedade do caminho

Borgatti define dois tipos de propriedade: distância e volume. A propriedade do tipo distância mais comum é a geodésica que é a quantidade de arcos presentes no menor caminho entre um ator e outro. Volume refere-se à quantidade de caminhos entre os nós. As centralidades por grau e *betweenness* são exemplos de métricas que usam o volume dos caminhos, enquanto que *closeness* usa distância.

O critério de escolha depende da aplicação. Para o nosso caso, queremos modelar o efeito da influência entre os pares na transmissão de informação. Nesse sentido, quanto mais atores próximos estiverem tentando convencer alguém de algo, maior será sua

influência sobre este (Watts and Dodds, 2007). Daí temos, naturalmente, a preferência sobre métricas relacionadas a volume, no lugar de distância, já que representa diretamente a quantidade de caminhos que existem entre os nós.

7.1.3 Redundância na proeminência

Vimos no Seção 6.3 como calcular a redundância das conexões entre representações disponíveis, porém também é interessante observar a relação que há entre os resultados da análise de cada rede. Para esse fim utilizaremos também a correlação linear de Pearson entre os vetores de proeminência calculado de forma que duas redes podem estar positivamente relacionadas, aparentemente independentes ou negativamente relacionadas. Um critério baseado na proeminência pode correlacionar redes estruturalmente bastante diferentes, mas que resultam no sobressaimento dos mesmos atores em termos de importância. Porém, se o objetivo for minizar a redundância do ponto de vista da ordenação final dos atores por grau de proeminência para a utilização em outros algoritmos como por exemplo, escolha de subconjunto ótimo para marketing viral, então talvez este critério seja o mais apropriado.

7.1.4 Relevância na proeminência

Quando tudo o que temos é a rede, para medirmos proeminência dos atores precisamos ser capazes também de medir sua aplicabilidade. Vimos que para alguns tipos de métricas a rede precisa apresetar determinado padrão de aglomeração. Por esta razão, podemos saber *a priori* quando uma representação vai produzir métricas significativas de proeminência. É sabido que redes sociais são formadas por diversas comunidades sobrepostas (Palla *et al.*, 2005) e por isso uma representação com alto grau de núcleo/periferia nessas condições é pouco provável. Uma abordagem seria quebrar a rede em suas comunidades para a partir daí encontrar os atores centrais de cada uma, existem diversas ferramentas de decomposição da rede em *k-cliques*, *k-plexes*, *k-cores* para grafos valorados e direcionados (Peay 1975; Doreian 1969; Freeman 1992; apud Wasserman and Faust 1994), mas todas elas tendem a ignorar atores pendentes que compõem grande parte da periferia.

Um modelo mais interessante para o nosso problema é a chamada comunidade de prática. A teoria das comunidades de prática não nasceu no meio sociométrico e é voltada ao estudo da gestão do conhecimento, porém o trabalho de Schenkel (2002) revisa as características estruturais típicas das comunidades de prática e sugere quatro delas muito similares ao que estamos procurando:

Conectada A rede possui apenas um componente;

Densa A média das conexões por ator é alta; o que é considerado alto varia a depender do tamanho da rede e, por tanto, da periferia;

Compacto A distância média entre os atores é menor do que o comum para *small-worlds*, cujo crescimento é logarítmico em relação ao crescimento da rede;

Núcleo/Periferia A rede exhibe um padrão de núcleo periferia, isto é, alto grau de *scale-freeness*.

Resta-nos encontrar maneiras de reconhecer essas comunidades e, mais importante, quais atores a compõem. Comunidades de prática foram definidas como: um grupo de pessoas informalmente e contextualmente conectadas a uma situação de trabalho em que estão empregando uma competência comum na perseguição de um objetivo comum (Wenger, 1999). Essas situações de trabalho, no entanto, podem ser generalizadas para situações que envolvem perícia, conhecimento técnico e por isso encontraremos comunidades de práticas não apenas no contexto organizacional, mas entre hobbistas também. O conhecimento em tais comunidades é compartilhado através de narrativas, conversação, *mentoring* e aprendizagem por experimentos (Brown and Duguid, 1991; Lave and Wenger, 1991), tais atividades podem ser transferidas para o meio digital (Hildreth *et al.*, 1998; Kimble *et al.*, 2001) e deixam rastros que podem ser mensurados para reconstruir a rede (Tyler *et al.*, 2005; Welser *et al.*, 2007).

É de se esperar, por tanto, que os membros de comunidades de prática se reúnam em torno de focos de interação onde possam compartilhar histórias, como por exemplo comunidades virtuais. Para diminuir a ambiguidade do termo, utilizaremos a palavra *coletividade* quando nos referirmos à comunidade virtual na qual os atores se afiliam digitalmente. Podemos construir uma rede de coletividades, onde um nó está ligado a outro se possuem membros em comum, essa rede é menos suscetível à existência de pendentes e por isso podemos aplicar algoritmos como os descritos em Palla *et al.* (2005) para reconhecer coletividades que se aglomeram em grupos. Essas coletividades poderiam ser usadas depois para filtrar a rede de atores e verificar se atendem às características de comunidades de prática, aumentando a relevância da representação para a análise da influência.

Outra possibilidade é tentar identificar o objetivo ou competência comum que reúne a comunidade através de mineração de texto. Possibilidades vão desde encontrar a distância relativa entre os atores baseados em quais palavras aparecem em seus textos (Reichling

et al., 2005), até criar uma ontologia dos termos propriamente dito e a partir daí estimar uma conexão entre os atores (Mori *et al.*, 2006). Spertus *et al.* (2005) propõem o uso da métrica de similaridade para termos extraídos do texto TF-IDF (Salton, 1989; Frakes and Baeza-Yates, 1992) e em MATSUO *et al.* (2007) encontramos uma versão aperfeiçoada que diminui a necessidade de um *corpus* da linguagem. A similaridade, dessa forma, pode indicar uma homofilia de interesses quando cruzada com as representações mensuradas das interações, essa homofilia poderia ser usada então para filtrar a rede de atores e, em verificando as características de comunidades de prática, também vir a aumentar a relevância da representação na análise da influência.

