UMA METODOLOGIA PARA A PREDIÇÃO DE AFINIDADES E FORMAÇÃO DE GRUPOS DE TRABALHO A PARTIR DE REDES SOCIAIS

DOUGLAS DONIZETI DE CASTILHO BRAZ

UMA METODOLOGIA PARA A PREDIÇÃO DE AFINIDADES E FORMAÇÃO DE GRUPOS DE TRABALHO A PARTIR DE REDES SOCIAIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais – Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Pedro Olmo Stancioli Vaz de Melo
Coorientador: Fabrício Benevenuto de Souza

Belo Horizonte

Fevereiro de 2014

Braz, Douglas Donizeti de Castilho

B827m Uma metodologia para a predição de afinidades e formação de grupos de trabalho a partir de redes sociais / Douglas Donizeti de Castilho Braz. — Belo Horizonte, 2014
xxii, 75 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Orientador: Pedro Olmo Stancioli Vaz de Melo
Coorientador: Fabrício Benevenuto de Souza

1. Computação – Teses. 2. Redes Sociais Online – Teses. 3. Relações Humanas – Teses. 4. Conflito interpessoal. – Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma metodologia para a predição de afinidades e formação de grupos de trabalho a partir de redes sociais

DOUGLAS DONIZETI DE CASTILHO BRAZ

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. ANTONIO ALFREDO FERREIRA LOUREIRO
Departamento de Ciência da Computação - UFMG

PROF. LUCIANO ANTONIO DIGIAMPIETRI
Escola de Artes Ciências e Humanidades - USP

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 27 de fevereiro de 2014.

À Maria Aparecida, Leandro, Lailiane e Dênis.

Agradecimentos

Deixo meus agradecimentos para todos que me encorajaram e contribuíram para o desenvolvimento deste trabalho. Em especial, gostaria de agradecer:

- Primeiramente à DEUS e à Nossa Senhora Aparecida, por estarem sempre presentes em meus trabalhos, pois sem a benção DELES não teria conseguido chegar até o fim.
- Aos meus pais, por acreditarem em mim, me apoiarem em todas as minhas decisões, sempre estarem ao meu lado me incentivando.
- À minha namorada, por estar presente em todos os momentos, nos bons e nos ruins, e sempre me ajudar a acreditar mais em mim mesmo e nunca desistir.
- Aos meus orientadores, Pedro Olmo e Fabrício Benevenuto, por acreditarem no trabalho e dedicarem seus esforços na sua construção.
- Aos meus amigos de república, pela companhia e pelos momentos de distração, que fazem parte, mesmo que indiretamente, da evolução dos trabalhos.
- Aos professores Humberto Brandão e Maria Regina Martinez, pela ajuda na concepção da metodologia empregada neste trabalho.

"Não existe nada de completamente errado no mundo,
mesmo um relógio parado, consegue estar certo duas vezes por dia."
(Paulo Coelho)

Resumo

Estudantes muitas vezes têm de se unir para realizar projetos de classe. Nesse processo, a nossa conjectura é que eles se agrupam não só a partir dos seus desempenhos em sala de aula (dado pelas notas, por exemplo), mas também a partir dos seus relacionamentos interpessoais (por exemplo, se eles confiam uns nos outros). Com o melhor de nosso conhecimento, não foram encontrados estudos na literatura sobre a relação entre a formação de equipes para projetos de classe e mídias sociais. Neste contexto, propomos uma nova metodologia para identificar afinidades de trabalho a partir de dados de mídias sociais. A metodologia proposta foi aplicada a um grupo de estudantes universitários de uma dada sala de aula. Primeiramente, através de um questionário, pedimos aos alunos para nos dizer com quem eles gostariam ou não de trabalhar nessa sala. Em seguida, via uma aplicação desenvolvida a partir da API do Facebook, coletamos dados de suas interações sociais online e, por fim, testamos alguns preditores para formação de equipe. Com isso, pôde-se concluir que o processo de escolha de colaboradores pode ser melhor estimado através de métricas sociais derivadas do Facebook para a força dos relacionamentos, popularidade, extroversão e homofilia, do que através de métricas relacionadas ao desempenho dos alunos. Estes resultados têm importantes implicações teóricas para a literatura de formação de equipes. Atualmente, empresas trabalham com dinâmicas cooperativas, onde o relacionamento interpessoal harmônico é de grande interesse das instituições. Práticas para agrupamentos harmônicos são utilizadas por departamentos de recursos humanos para maximizar o desempenho de equipes de trabalho. Além disso, existem implicações práticas para plataformas educacionais *online*, na realização de atividades colaborativas.

Palavras-chave: Redes Sociais *Online*, Relacionamento Negativo, Formação de Times, Teste Sociométrico, Capital Social.

Abstract

Students often have to team up for class projects. In this process, our conjecture is that they group based not only on the performance in the classroom (the grades of the students, for example), but also from their interpersonal relationships (for example, if they trust each other). With best of our knowledge, no studies were found on the relationship between team formation for class projects and social media. In this context, we propose a new methodology to identify work affinities from social media data. The proposed methodology was applied in a group of college students in a given classroom. First, through a questionnaire, we asked students to tell us who they would like to work or not in this room. Then, through an application developed from the Facebook API, we collected data from their online interactions, and finally tested some predictors for team building. Thus, it was concluded that self-organized team members selection can be better estimated from Facebook-derived proxies for the strength of the ties, popularity, extraversion and homophily than proficiency metrics. These results have important implications for the theoretical literature teaming. Currently, companies work with dynamic cooperative, where the harmonic interpersonal relationships is of great interest institutions. Practices for harmonic grouping are used by human resources departments to maximize the performance of work teams. Moreover, there are practical implications for online educational platforms in the performing of collaborative activities.

Keywords: Online Social Networks, Negative Relationships, Team Formation, Sociometric Test, Social Capital.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Motivação

Durante as nossas vidas, tarefas colaborativas são executadas em uma ampla e diversificada gama de atividades. De fato, é parte das nossas rotinas escolher ou ser escolhido por alguém para fazer uma tarefa colaborativa. Selecionar estudantes para participar de um projeto da escola, contratar funcionários para uma empresa, escolher jogadores para uma partida amistosa de futebol e selecionar colegas para abordar um problema de pesquisa são apenas uma pequena amostra de decisões envolvidas em atividades colaborativas que as pessoas eventualmente fazem em suas vidas. Diante deste contexto, surge a questão: *Quais fatores influenciam tais decisões, ou seja, quais fatores são determinantes para selecionar/repelir alguém para uma determinada tarefa colaborativa?* Pode-se responder a esta questão dizendo que a proficiência (ou a habilidade) de uma pessoa para fazer a tarefa determina se ela será selecionada. Apesar de acreditarmos que a proficiência desempenha um papel importante na decisão, pode-se fazer outra pergunta: *Existem informações registradas nas redes sociais que podem ajudar na formação de grupos de trabalho?* As redes sociais *online*, tais como o Facebook e o Google+, são capazes de imitar o ambiente social real em um virtual [Zhao et al., 2012]. Analisando como as pessoas se comportam nesses ambientes sociais virtuais podemos dizer como elas se comportam no real e, portanto, verificar se o seu comportamento social impacta nas suas decisões de colaboração.

Considere, por exemplo, a crescente expansão dos MOOCs (*Massive Open Online Course* - Cursos *Online* em Massa) [Daniel, 2012], que são cursos gratuitos ministrados através da Web que atraem um volume enorme de estudantes por todo o mundo.

Existem inúmeros sites que oferecem cursos *online*, e o *Coursera*[1] está entre os mais populares. Segundo Waldrop & Magazine, até março de 2013, o *Coursera* sozinho havia registrado cerca de 2,8 milhões de estudantes espalhados pelo globo. Como devem ser formadas as colaborações nessa modalidade de curso? Assim, surge uma perspectiva para a análise de dados das interações sociais online para responder a essa pergunta. Dado o grande número de pessoas no mundo que tais sites servem, uma forma eficaz de unir os estudantes em escala para realizar tarefas em equipe seria utilizar as características estudadas aqui.

1.2 Objetivos

Este trabalho tem como principal objetivo investigar o potencial das interações registradas em redes sociais *online* na predição de afinidades e formação de grupos de trabalho. Para verificar se o comportamento social impacta em tais decisões, propomos a seguinte metodologia. Primeiro deve-se extrair de um grupo de pessoas, o qual será insumo para realização do experimento, como é a relação de trabalho entre eles. Em seguida, deve-se coletar informações sobre as interações *online* deste grupo de pessoa, e por fim, estabelecer um mapeamento entre as interações *online* e as relações de trabalho deste grupo. Este mapeamento é feito através de características sociais que são extraídas dos dados coletados. Neste estudo, aplicamos essa metodologia em um cenário muito particular. Primeiro, foi aplicado um teste sociométrico em uma classe de alunos de graduação, em que os indivíduos foram questionados se gostariam de trabalhar com todos os outros alunos da classe. Em seguida, a partir de um aplicativo que desenvolvemos para o Facebook, foram coletados dados que contém uma série de características sociais sobre seus perfis e suas interações. Além disso, foram coletados as notas destes alunos, como forma de mensurar sua proficiência em desempenhar atividades de trabalho. Este cenário é apropriado porque os estudantes universitários muitas vezes têm que se unir para projetos de classe, e os mesmos se conhecem há no mínimo dois anos. Embora possa parecer simples e natural a forma como os estudantes escolhem seus grupos em uma sala de aula, acredita-se que o processo que determina as suas escolhas envolve uma mistura complexa de atributos sociais e habilidades, a fim de criar uma equipe que seja bem sucedida e agradável de trabalhar.

Para alcançar o objetivo principal, alguns objetivos específicos são necessários:

- Identificar características *online* que podem descrever a relação entre as pessoas.

[1] `https://www.coursera.org/about`

- Avaliar a capacidade de predição destas características para descrever as afinidades e formação de grupos de trabalho.

- Estabelecer um comparativo entre a capacidade de predição das características sociais e características da proficiência dos alunos.

- Avaliar quais características são mais importantes para descrever as afinidades e formação de grupos de trabalho.

1.3 Contribuições

A partir da análise dos dados coletados, revelamos uma série de conclusões interessantes. Primeiro, usando as notas dos alunos para inferir as habilidades individuais, foi descoberto que os estudantes mais qualificados nem sempre foram preferidos, indicando que o capital social tem um papel importante para determinar as suas escolhas. Então, foram investigadas uma série de características extraídas dos dados do Facebook relacionados à força da amizade, à popularidade do indivíduo no Facebook, se a pessoa é extrovertida, e sua similaridade com outros estudantes. A análise revela ao menos sete características extraídas do Facebook que são mais informativas do que as notas para determinar a disposição dos alunos para trabalhar em conjunto.

Embora as descobertas sejam retiradas de um cenário particular de sala de aula, elas têm implicações mais amplas. Para o problema de formação de equipe, os resultados mostram que os dados de uma rede social *online* podem indicar se dois indivíduos gostariam ou não de trabalhar juntos. Os resultados também podem ser aproveitados em várias aplicações *online*, como sistemas de recomendação de equipe e colaboração. Além disso, a metodologia proposta neste trabalho pode ser utilizada para a realização de outros experimentos com objetivo de estabelecer comparações entre ambiente *online* e *offline*.

Em resumo, as principais contribuições deste trabalho são:

- Criação de uma metodologia para a predição de afinidades e formação de grupos de trabalho a partir de redes sociais *online*.

- Caracterização de uma rede social de estudantes agrupada de acordo com as suas afinidades para execução de atividades colaborativas.

- Mapeamento entre ambientes *online* (Facebook) e *offline* (Teste Sociométrico);

1.4 Organização do Texto

O restante deste trabalho foi dividido em 6 capítulos, organizados da seguinte forma. No Capítulo 2 apresentamos uma visão geral sobre redes sociais e algumas análises que podem ser realizadas, e também alguns dos principais trabalhos relacionados ao tema desta dissertação. Em seguida, no Capítulo 3, descrevemos a metodologia para realização do experimento, assim como a organização dos dados utilizados. No Capítulo 4 é realizada a caracterização dos dados coletados, assim como a organização das informações dos dados do teste sociométrico. O Capítulo 5 apresenta a metodologia para extração de dados derivados da rede social Facebook. No Capítulo 6 investigamos alguns modelos para classificação das afinidades de trabalho através dos dados sociais, e também uma análise sobre quais são os principais fatores que descrevem esses relacionamentos. Finalizando, no Capítulo 7 apresentamos as conclusões e alguns direcionamentos para trabalhos futuros.

Capítulo 2

Referencial Teórico

Neste capítulo será apresentada uma visão geral dos principais conceitos e métricas aplicados na análise de redes sociais e que foram utilizados no desenvolvimento deste trabalho. Além disso, são revisados alguns estudos realizados anteriormente sobre relacionamentos negativos em redes sociais, formação de equipes, capital social em redes sociais, e comportamento *online vs offline*.

2.1 Preliminares

Redes sociais *online* [Mislove et al., 2007; Garton et al., 1997], tais como *Facebook* e *Google+*, têm se tornado uma forma popular para pessoas se expressarem, conectar com suas famílias, amigos e colegas, compartilhar informações entre si, tais como fotos e vídeos, e obter atualizações em tempo real de notícias e eventos. Esta popularização tem atraído cada vez mais usuários para estas redes, fazendo com que as relações estabelecidas entre as pessoas no mundo real sejam naturalmente refletidas na composição estrutural destas redes [Zhao et al., 2012].

Redes sociais são comumente representadas através de grafos direcionados ou não-direcionados. Um grafo é um formalismo matemático que serve para representar objetos e as relações entre eles. Esta simples estrutura é comumente usada para modelar uma grande diversidade de aplicações, tais como circuitos elétricos, estradas, relações sociais, dentre outras. Neste contexto, os vértices de um grafo representam pessoas e uma aresta entre dois vértices u e v representa um relacionamento entre eles [Easley & Kleinberg, 2010]. Muitas vezes, a abordagem utilizada para modelagem de uma rede social considera somente arestas que mapeiam relacionamentos positivos [Kunegis et al., 2009]. Porém, uma rede social não é composta somente por interações positivas, i.e., alguns relacionamentos entre usuários podem ser caracterizados como antipatia ou

inimizade [Leskovec et al., 2010a]. Dessa forma, um relacionamento negativo entre dois usuários u e v pode ser modelado através de uma aresta negativa (u, v) no grafo.

Estudos sobre redes e suas conexões, tais como a análise de redes complexas [Newman, 2003], são temas interdisciplinares que permeiam diversas áreas do conhecimento, como Ciência da Computação, Matemática, Física, Biologia e Sociologia. A análise destas redes nos permite a extração de novos conhecimentos sobre o domínio do sistema em questão [Newman, 2003]. Na seção seguinte serão apresentadas algumas métricas utilizadas na análise de redes sociais e que foram empregadas no desenvolvimento deste trabalho.

2.2 Análise de Redes Sociais

Nesta seção apresentamos métricas que tipicamente são utilizadas na análise de redes sociais. Algumas métricas podem ser aplicadas em redes complexas genéricas, sendo baseadas na topologia da rede. Outras são aplicadas especificamente no contexto de rede sociais.

2.2.1 Grau dos Vértices

O grau é uma das características mais fundamentais de um vértice [Dorogovtsev & Mendes, 2004]. Com base no grau dos vértices é possível derivar muitas medições para a rede. Na teoria dos grafos, o grau (ou valência) de um vértice v é o número de arestas incidentes em v. O grau de um vértice é comumente denotado como $deg(v)$ ou $grau(v)$ [Costa et al., 2007].

Considere o grafo direcionado $G_D(V, A)$, em que V é o conjunto de vértices e A o conjunto de arestas deste grafo. Sendo G_D um grafo direcionado, as arestas pertencentes ao conjunto A possuem direcionamento entre os vértices, ou seja, $v \rightarrow u$ indica que o vértice v alcança u, mas u não necessariamente alcança v. Assim, os vértices pertencentes ao conjunto V possuem $grau_{entrada}(v)$ e $grau_{saida}(v)$ que são, respectivamente, o conjunto das arestas que incidem sobre o vértice v e o conjunto de arestas que o vértice v emite.

Em grafos não direcionados, as arestas não possuem orientação, i.e., se existe uma aresta entre dois vértices u e v, $v \rightarrow u$ e $v \leftarrow u$. Portanto, para grafos não direcionados, $grau_{entrada}(v) = grau_{saida}(v)$. Como vértices unidos por uma aresta são denominados vértices adjacentes, o número de vértices adjacentes ao vértice v em um grafo não direcionado é equivalente ao $grau(v)$.

2.2.2 Centralidade

Métricas de centralidade em um grafo determinam a importância relativa de um vértice no grafo, por exemplo, o quanto uma pessoa é influente dentro de uma rede social [Newman, 2003]. Em outras palavras, métricas de centralidade são utilizadas para quantificar o quão vértices de um grafo são posicionados mais ao centro de sua estrutura do que outros. Existem inúmeras formas de mensurar a centralidade de um vértice em uma rede, que são baseadas em diferentes características do grafo, tais como os conceitos de distância entre vértices ou do grau dos vértices. Alguns exemplos de métricas amplamente explorados são *closeness*, *degree centrality* e *betweenness*.

A centralidade de grau (*degree centrality*) é conceitualmente a mais simples e também aquela que é amplamente utilizada neste trabalho, a qual é definida como o número de ligações incidentes sobre um vértice, i.e., o número de arestas que um nó possui [Costa et al., 2007]. Assim, $degree\ centrality(v) = grau(v)$. Utilizando o conceito de distância em grafos, temos a métrica *closeness* (proximidade), que é definida através do tamanho do caminho mínimo entre todos os pares de vértices. O distanciamento de um vértice v é definido como a soma do tamanho dos caminhos mínimos entre v e todos os outros vértices, e o *closeness* é o inverso dessa soma [Sabidussi, 1966]. Ainda sobre a estrutura da rede, temos a métrica de centralidade *betweenness*, que quantifica o número de vezes que um vértice aparece nos caminhos mínimos entre todos pares de vértices de um grafo. No contexto da Web, uma métrica de centralidade amplamente estudada é o $PageRank^{TM}$, que é a base da máquina de busca do *Google* [Page et al., 1999].

2.2.3 Assortatividade

A assortatividade (*assortative mixing*) é uma medida relacionada à conectividade entre vizinhos, e mensura a correlação entre os graus dos vértices que compartilham uma conexão. Em outras palavras, o coeficiente de assortatividade indica se vértices de grau similar tendem a conectar entre si (rede assortativa), ou se vértices com alto grau tendem a se conectar com vértices de baixo grau (rede não assortativa) [Antiqueira, 2011; Newman, 2002].

O coeficiente r que mede a assortatividade pode variar entre -1 e 1. Se a rede possui assortatividade negativa ($r < 0$), vértices que possuem grau elevado tendem a se conectar a vértices com menor grau. Se $r > 0$ (assortatividade positiva) indica que a rede possui propriedades assortativas, ou seja, vértices com graus semelhantes tendem a estabelecer conexões na rede. Quando $r = 0$, não há correlação entre os graus dos vértices. No caso de redes direcionadas, geralmente considera-se correlações

entre $grau_{entrada}$ *vs* $grau_{entrada}$ e $grau_{saida}$ *vs* $grau_{saida}$. Newman [2003] também define esta métrica como sendo uma forma de representar a *homofilia*, que é a tendência dos indivíduos associarem e estabelecer vínculo com outros semelhantes [McPherson et al., 2001]

2.2.4 Força dos Relacionamentos

A *força do relacionamento* mede o quão próximos dois indivíduos são. Mark Granovetter introduziu o conceito "força do relacionamento" (*tie strength*) com o trabalho "*Strength of Weak Ties*" [Granovetter, 1973]. A força de um relacionamento é uma combinação (provavelmente linear [Gilbert & Karahalios, 2009]) da quantidade de tempo, da intensidade emocional, da intimidade (confidência mútua), e dos serviços reciprocamente prestados, que caracterizam o relacionamento [Granovetter, 1973].