8

Conclusão e trabalhos futuros

Nas etapas finais do processo de descoberta do conhecimento, a avaliação e a utilização, cada domínio fará a aplicação apropriada. Alguns critérios de relevância para o caso do cálculo de proeminência foram discutidos na Seção 7.1.4. Nosso intuito com esse trabalho foi reunir método e técnicas para um processo pragmático de análise de influência de redes sociais digitais, oferecendo um passo a passo compreensivo do que pode ser feito, das dificuldades possíveis e das soluções existentes. Nesse sentido, sentimos que esse *framework* para a análise de influência em redes está ainda em sua infância e há muito o que se fazer. Em tópicos gerais, reunimos abaixo os principais pontos de extensão deste trabalho:

- Considerar a dimensão tempo na dinâmica da rede. Primeiro na evolução das conexões, seja através de séries temporais (Snijders, 1996); como intervalos de tempo (Butts and Pixley, 2004); ou como eventos pontuais no tempo (Butts, 2008). Depois na evolução da afiliação dos atores a grupos através do tempo (Berger-Wolf and Saia, 2006);
- Fazer melhor proveito das informações baseadas em afiliações, extendendo modelos p^* para redes dois-modos (*two-mode network*) (Field *et al.*, 2006). Em redes de dois-modos há mais de um tipo de nó, no nosso caso os atores seriam um tipo, as interações outro tipo. Essa visão hiperrelacional da interação permitiria diferenciar padrões de influência do tipo “um para muitos” dos do tipo “muitos de um pra um”;
- Integrar com modelos estatísticos da economia da atenção para tratar fenômenos como a competição pela atenção, sobrecarga de informação e outras propriedades de mercado (Falkinger, 2007);
- Aplicar o processo de mensuração descrito nesse trabalho a diferentes domínios, de

forma a melhor avaliar as propriedades de seus parâmetros, sua utilidade na análise de influência e facilidade de aplicação.

Não obstante, acreditamos que a singular reunião da literatura espalhada sobre o assunto em um *framework* pragmático facilitará a construção de ferramentas de mineração apropriadas para o problema. Esse, será o tópico da segunda parte deste trabalho.

Referências Bibliográficas

- ADAMIC, L. and ADAR, E. (2005). How to search a social network. *Social Networks*, **27**(3), 187–203.
- Adamic, L. A., Buyukkokten, O., and Adar, E. (2003). A social network caught in the Web.
- Anderson, M. J. and Robinson, J. (2001). Permutation Tests for Linear Models. *Australian New Zealand Journal of Statistics*, **43**(1), 75–88.
- Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. *International Conference on Knowledge Discovery and Data Mining*, page 44.
- Barros, F., Silva, E., Rabelo, J., and F (2002). Similar Documents Retrieval to Help Browsing and Editing in Digital Repositories. In *Communications, Internet and Information Technology*, page 376.
- Benassi, M., Greve, A., and Harkola, J. (1999). Looking for a network organization: The case of GESTO. *Journal of Market-Focused Management*, **4**(3), 205–229.
- Berger-Wolf, T. Y. and Saia, J. (2006). A framework for analysis of dynamic social networks. In *International Conference on Knowledge Discovery and Data Mining*.
- Blum, A. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**(1-2), 245–271.
- Blumstein, P. and Kollock, P. (1988). Personal relationships. *Annual Review of Sociology*, **14**, 467–490.
- Bonacich, P. (1987). Power and Centrality: A Family of Measures. *The American Journal of Sociology*, **92**(5), 1170 – 1182.
- Borgatti, S. and Everett, M. (2006). A graph-theoretic perspective on centrality. *Social Networks*, **28**(4), 466–484.
- Boyd, D. and Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *JOURNAL OF COMPUTER MEDIATED COMMUNICATION-ELECTRONIC EDITION-*, **13**(1), 210.
- Breiger, R. L. (1974). The Duality of Persons and Groups. *Social forces*, **53**(2), 181–190.

- Brelger, R., Carley, K., and Pattison, P. (2004). Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers.
- Brown, J., Broderick, A. J., and Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, **21**(3), 2–20.
- Brown, J. S. and Duguid, P. (1991). Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation. *Organization Science*, **2**(1), 40 – 57.
- Burnett, G. (2000). Information exchange in virtual communities: a typology.
- Burnett, G. and Buerkle, H. (2004). Information exchange in virtual communities: A comparative study. *Journal of Computer-Mediated*.
- Butts, C. T. (2008). A RELATIONAL EVENT FRAMEWORK FOR SOCIAL ACTION. *Sociological Methodology*, **38**(1), ???–???
- Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science (New York, N.Y.)*, **325**(5939), 414–6.
- Butts, C. T. and Pixley, J. E. (2004). A STRUCTURAL APPROACH TO THE REPRESENTATION OF LIFE HISTORY DATA. *Journal of Mathematical Sociology*, **28**(2), 81–124.
- Cios, K. J. and Kurgan, L. A. (2005). *Advanced Techniques in Knowledge Discovery and Data Mining*. Advanced Information and Knowledge Processing. Springer London, London.
- Clemons, E. K., Barnett, S., and Appadurai, A. (2007). The future of advertising and the value of social network websites: some preliminary examinations. *ACM International Conference Proceeding Series; Vol. 258*.
- Coleman, J. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, **94**, S95–S120.
- Coleman, J., Katz, E., and Menzel, H. (1966). *Medical innovation: A diffusion study*. Bobbs-Merrill, Indianapolis.