Basicamente, os relacionamentos podem ser classificados em dois tipos: fortes e fracos. Laços fortes caracterizam relacionamentos entre pessoas que se confiam e cujos círculos sociais sobrepõe firmemente entre eles. Laços fracos, por outro lado, caracterizam relacionamentos entre pessoas que são apenas conhecidas. No entanto, laços fracos muitas vezes fornecem acesso a novas informações, que normalmente não circulam na rede composta majoritariamente por laços fortes [Granovetter, 1973]. De maneira geral, a literatura sugere sete dimensões para força de um relacionamento: Intensidade, Intimidade, Duração, Reciprocidade de Serviços, Estrutural, Apoio Emocional e Distância Social [Gilbert & Karahalios, 2009].

2.3 Relacionamentos Negativos em Redes Sociais

Muitas vezes, a abordagem utilizada para modelagem de redes sociais considera somente arestas que mapeiam relacionamentos positivos. Porém, uma rede social não é composta somente por relacionamentos positivos. Alguns relacionamentos entre usuários podem ser caracterizados como antipatia ou inimizade [Kunegis et al., 2009].

A principal característica da base de dados utilizada neste trabalho está relacionada com a existência de arestas negativas [Easley & Kleinberg, 2010]. As arestas negativas, no contexto das relações de trabalho, indicam que um indivíduo não gostaria de realizar alguma atividade em grupo com o outro. Algumas redes sociais *online* possuem ferramentas que permitem a caracterização dos relacionamentos como negativos. Em Leskovec et al. [2010a] é realizado um estudo sobre redes sociais Epinions, Slashdot e Wikipedia. Essas redes sociais também têm como principal característica a ocorrência de relacionamentos positivos e negativos entre os usuários, e são alvo de vá-

rios trabalhos na literatura [Anchuri & Magdon-Ismail, 2012; Massa & Avesani, 2005; Maniu et al., 2011; Leskovec et al., 2010b]. Apesar do propósito destas redes ser diferente da rede utilizada neste trabalho, algumas análises podem ser aproveitadas. Como em Kunegis et al. [2009], onde é realizado um estudo sobre as relações dos usuários na rede social Slashdot, e as técnicas de análise da rede social foram adaptadas para o problema de arestas com pesos negativos. Mais especificamente, foram analisadas características globais da rede, como coeficiente de agrupamento, características dos vértices, tais como medida de centralidade e popularidade, e características das arestas, tais como distância e similaridade. Em Facchetti et al. [2011] é verificada a *Teoria do Equilíbrio Estrutural* [Easley & Kleinberg, 2010] nas três bases citadas, assim como é verificado para a base de dados deste trabalho (Seção 4.2).

Outro conjunto de análises geralmente realizadas em redes sociais *online* que possuem relacionamentos negativos é relacionado à previsão de links, ou seja, se um relacionamento será positivo ou negativo. Leskovec et al. [2010a] abordam o problema de previsão de links utilizando dados extraídos através da estrutura da rede, como grau dos vértices. Em Symeonidis et al. [2010] é proposta uma abordagem para recomendação de amizades baseada na estrutura da rede, inclusive quando esta possui relacionamentos negativos, evitando recomendação de amizades entre relacionamentos possivelmente negativos. Por outro lado, Kunegis et al. [2013] avaliam o valor dos relacionamentos negativos para previsão de outros relacionamentos negativos, propondo um modelo que utiliza somente informações estruturais da rede (centralidade e *closeness*) para prever relacionamentos negativos.

2.4 Formação de Times

Há uma ampla literatura relacionada com a formação de equipe. A maior parte tem-se centrado sobre o problema de como identificar os membros de um grupo que são coletivamente mais adequados para a resolução de uma tarefa específica. Em Wi et al. [2009], por exemplo, o problema é modelado como sendo um problema de programação inteira para encontrar uma correspondência ideal entre indivíduos e requisitos. Agustín-Blas et al. [2011], em vez disso, propuseram particionar a matriz pessoal de recursos de uma forma que todos os membros da equipe compartilhem a maioria do conhecimento dos recursos da equipe.

Essas abordagens, no entanto, não consideram se os membros da equipe tendem a apreciar as relações pessoais frutíferas. Para corrigir isso, pesquisadores propuseram incrementar as abordagens existentes a partir do temperamento dos membros [Fitz-

patrick & Askin, 2005] e com atributos interpessoais [Chen & Lin, 2004]. Não parece haver qualquer trabalho sobre a formação da equipe que se propõe a incrementar essas abordagens tradicionais com recursos *online* derivados de sites de mídia social, como este trabalho objetiva fazer. Em Sparrowe et al. [2001] é realizado um estudo de campo envolvendo 190 funcionários que forneceu evidências de que as redes sociais, definidas em termos de relações positivas e negativas, estão relacionadas com o desempenho individual e de grupo.

2.5 Capital Social

O termo capital social tem sido utilizado em uma variedade de contextos [Coleman et al., 1989]. Geralmente significa a capacidade das pessoas em garantir benefícios apenas por serem membros de grupos sociais específicos, ou ocupando posições vantajosas em uma rede social [Portes, 2000; Easley & Kleinberg, 2010]. Por exemplo, os indivíduos que pertencem a vários grupos tendem a transmitir informações valiosas de um grupo para outro.

Em estudos de sociologia e de marketing, o capital social tem sido muitas vezes usado para explicar por que determinados indivíduos são mais propensos a se deparar com novas oportunidades de emprego [Granovetter, 1973]. Mais recentemente, também tem sido associada com a eficácia de um grupo [Oh et al., 2006]. Em Mobius et al. [2004] são estabelecidas formas de mensurar o capital social de um indivíduo em redes sociais *online*, assim como Ellison et al. [2007] realizaram um estudo sobre os relacionamentos de usuários do Facebook, a formação e a manutenção do capital social. Burt [2000] apresentou uma revisão sobre as evidências da existência de uma conexão entre redes sociais *online* e capital social.

2.6 Comportamento *Online* vs *Offline*

Redes sociais tendem a refletir o comportamento e as relações interpessoais existentes no mundo real [Garton et al., 1997]. Uma grande quantidade trabalhos de pesquisa analisam até que ponto os dados de rede sociais *online* podem ser usados para inferir o comportamento *offline*. Jones et al. [2013] administraram uma pesquisa para os usuários do Facebook: eles pediram aos indivíduos para nomear os seus melhores amigos. Assim, relacionaram este levantamento de dados com o número de mensagens públicas e privadas entre os usuários e correspondentes melhores amigos no Facebook. Eles mostraram que a comunicação pública é tão informativa quanto mensagens privadas

para inferir a força de um relacionamento. Xiang et al. [2010] propuseram um modelo para prever a força de um relacionamento através de interações no Facebook e número de amigos comuns. Xie et al. [2012] estudaram as características comportamentais associadas aos usuários do Twitter que por acaso são colegas ou amigos na vida real. Mansson & Myers [2011] analisaram como os estudantes universitários expressam afeto a seus amigos mais próximos no Facebook, e identificaram 30 principais formas de expressão.

No contexto da extração de informações a partir de dados coletados de usuários através do Facebook, o trabalho de Zhao et al. [2012] propõe um framework para mensurar a força de um relacionamento entre dois indivíduos, assim como Gilbert & Karahalios [2009]. A abordagem apresentada nesse trabalho utiliza dados dos perfis dos usuários e interações entre os indivíduos em diferentes atividades, tais como trabalho, esporte e entretenimento. Assim, é estimada a força dos relacionamentos entre os usuários em diferentes atividades, diferentemente do trabalho de Gilbert & Karahalios [2009] que avalia a força de um relacionamento "em geral".

Capítulo 3

Metodologia e Base de Dados

Como estamos interessados em analisar características de comportamento em um ambiente *online* que podem descrever afinidades de trabalho (ambiente *offline*), propusemos uma metodologia que nos permite realizar esta análise. A Figura 3.1 apresenta o diagrama de atividades desta metodologia proposta. Nesta seção serão descritas algumas das atividades que compõem esta metologia, relacionadas principalmente à etapa de coleta dos dados. As demais atividades de análise dos dados serão detalhadas nas seções seguintes.

Primeiramente, é necessário identificar o conjunto de pessoas que farão parte do experimento. Assim. foi selecionada uma sala de aula de alunos de graduação de uma universidade anônima, de um país anônimo. Em seguida, é necessário identificar as afinidades de trabalho existentes entre esse grupo de pessoas. No nosso caso, isso foi feito através de um teste sociométrico, em que foi perguntado a cada aluno se ele gostaria de trabalhar com os demais alunos dessa mesma sala de aula. Esse mapeamento nos permite conhecer como são as afinidades de trabalho entre os participantes do experimento. Para analisar e compreender as suas respostas, é necessário, então, coletar informações relevantes sobre o seu desempenho em sala de aula, ou seja, as suas notas, e também sobre como os mesmos interagem socialmente com os outros alunos. Para essa última, coletamos um conjunto de interações *online* registradas pelo Facebook dos alunos, através de um aplicativo desenvolvido a partir da API (*Application Programming Interface*) para o Facebook. Estes conjuntos de dados são adequados para resolver as questões colocadas porque (i) cada aluno respondeu à pergunta sobre todos os outros alunos e (ii) todos eles se conhecem pessoalmente há no mínimo dois anos, se encontrando ao menos duas vezes por semana. Nas próximas seções estão descritos os detalhes desta metodologia para coleta de dados.

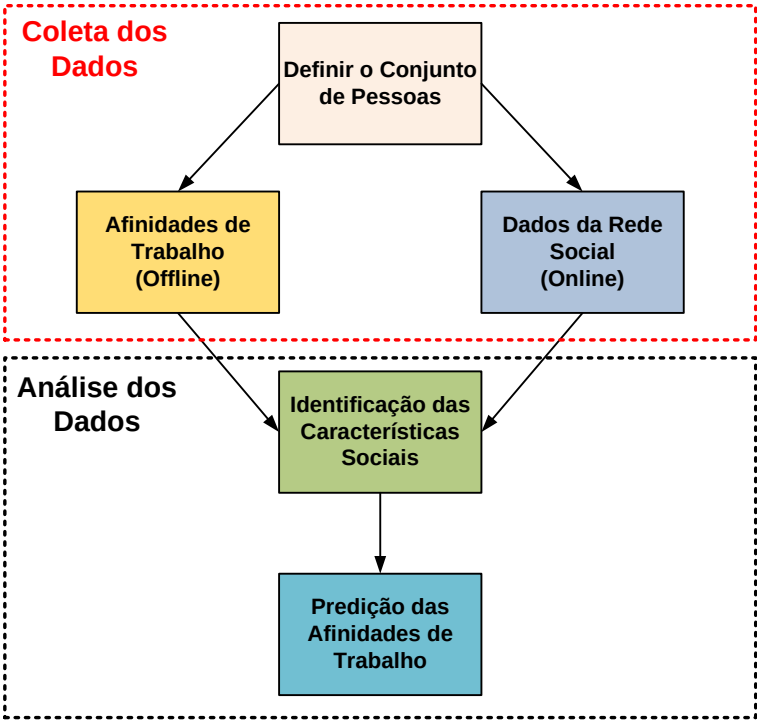


Figura 3.1: Diagrama de atividades da metodologia empregada neste trabalho.

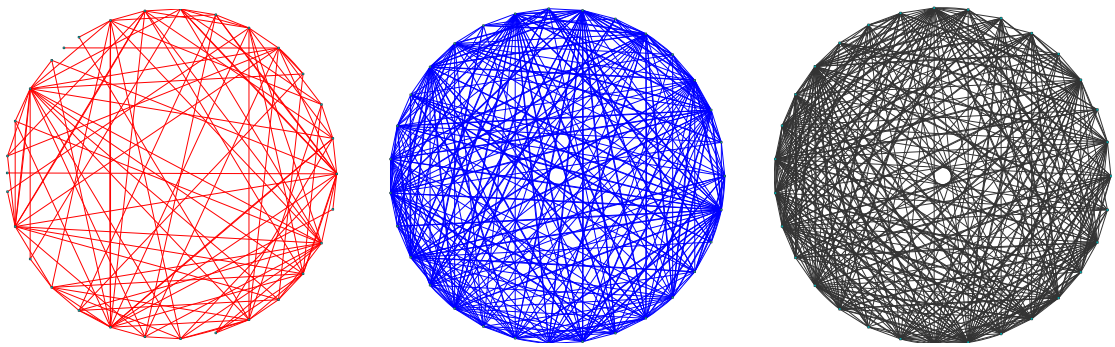
3.1 O Teste Sociométrico

A sociometria é um método quantitativo para medir as relações sociais [Moreno, 1953]. Ela foi desenvolvida pelo psicoterapeuta Jacob L. Moreno em seus estudos sobre a relação entre as estruturas sociais e de bem estar psicológico. O teste sociométrico pode ser aplicado em qualquer circunstância em que se deseja entender as relações dentro de um grupo. A partir deste conhecimento é possível, por exemplo, reorganizar as conexões, a distribuição de tarefas, definir novos líderes, entre outras aplicações [Bustos, 1979]. Em geral, o teste sociométrico consiste em um questionário para cada membro de um grupo de pessoas. A partir do questionário é construído o sociograma, que é basicamente o mapeamento da rede social do grupo [Agustín-Blas et al., 2011; Moreno, 1953].

Neste experimento, o teste sociométrico foi aplicado para entender a dinâmica existente em um grupo de pessoas quando eles deveriam colaborar para realizar tarefas em grupo. Para isso, foi selecionada uma sala de aula com 31 estudantes de graduação de uma universidade anônima. Em seguida, foi aplicado um questionário para cada aluno, contendo a seguinte pergunta: "Você gostaria de trabalhar em equipe com esta pessoa?". Após esta pergunta, o questionário mostra ao participante uma lista contendo os nomes de todos os colegas de classe. Na frente de cada nome existe um

espaço em branco onde o participante poderia assinalar uma das seguintes respostas: "SIM", "NÃO" ou "INDIFERENTE." Quando a resposta é "SIM" indica que o aluno estaria interessado em executar alguma atividade de grupo com o indivíduo em questão. Quando a resposta é "NÃO", o estudante rejeita a ideia de realizar alguma atividade em grupo com o indivíduo. Finalmente, quando a resposta é "INDIFERENTE", o aluno é indiferente a esse indivíduo em particular. Um exemplo do teste pode ser verificado no Apêndice A.

Assim, temos três tipos diferentes de relações $(i \to j)$ entre estudantes i e j. Em primeiro lugar, a relação pode ser positiva, ou seja, $(i \to j) = 1$, indicando o interesse do aluno i em trabalhar com o aluno j. Em segundo lugar, o relacionamento pode ser negativo, isto é, $(i \to j) = -1$, indicando que o indivíduo i não tem interesse em trabalhar com j. Finalmente, a relação pode ser neutra, ou seja, $(i \to j) = 0$, quando o indivíduo i é indiferente em relação ao indivíduo j. Uma vez que a pesquisa foi aplicada a todos os alunos em sala de aula, e cada aluno respondeu à pesquisa em relação a todos os outros alunos, o resultado é um sociograma completo, que consiste em 930 respostas entre os 31 estudantes.



(a) Negativo (b) Positivo (c) Neutro

Figura 3.2: Grafo baseado no resultado do teste sociométrico, separado pelas classes de arestas

Este sociograma completo também pode ser visto como um grafo completo direcionado valorado $G_S(V, E_S)$, em que o conjunto de vértices V é composto pelos alunos e o conjunto de arestas direcionadas E_S são as respostas do teste sociométrico, como pode ser visto na Figura 3.2. Além disso, na Figura 3.3 é mostrado o grau de saída e o grau de entrada de cada aluno em G_S agrupados pelo sinal da aresta. Pode-se notar que não existe um padrão claro na distribuição dos dados. Por exemplo, o aluno 1 possui o maior grau de entrada positivo da turma, e ao mesmo tempo é o aluno que possui maior grau de saída negativo, não querendo trabalhar com quase metade da turma. Outro exemplo interessante é o aluno 28, que possui o maior grau de entrada

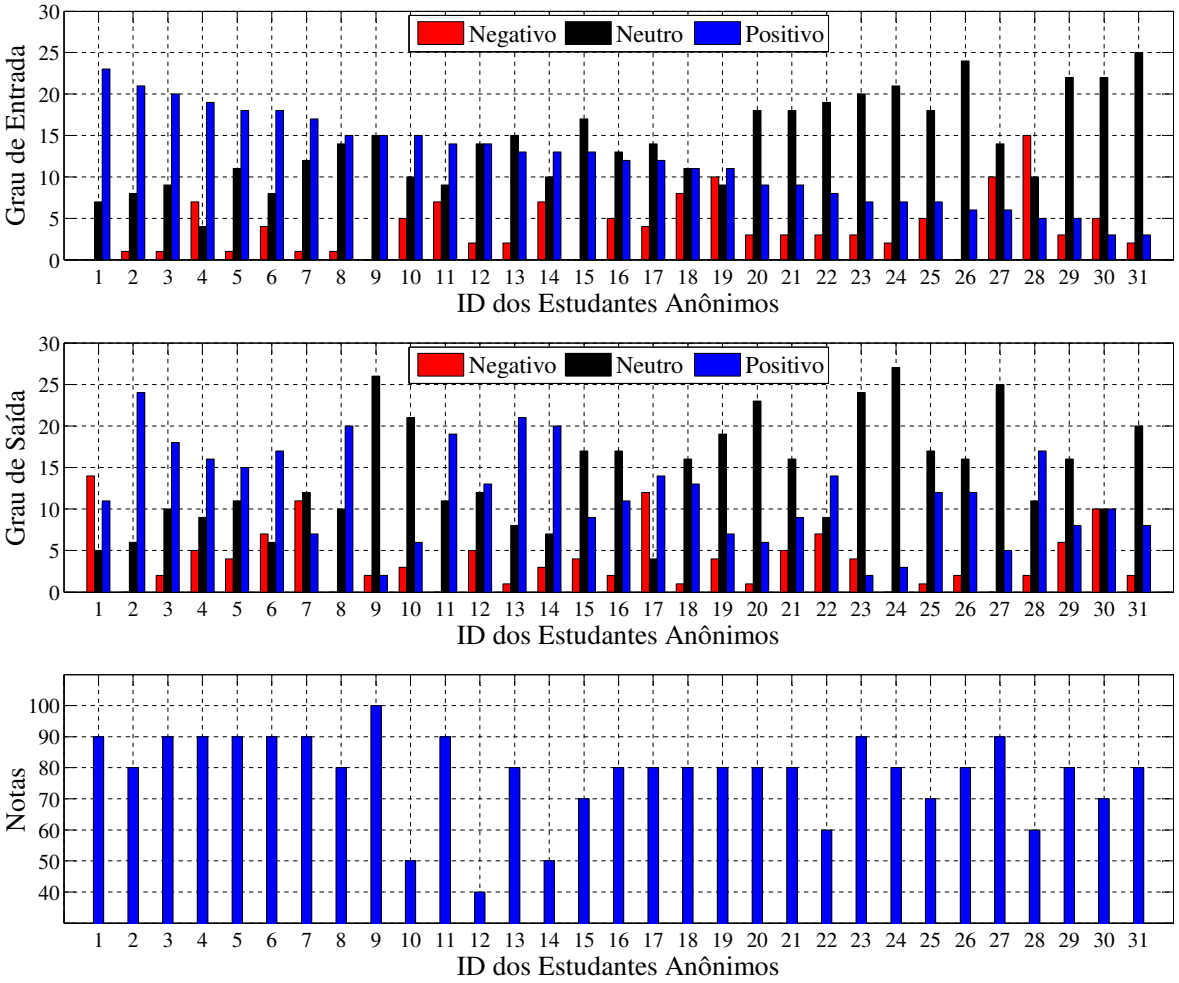


Figura 3.3: Notas dos alunos e total de respostas individuais atribuídas e recebidas no teste sociométrico

negativo, sendo que quase metade da turma não gostaria de trabalhar com ele. Em contrapartida, ele não trabalharia com poucas pessoas, apresentando grau de saída positivo equivalente a mais da metade da turma.