- Costa, R. A., Oliveira, R. Y. S. D., Silva, E. M. D., and Meira, S. R. L. (2008). A.M.I.G.O.S: Uma plataforma para Gestão de Conhecimento através de Redes Sociais. In *Simpósio Brasileiro de Sistemas Colaborativos*, pages 192–203. IEEE.
- Danet, B., Ruedenberg, L., and Rosenbaum-Tamari, Y. (1998). Hmmm. . . where’s that smoke coming from?:writing, play and performance on Internet relay chat. *Network and Netplay: virtual groups on the Internet*, page 41.
- Davenport, T. H. and Beck, J. C. (2001). *The Attention Economy: Understanding the New Currency of Business*. Harvard Business School Press, Boston.
- Dekker, D., Krackhardt, D., and Snijders, T. (2007). Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika*, **72**(4), 563–581.
- Dindia, K. and Canary, D. J. (1993). Definitions and Theoretical Perspectives on Maintaining Relationships. *Journal of Social and Personal Relationships*, **10**(2), 163–173.
- Ding, X. and Liu, B. (2007). The utility of linguistic rules in opinion mining. *Annual ACM Conference on Research and Development in Information Retrieval*, page 811.
- Domingos, P. and Richardson, M. (2001). Mining the network value of customers. *International Conference on Knowledge Discovery and Data Mining*.
- Dooley, R. and Levinsohn, S. (2001). *Analyzing discourse: A manual of basic concepts*. SIL International, Dallas.
- Doreian, P. (1969). A Note on the Detection of Cliques in Valued Graphs. *Sociometry*, **32**(2), 237 – 242.
- Ehrlich, K., Lin, C.-Y., and Griffiths-Fisher, V. (2007). Searching for experts in the enterprise: combining text and social network analysis. *Conference on Supporting Group Work*, pages 117–126.
- Falkinger, J. (2007). Attention economies. *Journal of Economic Theory*, **133**(1), 266–294.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**(11), 27.
- Field, S., Frank, K. A., Schiller, K., Riegle-Crumb, C., and Muller, C. (2006). Identifying positions from affiliation networks: Preserving the duality of people and events. *Social networks*, **28**(2), 97–123.
-

- Frakes, W. and Baeza-Yates, R. (1992). *Information retrieval: Data structures & algorithms*. Prentice Hall.
- Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social Networks*, **1**(3), 215–239.
- Freeman, L. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, **13**(2), 141–154.
- Freeman, L. C. (1992). The Sociological Concept of "Group": An Empirical Test of Two Models. *The American Journal of Sociology*, **98**(1), 152 – 166.
- Friedkin, N. (1980). A test of structural features of granovetter's strength of weak ties theory. *Social Networks*, **2**(4), 411–422.
- Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. *Conference on Human Factors in Computing Systems*, pages 211–220.
- Goldhaber, M. (1997). The attention economy and the net. *First Monday*, **2**(4).
- Goldhaber, M. (2006). The value of openness in an attention economy. *First Monday*.
- Granovetter, M. (1973). The strength of weak ties. *ajs*, **78**(6), 1360.
- Granovetter, M. (1995). *Getting a job: A study of contacts and careers*. University of Chicago Press.
- Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. *International World Wide Web Conference*, page 491.
- Gyarmati, D. and KYTE, D. (2004). Social capital, network formation and the Community Employment Innovation Project. *Policy Research Initiative*, **6**(3).
- Han, J. and Kamber, M. (2006). *Data mining: concepts and techniques*.
- Haythornthwaite, C. (2005). Social networks and Internet connectivity effects. *Information, Communication & Society*, **8**(2), 125–147.
- Herring, S. (1999). Interactional coherence in CMC. *Journal of Computer-Mediated Communication*, **4**(4).
- Herring, S. (2001). *Computer-mediated discourse*, pages 612–634.
-

- Herring, S. (2002). Computer-mediated communication on the Internet. *Annual review of information science*.
- Herring, S. (2007). A Faceted Classification Scheme for Computer Mediated Discourse. *Language@Internet*.
- Hildreth, P. M., Kimble, C., and Wright, P. (1998). Computer Mediated Communications and Communities of Practice.
- Humphreys, A. and Kozinets, R. (2009). The Construction of Value in Attention Economies. *Advances in Consumer Research*, **36**.
- Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist*, **11**(2), 37–50.
- Jain, A. and Mao, R. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine*, **22**(1), 4–37.
- Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine*, **19**(2), 153–158.
- Jensen, D. and Neville, J. (2003). Data mining in social networks. *Social Network Modeling and Analysis*.
- Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. *International Conference on Knowledge Discovery and Data Mining*.
- Kim, H., Kim, G. J., Park, H. W., and Rice, R. E. (2007). Configurations of Relationships in Different Media: FtF, Email, Instant Messenger, Mobile Phone, and SMS. *Journal of Computer-Mediated Communication*, **12**(4), 1183–1207.
- Kimble, C., Hildreth, P., and Wright, P. (2001). *Communities of practice: going virtual*, pages 220–234.
- Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T., and Fleck, M. (2009). Using social network analysis to enhance information retrieval systems. *Applications of Social Network Analysis (ASNA)*, Zurich, at <http://www.alexandria.unisg.ch/publications/46444>, accessed, **7**, 1–21.
-

- Knoke, D. and Burt, R. (1983). Prominence. *Applied network analysis*.
- Ko, K. (1996). Structural Characteristics of Computer-Mediated Language: A Comparative Analysis of InterChange Discourse. *Electronic Journal of Communication/La revue*.
- Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. *International Conference on Knowledge Discovery and Data Mining*, page 611.
- KURGAN, L. A. and MUSILEK, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, **21**(01), 1–24.
- Lave, J. (1991). Situating learning in communities of practice. *Perspectives on socially shared cognition*.
- Lave, J. and Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press, Cambridge.
- Li, L., Alderson, D., Doyle, J., and Willinger, W. (2005). Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, **2**(4), 431–523.
- Liljeros, F., Edling, C. R., Amaral, L. A., Stanley, H. E., and Aberg, Y. (2001). The web of human sexual contacts. *Nature*, **411**(6840), 907–8.
- Ma, H., Yang, H., Lyu, M. R., and King, I. (2008). Mining social networks using heat diffusion processes for marketing candidates selection. *Conference on Information and Knowledge Management*.
- Marsden, P. and Campbell, K. (1984). Measuring tie strength. *Social Forces*, **63**, 482–501.
- MARTÍNEZ, S. L. and Figueroa, M. V. (2000). Internet como medio y objeto de estudio en antropología. *Antropología e Internet*.
- MATHEWS, K. M., WHITE, M. C., LONG, R. G., SOPER, B., and VON BERGEN, C. W. (1998). ASSOCIATION OF INDICATORS AND PREDICTORS OF TIE STRENGTH. *Psychological reports*, **83**(3), 1459–1469.
- MATSUO, Y., MORI, J., HAMASAKI, M., NISHIMURA, T., TAKEDA, H., HASIDA, K., and ISHIZUKA, M. (2007). POLYPHONET: An advanced social network extraction system from the Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, **5**(4), 262–278.