3.2 Aproveitamento em Classe

Em tarefas colaborativas, talvez a estratégia mais utilizada (ou esperada) para escolher os colaboradores é selecionar aqueles que são os mais proficientes para realizar a atividade. Por exemplo, considere um cenário em que uma empresa está contratando funcionários ou dois capitães de futebol estão selecionando jogadores em uma partida entre amigos. Não é um absurdo dizer que a maioria das pessoas diria que os mais qualificados seriam selecionados primeiramente. Assim, a fim de verificar se e quanto

a proficiência dos alunos está relacionada com as respostas que eles dão e recebem no questionário, foi coletado as notas que os mesmos obtiveram para esta classe específica no semestre. Ainda na Figura 3.3 são apresentadas, juntamente com o grau de entrada e saída dos alunos, as notas obtidas por eles. Observe que no exemplo dado anteriormente, o aluno 28 possui um desempenho abaixo da média (menor que 60), o que poderia explicar seu grau de entrada negativo elevado. Na Figura 3.4 é apresentado o histograma das notas obtidas pelos alunos, na faixa de 0 (pior) a 100 (melhor). Observa-se que embora a maioria dos alunos têm notas entre 71 e 90, existem aqueles que falharam no curso (notas abaixo de 60) e aqueles que obtiveram um excelente desempenho (notas acima de 90).

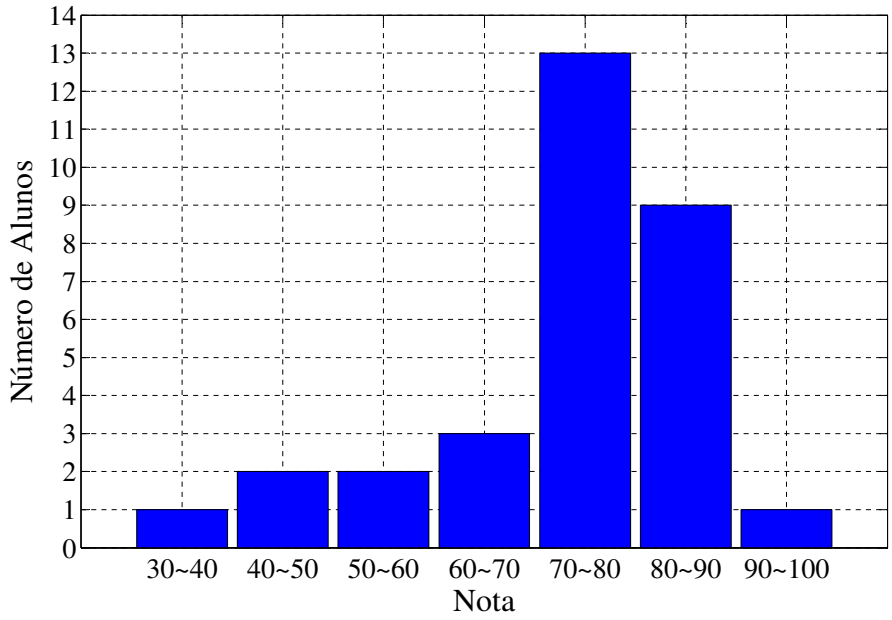


Figura 3.4: Histograma das notas

Para verificar o impacto das notas nas respostas dadas pelos alunos, foi calculado o coeficiente de correlação de *Spearman* [Fieller et al., 1957] entre o ranking dado pelas notas e o ranking dado pelo grau de entrada e saída dos alunos, agrupados pelo sinal da aresta. Será utilizado os termos $grau_{entrada}$ e $grau_{saida}$ para indicar o grau de entrada e o grau de saída de um determinado vértice, respectivamente. Além disso, é utilizado os símbolos $+$, 0 e $-$ para indicar os sinais positivos, neutros e negativos, respectivamente. Observe na Tabela 3.1 que há uma correlação significativa e de baixo *p-value* entre $grau^{+}_{entrada}$ e as notas dos alunos. Dizemos que a correlação é *significativa* quando os *p-values* são inferiores a $0,05$ (os números em negrito nas tabelas representam

correlações significativas). A partir disso, pode-se concluir que há uma tendência dos alunos mais proficientes atraírem mais respostas positivas na pesquisa, ou seja, mais alunos optam por trabalhar com ela/ele. No entanto, observando-se outras correlações, que não possuem valores significativos, ou seja, *p-values* altos, pode-se concluir que a nota de um aluno não possui relação causal com o número de respostas negativas e neutras que ele/ela recebe, e também para as respostas que ela/ele envia. Assim, embora as notas (ou a competência) dos alunos apresentarem um impacto em suas respostas, ainda há muita coisa que elas não podem explicar.

Tabela 3.1: Correlação entre notas e grau entrada/saída positivo, negativo e neutro

Grau	Coeficiente de Spearman	*p-value*
$grau^{+}_{entrada}$	**0,4727**	0,0073
$grau^{-}_{entrada}$	-0,2543	0,1674
$grau^{0}_{entrada}$	-0,3433	0,0586
$grau^{+}_{saida}$	-0,0363	0,8461
$grau^{-}_{saida}$	-0,0471	0,8014
$grau^{0}_{saida}$	-0,0373	0,8421

3.3 Dados da Rede Social – Facebook

Finalizando a etapa de coleta de dados proposta na metologia deste trabalho, apresentamos a atividade relacionada à obtenção dos dados de interações sociais online. Considere, por exemplo, as respostas positivas dadas por amigos próximos ou respostas negativas dadas entre grupos de estudantes que não se dão bem. Até que ponto uma resposta pode ser guiada por fatores semelhantes a estes? Para responder esta questão, foram coletadas interações no Facebook dos estudantes questionados na pesquisa. Para isso, foi desenvolvido um aplicativo que reúne diversas informações de suas contas do Facebook. O aplicativo foi desenvolvido com o único propósito de coletar algumas das inúmeras formas de interações entre os participantes, que concordaram em disponibilizar seus dados para coleta. É importante ressaltar que, por exigência do *Comitê de Ética para Pesquisas Humanas* da instituição onde foi realizado o experimento, todos os estudantes tiveram que aceitar participar da pesquisa, e apenas os dados relacionados a eles foram coletados. Assim, não foi coletada qualquer informação de pessoas que não estivessem participando do experimento. Um exemplo do *Termo de Consentimento Livre e Esclarecido* e da *Declaração de Aceite de Participação* podem ser vistos nos Apêndices B e C, respectivamente.

Os dados coletados são, em sua grande maioria, interações entre os participantes com os demais. As informações recuperadas através do Facebook compreendem:

1. *Amizade(i,j)* – aluno i é ou não amigo no Facebook do aluno j.
2. *Comentários em Foto(i,j)* – número de vezes que o aluno i comentou fotos pertencentes ao aluno j.
3. *Comentários em Links(i,j)* – número de vezes que o aluno i comentou links publicados pelo aluno j.
4. *Comentários em Status(i,j)* – número de vezes que o aluno i comentou atualizações de status publicadas pelo aluno j
5. *Comentários em Álbuns(i,j)* – número de vezes que o aluno i comentou álbuns de fotos pertencentes ao aluno j
6. *Conversas no Chat(i,j)* – número de mensagens inbox trocadas entre os alunos i e j.
7. *Filmes(i)* – filmes que o aluno i possui interesse.
8. *Músicas(i)* - músicas que o aluno i possui interesse.
9. *Interesses(i)* - interesses em geral que o aluno i possui.
10. *Grupos(i)* – grupos que o aluno i participa.
11. *Fotos Curtidas(i,j)* – número de vezes que o aluno i curtiu fotos pertencentes ao aluno j.
12. *Links Curtidos(i,j)* – número de vezes que o aluno i curtiu links compartilhados pelo aluno j.
13. *Status Curtidos(i,j)* – número de vezes que o aluno i curtiu atualizações de status do aluno j.
14. *Marcações(i,j)* - número de vezes que o aluno i marcou o aluno j em fotos ou comentários.

Dentre os dados coletados estão aqueles que pertencem exclusivamente ao aluno, como os filmes preferidos, as músicas preferidas, os grupos que participam e os interesses de cada um. Para utilizar essas informações, torna-se necessário estabelecer algum

Tabela 3.2: Dados coletados do Facebook, agrupados pelo sinal da aresta

Dados do Facebook	**Ocorrência por Aresta Positiva**	**Ocorrência por Aresta Negativa**	**Ocorrência por Aresta Neutra**
Número de Amigos	0,46	0,40	0,20
Conversas no Chat	672,47	109,39	166,06
Número de Marcações	0,13	0,08	0,01
Comentários em Fotos	1,34	0,34	0,38
Comentários em Links	0,69	1,43	0,77
Comentários em Atualizações	1,19	0,63	1,26
Comentário em Álbuns	0,01	0	0
Filmes em Comum	2,29	1,75	1,84
Grupos em Comum	7,71	6,94	5,79
Interesses em Comum	0,04	0,07	0,02
Músicas em Comum	1,85	2,13	2,12
Curtidas em Fotos	0,23	0,27	0,12
Curtidas em Links	0,17	0,13	0,05
Curtidas em Atualizações de Status	0,12	0,09	0,01

critério. Assim, foi utilizado o valor da intersecção ou da união entre estes conjuntos. Um resumo dos dados coletados a partir do Facebook pode ser visto na Tabela 3.2, todos agrupados pelo sinal da aresta.

Em primeiro lugar, observa-se que a ocorrência de amizade nas arestas neutras é significativamente menor do que nas arestas positivas e negativas. Além disso, é curioso ver que o número médio de comentários em links compartilhados entre arestas negativas é maior que nas positivas e neutras. No entanto, vemos que o número médio de conversas no *chat* trocadas nas arestas positivas é significativamente maior que nas arestas neutras e negativas. Finalmente, observa-se que o número de interesses comuns é muito baixo para as três classes de aresta. A partir dessas observações iniciais, percebe-se um potencial impacto das interações sociais nas respostas dadas pelos alunos. Este impacto será formalizado e quantificado nos capítulos seguintes.

3.4 Discussão

Em termos de limitações da bases de dados, alcançar uma boa representatividade é uma questão muito difícil neste estudo, como em muitas análises empíricas. Foi aplicado um teste sociométrico em uma turma de alunos de graduação, onde todos os estudantes concordaram em participar e todos eles possuem uma conta no Facebook, o que permite reunir suas interações sociais *online* através de um aplicativo. Trabalhos futuros podem

ser realizados utilizando o projeto de experimentos aqui proposto, que utilize grandes turmas de alunos, de diferentes origens e países. Além disso, embora os experimentos sejam limitados a uma turma de 31 estudantes, os objetos deste estudo são as relações entre esses alunos, o que corresponde a 930 arestas entre positivas, negativas e neutras. Para garantir que o tamanho da amostra não seja pequeno demais para tirar conclusões, em todas as análises foram aplicados testes estatísticos para verificar se os resultados são estatisticamente significativos. Outro tipo de trabalho futuro é a validação destes resultados em diferentes cenários, como empresas ou outras universidades.

Também é importante notar que o conjunto de dados do Facebook consiste em apenas estatísticas sobre as interações entre os alunos que aceitaram participar dos experimentos. A aplicação de Facebook não poderia recolher o conteúdo das mensagens trocadas pelos alunos, devido a limitações impostas pelo comitê de ética da universidade onde foi realizado o experimento. Isto impede a análise de uma série de recursos, por exemplo, os aspectos relacionados com o sentimento expresso nas mensagens trocadas entre os alunos.

Capítulo 4

Caracterização dos Dados

O conjunto de dados coletados, que compreendem os dados sociais *online* e os dados do relacionamento de trabalho, constituem um ambiente complexo de ser analisado. Assim, neste capítulo será realizado um estudo sobre as características existentes nos dados. Esta caracterização permite algumas conclusões, mas também mostra que este ambiente não é tão simples e intuitivo de ser analisado. Como este tipo de rede é baseada em afinidades de trabalho, serão realizadas análises que são geralmente aplicadas em ambientes de redes complexas [Costa et al., 2007].

4.1 Análise de Reciprocidade

A análise de reciprocidade é realizada na rede *offline* extraída do teste sociométrico. Ela consiste em avaliar como é a afinidade entre todos os pares de vértices i e j pertencentes ao grafo $G_S(V, E_S)$, ou seja, a reciprocidade entre o vértice i e o vértice j, pois ambos vértices podem ter opiniões semelhantes ou distintas entre si quanto à afinidade de trabalho.

Para realizar esta análise, as diferentes formas de reciprocidade foram modeladas da seguinte forma: se o aluno i assinalou negativo para o aluno j, e o aluno j assinalou positivo para i, então esta reciprocidade se enquadra no padrão $Positivo + Negativo$; se ambos alunos i e j responderam neutro, sua reciprocidade será o padrão $Neutro + Neutro$. Isso permite que o grafo direcionado G_S seja analisado através de um perspectiva não direcionada, definida por $G_{SND}(V, E_R)$, onde E_R representa o conjunto das arestas não direcionadas obtidas pela padronização da reciprocidade. Esta análise, aplicada ao grafo G_S, apresenta a reciprocidade existente através de 6 padrões. O comportamento da turma pode ser visto na Tabela 4.1, em que é apresentada a proporção dos relacionamentos que se enquadram em cada pa-

drão. Ainda, expandindo a análise da reciprocidade para o nível da rede social *online*, é apresentado o percentual de amizades no Facebook dentro dos conjuntos obtidos por cada padrão. A Tabela 4.1 apresenta esses valores ordenados por esse percentual.

Tabela 4.1: Reciprocidade comparada à amizade no Facebook

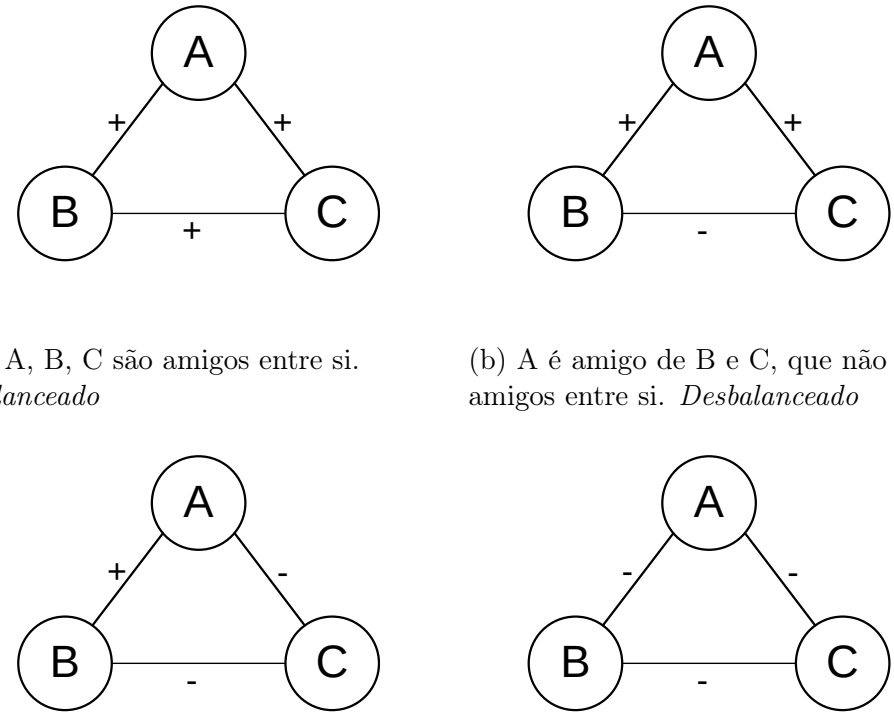
Reciprocidade	**Proporção da Turma**	**Percentual de Amizade**
$Positivo + Negativo$	10,11%	97,87%
$Positivo + Positivo$	23,66%	97,27%
$Negativo + Negativo$	2,15%	90,00%
$Neutro + Positivo$	21,94%	70,59%
$Neutro + Negativo$	11,40%	62,26%
$Neutro + Neutro$	30,75%	27,97%

Algumas conclusões preliminares podem ser feitas através da Tabela 4.1. Primeiro, observe que o número de padrões de reciprocidade $Positivo + Negativo$ é baixo se comparado com os demais padrões, sendo o segundo com menor valor de ocorrência dentro da turma. Porém, ele apresenta alto grau de amizade no Facebook, assim como o padrão $Positivo + Positivo$, com $97,27\%$ das relações contendo amizade. Outro resultado interessante acontece com o padrão $Neutro + Neutro$. Este padrão representa cerca de $30,75\%$ da base de dados, mas é o que possui menor número de amizades no Facebook, contendo apenas $27,97\%$ de amizades. Isso nos mostra que a ocorrência de amizades no Facebook é baixa entre pessoas que opinam mutuamente neutro sobre afinidades de trabalho. Uma justificativa seria que esses pares de pessoas não se conhecem muito bem no mundo real, e por isso, não possuem opinião concreta sobre querer ou não trabalhar entre eles. Outro fato interessante é o padrão $Negativo + Negativo$ ter menor número de ocorrências, mas com alto percentual de amizade no Facebook, com aproximadamente $90,0\%$.

4.2 Equilíbrio Estrutural da Rede

A *Teoria do Equilíbrio Estrutural* foi proposta inicialmente por Heider [1946] a fim de compreender a estrutura e origem de tensões e conflitos em uma rede de indivíduos, considerando que os mesmos podem ter relações de amizade ou hostilidade. A modelagem desta teoria em termos de grafos ponderados, em que os nós de um grafo representam os usuários e as arestas entre eles representam uma relação de amizade/inimizade, foi proposta em Cartwright & Harary [1956]. De forma resumida, a teoria está relacionada com o seguinte princípio: a potencial fonte de tensões em uma rede social são os ciclos

do grafo (isto é, os caminhos fechados começando e terminando no mesmo nó), mais especificamente os ciclos de paridade negativa (ou seja, aqueles que possuem um número ímpar de arestas negativas). Isso mostra que, para ciclos contendo três pessoas, temos grafos balanceados quando "*o amigo do meu amigo é meu amigo*", "*o amigo do meu inimigo é meu inimigo*", "*o inimigo do meu amigo é meu inimigo*"e "*o inimigo do meu inimigo é meu amigo*". A ideia básica da teoria do equilíbrio estrutural é apresentada na Figura 4.1 [Easley & Kleinberg, 2010].



(a) A, B, C são amigos entre si. *Balanceado*

(b) A é amigo de B e C, que não são amigos entre si. *Desbalanceado*

(c) A e B são amigos e mutuamente inimigos de C. *Balanceado*

(d) A, B e C são inimigos entre si. *Desbalanceado*

Figura 4.1: Teoria do equilíbrio estrutural

Em particular, um grafo ponderado é exatamente balanceado se, e somente se, todos os ciclos dentro deste grafo são balanceados [Facchetti et al., 2011]. Seguindo esta teoria, se um grafo apresentar algum ciclo contendo um número ímpar de arestas negativas, ele estará desbalanceado.

Para analisar se a teoria do equilíbrio estrutural se aplica a G_S, foi realizada uma análise sobre todos os ciclos mínimos do grafo seguindo o princípio supracitado. Porém, no contexto de G_S, as arestas são direcionadas. Uma aresta $i \to j$ representa a afinidade do aluno i para o aluno j. Tal fato pode afetar a teoria do equilíbrio

em alguns ambientes, mas não neste caso em particular, pois os ciclos existem no grafo mesmo com as arestas sendo direcionadas. Por exemplo, se o aluno A quer trabalhar com o aluno B, e o aluno B quer trabalhar com o aluno C, o ciclo estará completo com a opinião do aluno C para o aluno A. A análise da teoria do equilíbrio foi realizada seguindo este fluxo de informações citado como exemplo. Assim, para um grupo de 3 participantes A, B e C, apesar de termos 6 formas equivalentes de analisar a existência da teoria do equilíbrio (ABC, ACB, BAC, BCA, BAC e CBA), será computado somente 1 triângulo, i.e., o triângulo que contém o ciclo $A \rightarrow B \rightarrow C \rightarrow A$, sendo $A < B < C$. Para um grafo contendo n vértices, o número total de triângulos $total(n)$ é dado por $total(n) = n * (n-1) * (n-2)/6$.

A Tabela 4.2 apresenta os resultados obtidos através do teste para verificação da teoria do equilíbrio. O algoritmo para realizar este procedimento pode ser encontrado em Easley & Kleinberg [2010].

Tabela 4.2: Análise da teoria do equilíbrio em G_S

Arestas Negativas	**Quantidade**	**Percentual**
0	338	7,52%
1	370	8,23%
2	102	2,27%
3	4	0,09%
Não Formam Triângulo	3681	81,89%

A Tabela 4.2 mostra que existem aproximadamente 10% de triângulos balanceados, resultado da união dos triângulos que possuem zero e duas arestas negativas (ambos triângulos balanceados). Com a existência do valor neutro em algumas arestas, a não formação de triângulos é grande, assumindo $81,89\%$ de toda rede G_S proveniente do teste sociométrico. Com a existência de triângulos desbalanceados, presentes nas classes com uma e três arestas negativas, pode-se concluir que o grafo é não balanceado.

Outra análise baseada em triângulos pode ser realizada, levando-se em consideração a metodologia de análise de reciprocidade apresentada na seção anterior. Cada triângulo pode ser caracterizado através das 6 classes de reciprocidade de arestas propostas na Tabela 4.3. O número de combinações de arestas é equivalente a 216 triângulos. Assim, nesta análise foram detectados os padrões de ocorrência de triângulos formados pelas seis classes de reciprocidade de arestas, ou seja, um triângulo pode conter quaisquer das 6 classes propostas (000, 123, 555, etc).

Além da detecção dos triângulos em G_S, foi realizada uma análise aleatória de distribuição das arestas em um grafo contendo a mesma quantidade de vértices. Se-

Tabela 4.3: Padrões de reciprocidade

Reciprocidade	**Padrão**
$Negativo + Negativo$	0
$Negativo + Neutro$	1
$Neutro + Neutro$	2
$Neutro + Positivo$	3
$Positivo + Negativo$	4
$Positivo + Positivo$	5

guindo a proporção de arestas positivas, negativas e neutras em G_S (369, 120, 441 respectivamente), foram gerados 1000 grafos aleatórios contendo esta mesma proporção. Em cada grafo aleatório, também foram computados os triângulos de acordo com a reciprocidade proposta. Assim, podemos identificar como as interações entre as pessoas influenciam na formação dos triângulos de reciprocidade. O resultado desta análise pode ser visto nas Tabela 4.4.

Tabela 4.4: Análise de reciprocidade em triângulos

Padrão	**Qtd.**	**Razão**	**Qtd. Aleatória**	**Razão Aleatória**	**Qtd. / Qtd. Aleatória**
555	141	3,14%	17,06	0,38%	8,27
055	26	0,58%	5,46	0,12%	4,76
455	132	2,94%	33,77	0,75%	3,91
222	186	4,14%	49,99	1,11%	3,72
005	2	0,04%	0,57	0,01%	3,49
445	72	1,60%	22,01	0,49%	3,27
444	15	0,33%	4,72	0,10%	3,18
044	7	0,16%	2,24	0,05%	3,12
225	333	7,41%	107,00	2,38%	3,11
045	19	0,42%	7,07	0,16%	2,69

Esta análise apresentada na Tabela 4.4 evidencia a existência de uma força que influencia diretamente nas escolhas das pessoas. Nesta análise é apresentado somente os 10 resultados mais relevantes, ordenados pela razão entre a *Quantidade* de ocorrência do padrão em G_S, e a *Quantidade Aleatória* de ocorrência do padrão nos grafos aleatórios. O valor da coluna *Quantidade Aleatória* mostra a média de ocorrência do padrão nos 1000 grafos gerados aleatoriamente. Observe que o padrão 555, que representa um triângulo composto somente por arestas positivas, aparece 141 vezes em G_S, sendo este o que possui maior razão, pois em grafos aleatórios aparece significativamente menos. Isso mostra que este comportamento no mundo real é induzido por algum fator que

não existe no universo dos grafos aleatórios. O mesmo acontece com o padrão 044, que é um padrão que possivelmente gera um desbalanceamento por ter a aparição da classe 0, que é composto por $Negativo + Negativo$. Todos os padrões apresentados na Tabela 4.4 aparecem ao menos 2 vezes mais que em grafos aleatórios.

4.3 Análise de Dispersão do Dados

Nesta seção, serão apresentadas algumas análises sobre a dispersão da base de dados do teste sociométrico. Para realizar esta análise, foram utilizados os graus de entrada e saída de cada aluno. O $grau^{-}_{entrada}$ é relacionado com o $grau^{-}_{saida}$ do aluno, o $grau^{0}_{entrada}$ é relacionado com o $grau^{0}_{saida}$, e consequentemente, o $grau^{+}_{entrada}$ é relacionado com o $grau^{+}_{saida}$. Neste contexto, é obtida uma coordenada (x, y) de cada aluno em um plano cartesiano, avaliando as arestas negativas, positivas e neutras. Isso permite a caracterização de cada aluno quanto ao número de escolhas que ele faz, e o quanto ele é escolhido pela turma. O resultado desta análise pode ser visto na Figura 4.2.

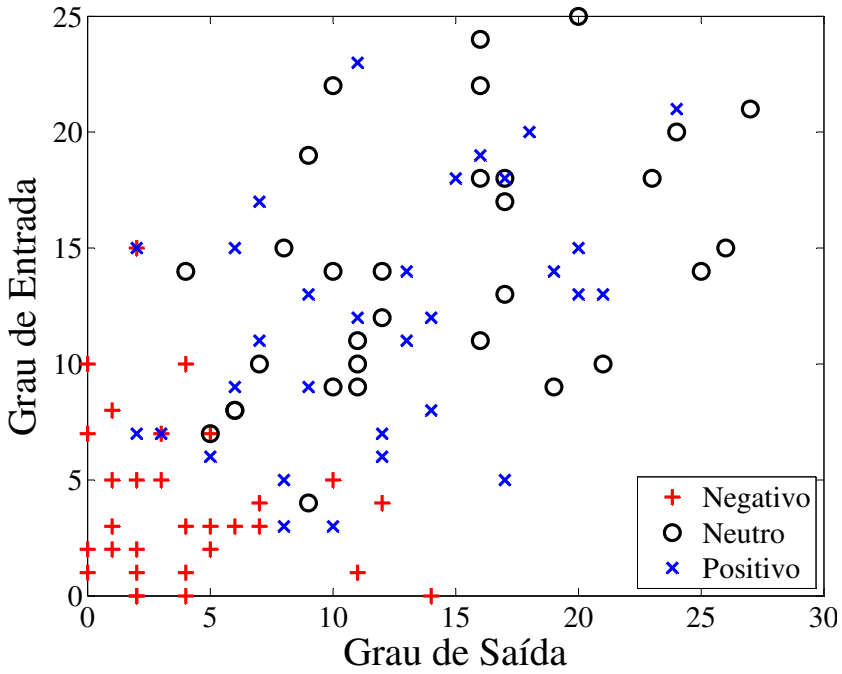


Figura 4.2: Caracterização dos alunos através do grau de entrada e saída

A Figura 4.2 mostra a diferença comportamental entre as escolhas positivas e as escolhas negativas. Entre as positivas e as neutras, este comportamento é semelhante. Analisando somente as escolhas negativas, pode-se ressaltar a existência de um aluno que não deseja trabalhar com quase metade da classe ($grau^{-}_{saida}$ 14), porém, este não

recebe nenhuma aresta negativa. Cruzando estes resultados com a análise da Figura 3.3, vemos que este aluno possui um $grau^{+}_{entrada}$ alto, aproximadamente 23 arestas. Em um caso oposto, a análise mostra a existência de um aluno que possui $grau^{-}_{entrada}$ alto (15 arestas negativas), mostrando que quase metade da classe não quer trabalhar com ele, mesmo ele respondendo que não quer trabalhar com poucas pessoas na classe ($grau^{-}_{saida}$ equivalente a aproximadamente 2). Essas duas pessoas provavelmente possuem perfis completamente diferentes dentro da classe, e seu comportamento interfere diretamente nas suas escolhas, e na escolha das outras pessoas. Seguindo a metodologia proposta na Figura 3.1, a análise do perfil e os fatores que podem interferir nestas escolhas serão analisadas nas seções seguintes.

4.4 Discussão

Em suma, os resultados mostrados neste capítulo sugerem que o cenário que estamos analisando não segue um padrão claro. Mesmo com um conjunto de dados limitado a um grupo de 31 alunos, a base de dados é bastante rica e com inúmeras informações a serem pesquisadas. Além disso, as preferências de colaboração não são coerentes com grafos sinalizados tradicionais, que usualmente apresentam uma configuração de triângulos próxima do balanceamento [Easley & Kleinberg, 2010]. De uma forma geral, esta caracterização sugere que a dinâmica por trás das escolhas de parceiros de trabalho pode vir a ser analisada como um ambiente social, mas os resultados mostram que não seguem os mesmos princípios. Através desta análise da estrutura da rede conseguimos abstrair algumas informações interessantes sobre as redes *online* e *offline*, mas também sugere uma análise mais aprofundada dos dados.

Capítulo 5

Características Sociais

Apesar da proficiência do estudante estar relacionada com a decisão de selecioná-lo ou não para atividades colaborativas, ela não é capaz de explicar boa parte das respostas. Além disso, foi mostrado que determinadas interações do Facebook são mais (ou menos) presentes em certos grupos de arestas valoradas, indicando que o comportamento social pode também impactar nas respostas do teste sociométrico.

De acordo com a metodologia proposta na Figura 3.1, entraremos na etapa de *Análise dos Dados*. Mais especificamente, neste capítulo abordaremos a atividade de "Identificação das Características Sociais", em que serão descritas várias características sociais que podem influenciar a decisão de escolher uma pessoa para realizar atividades colaborativas. Tais características foram extraídas diretamente dos dados coletados do Facebook. Estes dados foram modelados em um grafo não direcionado $G_F(V, E_F)$, em que o conjunto de vértices V representa o conjunto dos alunos (o mesmo conjunto de $G_S(V, E_S)$). As arestas em G_F existem entre dois alunos caso eles sejam amigos no Facebook. As características sociais foram divididas em dois grupos:

- **Atributos dos Atores**, que caracterizam os estudantes;
- **Atributos dos Relacionamentos**, que caracterizam a relação entre dois alunos. É importante ressaltar que mesmo que dois alunos não sejam amigos no Facebook, a sua relação terá um valor para o atributo.

Para os atributos dos atores, foi verificada e quantificada a influência utilizando a mesma metodologia utilizada para identificar o impacto das notas nas respostas do teste sociométrico, ou seja, foi calculado o coeficiente de correlação de *Spearman* entre o ranking dado pelo grau de entrada e saída dos alunos em G_S, agrupados pelo sinal da aresta, e o ranking dado pelos atributos analisados. Para os atributos dos relaciona-

mentos, foi verificada e quantificada a influência pelo cálculo da Função de Distribuição Cumulativa (*Cumulative Distribution Function* - CDF) dos atributos agrupados pelo sinal da aresta. A CDF descreve a probabilidade de uma variável aleatória X ser encontrada com um valor igual ou inferior a x. Esta função permite avaliar o comportamento da distribuição dos atributos quando agrupados pelo sinal da aresta de G_S [Smirnov, 1948]. Se duas CDFs (por exemplo, a CDF para arestas negativas e positivas) são significativamente distintas, então tem-se um forte indício de que o atributo é capaz de influenciar as respostas.

5.1 Atributos dos Atores

5.1.1 Popularidade

Nesta seção será investigado se os alunos populares da classe tendem a atrair um tipo específico de resposta, por exemplo, arestas positivas. Foi calculada a popularidade de um aluno de duas formas:

- $popularidade_1(i)$, como sendo a centralidade do aluno no Facebook, ou seja, no grafoG_F. Dentre as métricas para calcular a centralidade citadas na seção 2.2.2, será utilizada a *Degree Centrality*, calculada através do número de alunos em sala que o aluno i é amigo no Facebook, i.e., $popularidade_1(i) = grau(i) \in G_F$.
- $popularidade_2(i)$, como sendo o número distinto de estudantes que postaram atividades na página do Facebook do estudante i, por exemplo, comentários sobre seus links compartilhados, curtidas em suas fotos, entre outros.

Na Tabela 5.1 é apresentado o coeficiente de correlação de *Spearman* entre o ranking produzido pelas métricas de popularidade e o ranking dado pelo grau de entrada e saída dos alunos em G_S, agrupados pelo sinal da aresta. Em primeiro lugar, observa-se que tanto a métrica $popularidade_1$ quanto $popularidade_2$ apresentam correlações significativas com o grau dos alunos para vários sinais e em ambas as direções. Em ambos os casos, a correlação mais forte é vista para o $grau^0_{entrada}$, ou seja, o número de arestas neutras que chegam em um vértice. Uma vez que a correlação é negativa, isso indica que os alunos que não são populares tendem a receber mais arestas neutras, ou seja, as pessoas geralmente são indiferentes para com eles. Além disso, uma vez que a correlação $grau^0_{saida}$ é significativa para a métrica $popularidade_1$, também pode-se inferir que os alunos que não são populares também tendem a votar "INDIFERENTE". Por outro lado, observando-se a correlação para $grau^+_{entrada}$e $grau^-_{entrada}$, é curioso que

Tabela 5.1: Impacto das métricas de popularidade na escolha de parceiros para atividades colaborativas

	$popularidade_1$	*p-value*	$popularidade_2$	*p-value*
$grau^{+}_{entrada}$	**0,46**	0,009	**0,49**	0,004
$grau^{-}_{entrada}$	**0,36**	0,04	0,18	0,32
$grau^{0}_{entrada}$	**-0,74**	0,00001	**-0,66**	0,00003
$grau^{+}_{saida}$	**0,58**	0,0007	**0,37**	0,03
$grau^{-}_{saida}$	0,12	0,52	0,21	0,24
$grau^{0}_{saida}$	**-0,64**	0,0001	**-0,46**	0,008

quanto mais popular é um estudante, mais ele/ela tende a receber marcações negativas e positivas dos demais. Isso mostra que os estudantes populares são bem conhecidos pela classe, por isso é mais fácil de tomar uma decisão extrema (decisão positiva ou negativa) sobre eles. Por fim, ao analisar $grau^{+}_{saida}$, é possível inferir que os alunos populares curiosamente tendem a votar mais positivamente

5.1.2 Extroversão

Outra característica que pode ter impacto na decisão dos alunos é o seu nível de extroversão. Pessoas extrovertidas tendem a apreciar as interações humanas e ser entusiasmadas, comunicativas, assertivas e sociáveis [Eysenck, 1970]. Neste trabalho, estudantes extrovertidos são definidos com base no número de interações públicas que realizam no mural de outros alunos. Assumindo que esta medida é o quanto um indivíduo interage publicamente com outras pessoas no Facebook, esta métrica mensura quanta atenção social este indivíduo está buscando, o que representa a característica central de pessoas extrovertidas [Ashton et al., 2002]. Foram definidas duas formas de medir se um aluno é extrovertido:

- $extroversao_1(i)$, dada pelo número de interações públicas que o aluno i publicou em outras páginas no Facebook, por exemplo, comentários sobre os links dos outros estudantes, curtidas em fotos, entre outros.

- $extroversao_2(i)$, definida como sendo o número de estudantes distintos que o aluno i postou atividades públicas.

Enquanto a métrica $extroversao_1(i)$ mede o volume de interações realizadas, a $extroversao_2(i)$ enumera quantos alunos distintos receberam interações do aluno i. Na Tabela 5.2 é apresentado o coeficiente de correlação de *Spearman* entre o ranking produzido pelas métricas de extroversão e o ranking dado pelo grau de entrada e saída dos

Tabela 5.2: Impacto das métricas de extroversão na escolha de parceiros para atividades colaborativas

	$extroversao_1$	*p-value*	$extroversao_2$	*p-value*
$grau^{+}_{entrada}$	0,035	0,85	**0,516**	0,002
$grau^{-}_{entrada}$	**0,417**	0,01	0,230	0,212
$grau^{0}_{entrada}$	-0,266	0,14	**-0,726**	0,0003
$grau^{+}_{saida}$	0,093	0,61	0,386	0,031
$grau^{-}_{saida}$	0,225	0,22	0,262	0,153
$grau^{0}_{saida}$	-0,225	0,22	**-0,521**	0,002

alunos em G_S. Para a maioria dos resultados sobre a métrica $extroversao_1(i)$, pode-se notar baixas correlações e altos *p-values*. No entanto, observa-se que há uma correlação significativa entre a métrica $extroversao_1(i)$ e o $grau^{-}_{entrada}$, o que pode indicar que quanto mais um indivíduo posta no mural de outros, menos estes alunos querem trabalhar com ele. Isto sugere que os estudantes que realizam excessivamente interações públicas no Facebook podem ser também intrusivos, gerando reações negativas nos demais usuários. Outra conjectura é que quando um aluno posta um número excessivo de mensagens para outros, ele pode deixar a impressão de que desperdiça tempo demais no Facebook e, por essa razão, não seria um bom companheiro de projeto.

Em relação à métrica $extroversao_2(i)$, pode-se observar que existem correlações negativas significativas para o $grau^{0}_{entrada}$e $grau^{0}_{saida}$. Isto sugere que os estudantes que não são publicamente ativos no Facebook geralmente não são bem conhecidos pelos outros, geralmente atraindo e gerando reações neutras. Além disso, pode-se observar que os estudantes que postam comentários públicos em um grande número de páginas do Facebook tendem a atrair tanto reações positivas quanto negativas, em sua maioria positivas, uma vez que a correlação é significativamente positiva com o $grau^{+}_{entrada}$.

5.2 Atributos dos Relacionamentos

5.2.1 Força do Relacionamento

A *força do relacionamento* mede o quão próximos dois indivíduos são. Como mencionado anteriormente, existem várias formas de calcular essa força quando os dados de redes sociais *online* estão disponíveis. Neste trabalho, serão consideradas quatro métricas:

- $forcaRelacionamento_1(i, j)$, definida como sendo o número total de mensagens privadas de *chat* trocadas entre os estudantes i e j.