- Milgram, S. (1967). The small world problem. *Psychology today*, **1**, 61–67.
- Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Internet Measurement Conference*.
- Mitchell, J. (1987). The components of strong ties among homeless women. *Social Networks*, **9**(1), 37–47.
- Mitzenmacher, M. (2004). A brief history of lognormal and power law distributions. *Internet Mathematics*.
- Mori, J., Tsujishita, T., Matsuo, Y., and Ishizuka, M. (2006). *Extracting Relations in Social Networks from the Web Using Similarity Between Collective Contexts*, volume 4273 of *Lecture Notes in Computer Science*, pages 487–500. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Nakao, K. (1990). Distributions of measures of centrality: enumerated distributions of freeman’s graph centrality measures. *Connections*, **13**(3), 10–22.
- Neville, J., Şimşek, O., Jensen, D., and Komoroske, J. (2005). Using relational knowledge discovery to prevent securities fraud. *Knowledge discovery*.
- Newman, M., Barabasi, A., and Watts, D. (2006). *The structure and dynamics of networks*. Princeton University Press.
- Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043), 814–8.
- Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, **2**(1), 1–135.
- Peay, E. R. (1975). Grouping by cliques for directed relationships. *Psychometrika*, **40**(4), 573–574.
- Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, **27**(8), 1226–38.

- Perlman, D. and Fehr, B. (1987). The development of intimate relationships. *Intimate relationships: Development, dynamics, and deterioration*, pages 13–42.
- Petróczi, A., Nepusz, T., and Bazsó, F. (2006). Measuring tie-strength in virtual social networks. *Connections*, **27**(2), 39–52.
- Recuero, R. (2008). PRÁTICAS DE SOCIABILIDADE EM SITES DE REDES SOCIAIS. *pontomidia.com.br*, pages 1–16.
- Reichling, T., Schubert, K., and Wulf, V. (2005). Matching human actors based on their texts. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work - GROUP '05*, page 61, New York, New York, USA. ACM Press.
- Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. *International Conference on Knowledge Discovery and Data Mining*.
- Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research*, **3**(2), 131–141.
- ROBINS, G., PATTISON, P., KALISH, Y., and LUSHER, D. (2007a). An introduction to exponential random graph (p) models for social networks. *Social Networks*, **29**(2), 173–191.
- ROBINS, G., SNIJDERS, T., WANG, P., HANDCOCK, M., and PATTISON, P. (2007b). Recent developments in exponential random graph (p) models for social networks. *Social Networks*, **29**(2), 192–215.
- Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science (New York, N.Y.)*, **311**(5762), 854–6.
- Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison Wesley, Reading, MA.
- Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, **7**(2).
- Sassen, S. (2002). Towards a Sociology of Information Technology. *Current Sociology*, **50**(3), 365–388.
-

- Schenkel, A. J. (2002). COMMUNITIES OF PRACTICE OR COMMUNITIES OF DISCIPLINE MANAGING DEVIATIONS AT THE fjRESUND BRIDGE. *Managing*, (december).
- Seidman, S. and Foster, B. (1978). A graph-theoretic generalization of the clique concept. *Journal of Mathematical Sociology*, **6**, 139–254.
- Simon, H. (1996). Designing organizations for an information-rich world. *LIBRARY OF CRITICAL WRITINGS IN ECONOMICS*.
- Skå geby, J. (2009). Exploring Qualitative Sharing Practices of Social Metadata: Expanding the Attention Economy. *The Information Society*, **25**(1), 60–72.
- Snijders, T. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, **21**, 149–172.
- Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). NEW SPECIFICATIONS FOR EXPONENTIAL RANDOM GRAPH MODELS. *Sociological Methodology*, **36**(1), 99–153.
- Song, X., Chi, Y., Hino, K., and Tseng, B. (2007). Identifying opinion leaders in the blogosphere. *Conference on Information and Knowledge Management*, pages 971–974.
- Spertus, E., Sahami, M., and Buyukkokten, O. (2005). Evaluating similarity measures: a large-scale study in the orkut social network. *International Conference on Knowledge Discovery and Data Mining*, page 678.
- Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, **11**(1), 1–37.
- Tyler, J., Wilkinson, D., and Huberman, B. (2005). E-Mail as Spectroscopy: Automated Discovery of Community Structure within Organizations. *The Information Society*, **21**(2), 143–153.
- Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge Univ Pr.
- Watts, D. J. and Dodds, P. S. (2007). Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, **34**(4), 441–458.
-

- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**(6684), 440–2.
- Wellman, B. (1982). *Studying personal communities*, pages 61–80. Sage, Beverly Hills, CA.
- Wellman, B. and Wortley, S. (1990). Different Strokes from Different Folks: Community Ties and Social Support. *The American Journal of Sociology*, **96**(3), 558 – 588.
- Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., and Haythornthwaite, C. (1996). Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community. *Annual Review of Sociology*, **22**(1), 213–238.
- Welser, H., Gleave, E., Fisher, D., and Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, **8**(2), 1–32.
- Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.
- WERRY, C. C. (1996). Linguistic and interactional features of Internet relay chat. *Pragmatics & beyond. New series*, **39**, 47–63.
- Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Human Language Technology Conference*, page 347.
- Wu, F. and Huberman, B. (2009). Persistence and Success in the Attention Economy. *Arxiv preprint arXiv:0904.0489*.
- Xiang, R., Neville, J., and Rogati, M. (2010). Modeling relationship strength in online social networks. *International World Wide Web Conference*, pages 981–990.

Appendices



Estudo de caso: A.M.I.G.O.S.

Nosso estudo de caso aplica-se sobre um ambiente digital de interação desenvolvido para a gestão do conhecimento em rede social. Durante o estudo de caso aplicamos a metodologia proposta no Capítulo 3 e seu passo a passo será descrito nas seções A.1, A.2, A.3 e A.4. Nos resultados avaliaremos a utilidade do método e das ferramentas desenvolvidas no processo.

A.1 Entendendo o domínio

Nosso objetivo neste processo de mineração é analisar uma “fotografia” do ambiente digital do a.m.i.g.o.s. e extrair dela um grupo de atores-chaves; aqueles que se destacam dos demais pelo seu posicionamento ou influência sobre a rede. Para darmos prosseguimento a análise, precisamos entender o ambiente e decidir sobre a ferramenta de análise que usaremos. Sobre o ambiente temos:

(Costa *et al.*, 2008) Acrônimo de Ambiente Multimídia para Integração de Grupos e Organizações Sociais, o a.m.i.g.o.s tem por objetivo prover a infra-estrutura necessária para a criação de redes sociais virtuais para os mais diversos fins. Dentre estes, pode-se destacar o seu uso para estimular a criação e compartilhamento do conhecimento pelos seus diversos membros, podendo estes estarem relacionados a uma organização social. Nele é permitida a criação explícita das redes sociais através dos usuários e seus contatos. Cada contato é explicitamente adicionado por cada usuário, mesmo que dentro de uma mesma organização, e este relacionamento é navegável por qualquer outro membro da rede social. Nas próximas linhas são apresentadas as principais funcionalidades com suas características e possíveis usos.

Perfis Cada usuário possui um perfil no a.m.i.g.o.s. Este perfil consiste de um conjunto de dados preenchidos na forma de cadastro, que definem algumas propriedades simples do usuário, como local de residência, idiomas que possui conhecimento, endereço de e-mail, identificadores de aplicações de mensagem instantânea (Windows Live Messenger, Skype, Google Talk, dentre outros), e uma descrição de suas áreas de interesse. Porém a parte mais relevante do perfil não é preenchida pelo usuário, e sim inferida pelo sistema. *[Para cada usuário, um índice de atividade para a produção e consumo de conteúdos também é automaticamente calculado]*

Histórias Histórias são destinadas ao registro, compilação e apresentação de conhecimentos emergentes entre os participantes da rede. Construído de forma gradual, através de contribuições espontâneas ou induzidas, qualquer usuário do sistema pode inserir no ambiente suas próprias histórias de sucesso ou dilemas, à medida que as considere relevantes para o objetivo da rede social. Adicionalmente as histórias podem estar associadas a uma ou mais comunidades, o que indica que, apesar do autor ser um usuário em específico, o conhecimento construído encontra-se de alguma forma relacionado a estas comunidades.

Cada usuário do sistema poderá, adicionalmente, atuar como um revisor do conteúdo inserido por seus pares, avaliando qualitativamente as contribuições disponibilizadas neste ambiente. Esta avaliação pode ser realizada de uma das duas formas:

- Adição de comentários que contribuam para a evolução da história, criando-se assim uma história mais rica, com mais participantes e novos conhecimentos. À medida que a história for acrescida de comentários, é criado um diálogo associado ao conhecimento em construção;
- Atribuição de uma nota, variando de uma (1) a cinco (5) estrelas, às histórias que lê. Permitindo que este conhecimento, expresso através de histórias, possa ser apresentado através de um ranking que indique as mais relevantes para os membros daquela rede social.

Relacionamentos O a.m.i.g.o.s dá suporte a praticamente todos os mecanismos de relacionamentos existentes nas atuais redes sociais. Nele

cada usuário pode adicionar a sua lista de contatos qualquer outro usuário também membro da rede social. Esta lista de contatos pode ser agrupada em grupos, facilitando a organização dos contatos pelo seu usuário.

Comunidades Virtuais Comunidades podem ser vistas como agregações de pessoas com objetivos em comum. O a.m.i.g.o.s dá suporte à criação de manutenção de comunidades por parte de seus usuários, podendo estes convidarem membros de sua lista de contatos a participar das discussões ou atividades a serem realizadas no âmbito da comunidade. Cada comunidade possui uma série de mecanismos para a criação e compartilhamento do conhecimento. O principal mecanismo de criação e compartilhamento do conhecimento é o fórum de discussão, onde os membros da comunidade podem iniciar discussões sobre os mais diversos assuntos.

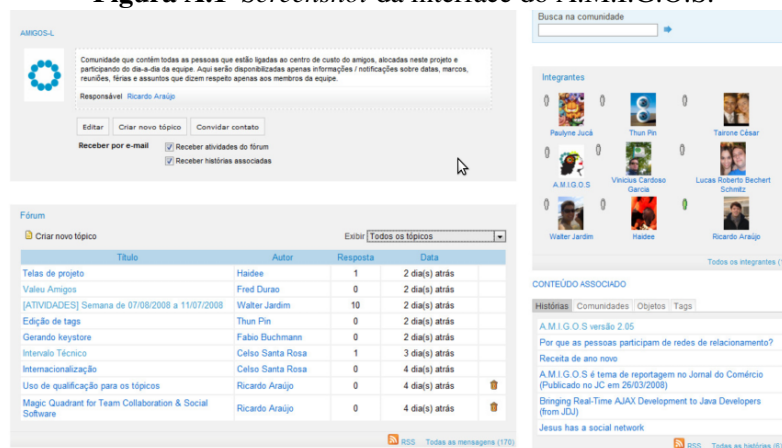
Um segundo mecanismo de compartilhamento do conhecimento é a associação de histórias à comunidade. Esta associação pode ser realizada por qualquer membro da comunidade ao criar uma história no sistema. Caso deseje-se, é possível até mesmo que a história seja visível apenas pelos membros das comunidades relacionadas.

Recomendações Como mecanismo de disseminação do conhecimento, o a.m.i.g.o.s possui suporte a recomendações. Estas recomendações são sempre direcionadas a usuários do sistema e podem ser referentes a histórias, comunidades, tópicos de um fórum ou outros usuários. Existem basicamente dois tipos de recomendação, uma feita manualmente por um usuário para seus contatos, e a outra realizada automaticamente pelo sistema para um usuário a partir da probabilidade do interesse deste no conteúdo recomendado.

Para que a recomendação automática seja possível, o sistema vai montando o perfil do usuário à medida que este utiliza o sistema, baseado no que é lido ou escrito por ele. Para isto, o sistema faz uso do *Vector Space Model* (Barros *et al.*, 2002), varrendo o conteúdo textual disponível em cada elemento, calculando então o centróide do conteúdo e conseqüentemente o centróide do usuário, este composto pela soma vetorial do centróide de cada um dos seus conteúdos. Em seguida o sistema tenta identificar outros usuários ou conteúdos com centróides

similares, recomendando-os ao usuário em questão sempre que esta similaridade for maior que um limiar configurado.