- $forcaRelacionamento_2(i, j)$, dada pelo número total de interações públicas que os estudantes i e j trocaram, i.e., foram contadas toda atividade pública que o estudante i postou na página do aluno j e vice-versa. Em Jones et al. [2013] também é utilizado esta métrica para mensurar a força de um relacionamento, onde é mostrado que as interações públicas são mais informativas que as mensagens privadas (utilizada na métrica anterior).

- $forcaRelacionamento_3(i, j)$, definida por Gilbert & Karahalios [2009] para medir a força de um relacionamento. Neste caso, foram utilizados os mesmos coeficientes descritos em Gilbert & Karahalios [2009], mas considerando apenas os dados disponíveis, apresentados na Tabela 5.3, onde a coluna *Impacto na Força* representa o fator de multiplicação para obtenção da força da relação.

Tabela 5.3: Dados para cálculo da $forcaRelacionamento_3(i, j)$

Variáveis	**Impacto na Força**	**Características do Facebook**
Variáveis Estruturais	0,045	Músicas em Comum Grupos em Comum Interesses em Comum Filmes em Comum Amigos em Comum
Variáveis de Intensidade	0,1970	Comentários em Fotos Comentários em Links Comentários em Atualização de Status Comentários em Álbuns Curtidas em Fotos Curtidas em Links Curtidas em Atualização de Status Mensagens de *Chat* Privadas
Variáveis de Intimidade	0,3280	Marcações

- $forcaRelacionamento_4(i, j)$, definida com sendo uma variável binária, que recebe 1 se os estudantes i e j são amigos no Facebook e 0 caso contrário.

Na Figura 5.1 é mostrado as CDFs para os três primeiros indicadores de força dos relacionamentos agrupados pelo sinal da aresta. Como apresentado, pode-se observar que a métrica $forcaRelacionamento_1$ não consegue distinguir muito bem a distribuição das três curvas, mas pode-se observar, no entanto, que a distribuição neutra tem cerca de 50% das arestas com menos de 100 conversas enquanto para a distribuição negativa e positiva esse valor é 30%. A $forcaRelacionamento_2$ é capaz de distinguir

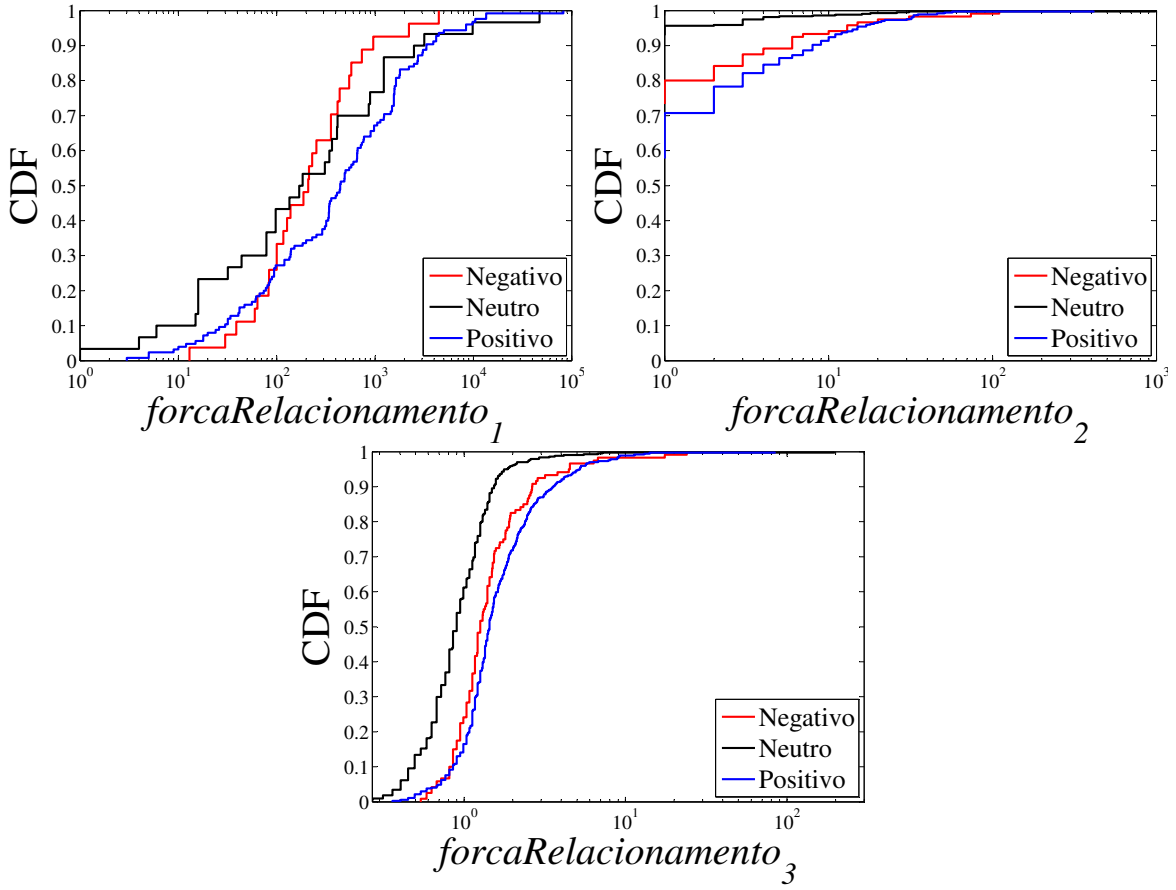


Figura 5.1: CDFs para as métricas de força do relacionamento, agrupadas pelo sinal da aresta.

melhor as três distribuições. Nota-se, por exemplo, que quase 96% da distribuição neutra tem zero interações públicas. Além disso, para as distribuições positivas e negativas, este número também é bastante elevado, com valores de 71% e 80%, respectivamente. Analisando a $forcaRelacionamento_3$, as três distribuições têm um comportamento semelhante, sendo que a neutra é razoavelmente distante das outras. Cerca de 70% das arestas neutras possuem valores de $forcaRelacionamento_3$ menores do que um valor de intensidade de 1, enquanto que para as distribuições positivas e negativas estes valores representam aproximadamente 18% das arestas. Para a métrica $forcaRelacionamento_4$, uma vez que é binária, foi calculado a proporção de arestas que têm valores de $forcaRelacionamento_4 = 1$, ou seja, são amigos no Facebook. Para as arestas negativas, a proporção é de 40%, enquanto que para as arestas positivas, a proporção é de 46%. Estes valores são maiores do que as arestas neutras, que é de 20%. Isso pode indicar que $forcaRelacionamento_4$ tem um potencial para diferenciar as arestas neutras das positivas e negativas.

5.2.2 Homofilia

A homofilia é a tendência dos indivíduos se associarem e estabelecer vínculos com outros semelhantes, i.e., indivíduos em relacionamentos homofílicos compartilham características comuns [McPherson et al., 2001]. Para investigar a homofilia no contexto deste trabalho, foram definidas três métricas distintas para medir a similaridade no Facebook:

- $similaridade_1(i, j)$, utilizada para medir a similaridade entre dois indivíduos em termos da topologia da rede no Facebook. Para isso, foi utilizado o Coeficiente de Jaccard para medir o grau de sobreposição entre os vetores de nós vizinhos de cada estudante [Symeonidis et al., 2010]. Dado dois vetores de nós r_i e r_j representando os vizinhos dos estudantes i e j em G_F, a $similaridade_1$ é definida como:

$$similaridade_1(i, j) = \frac{|r_i \cap r_j|}{|r_i \cup r_j|}$$

  onde r_i e r_j representam o conjunto de amigos que os estudantes i e j possuem no Facebook, respectivamente.

- $similaridade_2(i, j)$, definida como sendo a medida das características que dois estudantes possuem em comum no Facebook. Para isto, foram utilizadas as informações sobre os filmes que os estudantes possuem interesse e os grupos do Facebook que eles participam. Não foram utilizadas informações sobre as músicas e os interesses dos alunos porque estes dados são muito esparsos, com a cardinalidade da intersecção apresentando valores muito pequenos, o que acaba não agregando informação nesta análise. Dados dois vetores de características $filme_i$ e $grupo_i$ representando os filmes e os grupos que um estudante i possui interesse, respectivamente, a métrica $similaridade_2(i, j)$ é definida como:

$$soma(i, j) = \frac{|filme_i \cap filme_j|}{|filme_i \cup filme_j|} + \frac{|grupo_i \cap grupo_j|}{|grupo_i \cup grupo_j|}$$

$$similaridade_2(i, j) = \frac{soma(i, j)}{2}$$

  onde foi aplicado o Coeficiente de Jaccard entre estes dois vetores de cada estudante i e j, armazenado na variável $soma(i, j)$, e então calculado a média aritmética entre esses dois valores.

- $similaridade_3(i, j)$, definida como sendo o número de amigos comuns que dois alunos distintos possuem no Facebook. Estas informações não foram adicionadas à métrica $similaridade_2$ anterior porque estes dados pertencem à estrutura da rede, não sendo uma informação que os alunos compartilham através do Facebook. Essa métrica representa o número de amigos comuns, diferentemente também da métrica $similaridade_1$, que retorna um valor entre 0 e 1, i.e., o valor do coeficiente de Jaccard, considerando o conjunto de amigos que cada aluno possui.

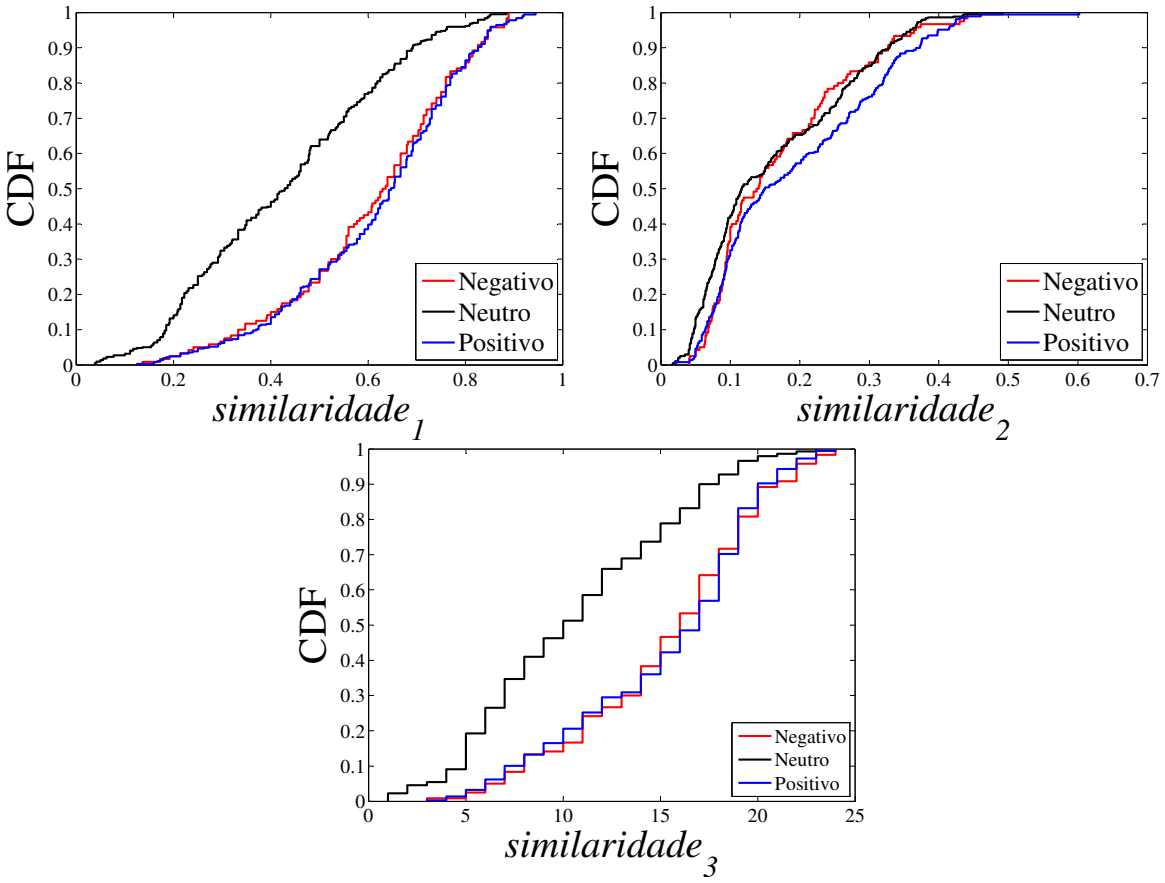


Figura 5.2: CDFs para as métricas de similaridades, agrupados pelo sinal da aresta

Na Figura 5.2 são apresentadas as CDFs para as três métricas de homofilia, agrupadas pelo sinal da aresta. Observa-se que a partir da $similaridade_1$ e da $similaridade_3$ podemos diferenciar claramente a distribuição neutra das outras duas. Isto mostra que, em nível de estrutura da rede, as relações neutras têm um comportamento diferente das relações positivas e negativas. Através destas métricas não é possível separar os relacionamentos negativos dos positivos, pois eles têm um comportamento semelhante.

Para a métrica $similaridade_2$ as distribuições das três relações têm um comportamento muito semelhante, o que torna difícil a sua separação.

5.3 Discussão

Neste capítulo foram apresentadas algumas métricas sociais, derivadas da base de dados extraída do Facebook. Algumas métricas estão relacionadas aos estudantes somente, como a popularidade e a extroversão. Outras métricas estão vinculadas ao relacionamento entre dois indivíduos, como a força da relação e a homofilia. Em resumo, alguns resultados obtidos sugerem que: alunos populares tendem a atrair mais arestas positivas; alunos pouco populares tendem a atrair e emitir mais arestas neutras; alunos extrovertidos atraem mais arestas positivas; alunos pouco extrovertidos tendem a atrair mais arestas neutras.

A avaliação isolada de cada métrica mostra o quanto ela é capaz de explicar as afinidades de trabalho. Neste contexto, surge o seguinte questionamento: "*A combinação dos fatores sociais e da proficiência conseguem explicar a afinidade de trabalho entre os estudantes? Os fatores sociais são melhores que a proficiência para explicar as afinidades de trabalho?*" Para responder estas perguntas, será apresentado no capítulo seguinte uma análise envolvendo aprendizagem de máquina para predição das afinidades de trabalho.

Capítulo 6

Classificação dos Relacionamentos

A fim de explicar o comportamento das afinidades de trabalho através das interações sociais e da proficiência, será apresentado neste capítulo uma análise utilizando algoritmos de aprendizagem de máquina para classificação dos relacionamentos. Este capítulo está relacionado com a última atividade na etapa de *Análise dos Dados*, proposta na metodologia apresentada na Figura 3.1, a atividade "Predição das Afinidades de Trabalho". A partir das características apresentadas nas seções anteriores, classificaremos os relacionamentos dos estudantes de acordo com as três classes de afinidade de trabalho: negativa, neutra e positiva. Estas classes correspondem às respostas informadas pelos estudantes no teste sociométrico. Estas análises serão realizadas através dos algoritmos apresentados na seção seguinte.

6.1 Classificadores

Os métodos utilizados neste trabalho para classificação são algoritmos implementados e disponibilizados através do software *Weka*[1]. Este é um software gratuito, registrado sobre a licença *GNU General Public License*, desenvolvido pela Universidade de Waikato, Nova Zelândia. Ele representa uma coleção de algoritmos de aprendizagem de máquina, além de métodos para análises estatísticas [Hall et al., 2009; Witten & Frank, 2005]. Serão utilizados cinco modelos de aprendizagem de máquina presentes no software *Weka*: *Multilayer Perceptron* (MLP), *Random Forest* (RF), *Naive Bayes* (NB), *k-Nearest Neighbors* (KNN) e *Support Vector Machines* (SVM). Estes algoritmos são baseados em aprendizagem supervisionada, e têm como entrada um arquivo estruturado como uma matriz e organizado da seguinte maneira: cada linha do arquivo corresponde

[1] http://www.cs.waikato.ac.nz/ml/weka/

a uma observação e cada coluna do arquivo corresponde a uma característica da observação, sendo que a última coluna representa aquilo que queremos classificar, ou seja, o rótulo da observação.

Multilayer Perceptron - MLP (Perceptron de Múltiplas Camadas) consiste de várias camadas de elementos simples (ou dois estados) de processamento sigmoidal, ou neurônios, que interagem usando conexões ponderadas [Fahlman & Hinton, 1987]. Depois de uma camada de entrada mais baixa, existe normalmente qualquer número de camadas intermédias, ou escondidos, seguidas por uma camada de saída na parte superior [Pal & Mitra, 1992; Rosenblatt, 1961]. MLP utiliza uma técnica chamada de aprendizado supervisionado de retro propagação para treinamento da rede. A aprendizagem ocorre no perceptron (neurônio) alterando pesos de conexão depois que cada partes de dados é processado, com base na quantidade de erro na saída em comparação com o resultado esperado. Este é um exemplo de aprendizagem supervisionada, e é realizada através de retro propagação [Rumelhart et al., 1985]. No software Weka, este método recebe os seguintes parâmetros de execução: (i) Taxa de aprendizado = 0,3; (ii) Momentum= 0,2; (iii) Número de Épocas = 1.000; (iv) Número de nós na camada escondida = $(atributos + classes)/2$

Random Forest - RF (Floresta Aleatória) é um método de aprendizado para classificação, que executa através da construção de um grande número de árvores de decisão no período de treinamento, produzindo a classe que representa a maioria das saídas das árvores individuais [Breiman, 2001]. Possui a característica de expandir muitas árvores de decisão (produzindo uma *floresta*) que são usadas para classificar novos objetos. Cada árvore de decisão é construída a partir de um subconjunto aleatório do conjunto de dados de treinamento [Costa et al., 2013]. Esta técnica é baseada em algoritmos genéticos. O software *Weka* traz a implementação de *Random Forest* proposta por Breiman [2001], e possui mecanismos para calcular o erro de convergência, que tende a um limite quando o número de árvores na floresta torna-se grande. O erro generalizado de uma árvore classificadora depende da força de uma árvore individual e da correlação entre elas. Os parâmetros utilizados neste modelo para experimentação foram: (i) *numFeatures* = 15 (usado na seleção randômica de atributos); (ii) *numTree* = 61 (número de árvores geradas).

Naive Bayes - NB é um classificador probabilístico simples baseado na aplicação de teorema de Bayes [Hall et al., 2009]. Também conhecidas como redes de opinião, redes causais e gráficos de dependência probabilística, são modelos gráficos para raciocínio

baseado na incerteza, onde os nós representam as variáveis (discretas ou contínuas), e as arestas representam a conexão direta entre eles [Korb & Nicholson, 2003]. Um classificador *Naive Bayes* assume que a presença ou ausência de uma característica particular está relacionada com a presença ou ausência de qualquer outro elemento. A utilização deste modelo no software *Weka* não requer nenhuma parametrização para seu funcionamento.

k-Nearest Neighbors - KNN é um dos mais simples de todos os algoritmos de aprendizado de máquina, em que o modelo de classificação é apenas aproximado localmente, exigindo cálculo das distâncias entre um padrão de teste e todos os padrões no conjunto de treinamento, sendo que a computação é adiada até o processo real de classificação [Jain et al., 2000]. É um algoritmo não parametrizado, utilizado tanto para classificação quanto para regressão. Na classificação, a saída é a associação de uma classe. Um objeto é classificado pelo voto da maioria de seus vizinhos, com o objeto que está sendo atribuído à classe mais comum entre os seus k vizinhos mais próximos. Se $k = 1$, então o objeto é simplesmente rotulado com a mesma classe do seu único vizinho mais próximo. A distância entre os objetos pode ser realizada de várias formas, como por exemplo a distância euclidiana, que é comumente utilizada quando as variáveis são contínuas [Hamming, 1950]. No *Weka*, foi aplicada a parametrização $k = 3$, i.e., foram utilizados os 3 vizinhos mais próximos do objeto em análise, sendo a distância entre os objetos obtida através da distância euclidiana.