Figura A.1 Screenshot da interface do A.M.I.G.O.S.



Considerando os critérios estabelecidos no Capítulo 4, temos que a representação da rede precisa ser não-binária e envolver a maior quantidade de dimensões da força do laço possível; Por esta razão focaremos nas interações textuais do a.m.i.g.o.s. usando o conceito de atenção apresentado na Seção 5.2. Para a análise de influência, usaremos as métricas de *Flow betweenness* de Freeman e o poder de Bonacich; as razões dessa escolha são a de que atendem aos critérios de serem sensíveis tanto à estrutura macro da rede, quanto a configuração local do nó, e também por simplicidade de implementação.

A.2 Entendendo os Dados

Como para os fins desse estudo nos foi cedido o acesso a uma “fotografia” do banco de dados do ambiente, então será uma análise intrínseca; não precisaremos de *crawlers*. Devido à complexidade de implementar algoritmos de análise de sentimento, seguiremos uma abordagem ingênua e não consideraremos a intenção da interação. Como se trata de interações em um contexto organizacional fechado, é pouco provável que tenhamos disruptores e históricos longos de adversidade, de forma que podemos considerar que muitas interações entre dois membros é uma bom indicativo da positividade de suas intenções.

O poder da abordagem apresentada na Seção 5.2 é a integração de vários tipos de interação, conquanto sejam textuais. Então utilizaremos quatro medianeiros textuais

reconhecidos no ambiente e classificaremos de acordo com a tipologia apresentada na Seção 5.1:

Tópicos são mensagens que **iniciam** discussões nos fóruns das comunidades. A abrangência é individual desenvolvida, já que é direcionada aos membros daquela comunidade apenas.

Respostas compõem o resto das mensagens nas discussões dos fóruns e seguem naturalmente os tópicos e umas às outras. A abrangência também é individual desenvolvida.

Histórias que são os textos de propósito geral que podem ser visualizados por qualquer pessoa dentro do sistema, com algumas exceções. A abrangência, nesse caso, pode ser individual desenvolvida quando relacionada a uma comunidade apenas ou individual generalizada quando para todo o ambiente.

Comentários relativos às histórias publicadas. Cada história tem um espaço público onde qualquer membro pode ler e publicar sua opinião, trata-se portanto de uma interação de abrangência comum.

É importante denotar a abrangência de cada medianeiro pois que afeta diretamente o cálculo da atenção residual que cada autor tem com seus leitores. Usaremos o coeficiente β igual ao inverso multiplicativo da quantidade de membros da abrangência, assim cada autor cede parcelas iguais a sua audiência por cada interação. Para interações de abrangência generalizada para todo o ambiente, como é o caso da grande maioria das histórias cadastradas, temos que $\beta = 0$, por tanto, o autor não cede atenção residual para ninguém.

Como nossa análise é intrínseca, pela disposição da informação no banco de dados podemos saber exatamente qual membro acessou e leu quais tópicos e histórias. Por esta razão, definimos o coeficiente α da seguinte maneira: 0, caso não tenha lido; 0.2, caso tenha. Sendo assim, membros da abrangência que não leram não cedem atenção ao autor, membros que leram cedem 20% do total possível e aqueles que leram e responderam ou comentaram cedem 100% da atenção.

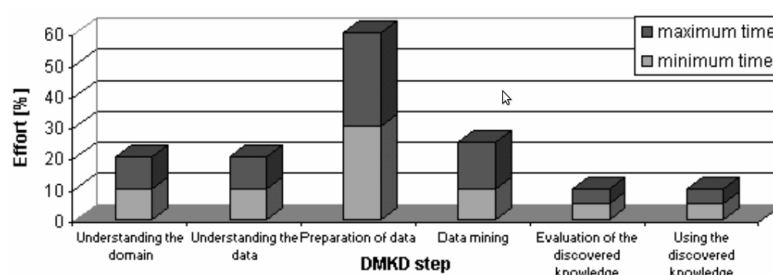
Fizemos $\gamma = 1/2$ de forma que a atenção transitiva decai para a metade a cada elo da *thread*. Infelizmente, mesmo a análise sendo intrínseca, não tivemos acesos a interações passivas do usuário como, por exemplo, o tempo que ele passou em cada leitura ou logado no sistema. Dessa forma é complicado estimar uma função de participação $E(\cdot)$,

porém nós podemos considerar o próprio índice de atividade calculado pelo sistema como aproximação desse parâmetro.

A.3 Preparando os dados

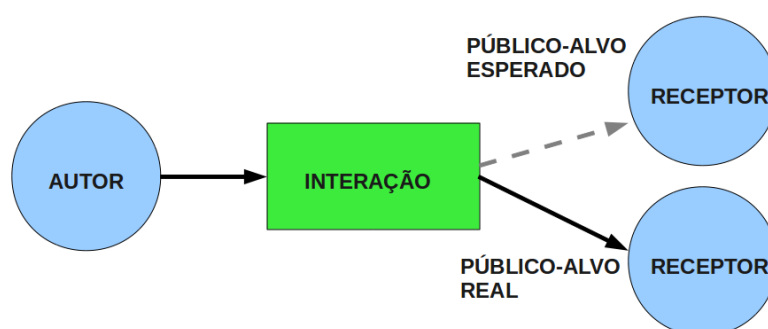
Segundo Cios and Kurgan (2005), em torno de 45% do esforço total empregado no processo de mineração de dados é usado no passo de preparação dos dados, como podemos ver na Figura A.2. Por esta razão produzimos como subproduto dessa atividade um *framework* enxuto para a extração e manipulação dos dados. O banco original é acessível via *SQL*, através do qual extraímos as interações e armazenamos em outro banco com controle de data de quando cada interação foi capturada; em cima desse banco criamos uma estrutura que abstrai a camada de dados da aplicação, permitindo ao pesquisador utilizar se concentrar na exploração do grafo.