Support Vector Machines - SVM (Máquinas de Vetores de Suporte) constituem uma técnica de aprendizado de máquinas supervisionado binária, que toma como entrada um conjunto de dados e prediz a qual classe, entre duas possíveis, a entrada faz parte. As SVMs são embasadas pela teoria de aprendizado estatístico, desenvolvida por Cortes & Vapnik [1995]. Um modelo SVM é uma representação dos padrões de formação mapeados como pontos num espaço p-dimensional, em que p é o número de características dos padrões. Os novos exemplos são então mapeados no mesmo espaço e preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles são colocados [Burges, 1998]. A implementação deste algoritmo no *Weka* utiliza a biblioteca LIBSVM, uma biblioteca para máquinas de vetores de suporte [Chang & Lin, 2011]. A parametrização utilizada foi: $gama = 0,1$ e $c = 16$.

6.2 Tratamento dos Dados

A fim de realizar a classificação das afinidades de trabalho através de dados sociais e de proficiência dos participantes, foi necessário a realização de um tratamento prévio na base de dados. Os dados do teste sociométrico, que contêm as afinidades de trabalho, possuem arestas rotuladas como positivas, negativas e neutras. Porém, existe significativa diferença entre o número de arestas de cada classe, como mostrado na Figura 6.1.

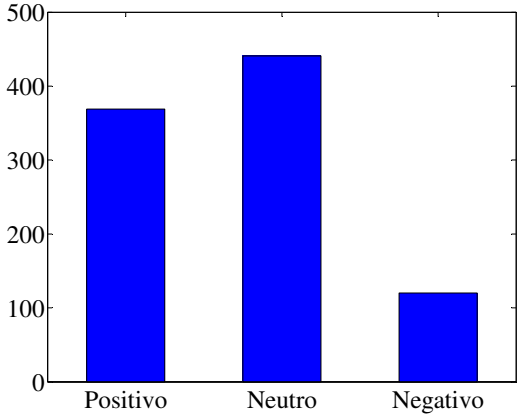


Figura 6.1: Diferença entre o número de arestas positivas, negativas e neutras.

Uma base de dados está desbalanceada se as categorias para classificação não estão igualmente representadas, i.e., existem muito mais instâncias de algumas classes do que outras [Chawla et al., 2002]. Existem uma série de soluções para o problema de classes desbalanceadas, tanto em nível de dados quanto em nível de algoritmos. Em nível de dados, estas soluções incluem diferentes formas de amostragem (*resampling*), como sobre-amostragem (*oversampling*) e a sub-amostragem (*undersampling*) [Chawla et al., 2004]. A técnica *undersampling* consiste em escolher seletivamente um subconjunto de padrões da classe majoritária, mantendo a população original da classe minoritária [Kubat & Matwin, 1997]. A abordagem da técnica *oversampling* propõe que a classe minoritária tenha uma sobre-amostragem, criando exemplos "sintéticos", como um meio eficaz de aumentar a sensibilidade de um classificador para a classe minoritária [Chawla et al., 2002]. Neste trabalho serão empregadas, além da análise utilizando a base de dados desbalanceada, duas técnicas para tratamento do problema de desbalanceamento:

- *Random Undersampling*: *undersampling* aleatório é uma abordagem simples para amostragem. Padrões da classe majoritária no conjunto de treinamento são eli-

minados aleatoriamente, até que a proporção entre a classe minoritária e a majoritária esteja em um nível desejado. Apesar da sua simplicidade, o *undersampling* aleatório tem sido empiricamente demonstrado ser um dos métodos mais eficazes de amostragem [Liu, 2004].

- *SMOTE Oversampling*: o algoritmo SMOTE (***S**ynthetic **M**inority **O**ver-sampling **TE**chnique* - tradução livre "Técnica de Sobre-Amostragem Minoritária Sintética") é simples, mas eficaz, superando *oversampling* aleatório em vários problemas de baixa dimensionalidade. Em particular, SMOTE é uma das poucas técnicas de amostragem que tem consistentemente melhor resultado do que *undersampling* aleatório e *oversampling* aleatório em termos de desempenho [Liu, 2004]. No algoritmo SMOTE, a classe minoritária é aumentada tomando-se cada amostra desta classe e introduzindo exemplos sintéticos ao longo do conjunto de dados. Dependendo de quanto se deseja aumentar o número de padrões da classe minoritária, k vizinhos de cada amostra são selecionados. A implementação básica comumente utiliza 5 vizinhos mais próximos. Por exemplo, se é necessário aumentar a amostragem da classe minoritária em 200%, somente 2 dos 5 vizinhos mais próximos são escolhidos e uma amostra sintética é gerada [Chawla et al., 2002]. Um padrão sintético é gerado da seguinte forma: encontra-se a diferença entre o vetor de características sob análise e seus vizinhos mais próximos; multiplica-se essa diferença por um valor aleatório entre 0 e 1; e adiciona-se esse valor ao vetor de características sintético em questão [Chawla et al., 2002].

Undersampling e *oversampling* são dois métodos utilizados que amenizam o problema do desbalanceamento dos dados através da sub-amostragem da classe majoritária e repetindo instâncias da classe minoritária, respectivamente [Drummond et al., 2003]. O nível de desbalanceamento é reduzido em ambos métodos, com o intuito de que uma base de dados mais balanceada pode gerar melhores resultados. Enquanto o *undersampling* requer tempo de treinamento mais curto, porém com o custo de ignorar dados potencialmente úteis, o *oversampling* aumenta o tamanho do conjunto de treinamento e, portanto, requer mais tempo de treinamento [Liu et al., 2009]. A literatura não define qual das duas técnicas é "melhor" para ser aplicada, mas mostra que ambas, quando aplicadas juntamente com algoritmos de classificação apropriados, empregadas em diferentes contextos, produzem resultados satisfatórios [Liu, 2004; Chawla et al., 2004; Liu et al., 2009; Chawla et al., 2002]. Como não é objetivo deste trabalho esgotar as tentativas de melhoria contínua dos resultados de classificação, serão utilizadas as duas técnicas de amostragem supracitadas, assim como a base de dados desbalanceada, a fim de mostrar o potencial existente na aplicação de cada uma delas.

Ao todo, temos 930 tuplas (ou relacionamentos) para construção da matriz de entrada de dados. Os atributos relacionados ao ator foram processados da seguinte forma: para um relacionamento $(i \rightarrow j)$, o estudante *origem* i respondeu sobre sua afinidade de trabalho para o estudante *destino* j. Assim, foram colocados os valores destas características tanto para o estudante *origem* quanto para o estudante *destino* como atributos na matriz de entrada de dados. As características pertencentes ao relacionamento entre os alunos i e j (*origem* e *destino*) foram colocadas normalmente como colunas individuais na matriz de entrada. A saída para os modelos de classificação são os rótulos dos relacionamentos $(i \rightarrow j)$ resultantes do teste sociométrico. Além disso, os valores de entrada para os modelos foram normalizados utilizando a seguinte fórmula:

$$norm_i = \frac{x_i - min}{max - min}$$

onde min é o menor valor em C_k, sendo C_k o conjunto de todos os valores que característica k possui, max é o maior valor em C_k e x_i é o valor do elemento em análise. O arquivo aplicado aos modelos seguem o formato mostrado na Tabela 6.1.

6.3 Classificação dos Relacionamentos

Os resultados obtidos pelos métodos de classificação são apresentados nesta seção. Seguindo a metodologia descrita previamente, os experimentos são divididos em 3 ambientes, que possuem arquivos de entrada com conjuntos de características distintos:

1. **Fatores Sociais** - serão utilizadas somente características derivadas das interações sociais coletadas do Facebook, tais como Força do Relacionamento, Homofilia, Extroversão e Popularidade.

2. **Proficiência** - serão utilizadas somente características descritivas da proficiência de cada aluno, ou seja, as notas.

3. **Fatores Sociais e Proficiência** - serão utilizadas todas as características, tanto as sociais quanto as descritivas da proficiência.

O processo de classificação será realizado a partir da técnica *k-fold cross-validation* (validação cruzada de k amostras). No *k-fold cross validation*, às vezes chamado de *rotation estimation* (estimativa de rotação), o conjunto de dados D é dividido aleatoriamente em k subconjuntos mutuamente exclusivos $D_1, D_2, ..., D_k$ de aproximadamente igual tamanho, conhecidos como *folds* [Geisser, 1993]. O classificador é treinado e

Tabela 6.1: Resumo das características utilizadas nos modelos de classificação

Característica	**Categoria**	**Descrição**
nota (origem)	Proficiência	Nota do estudante de origem
nota (destino)	Proficiência	Nota do estudante de destino
$popularidade_1$ *(origem)*	Popularidade	Número de amigos que o estudante de origem possui no Facebook
$popularidade_1$ *(destino)*	Popularidade	Número de amigos que o estudante de destino possui no Facebook
$popularidade_2$ *(origem)*	Popularidade	Número de estudantes distintos que postaram atividades no mural do estudante de origem
$popularidade_2$ *(destino)*	Popularidade	Número de estudantes distintos que postaram atividades no mural do estudante de destino
$extroversao_1$ *(origem)*	Extroversão	Interações públicas que o estudante publicou nos murais dos outros
$extroversao_1$ *(destino)*	Extroversão	Interações públicas que o estudante publicou nos murais dos outros
$extroversao_2$ *(origem)*	Extroversão	Número de estudantes distintos que o estudante de origem postou atividades no mural
$extroversao_2$ *(destino)*	Extroversão	Número de estudantes distintos que o estudante de destino postou atividades no mural
$forcaRelacionamento_1$	Força dos Relacionamentos	Número total de mensagens de chat privado trocadas entre dois estudantes
$forcaRelacionamento_2$	Força dos Relacionamentos	Número total de interações públicas trocadas entre dois estudantes
$forcaRelacionamento_3$	Força dos Relacionamentos	Métrica força de um relacionamento proposta em Gilbert & Karahalios [2009]
$forcaRelacionamento_4$	Força dos Relacionamentos	1 se dois estudantes são amigos no Facebook, 0 caso contrário
$similaridade_1$	Similaridade	Similaridade entre o conjunto de amigos de dois estudantes
$similaridade_2$	Similaridade	Similaridade entre o conjunto de filmes e grupos de dois estudantes
$similaridade_3$	Similaridade	Número de amigos em comum que dois amigos possuem no Facebook

testado k vezes: cada vez $t \in 1, 2, ..., k$ é treinado em $D - D_t$ e testado em D_t. As técnicas de *undersampling* e *oversampling* são aplicadas no conjunto $D - D_t$ durante o período de treinamento. A acurácia da validação cruzada pode ser estimada através do número total de classificações corretas, dividido pelo número de instâncias no conjunto de dados [Kohavi, 1995]. Nos experimentos realizados, a base de dados original é particionada em 5 subconjuntos. Ao todo, o *5-fold cross validation* foi repetido 50 vezes, com diferentes sementes utilizadas para embaralhar a base de dados original. Assim, os resultados apresentados são as médias aritméticas destas 50 execuções. Além disso, são apresentadas as oscilações para o intervalo de confiança de 95% do conjunto de resultados obtidos.

Para avaliar os resultados dos modelos de classificação serão utilizadas as seguintes métricas [Baeza-Yates et al., 1999]:

- **Matriz de Confusão**: é uma tabela específica que proporciona uma medida efetiva de um modelo de classificação, ao mostrar o número de classificações corretas versus as classificações preditas para cada classe. Cada coluna da matriz representa as instâncias de uma classe que foi prevista, enquanto cada linha representa as instâncias reais da classe [Townsend, 1971]. Um exemplo de uma matriz de confusão, relacionada com o contexto deste trabalho, pode ser visto na tabela abaixo.

Tabela 6.2: Matriz de confusão para classificação das afinidades de trabalho

		Previsto		
		Negativo	**Neutro**	**Positivo**
	Negativo	**a**	b	c
Real	**Neutro**	d	**e**	f
	Positivo	g	h	**i**

Os valores ***a***, ***e*** e ***i*** mostrados na Tabela 6.2 representam os dados reais que foram classificados corretamente para as classes negativa, neutra e positiva, respectivamente.

- **Precisão**: proporção de previsões feitas corretamente, em relação ao número total de previsões feitas para uma classe. Esta métrica é descrita pela fórmula:

$$precisao = \frac{VP}{VP + FP}$$

onde FP (Falso Positivo) é o conjunto de casos da classe α incorretamente classificados como sendo desta classe, e VP (Verdadeiro Positivo) é o conjunto de casos da classe α corretamente classificados.

- **Revocação**: avalia o número de previsões feitas corretamente pelo número ideal de previsões que deveriam ter sido feitas. Esta métrica é descrita pela fórmula:

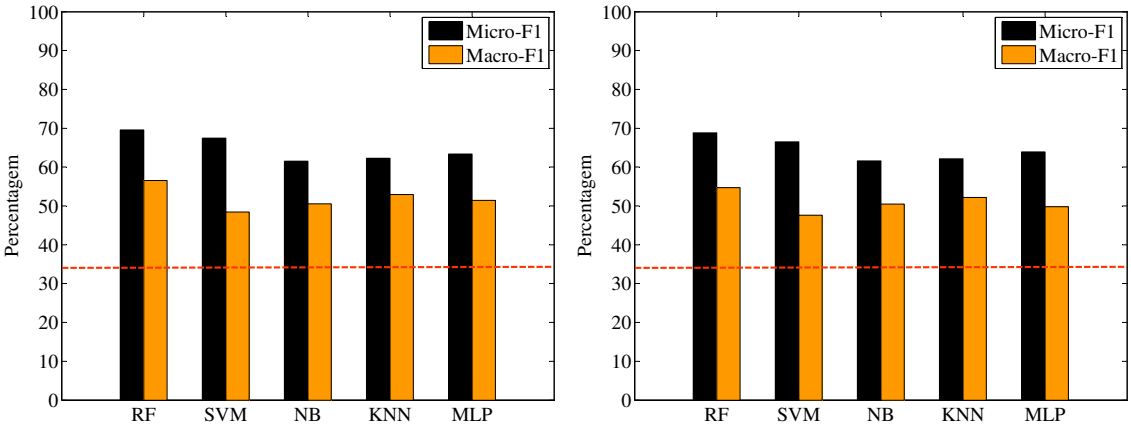
$$revocacao = \frac{VP}{VP + FN}$$

onde FN (Falso Negativo) corresponde ao conjunto de casos relevantes da classe α contidos na base de dados, mas incorretamente classificados.

Um exemplo do poder de avaliação das métricas precisão e revocação pode ser definido como: revocação é 100% quando todos os casos relevantes da classe α são mostrados; precisão é 100% quando todos os casos classificados como sendo da classe α são relevantes.

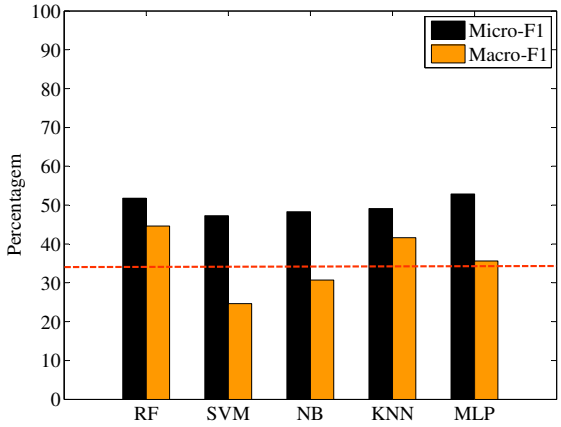
- **Micro-F1** e **Macro-F1**: Primeiramente, definimos a métrica F1 (*F-Score* ou *F-Measure*) como sendo a média harmônica entre a precisão e o revocação, definida por $F1 = 2pr/(p + r)$, onde p é precisão e r é revocação [Yang, 1999]. Assim, a Micro-F1 é calculada computando-se primeiramente os valores de precisão e revocação globais, e em seguida, calcular F1. Valores da métrica Macro-F1 são computadas primeiramente obtendo-se os resultados de F1 para cada classe de forma isolada, como exemplificado acima na Tabela 6.2, e em seguida uma média destes resultados [Benevenuto et al., 2010]. A Macro-F1 considera cada classe sendo igualmente importante, enquanto a Micro-F1 está relacionada somente com a quantidade de padrões classificados corretamente pelo modelo, independente da proporção existente em cada uma delas.

As Figuras 6.2, 6.3 e 6.4 apresentam uma comparação entre as métricas Micro-F1 e Macro-F1 para os experimentos realizados utilizando a base de dados desbalanceada e as técnicas de *undersampling* e *oversampling*, respectivamente. Cada figura apresenta os resultados dos classificadores utilizando como entrada: somente dados sociais; somente notas dos alunos; e utilizando todos os dados (sociais e notas). Os valores para a Micro-F1 são equivalentes à acurácia de cada classificador, enquanto a Macro-F1 é mais sensível à taxa de acerto de cada classe. Nas três figuras podemos ver que, em geral, os resultados são maiores que um algoritmo aleatório, que sorteia aleatoriamente qual o rótulo de cada aresta, representado pela linha tracejada nos gráficos.