Figura A.2 Esforço relativo para cada passo do KDDM (Cios and Kurgan, 2005)



O *framework* foi escrito em *Python* e utiliza uma *API* de *ORM* para esconder o serviço de banco de dados utilizado. Através dessa modelagem o pesquisador tem simples acesso ao conjunto de membros da rede, ao conjunto de suas interações (e propriedades) e a abrangência de cada. Nesse quesito, precisamos separar o público-alvo esperado e o real (ver Figura A.3), com isso, a atenção reflexiva é calculada em cima do público-alvo esperado, enquanto que só o público-alvo direto cede atenção direta para o autor. Com isso é possível que autores cedam mais do que recebem, à medida que produzem para uma comunidade onde poucos ou nenhum de seus membros consomem este conteúdo.

Implementamos o algoritmo de cálculo de atenção como apresentado na Seção 6.2.1, e obtivemos o resultado separado para os dois grupos de interações textuais: tópicos e resposta; histórias e comentários. O resultado que encontramos é apresentado na Figura A.4, onde a distribuição da carga de atenção parece seguir uma lei de potência. Essa característica é interessante porque mimetiza a propriedade *scale-free* das redes sociais.

Figura A.3 Exemplo de modelagem no *framework*

As Figuras A.5, A.6, A.7 e A.8 apresentam visualizações das quatro redes concebidas: uma não textual, que é a de contatos; três textuais, ou seja, utilizando nosso algoritmo de atenção, sobre apenas tópicos e suas repostas, apenas histórias e seus comentários e uma completa considerando tópicos e histórias. Na Tabela A.1 apresentamos algumas estatísticas sobre as quatro redes, indicando grau médio de reciprocidade em todas as representações. Alto grau de reciprocidade era esperado para as redes de atenção devido à atenção residual, assim toda interação gera uma conexão da audiência para o autor e sua recíproca do autor para a audiência, mesmo que em menor intensidade. O resultado encontrado, principalmente na rede de tópicos, indica que apesar do autor da mensagem visar obter a atenção de toda uma comunidade, poucos foram os que efetivamente se prestaram a isso.

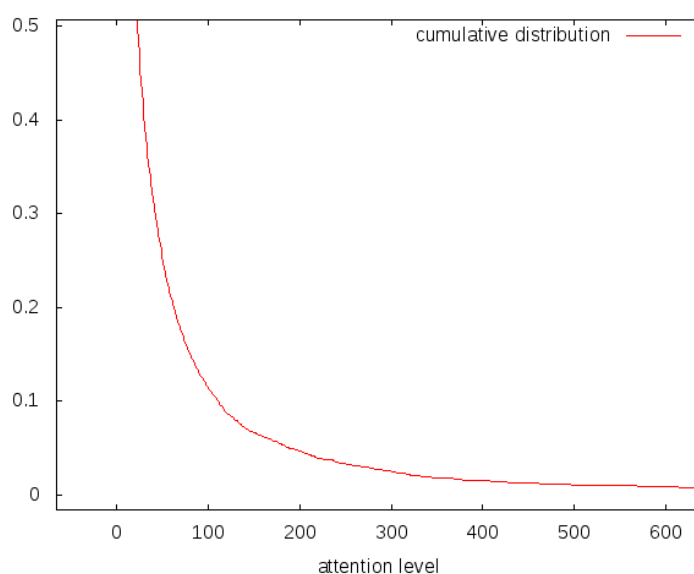
Figura A.4 Distribuição da Probabilidade Cumulativa da Atenção

Figura A.5 Visualização da rede de contatos do a.m.i.g.o.s.

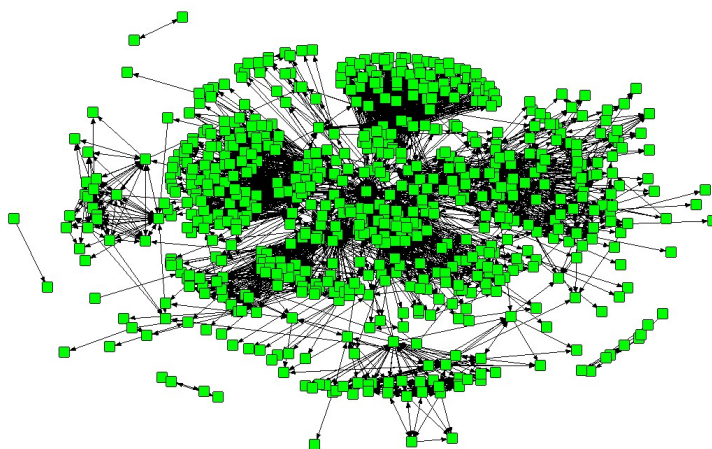


Figura A.6 Visualização da rede de tópicos do a.m.i.g.o.s.

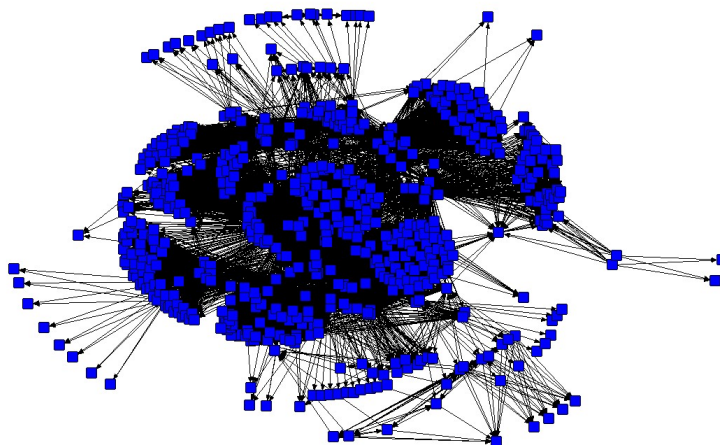


Figura A.7 Visualização da rede de histórias do a.m.i.g.o.s.