(a) Sociais e Notas

(b) Sociais

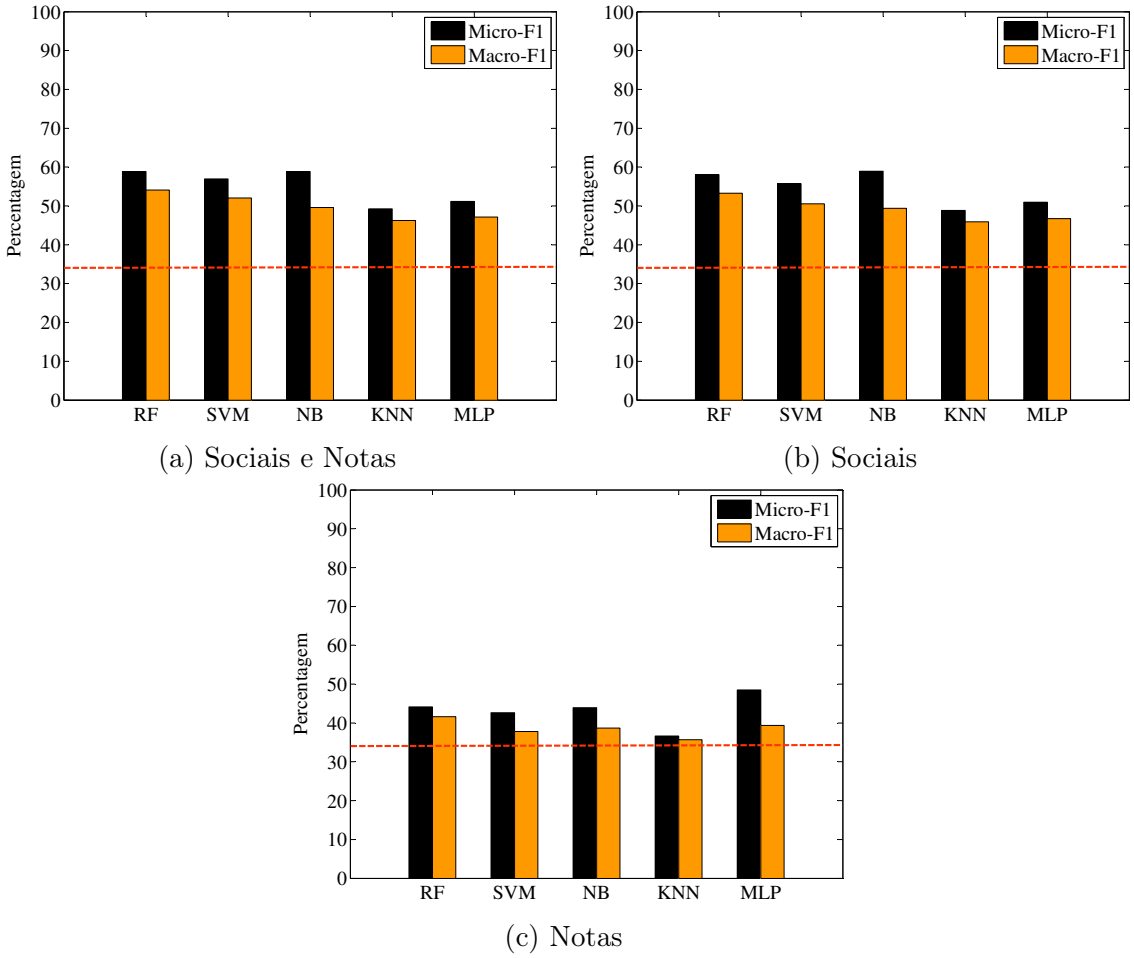


(c) Notas

Figura 6.2: Avaliação dos classificadores para a base de dados desbalanceada

Tabela 6.3: Micro e Macro-F1 para a base de dados desbalanceada

Método	**Dados Utilizados**	**Micro-F1**	**Macro-F1**
Random Forest	Sociais + Notas	$0,695 \pm 0,002$	$0,565 \pm 0,004$
	Notas	$0,517 \pm 0,004$	$0,445 \pm 0,005$
	Sociais	$0,688 \pm 0,002$	$0,546 \pm 0,003$
SVM	Sociais + Notas	$0,674 \pm 0,001$	$0,484 \pm 0,001$
	Notas	$0,472 \pm 0,001$	$0,246 \pm 0,002$
	Sociais	$0,665 \pm 0,001$	$0,475 \pm 0,001$
Naive Bayes	Sociais + Notas	$0,615 \pm 0,002$	$0,505 \pm 0,003$
	Notas	$0,482 \pm 0,002$	$0,306 \pm 0,001$
	Sociais	$0,615 \pm 0,003$	$0,505 \pm 0,002$
KNN	Sociais + Notas	$0,622 \pm 0,002$	$0,529 \pm 0,003$
	Notas	$0,491 \pm 0,003$	$0,415 \pm 0,004$
	Sociais	$0,621 \pm 0,003$	$0,521 \pm 0,003$
Multilayer Perceptron	Sociais + Notas	$0,633 \pm 0,003$	$0,514 \pm 0,004$
	Notas	$0,528 \pm 0,002$	$0,355 \pm 0,002$
	Sociais	$0,638 \pm 0,004$	$0,497 \pm 0,004$



(a) Sociais e Notas

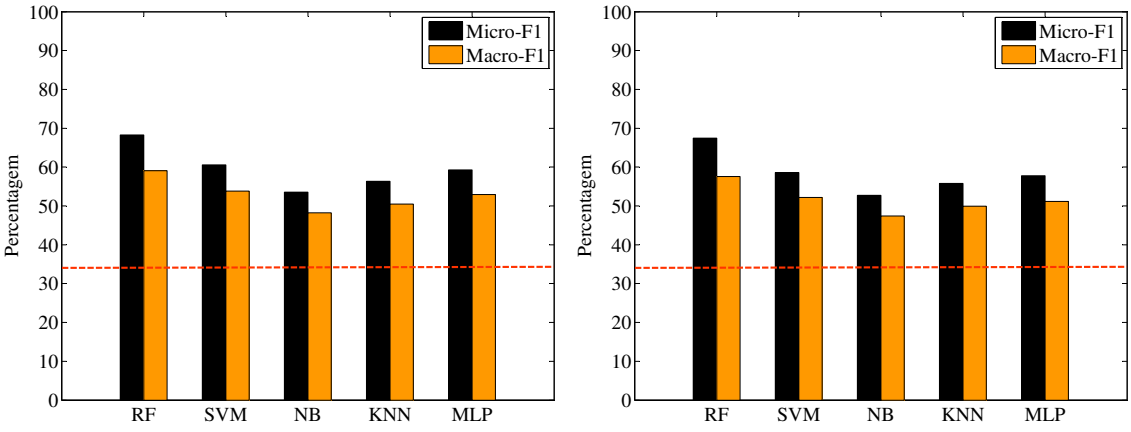
(b) Sociais

(c) Notas

Figura 6.3: Avaliação dos classificadores para a base de dados com *undersampling*

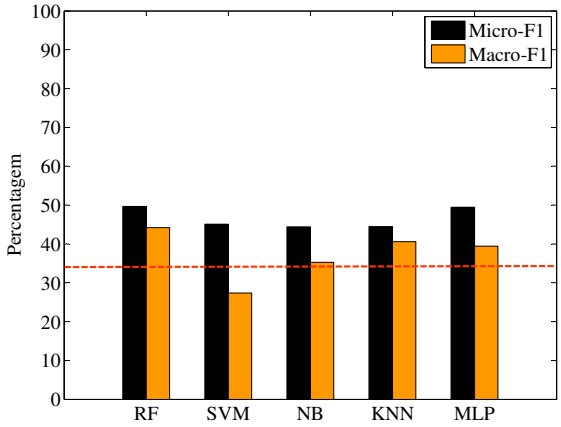
Tabela 6.4: Micro e Macro-F1 para a base de dados *undersampled*

Método	**Dados Utilizados**	**Micro-F1**	**Macro-F1**
Random Forest	Sociais + Notas	$0,588 \pm 0,005$	$0,54 \pm 0,005$
	Notas	$0,441 \pm 0,004$	$0,415 \pm 0,004$
	Sociais	$0,58 \pm 0,004$	$0,532 \pm 0,004$
SVM	Sociais + Notas	$0,569 \pm 0,004$	$0,52 \pm 0,004$
	Notas	$0,426 \pm 0,007$	$0,377 \pm 0,007$
	Sociais	$0,557 \pm 0,005$	$0,505 \pm 0,004$
Naive Bayes	Sociais + Notas	$0,588 \pm 0,008$	$0,496 \pm 0,006$
	Notas	$0,439 \pm 0,006$	$0,386 \pm 0,005$
	Sociais	$0,589 \pm 0,008$	$0,494 \pm 0,006$
KNN	Sociais + Notas	$0,492 \pm 0,004$	$0,462 \pm 0,004$
	Notas	$0,366 \pm 0,006$	$0,356 \pm 0,006$
	Sociais	$0,488 \pm 0,005$	$0,459 \pm 0,004$
Multilayer Perceptron	Sociais + Notas	$0,511 \pm 0,005$	$0,471 \pm 0,004$
	Notas	$0,485 \pm 0,006$	$0,393 \pm 0,008$
	Sociais	$0,51 \pm 0,005$	$0,467 \pm 0,004$



(a) Sociais e Notas

(b) Sociais



(c) Notas

Figura 6.4: Avaliação dos classificadores para a base de dados com *oversampling*

Tabela 6.5: Micro e Macro-F1 para a base de dados *oversampled*

Método	**Dados Utilizados**	**Micro-F1**	**Macro-F1**
Random Forest	Sociais + Notas	$0,682 \pm 0,003$	$0,59 \pm 0,005$
	Notas	$0,496 \pm 0,004$	$0,441 \pm 0,005$
	Sociais	$0,674 \pm 0,003$	$0,575 \pm 0,004$
SVM	Sociais + Notas	$0,605 \pm 0,003$	$0,538 \pm 0,003$
	Notas	$0,451 \pm 0,003$	$0,273 \pm 0,004$
	Sociais	$0,585 \pm 0,002$	$0,521 \pm 0,003$
Naive Bayes	Sociais + Notas	$0,535 \pm 0,003$	$0,482 \pm 0,002$
	Notas	$0,443 \pm 0,002$	$0,352 \pm 0,003$
	Sociais	$0,527 \pm 0,002$	$0,474 \pm 0,002$
KNN	Sociais + Notas	$0,563 \pm 0,003$	$0,504 \pm 0,003$
	Notas	$0,444 \pm 0,003$	$0,406 \pm 0,003$
	Sociais	$0,558 \pm 0,003$	$0,499 \pm 0,003$
Multilayer Perceptron	Sociais + Notas	$0,593 \pm 0,006$	$0,529 \pm 0,006$
	Notas	$0,494 \pm 0,008$	$0,394 \pm 0,005$
	Sociais	$0,577 \pm 0,006$	$0,511 \pm 0,005$

Na Figura 6.2, a diferença entre as métricas Micro-F1 e Macro-F1 nos indica que o classificador é capaz de indicar melhor algumas classes do que outras, i.e., o classificador é capaz de acertar mais relacionamentos neutros do que os demais, por exemplo. Isso acontece pois a base de dados possui mais arestas neutras do que positivas e negativas, ocasionando o seguinte problema: mesmo que o classificador indicasse que todas as arestas são neutras, o valor para Micro-F1 seria alto, pois acertaria todo o conjunto das arestas neutras, o que representa cerca de $47,4\%$ da base de dados. Já com a métrica Macro-F1, que é mais sensível à taxa de acerto de cada classe, os valores são menores que a Micro-F1. Essa diferença entre as duas métricas diminui nos resultados das Figuras 6.3 e 6.4, pois com a aplicação das técnicas *oversampling* e *undersampling*, a base de dados tende a ficar mais equilibrada. Nestas análises, os dados sociais geralmente possuem melhores resultados que somente dados da proficiência para explicar as afinidades de trabalho. Através das Figuras 6.3 e 6.4 podemos observar também que, de uma forma geral, os resultados dos classificadores utilizando a técnica *oversampling* possuem taxas de acerto maiores do que utilizando a técnica *undersampling*. Essa comparação nos mostra o potencial existente na aplicação de cada uma das técnicas.

As Tabelas 6.3, 6.4 e 6.5 apresentam os valores das métricas Micro-F1 e Macro-F1 para os resultados obtidos pelos classificadores, juntamente com a variação, utilizando-se um intervalo de confiança de 95%, para os experimentos realizados com a base de dados desbalanceada e as técnicas de *undersampling* e *oversampling*, respectivamente. As tabelas correspondem aos resultados apresentados nas Figuras 6.2, 6.3 e 6.4, respectivamente. Em geral, podemos ver que os valores de Micro-F1 e Macro-F1 não variam mais que 1% em relação à media, para o conjunto de experimentos realizados.

As Tabelas 6.6, 6.7 e 6.8 apresentam os valores das métricas Precisão e Revocação para os resultados obtidos pelos classificadores, juntamente com a variação, utilizando-se um intervalo de confiança de 95%, para os experimentos realizados com a base de dados desbalanceada e as técnicas de *undersampling* e *oversampling*, respectivamente. Pode-se observar que os valores não variam em mais que $0,5\%$ para os experimentos realizados. Podemos observar que obtivemos as maiores taxas de Precisão com o algoritmo RF, que possui valores próximos para as três configurações de dados (desbalanceados, *undersampled* e *oversampled*) utilizando dados Sociais e Notas.

A Tabela 6.9 apresenta a matriz de confusão para o classificador RF, que possui melhores resultados para a base de dados que utiliza dados sociais e proficiência, sem aplicação de nenhuma técnica para desbalanceamento. A Tabela 6.10 apresenta a matriz de confusão para o classificador RF, que possui melhores resultados para a base de dados utilizando dados sociais e proficiência, com aplicação da técnica *undersampling*, e a Tabela 6.11 apresenta a matriz de confusão para o classificador RF, que possui

Tabela 6.6: Resultados das métricas para classificação utilizando a base de dados desbalanceada

Método	**Dados Utilizados**	**Precisão**	**Revocação**
Random Forest	Sociais + Notas	$0,672 \pm 0,003$	$0,696 \pm 0,003$
	Notas	$0,513 \pm 0,005$	$0,518 \pm 0,005$
	Sociais	$0,658 \pm 0,003$	$0,688 \pm 0,003$
SVM	Sociais + Notas	$0,625 \pm 0,013$	$0,675 \pm 0,001$
	Notas	$0,403 \pm 0,008$	$0,473 \pm 0,002$
	Sociais	$0,608 \pm 0,011$	$0,665 \pm 0,002$
Naive Bayes	Sociais + Notas	$0,602 \pm 0,003$	$0,615 \pm 0,003$
	Notas	$0,42 \pm 0,003$	$0,483 \pm 0,002$
	Sociais	$0,601 \pm 0,002$	$0,616 \pm 0,003$
KNN	Sociais + Notas	$0,63 \pm 0,003$	$0,623 \pm 0,003$
	Notas	$0,488 \pm 0,004$	$0,491 \pm 0,004$
	Sociais	$0,629 \pm 0,003$	$0,621 \pm 0,003$
Multilayer Perceptron	Sociais + Notas	$0,612 \pm 0,004$	$0,634 \pm 0,004$
	Notas	$0,466 \pm 0,004$	$0,529 \pm 0,002$
	Sociais	$0,606 \pm 0,004$	$0,639 \pm 0,005$

Tabela 6.7: Resultados das métricas para classificação utilizando a base de dados *undersampled*

Método	**Dados Utilizados**	**Precisão**	**Revocação**
Random Forest	Sociais + Notas	$0,657 \pm 0,005$	$0,588 \pm 0,005$
	Notas	$0,499 \pm 0,005$	$0,442 \pm 0,004$
	Sociais	$0,651 \pm 0,004$	$0,581 \pm 0,005$
SVM	Sociais + Notas	$0,651 \pm 0,004$	$0,569 \pm 0,005$
	Notas	$0,452 \pm 0,006$	$0,427 \pm 0,007$
	Sociais	$0,645 \pm 0,004$	$0,557 \pm 0,005$
Naive Bayes	Sociais + Notas	$0,601 \pm 0,004$	$0,588 \pm 0,008$
	Notas	$0,467 \pm 0,005$	$0,439 \pm 0,006$
	Sociais	$0,599 \pm 0,004$	$0,589 \pm 0,009$
KNN	Sociais + Notas	$0,625 \pm 0,005$	$0,493 \pm 0,005$
	Notas	$0,479 \pm 0,008$	$0,366 \pm 0,007$
	Sociais	$0,625 \pm 0,005$	$0,488 \pm 0,005$
Multilayer Perceptron	Sociais + Notas	$0,601 \pm 0,005$	$0,511 \pm 0,005$
	Notas	$0,48 \pm 0,007$	$0,485 \pm 0,006$
	Sociais	$0,6 \pm 0,006$	$0,51 \pm 0,005$

melhores resultados para a base de dados que utiliza dados sociais e proficiência, com aplicação da técnica *oversampling*. Os valores em negrito representam os padrões que foram classificados corretamente. As três tabelas mostram que, tanto para a base de dados desbalanceada quanto para as técnicas *oversampling* e *undersampling*, as taxas de acerto das arestas neutras são as maiores, seguidas pelas taxas de acerto das arestas positivas e negativas. Com a técnica *undersampling* é obtido o maior número de acertos para arestas negativas, aproximadamente 49% do conjunto é classificado corretamente.

Tabela 6.8: Resultados das métricas para classificação utilizando a base de dados *oversampled*

Método	Dados Utilizados	Precisão	Revocação
Random Forest	Sociais + Notas	$0,675 \pm 0,004$	$0,683 \pm 0,004$
	Notas	$0,504 \pm 0,004$	$0,497 \pm 0,004$
	Sociais	$0,667 \pm 0,003$	$0,674 \pm 0,003$
SVM	Sociais + Notas	$0,655 \pm 0,002$	$0,606 \pm 0,004$
	Notas	$0,421 \pm 0,009$	$0,451 \pm 0,003$
	Sociais	$0,646 \pm 0,002$	$0,586 \pm 0,003$
Naive Bayes	Sociais + Notas	$0,627 \pm 0,003$	$0,536 \pm 0,003$
	Notas	$0,457 \pm 0,004$	$0,444 \pm 0,003$
	Sociais	$0,627 \pm 0,003$	$0,527 \pm 0,003$
KNN	Sociais + Notas	$0,633 \pm 0,003$	$0,563 \pm 0,004$
	Notas	$0,489 \pm 0,004$	$0,445 \pm 0,004$
	Sociais	$0,631 \pm 0,003$	$0,558 \pm 0,003$
Multilayer Perceptron	Sociais + Notas	$0,636 \pm 0,005$	$0,593 \pm 0,006$
	Notas	$0,487 \pm 0,005$	$0,495 \pm 0,008$
	Sociais	$0,624 \pm 0,005$	$0,578 \pm 0,006$

Tabela 6.9: Matriz de confusão para RF sem balanceamento para dados sociais e proficiência

		Previsto		
		Negativo	**Neutro**	**Positivo**
Real	**Negativo**	**14,78%**	39,94%	45,28%
	Neutro	2,75%	**79,76%**	17,49%
	Positivo	3,03%	21,78%	**75,19%**

Tabela 6.10: Matriz de confusão para RF utilizando *undersampling* para dados sociais e proficiência

		Previsto		
		Negativo	**Neutro**	**Positivo**
Real	**Negativo**	**49,03%**	23,81%	27,17%
	Neutro	20,32%	**62,83%**	16,84%
	Positivo	27,34%	15,42%	**57,24%**

Tabela 6.11: Matriz de confusão para RF utilizando *oversampling* para dados sociais e proficiência

		Previsto		
		Negativo	**Neutro**	**Positivo**
Real	**Negativo**	**28,89%**	35,86%	35,25%
	Neutro	6,81%	**77,28%**	15,91%
	Positivo	10,42%	19,30%	**70,28%**

6.4 Os Fatores Mais Importantes

Depois de ter analisado as características derivadas do Facebook separadamente, e também como elas explicam coletivamente a formação de links relacionados à afinidades de trabalho, será realizada uma análise de quais são as características mais preditivas para explicar estes relacionamentos de trabalho. Neste contexto, foram utilizadas dois algoritmos: o *Information Gain* e χ^2 (Chi-Squared) [Yang & Pedersen, 1997]. Ambos são métodos de seleção de características, amplamente utilizados para identificar o subconjunto dos recursos que são mais preditivos em uma classificação.

Nesta análise, foram utilizados os dados apresentados na Seção 6.2. O ranqueamento foi realizado utilizando a classificação de todas as arestas relacionadas com o teste sociométrico, ou seja, arestas positivas, negativas e neutras. Assim, este resultado mostra quais os fatores mais relevantes para explicar a formação de links positivos, negativos e neutros. Os resultados obtidos são apresentados na Tabela 6.12.

Tabela 6.12: Ranking das características mais importantes, apresentadas pelo ranking gerado pelo IG (*Information Gain*) e pelo χ^2 (*Chi-Squared*)

Descrição	**Rank IG**	**Valor IG**	**Rank χ^2**	**Valor χ^2**
$forcaRelacionamento_3$	1	0.194	1	226.01
$forcaRelacionamento_4$	2	0.181	2	220.00
$similaridade_1$	3	0.151	3	184.51
$similaridade_3$	4	0.150	4	181.10
$forcaRelacionamento_2$	5	0.098	5	120.01
$popularidade_1(origem)$	6	0.084	8	100.93
$extroversao_1(destino)$	7	0.084	6	116.81
$forcaRelacionamento_1$	8	0.083	9	98.95
$popularidade_2(destino)$	9	0.079	10	96.17
$nota(destino)$	10	0.075	7	104.31
$extroversao_2(destino)$	11	0.073	11	91.05
$extroversao_2(origem)$	12	0.069	13	82.56
$nota(origem)$	13	0.065	12	89.44
$popularidade_1(destino)$	14	0.048	14	62.14
$extroversao_1(origem)$	15	0.040	15	46.70
$popularidade_2(origem)$	16	0.035	16	44.39
$similaridade_2$	17	0.022	17	27.30

Os dois métodos ranqueiam a relevância das características de forma bastante similar. O coeficiente de correlação de *Spearman* entre os rankings gerados pelos dois métodos é $0,9810$, com *p-value* $4,2894 * 10^{-12}$.

O resultado mais importante desses ranqueamentos é que as notas (ou proficiência) dos alunos não são tão relevantes, i.e., as notas não ocupam posições superiores à

sétima e à décima nas listas de relevância de características. Por outro lado, a característica mais preditiva é a métrica para a força do relacionamento, que foi proposta e amplamente avaliada na literatura [Gilbert & Karahalios, 2009], o que indica a validade externa dos nossos resultados e mostra que a força de um relacionamento influencia significativamente na escolha dos parceiros para atividades colaborativas. Além disso, as características sociais, tais como similaridade entre os usuários, são mais preditivas do que notas, corroborando com os resultados mostrados na Seção 6.3.

6.5 Discussão

Foram utilizados 5 algoritmos de aprendizado de máquina distintos para resolver o problema de classificação das afinidades de trabalho. Como a base de dados é desbalanceada (120 respostas negativas, 369 positivas e 441 neutras no teste sociométrico), foram aplicadas duas técnicas para amenizar os efeitos deste desnível entre o número de padrões de cada classe: *undersampling* e *oversampling*. O *undersampling* utilizado consiste na obtenção de um subconjunto da classe majoritária, através da seleção aleatória, de modo que a diferença entre o número de padrões das classes seja minimizado. O *oversampling* é realizado através da técnica SMOTE, e consiste na geração de dados sintéticos para aumentar o volume de padrões da classe minoritária, também visando a minimização da diferença entre o tamanho das classes. Nestes três contextos, foram realizados experimentos utilizando diferentes conjuntos de dados como entrada dos algoritmos: utilizando somente dados sociais; utilizando somente as notas dos alunos; e utilizando as notas e os dados sociais.

O desempenho de cada algoritmo foi computado, e as análises mostram que características sociais são mais relevantes que a proficiência para explicar a afinidade de trabalho entre os alunos. Apesar da proficiência ter impacto nas escolhas para atividades colaborativas, ela não é a mais importante. Fatores sociais impactam mais nas escolhas do que somente a proficiência para realizar uma atividade de colaboração. Isso mostra que relações sociais bem estruturadas podem contribuir mais em um grupo de trabalho do que simplesmente a capacidade de execução individual de cada membro. Por exemplo, o fato de um estudante querer trabalhar com aqueles que ele possui um "forte" relacionamento é uma evidência que pode sustentar essa afirmação.

Capítulo 7

Considerações Finais

7.1 Conclusões

Neste trabalho foi proposta uma metodologia para predição de afinidades e formação de grupos de trabalho a partir de interações realizadas em redes sociais *online*. A metodologia proposta consiste de quatro passos bem definidos. Primeiro deve-se extrair de um grupo de pessoas, o qual será insumo para realização do experimento, como é a relação de trabalho entre eles. Em seguida, deve-se coletar informações sobre as interações *online* deste grupo de pessoa, e por fim, estabelecer um mapeamento entre as interações *online* e as relações de trabalho deste grupo. Este mapeamento é feito através de características sociais que são extraídas dos dados coletados. Para isso, foi realizado um experimento em um cenário muito particular. Primeiro, foi aplicado um teste sociométrico em uma classe de alunos de graduação, em que os indivíduos foram questionados se gostariam de trabalhar com todos os outros alunos da classe. Em seguida, a partir de um aplicativo que desenvolvemos para o Facebook, foram coletados dados que contém uma série de características sociais sobre seus perfis e suas interações. Além disso, foram coletados as notas destes alunos, como forma de mensurar sua proficiência em desempenhar atividades de trabalho. O mapeamento entre os dados sociais e as relações de trabalho foi realizado a partir da extração de caraterísticas sociais do conjunto de dados coletados, e também da classificação das relações de trabalhos a partir destas características sociais.

Mesmo com um conjunto de dados limitado a um grupo de 31 alunos, a base de dados é rica e com inúmeras informações a serem pesquisadas, pois este número chega a 930 relacionamentos. O confronto entre dados *online* e *offline* nos permite avaliar diretamente como o comportamento social *online* interfere em relacionamentos de trabalho. Comparado às notas da classe, mostramos que as características sociais

propostas derivadas do Facebook são mais preditivas sobre quem os estudantes desejam trabalhar. A mais importante dessas características é a aproximação para a força dos relacionamentos, proposta em Gilbert & Karahalios [2009], sugerindo a importância da construção do capital social na formação de uma colaboração.

7.2 Aplicação dos Resultados

Atualmente empresas trabalham com dinâmicas cooperativas, onde o relacionamento interpessoal harmônico é de grande interesse das instituições. Práticas para agrupamentos harmônicos são utilizadas por departamentos de recursos humanos para maximizar o desempenho de equipes de trabalho em instituições, aumentando consequentemente a produtividade destas empresas [Moscovici, 1996; Sparrowe et al., 2001]. A dinâmica social existente em grupo de pessoas pode ser encontrada e modelada através da metodologia proposta por Moreno [1953], denominado teste sociométrico. Este método consiste em um questionário aplicado a cada membro de um grupo de pessoas. A partir do questionário é construído um sociograma, que é basicamente o mapeamento da rede social do grupo. Mesmo a realização do teste sociométrico para formação de equipes sendo eficiente e eficaz, este método pode ser, em algumas circunstâncias, invasiva, e até mesmo desconfortável. Neste contexto, o uso de técnicas de análise e predição dos relacionamentos de trabalho podem fornecer evidências sobre o comportamento deste grupo. Assim, evita-se o emprego do teste sociométrico, e pode-se fornecer, com maior confiança, informações das relações de trabalho dentro de um grupo, melhorando assim o desempenho na execução de tarefas colaborativas.

Em outro contexto para aplicação dos resultados, considere que o Facebook é uma modalidade de comunicação remota, pois os usuários são separados no espaço e no tempo. Ainda assim, os resultados sugerem que o site de rede social se assemelha à comunicação próxima entre os alunos inseridos em sala de aula, e que está em consonância com estudos recentes sobre a relação entre interações *online* e *offline*. Estes resultados também são de importância prática. Por exemplo, sites educacionais como *Coursera*[1], em parceria com as melhores universidades, oferecem cursos gratuitos *online*. Dado o grande número de pessoas no mundo que tais sites servem, uma forma eficaz para unir os estudante em escala é usar as mesmas características estudadas aqui. Além disso, a metodologia proposta pode servir para análises em diversos contexto, onde é desejado estabelecer um mapeamento entre ambientes *online* e *offline*.

[1] `https://www.coursera.org/about`

7.3 Trabalhos Futuros

A seguir, serão listadas alguns direcionamentos para realização de trabalhos futuros.

- Repetir estudos semelhantes em classes de diferentes países, para explorar os efeitos interculturais.
- Realizar as análises propostas em ambientes diferente da sala de aula, como organizações e empresas.
- Construir uma aplicação para recomendação de formação de equipes através de contas dos usuários no Facebook. Mesmo a realização do teste sociométrico para formação de equipes sendo eficiente e eficaz, este método pode ser, em algumas circunstâncias, invasivo, e até mesmo desconfortável. Assim, o desenvolvimento de uma aplicação para o Facebook tornaria esta análise mais discreta e menos constrangedora.

Referências Bibliográficas

Agustín-Blas, L. E.; Salcedo-Sanz, S.; Ortiz-García, E. G.; Portilla-Figueras, A.; Pérez-Bellido, Á. M. & Jiménez-Fernández, S. (2011). Team formation based on group technology: A hybrid grouping genetic algorithm approach. *Computers & Operations Research*, 38(2):484--495.

Anchuri, P. & Magdon-Ismail, M. (2012). Communities and balance in signed networks: A spectral approach. Em *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pp. 235--242. IEEE.

Antiqueira, L. (2011). *Relações da estrutura de redes complexas com as dinâmicas do passeio aleatório, de transporte e de sincronização*. Tese de doutorado, Universidade de São Paulo.

Ashton, M. C.; Lee, K. & Paunonen, S. V. (2002). What is the central feature of extraversion? social attention versus reward sensitivity. *Journal of Personality and Social Psychology*, 83(1).

Baeza-Yates, R.; Ribeiro-Neto, B. et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Benevenuto, F.; Magno, G.; Rodrigues, T. & Almeida, V. (2010). Detecting spammers on twitter. Em *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, p. 12.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121--167.

Burt, R. S. (2000). The network structure of social capital. *Research in organizational behavior*, 22:345--423.

Bustos, D. M. (1979). *The Sociometric Testing: fundamentals, techniques and applications*. Brasilense Publisher.

Cartwright, D. & Harary, F. (1956). Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277.

Chang, C.-C. & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O. & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*.

Chawla, N. V.; Japkowicz, N. & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1--6.

Chen, S.-J. & Lin, L. (2004). Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. *Engineering Management, IEEE Transactions on*, 51(2):111--124.

Coleman, J. S. et al. (1989). *Social capital in the creation of human capital*. University of Chicago Press.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273--297.

Costa, H.; Benevenuto, F. & Merschmann, L. H. (2013). Detecting tip spam in location-based social networks. Em *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pp. 724--729. ACM.

Costa, L. d. F.; Rodrigues, F. A.; Travieso, G. & Villas Boas, P. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167--242.

Daniel, J. (2012). Making sense of moocs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*.

Dorogovtsev, S. N. & Mendes, J. F. (2004). The shortest path to complex networks. *arXiv preprint cond-mat/0404593*.

Drummond, C.; Holte, R. C. et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Em *Workshop on Learning from Imbalanced Datasets II*, volume 11. Citeseer.

Easley, D. & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a highly connected world*, volume 8. Cambridge Univ Press.

Ellison, N. B.; Steinfield, C. & Lampe, C. (2007). The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143--1168.

Eysenck, H. (1970). *Readings in Extraversion-introversion: Theoretical and methodological issues*. Readings in Extraversion-introversion. Staples Press.

Facchetti, G.; Iacono, G. & Altafini, C. (2011). Computing global structural balance in large-scale signed social networks. *Proceedings of the National Academy of Sciences*, 108(52):20953--20958.

Fahlman, S. E. & Hinton, G. E. (1987). Connectionist architectures for artificial intelligence. *Computer;(United States)*, 20(1).

Fieller, E.; Hartley, H. & Pearson, E. (1957). Tests for rank correlation coefficients. i. *Biometrika*, 44(3/4):470--481.

Fitzpatrick, E. L. & Askin, R. G. (2005). Forming effective worker teams with multi-functional skill requirements. *Computers & Industrial Engineering*, 48(3):593--608.

Garton, L.; Haythornthwaite, C. & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1):0--0.

Geisser, S. (1993). *Predictive inference*, volume 55. CRC Press.

Gilbert, E. & Karahalios, K. (2009). Predicting tie strength with social media. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 211--220. ACM.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, pp. 1360--1380.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10--18.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147--160.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107--112.

Jain, A. K.; Duin, R. P. W. & Mao, J. (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4--37.

Jones, J. J.; Settle, J. E.; Bond, R. M.; Fariss, C. J.; Marlow, C. & Fowler, J. H. (2013). Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Em *IJCAI*, volume 14, pp. 1137--1145.

Korb, K. B. & Nicholson, A. E. (2003). *Bayesian artificial intelligence*. cRc Press.

Kubat, M. & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. Em *ICML*, volume 97, pp. 179--186.

Kunegis, J.; Lommatzsch, A. & Bauckhage, C. (2009). The slashdot zoo: mining a social network with negative edges. Em *Proceedings of the 18th international conference on World wide web*, pp. 741--750. ACM.

Kunegis, J.; Preusse, J. & Schwagereit, F. (2013). What is the added value of negative links in online social networks? Em *Proceedings of the 22nd international conference on World Wide Web*, pp. 727--736. International World Wide Web Conferences Steering Committee.

Leskovec, J.; Huttenlocher, D. & Kleinberg, J. (2010a). Predicting positive and negative links in online social networks. Em *Proceedings of the 19th international conference on World wide web*, pp. 641--650. ACM.

Leskovec, J.; Huttenlocher, D. & Kleinberg, J. (2010b). Signed networks in social media. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1361--1370. ACM.

Liu, A. Y.-c. (2004). *The effect of oversampling and undersampling on classifying imbalanced text datasets*. Tese de doutorado, Citeseer.

Liu, X.-Y.; Wu, J. & Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(2):539--550.

Maniu, S.; Cautis, B. & Abdessalem, T. (2011). Building a signed network from interactions in wikipedia. Em *Databases and Social Networks*, pp. 19--24. ACM.

Mansson, D. H. & Myers, S. A. (2011). An initial examination of college students' expressions of affection through facebook. *Southern Communication Journal*, 76(2):155--168.

Massa, P. & Avesani, P. (2005). Controversial users demand local trust metrics: An experimental study on epinions. com community. Em *Proceedings of the National Conference on artificial Intelligence*, volume 20, p. 121. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

McPherson, M.; Smith-Lovin, L. & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pp. 415--444.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P. & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. Em *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29--42. ACM.

Mobius, M. M.; Quoc-Anh, D. & Rosenblat, T. S. (2004). Social capital in social networks. *Retrieved March*, 3:2009.

Moreno, J. L. (1953). *Who shall survive?: A new approach to the problem of human interrelations*. Beacon House Inc.

Moscovici, F. (1996). *Desenvolvimento interpessoal: treinamento em grupo*. José Olympio.

Newman, M. E. (2002). Assortative mixing in networks. *Physical review letters*, 89(20):208701.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167--256.

Oh, H.; Labianca, G. & Chung, M.-H. (2006). A multilevel model of group social capital. *Academy of Management Review*, 31(3):569--582.

Page, L.; Brin, S.; Motwani, R. & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

Pal, S. K. & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *Neural Networks, IEEE Transactions on*, 3(5):683--697.

Portes, A. (2000). Social capital: Its origins and applications in modern sociology. *LESSER, Eric L. Knowledge and Social Capital. Boston: Butterworth-Heinemann*, pp. 43--67.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Relatório técnico, DTIC Document.

Rumelhart, D. E.; Hinton, G. E. & Williams, R. J. (1985). Learning internal representations by error propagation. Relatório técnico, DTIC Document.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581--603.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2).

Sparrowe, R. T.; Liden, R. C.; Wayne, S. J. & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. *Academy of management journal*, 44(2):316--325.

Symeonidis, P.; Tiakas, E. & Manolopoulos, Y. (2010). Transitive node similarity for link prediction in social networks with positive and negative links. Em *Proceedings of the fourth ACM conference on Recommender systems*. ACM.

Townsend, J. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 9(1):40--50.

Waldrop, M. M. & Magazine, N. (2014). Massive open online courses, aka moocs, transform higher education and science.

Wi, H.; Oh, S.; Mun, J. & Jung, M. (2009). A team formation model based on knowledge and collaboration. *Expert Systems with Applications*, 36(5):9121--9134.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xiang, R.; Neville, J. & Rogati, M. (2010). Modeling relationship strength in online social networks. Em *Proceedings of the 19th international conference on World wide web*, pp. 981--990. ACM.

Xie, W.; Li, C.; Zhu, F.; Lim, E.-P. & Gong, X. (2012). When a friend in twitter is a friend in life. Em *Proceedings of the 3rd Annual ACM Web Science Conference*, pp. 344--347. ACM.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69--90.

Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Em *ICML*, volume 97, pp. 412--420.

Zhao, X.; Yuan, J.; Li, G.; Chen, X. & Li, Z. (2012). Relationship strength estimation for online social networks with the study on facebook. *Neurocomputing*, 95:89--97.

Apêndice A

Teste Sociométrico

Assinale com um X a resposta para a seguinte pergunta: *Você gostaria de trabalhar em equipe com esta pessoa ?*

NOME DO VOLUNTÁRIO	SIM	NÃO	INDIFERENTE
$estudante_1$			
$estudante_2$			
$estudante_3$			
$estudante_4$			
$estudante_5$			
$estudante_6$			
$estudante_7$			
(...)			
$estudante_n$			

Apêndice B

Termo de Consentimento

1. NOME DA PESQUISA: Um arcabouço para a predição de afinidades e formação de grupos de trabalhos a partir de redes sociais

2. PESQUISADOR RESPONSÁVEL: Douglas Donizeti de Castilho Braz

3. INSTITUIÇÃO PROMOTORA: *Anônima*

4. OBJETIVO, JUSTIFICATIVA E PROCEDIMENTO DE COLETA DE DADOS: O objetivo deste experimento é coletar dados sobre afinidades de trabalho dos participantes, além de dados das suas interações online através do Facebook. Mediante assinatura de termos de sigilo e confidencialidade, o software por esta pesquisa desenvolvido poderá ser auditado em qualquer fase do seu desenvolvimento. A participação é voluntária. A coleta de dados se dará pela resposta de um questionário e acesso a um aplicativo disponibilizado no Facebook. Esta participação não implicará em qualquer risco, despesa ou desconforto. O sigilo da identidade dos participantes será assegurado. Dados de terceiros disponibilizados pelo Facebook não serão coletados, pois o software será programado para a filtragem desses dados, de modo que somente serão coletados dados dos sujeitos que tiverem aceitado participar da pesquisa. A coleta de dados digitais obedecerá a Política de Privacidade prevista no Facebook.

5. BENEFÍCIOS E DIVULGAÇÃO DOS RESULTADOS: Não é previsto nenhum benefício financeiro ou de outra natureza aos sujeitos da pesquisa. Os autores reservam-se o direito de publicar e apresentar os dados em meios de divulgação científica como meio de gerar informações importantes para o desenvolvimento da área de conhecimento.

Apêndice C

Declaração de Aceite

DECLARO para fins de participação em pesquisa, na condição de sujeito da mesma, que fui devidamente esclarecido do projeto intitulado: **Um arcabouço para a predição de afinidades e formação de grupos de trabalhos a partir de redes sociais**, sob a responsabilidade do pesquisador *Douglas Donizeti de Castilho Braz*, quanto aos seguintes aspectos:

- Justificativa, objetivos e procedimentos que serão utilizados na pesquisa;
- Desconfortos e riscos possíveis;
- Liberdade de me recusar a participar da pesquisa, sem penalização alguma e sem prejuízo;
- Garantia de sigilo quanto aos dados confidenciais envolvidos na pesquisa, assegurando-me absoluta privacidade;
- Esclarecido que, talvez, não terei benefício direto com a pesquisa, porém poderei contribuir com outras pessoas, pois se espera que este estudo traga informações importantes e que o conhecimento produzido possa ser divulgado em eventos e revistas científicas;

Após convenientemente esclarecido, concordo, voluntariamente, em participar desta pesquisa.

NOME: _____

RG: _____

ASSINATURA: _____