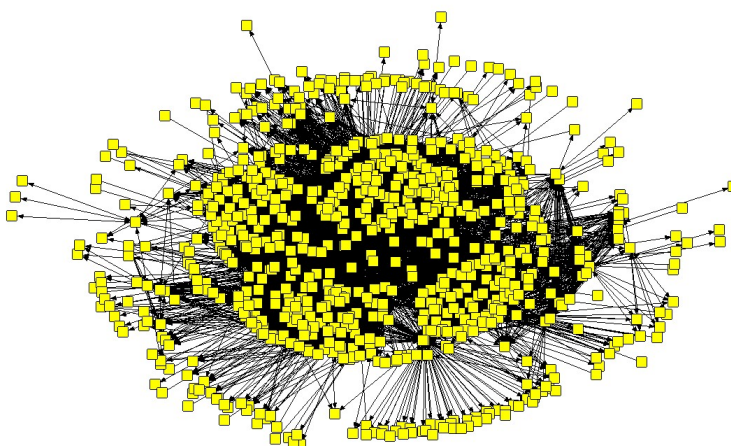
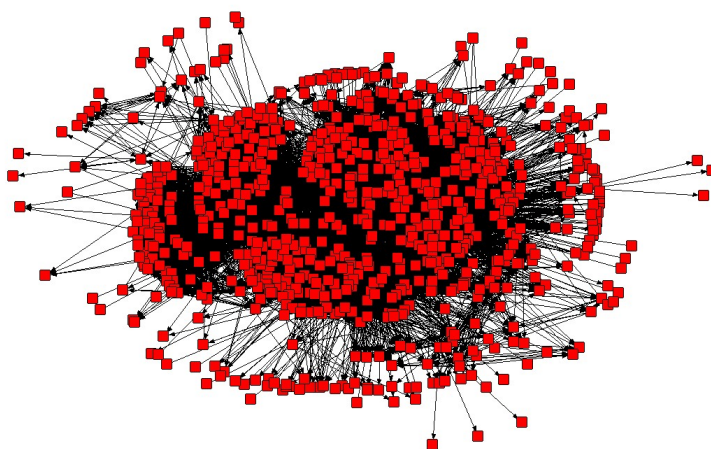


Figura A.8 Visualização da rede de atenção completa do a.m.i.g.o.s.**Tabela A.1** Tamanho e propriedades das redes

Rede	Tam. Componente	Densidade	Desv. Padrão	Reciprocidade
Contatos	680	0,01	-	0,55
Tópicos	618	1,57	64,72	0,42
Narrativa	815	2,37	38,04	0,53
Completa	838	3,09	63,93	0,57

Antes de prosseguir com a análise da influência, podemos avaliar o quanto de informação mútua existe nas redes mensuradas. Como foi mostrado na Seção 6.3, redes com alto grau de sobreposição podem ser descartadas uma em favor da outra para reduzir o conjunto a ser analisado na próxima etapa. Na Tabela A.2 temos o resultado para as quatro redes mensuradas. Se considerarmos separadamente tópicos e histórias, veremos que eles tem pouca coisa em comum; mas a rede completa de atenção por sua vez é fortemente correlacionada com ambas, sugerindo sua utilização no lugar das primeiras. Também notamos uma leve para moderada correlação positiva entre a rede de contatos e a de atenção, apesar de não estarem relacionadas diretamente. Essa diferença entre as redes justificaria também uma forte diferença na análise da influência?

A.4 Análise da influência

Após a mensuração das redes, apresentada na seção anterior, passamos à análise de influência propriamente dita. Havíamos decidido usar o *Flow Betweenness* no passo 1 desse projeto, devido à natureza valorada da representação, porém não nos foi possível

Tabela A.2 QAP aplicado as quatro redes ($p. < 0,001$)

QAP corr	completa	contatos	histórias	tópicos
completa	1	0,22	0,67	0,82
contatos	0,22	1	0,23	0,12
histórias	0,67	0,23	1	0,13
tópicos	0,82	0,12	0,13	1

alcançar nosso intento devido ao alto custo computacional que se mostrou para essa métrica. Sendo assim, substituímos pela métrica simples de *Betweenness* que dicotomiza a rede pela média, isto é, os laços cuja força seja maior ou igual que a média continuam (1) e os abaixo são removidos (0). A métrica de Bonacich foi calculada como planejada.

As Tabelas A.3 e A.4 apresentam os resultados para a métrica de *Betweenness* nas rede de contatos e na de atenção. No cabeçalho temos o índice de centralização de Freeman que indica o quão próximo a rede está de um gráfico do tipo estrela. É interessante notar que a rede de atenção formada apenas com os tópicos e suas resposta apresentou o maior grau de centralização: 37,18%; enquanto que as histórias apresentaram: 24,20%. Destacamos em negrito os membros que se repetiram nas quatro representações: contatos, tópicos, histórias e completa. É interessante notar que apesar não haver nenhuma relação direta entre a rede completa e a de contatos, aquela manteve quase que inalterada a ordenação do grupo que intersecta ambas.

Tabela A.3 Os 9 membros mais bem posicionados na rede de Contatos

Contatos	Centralização=30,81%
Membro	<i>Betweenness</i>
3	31,06
8	13,48
135	11,44
2	8,68
200	5,14
201	4,84
796	4,82
601	4,57
665	4,55

Tabela A.4 Os 9 membros mais bem posicionados na rede de Atenção

Completa	Centralização=23,07%
Membro	<i>Betweenness</i>
3	23,18
2	12,53
8	8,99
200	3,65
796	3,37
6	1,79
665	1,7
397	1,61
201	1,5

Quando analisamos a correlação entre os resultados de centralidade das quatro redes mensuradas, apresentada na Tabela A.5, podemos perceber o impacto da decisão em usar a métrica de *Betweenness* no lugar da original. A dicotomização da rede quase que totalmente mimetiza o padrão de posicionamento encontrado na rede de contatos; i.e., removido o peso dos laços, parece-nos que os atores fazem as mesmas escolhas em termos de quem interagir e de quem adicionar aos seus contatos. Entretanto, o índice de Bonacich leva em consideração este peso para a rede de atenção, logo seus resultados devem se diferenciar bastante da rede não-textual.

Tabela A.5 Correlação do índice de centralidade

<i>Betweenness</i>	completa	contatos	histórias	tópicos
completa	1	0,89	0,96	0,94
contatos	0,89	1	0,91	0,87
histórias	0,96	0,91	1	0,93
tópicos	0,94	0,87	0,93	1